The Annals of Applied Statistics
2026, Vol. 20, No. 2, 1736–1765
https://doi.org/10.1214/26-AOAS2183

A STRUCTURED ESTIMATOR FOR LARGE COVARIANCE MATRICES IN THE PRESENCE OF PAIRWISE AND SPATIAL COVARIATES

BY MARTIN METODIEV[1,a], MARIE PERROT-DOCKÈS[2,c], SARAH OUADAH[3,d], BAILEY K. FOSDICK[4,f], STÉPHANE ROBIN[3,e], PIERRE LATOUCHE[1,5,b] AND ADRIAN E. RAFTERY[6,g]

[1]*Laboratoire de Mathématiques Blaise Pascal, Université Clermont Auvergne, CNRS,* [a]*martin.metodiev@doctorant.uca.fr,* [b]*Pierre.LATOUCHE@uca.fr*

[2]*Université Paris Cité, CNRS, MAP5,* [c]*marie.perrot-dockees@u-paris.fr*

[3]*Sorbonne Université, Université Paris Cité, CNRS, Laboratoire de Probabilités, Statistique et Modélisation (LPSM),* [d]*sarah.ouadah@sorbonne-universite.fr,* [e]*stephane.robin@sorbonne-universite.fr*

[4]*Department of Biostatistics & Informatics, Colorado School of Public Health,* [f]*bailey.fosdick@cuanschutz.edu*

[5]*Institut Universitaire de France (IUF)*

[6]*Department of Statistics, University of Washington,* [g]*raftery@uw.edu*

We consider the problem of estimating a high-dimensional covariance matrix from a small number of observations when covariates on pairs of variables are available and the variables can have spatial structure. This is motivated by the problem arising in demography of estimating the covariance matrix of the total fertility rate (TFR) of 195 different countries when only 11 observations are available. We construct an estimator for high-dimensional covariance matrices by exploiting information about pairwise covariates, such as whether pairs of variables belong to the same cluster, or spatial structure of the variables, and interactions between the covariates. We reformulate the problem in terms of a mixed effects model. This requires the estimation of only a small number of parameters, which are easy to interpret and which can be selected using standard procedures. The estimator is consistent under general conditions, and asymptotically normal. It works if the mean and variance structure of the data is already specified or if some of the data are missing. Using simulations, we assess its performance under our model assumptions as well as under model misspecification. We find that it outperforms several popular alternatives. We apply it to the TFR dataset and draw some conclusions.

1. Introduction. We consider the problem of estimating a large covariance matrix from a small number of data points when covariates on pairs of variables and their spatial structure are available. This is motivated by the problem, arising in demographic research, of estimating the covariance matrix of the total fertility rate for a large number of countries from data at a small number of time points.

Our specific goal is to estimate the covariance matrix of a model for the total fertility rate (TFR) used by the United Nations (U.N.) for 195 different countries. The dataset is denoted Y and is made up of measurements of the TFR at $T = 11$ time points (observations) for $d = 195$ countries (variables). Each time point is a five-year period. The large dimension of the covariance matrix makes the use of a standard estimator, such as the sample covariance matrix, untenable, and necessitates the use of additional information. Fortunately, for the TFR dataset, we have been able to obtain information about pairwise, time-invariant, covariate structures. In particular, these include a spatial structure (e.g., countries being contiguous to each other) and structures determined by cluster membership (e.g., countries belonging to the

Received July 2025; revised April 2026.
Key words and phrases. Large covariance matrix estimation, pairwise covariates, spatial effects.

same region, countries having the same common colonizer). The framework we propose also allows us to include interactions when modeling a covariance matrix.

The problem arises in the context of the production of the United Nations' *World Population Prospects* (WPP) (United Nations (2010, 2024)), which issues the U.N.'s official estimates and projections of population for all the countries of the world. These are widely used by governments at all levels, in the private sector, and by researchers, especially in the health and social sciences. They are also used to inform policymaking by international organizations about issues, such as food security, and to monitor progress toward targets, such as the U.N.'s Sustainable Development Goals for 2030. The WPP provides estimates of past and current population by age and sex from 1950 to the date of publication and projections for the period from then to the year 2100. It also contains estimates and projections of the components of population change, namely, fertility, mortality, and international net migration. The WPP also includes estimates and projections for about 300 aggregates of different countries by geography (e.g., the 22 U.N. regions, such as Western Europe or Eastern Africa), development level (e.g., Lower and Middle-Income Countries), or other criteria. The WPP is revised every two or three years in light of the most recent data and to use the most current methodology.

Until 2006, the U.N. used a deterministic method for projecting future population that combined the cohort-component method of population projection with subjective expert opinion about future fertility, mortality, and migration (Preston, Heuveline and Guillot (2001)). This is similar to methods used by most other agencies doing population projections at least since the 1930s. In this method, uncertainty was not expressed using standard statistical methods, such as confidence intervals or standard errors, but rather scenarios in which, for example, fertility was increased or decreased relative to the main, expert-based projection by user-specified amounts. This was widely criticized as having no probabilistic or statistical interpretation and generating poorly calibrated uncertainty statements, which are implausible over multiple projection periods or countries. As a result, starting in 2006, the U.N. engaged in a collaborative research effort to develop better, statistically-based methods for probabilistic population projections, involving statisticians at the University of Washington. These methods were thoroughly evaluated and adopted for the first time for the U.N.'s official projections in 2015.

A critical part of this new method is a method for probabilistic forecasting of the total fertility rate (TFR). This rate is specific to a given country and time period and is defined as the expected number of children a woman would bear if she survived the reproductive interval (usually defined as up to age 49) and at each age experienced the age-specific fertility rates for that age and period. This method, proposed by Alkema et al. (2011a), is based on the observation that the evolution of TFR over time everywhere has been divided into three phases: phase I, usually in preindustrial societies, generally consisting of fluctuations around a high level in the range of four to seven children per woman, phase II, usually following the start of industrialization in which fertility declines until it reaches some point below the replacement rate of 2.1, and phase III in which TFR again fluctuates but remains low.

Phase II is usually called the fertility transition and is the most critical phase for forecasting future population. It is modeled by a Bayesian hierarchical model, defined for each country as a random walk with a state-dependent drift. The drift, or expected fertility decline, is assumed to follow a double-logistic function of the current fertility level. This typically yields a fertility decline that starts slowly, accelerates, slows, and then stops at some level below the replacement level. Alkema et al. (2011a) evaluated the model thoroughly and found it to give both good point forecasts and well-calibrated interval forecasts.

TFR tends to be highly correlated across countries because it evolves similarly in similar countries. However, once one takes account of the country-specific trend through the double-logistic fertility decline model, the correlation is much lower, but still positive. This matters

little for probabilistic forecasting for individual countries. However, it can matter a great deal for aggregates of countries. Typically, between-country TFR correlations are positive, and so if they are ignored in forming probabilistic forecasts, they will tend to underestimate the uncertainty associated with such aggregate projections. Thus estimating the between-country correlations is critical for such purposes.

This is made challenging by the fact that the number of countries is typically much larger than the number of time points available for estimation. For example in the present work, which is motivated by the 2010 WPP revision, there are $d = 195$ countries, but only $T = 11$ time points, each representing a different five-year time period. Thus, any individual correlation estimate based on these 11 time points will tend to be very prone to random sampling error; for example, the standard error of a standard Pearson correlation estimate from T time points is close to $\frac{1}{\sqrt{T}}$ when the true correlation is close to zero. Thus, an empirical estimate of the correlation matrix will tend to be very noisy when there are only 11 time points.

Fosdick and Raftery (2014) developed a method to overcome this problem for countries and time points in the critical phase II period. This is a two-stage method in which first the one-step ahead residuals are standardized by their standard errors. Then the individual correlations are modeled as linear combinations of country pair-specific covariates. Many such covariates were considered, and three were found useful: whether the two countries are neighbors, whether they are both in the same U.N. region, and whether they shared a colonizer in 1945. The resulting model fits the correlations well and is heavily regularized, but it has a drawback, namely, that it does not yield a positive semidefinite matrix; whereas in fact, correlation matrices are positive semidefinite. Thus, for example, maximum likelihood estimation is not possible, because the likelihood is not well defined.

Fosdick and Raftery (2014) got around this by using maximum pairwise likelihood estimation (Cox and Reid (2004), Varin, Reid and Firth (2011)). This still did not yield a positive semidefinite result. Fosdick and Raftery (2014) then used the positive semidefinite matrix closest in Frobenius norm to the resulting pairwise likelihood estimate, obtained by performing an eigenvalue decomposition of the pairwise likelihood estimate, setting any negative eigenvalues to a value close to zero, and then reconstructing the matrix. The resulting estimate validated well. Crucially, it also yielded much improved estimates of uncertainty for aggregate regions, with predictive variances that were up to three times larger than those that resulted from assuming independence.

While the method of Fosdick and Raftery (2014) generally performed well, it has some ad hoc aspects, and we wished to develop a more statistically principled method. That is our goal in the present manuscript. Here we propose a more standard statistical approach to correlation estimation by reexpressing it in terms of a random effects model for which we can carry out standard statistical estimation. This yields a positive semidefinite estimator of the correlation matrix, without the need for pairwise likelihoods or post hoc adjustments to ensure positive semidefiniteness. While the random effects model that we use is standard, its use for estimating correlation matrices with far more variables than observations using variable-pair-specific covariates is new, as far as we know.

The manuscript is organized as follows. In Section 2 we give a brief review of the relevant literature on covariance matrix estimation. In Section 3 we give an overview of the dataset and give the motivation for our model assumptions, followed by the general setting for the estimator we propose. In Section 4 we define our estimator and derive its properties. We propose an estimation algorithm (Section 4.3) and show how to adjust our estimator in the case that the model assumptions do not hold (Section 4.4). In Section 5 we assess its performance under a variety of different settings by simulation and compare it to some popular alternative methods. Then we use these techniques to derive estimates of the covariance matrix of the TFR dataset, whose correlations we compare in Section 6. We close with a discussion of our

work. The code for this paper for scientific dissemination is available in the Supplementary Material and also on Github https://github.com/m-metodiev/SCE/tree/main. The estimators used in this article are made available in the R-package scov (Metodiev, Perrot-Dockès and Robin (2025)).

2. Covariance matrix estimation. As we shall see, the performance of standard covariance matrix estimators is not excellent in the setting we consider, because they cannot handle pairwise and spatial covariates. These include shrinkage estimators, such as the Ledoit–Wolf estimator (Ledoit and Wolf (2004)), which shrink covariance matrices toward sparse matrices or the identity matrix (for a review, see Pourahmadi (2013), Ledoit and Wolf (2022)) as well as estimators, such as the graphical lasso (Friedman, Hastie and Tibshirani (2008)) which assumes sparsity of the inverse of the covariance matrix.

The most commonly used approaches, which assume dependency structures within the model, are approaches that use factor models (for an overview of these approaches, see Fan, Liao and Liu (2016)). There are indeed estimation procedures that use factor models and which can also allow for the encoding of spatial information, as in Christensen and Amemiya (2003), Wang and Wall (2003), Lopes, Salazar and Gamerman (2008), Lopes, Gamerman and Salazar (2011), Thorson et al. (2015). Clustered structures, on the other hand, are estimated via multilevel factor models (Longford and Muthén (1992)). While we will show that the correlation structures of the different cluster effects present in the data can each be interpreted in the framework of multilevel factor models, none of the models available in the literature can combine several multilevel factor models with a spatial structure. Moreover, the number of samples we are facing is too small to use techniques such as parallel factor analysis (Harshman and Lundy (1994)).

Interpretability. Interpretability of the model parameters has been stressed to be an important feature of covariance matrix parametrization by Pourahmadi (1999, 2011), but while that work does provide solutions, they do not incorporate pair-specific covariates into the covariance matrix estimation procedure. Approaches that do include pairwise covariates into the model have involved evaluating linear combinations of known matrices on the scale of the covariance matrix (Anderson (1973)). Sums of covariance matrices also appear in linear mixed-effect models when adding crossed effects (Gałecki and Burzykowski (2013)).

However, sums of covariance matrices are, in practice, hard to interpret. To illustrate this problem, suppose that we set the covariance matrix of the data to the sum of the following two covariance matrices:

$$\Sigma := \begin{pmatrix} 18 & 0 \\ 0 & 18 \end{pmatrix} + \begin{pmatrix} 2 & 0.8 \\ 0.8 & 2 \end{pmatrix}, \tag{1}$$

and that a two-dimensional Gaussian model with covariance matrix Σ is used to model the data. The first matrix assumes independence between the two variables considered, while the second matrix characterizes variables which are clearly correlated. Individually, the correlations of the second matrix tell us nothing about the correlation matrix of the data, since they might be overwhelmed by the entries of the first. Indeed, while the correlations of the two variables associated with the second matrix are relatively high (0.4), the correlation matrix associated with Σ is given by

$$R = \begin{pmatrix} 1 & 0.04 \\ 0.04 & 1 \end{pmatrix}, \tag{2}$$

meaning that the correlation between the two variables is small in the resulting dataset.

Different scales, such as the matrix logarithm (Chiu, Leonard and Tsui (1996)), have been suggested, but they suffer from a similar problem. Bonat and Jørgensen (2016) generalized

these approaches into a framework for nonnormal multivariate data, called multivariate covariance generalized linear models (McGLM). A McGLM was used by Bonat and Jørgensen (2016) to include spatiotemporal covariates in the covariance estimation. It is also possible to use de Freitas et al. (2022) to incorporate information about variables belonging to a similar cluster. However, while the framework of McGLM is the closest to ours, it is not clear how to use it to get interpretable results or how to combine spatial structures and clusters in the same model.

Some Bayesian approaches get interpretable parameters by separating the structure of the variance parameters from the structure of the correlation parameters. This separation strategy has been used successfully in Bayesian inference in a variety of approaches (see, for instance, Barnard, McCulloch and Meng (2000), Lewandowski, Kurowicka and Joe (2009), Qian (2009), Tokuda et al. (2011)), but none of these approaches incorporate pairwise covariates into the covariance matrix estimation. There are also some Bayesian approaches for covariance matrix estimation which use information about different clusters of variables, such as Karolyi (1992, 1993), Aguilar and West (2000), West (2003), Liechty, Liechty and Müller (2004), but none of these combine this information with a spatial structure.

