*The Annals of Applied Statistics*
2018, Vol. 12, No. 2, 940–970
https://doi.org/10.1214/18-AOAS1175


# ESTIMATING LARGE CORRELATION MATRICES FOR INTERNATIONAL MIGRATION

BY JONATHAN J. AZOSE[*,†] AND ADRIAN E. RAFTERY[†]

*Pacific Northwest National Laboratory[*] and University of Washington[†]*

The United Nations is the major organization producing and regularly updating probabilistic population projections for all countries. International migration is a critical component of such projections, and between-country correlations are important for forecasts of regional aggregates. However, in the data we consider there are 200 countries and only 12 data points, each one corresponding to a five-year time period. Thus a $200 \times 200$ correlation matrix must be estimated on the basis of 12 data points. Using Pearson correlations produces many spurious correlations. We propose a maximum *a posteriori* estimator for the correlation matrix with an interpretable informative prior distribution. The prior serves to regularize the correlation matrix, shrinking *a priori* untrustworthy elements towards zero. Our estimated correlation structure improves projections of net migration for regional aggregates, producing narrower projections of migration for Africa as a whole and wider projections for Europe. A simulation study confirms that our estimator outperforms both the Pearson correlation matrix and a simple shrinkage estimator when estimating a sparse correlation matrix.

**1. Introduction.** International migration is a major contributor to population change, but is hard to project, making proper quantification of uncertainty especially important. Existing global models for migration are well calibrated marginally, that is, for individual countries [Azose and Raftery (2015)], but typically rely on an unrealistic modeling assumption that forecast errors are uncorrelated across countries. If correlations exist, but are not modeled, the resulting projections may still be well calibrated for countries individually, but can under or overestimate variance in projections of migration for regions that span multiple countries. We present a method for estimating a correlation matrix from a small number of data points that uses informative priors, shrinking elements of the correlation matrix which we expect *a priori* to be small. In applying this method to migration, we choose priors based on empirical evidence of nonzero correlations among classes of countries which are "close" to one another according to a variety of distance covariates. Our method improves projections of migration for regional aggregates while mitigating the issue of spurious correlations that arises from trying to estimate a large correlation matrix based on many short time series.

Received November 2017; revised April 2018.

*Key words and phrases.* Correlation estimation, international migration, maximum a posteriori estimation, high-dimension.

The primary challenge of this application is how to produce principled estimates of the correlation matrix which include both the empirical information from the data and external knowledge not captured in the data themselves. This challenge is the realm of Bayesian statistics, and indeed our estimates take the form of a *maximum a posteriori* (MAP) estimator, balancing a data-based likelihood against an informative prior based on external knowledge. The novelty of our method is the ability to express world knowledge in the form of a simple, interpretable prior distribution which is suitable for use in elicitation from domain experts without deep statistical expertise—namely, a prior placed directly on elements of the correlation matrix.

The idea of Bayesian estimation of covariance structure is by no means new. One popular existing method is the graphical lasso [Friedman, Hastie and Tibshirani (2008)], which is equivalent to a MAP estimate under a simple prior distribution on the inverse covariance matrix. Prior distributions can be placed on other transformations of the covariance matrix as well, including eigenvalue decompositions [Chi and Lange (2014)] or the covariance matrix itself [Bien and Tibshirani (2011)]. Although the existing methods are appropriate for many applications, they generally do not lend themselves well to incorporation of informative prior information, especially if that information is to be elicited from experts rather than selected empirically. This article's methodological contribution is to demonstrate a correlation estimator which admits an informative prior that is easily interpretable and consequently well suited to expert elicitation.

1.1. *Illustrative example*. In this section, we focus on six selected countries—Estonia, Latvia, Lithuania, South Africa, Zimbabwe, and Zambia—to highlight the need for regularization of the correlation matrix.

Migration rates in Estonia, Latvia, and Lithuania over the period from 1950 to 2010 look quite similar (top row of Figure 1). All three countries shared a spike in out-migration during the 1990–1995 time period, which appears as a large negative forecast error in a first-order autoregressive [AR(1)] model. This sudden jump in out-migration among the Baltic states shares a common cause, namely the fall of the Soviet Union, which both induced westward migration and prompted many ethnic Russians to move to Russia [Fassmann and Munz (1994), Okolski (1998)].

Meanwhile, several countries in Southern Africa also experienced big shifts in migration rates during the 1990–1995 time period (bottom row of Figure 1). From 1990 to 1995, South Africa received substantially more in-migration than it had in previous decades, while Zimbabwe and Zambia both switched from being net receivers of migrants to net senders. For these three countries, at least some of the change in migration was due to political shifts related to the end of South Africa's apartheid regime. For example, the number of legal entrants to South Africa who overstayed their visas grew dramatically during the 1990s, with many such entrants coming from other countries of the Southern African Development Community [Crush (1999)].

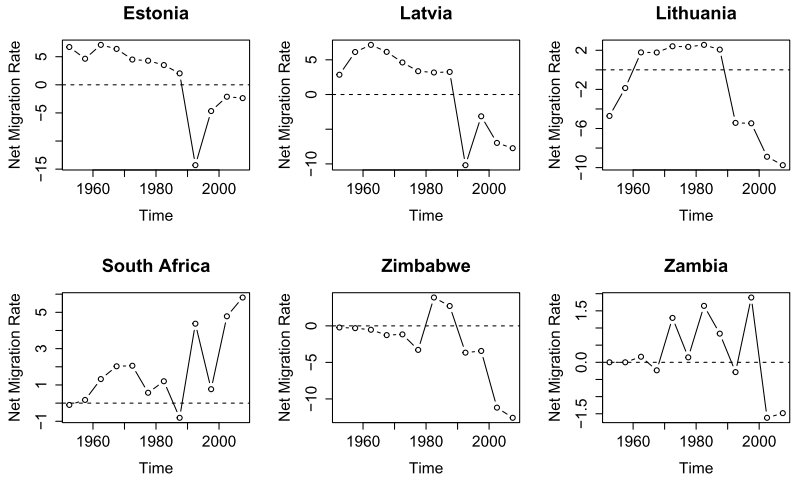


FIG. 1. *Net migration rates (net annual migrants per thousand individuals) for six countries.*

Because all six countries experienced pronounced changes in migration rates during the same time period, the usual Pearson estimates of the correlation in forecast errors are relatively large for these six countries (left panel of Figure 2). Knowledge of world affairs, however, suggests that some of these correlations may be spurious. There are plausible explanations for the correlations within the three Baltic nations and within the Southern African nations, but the cross-regional correlations are suspect. In fact, the cross-regional correlations seem to have arisen largely from a coincidental synchrony in the timing of different geopolitical events,

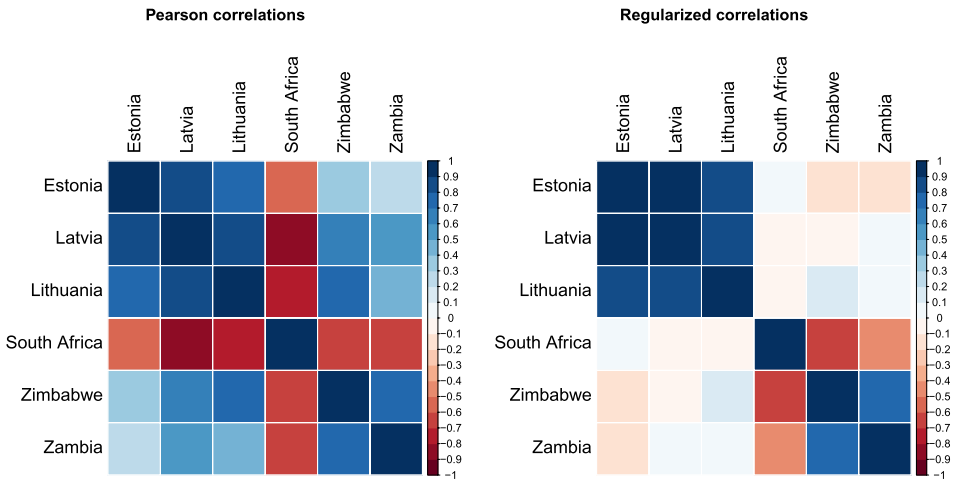


FIG. 2. *Estimated correlations among forecast errors for migration. The left panel shows the Pearson correlation estimates. The right panel shows our regularized estimates.*

and do not represent correlations that we would expect to continue to exist in future migration data. Our method is designed to shrink these seemingly spurious cross-regional correlations, producing the estimated correlation matrix shown in the right panel of Figure 2. Cross-regional correlations decrease substantially in magnitude, while correlations within regions remain largely unchanged.

Note that from the standpoint of interpretability, correlations are the most natural unit of analysis in this example. Elements of the inverse covariance matrix are interpretable as statements about conditional independence, which is not of primary interest here. Elements of the covariance matrix itself are difficult to parse, as they are not normalized for the variances of the associated countries. Various other transformations of the covariance matrix are available for mathematical analysis (e.g., the eigenvalue decomposition or the Cholesky decomposition), but their values are generally only interpretable by those with specialized knowledge of linear algebra. Consequently, it is most natural in this application to express prior beliefs directly on correlations. We provide the methodology for doing so in Section 2.

1.2. *Background*. Country-specific projections of international migration are an important input in policy-making decisions [Bijak et al. (2007), Brown and Bean (2012)]. Projected migration figures are commonly used in long-term planning of social welfare programs [U. S. Social Security Administration (2013), Wright (2010)]. However, projection of migration is difficult [Bijak and Wiśniowski (2010) describe migration as "barely predictable"] and global modeling of migration remains somewhat rudimentary. The United Nations Population Division produces global projections of fertility, mortality, and migration for all countries [United Nations (2012)]. For most countries, the 2012 revision of the World Population Prospects (WPP) deterministically projects net migration to persist at current levels until 2050 and decline linearly thereafter.

Reliable analysis of migration for multi-country regions is a topic of growing importance, as global migration governance begins to incorporate more multi-party policy agreements, in contrast to the largely unilateral and bilateral migration policies of the past. Europe has been at the forefront of implementing multi-lateral migration policies. Two such examples are the European Neighbourhood Policy, which establishes shared border management practices between the European Union and its neighbors in Eastern Europe and the Mediterranean [Barbé and Johansson-Nogués (2008)], and the Common European Asylum System, which ensures standard processes for asylum applicants across EU countries [Thielemann (2008)]. Furthermore, the International Organization for Migration (IOM) is increasingly advocating for international migration governance [International Organization for Migration (2015)], and was named in 2016 as a related organization to the United Nations [United Nations (2016)]. Greater support from the United Nations provides a possible approach to expanding multi-lateral migration treaties to more regions of the world.