At the core of the covariance matrix estimation strategy we propose is the idea of combining the separation strategy with the approach of Anderson (1973). This allows us to normalize the covariance matrices into correlation matrices such that the parameters of the random effects are interpretable. Going back to equations (1) and (2), we recommend instead separating the variance estimation and modeling R directly as a convex combination of two correlation matrices. In our previous toy example, a possible choice could be

$$R = 0.9\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix} + 0.1\begin{pmatrix}1 & 0.4\\ 0.4 & 1\end{pmatrix}.$$

The importance of each matrix is given by its linear coefficient. Independently of the values of the other matrix, we will know that the correlation between the two variables is such that 10% of the correlation structure is explained by the second correlation matrix.

In the TFR dataset that motivates our work, a variety of effects and their interactions are to be taken into account. We will show that the setting of the TFR dataset allows us to give useful estimates that work well, in the sense that useful properties (identifiability, consistency, asymptotic normality in the number of data points T *and* in the dimension d, available confidence regions) hold under some assumptions. We will also adjust our estimator in such a way that it gives consistent estimates, without having to assume that specific model assumptions hold.

3. Modeling covariance matrices with known pairwise and spatial covariates.

3.1. *The total fertility rate data.* The dataset we analyze consists of the total fertilty rate (TFR) of 195 countries in successive *five*-year periods from 1950 to 2010. One country as well as one time period were removed for reasons on which we will elaborate further in this section and Section 6. The dataset is denoted by Y and is made up of $T = 11$ time points (observations) and $d = 195$ countries (variables). The element $Y_{t,j}$ of Y is the TFR of country j at time t, and $Y_t^\intercal := (Y_{t,1}, \ldots, Y_{t,d}) \in \mathbb{R}^d$ denotes row t of matrix Y. Note that the index t is used to refer to different observations because of the specific nature of the dataset. There is previous work on this dataset on which we build. We will first summarize this work.

The United Nations (U.N.) predicted the total fertility rate (TFR) of different countries via the Bayesian hierarchical model of Alkema et al. (2011a) in the 2010 world population prospects (United Nations (2010)). The TFR was modeled as

$$Y_t = \mu_t + \varepsilon_t^0, \quad \varepsilon_t^0 | Y_{t-1} \sim \mathrm{MVN}_d\big(0_d, \mathrm{diag}(\sigma_t)\ I_d\ \mathrm{diag}(\sigma_t)\big),$$

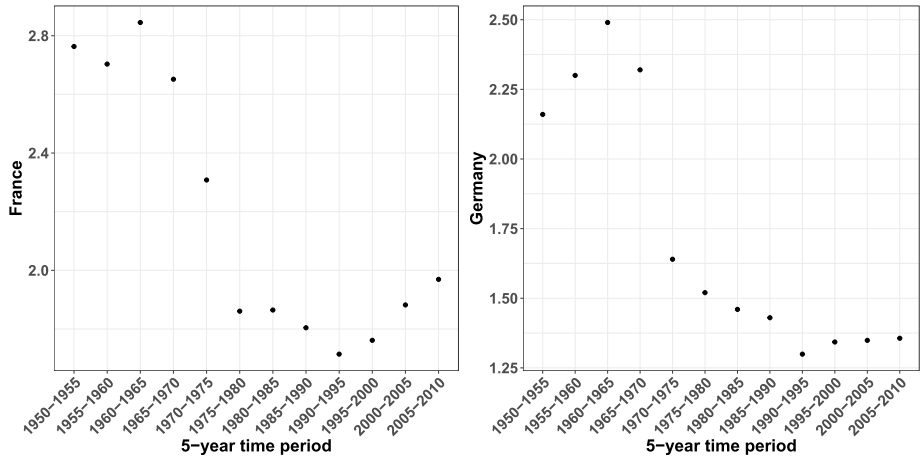

FIG. 1. *TFR values of France and Germany.*

with independent forecast errors ε_t^0, $\mu_t^\mathsf{T} := (\mu_{t,1}, \ldots, \mu_{t,d})$, where $\mu_{t,j}$ denotes the expected TFR (conditional on Y_{t-1}) for country j in period t for $j \in [\![1, d = 195]\!]$ and $\sigma_{t,j}^2$ the variance (conditional on Y_{t-1}), $t \in [\![1, T = 11]\!]$. Note that $\operatorname{diag}(\sigma_t)$ is a diagonal matrix with entries σ_t, 0_d is the null-vector, I_d the identity matrix, and $\mathrm{MVN}_d(m, S)$ denotes the multivariate normal distribution of dimension d with mean m and covariance matrix S.

The parameters μ_t, σ_t are time dependent, going through three stages: pretransition, transition, and post-transition. If country j is in phase II (transition), $\mu_{t,j}$ is on average decreasing in t with $(\mu_{t,j}, \sigma_{t,j})$ following a Bayesian hierarchical model. If it is in phase III (post-transition), $\mu_{t,j}$ fluctuates around a factor below 2.1, the replacement fertility level, with a constant variance $\sigma_{t,j}^2$. Phase I is not included in the model since all countries have already completed the pretransition, eliminating the need for predictions. While this model was build on a solid demographic background, the independence assumption within each observation of TFR values, implied by the fact that the covariance matrices are set to be diagonal, remains questionable. After all, one would not expect two countries that share similar attributes to have independent standardized errors. As an example, consider the TFR values of Germany and France, which are shown in Figure 1. Their high values around 1950 and 1960 have similar explanations, their social and economic recovery after World War II leading to the postwar baby boom, so one would expect their standardized errors to be positively correlated.

Fosdick and Raftery (2014) found that between-country dependencies that are not included in the model of Alkema et al. (2011a) yielded prediction intervals for forecasting regions consisting of multiple countries that were too narrow. To tackle this, Fosdick and Raftery proposed estimating the covariance matrix of the joint distribution of the countries' TFRs. They did this by modeling the covariance matrix using time-invariant pairwise information, namely, whether two countries had the same common colonizer after 1945, whether two countries are contiguous, and whether two countries are in the same U.N. region.[1]

A first model was built as

$$Y_t = \mu_t + \varepsilon_t^0, \quad \varepsilon_t^0 | Y_{t-1} \sim \mathrm{MVN}_d(0_d, \Sigma_t),$$

where Σ_t denotes the covariance matrix (conditional on Y_{t-1}), $t \in [\![1, T = 11]\!]$.

[1]We identify regions by the U.N. subdivision of the sustainable development goal (SDG) regions into 21 geographic subregions. We identify the 10 different colonizers and their colonial relationships after 1945 as well as their neighborhood structure, via the database of the Centre d'Études Prospectives et d'Informations Internationales (CEPII) (Mayer and Zignago (2006)).

Fosdick and Raftery (2014) also decomposed the covariance structure of the model into the standard deviation vector at time t, $\sigma_t \in \mathbb{R}_+^d$, as well as a correlation structure R_t. Conditional on Y_{t-1}, we have that

$$\varepsilon_t^0 | Y_{t-1} \sim \text{MVN}_d(0_d, \text{diag}(\sigma_t) R_t \, \text{diag}(\sigma_t)).$$

With the method of Alkema et al. (2011a) already providing accurate estimates of μ_t and σ_t, Fosdick and Raftery chose to model ε_t, the standardized version of ε_t^0, such that

$$\varepsilon_t | Y_{t-1} \sim \text{MVN}_d(0_d, R_t), \tag{3}$$

and focused on the estimation of R_t. For all countries that are in phase II or III of the model of Alkema et al. (2011a), they set

$$(R_t)_{i,j} = \begin{cases} 1 & \text{if } i = j, \\ \alpha_0^{(1)} + \alpha_1^{(1)} \text{contig}_{i,j} & \\ \quad + \alpha_2^{(1)} \text{comcol}_{i,j} + \alpha_3^{(1)} \text{region}_{i,j} & Y_{t-1,i} < \kappa, Y_{t-1,j} < \kappa, \\ \alpha_0^{(2)} + \alpha_1^{(2)} \text{contig}_{i,j} & \\ \quad + \alpha_2^{(2)} \text{comcol}_{i,j} + \alpha_3^{(2)} \text{region}_{i,j} & \text{otherwise}, \end{cases} \tag{4}$$

where κ is a threshold parameter, $\text{contig}_{i,j} = 1$ if countries i and j are contiguous, $\text{comcol}_{i,j} = 1$ if they had a common colonizer after 1945, $\text{region}_{i,j} = 1$ if they are in the same U.N. region, and 0 otherwise.

Fosdick and Raftery (2014) selected these covariates by Bayesian model selection using the BMA package (Raftery et al. (2013a)) on the database of Mayer and Zignago (2006), which includes a variety of pairwise covariates. The authors selected the three covariates mentioned (the "contig," "comcol," and "region" covariate), based on the fact that their posterior inclusion probabilities are above 50%, which is predictively optimal under some assumptions according to Barbieri and Berger (2004). To exhibit the need for the inclusion of these covariates, we visualized the standardized errors of five different countries in Figure 2. Linear regression lines are added for all country pairs. Germany, France, Switzerland, and Luxembourg are all spatially close to each other and in the same U.N. region (Western Europe), and their regression lines do indeed point in the same direction. On the other hand, the Republic of Korea shares none of the pairwise covariates listed with these other countries, and its relationship with them is clearly weaker.

The problem with this approach is that the value of R_t that it yields is not necessarily positive definite for all values of $\alpha_0^{(k)}, \ldots, \alpha_3^{(k)}$. To address this, Fosdick and Raftery (2014) mapped their estimate of R_t onto the positive definite correlation matrix that was closest to it with regards to the Frobenius norm. This is quite simply done computationally, by taking the eigenvalue decomposition of the estimated matrix, setting negative eigenvalues to a small positive value, and reconstituting the matrix. However, this method comes with no statistical guarantees. We seek instead a statistically principled approach to estimating the covariance matrix.

Here we propose a new approach that models the correlation matrix of the TFR directly, as a time-independent correlation matrix R, which is a linear combination of several correlation matrices. These correlation matrices are in turn entirely determined by known covariates, independent of the thresholding level κ, and in particular of the time point t, which is why we do not use the index t for R in our model. Therefore, R will be positive definite if at least one of these correlation matrices is positive definite.

For each time point t, we follow Fosdick and Raftery (2014) in only modeling the TFR Y_t of the countries that are in phase II or III of the model of Alkema et al. (2011a). This occurs only after the first time period, which is why one time period was removed to obtain $T = 11$.

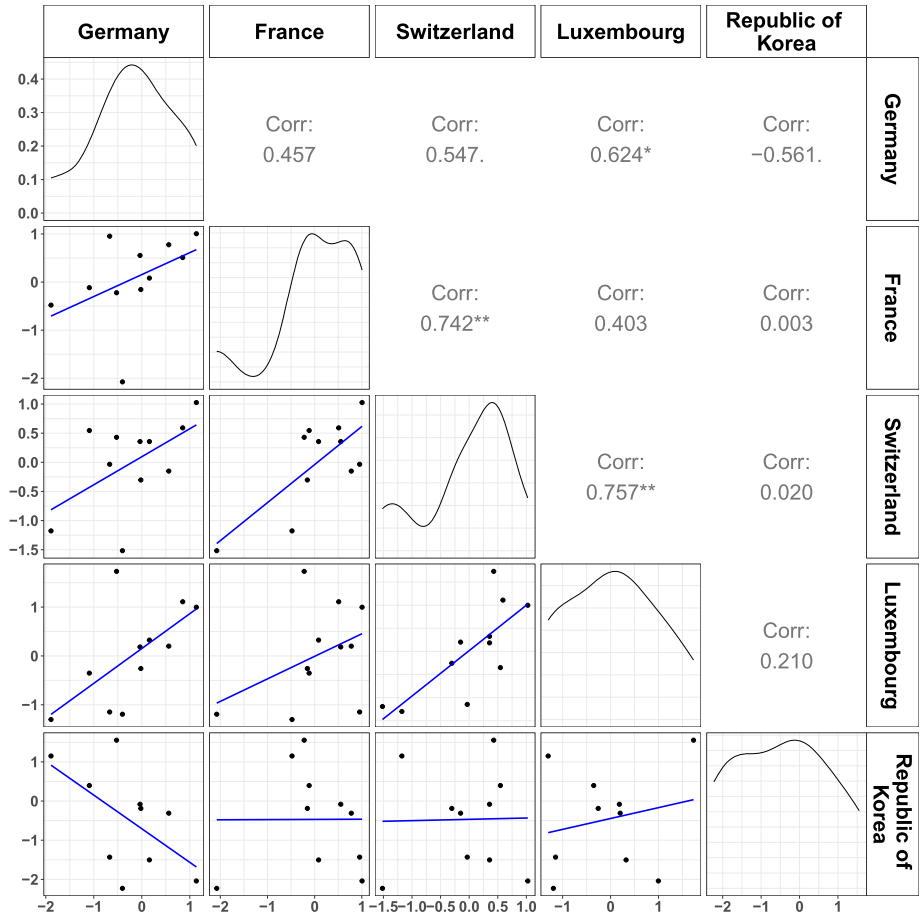

FIG. 2. *Scatterplot matrix of the standardized errors $\varepsilon_1, \ldots, \varepsilon_T$ for five different countries; a regression line is added to each panel. Density plots are shown on the diagonal, and correlation estimates are displayed on the upper right panels.*

Our methodology relies on the decomposition of the stanadardized errors ε_t as a weighted sum of standardized, independent effects. Thus, for every time point $1 \le t \le T$, we consider

$$Y_t = \mu_t + \operatorname{diag}(\sigma_t)\varepsilon_t, \quad \varepsilon_t \sim \mathrm{MVN}_d(0_d, R).$$

We model the residuals as follows:

$$\begin{aligned}
&\varepsilon_t = A_t + B_t + C_t + D_t + E_t, \quad A_t, B_t, C_t, D_t, E_t \in \mathbb{R}^d \text{ where} \\
&A_t \overset{\text{i.i.d}}{\sim} \mathrm{MVN}_d(0_d, \alpha_A F_A), \qquad B_t \overset{\text{i.i.d}}{\sim} \mathrm{MVN}_d(0_d, \alpha_B F_B), \\
&C_t \overset{\text{i.i.d}}{\sim} \mathrm{MVN}_d(0_d, \alpha_C F_C), \qquad D_t \overset{\text{i.i.d}}{\sim} \mathrm{MVN}_d\big(0_d, \delta_D \Gamma(\beta_D)^{-1}\big), \\
&E_t \overset{\text{i.i.d}}{\sim} \mathrm{MVN}_d(0_d, \alpha_E I_d).
\end{aligned} \tag{5}$$

Note that the covariance matrix of the standardized errors is equal to the correlation matrix of the data, conditional on the values at the previous time points, since by construction each $\varepsilon_{t,j}$ has variance 1.