However, migration policy at the level of multi-country regions faces several substantial hurdles before it can achieve greater scope. One issue, which we attempt to address in this paper, is that the state of quantitative knowledge of regional migration dynamics is weak in most regions. The ability to estimate the migration flow between any given pair of countries is a fairly recent innovation [Abel (2013)], and the methodology to produce such flow estimates relies on constraints which can obscure regional dynamics. A key goal of our work is to provide additional insight into migration dynamics for arbitrary collections of countries, which may be used to provide a quantitative basis for policy decisions. Estimates of between-country correlations in net migration and the resulting plausible range of values for net migration within any given group of countries may be used as a quantitative basis for both the range of likely net migration values, and the natural apportionment of that net migration (i.e., the expected distribution of net migration to all countries in the absence of new policies).

In addition to providing regional migration projections for any desired collection of countries, a second goal of our work is to improve the migration component of probabilistic population projections. To produce fully probabilistic population projections, one must incorporate probabilistic projections of fertility and mortality with a global probabilistic model of migration. It follows from the demographic balancing equation that the contribution of migration to population change is given by *net* migration (i.e., in-migration minus out-migration). Probabilistic models exist for both net migration [Azose and Raftery (2015), Azose, Ševčíková and Raftery (2016)] and in- and out-migration separately [Wiśniowski et al. (2015)]. Both of these models are Bayesian hierarchical autoregressive models which treat forecast errors in migration as independent across countries, conditional on model parameters. This leads to projections that are well calibrated for individual countries, but may not be for multi-country aggregates. Our method aims to relax this independence assumption.

It is worth noting that a strong correlation in migration rates themselves need not translate to a strong correlation in forecast errors. For example, from 1960 through 2000, Mexico was consistently either the largest or second-largest source of migration flows to the US, with nearly 5 million Mexicans or more migrating to the US during the 1990s [Abel (2013)]. While we estimate that net migration rates for the USA and Mexico have a correlation of $-0.56$ based on quinquennial WPP data from 1950–2010, we estimate a correlation in forecast errors of only $-0.07$. That is, most of the relationship between the USA and Mexico is already captured by the autoregressive model parameters, and the "random" components of migration rates for the two countries are nearly independent conditional on the AR(1) model.

In this high-dimensional setting with short time series, the empirical correlation matrix is a poor estimator, in that it can include many spuriously large estimated correlations. Our goal is to use regularization to improve an empirical correlation

matrix for forecast errors in migration. There is a large body of literature on regularized estimation of covariance matrices, with applications in genomics, image processing, and finance, among other fields [Fan, Han and Liu (2014)]. The novelty of our method is that it allows the incorporation of available prior information in an easily interpretable way.

Existing covariance estimators based on penalized likelihood maximization are typically maximum *a posteriori* (MAP) estimates under some prior distribution of covariance, but these formulations are not well suited to specifying beliefs directly about elements of the correlation matrix. Perhaps the most similar method to ours is that of Bien and Tibshirani (2011), which allows informative priors on elements of the covariance matrix rather than the correlation matrix. Their method is not directly applicable to our setting, as our goal is to augment existing marginal variances with a suitable correlation structure. Other proposed MAP estimators include the graphical lasso [Friedman, Hastie and Tibshirani (2008)], which can be used to place an informative prior on the inverse covariance, and the method of Chi and Lange (2014), which penalizes covariance estimates that have very large or very small eigenvalues. An extreme example is given by Chaudhuri, Drton and Richardson (2007), who provide a method for covariance estimation in the presence of known zeroes. Zhang and Zou (2014) propose a variant on penalized likelihood maximization that replaces the negative log likelihood with a simpler loss function.

A related class of covariance estimators relies on shrinkage of an empirical covariance matrix towards a simpler estimator, typically trading some bias for lower mean squared error [Ledoit and Wolf (2003, 2004, 2012)]. A strength of these methods is that as long as the empirical covariance matrix is positive semi-definite and the shrinkage target is positive definite, a linear combination of the two will naturally be positive definite. Applying a shrinkage method to the migration setting would be difficult, as the elements we would like to penalize do not define a positive definite shrinkage target.

A form of regularization that is straightforward to implement is applying thresholding directly to elements of a covariance or correlation matrix [Bickel and Levina (2008a), El Karoui (2008)]; these authors show that a hard-thresholded covariance matrix is consistent in operator norm. Generalized thresholding [Antoniadis and Fan (2001)], developed in the context of wavelet applications, provides a class of related regularized estimators. A key difficulty with such estimators is that care must be taken to ensure that the resulting estimator is positive definite. In some problems, this can be handled by selecting a thresholding constant from an appropriate range [Fan, Liao and Mincheva (2013)]. Unfortunately, such an approach is not easily adapted to our problem. The structure of the elements we wish to penalize is such that we can tolerate only a small amount of shrinkage of all penalized elements before our estimated correlation matrix loses positive definiteness.

One fully Bayesian treatment is proposed by Liechty, Liechty and Müller (2004), who include substantive prior information by specifying clusters of correlations which they expect to be similar. This is unfortunately unsuitable to our

setting, since geographical and cultural proximity can give rise to either positive or negative correlations. Huang and Wand (2013) describe a computationally attractive *noninformative* prior on covariances, which does not easily extend to the informative priors we would like to include. Other fully Bayesian treatments are given by Barnard, McCulloch and Meng (2000), who propose a prior on the correlation matrix that is either marginally or jointly uniform, and Leonard and Hsu (1992) and Deng and Tsui (2013), who propose Bayesian estimation of the logarithm of the covariance matrix, which is unfortunately hard to interpret.

In scenarios where there is a natural ordering to the variables, it is often reasonable to make the assumption that large values of $|i - j|$ imply near independence or conditional independence. When this is the case, one can regularize by banding or tapering of the covariance or inverse covariance matrix [Bickel and Levina (2008b), Chen, Xu and Wu (2013), Fan, Huang and Li (2007), Furrer and Bengtsson (2007), Levina, Rothman and Zhu (2008)]. These approaches are not suitable to our problem, as there is no natural ordering of countries.

Good overviews of other methods in covariance estimation are given by Fan, Liao and Liu (2016) and Pourahmadi (2011).

**2. Methods.** We start with an established, well calibrated autoregressive model on net migration rates for all countries [Azose and Raftery (2015)], where net migration in a five-year period is defined as number of in-migrants minus number of out-migrants per year per 1,000 population. This model has the form

$$\begin{aligned}
\boldsymbol{g}_t - \boldsymbol{\mu} &= \operatorname{diag}(\boldsymbol{\phi})(\boldsymbol{g}_{t-1} - \boldsymbol{\mu}) + \boldsymbol{\varepsilon}_t, \\
\boldsymbol{\varepsilon}_t &\overset{\text{iid}}{\sim} \mathcal{N}_C\big(\mathbf{0}, \operatorname{diag}(\boldsymbol{\sigma}) \cdot I_C \cdot \operatorname{diag}(\boldsymbol{\sigma})\big), \\
\phi_c &\overset{\text{iid}}{\sim} U(0, 1), \\
\mu_c &\overset{\text{iid}}{\sim} N\big(\lambda, \tau^2\big), \\
\sigma_c^2 &\overset{\text{iid}}{\sim} IG(a, b).
\end{aligned}$$

Notationally, $\boldsymbol{g}_t$ is a length-$C$ vector of net migration rates for all countries during the time period from $t$ to $t + 1$, where $C$ is the number of countries analyzed. The quantities $\boldsymbol{\mu}$, $\boldsymbol{\phi}$, and $\boldsymbol{\sigma}$ are vectors of model parameters, each of length $C$. The $c$th entry in each parameter vector (i.e., $\mu_c$, $\phi_c$, and $\sigma_c^2$) gives the scalar parameter value corresponding to country $c$, while $\mathbf{0}$ is a length-$C$ vector of zeroes. The distributions used are as follows: $U$ denotes a uniform distribution parametrized by its upper and lower bounds, $N$ denotes a univariate normal distribution parametrized by its mean and variance, $\mathcal{N}_C$ denotes a $C$-dimensional multi-variate normal parametrized by its mean vector and covariance matrix, while $IG$ denotes an inverse gamma distribution parametrized by its shape and scale. [We have omitted

the specifics of hyperpriors on $a$, $b$, $\lambda$, and $\tau$, which Azose and Raftery (2015) selected to reflect the ranges of plausible values.]

Notably, forecast errors in this model are treated as conditionally independent, given the model's other parameters. Our method augments this model with an estimated correlation structure. Although the present paper focuses on the migration context, the same technique could be applied to probabilistic models of other demographic indicators.

From this point forward, we refer to the model of Azose and Raftery (2015) as the Bayesian Hierarchical Model with Independent Forecast Errors (BHM+IFE). In principle, the methodology we describe here provides a means of estimating a correlation matrix to be adjoined to any probabilistic model with conditionally independent forecast errors.

The outline of our procedure for estimating a correlation matrix is as follows:

1. From the BHM+IFE model, draw a posterior sample of $m$ realizations of model parameters, $\boldsymbol{\mu}^{(1)}, \boldsymbol{\phi}^{(1)}, \boldsymbol{\sigma}^{(1)}, \ldots, \boldsymbol{\mu}^{(m)}, \boldsymbol{\phi}^{(m)}, \boldsymbol{\sigma}^{(m)}$.
2. Convert the estimated forecast errors from the posterior sample of model parameters to a single empirical correlation matrix, $\tilde{R}$.
3. Combine the empirical correlation matrix with informative priors on correlations to obtain a maximum *a posteriori* (MAP) correlation estimate, $\hat{R}$.

This procedure can be viewed as performing a single step of the Monte Carlo EM (MCEM) algorithm [Wei and Tanner (1990)].

The posterior sampling in stage 1 can be performed using any reasonable sampling procedure. In practice, we performed our posterior sampling with a combination of Gibbs sampling and Metropolis–Hastings steps.

In the following sections, we first discuss the details of obtaining a MAP estimator (Section 2.1), including an algorithm for computing this estimator. This is followed by a section discussing practical considerations (Section 2.2), including initialization and the selection of an empirical correlation matrix and a regularization parameter.

2.1. *A MAP estimator for correlation.* Our goal is to estimate the correlation structure, $R$, of forecast errors, $\boldsymbol{\varepsilon}_t$. We assume a model of the form

$$\boldsymbol{\varepsilon}_t \overset{\text{iid}}{\sim} \mathcal{N}_C(\mathbf{0}, \Sigma), \tag{1}$$

where the variance matrix, $\Sigma$, decomposes into a vector of standard deviations, $\boldsymbol{\sigma}$, and a correlation matrix, $R$, as $\Sigma = \operatorname{diag}(\boldsymbol{\sigma}) \cdot R \cdot \operatorname{diag}(\boldsymbol{\sigma})$.