The random vectors, $A_{t,j}$, $B_{t,j}$, $C_{t,j}$, $D_{t,j}$ denote the random effects on country j at time t, corresponding to the effect of having a common colonizer, belonging to the same region, the global effect, the contiguity effect, respectively. The random vector E_t denotes

i.i.d. Gaussian noise for all countries at time t (its correlation matrix is the identity matrix of dimension d, I_d). Its presence ensures that the correlation matrix of ε_t is positive definite, since at least one of the correlation matrices of the standardized effects is positive definite. The random effects are standardized, in the sense that the matrices F_A, F_B, F_C, $\Gamma(\beta_D)^{-1}$ are correlation matrices, that is, positive semidefinite matrices with diagonal entries equal to 1. They are also weighted by positive constants $\alpha_A, \ldots, \alpha_E, \delta_D$, which must sum to one. This ensures that the impact of each effect is measurable via its linear coefficient: for every pair of countries (i, j), its correlation is an average of the individual entries $(F_A)_{i,j}$, $(F_B)_{i,j}$, $(F_C)_{i,j}$, $\Gamma(\beta_D)^{-1}_{i,j}$ weighted by their respective weights.

We use two separate approaches for modeling the correlation matrices of the random effects. Matrices with the capital letter F naturally impose a block structure on the model, in the sense that they partition the countries based on which cluster (i.e., region, colonizer) they belong to. Let $f_A \in \{0,1\}^{d\times 10}$, $f_B \in \{0,1\}^{d\times 21}$ denote the covariates corresponding to the regional and common colonizer clusters, meaning that $(f_B)_{j,r} = 1$ if country j belongs to region r and 0 otherwise (analogously for f_A). We can easily transform these covariates into correlation matrices by setting $F_A = f_A f_A^\top$, $F_B = f_B f_B^\top$, meaning that $(F_A)_{i,j}$, $(F_B)_{i,j}$ are equal to 1 if countries i and j have the same common colonizer or belong to the same U.N. region, respectively. The global effect applies equally to every country pair such that $F_C = 1_d 1_d^\top$, where $(1_d)_j = 1$ for all j, meaning that $(F_C)_{i,j}$ is always 1. Note that the way these matrices are modeled corresponds to the way in which correlation matrices are modeled in multilevel factor models, only that in our case the factor loadings and group means are known. They can also be represented via the design matrix of a linear mixed-effects model. We elaborate on these points in Appendix D (Metodiev et al. (2026)).

The contiguity effect was modeled using a conditional autoregressive (CAR) model. There are many different CAR models, each of which is usually specified by a small number of parameters (Wall (2004), Kyung and Ghosh (2010), MacNab (2011), Tastu, Pinson and Madsen (2013), Ver Hoef, Hanks and Hooten (2018)). We parametrize the CAR model via only one autocorrelation parameter, β_D, in a Gaussian Markov random field (GMRF). We follow Besag and Kooperberg (1995) and Besag, York and Mollié (1991) in setting

$$\Gamma(\beta_D) = Q_{\beta_D} M_2 (I_d - \beta_D M_1) Q_{\beta_D},$$

$$(M_2)_{i,j} = \begin{cases} \sum_{e=1}^{d} M_{i,e} & i = j, \\ 0 & i \neq j, \end{cases} \qquad (M_1)_{i,j} = \frac{M_{i,j}}{(M_2)_{i,j}}, \tag{6}$$

with the parameter $\beta_D \in (0,1)$, where M is the adjacency matrix induced by the neighborhood structure of the underlying geography. The expression for Q_{β_D} in equation (6) is different from the most commonly used form. This is because here we are modeling the correlation matrix rather than the covariance matrix. The matrix Q_{β_D} is a nonnegative diagonal matrix chosen such that all diagonal entries of $\Gamma(\beta_D)^{-1}$ are equal to 1. In the model we propose, a simple analytic expression of Q_{β_D} exists, as illustrated in Appendix A (Metodiev et al. (2026)).

It can happen that components of the CAR model are unconnected and that there are countries that have no neighbors at all (e.g., countries that consist of islands). The latter case is of particular importance to us since the standard CAR model is not defined for unconnected nodes. However, the literature on how to deal with this case is sparse. We follow Freni-Sterrantino, Ventrucci and Rue (2018) in assuming that the correlation of the spatial effect is $\Gamma(\beta_D)^{-1}_{j,j} = 1$ for every country j and $\Gamma(\beta_D)^{-1}_{i,j} = 0$ if country i or country j is set on an island.

Note that the assumptions made in equation (5) are general: we assume only that there are some independent, standardized effects through which we can express the impact of our covariates on the correlation matrix of the data.

3.2. *Correlation structure*. We now develop a more general framework for correlation estimation. Suppose that there are K known correlation matrices $F_1, \dots, F_K$ defined with no parameters (these may correspond, e.g., to correlation matrices derived from clusters, such as regions and colonizers), one known positive definite correlation matrix F_0 (this will usually correspond to i.i.d. Gaussian noise) and one known correlation matrix defined via a number of parameters $\beta_1, \dots, \beta_G$ (this will usually correspond to the correlation matrix of a spatial effect). Let $Y_1, \dots, Y_T$ be a Markov process of Gaussian vectors with mean vectors $\mu_t = \mathrm{E}[Y_t | Y_{t-1}]$, variance vectors $\sigma_t^2 = \mathrm{Var}[Y_t | Y_{t-1}]$, and a correlation matrix $R = \mathrm{Cor}(Y_t | Y_{t-1})$, which does not depend on the time point t. We model the correlation matrix R as follows:

$$R(\alpha, \beta, \delta) = \Phi(\alpha) + \delta \cdot \Gamma(\beta)^{-1}, \quad \text{where } \Phi(\alpha) = \sum_{k=0}^{K} \alpha_k F_k, \tag{7}$$

with the constraints

$$\alpha_k, \delta > 0, \quad \delta + \sum_{k=0}^{K} \alpha_k = 1. \tag{8}$$

Since the parameters must sum to 1, we restrict our estimation procedure to only estimating $(\alpha, \beta, \delta) := (\alpha_1, \dots, \alpha_K, \beta_1, \dots, \beta_G, \delta)$, because one can easily solve for $\alpha_0 = 1 - \sum_{k=1}^{K} \alpha_k - \delta$.

Here we focus on estimating the correlation matrix R by modeling it as a sum of correlation matrices. One could also estimate the covariance matrix Σ by modeling it as a sum of covariance matrices. The same kind of methodology can still be applied in this case, with some minor changes.

Few assumptions about $\Gamma(\beta)$ are needed, other than assumptions on its differentiability. The order of differentiability required for our algorithm to work well will be discussed in the next section. Ideally, $\Gamma(\beta)^{-1}$ should be twice differentiable with bounded first derivatives and continuous second derivatives.

Model (7) is a generalization of model (5): due to the assumption of independence, we know that the correlation matrix of the standardized errors is a linear combination of the correlation matrices of the random effects. Also, the diagonal entries of a correlation matrix are equal to 1. The restriction (8) ensures that this is the case for the model we propose.

The structure is defined on the scale of the correlation matrix. No assumptions are made on the structure of the variance vectors σ_t^2 or the mean vectors μ_t. They may even depend on the data, but only on the data of their predecessor, Y_{t-1}. Thus the data follow the following distribution:

$$\begin{aligned} (Y_1, \dots, Y_T) &\sim \prod_{t=1}^{T} \mathrm{MVN}_d\big(\mu_t, \Sigma(\sigma_t, \alpha, \beta, \delta)\big), \\ \Sigma(\sigma_t, \alpha, \beta, \delta) &= \mathrm{diag}(\sigma_t) R(\alpha, \beta, \delta) \,\mathrm{diag}(\sigma_t), \end{aligned} \tag{9}$$

where $R(\alpha, \beta, \delta)$ is defined in equation (7).

Note that the description of the data was chosen to be as general as possible, allowing dependencies between the data points, as well as mean- and variance structures that vary with time. This allows us to add a correlation structure to models which are potentially quite complicated, but already have well-known mean- and variance estimators available.

This is the case for our model in which the countries marginally follow a random walk model with nonlinear drift and for which the marginal mean- and variance parameters have already been estimated efficiently in Alkema et al. (2011a). This is also why we are primarily interested in the properties of our estimator on the scale of the correlation matrix.

However, one could also use our model in the classical case, where all data points are i.i.d. and use standard estimators for the mean- and variance parameters, in which case consistency on the scale of the correlation matrix carries over to consistency on the scale of the covariance matrix. We implement an algorithm for this case and test it in Section 5.

4. Inference about the covariance matrix.

4.1. *Maximum likelihood estimation.* The likelihood associated with the proposed model is

$$L(Y;\mu,\sigma,\alpha,\beta,\delta)=\prod_{t=1}^{T}\mathrm{MVN}_d\big(Y_t;\mu_t,\Sigma(\sigma_t,\alpha,\beta,\delta)\big), \tag{10}$$

where $\mu=(\mu_t)_t$, $\sigma=(\sigma_t)_t$. In our context, accurate estimates of μ and σ are already available and denoted by $\hat{\mu}$ and $\hat{\sigma}$, respectively. We denote the maximum likelihood estimator of $\Sigma(\sigma_t,\alpha,\beta,\delta)$ by

$$\hat{\Sigma}_t^{\mathrm{SCE}}:=\Sigma(\hat{\sigma}_t,\hat{\alpha}_{\mathrm{SCE}},\hat{\beta}_{\mathrm{SCE}},\hat{\delta}_{\mathrm{SCE}}),$$

where $(\hat{\alpha}_{\mathrm{SCE}},\hat{\beta}_{\mathrm{SCE}},\hat{\delta}_{\mathrm{SCE}})$ is the output of an optimization algorithm that solves

$$(\hat{\alpha}_{\mathrm{SCE}},\hat{\beta}_{\mathrm{SCE}},\hat{\delta}_{\mathrm{SCE}})=\operatorname*{argmax}_{(\alpha,\beta,\delta)}L(Y;\hat{\mu},\hat{\sigma},\alpha,\beta,\delta).$$

We call this the "structured covariance estimator" (SCE), because it is meant to capture known pairwise dependency structures in the data. Note that if one does not already have accurate estimates of μ and σ, it is also possible to optimize the likelihood in equation (10) jointly over $(\mu,\sigma,\alpha,\beta,\delta)$.

Identifiability. Before computing the SCE, one might check that the model is identifiable. This is relatively easy to do if the dimension of β is small (note that in our case β is one-dimensional). This is because for fixed β, identifiability can be checked by solving a linear program.

THEOREM 4.1. *The model is identifiable if the linear program*

$$\begin{aligned}\max_{\alpha,\alpha',\delta,\delta'}\quad & \delta+\delta'\\ \text{such that}\quad & \sum_{k=0}^{K}(\alpha_k-\alpha_k')F_k+\delta\Gamma(\beta)^{-1}=\delta'\Gamma(\beta')^{-1}\\ & \sum_{k=0}^{K}\alpha_k+\delta=\sum_{k=0}^{K}\alpha_k'+\delta'=1,\quad \alpha_0,\dots,\alpha_K,\alpha_0',\dots,\alpha_K',\delta,\delta'\geq 0\end{aligned}$$

has output 0 *and the matrices* $F_0,\dots,F_K,\Gamma(\beta)^{-1}$ *are linearly independent for every* $\beta\neq\beta'$.

The proofs of this and any other theorem in this article can be found in Appendix B (Metodiev et al. (2026)). The theorem can be used to verify identifiability, since its conditions can be checked with a grid-search over β, β'. Let us note that this check will return parameter values for which the model is not identifiable, if it finds any.

Initialization. In order to initialize the procedure, we consider

$$(\alpha^{(0)}, \beta^{(0)}, \delta^{(0)}) := \operatorname*{argmin}_{(\alpha,\beta,\delta)} \|R(\alpha, \beta, \delta) - \hat{R}\|_F, \tag{11}$$

where $R(\alpha, \beta, \delta)$ is defined in equation (7), $\|.\|_F$ denotes the Frobenius norm, and $\hat{R}$ is the Pearson correlation matrix of the estimated standardized errors, $\hat{\varepsilon}_t = \operatorname{diag}(\hat{\sigma}_t)^{-1}(Y_t - \hat{\mu}_t)$. This particular initialization is useful because $R(\alpha^{(0)}, \beta^{(0)}, \delta^{(0)})$ itself is a consistent estimator as long as $\hat{R}$ is consistent.

For fixed β the initialization is an optimization problem that can be solved in polynomial time, as long as the conditions of Theorem 4.1 hold, using a convex quadratic optimization method from Goldfarb and Idnani (1983a, 2006), implemented in the R-package quadprog (Turlach and Weingessel (2019)). Thus, we can solve this problem via a grid-search over β.

Note that the optimal solution of equation (11) may lie on the edge of the parameter space. This is a problem since parameter values, such as $\alpha_1 = 1$, are not allowed. In this case, additional constraints are added to ensure that the optimal solution is feasible. The complete procedure used is defined in Appendix C (Metodiev et al. (2026)).

Gradient descent. It is equivalent, and computationally more convenient, to maximize a specific transformation of the likelihood,

$$l(\alpha, \beta, \delta) := \frac{1}{T} \log \prod_{t=1}^{T} \frac{1}{|R(\alpha, \beta, \delta)|} \exp\bigl(-\hat{\varepsilon}_t^{\top} R(\alpha, \beta, \delta)^{-1} \hat{\varepsilon}_t\bigr),$$

where $|R(\alpha, \beta, \delta)|$ denotes the determinant of $R(\alpha, \beta, \delta)$. We do this by using a quasi-Newton algorithm (BFGS) (Fletcher and Reeves (1964), Broyden (1970), Goldfarb (1970), Shanno (1970)). This algorithm requires derivatives of l, which are easy to obtain in our case since there is a simple expression for the spatial effect matrix $\Gamma(\beta_D)^{-1}$. A list of these derivatives is given in Appendix A (Metodiev et al. (2026)).