2.1.1. *Expression for the MAP correlation estimator.* To determine a MAP estimator for $R$, we express the posterior distribution for $R$ as a product of likelihood and prior.

*Data likelihood.* Equation (1) implies a likelihood function for $R$ of the form

$$p(\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_{T-1} | R, \boldsymbol{\sigma}) \propto_R \det(R)^{-(T-1)/2} \times \exp\left( -\frac{1}{2} \sum_{t=1}^{T-1} \boldsymbol{\varepsilon}_t' \operatorname{diag}(\boldsymbol{\sigma})^{-1} R^{-1} \operatorname{diag}(\boldsymbol{\sigma})^{-1} \boldsymbol{\varepsilon}_t \right),$$

restricted to the space $\Omega$ of valid correlation matrices (i.e., positive semi-definite matrices with ones on the diagonal). Matrix trace identities simplify this likelihood to

$$p(\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_{T-1} | R, \boldsymbol{\sigma}) \propto_R \det(R)^{-(T-1)/2} \exp\left( -\frac{1}{2} \operatorname{tr}\big(R^{-1} \tilde{R}\big) \right),$$

where

$$\tilde{R} := \frac{1}{T-1} \sum_{t=1}^{T-1} \operatorname{diag}(\boldsymbol{\sigma})^{-1} \boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t' \operatorname{diag}(\boldsymbol{\sigma})^{-1}.$$

The evidence from the data is encapsulated in $\tilde{R}$, which is something akin to an empirical correlation matrix. Note that $\tilde{R}$ would be a sufficient statistic for $R$ if the $\boldsymbol{\varepsilon}_t$'s and $\boldsymbol{\sigma}$ were known. In fact, neither the $\boldsymbol{\varepsilon}_t$'s nor $\boldsymbol{\sigma}$ are known, and $\tilde{R}$ must be replaced with a sensible estimate in order to proceed. Details of the estimation of $\tilde{R}$ are given in Section 2.2.1.

*Prior.* Our choice of prior distribution on $R$ is motivated by a desire to incorporate informative prior beliefs about which country pairs are likely to be nearly uncorrelated. As such, we choose a prior of the form

$$\pi(R) \propto_R \prod_{0 \le i < j \le C} \exp\big(-\lambda P_{ij} |R_{ij}|\big),$$

again restricted to $\Omega$. The matrix $P$ with entries $P_{ij}$ is a penalty matrix that encodes the extent to which we believe that countries $i$ and $j$ may be correlated. In our application to migration, we constrain all the entries in $P$ to be equal to 0 or 1, although in general $P$ may be allowed to have arbitrary nonnegative entries. The parameter $\lambda$ is an overall regularization parameter that encodes how strongly we want to penalize correlations.

The key benefit of this prior is its ease of interpretability. Setting $P_{ij} = 1$ expresses a belief that $R_{ij}$ should be close to zero, with the strength of that belief controlled by $\lambda$. Setting $P_{ij} = 0$ implies that all values of $R_{ij}$ are equally believable, *a priori*. Other penalized likelihood estimators have been proposed, corresponding to MAP estimators under implied priors on precision [Friedman, Hastie and Tibshirani (2008)], covariance [Bien and Tibshirani (2011)], or eigenvalues of the covariance matrix [Chi and Lange (2014)]. None of these allow one to specify prior beliefs about correlations directly.

Note that under this specification, the prior distribution of the correlation $R_{ij}$ is either uniform or truncated Laplace conditionally on the rest of the correlation matrix, but marginal distributions will not be uniform or double exponential. Although it is possible to specify a marginally uniform prior on all elements of the correlation matrix [Barnard, McCulloch and Meng (2000)], we know of no way to specify a distribution that is marginally uniform for some elements and marginally peaked at zero for others.

Because the prior density is a product of Laplace densities on correlations, we will refer to our eventual correlation estimator as the LPoC (Laplace Prior on Correlations) estimator. Augmenting the BHM+IFE with the LPoC correlation estimate produces the BHM+LPoC model.

*Posterior.* Combining the likelihood and prior, we obtain the log posterior distribution for $R$, equal to

$$\log p(R|\boldsymbol{\varepsilon}_1,\ldots,\boldsymbol{\varepsilon}_{T-1},\boldsymbol{\sigma}) = \text{const.} - \frac{T-1}{2}\log\det(R) - \frac{T-1}{2}\operatorname{tr}\bigl(R^{-1}\tilde{R}\bigr) - \frac{\lambda}{2}\|P * R\|_1 + c(\boldsymbol{\varepsilon}_1,\ldots,\boldsymbol{\varepsilon}_{T-1},\boldsymbol{\sigma}),$$

where $*$ denotes elementwise matrix multiplication, and $\|\cdot\|_1$ gives the sum of the absolute value of the elements of a matrix. Thus, finding the MAP estimator for $R$ is equivalent to solving the minimization problem

$$\underset{R\in\Omega}{\text{Minimize}}\left\{\log\det(R) + \operatorname{tr}\bigl(R^{-1}\tilde{R}\bigr) + \frac{1}{T-1}\lambda\cdot\|P * R\|_1\right\}. \tag{2}$$

Algorithmic details of a numerical solution are given in Section 2.1.2.

Note that if the penalty parameter, $\lambda$, is zero, then this minimization problem yields the maximum likelihood estimator (MLE) of R conditional on $\boldsymbol{\sigma}$. As long as $\tilde{R}$ is itself positive definite, this MLE is just $\tilde{R}$, the empirical correlation matrix. Similarly, if $\lambda$ is held fixed as $T$ grows, the penalty term in (2) converges to zero and the LPoC estimator converges to $\tilde{R}$. As long as $\tilde{R}$ is consistent for $R$, the LPoC estimator is also consistent.

2.1.2. *Solving the minimization problem.* We apply a majorize-minimize algorithm similar to that used by Bien and Tibshirani (2011) to the minimization problem in (2). We establish a basic outline for this algorithm before providing full details.

The function being minimized over is the sum of a convex and a concave component. Broadly speaking, the majorize-minimize algorithm repeatedly iterates through the following steps:

1. Replace the concave component with its tangent plane to obtain a fully convex function.
2. Find the global minimum of the convex function from Step 1.
3. Return to Step 1, computing the new tangent plane at the current location.

Notationally, we label our starting point for this algorithm as $R_0$ and subsequent iterations of this majorize-minimize algorithm are denoted with subscripts $R_1, R_2, \ldots$.

In (2), the concave component is $\log\det(R)$, which we replace with the tangent plane $\log\det R_i + \operatorname{tr}(R_i^{-1}(R - R_i))$. After simplifying and removing terms which are constant in $R$, the convex minimization problem in the $i$th iteration of the algorithm is

$$\underset{R\in\Omega}{\text{Minimize}}\big\{\operatorname{tr}\big(R_i^{-1}R\big) + \operatorname{tr}\big(R^{-1}\tilde{R}\big) + \lambda\|P * R\|_1\big\}. \tag{3}$$

In this inner minimization problem (3), all of the terms in the objective function are convex, and all but $\lambda\|P * R\|_1$ are differentiable, so we can apply the generalized gradient descent algorithm [Beck and Teboulle (2009)]. Each generalized gradient descent step from initial location $R_{\text{old}}$ takes the form

$$\underset{\omega\in\Omega}{\operatorname{argmin}}\left\{(2t)^{-1}\|\omega - \big(R_{\text{old}} - t\big(R_i^{-1} - R_{\text{old}}^{-1}\tilde{R}R_{\text{old}}^{-1}\big)\big)\|_F^2 + \frac{\lambda}{T-1}\|P * \omega\|_1\right\}, \tag{4}$$

where $t$ is a step size parameter. Equation (4) is itself another minimization problem, but a tractable one, with one robust solution given by Cui, Leng and Sun (2016).

Thus, the complete algorithm for finding the MAP estimator is shown in Algorithms 1 and 2.

### 2.2. *Practical considerations.*

2.2.1. *Estimating* $\tilde{R}$. Since the forecast errors and model parameters of the BHM+IFE model are unknown, we do not have access to the true value of $\tilde{R}$. Instead we use an estimate of $\tilde{R}$. For practical reasons, we would prefer to have $\tilde{R}$ itself be a valid correlation matrix so that (2) will have a known analytic solution in the limiting scenarios where $T$ grows or $\lambda$ goes to zero. Accordingly, we might choose an estimator $\tilde{R}^{\text{basic}}$ with elements defined by

$$\tilde{R}_{ij}^{\text{basic}} := \frac{\sum_{t=1}^{T-1}\hat{\varepsilon}_{i,t}\hat{\varepsilon}_{j,t}}{\sqrt{\sum_{t=1}^{T-1}\hat{\varepsilon}_{i,t}^2}\sqrt{\sum_{t=1}^{T-1}\hat{\varepsilon}_{j,t}^2}},$$

**Algorithm 1:** Outer loop: Majorize-minimize algorithm to solve equation (2)

```
Initialize R_i at correlation matrix R_0 ∈ Ω
repeat
    Set R_{i+1} = argmin_{R∈Ω}{tr(R_i^{-1}R) + tr(R^{-1}R̃) + λ‖P * R‖_1}.
    // Apply Algorithm 2 to solve this inner
       minimization problem
until ‖R_{i+1} − R_i‖_∞ is sufficiently small
```

**Algorithm 2:** Inner loop: Generalized gradient descent to solve equation (3)

Initialize $R_{\text{old}}$ at correlation matrix $R_i \in \Omega$
**repeat**
  Propose an update $R_{\text{new}} =$
  $\text{argmin}_{\omega\in\Omega}\{(2t)^{-1}\|\omega-(R_{\text{old}}-t(R_i^{-1}-R_{\text{old}}^{-1}\tilde{R}R_{\text{old}}^{-1}))\|_F^2+\frac{\lambda}{T-1}\|P*\omega\|_1\}$.
  `// Appeal to Cui, Leng and Sun (2016) to solve this minimization problem`
  **if** *objective function in (3) is lower at* $R_{\text{new}}$ *than* $R_{\text{old}}$ **then**
    Set $R_{\text{old}} = R_{\text{new}}$
    Adjust step size according to procedure in Appendix A.
  **else**
    Adjust step size according to procedure in Appendix A.
  **end**
**until** *improvement in objective function from* $R_{\text{old}}$ *to* $R_{\text{new}}$ *is sufficiently small*

where $\hat{\boldsymbol{\varepsilon}}_t$ is the posterior mean of $\boldsymbol{\varepsilon}_t$ from the BHM+IFE model. This estimate, $\tilde{R}^{\text{basic}}$, is the MLE for estimating the correlation matrix of a multivariate normal random variable with mean known to be zero and unknown marginal variance terms. By construction, $\tilde{R}^{\text{basic}}$ is guaranteed to be positive semi-definite and to have ones on the diagonal.

However, in our application, $\tilde{R}^{\text{basic}}$ is of low rank, since $T$ is small relative to the dimension of the matrix. For computational reasons, we would prefer to have a strictly positive definite matrix, so we estimate $\tilde{R}$ by

$$\tilde{R}^{\text{PD}} = 0.99 \cdot \tilde{R}^{\text{basic}} + 0.01 \cdot I_C.$$

This change can be viewed as augmenting our estimates of $\boldsymbol{\varepsilon}_t$ with a small amount of additional uncorrelated data.

2.2.2. *Selecting the regularization parameter* $\lambda$. Although the penalty matrix $P$ can be selected on the basis of world knowledge, we are less likely to have genuine prior beliefs about the value of the regularization parameter $\lambda$. Accordingly, we need some procedure for selecting a value for $\lambda$. In regularization problems, it is common to select the regularization parameter via cross-validation [Bien and Tibshirani (2011), Chi and Lange (2014), Huang et al. (2006)]. This approach is too computationally intensive to be feasible for our application. Among shrinkage estimators, it is common to choose the amount of shrinkage in order to minimize an expected loss function [James and Stein (1961), Ledoit and Wolf (2003)]. However, no suitable analytic result exists that allows us to approximately minimize expected loss in our scenario.

Consequently, we developed a heuristic criterion that selects $\lambda$ in a way that aligns with the goal of our regularization process. Our method's intent is to shrink

the magnitude of penalized elements of the correlation matrix while leaving unpenalized elements more or less unchanged. In practice, although we succeed in shrinking penalized elements towards zero, this shrinkage usually comes at the cost of inflating other elements. We have observed that this inflation tends to grow more pronounced as $\lambda$ grows. For very large values of $\lambda$, our estimated correlation matrix may shrink nearly all penalized entries to zero at the expense of inflating a few elements (both penalized and unpenalized) to nearly $\pm 1$. This is not a desirable outcome.

Although it may seem counterintuitive at first, the observed inflation is not an artifact of a coding error or poor convergence of our algorithm. A simple reproducible example of inflation in a $3 \times 3$ matrix is provided in Appendix B. In this low-dimensional setting, standard numerical optimization routines agree with the results from our code and both display inflation of unpenalized elements.

Our criterion for selecting $\lambda$ compares the off-diagonal elements of $\tilde{R}$ and $\hat{R}(\lambda)$. We choose the value of $\lambda$ which maximizes the difference between average shrinkage and average inflation. Formally, our criterion is defined by

$$k(\tilde{R}, \lambda) = \underset{i,j \text{ s.t. } |\hat{R}(\lambda)_{ij}| < |\tilde{R}_{ij}|}{\text{mean}} \left(|\tilde{R}_{ij}| - |\hat{R}(\lambda)_{ij}|\right) - \underset{i,j \text{ s.t. } |\hat{R}(\lambda)_{ij}| > |\tilde{R}_{ij}|}{\text{mean}} \left(|\hat{R}(\lambda)_{ij}| - |\tilde{R}_{ij}|\right).$$

Large positive values of $k$ are desirable, as they correspond to values of $\lambda$ for which we induce a lot of shrinkage and not much inflation.

2.2.3. *Initialization*. Although our algorithm is guaranteed to find a locally optimal solution, there is no guarantee of finding a global optimum because the objective function is not convex. Because of computational limitations, we restrict ourselves to searching for a locally optimal solution near $\tilde{R}$, rather than performing a broader search of the parameter space for better local minima.

Initialization at $\tilde{R}$ is intuitively appealing, since $\tilde{R}$ is the known optimum in the unpenalized ($\lambda = 0$) case. When taken in concert with our procedure for selection of $\lambda$, it also suggests an iterative process for initialization. Our procedure for selecting $\lambda$ requires us to compute $\hat{R}(\lambda)$ for a range of $\lambda$ values. If we slowly increase $\lambda$ away from zero, we can typically select good initial values by using the estimates computed for previous values of $\lambda$. For instance, at $\lambda = 0$, we appeal to the known solution of $\hat{R}(\lambda = 0) = \tilde{R}$. At $\lambda = 0.1$, we start our search at $R_0 = \hat{R}(\lambda = 0) = \tilde{R}$. At $\lambda = 0.2$, we start our search at $R_0 = \hat{R}(\lambda = 0.1)$, and so on.

2.2.4. *Alternative solution to innermost minimization problem*. While Cui, Leng and Sun (2016) present a guaranteed solution to the minimization problem in equation (4), theirs is an iterative algorithm which can be slow under common circumstances. In practice, results in this paper have backed off to a simpler algorithm which may not find the optimum under some conditions.

We note that if the restriction to $\Omega$ were not present, equation (4) would have a simple analytic solution, given by

$$R_{\text{new}} = \mathcal{S}\left(R_{\text{old}} - t\left(R_i^{-1} - R_{\text{old}}^{-1}\tilde{R}R_{\text{old}}^{-1}\right), \frac{\lambda}{T-1}tP\right), \tag{5}$$

where $\mathcal{S}$ is the elementwise soft-thresholding operator defined by

$$\mathcal{S}(X,\alpha)_{ij} = \operatorname{sign}(X_{ij}) \cdot \left(|X_{ij}| - \alpha_{ij}\right) \cdot \mathbb{1}\left(|X_{ij}| > \alpha_{ij}\right).$$

(The updates are actually restricted to the off-diagonal elements only, as the diagonal elements of a correlation matrix are constrained to equal 1.) Thus, if there were no positive definiteness constraint, each update step would consist of a gradient descent step according to the gradient of the differentiable component followed by soft-thresholding the result.

Although we do have to satisfy a positive definiteness constraint, in our approximate algorithm we start by trying the update step in (5). If this update results in a valid correlation matrix, then that matrix is our solution to (4), and we take this as our proposed $R_{\text{new}}$. However, sometimes the soft-thresholded gradient step results in a matrix that is not positive definite. In that case, rather than applying an iterative solution, in practice we have simply reduced the step size and returned to the top of the loop.

We expect this approximation to have two major impacts. First, if the true local minimum is on the boundary of positive definite space, then our modified algorithm can converge to points which are not local minima. (In our application, we ruled out this scenario by confirming that all eigenvalues of our final solution were strictly positive.) Second, it typically forces small step sizes to avoid generalized gradient descent steps which would land outside $\Omega$. Because we are explicitly looking for a solution near $\tilde{R}$, and doing so by applying small, incremental increases to $\lambda$, we do not believe this to be an issue in our application. However, it would take a broader search of parameter space to conclusively state that this has not influenced our results.

In any case, this algorithmic modification is an approximation implemented for the sake of computational speed, and the full procedure is available for those with more computational power.

2.2.5. *Step size selection.* Step size selection has a large impact on performance and convergence of this algorithm. Details of step size selection are discussed in Appendix A.

**3. Results.** In this section, we first report results from applying our method to global migration data in Section 3.1. Section 3.2 then provides a simulation study which demonstrates that our method outperforms Pearson correlations and the Ledoit–Wolf shrinkage estimator [Ledoit and Wolf (2003)] in the situation where the penalty matrix $P$ is appropriate to the true correlation structure.

3.1. *Application to migration.*

3.1.1. *Data.* We use data on net migration from the 2012 revision of the World Population Prospects (WPP) [United Nations (2012)]. The WPP contains estimates of net migration for all countries in five-year time periods from 1950 until 2010, a total of 12 time periods. We compute the net migration rate $g_{c,t}$ as the net number of migrants in country $c$ over the five-year period starting at time $t$, divided by thousands of individuals in country $c$ at time $t$.

Because we want to express prior beliefs as a function of distance covariates, we restrict the set of modeled countries to the 191-country overlap between the WPP 2012 and the set of countries included in CEPII's GeoDist database, a database of bilateral distance covariates defined on pairs of countries [Mayer and Zignago (2011)].

3.1.2. *Selection of $P$.* Our estimation technique requires that we choose a penalty matrix, $P$, that reflects our prior beliefs about which country pairs are likely to be correlated. Although it would be possible to elicit expert opinion about each of the roughly 18,000 country pairs, we instead choose a $P$ that can be characterized in terms of just a few covariates. Our matrix $P$ penalizes a pair of countries if *none* of the following conditions is met:

1. The two countries are contiguous.
2. The two countries' most important cities are located less than 3000 km apart.
3. The two countries are in the same region according to the United Nations Population Division's division of the world into 22 regions, based on both geographical contiguity and cultural affinity [United Nations (2012)].
4. The two countries are currently in a colonial relationship.

This definition of $P$ is in line with migration theory, which suggests that migrant flows are more likely when monetary and social costs of movement are low [Harris and Todaro (1970), Lee (1966), Sjaastad (1962), Stark and Bloom (1985)], as will be the case with countries which are geographically proximate or share administrative ties. This definition penalizes 85% of country pairs, leaving 15% unpenalized. The average country is considered to be "close" to 29 other countries, and "distant" from the remaining 161.

In selecting these conditions, we examined nine candidate distance covariates. The first eight such covariates come from CEPII's GeoDist database [Mayer and Zignago (2011)], while the ninth is derived from the United Nations division into 22 regions. The left column of Table 1 gives the complete list of covariates considered. As an empirical basis for determining which criteria to include in defining our penalty matrix, we examined the elements of the sample correlation matrix for all pairs of countries meeting each criterion. Using a Kolmogorov–Smirnov test, we tested whether the distribution of these sample correlations was different from

TABLE 1

*Results of Kolmogorov–Smirnov test that empirical correlations are significantly different from the distribution of elements of a sample correlation matrix when the true error structure is uncorrelated. p-values lower than* 0.05 *are bolded*

| **Covariate** | ***p*-value** |
|---|---|
| Contiguous | **0.019** |
| Common language (official) | 0.23 |
| Common language (spoken by 9% of pop.) | 0.58 |
| Geodesic distance less than 3000 km | **0.0003** |
| Colonial relationship after 1945 | 0.57 |
| Common colonizer after 1945 | 0.11 |
| Current colonial relationship | **0.035** |
| Ever had a colonial link | 0.36 |
| Same UN Region | **0.036** |

the distribution of elements of the sample correlation matrix under a null hypothesis of uncorrelated errors. The right column of Table 1 shows the $p$-values from these Kolmogorov–Smirnov tests. Our definition of the penalty matrix $P$ includes all covariates with a $p$-value less than 0.05.