4.2. *MLE properties*. In Theorem 4.2 below, we prove several desirable properties of this estimator, such as consistency and asymptotic normality in the number of time points T. Interestingly, it turns out that, under mild conditions, we also have consistency and asymptotic normality in the dimension of the data, d, even if there is just one time point. This result is given in Theorem 4.4. This is important because in the case of the TFR dataset the data are observed for at most $d = 195$ countries and $T = 11$ time points.

THEOREM 4.2. *Let* $(\alpha^\star, \beta^\star, \delta^\star)$ *be the true model parameter and* $\Sigma_t^\star$ *denote the true covariance matrix*, $R^\star$ *the true correlation matrix. Suppose that*:

(A1) *the model given by equation* (7) *is identifiable*,

(B1) $\Gamma(\beta)^{-1}$ *is a uniformly continuous function in* β *with open, convex and bounded domain*,

(C1) *the squared estimated standardized errors* $S_T = \frac{1}{T}\sum_{t=1}^{T} \hat{\varepsilon}_t \hat{\varepsilon}_t^{\top}$ *almost surely converge to* $R^\star$.

Then the normalized SCE, $\hat{R}_{\mathrm{SCE}} = R(\hat{\alpha}_{\mathrm{SCE}}, \hat{\beta}_{\mathrm{SCE}}, \hat{\delta}_{\mathrm{SCE}})$, *is strongly consistent in the number of time points* T.

If assumptions (A1)–(C1) *hold and*:

(A2) $(\hat{\alpha}_{\mathrm{SCE}}, \hat{\beta}_{\mathrm{SCE}}, \hat{\delta}_{\mathrm{SCE}})$ *is a stationary point of* l,

(B2) $\Gamma(\beta)$ *is at least twice differentiable in* β *with continuous partial derivatives*,

(C2) *$Z_T := \sqrt{T}(S_T - R^\star)$ converges in distribution to Z, a random matrix for which the vector $(Z_{1,1}, Z_{1,2}, \ldots, Z_{d,d})$ follows a multivariate normal distribution with $E[Z_{i,j}] = 0$ and $\mathrm{Cov}(Z_{i_1,j_1}, Z_{i_2,j_2}) = \mathrm{Cor}((Y_1 Y_1^\mathsf{T})_{i_1,j_1}, (Y_1 Y_1^\mathsf{T})_{i_2,j_2})$, where Y_1 denotes the vector of the first observation,*

then the vector $\sqrt{T}((\hat{\alpha}_{\mathrm{SCE}}, \hat{\beta}_{\mathrm{SCE}}, \hat{\delta}_{\mathrm{SCE}}) - (\alpha^\star, \beta^\star, \delta^\star))$ is asymptotically normally distributed with means 0 and covariance matrix equal to the inverse of the Fisher information matrix

$$I(\alpha, \beta, \delta)_{i,j} = \frac{1}{2}\mathrm{tr}\left(R(\alpha, \beta, \delta)^{-1} \frac{\partial R}{\partial (\alpha, \beta, \delta)_i}(\alpha, \beta, \delta) R(\alpha, \beta, \delta)^{-1} \frac{\partial R}{\partial (\alpha, \beta, \delta)_j}(\alpha, \beta, \delta) \right)$$

evaluated at the true parameter, in the number of time points T.

REMARK 4.3. The correlation matrix of the CAR-model, which we use in the experiments section (Section 5) of this article, fulfills Conditions (B1) and (B2). We can check for identifiability (Condition A1), using Theorem 4.1. As time passes, the countries eventually enter phase III of the model of Alkema et al. (2011a). Thus, the majority of the mean and variance estimators used by Fosdick and Raftery (2014) are, for T sufficiently large, approximating MLEs of the autocorrelation and variance parameter of an AR(1) model, since only phase III of their model is relevant for consistency in T. These types of estimators are strongly consistent and asymptotically normal for a very general class of models that use time series (Hannan (1973)) and the strong law of large numbers still holds for weakly correlated observations (Lyons (1988)), so there is an argument to be made that Conditions (C1) and (C2) hold as well. Additionally, asymptotic confidence regions can be obtained by calculating the inverse of the Fisher information matrix. We can observe the properties of Theorem 4.2 in the experiments section.

In our case we are going to show in Section 5.2 of this article that the number of time points necessary for convergence is quite small, as long as the dimension d of the data is sufficiently large. In fact, thanks to the following theorem, we have convergence, even if $T = 1$.

THEOREM 4.4. *Suppose that the estimated standardized errors follow an i.i.d. normal distribution with mean 0_d, unit variances and correlation matrix R, and that a global maximum of the likelihood is given by $(\hat{\alpha}_{\mathrm{SCE}}, \hat{\beta}_{\mathrm{SCE}}, \hat{\delta}_{\mathrm{SCE}})$. Suppose also that*:

(A) *$\Gamma(\beta)$ is two times differentiable with respect to β with continuous partial derivatives and open domain,*

(B) *the correlation matrix of the Fisher information matrix, $\mathrm{Cor}(I(\alpha, \beta, \delta)^{-1}) = Q_I(\alpha, \beta, \delta) I(\alpha, \beta, \delta)^{-1} Q_I(\alpha, \beta, \delta)$, where Q_I is the unique nonnegative diagonal matrix for which this product is a matrix with a diagonal of ones, exists and is nonsingular,*

(C) *$F_0 = I_d$, the sum of each row of the elementwise absolute values $|F_1|, \ldots, |F_K|$, $|\Gamma(\beta)^{-1}|, |\frac{\partial}{\partial \beta_g}|\Gamma(\beta)^{-1}|, |\frac{\partial^2}{\partial \beta_g \partial \beta_w}|\Gamma(\beta)^{-1}|$ is bounded in d (i.e., it is $O(1)$) and for a proportion of at least $p \in (0, 1)$ of the rows there exists a $\tau > 0$ such that each sum of squares of the nondiagonal row-elements has a lower-bound of τ.*

Then $\sqrt{I(\hat{\alpha}_{\mathrm{SCE}}, \hat{\beta}_{\mathrm{SCE}}, \hat{\delta}_{\mathrm{SCE}})}((\alpha^\star, \beta^\star, \delta^\star) - (\hat{\alpha}_{\mathrm{SCE}}, \hat{\beta}_{\mathrm{SCE}}, \hat{\delta}_{\mathrm{SCE}}))$ converges in distribution in d to a standard normal random variable, and the estimators $(\hat{\alpha}_{\mathrm{SCE}}, \hat{\beta}_{\mathrm{SCE}}, \hat{\delta}_{\mathrm{SCE}})$ are consistent.

REMARK 4.5. The theorem holds for any value of T, the number of time points, even for $T = 1$. Conditions (A) and (C) are fulfilled by the CAR model, as long as the size of the largest connected component of the underlying spatial structure is bounded in d and there are fewer island-countries than nonisland-countries. In the case of the TFR dataset, Conditions

(B)–(C) boil down to conditions on the data having multiple distinct clusters, each of which is limited in its size but has more than one component. This is mostly the case, as there are several regions, colonizers, and (approximately) separate connected components (i.e., continents). It is not fulfilled by the global effect, since its cluster is of size d. However, we found that the global effect only made a small impact in practice and thus should not affect performance too much. In addition, the assumption that the mean- and standard deviation estimators are exact is crucial, since it only occurs approximately in practice. We analyze the convergence of our model in the number of countries, d, and discuss the effects of the mean- and standard deviation estimators in Section 5.3.

REMARK 4.6. It may be of interest to include *time-dependent* covariates into the estimation procedure. In principle, this does not change our estimation procedure, but results with respect to the asymptotic behaviour of our estimator as T goes to infinity do not hold in this case. However, Theorem 4.4 makes no assumptions on T, and so convergence in d will hold, even if the correlation matrix R was to vary with T.

4.3. *Model selection.* All the available matrices may not be relevant for modeling the covariance of the data. If this is the case, only a subset of the matrices $\{\Gamma(\beta)^{-1}, F_1, \dots, F_K\}$ should be used to construct the model.

For an index set $J \subseteq \{-1, 0, 1, \dots, K\}$, we define

$$R_J(\alpha, \beta, \delta) = \Phi_J(\alpha) + \delta \cdot \Gamma(\beta)^{-1} \cdot \mathbb{1}_{-1\in J}, \quad \text{where } \Phi_J(\alpha) = \sum_{k\in J\setminus\{-1\}} \alpha_k F_k.$$

Then we conduct model selection via the Bayesian information criterion (BIC). For any given index set J, we define the BIC as

$$-2\log\big(L_J(Y; \hat{\mu}, \hat{\sigma}, \hat{\alpha}_{\text{SCE}}, \hat{\beta}_{\text{SCE}}, \hat{\delta}_{\text{SCE}})\big) + \big(|J| + G \cdot \mathbb{1}_{-1\in J}\big)\log(T),$$

with L_J defined like L in equation (10) and R replaced by R_J in the definition of Σ. The asymptotic normality of the model parameters, guaranteed by Theorems 4.2 and 4.4, ensures that, under mild conditions, the posterior distribution is approximately normal, which justifies the use of the BIC (Raftery (1995)).

4.4. *Model misspecification.* In some cases, only parts of the correlation coefficients are explained by known covariates. In such cases, equation (7) does not hold, and the SCE is not consistent. However, the available covariates may still be useful to improve the efficiency of the covariance matrix estimation.

To capture both the part of the correlation that is due to the known covariates and the part that is not, we combine $\hat{R}_{\text{SCE}}$ with another correlation estimator that does not make any assumption on the covariance structure.

This is the case for the Pearson correlation estimator. In our specific setting, we have access to $\hat{\mu}$ and $\hat{\sigma}$, so we propose using a Pearson-type estimator,

$$(\hat{R}_{\text{Pearson}})_{i,j} = \begin{cases} \dfrac{1}{T-1}\displaystyle\sum_{t=1}^{T} \frac{(Y_{t,i} - \hat{\mu}_{t,i})(Y_{t,j} - \hat{\mu}_{t,j})}{\hat{\sigma}_{t,i}\hat{\sigma}_{t,j}} & i \neq j, \\ 1 & i = j. \end{cases}$$

Note that if $\hat{\mu}$ and $\hat{\sigma}$ correspond to the sample mean and variance, $\hat{R}_{\text{Pearson}}$ is equal to the Pearson correlation matrix. If not, it may not be positive definite in which case we map it to the positive definite correlation matrix that is closest in Frobenius norm, using the algorithm of Cheng and Higham (1998), implemented in the "Matrix" R package (Bates, Maechler and Jagan (2024)).

The convex combination of this estimator and $\hat{R}_{\text{SCE}}$ gives the estimator

$$\begin{aligned}\hat{\Sigma}_t^{\text{WSCE}} &= \operatorname{diag}(\hat{\sigma}_t)\hat{R}_{\text{WSCE}}\operatorname{diag}(\hat{\sigma}_t),\\ \hat{R}_{\text{WSCE}} &= (1-\hat{\lambda}_{\text{WSCE}})\hat{R}_{\text{SCE}} + \hat{\lambda}_{\text{WSCE}}\hat{R}_{\text{Pearson}}.\end{aligned} \tag{12}$$

$\hat{\Sigma}_t^{\text{WSCE}}$ is consistent as long as $\hat{\lambda}_{\text{WSCE}} \in [0,1]$ approaches one, which is the case for the optimal value that we propose. Ledoit and Wolf (2003) make a very similar argument and give optimality conditions for the optimal mixing constant between two estimators, one of which may not be consistent. The same results hold in our setting, under the following conditions.

THEOREM 4.7. *Suppose that all conditions of Theorem 4.2 hold and*:

(A3) *the model is misspecified, meaning that* $R^\star \neq R(\alpha,\beta,\delta)$ *for all* (α,β,δ), *where* $R^\star$ *is the true, nondegenerate correlation matrix,*

(B3) $(\hat{\alpha}_{\text{SCE}}, \hat{\beta}_{\text{SCE}}, \hat{\delta}_{\text{SCE}})$ *converges almost surely to its limit,* $(\alpha^\star, \beta^\star, \delta^\star)$, *and the* L^2 *limit of* $\sqrt{T}((\hat{\alpha}_{\text{SCE}}, \hat{\beta}_{\text{SCE}}, \hat{\delta}_{\text{SCE}}) - (\alpha^\star, \beta^\star, \delta^\star))$ *exists,*

(C3) $\beta^\star$ *is in the domain of* Γ,

(D3) $Z_T := \sqrt{T}(S_T - R^\star)$ *converges in* L^2 *to* Z, *a random matrix for which the vector* $(Z_{1,1}, Z_{1,2}, \ldots, Z_{d,d})$ *follows a multivariate normal distribution with* $E[Z_{i,j}] = 0$ *and* $\operatorname{Cov}(Z_{i_1,j_1}, Z_{i_2,j_2}) = \operatorname{Cor}((Y_1Y_1^\mathsf{T})_{i_1,j_1}, (Y_1Y_1^\mathsf{T})_{i_2,j_2})$, *where* Y_1 *denotes the vector of the first observation.*

Then the constant $\hat{\lambda}^*_{\text{WSCE}}$ *in equation* (12), *which minimizes the expected squared error of* $\hat{R}_{\text{WSCE}}$ *in the Frobenius norm, is given by*

$$1 - \hat{\lambda}^*_{\text{WSCE}} = \frac{1}{T}\cdot\frac{\pi-\rho}{\gamma} + O\left(\frac{1}{T^2}\right) = O\left(\frac{1}{T}\right),$$

where

$$\begin{pmatrix}\pi\\ \rho\\ \gamma\end{pmatrix} = \sum_{i,j}^{d}\begin{pmatrix}\operatorname{Var}(\sqrt{T}(\hat{R}_{\text{Pearson}})_{i,j})\\ \operatorname{Cov}(\sqrt{T}(\hat{R}_{\text{SCE}})_{i,j}, \sqrt{T}(\hat{R}_{\text{Pearson}})_{i,j})\\ (\mathrm{E}[(\hat{R}_{\text{SCE}})_{i,j}] - R^\star_{i,j})^2\end{pmatrix}.$$

$\hat{\lambda}^*_{\text{WSCE}}$ can be thought of as a shrinkage constant. If the asymptotic expected error of the SCE, γ, is small and if the asymptotic variance of the Pearson-type correlation matrix, π, is large, the WSCE is shrunk toward the SCE. Otherwise, it is shrunk toward the Pearson-type correlation matrix. $\hat{\lambda}^*_{\text{WSCE}}$ approaches 1 as the sample size T increases, which in turn ensures consistency of the WSCE due to the consistency of the Pearson-type correlation matrix.

Empirical estimates for the asymptotic expectations and variances used to define π and γ are readily available via standard estimation procedures (we give our choice in Appendix C, Metodiev et al. (2026)). However, ρ is defined via the covariances of $\hat{R}_{\text{SCE}}$ and the Pearson-type correlation matrix. These may be estimated via a bootstrap approach, but this would require a repeated calculation of the SCE over a large simulated sample, which is computationally expensive. Alternatively, one could approximate an upper bound of ρ, ρ_U, which is given by the Cauchy–Schwartz inequality,

$$\rho_U = \sum_{i,j}^{d}\sqrt{\operatorname{Var}[\sqrt{T}(\hat{R}_{\text{SCE}})_{i,j}]}\sqrt{\operatorname{Var}[\sqrt{T}(\hat{R}_{\text{Pearson}})_{i,j}]}.$$

This upper bound approaches ρ as the correlation between the two estimators increases. It is also easy to estimate, since estimators of the asymptotic variances of the Pearson correlation

are well known, and the Fisher information can still be used to compute the variance of the SCE under model misspecification. One can thus make a choice between setting

$$\hat{\lambda}^{\mathrm{U}}_{\mathrm{WSCE}} = 1 - \frac{1}{T} \cdot \frac{\hat{\pi} - \hat{\rho}_U}{\hat{\gamma}}, \quad \text{or} \quad \hat{\lambda}_{\mathrm{WSCE}} = 1 - \frac{1}{T} \cdot \frac{\hat{\pi} - \hat{\rho}}{\hat{\gamma}},$$

where the former expression can be computed efficiently, while the latter is more precise, but also more expensive to compute. In either case, $\hat{R}_{\mathrm{WSCE}}$ is a consistent estimator of the correlation matrix of the data, under no assumptions on its correlation structure.

Note that it is possible that $\hat{\lambda}^{\mathrm{U}}_{\mathrm{WSCE}}$ and/or $\hat{\lambda}_{\mathrm{WSCE}}$ lie outside of $[0, 1]$. In this case they are rounded to 0 or 1. $\hat{\lambda}^{\mathrm{U}}_{\mathrm{WSCE}}$ is also an upper bound on the optimal shrinkage constant, meaning that the WSCE can be shrunk too highly toward the Pearson-type correlation matrix. This can be avoided by computing $\hat{\lambda}_{\mathrm{WSCE}}$ by using a bootstrap algorithm. We used this approach when the dataset contained missing values or when the mean and variance estimators were inprecise. However, we found that the upper bound that we suggested is quite accurate when the mean and variance vectors of the model are considered known and there are no missing values present, as illustrated in Section 5.4 below.

5. Numerical experiments. This section presents numerical experiments to emphasize the advantages and limits of the proposed estimators compared to several state-of-the-art methods. First, we compare their performances within the model assumptions, with varying covariate structures and sizes for the datasets. Then we test the robustness of the model under missing values, model misspecification and perform a sensitivity analysis of the assumption that the underlying distribution of the data is multivariate Gaussian.

5.1. *Simulation settings*.

Sample distribution. For each scenario, we simulated 40 independent datasets. Each dataset contains $T = 11$ samples drawn independently such that

$$Y_t \sim \mathrm{MVN}_d(2.1_d, R) \text{ for } 1 \le t \le T, \tag{13}$$

with R denoting a correlation structure from equation (5). Details of all simulation settings can be found in Appendix E (Metodiev et al. (2026)).

Depending on the scenario, we considered one of, or a combination of, the following two settings:

- *Fully simulated setting (FSS):* For each country the membership vector indicating its region (resp., colonizer) is drawn from a multinomial distribution. The adjacency matrix of the spatial structure is simulated from an Erdős–Rényi random graph model (Erdős and Rényi (1959)) with connection probability $\log(d)/d$.
- *Using the TFR data (TFR):* F_A, F_B, F_C, and the adjacency matrix that defines the function $\Gamma(\beta_D)$ are chosen from the real data.

REMARK 5.1. In the TFR dataset, the data points are not i.i.d. However, the standardized errors, which we use to estimate the covariance matrix, are approximately i.i.d., so this general setting is appropriate. Moreover, simulating the data as i.i.d. allowed us to compare our estimators to many standard estimators which can only work in an i.i.d. setting.

External information. The SCE and the WSCE depend on estimates of μ and σ. These can either be computed beforehand, using external information, or via standard estimators (the sample mean and variance), or in direct combination with the parameters of the SCE, by maximizing the likelihood function jointly over $(\mu, \sigma, \alpha, \beta, \delta)$. Thus, we distinguish three cases:

- *known*, where the true μ and σ were used to calculate the SCE and the WSCE,
- *unknown (one-step)*, where the mean, variance and the SCE parameters were estimated jointly in one single estimation step,
- *unknown (two-step)*, where μ and σ were estimated by standard estimators. In turn, these estimators were then used in the estimation procedure of the SCE parameters (α, β, δ), resulting in a two-step procedure.

REMARK 5.2. Alkema et al. (2011a) provide accurate estimates of μ and σ for the TFR dataset. Within their model, μ and σ were estimated with only $15 + 3 = 18$ parameters (hyperparameters of the Bayesian hierarchical model of phase II (Alkema et al. (2011b)) plus parameters of the AR(1)-model of phase III) and not $2 \cdot d = 390$, which are needed for the empirical estimators in the unknown setting. Thus, the performance of the SCE (resp., WSCE) is likely to be in between the known and unknown case.

Comparison estimators. For comparison, we considered the initial value estimator (IVE),

$$\hat{\Sigma}_t^{\mathrm{IVE}} = \Sigma\big(\hat{\sigma}_t, \alpha^{(0)}, \beta^{(0)}, \delta^{(0)}\big), \tag{14}$$

where $(\alpha^{(0)}, \beta^{(0)}, \delta^{(0)})$ are determined from the initialization step.

We also computed Pearson's correlation matrix estimator as well as the Ledoit–Wolf estimator (Ledoit and Wolf (2004)), an estimator that uses factor models (Zhou and Palomar (2024)), where the number of factors was chosen to be equal to the number of variables, and the glasso estimator, an estimator for a sparse precision matrix, implemented in Galloway (2018), where the hyperparameter is chosen using cross-validation.

Performance evaluation. To compare the performances of the different estimators, we compare their mean absolute error (MAE) evaluated on the scale of the correlation matrix,

$$\frac{1}{d^2} \sum_{i=1}^{d} \sum_{j=1}^{d} \big| R^{\star}_{i,j} - \bar{R}_{i,j} \big|, \tag{15}$$

where $R^{\star}$ denotes the true correlation matrix and $\bar{R}$ denotes the estimated correlation matrix.

5.2. *Settings within the model assumptions*.

Description. In this section we compare the performance of the different estimators within the model assumption, in both *FSS* with $d = 200$ and *TFR* with $d = 195$. We also study the impact of external information on the parameters.

Results. The results are shown in Figure 3. In all settings, the WSCE, SCE, and the IVE outperformed the other estimators. Since these three estimators are the only ones that can take advantage of knowing μ and σ, they are the only ones that change between the known and unknown cases. Knowing the parameters has little impact on the IVE but improves the performances of both the SCE and WSCE. Thus, they outperformed the IVE in such a scenario. In the unknown case, the one-step procedure seems to improve the performance of the WSCE and SCE, though the improvement is rather minor. In particular, it seems that negligible bias is produced when applying the two-step instead of the one-step procedure. This result is important for the case of the TFR dataset, since the two-step approach is used to estimate its covariance matrix. Note that, whenever the mean and variance parameters are unknown, the simple version of $\hat{\lambda}_{\mathrm{WSCE}}$, $\hat{\lambda}^{U}_{\mathrm{WSCE}}$, was unstable. Thus we used the parametric bootstrap to calculate $\hat{\lambda}_{\mathrm{WSCE}}$ (see Appendix C of Metodiev et al. (2026) for more details).

5.3. *Performance with varying dimensions of the data*. We now describe the performance of the WSCE for different values of d.

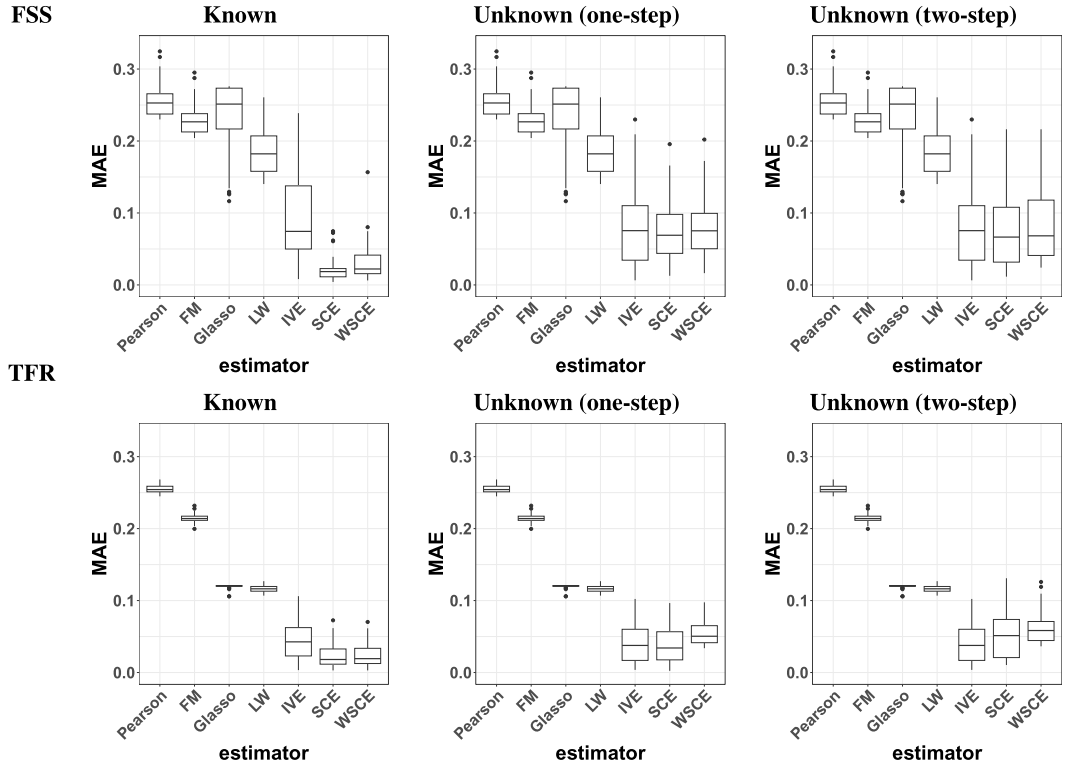

FIG. 3. *Boxplots of the mean absolute error* (*MAE*) *for the Pearson correlation matrix* (*Pearson*), *the estimator that uses factor models* (*FM*), *the glasso estimator* (*Glasso*), *the Ledoit–Wolf estimator* (*LW*), *and the IVE*, *SCE*, *and WSCE for* 40 *independent simulations*. *Estimators were evaluated in the fully simulated setting* (*FSS*) *and the setting of the TFR dataset* (*TFR*): Left: *The case when the means and variances are known*. Middle: *The case when they are unknown and estimated jointly with the other SCE parameters*. Right: *The case when they are unknown and estimated by using the sample mean and variance*. *The errors of the IVE*, *SCE*, *and WSCE are the lowest*, *with the SCE and WSCE outperforming the IVE when the means and variances are known*.

Description. We consider the *TFR* setting with known mean and variance parameters. We selected subsets of the data of sizes $d = 14, 32, 65, 115, 195$. These subsets are created by cumulatively including countries in Southern and Middle Africa, Eastern Africa, Western and Northern Africa, Asia, America, and all other countries, respectively.

Results. The results are shown in Figure 4. $T = 11$ was fixed, and 40 different datasets were simulated independently for each value of d. The mean absolute error decreases with the number of countries, d. This was expected due to the consistency in d of the parameters (Theorem 4.4). In other words, adding more countries to the dataset improved the performance of the WSCE. This effect prevailed—although much less pronounced—even in the case that the mean- and variance parameters had to be estimated before computing the WSCE.

5.4. *Performance under model misspecification and the presence of missing values*.

Settings. Here we study the impact of model misspecification on our estimators. We replace R by R_{miss} in equation (13), where

$$R_{\text{miss}} = \xi R(\alpha, \beta, \delta) + (1 - \xi)\tilde{R},$$

with $\xi \in [0, 1]$, $R(\alpha, \beta, \delta)$ denoting the correlation structure from the *TFR* setting and $\tilde{R}$ being simulated using an additional matrix F_{miss} from the *FSS* setting, which is not given to

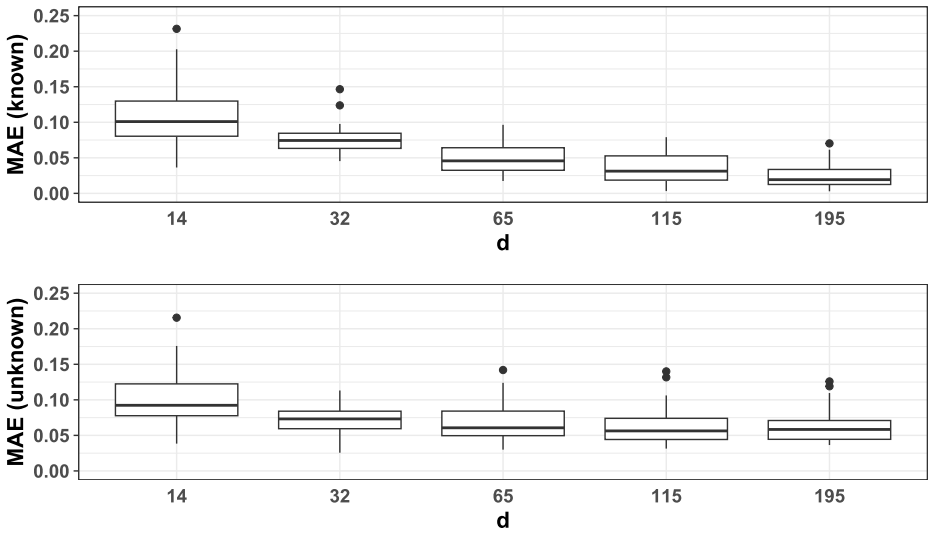

FIG. 4. *Boxplots of the mean absolute error (MAE) of the WSCE with different values of d, repeated for* 40 *independent simulations in the case that the means and variances are known and in the case that the means and variances are unknown and have to be estimated. The values of d correspond to the number of countries in the following regions (added cumulatively): Southern and Middle Africa, Eastern Africa, Western and Northern Africa, Asia, and America, the remaining countries in the dataset.*

the model. Thus, the model assumptions hold when $\xi = 0$, and the correlation structure does not depend on the observed covariates when $\xi = 1$.