3.1.3. *Selection of the regularization parameter,* $\lambda$. We computed values of $\hat{R}(\lambda)$ for all values of $\lambda$ from 0 to 3 in increments of 0.1. Figure 3 shows the value of $k(\tilde{R}, \lambda)$ over a range of $\lambda$ values. We found that $k(\tilde{R}, \lambda)$ peaked at $\lambda = 0.6$, where we find average shrinkage of 0.13 compared with average inflation of 0.07. Increasing $\lambda$ from 0.6 to 0.7 induces additional shrinkage, but at the cost of greatly inflating some correlations. Accordingly, we choose $\hat{R}(0.6)$ as our estimate of $R$.

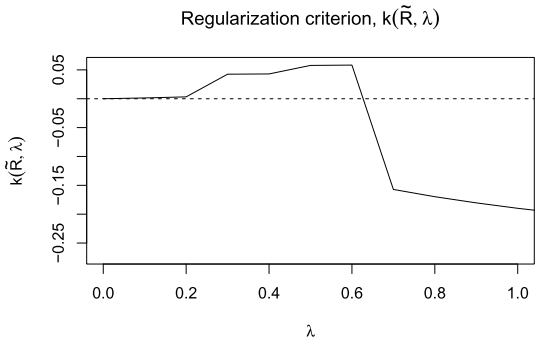


FIG. 3. *Regularization criterion,* $k(\tilde{R}, \lambda)$ *as a function of* $\lambda$. *The regularization criterion is the difference between the average shrinkage among shrunk elements of* $\hat{R}(\lambda)$ *and average inflation among inflated elements.*

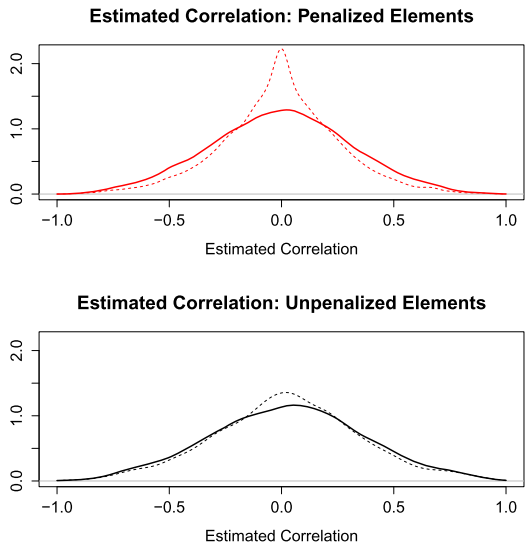


FIG. 4. *Comparison of elements of the correlation matrix before regularization* (*solid curves*) *and after* (*dashed curves*). *Top panel shows penalized elements*; *bottom panel shows unpenalized elements*.

Figure 4 shows the impact of regularization on the correlation matrix. Among penalized elements (top panel), we see substantial shrinkage towards zero, although many penalized elements remain large in magnitude, even after regularization. The bottom panel shows the unpenalized elements of the correlation matrix before regularization (solid curve) and after (dashed curve). On average we induce some shrinkage in the unpenalized elements, but the distribution is largely unchanged.

3.1.4. *Projection and evaluation*. We augment the BHM+IFE model with the LPoC estimate $\hat{R}(0.6)$ to produce probabilistic projections of migration for any collection of countries. Figure 5 contains medians and 80% prediction intervals of projected migration for all continents. In Africa, negative correlations narrow our projections. In Europe, positive correlations cause forecasts to widen. For the other continents, we see little change in projected migration.

For evaluation, we compare true migration rates for regional aggregates in 1995–2010 with projections of the same regional aggregates based only on migration data from 1950–1995. This procedure entails re-estimation of the BHM+IFE model using only the 1950–1995 data, followed by construction of an empirical correlation matrix, selection of $\lambda$, and extraction of $\hat{R}(\lambda)$. We compare the performance of the BHM+IFE model on regional aggregates to a model using the same sampled values of $\boldsymbol{\mu}$, $\boldsymbol{\phi}$, and $\boldsymbol{\sigma}$, but augmented with $\hat{R}(\lambda)$.

As an evaluation metric, we use the negatively oriented continuous ranked probability score (CRPS) [Gneiting and Raftery (2007), Hersbach (2000)]. The CRPS

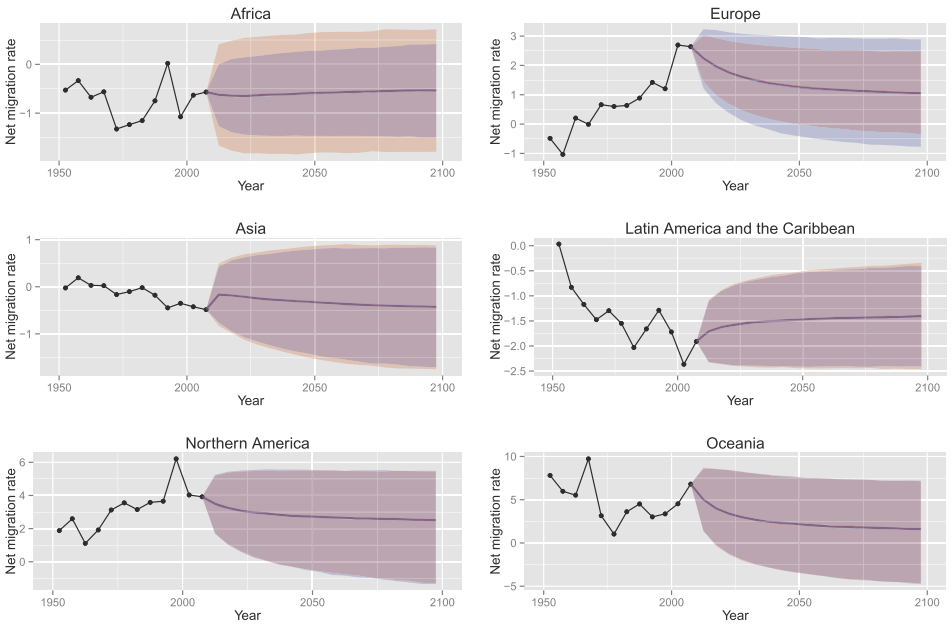


FIG. 5. *Medians and* 80% *prediction intervals for net migration among continents. Projections from the Bayesian hierarchical model with independent forecast errors* (*BHM+IFE*) *are given in red. Projections using our estimated correlation matrix* (*BHM+LPoC*) *are in blue. Overlap is in purple.*

compares the cumulative distribution function, $F$, of a probabilistic forecast to an observation, $x$, and is defined by

$$\mathrm{CRPS}(F, x) = \int_{-\infty}^{\infty} \big(F(y) - \mathbb{1}\{x \le y\}\big)^2 \, dy.$$

In our application, the two probabilistic forecasts under consideration have the same mean as one other, by design. One approximate way of looking at CRPS in this setting is that when $g_{c,t}$ is close to the mean of the forecast, we reward $F$ for having low variance; when $g_{c,t}$ is far from the mean, we reward $F$ for having high variance.

Table 2 gives CRPS for projections of aggregate migration for the six continents. Our model improves the quality of projections in Africa and Europe, while projections for the other four continents are more or less unchanged. Figure 6 illustrates the change in projections of net migration in 1995–2010 for four subregions of Africa and Europe. Projections from the BHM+IFE model are in red; projections from BHM+LPoC are in blue. Our method narrows prediction intervals in Eastern and Western Africa, bringing the width of the 80% prediction intervals more into line with the range of observed variability. In both regions, true migration rates for the projected period stayed within our narrower intervals. In contrast, our method

TABLE 2

*Continuous ranked probability score for all continents evaluated on projections of* 1995–2010, *where lower is better. Left column*: *Projections based on the Bayesian hierarchical model with independent correlation structure* (*BHM+IFE*). *Right column*: *Projections based on the Bayesian hierarchical model with our regularized correlation estimate* (*BHM+LPoC*). *Bolded entry in each row indicates the lower value*

| | **IFE** | **LPoC** |
|---|---|---|
| Africa | 1.66 | **1.49** |
| Asia | **0.73** | 0.74 |
| Europe | 3.92 | **3.76** |
| Latin America and the Caribbean | **1.62** | **1.62** |
| Northern America | 5.02 | **4.99** |
| Oceania | 8.53 | **8.49** |

widens projections in Northern and Western Europe, where the 80% intervals from the BHM+IFE model either miss or nearly miss capturing some of the observed data points.

This regional analysis suggests several policy considerations. For Europe as a whole and Northern and Western Europe in particular, our approach uncovers additional variability in the aggregated net migration rate which is not present in the model with independent forecast errors. If it is desirable to avoid periods of either extreme in- or out-migration, Europe could take measures to dampen this variance. Two practical options would be shared migration quotas to control in-migration or greater incentivizing of within-Europe migration during economic recessions.

For Africa, the situation is nearly the opposite. Many of the strongest negative correlations we observed can be traced back to within-region refugee flows. In Africa, at least, when a country undergoes a shock that results in a sudden spike in out-migration, that out-migration is likely to disperse in-migrants elsewhere in the region. A primary regional policy consideration should be planning for equitable refugee uptake and, if desirable, repatriation. Quantification of likely scenarios in the absence of new policy interventions can be guided by the trajectories from our model.