In the TFR dataset, the values of the standardized errors are missing for countries that have not yet entered phases II or III of the model of Alkema et al. (2011a). Thus, in this scenario we consider two settings: one without missing values and the other one with missing values. In the latter scenario, we set the values of Y_t that were missing in the TFR dataset to be missing. The IVE, SCE, and WSCE need to be adapted under the presence of missing values. The corresponding procedure is described in Appendix E (Metodiev et al. (2026)).

Results. The results are shown in Figure 5. Ten independent datasets were simulated for each value of ξ, and estimators were evaluated on the scale of the correlation matrix. The SCE and the WSCE perform similarly and outperform the other estimators in the scenarios that are not too far away from the model assumptions. As expected, the performances of both the SCE and the WSCE deteriorate when the value of ξ increases, meaning that the scenario gets further away from the model assumptions. However, when the scenario is very different from the model assumptions ($\xi > 0.5$), the WSCE outperforms the SCE and performs at least as well as the other estimators.

5.5. *Sensitivity analysis of the assumption of multivariate normality.*

Settings. To show how our estimators perform in the case that the data are non-Gaussian, we simulate from a truncated multivariate t-distribution. The t-distribution is truncated at 0 such that all simulated values are positive, since the definition of the TFR makes negative values impossible. The degrees of freedom of this distribution are varied between 10, where the t-distribution is quite far away from a normal distribution, and 1,000, where they are fairly similar.

Results. The results are shown in Figure 6. Forty independent datasets were simulated for each value of the degrees of freedom, and estimators were evaluated on the scale of the correlation matrix. In a way, these results mimic the results from the previous section. While

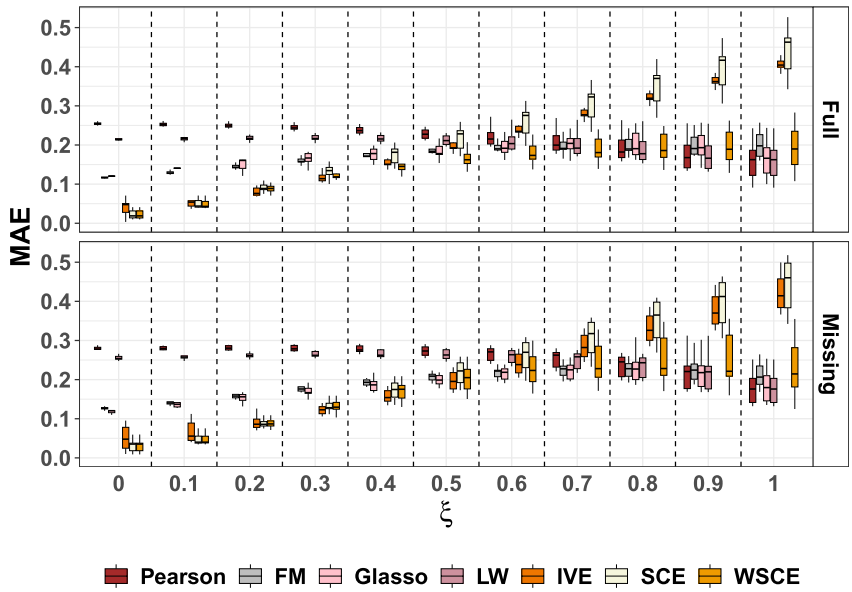

FIG. 5. *Boxplots of the mean absolute error* (*MAE*) *of the Pearson correlation matrix* (*Pearson*), *the Ledoit–Wolf estimator* (*LW*), *the estimator that uses factor models* (*FM*), *the glasso estimator* (*Glasso*), *IVE*, *SCE*, *and WSCE*, *for* $\xi = 0, 0.1, \ldots, 0.9, 1$. *Ten independent simulations were calculated by adding additional structure to the correlation*, *which is unknown and randomly simulated. At* $\xi = 1$, *the correlation structure does not depend on the observed covariates*: *Full*: *No missing values*. *Missing*: *Values that were missing in the original dataset were also set to be missing*. *The results do not change much between these two settings*. *This indicates that the missing-value imputation used is quite robust*.

the SCE performs well when the distribution of the the data is close to a multivariate normal, it is struggling when this distribution is farther away. The IVE, however, does not suffer from this problem. This is not surprising, since the construction of the IVE only makes use of the correlation structure of the data but does not require the data to follow any specific distribution, as long as its correlation structure is accurately specified. It follows that the IVE is a semiparametric estimator and thus not impacted by the change in distributions. The WSCE captures the fact that the distribution of the data is misspecified and thus stays close to the other covariance matrix estimators, not suffering from the same shortcomings as the SCE. Overall, when comparing all settings of the simulation section, we can thus conclude that the WSCE outperformed or showed similar performance to all of the other estimators in the vast majority of those settings.

6. Covariance estimates for the TFR dataset. In this section we will study the TFR dataset. We start by describing the data in detail. Then we calculate our estimators with and without interaction terms, and we finish by performing model selection.

6.1. *TFR data description*. As described in Section 3.1, we want to estimate the covariance matrix of the total fertility rate (TFR) for 195 countries. Since we are in a Markov model with dependent observations, we estimate the covariance matrix of the TFR conditional on its preceding values, where the covariance matrix can vary with time. We assume that we are in the model described by equation (5).

We used the estimates of Fosdick and Raftery (2014) for the mean- and standard deviation vectors in the setting of Alkema et al. (2011a), meaning that it was assumed that

$$\mu_{t,j} = Y_{t-1,j} - \tilde{\mu}(\theta_j, Y_{t-1,j}), \qquad \sigma_{t,j} = \tilde{\sigma}(\theta_j, Y_{t-1,j}),$$

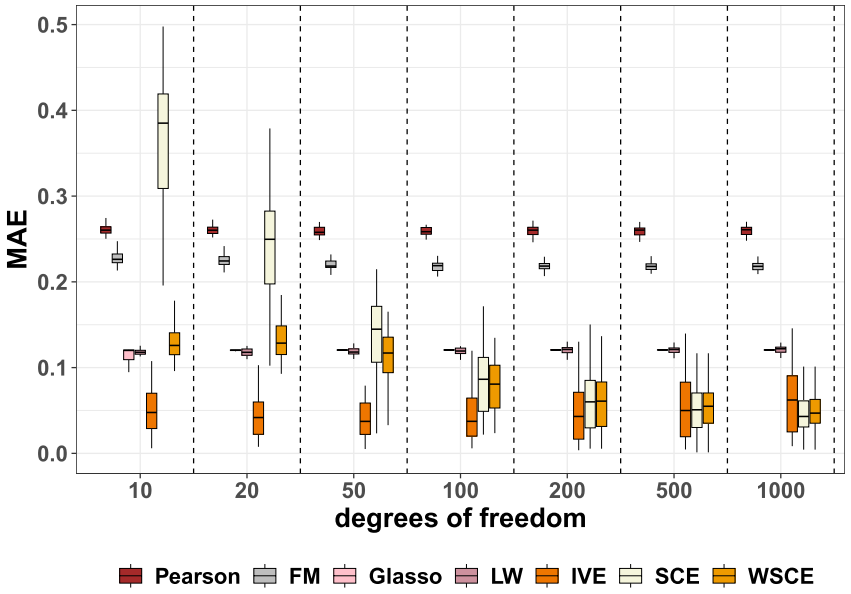

FIG. 6. *Boxplots of the mean absolute error* (*MAE*) *of the Pearson correlation matrix* (*Pearson*), *the Ledoit–Wolf estimator* (*LW*), *the estimator that uses factor models* (*FM*), *the glasso estimator* (*Glasso*), *IVE*, *SCE*, *and WSCE*, *for varying degrees of freedom of the trunacted multivariate t distribution. Forty independent simulations were calculated by sampling from this truncated multivariate t-distribution. At* 1,000 *degrees of freedom*, *this distribution is quite close to a normal distribution*; *at* 10 *degrees of freedom it is much further away.*

during phase II and that during phase III

$$\mu_{t,j} = 2.1 + 0.8859(Y_{t-1,j} - 2.1), \qquad \sigma_{t,j} = 0.1016.$$

Here $\tilde{\mu}$ describes a double logistic function, and $\tilde{\sigma}$ denotes the corresponding standard deviation function, which depend on a vector of parameters $\theta = (\theta_1, \ldots, \theta_d)$. Note that this vector is not estimated directly but via a Bayesian hierarchical model with 15 hyperparameters (Alkema et al. (2011b)). Estimates of the standard deviation and mean vector were computed by simulating an MCMC sample and averaging over the parameter-specific forecast errors. These estimates were used to normalize the data and obtain the standardized residuals, to which our method will be applied.

Let us recall that, in the model described by equation (5), the data $(Y_1, \ldots, Y_T)$ denotes the TFR at time $1, \ldots, T$. We fit the model described in equation (9) with

$$R(\alpha, \beta, \delta) = \alpha_A F_A + \alpha_B F_B + \alpha_C F_C + \delta_D \Gamma(\beta_D)^{-1} + \alpha_E I_d, \tag{16}$$

where $(F_A)_{i,j}$, $(F_B)_{i,j}$ are equal to 1 if country i and j have the same common colonizer or belong to the same U.N. region, respectively, $(F_C)_{i,j}$ is always 1, and $\Gamma(\beta_D)$ is described by equation (6).

Selected countries. Vanuatu was reportedly colonized by two countries. This nonunique cluster membership could be modeled by splitting the common colonizer covariate into multiple distinct covariates for each respective cluster. However, this would unreasonably increase the number of parameters that we would need to estimate. Thus, we removed the corresponding variable and worked with the remaining $d = 195$ countries.

Missing values. We aim at estimating the covariance matrix for country pairs where both countries have entered either phase II or III of the model of Alkema et al. (2011a). The TFR values of the countries that are still in phase I are thus treated as missing values. As

TABLE 1

The number of countries in phase I of the model of Alkema et al. (2011a) for each of the successive five-year periods used in the TFR dataset. They are decreasing from the period 1950–1955, *during which all countries were in phase I, to the period* 2005–2010, *during which all countries are no longer in phase I*

5-year period	1950–1955	1955–1960	1960–1965	1965–1970	1970–1975	1975–1980
Countries in phase I	196	121	98	74	56	35
5-year period	1980–1985	1985–1990	1990–1995	1995–2000	2000–2005	2005–2010
Countries in phase I	26	7	4	2	0	0

time passes, more countries go from phase I to phase II. Thus, the number of missing values changes between the observations. We give the number of countries that are in phase I at each time period in Table 1. The values of the standardized errors $\varepsilon_{t,j}$, which are used to estimate the covariance matrix, are assumed to be missing at random everywhere. Thus, covariance estimation is appropriate on the marginal distribution of the nonmissing values (Rubin (1976), Seaman et al. (2013)).

REMARK 6.1. We assume that the standardized residuals are missing at random everywhere. This assumption should be fulfilled approximately, since these standardized residuals are approximately independent from the fitted values, which roughly determine the start of phase II. The start of phase II is determined deterministically by first identifying the local maximum within 0.5 absolute distance below the global maximum of the TFR values, setting it as the start of phase II if it is above 5.5, and setting the first observation of the dataset to the start point otherwise (Alkema et al. (2011a)).

6.2. *Covariance without interaction*. In this section we compute the SCE by estimating the parameters of equation (16).

Impact of the covariate. We introduce the concept of average effects to compare the effects of the different covariates in the model. The average effect of a covariate is the average correlation in the data that is due to this covariate. For direct effects this corresponds to the value of the linear coefficients ("common colonizer," "same region"). For the contiguity effect, we take the overall mean effect of country pairs which are direct neighbors of each other,

$$\eta_{\text{contig}} = \frac{1}{\sum_{i\neq j} M_{i,j}} \sum_{i\neq j \text{ s.t. } M_{i,j}=1} \delta_D \Gamma(\beta_D)^{-1}_{i,j}.$$

REMARK 6.2. η_{contig} is not equal to δ_D because, contrary to the values of F_A or F_B, $\Gamma(\beta_D)^{-1}_{i,j}$ is not always equal to 0 or 1.

The rationale is that if one adds all the pertinent coefficients for a given covariate, one gets the mean of the correlations for data points that have this covariate in common. For instance,

$$\alpha_B + \eta_{\text{contig}}$$

gives the mean of the estimated correlations of countries that are neighbors and in the same region, but not with the same colonizer.

The estimated average effects are given in Table 2. Since we model the correlation- and not the covariance matrix of the TFR, the average effects are directly interpretable: they vary between 0 (no correlation) and 1 (full correlation). In our specific case, we can see in Table 2 that the contiguity effect explains most of the correlation captured by the model, with an average correlation of 0.162, while the common colonizer effect is comparatively small, with an average effect of 0.038.

TABLE 2
Average effects of the model in which all effects are included. All effects were rounded to the third decimal place

	comcol	sameRegion	intercept	contig
avg. effect	0.038	0.044	0.06	0.162

6.3. *Interaction effects and model selection.* We can see in Table 2 that at least two effects needed to overlap for countries to have a correlation higher than 0.2. This was not the case in Fosdick and Raftery (2014), where, for example, the contiguity effect alone accounted for a correlation of 0.26 for all country pairs with TFRs below 5. We wanted to check if there were interaction effects. Indeed, the neighborhood effect, for instance, may be different if you are in the same region or not.

Thus, in addition to the intercept, "common colonizer," "same region," and the spatial effect, we add their interactions by adding random effects with correlation matrices equal to the Hadamard product of the correlation matrices of the individual effects. Martini et al. (2020) do that with covariance matrices, but since we separate the variance from the correlation matrix estimation, we use correlation matrices instead.

We can include up to three interaction effects, with correlation matrices

$$F_{A,B} = F_A \odot F_B, \quad \text{common colonizer and same region effect,}$$

$$F_{A,D} = F_A \odot \Gamma(\beta_D)^{-1}, \quad \text{common colonizer and spatial effect,}$$

$$F_{B,D} = F_B \odot \Gamma(\beta_D)^{-1}, \quad \text{same region and spatial effect.}$$

This gives up to eight effects. We select which effect we should include by computing the BIC for each of the possible $128 = 2^8$ models. However, we exclude interaction effects whenever one of their individual component effects is excluded. This reduces the scope of our model selection to 35 models; the results of which are plotted in Figure 7.

The BIC was centered in the base model described in equation (16). Interestingly, the only models with a better (lower) BIC than the base model are the ones that do include interaction effects. This is in agreement with Fosdick and Raftery (2014), which kept all these effects but did not try to add interactions. The model with the lowest BIC is the model that includes all but one interaction effect, the effect that accounts for the interaction between "common colonizer" and the "same region" effect.

Just as for the previous model, we compared the average effects in this model. For interactions between direct effects (common colonizer and same region), the average effect is the coefficient $\alpha_{A,B}$. For interaction effects that involve the contiguity effect, we take the mean effect of all pairs of countries that are neighbors and in the same region (resp., have the same colonizer),

$$\eta_{\text{contig},B} = \frac{1}{\sum_{i\neq j}(F_B \odot M)_{i,j}} \sum_{i\neq j \text{ s.t. } M_{i,j}=1 \,\&\, (F_B)_{i,j}=1} \alpha_{B,D}\big(F_B \odot \Gamma(\hat{\beta}_{\text{SCE}})^{-1}\big)_{i,j}.$$

Table 3 gives the average effects of the selected model. Effects are much higher when two attributes overlap. This shows that the correlation is particularly high if the contiguity and either the common colonizer or the same region effect applies.

The correlation matrix obtained with these coefficients is plotted in Figure 8. Except for some clusters of countries, the TFRs of two countries are mostly estimated to not be highly correlated, given their previous TFRs.

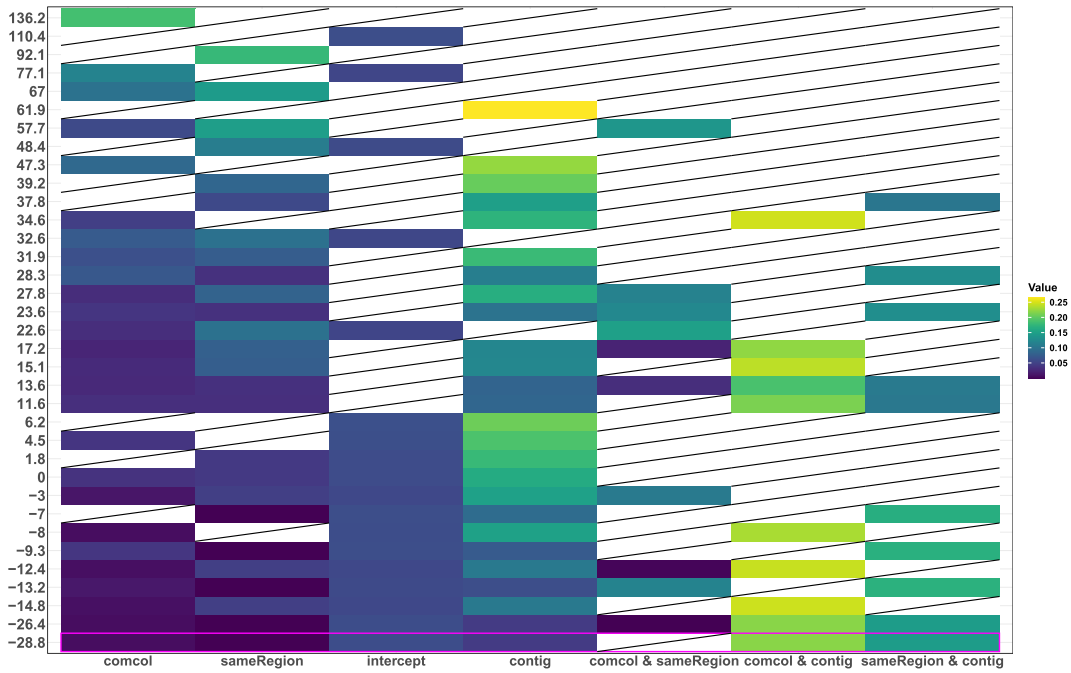

FIG. 7. *On the y-axis*: *Values of the BIC for all* 35 *models tested*; *the BIC is centered such that the model which includes all but interaction effects shows a value of* 0; *squares that were crossed out indicate effects that are not included in their respective model. On the x-axis*: *Different average effects*; *we can see that only models which include interaction effects have a lower BIC than this model*; *the model with the lowest BIC set only the combined effect of the "regional" and "common-colonizer" covariates to* 0.

To check if we do not miss correlations that may come from an effect that is not included in the model, we computed the WSCE. However, it was equal to the SCE. In our simulation settings, this corresponds to the case where our model assumptions were correct. Thus, we can find no evidence against our model assumptions.

7. Discussion. We introduced the structured covariance estimator (SCE) and the weighted structured covariance estimator (WSCE), estimators for large covariance matrices in the presence of pairwise covariates. We showed consistency and asymptotic normality of these estimators in the dimension of the data and in the number of data points and gave a procedure for estimating their confidence regions, under mild assumptions. Furthermore, we tested our estimators in scenarios in which some part of our model was misspecified, where the WSCE performed well. We applied the WSCE and SCE to estimate the covariance matrix of a model for the total fertility rate (TFR) used by the United Nations (U.N.) for 195 different countries. Our results could be used by the U.N. to better adjust their prediction intervals for forecasting regions consisting of multiple countries, since these were too narrow

TABLE 3

Average effects of the model chosen in Figure 7 *in which all effects and their interactions, but the interaction of the "common colonizer" and the "same region" effect, are included. All effects but the regional effect were rounded to the third decimal place*

	comcol	sameRegion	intercept	contig	comcol and contig	sameRegion and contig
avg. effect	0.008	1e−06	0.061	0.046	0.218	0.145

FIG. 8. *A heatmap of the correlation matrix calculated from the WSCE* (*diagonal entries were set to* 0 *to improve the visualization*); *on the x-axis*: *colonizers* (*"comcols"*, *countries with no colonizer were set to belong to colonizer "."*); *on the y-axis*: *regions*; *countries are particularly correlated if they are close to each other and are either in the same region or share the same common colonizer. The matrix was sorted using the order given by hierarchical clustering with the ward distance on the dissimilarity matrix* $1 - \hat{R}_{WSCE}$, *using the option "h-clust" in the "corrplot" package* (*Wei and Simko* (2021)). *A .html file with the labels of the individual estimates is provided under the following link.*

in previous projections (Fosdick and Raftery (2014)). In addition, the estimates themselves provide information, showing that two countries' standardized residuals are highly correlated if they are neighbors and either shared the same common colonizer after 1945 or are in the same region.

We incorporated pairwise information into the covariance matrix estimation by modeling the standardized errors as a sum of weighted and standardized random effects. A very different approach from ours would be to penalize the estimator of the covariance matrix directly. This was done by Liu, Wang and Zhao (2014) and in a Bayesian way by Azose and Raftery (2018), both of which use a weight matrix to penalize individual matrix entries. However, this requires direct estimation of all parameters of the covariance matrix, of which there are $195 \times 194/2 = 18{,}915$ in our case. We do not need to do this in our model, where we only require efficient estimation of a small number of parameters. In fact, Azose and Raftery (2018) pointed out that the high dimension of their parameter space made it impractical for them to carry out a MCMC simulation in their Bayesian setting. In contrast, an extension of our method to the Bayesian paradigm would be straightforward, since we only need to simulate a vector of dimension 6. Combining this with the number of parameters that needed to be estimated in the model of Alkema et al. (2011a), we get $18 + 6 = 24$ parameters needed for estimation, which is well below the total number of $195 \times 11 = 2{,}145$ (number of rows $\times$ number of columns of the dataset) individual data points.

The initial value of our estimator, the IVE, performed well in our simulation study. The WSCE is an asymptotically optimal interpolation between the SCE and the Pearson correlation matrix, as long as the underlying distribution of the data is Gaussian. However, if the distribution is very different from a Gaussian distribution, it might perform suboptimally. In this case, one could instead use a linear interpolation between the IVE and the Pearson correlation matrix, since the IVE is consistent and asymptotically normal if our correlation structure holds, but the data are non-Gaussian. In the specific case of the TFR dataset, we decided to make use of the assumption of multivariate normality of the data, since the validity of this model has been checked thoroughly in Alkema et al. (2011a), where the estimation approach based on this model performed well in terms of out-of-sample predictive performance.

Another assumption of our approach is the time-independence of the correlation matrix. For the TFR dataset, this model has been validated by Fosdick and Raftery (2014), since they showed that the inclusion of the selected time-independent covariates gives more precise prediction intervals than the current model of Alkema et al. (2011a)—the purpose for which this estimator was built. While we are not aware of any other time-independent covariates that could be used to improve our predictions, the choice of using time-independent covariates instead of time-dependent covariates is still largely prescriptive rather than data-driven. This choice may well lead to an incomplete description, since it is well known that time-dependent covariates, such as the unemployment rate, influence the TFR. On one hand, the primary use of our estimates is the inclusion of these estimates in TFR projection that span multiple time-periods, usually multiple decades. The inclusion of time-dependent covariates, or in general any exogenous variables, into models that create such projections is problematic (see De Beer (2000) for a review). This is because a projection of the TFR using time-dependent covariates would require the projection of the covariates themselves, which could be quite difficult due to the uncertainty associated with these covariates. For example, the unemployment rate is a covariate that certainly does have an impact on the TFR, but may well be much harder to predict than the TFR itself. This is why we excluded time-dependent covariates and only included time-independent covariates, since these covariates do not need to be estimated when making future projections. Note that this choice is standard not only in the TFR projections of Alkema et al. (2011a) and Fosdick and Raftery (2014) but also in projections of demographic variables in general, such as projections of life expectancy (Raftery et al. (2012)) and population (Raftery et al. (2013b)), in the sense that none of these projections include time-dependent covariates into their respective model. On the other hand, time-dependend covariates are somewhat included implicitly in our model due to the fact that the model is

trained on TFR data from the past, which is itself influenced by such time-dependent covariates. Thus, the influence of these time-dependent covariates is partially captured by the autoregressive part of our model.

There is also the question of whether we should be using the Pearson correlation matrix in constructing the WSCE. We obtained decent results, even in scenarios far from our model assumptions, but note that the Pearson correlation matrix is known to not behave well in settings with small sample size and a high number of variables. Depending on the setting, one can decide to use a more adapted estimator in the construction of the WSCE, such as the Ledoit–Wolf estimator or Glasso, for instance.

A similar argument can be made for our mean- and variance estimators: we used the fact that accurate estimates of the mean and variance of the data were already provided. This is convenient, but not necessary. In cases where these estimates are not provided we recommend estimating the mean-, variance-, and correlation structure jointly if computationally feasible. In our case the combination of the inference algorithm used for the original model of Alkema et al. (2011a), which is a Bayesian hierarchical model that requires a high-dimensional Markov chain Monte Carlo (MCMC) algorithm for parameter estimation, with our algorithm, which is based on a frequentist approach that mainly requires gradient ascent, is out of the scope of this article. Indeed, our article focuses on modeling the covariance matrix via known pairwise and spatial covariates, not on the combination of frequentist and Bayesian approaches for parameter estimation. However, we do believe that such a joint parameter estimation would indeed be more coherent conceptually and may well reduce the bias created by using one model for the estimation of the mean and variance, and another for the correlation structure. A creation of such a joint estimation procedure could thus be interesting future work.

Finally, the WSCE could be adapted to very general settings such as generalized linear models (GLMs), in a similar vein to Bonat and Jørgensen (2016).

The name of the WSCE refers to the fact that we have a weighted combination of correlation structures defined via known covariates. It is not to be confused with techniques that have similar names but only focus on the estimation of one given matrix structure, such as Burg, Luenberger and Wenger (1983), Sun, Babu and Palomar (2016), Lopuhaä, Gares and Ruiz-Gazen (2023).

Acknowledgements. The authors would like to thank the anonymous referees, an Associate Editor, and the Editor for their constructive comments. The authors would also like to thank Daniel Suen for recommending the re-use of the initialization step of the estimator in the bootstrapping procedure (instead of reevaluating it in each sample that was simulated) and Julie Josse for her recommendation of Josse and Husson (2016). The research of Marie Perrot-Dockès was supported by Université Paris Cité, CNRS, MAP5, F-75006 Paris, France.

Funding. Raftery's research was supported by NIH grant R01 HD-070936, the Fondation des Sciences Mathématiques de Paris (FSMP), and Université Paris Cité (UPC). He thanks the Laboratoire MAP5 at UPC for warm hospitality.

SUPPLEMENTARY MATERIAL

Supplementary material (DOI: 10.1214/26-AOAS2183SUPPA; .pdf). Contains Appendix A (derivatives of the likelihood), Appendix B (proofs), Appendix C (details of the algorithm), Appendix D (alternative model justifications), and Appendix E (simulation settings details).

Code (DOI: 10.1214/26-AOAS2183SUPPB; .zip). Code for scientific dissemination.

REFERENCES

AGUILAR, O. and WEST, M. (2000). Bayesian dynamic factor models and portfolio allocation. *J. Bus. Econom. Statist.* **18** 338–357.

ALKEMA, L., RAFTERY, A. E., GERLAND, P., CLARK, S. J., PELLETIER, F., BUETTNER, T. and HEILIG, G. K. (2011a). Probabilistic projections of the total fertility rate for all countries. *Demography* **48** 815–839.

ALKEMA, L., RAFTERY, A. E., GERLAND, P., CLARK, S. J., PELLETIER, F., BUETTNER, T. and HEILIG, G. K. (2011b). Probabilistic projections of the total fertility rate for all countries, Online Resource 1. *Demography* **48** 815–839.