The few policy considerations discussed above are by no means exhaustive. A recent report by the IOM and McKinsey & Company [International Organization for Migration and McKinsey & Company (2018)] lays out a case for the many areas in which improvements in migration data would be valuable to sending countries, receiving countries, and migrants themselves. Many of their cited planning considerations would benefit from better understanding of correlations

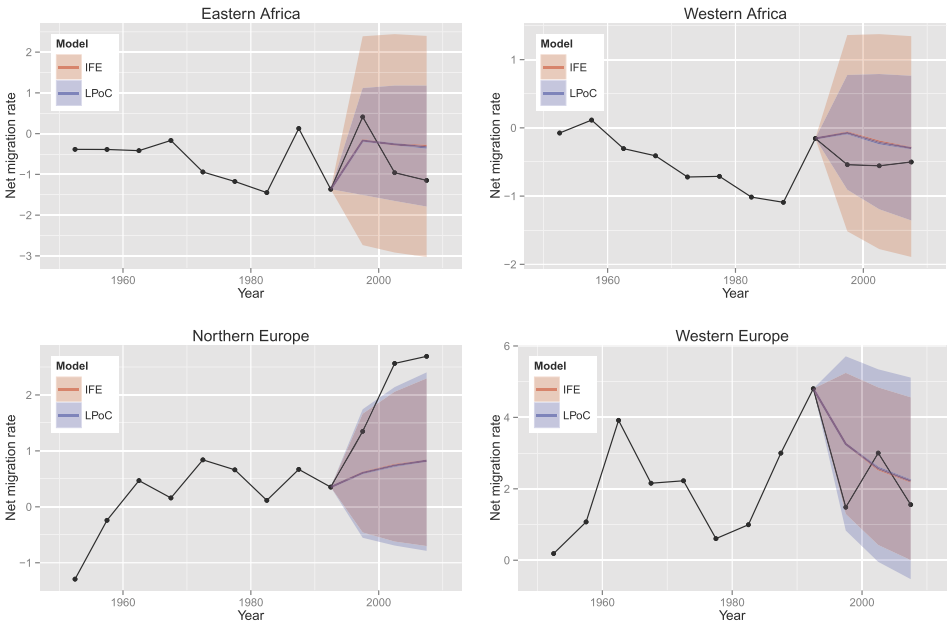


FIG. 6. *Medians and* 80% *prediction intervals for projections of net migration rates for regional aggregates. Projections from the Bayesian hierarchical model with independent forecast errors* (*BHM+IFE*) *are given in red. Projections using our estimated correlation matrix* (*BHM+LPoC*) *are in blue. Overlap is in purple.*

in migration. These considerations include setting policies to incentivize immigration among groups with relevant skills, forecasting loads on health and education systems in receiving countries (which may differ according to migrants' country of origin), and selecting appropriate locations to cater to basic needs of temporary refugees.

3.2. *Simulation study.* In this section, we show by simulation that our regularization procedure improves correlation estimates in a low-dimensional setting. To match the application of interest, we simulate 12 observed time points from an AR(1) process with correlated errors. For computational tractability, we decrease the number of simulated countries from 191 in the real data to 9 in the simulation. For each of 100 simulations, we perform the following procedure:

1. Generate a set of simulated migration rates $\boldsymbol{g}_1, \ldots, \boldsymbol{g}_{12}$ from an AR(1) process with errors correlated as described below.
2. Produce point estimates of $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_{11}$ via MCMC sampling of $\boldsymbol{\mu}$, $\boldsymbol{\phi}$, and $\boldsymbol{\sigma}$.
3. Convert $\boldsymbol{\varepsilon}_t$'s to a matrix $\tilde{R}$ using the procedure in Section 2.2.1.
4. Solve the minimization problem (2) to obtain a regularized estimate for the correlation matrix.

Since the procedure for selecting $\lambda$ is computationally intensive, we perform this procedure only once and use the same value of $\lambda$ for all subsequent simulations.

3.2.1. *Simulation details.* We simulate a collection of nine countries with true migration rates governed by the AR(1) process:

$$\boldsymbol{g}_t - \boldsymbol{\mu} = \operatorname{diag}(\boldsymbol{\phi})(\boldsymbol{g}_{t-1} - \boldsymbol{\mu}) + \boldsymbol{\varepsilon}_t.$$

For simplicity, we take $\boldsymbol{\mu} = \mathbf{0}$, $\boldsymbol{\phi} = \frac{1}{2}\mathbf{1}$, and

$$\boldsymbol{\varepsilon}_t \overset{\text{iid}}{\sim} \mathcal{N}_9(\mathbf{0}, \Sigma).$$

We fix $\Sigma$ to be block diagonal. The correlation structure within each $3 \times 3$ block is given by

$$\Sigma_{3\times 3} = \begin{pmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix},$$

and the full covariance matrix by

$$\Sigma = \begin{pmatrix} \Sigma_{3\times 3} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \Sigma_{3\times 3} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \Sigma_{3\times 3} \end{pmatrix}.$$

We then simulate observations $\boldsymbol{g}_1, \ldots, \boldsymbol{g}_{12}$ and attempt to make inference on the correlation structure of $\Sigma$.

Because we are basing inference on a small number of time points, Pearson estimates of correlation are highly variable. Solid curves in Figure 7 show the distributions of the off-diagonal elements of the unregularized Pearson correlation matrix in the ideal scenario where the values of $\boldsymbol{\varepsilon}_t$ can be perfectly estimated. The top panel shows the distribution of the elements for which the true correlation is zero. The bottom panel shows elements for which the true correlation is 0.5. In both cases, high variability makes inference difficult. Our method is designed to decrease variability among estimated correlations for those country pairs where prior knowledge suggests that the correlation should be close to zero.

To illustrate a best case scenario, we choose a penalty matrix $P$ which is well suited to the true correlation structure. The simplest such $P$ is the one which penalizes the off-diagonal elements of the correlation matrix if and only if the true correlation is zero, namely

$$P = \begin{pmatrix} \mathbf{0}_{3\times 3} & \mathbf{1}_{3\times 3} & \mathbf{1}_{3\times 3} \\ \mathbf{1}_{3\times 3} & \mathbf{0}_{3\times 3} & \mathbf{1}_{3\times 3} \\ \mathbf{1}_{3\times 3} & \mathbf{1}_{3\times 3} & \mathbf{0}_{3\times 3} \end{pmatrix}.$$

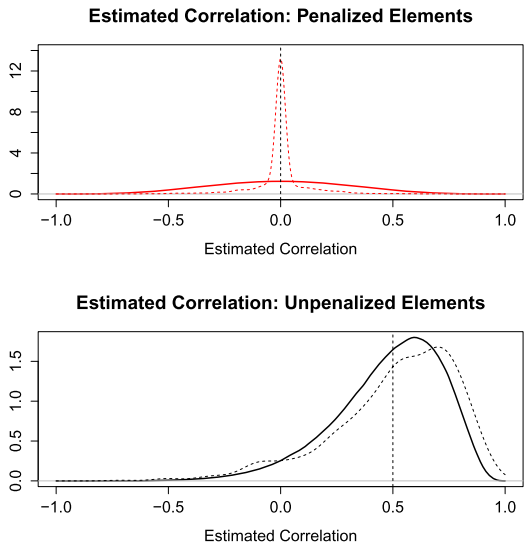


FIG. 7. *Simulation study results*: *Comparison of elements of the correlation matrix before regularization* (*solid curves*) *and after* (*dashed curves*). *Top panel shows penalized elements*; *bottom panel shows unpenalized elements*. *True correlations are indicated with dashed vertical lines*.

3.2.2. *Initial run to select* $\lambda$. Our procedure to select $\lambda$ is computationally expensive, as it requires us to compute $\hat{R}(\lambda)$ repeatedly as $\lambda$ varies. We therefore perform this procedure only once and use the same $\lambda$ for estimation of $R$ in all subsequent simulated data sets. Figure 8 plots our $\lambda$-selection criterion based on a single simulated data set over the range $\lambda = 0, 0.1, 0.2, \ldots, 10$. The exact curve, shown in black, exhibits some jumpiness in this low-dimensional setting, a problem which naturally becomes less severe in the high-dimensional setting of inter-

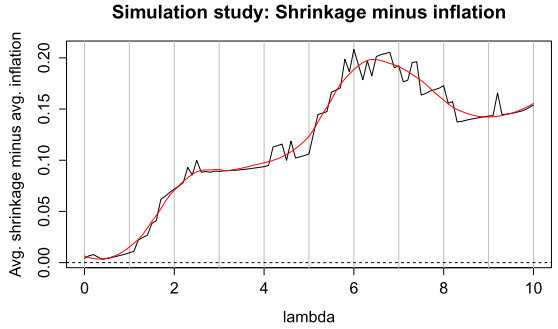


FIG. 8. *Average shrinkage minus average inflation of elements of* $\hat{R}(\lambda)$ *as* $\lambda$ *varies from* 0 *to* 10. *Exact curve in black*, *Lowess-smoothed curve in red*.

est. Because of this jumpiness, we base our selection of $\lambda$ on a Lowess-smoothed curve, selecting the maximizing value of $\lambda = 6.4$.

This curve suggests the possibility of another local or global maximum at a value higher than $\lambda = 10$. However, we have selected the first clear local maximum in order to mimic as well as possible the procedure performed in the high-dimensional setting. For matrices on the order of $200 \times 200$, as in our actual application, recomputing $\hat{R}(\lambda)$ for each new value of $\lambda$ can take on the order of hours of real time, depending on available hardware. Because of this, for the migration data we continued updating $\lambda$ until we felt sufficiently confident of having found a local optimum of this criterion, and then we stopped. There are no guarantees of global optimality surrounding this procedure; it is used for practical reasons, and we adopt it in the present simulation setting.

3.2.3. *Evaluation of repeated estimation of $R$.* We produced 100 estimates of $\hat{R}(\lambda = 6.4)$ from 100 different sets of simulated migration rates, all using the same block diagonal correlation structure. The dashed lines in Figure 7 show the distribution of off-diagonal elements of $\hat{R}$, split into those elements where the true correlation is 0 and elements where the true correlation is 0.5 (top and bottom panel, resp.).

Our method succeeded in shrinking penalized elements towards zero. Among elements where the true correlation is zero, we correctly estimated an exact zero in 62% of cases in this simulation. Among unpenalized elements, our method produced estimates with slightly more variability (the standard deviation was 0.256 for Pearson correlations versus 0.272 for our estimates). Both methods produced estimates for unpenalized elements that are within two standard errors of the true mean value of 0.5. The mean estimated correlation was 0.489 for Pearson correlations (standard error 0.009) versus 0.514 for our estimates (standard error of 0.009). On the whole, the LPoC estimator greatly improved estimates of penalized elements at the expense of slightly increasing variability in unpenalized elements.

Table 3 compares mean absolute error and mean squared error from our method with two competing estimators. We compare our results against both Pearson correlation matrices and correlation matrices that have been regularized using the Ledoit–Wolf method, which shrinks Pearson estimates towards a spherical correlation structure [Ledoit and Wolf (2003)]. In the top panel, we estimate $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_{11}$ with a Bayesian hierarchical model, as is done in our real application to migration.

In the bottom panel, we assume instead a scenario where we have direct access to $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_{11}$, as would be suitable in other applications where the interest is in estimating correlations of directly observed quantities. In both cases, our method provides an overall reduction in mean squared error by at least two thirds when compared against the Pearson sample correlation matrix. A large reduction in error from shrinking penalized elements is offset by a mild increase in error among unpenalized elements. We also outperform the Ledoit–Wolf estimator in terms of overall error.