ANDERSON, T. W. (1973). Asymptotically efficient estimation of covariance matrices with linear structure. *Ann. Statist.* **1** 135–141. MR0331612

AZOSE, J. J. and RAFTERY, A. E. (2018). Estimating large correlation matrices for international migration. *Ann. Appl. Stat.* **12** 940–970. MR3834291 https://doi.org/10.1214/18-AOAS1175

BARBIERI, M. M. and BERGER, J. O. (2004). Optimal predictive model selection. *Ann. Statist.* **32** 870–897. MR2065192 https://doi.org/10.1214/009053604000000238

BARNARD, J., MCCULLOCH, R. and MENG, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statist. Sinica* **10** 1281–1311. MR1804544

BATES, D., MAECHLER, M. and JAGAN, M. (2024). Matrix: Sparse and dense matrix classes and methods. R package version 1.7-0.

BESAG, J. and KOOPERBERG, C. (1995). On conditional and intrinsic autoregressions. *Biometrika* **82** 733–746. MR1380811 https://doi.org/10.1093/biomet/82.4.733

BESAG, J., YORK, J. and MOLLIÉ, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Ann. Inst. Statist. Math.* **43** 1–59. MR1105822 https://doi.org/10.1007/BF00116466

BONAT, W. H. and JØRGENSEN, B. (2016). Multivariate covariance generalized linear models. *J. R. Stat. Soc., Ser. C, Appl. Stat.* **65** 649–675. MR3564998 https://doi.org/10.1111/rssc.12145

BROYDEN, C. G. (1970). The convergence of a class of double-rank minimization algorithms I. General considerations. *IMA J. Appl. Math.* **6** 76–90. https://doi.org/10.1093/imamat/6.1.76

BURG, J. P., LUENBERGER, D. G. and WENGER, D. L. (1983). Estimation of structured covariance matrices. *Proc. IEEE* **70** 963–974.

CHENG, S. H. and HIGHAM, N. J. (1998). A modified Cholesky algorithm based on a symmetric indefinite factorization. *SIAM J. Matrix Anal. Appl.* **19** 1097–1110. MR1636528 https://doi.org/10.1137/S0895479896302898

CHIU, T. Y. M., LEONARD, T. and TSUI, K.-W. (1996). The matrix-logarithmic covariance model. *J. Amer. Statist. Assoc.* **91** 198–210. MR1394074 https://doi.org/10.2307/2291396

CHRISTENSEN, W. F. and AMEMIYA, Y. (2003). Modeling and prediction for multivariate spatial factor analysis. *J. Statist. Plann. Inference* **115** 543–564. MR1985883 https://doi.org/10.1016/S0378-3758(02)00173-8

COX, D. R. and REID, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika* **91** 729–737. MR2090633 https://doi.org/10.1093/biomet/91.3.729

KAROLYI, G. A. (1993). A Bayesian approach to modeling stock return volatility for option valuation. *J. Financ. Quant. Anal.* **28** 579–594.

DE BEER, J. (2000). *Dealing with Uncertainty in Population Forecasting*. Statistics Netherlands, Department of Population, Voorburg.

DE FREITAS, L. A. C., CARLOS, LDO., CAMPOS, A. C. L. and BONAT, W. H. (2022). Hypothesis tests for multiple responses regression: Effect of probiotics on addiction and binge eating disorder. arXiv e-prints, arXiv:2208.00027.

ERDŐS, P. and RÉNYI, A. (1959). On random graphs. I. *Publ. Math. Debrecen* **6** 290–297. MR0120167 https://doi.org/10.5486/pmd.1959.6.3-4.12

FAN, J., LIAO, Y. and LIU, H. (2016). An overview of the estimation of large covariance and precision matrices. *Econom. J.* **19** C1–C32. MR3501529 https://doi.org/10.1111/ectj.12061

FLETCHER, R. and REEVES, C. M. (1964). Function minimization by conjugate gradients. *Comput. J.* **7** 149–154. MR0187375 https://doi.org/10.1093/comjnl/7.2.149

FOSDICK, B. K. and RAFTERY, A. E. (2014). Regional probabilistic fertility forecasting by modeling between-country correlations. *Demogr. Res.* **30** 1011.

FRENI-STERRANTINO, A., VENTRUCCI, M. and RUE, H. (2018). A note on intrinsic conditional autoregressive models for disconnected graphs. *Spat. Spatio-Tempor. Epidemiol.* **26** 25–34.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

GAŁECKI, A. and BURZYKOWSKI, T. (2013). *Linear Mixed-Effects Models Using R: A Step-by-Step Approach*. *Springer Texts in Statistics*. Springer, New York. MR3024843 https://doi.org/10.1007/978-1-4614-3900-4

GALLOWAY, M. (2018). CVglasso: Lasso Penalized Precision Matrix Estimation. R package version 1.0.

GOLDFARB, D. (1970). A family of variable-metric methods derived by variational means. *Math. Comp.* **24** 23–26. MR0258249 https://doi.org/10.2307/2004873

GOLDFARB, D. and IDNANI, A. (1983a). A numerically stable dual method for solving strictly convex quadratic programs. *Math. Program.* **27** 1–33. MR0712108 https://doi.org/10.1007/BF02591962

GOLDFARB, D. and IDNANI, A. (2006). Dual and primal-dual methods for solving strictly convex quadratic programs. In *Numerical Analysis*: *Proceedings of the Third IIMAS Workshop Held at Cocoyoc*, *Mexico*, *January* 1981, pp. 226–239. Springer, Berlin.

HANNAN, E. J. (1973). The asymptotic theory of linear time-series models. *J. Appl. Probab.* **10** 130–145, corrections, ibid. 10 (1973), 913. MR0365960 https://doi.org/10.1017/s0021900200042145

HARSHMAN, R. A. and LUNDY, M. E. (1994). Parafac: Parallel factor analysis. *Comput. Statist. Data Anal.* **18** 39–72.

JOSSE, J. and HUSSON, F. (2016). missMDA: A package for handling missing values in multivariate data analysis. *J. Stat. Softw.* **70** 1–31.

KAROLYI, G. A. (1992). Predicting risk: Some new generalizations. *Management Science* **38** 57–74.

KYUNG, M. and GHOSH, S. K. (2010). Maximum likelihood estimation for directional conditionally autoregressive models. *J. Statist. Plann. Inference* **140** 3160–3179. MR2659845 https://doi.org/10.1016/j.jspi.2010.04.012

LEDOIT, O. and WOLF, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *J. Empir. Finance* **10** 603–621.

LEDOIT, O. and WOLF, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivariate Anal.* **88** 365–411. MR2026339 https://doi.org/10.1016/S0047-259X(03)00096-4

LEDOIT, O. and WOLF, M. (2022). The power of (non-) linear shrinking: A review and guide to covariance matrix estimation. *J. Financ. Econom.* **20** 187–218.

LEWANDOWSKI, D., KUROWICKA, D. and JOE, H. (2009). Generating random correlation matrices based on Vines and extended onion method. *J. Multivariate Anal.* **100** 1989–2001. MR2543081 https://doi.org/10.1016/j.jmva.2009.04.008

LIECHTY, J. C., LIECHTY, M. W. and MÜLLER, P. (2004). Bayesian correlation estimation. *Biometrika* **91** 1–14. MR2050456 https://doi.org/10.1093/biomet/91.1.1

LIU, H., WANG, L. and ZHAO, T. (2014). Sparse covariance matrix estimation with eigenvalue constraints. *J. Comput. Graph. Statist.* **23** 439–459. MR3215819 https://doi.org/10.1080/10618600.2013.782818

LONGFORD, N. T. and MUTHÉN, B. O. (1992). Factor analysis for clustered observations. *Psychometrika* **57** 581–597. MR1246752 https://doi.org/10.1007/BF02294421

LOPES, H. F., GAMERMAN, D. and SALAZAR, E. (2011). Generalized spatial dynamic factor models. *Comput. Statist. Data Anal.* **55** 1319–1330. MR2741417 https://doi.org/10.1016/j.csda.2010.09.020

LOPES, H. F., SALAZAR, E. and GAMERMAN, D. (2008). Spatial dynamic factor analysis. *Bayesian Anal.* **3** 759–792. MR2469799 https://doi.org/10.1214/08-BA329

LOPUHAÄ, H. P., GARES, V. and RUIZ-GAZEN, A. (2023). S-estimation in linear models with structured covariance matrices. *Ann. Statist.* **51** 2415–2439. MR4682703 https://doi.org/10.1214/23-aos2334

LYONS, R. (1988). Strong laws of large numbers for weakly correlated random variables. *Michigan Math. J.* **35** 353–359. MR0978305 https://doi.org/10.1307/mmj/1029003816

MACNAB, Y. C. (2011). On Gaussian Markov random fields and Bayesian disease mapping. *Stat. Methods Med. Res.* **20** 49–68. MR2767372 https://doi.org/10.1177/0962280210371561

MARTINI, J. W., CROSSA, J., TOLEDO, F. H. and CUEVAS, J. (2020). On Hadamard and Kronecker products in covariance structures for genotype× environment interaction. *The Plant Genome* **13** e20033.

MAYER, T. and ZIGNAGO, S. (2006). Notes on CEPII's distances measures. Electronic resource. https://mpra.ub.uni-muenchen.de/26469/1/MPRA_paper_26469.pdf.

METODIEV, M., PERROT-DOCKÈS, M., OUADAH, S., FOSDICK, B. K, ROBIN, S., LATOUCHE, P. and RAFTERY, A. E (2026). Supplement to "A structured estimator for large covariance matrices in the presence of pairwise and spatial covariates." https://doi.org/10.1214/26-AOAS2183SUPPA, https://doi.org/10.1214/26-AOAS2183SUPPB

METODIEV, M., PERROT-DOCKÈS, M. and ROBIN, S. (2025). scov: Structured Covariances Estimators for Pairwise and Spatial Covariates. R package version 2.0.0.

UNITED NATIONS (2010). *World Population Prospects: the 2010 Revision, Volume I: Comprehensive Tables*. New York, N.Y.

UNITED NATIONS (2024). *World Population Prospects: the 2024 Revision, Volume I: Comprehensive Tables*. New York, N.Y.

POURAHMADI, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika* **86** 677–690. MR1723786 https://doi.org/10.1093/biomet/86.3.677

POURAHMADI, M. (2011). Covariance estimation: The GLM and regularization perspectives. *Statist. Sci.* **26** 369–387. MR2917961 https://doi.org/10.1214/11-STS358

POURAHMADI, M. (2013). *High-Dimensional Covariance Estimation*. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR3235948 https://doi.org/10.1002/9781118573617

PRESTON, S. H., HEUVELINE, P. and GUILLOT, M. (2001). *Measuring and Modeling Population Processes*. Blackwell Publishers, Oxford, U.K.

QIAN, L. (2009). Bayesian semiparametric correlation models for longitudinal data with applications to an HIV/AIDS biomarker study Phd thesis, Univ. California. Available at https://www.proquest.com/docview/304854584?pq-origsite=gscholar&fromope%nview=true&sourcetype=Dissertations%20&%20Theses.

RAFTERY, A. E. (1995). Bayesian model selection in social research. *Sociol*. *Method*. 111–163.

RAFTERY, A., HOETING, J., VOLINSKY, C., PAINTER, I. and YEUNG, K. Y. (2013a). BMA: Bayesian model averaging. R package version 3.16.1.

RAFTERY, A. E., CHUNN, J. L., GERLAND, P. and ŠEVČÍKOVÁ, H. (2013b). Bayesian probabilistic projections of life expectancy for all countries. *Demography* **50** 777–801.

RAFTERY, A. E., LI, N., ŠEVČÍKOVÁ, H., GERLAND, P. and HEILIG, G. K. (2012). Bayesian probabilistic population projections for all countries. *Proc*. *Natl*. *Acad*. *Sci*. *USA* **109** 13915–13921.

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592. MR0455196 https://doi.org/10.1093/biomet/63.3.581

SEAMAN, S., GALATI, J., JACKSON, D. and CARLIN, J. (2013). What is meant by "missing at random"? *Statist*. *Sci*. **28** 257–268. MR3112409 https://doi.org/10.1214/13-sts415

SHANNO, D. F. (1970). Conditioning of quasi-Newton methods for function minimization. *Math*. *Comp*. **24** 647–656. MR0274029 https://doi.org/10.2307/2004840

SUN, Y., BABU, P. and PALOMAR, D. P. (2016). Robust estimation of structured covariance matrix for heavy-tailed elliptical distributions. *IEEE Trans*. *Signal Process*. **64** 3576–3590. MR3515702 https://doi.org/10.1109/TSP.2016.2546222

TASTU, J., PINSON, P. and MADSEN, H. (2013). Space-time scenarios of wind power generation produced using a Gaussian copula with parametrized precision matrix.

THORSON, J. T., SCHEUERELL, M. D., SHELTON, A. O., SEE, K. E., SKAUG, H. J. and KRISTENSEN, K. (2015). Spatial factor analysis: A new tool for estimating joint species distributions and correlations in species range. *Methods Ecol*. *Evol*. **6** 627–637.

TOKUDA, T., GOODRICH, B., VAN MECHELEN, I., GELMAN, A. and TUERLINCKX, F. (2011). Visualizing distributions of covariance matrices. Columbia Univ, New York. Tech. Rep, 18–18.

TURLACH, B. A. and WEINGESSEL, A. (2019). quadprog: Functions to solve quadratic programming problems. R package version 1.5-8.

VARIN, C., REID, N. and FIRTH, D. (2011). An overview of composite likelihood methods. *Statist*. *Sinica* **21** 5–42. MR2796852

VER HOEF, J. M., HANKS, E. M. and HOOTEN, M. B. (2018). On the relationship between conditional (CAR) and simultaneous (SAR) autoregressive models. *Spat*. *Stat*. **25** 68–85. MR3809256 https://doi.org/10.1016/j.spasta.2018.04.006

WALL, M. M. (2004). A close look at the spatial structure implied by the CAR and SAR models. *J*. *Statist*. *Plann*. *Inference* **121** 311–324. MR2038824 https://doi.org/10.1016/S0378-3758(03)00111-3

WANG, F. and WALL, M. M. (2003). Generalized common spatial factor model. *Biostatistics* **4** 569–582.

WEI, T. and SIMKO, V. (2021). R package 'corrplot': Visualization of a correlation matrix. (Version 0.92).

WEST, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. *Bayesian Statistics* **7** 733–742.

ZHOU, R. and PALOMAR, D. P. (2024). covFactorModel: Covariance matrix estimation via factor models. R package version 0.1.0.