TABLE 3

*Evaluation of correlation matrix estimates from simulation study. "LPoC" refers to our estimator, which uses Laplace priors on correlations. MAE is mean absolute error. MSE is mean squared error. Averages over "all elements" exclude diagonal elements, which are fixed at zero by all methods. The lowest (best) values are shown in bold*

| | **Estimator** | **MAE** | **MSE** |
|---|---|---|---|
| Values of $\boldsymbol{\varepsilon}_t$ estimated with MCMC | | | |
| All elements | Pearson | 0.253 | 0.098 |
| | Ledoit–Wolf | 0.193 | 0.055 |
| | LPoC | **0.090** | **0.028** |
| True correlation = 0 | Pearson | 0.270 | 0.109 |
| | Ledoit–Wolf | 0.190 | 0.053 |
| | LPoC | **0.049** | **0.012** |
| True correlation = 0.5 | Pearson | 0.201 | 0.066 |
| | Ledoit–Wolf | **0.200** | **0.060** |
| | LPoC | 0.214 | 0.074 |
| True values of $\boldsymbol{\varepsilon}_t$ used | | | |
| All elements | Pearson | 0.227 | 0.079 |
| | Ledoit–Wolf | 0.182 | 0.047 |
| | LPoC | **0.078** | **0.022** |
| True correlation = 0 | Pearson | 0.244 | 0.089 |
| | Ledoit–Wolf | 0.162 | 0.039 |
| | LPoC | **0.041** | **0.010** |
| True correlation = 0.5 | Pearson | **0.176** | **0.051** |
| | Ledoit–Wolf | 0.243 | 0.073 |
| | LPoC | 0.190 | 0.058 |

**4. Discussion.** Our method augments probabilistic projections of migration that are well calibrated for individual countries, with a correlation structure that reflects prior knowledge of between-country correlations. By combining a high-dimensional empirical correlation matrix with an informative prior that shrinks spurious correlations, we produce an estimated correlation matrix that is in line with migration theory and improves projections of regional aggregates. When compared with a simple model that assumes uncorrelated forecast errors, our method narrows projections of net migration for Africa and widens projections for Europe. Out-of-sample evaluation confirms that these changes produce better probabilistic forecasts as measured by continuous ranked probability score. Mechanically, the novelty of our method is our prior on correlations, which benefits from being interpretable and simple in form, and converts MAP estimation to an $\ell_1$-penalized regularization problem which is computationally tractable.

Our analysis focuses on modeling net migration, rather than immigration and emigration or a complete matrix of migration flows. Although net migration is sufficient for computing population change, it is not ideal in that it obscures the relationships between in- and out-migration [Rogers (1990)]. However, net migration is attractive in that it can be estimated using residual methods as long as good estimates are available for births, deaths, and population change. Indeed, this is typically the approach taken to produce the net migration estimates for the WPP, even in countries which produce official estimates of immigration and emigration [United Nations (2012)]. Emigration is known to be particularly difficult to estimate; de Beer et al. (2010) have documented pervasive under-counting in official estimates of emigration among European countries, theorizing that this is due to the difficulty of incentivizing individuals to report when they leave the country.

If sufficient data were to become available, an attractive alternative to our method would be to model a full matrix of bilateral migration flows. Such a model would naturally imply correlations in migration—if out-migrants from country $i$ tend to go to country $j$, then net migration in countries $i$ and $j$ will be negatively correlated. However, modeling the global bilateral flow matrix is currently not feasible. Abel (2013) produces global estimates of migration flows based on migrant stock data, but for only a small number of time periods at which migrant stock data exist. His method involves minimizing the total number of migrants subject to the available data on migrant stocks. This induces many structural zeroes in his estimates, making modeling difficult.

Although our method produces a MAP estimator in the presence of informative priors, we are not able to leverage the usual Bayesian machinery to produce a sample from the posterior distribution. While it would in theory be possible to use MCMC methods to produce a posterior sample by updating one element of the correlation matrix at a time, an updating procedure would need to iterate through some 18,000 elements of the correlation matrix, checking for positive definiteness after each proposed step. Such an algorithm is therefore likely to move around the parameter space too slowly to be of any use. In some settings, a Laplace approximation centered at the posterior mode can provide a good approximation of marginal posterior distributions [Tierney and Kadane (1986)]. However, the double-exponential priors in our setting render this procedure impracticable. Within each orthant of the parameter space, a quadratic approximation to the log likelihood is reasonable, but because of the $\ell_1$ penalty term, a different quadratic approximation would be required for each of the roughly $2^{18,000}$ orthants, which is not feasible.

Given our interest in combining data with prior beliefs, an inverse Wishart prior on covariance is tempting because it allows easy sampling from the full posterior. However, the inverse Wishart distribution is restrictive in form [Barnard, McCulloch and Meng (2000)] and does not provide a straightforward way to describe prior beliefs about correlations.

Another tempting alternative is that of Liu, Wang and Zhao (2014), who gave a simple thresholding method for producing a penalized correlation matrix that is guaranteed to be positive definite. Their estimator solves

$$\underset{\omega \succ \delta \cdot I}{\operatorname{argmin}} \frac{1}{2} \|\tilde{R} - \omega\|_F^2 + \lambda \|W * \omega\|_{1,\text{off}},$$

to produce an estimator among the set of valid correlation matrices with minimum eigenvalue no smaller than $\delta$. Although the weight matrix, $W$, is in principle arbitrary, they use $W$ to induce greater shrinkage where empirical correlations are weakest, not as a means of conveying prior information. We would be hesitant to replace $W$ with our penalty matrix $P$, as that use of their method would not incorporate prior information in a principled way.

In this work, we chose a fairly simple penalty matrix $P$, in which all entries were constrained to be either zero or one. Several straightforward generalizations of this penalty matrix are possible with only minimal methodological changes. First, the entries in $P$ can in general take any nonnegative values, reflecting prior beliefs about individual correlations which are allowed to vary in strength. Eliciting some 18,000 $P_{ij}$ values individually is not realistically feasible, but it may be possible to elicit expert priors in some useful parametric form, for example, taking $P_{ij}$ values to be the outputs of a linear regression, with experts expressing prior beliefs on the association between identified covariates and expected strength of correlations. Work on estimating migration flows within Europe [Raymer et al. (2013)] has made use of expert elicitation of informative priors on parameters in a migration model, and found it to be a practicable solution. Second, our method can be generalized to shrink estimated correlations towards nonzero values by replacing the penalty term $\lambda \|P * R\|_1$ with $\lambda \|P * (R - S)\|_1$ for some target matrix $S$. This may be desirable in cases where heavily structured estimates of correlations are available, as is the case for modeling of fertility [Fosdick and Raftery (2014)].

Note that we have used the 2012 revision of the WPP here [United Nations (2012)]. The more recent 2017 revision [United Nations (2017)] contains one additional data point. It would be of interest to redo the analysis with the newer data, but we expect the results would be similar.

## APPENDIX A: DETERMINING STEP SIZE

Step size selection is necessary in high dimensions for the general gradient descent algorithm to converge quickly enough to be useful. Complex methods for step size selection are available, but we obtained reasonable results with the backtracking line search algorithm, which starts with a large step size and decreases step size whenever a proposed step results in too little improvement in the objective function.

Say we have an objective function $f(x)$ which we are trying to minimize. The core of the backtracking line search algorithm is as follows [Nocedal and Wright (2006)]:

1. Fix a backtracking coefficient $\beta \in (0, 1)$, a starting step size $\alpha_0$, and a starting location $x_0$.

2. Propose a step of length $\alpha_k$ in direction $p_k$. (The backtracking line search algorithm is a generic algorithm that will work regardless of how the direction $p_k$ is determined.)

3. If the improvement in the objective function is enough to meet the Armijo condition given in (6) below, then take the proposed step, that is, take $x_{k+1} = x_k + \alpha_k p_k$. Keep the step size constant (i.e., $\alpha_{k+1} = \alpha_k$).

4. Otherwise, if there is any improvement in the objective function, take the proposed step, but also decrease the step size for the next iteration (specifically, set $\alpha_{k+1} = \beta\alpha_k$).

5. Otherwise, there must have been no improvement in the objective function. Do not take a step, but do decrease step size. ($x_{k+1} = x_k$ and $\alpha_{k+1} = \beta\alpha_k$.)

6. Repeat steps 2–5 until convergence.

The Armijo condition, which is used to determine whether to decrease step size, is as follows. The Armijo condition is met if the following inequality is satisfied:

$$f(x_k + \alpha_k p_k) \le f(x_k) + c_1\alpha_i \nabla f_k^T p_k. \tag{6}$$

[$c_1$ is a constant chosen from $(0, 1)$ that controls how strictly the change in $f$ must match the gradient at $x_k$.]

In our application, there is a missing component—we cannot actually compute the gradient of our objective function. The relevant objective function is given by

$$f(R) = \operatorname{tr}\big(R_i^{-1} R\big) + \operatorname{tr}\big(R^{-1}\tilde{R}\big) + \lambda \| P * R \|_1.$$

The first two terms in the sum are differentiable, but the third is not.

We rewrite the Armijo condition as

$$f(x_k + \alpha_k p_k) \le f(x_k) + c_1\alpha_i p_k^T p_k - c_1\alpha_i (p_k - \nabla f_k)^T p_k,$$

and then approximate $(p_k - \nabla f_k)$ by

$$-2 \cdot \nabla\big(\operatorname{tr}\big(R_i^{-1} R\big) + \operatorname{tr}\big(R^{-1}\tilde{R}\big)\big).$$

## APPENDIX B: INFLATION OF CORRELATION ESTIMATES

We provide here an example of our correlation estimation procedure which produces inflation in some unpenalized elements of the correlation matrix. We solved the minimization problem in (2) with three different methods, finding identical answers each time, up to small numerical tolerances. Those methods are:

1. Estimate $R$ using our code, which appeals to the generalized gradient descent algorithm.

2. Estimate $R$ using a black-box numerical optimization algorithm, which has access to the function we're minimizing, but not its derivative.

3. Estimate $R$ by finding an analytic expression for the gradient of the function we're minimizing, and solve for a point where the gradient is zero.

One case in which inflation manifests if we take our evidence from the data to be given by

$$\tilde{R} = \begin{pmatrix} 1 & 0.8 & 0.5 \\ 0.8 & 1 & 0.1 \\ 0.5 & 0.1 & 1 \end{pmatrix}$$

and the penalty matrix by

$$P = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

We denote the unknown true correlation matrix by

$$R = \begin{pmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_3 \\ \rho_2 & \rho_3 & 1 \end{pmatrix}.$$

We fix the regularization parameter at $\lambda = 0.5$. The problem is then to estimate the three parameters $\rho_1$, $\rho_2$, and $\rho_3$.

With all three methods we find an estimate of

$$\hat{\boldsymbol{\rho}} = \begin{pmatrix} \hat{\rho}_1 \\ \hat{\rho}_2 \\ \hat{\rho}_3 \end{pmatrix} = \begin{pmatrix} 0.8211 \\ 0.1542 \\ -0.1813 \end{pmatrix}.$$

Note that the second element, which is penalized, experiences shrinkage towards zero, as expected. The first element is inflated, while the third is both inflated and changes sign.

## REFERENCES


ABEL, G. (2013). Estimating global migration flow tables using place of birth data. *Demogr. Res.* **28** 505–546.

ANTONIADIS, A. and FAN, J. (2001). Regularization of wavelet approximations. *J. Amer. Statist. Assoc.* **96** 939–967. With discussion and a rejoinder by the authors. MR1946364

AZOSE, J. J. and RAFTERY, A. E. (2015). Bayesian probabilistic projection of international migration. *Demography* **52** 1627–1650.

AZOSE, J. J., ŠEVČÍKOVÁ, H. and RAFTERY, A. E. (2016). Probabilistic population projections with migration uncertainty. *Proc. Natl. Acad. Sci. USA* **113** 6460–6465.

BARBÉ, E. and JOHANSSON-NOGUÉS, E. (2008). The EU as a modest 'force for good': The European Neighbourhood Policy. *Int. Aff.* **84** 81–96.

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

BECK, A. and TEBOULLE, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.* **2** 183–202. MR2486527

BICKEL, P. J. and LEVINA, E. (2008a). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604. MR2485008

BICKEL, P. J. and LEVINA, E. (2008b). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199–227. MR2387969

BIEN, J. and TIBSHIRANI, R. J. (2011). Sparse estimation of a covariance matrix. *Biometrika* **98** 807–820. MR2860325

BIJAK, J. and WIŚNIOWSKI, A. (2010). Bayesian forecasting of immigration to selected European countries by using expert knowledge. *J. Roy. Statist. Soc. Ser. A* **173** 775–796. MR2759965

BIJAK, J., KUPISZEWSKA, D., KUPISZEWSKI, M., SACZUK, K. and KICINGER, A. (2007). Population and labour force projections for 27 European countries, 2002–2052: Impact of international migration on population ageing. *Eur. J. Popul.* **23** 1–31.

BROWN, S. K. and BEAN, F. D. (2012). Population growth. In *Debates on U.S. Immigration* (J. Gans, E. M. Replogle and D. J. Tichenor, eds.). SAGE, Thousand Oaks, CA.

CHAUDHURI, S., DRTON, M. and RICHARDSON, T. S. (2007). Estimation of a covariance matrix with zeros. *Biometrika* **94** 199–216. MR2307904

CHEN, X., XU, M. and WU, W. B. (2013). Covariance and precision matrix estimation for high-dimensional time series. *Ann. Statist.* **41** 2994–3021. MR3161455

CHI, E. C. and LANGE, K. (2014). Stable estimation of a covariance matrix guided by nuclear norm penalties. *Comput. Statist. Data Anal.* **80** 117–128. MR3240481

CRUSH, J. (1999). Fortress South Africa and the deconstruction of apartheid's migration regime. *Geoforum* **30** 1–11.

CUI, Y., LENG, C. and SUN, D. (2016). Sparse estimation of high-dimensional correlation matrices. *Comput. Statist. Data Anal.* **93** 390–403. MR3406221

DE BEER, J., RAYMER, J., VAN DER ERF, R. and VAN WISSEN, L. (2010). Overcoming the problems of inconsistent international migration data: A new method applied to flows in Europe. *Eur. J. Popul.* **26** 459–481.

DENG, X. and TSUI, K.-W. (2013). Penalized covariance matrix estimation using a matrix-logarithm transformation. *J. Comput. Graph. Statist.* **22** 494–512. MR3173726

EL KAROUI, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann. Statist.* **36** 2717–2756. MR2485011

FAN, J., HAN, F. and LIU, H. (2014). Challenges of big data analysis. *Nat. Sci. Rev.* **1** 293–314.

FAN, J., HUANG, T. and LI, R. (2007). Analysis of longitudinal data with semiparametric estimation of convariance function. *J. Amer. Statist. Assoc.* **102** 632–641. MR2370857

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

FAN, J., LIAO, Y. and MINCHEVA, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 603–680. With 33 discussions by 57 authors and a reply by Fan, Liao and Mincheva. MR3091653

FASSMANN, H. and MUNZ, R. (1994). European East–West migration, 1945–1992. *Int. Migr. Rev.* **28** 520–538.

FOSDICK, B. K. and RAFTERY, A. E. (2014). Regional probabilistic fertility forecasting by modeling between-country correlations. *Demogr. Res.* **30** 1011–1034.

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

FURRER, R. and BENGTSSON, T. (2007). Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. *J. Multivariate Anal.* **98** 227–255. MR2301751

GNEITING, T. and RAFTERY, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** 359–378. MR2345548

HARRIS, J. R. and TODARO, M. P. (1970). Migration, unemployment and development: A two-sector analysis. *Am. Econ. Rev.* **60** 126–142.

HERSBACH, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather Forecast.* **15** 559–570.

HUANG, A. and WAND, M. P. (2013). Simple marginally noninformative prior distributions for covariance matrices. *Bayesian Anal.* **8** 439–451. MR3066948

HUANG, J. Z., LIU, N., POURAHMADI, M. and LIU, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika* **93** 85–98. MR2277742

INTERNATIONAL ORGANIZATION FOR MIGRATION (2015). *Migration Governance Framework* (*C*/106/40). International Organization for Migration, Geneva. Available at https://governingbodies.iom.int/system/files/en/council/106/C-106-40-Migration-Governance-Framework.pdf.

INTERNATIONAL ORGANIZATION FOR MIGRATION and MCKINSEY & COMPANY (2018). *More than Numbers*: *How Migration Data Can Deliver Real-Life Benefits for Migrants and Governments*. International Organization for Migration, Geneva. Available at https://publications.iom.int/system/files/pdf/more_than_numbers.pdf.

JAMES, W. and STEIN, C. (1961). Estimation with quadratic loss. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I* 361–379. Univ. California Press, Berkeley, CA. MR0133191

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

LEDOIT, O. and WOLF, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Ann. Statist.* **40** 1024–1060. MR2985942

LEE, E. S. (1966). A theory of migration. *Demography* **3** 47–57.

LEONARD, T. and HSU, J. S. J. (1992). Bayesian inference for a covariance matrix. *Ann. Statist.* **20** 1669–1696. MR1193308

LEVINA, E., ROTHMAN, A. and ZHU, J. (2008). Sparse estimation of large covariance matrices via a nested Lasso penalty. *Ann. Appl. Stat.* **2** 245–263. MR2415602

LIECHTY, J. C., LIECHTY, M. W. and MÜLLER, P. (2004). Bayesian correlation estimation. *Biometrika* **91** 1–14. MR2050456

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

MAYER, T. and ZIGNAGO, S. (2011). Notes on CEPII's distances measures: The GeoDist database.

NOCEDAL, J. and WRIGHT, S. J. (2006). *Numerical Optimization*, 2nd ed. Springer, New York. MR2244940

OKOLSKI, M. Regional dimension of international migration in Central and Eastern Europe. *Genus* **54** 11–36.

POURAHMADI, M. (2011). Covariance estimation: The GLM and regularization perspectives. *Statist. Sci.* **26** 369–387. MR2917961

RAYMER, J., WIŚNIOWSKI, A., FORSTER, J. J., SMITH, P. W. F. and BIJAK, J. (2013). Integrated modeling of European migration. *J. Amer. Statist. Assoc.* **108** 801–819. MR3174664

ROGERS, A. (1990). Requiem for the net migrant. *Geogr. Anal.* **22** 283–300.

SJAASTAD, L. A. (1962). The costs and returns of human migration. *J. Polit. Econ.* **70** 80–93.

STARK, O. and BLOOM, D. E. (1985). The new economics of labor migration. *Am. Econ. Rev.* **75** 173–178.

THIELEMANN, E. (2008). The future of the common European asylum system. *Eur. Policy Anal.* **1** 1–8.

TIERNEY, L. and KADANE, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc.* **81** 82–86. MR0830567

U. S. SOCIAL SECURITY ADMINISTRATION (2013). The 2013 Annual Report of the Board of Trustees of the Federal Old-age and Survivors Insurance and Federal Disability Insurance Trust Funds. Board of Trustees, Federal Old-Age and Survivors Insurance and Federal Disability Insurance Trust Funds.

UNITED NATIONS (2012). *World Population Prospects*: *The* 2012 *Revision*. United Nations, New York.

UNITED NATIONS (2016). *Agreement Concerning the Relationship Between the United Nations and the International Organization for Migration* (*A/RES/*70/976). United Nations, New York. Available at https://digitallibrary.un.org/record/837208/files/A_RES_70_296-EN.pdf.

UNITED NATIONS (2017). *World Population Prospects*: *The* 2017 *Revision*. United Nations, New York.

WEI, G. C. and TANNER, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J. Amer. Statist. Assoc*. **85** 699–704.

WIŚNIOWSKI, A., SMITH, P. W., BIJAK, J., RAYMER, J. and FORSTER, J. J. (2015). Bayesian population forecasting: Extending the Lee–Carter method. *Demography* **52** 1035–1059.

WRIGHT, E. (2010). 2008-based national population projections for the United Kingdom and constituent countries. *Popul. Trends* **139** 91–114.

ZHANG, T. and ZOU, H. (2014). Sparse precision matrix estimation via lasso penalized D-trace loss. *Biometrika* **101** 103–120. MR3180660

PACIFIC NORTHWEST NATIONAL LABORATORY
1100 DEXTER AVENUE N
SUITE 500
SEATTLE, WASHINGTON 98109
USA
AND
DEPARTMENT OF STATISTICS
UNIVERSITY OF WASHINGTON
BOX 354322
SEATTLE, WASHINGTON 98195-4322
USA
E-MAIL: jonathan.azose@pnnl.gov

DEPARTMENT OF STATISTICS
UNIVERSITY OF WASHINGTON
BOX 354322
SEATTLE, WASHINGTON 98195-4322
USA
E-MAIL: raftery@u.washington.edu