Analysing international migration flows: a Bayesian network approach

Federico Castelletti and Emanuela Furfaro

Abstract In this paper we investigate the non-economic determinants of international migration flows. We approach this problem using a Bayesian graphical model methodology which performs variable selection among a set of covariates that are potentially related to migration flows. We consider inflows to Italy from 171 different countries of origin on which we measure demographic and geographic characteristics. While results are coherent with the most recent literature, our method also provides measures of uncertainty around the dependence structure between covariates and migration flows.

Key words: migration flows, model selection, Bayesian networks.

1 Introduction

In the last decades, international migrations towards developed countries have significantly grown thus making prediction of migratory flows a non-negligible component in population size projections [1]. Determinants of international migrations firstly include economic and political issues. Coherently, most of the recent literature has focused on the effects of economic and political variables [4], while only few studies have also accounted for non-economic (e.g. demographic) factors [5]. However, demographic variables are quite easy to predict and thus an extended analysis that also incorporates such characteristics can significantly improve the prediction of international migrations [5].

Federico Castelletti
Dipartimento di Scienze Statistiche, Largo Gemelli 1, Milano
e-mail: federico.castelletti@unicatt.it

Emanuela Furfaro
Dipartimento di Scienze Statistiche, Largo Gemelli 1, Milano
e-mail: emanuela.furfaro@unicatt.it

In this paper we investigate the relationship between non-economic variables and international migrations using a graphical model based approach. Graphical models represent a promising and effective tool for discovering dependence relationships among potentially many variables. From a statistical point of view, we consider a problem of covariate selection for a response variable inferring a graphical structure which jointly models the dependence relationships within covariates and between covariates and response.

2 Methods

We first introduce some general notation. A graph $\mathscr{G}$ is a pair (V,E) where $V = \{1,\ldots,p\}$ is a set of vertices (or nodes) and $E \subseteq V \times V$ a set of edges. Let $u,v \in V$, $u \neq v$. If $(u,v) \in E$ and $(v,u) \notin E$ we say that $\mathscr{G}$ contains the directed edge $u \to v$. If instead $(u,v) \in E$ and $(v,u) \in E$ we say that $\mathscr{G}$ contains the undirected edge $u - v$. Two vertices u,v are adjacent if they are connected by an edge (directed or undirected). For any pair of distinct nodes $u,v \in V$, we say that u is a *parent* of v if $u \to v$. Conversely, we say that v is a *son* of u. The set of all parents of u in $\mathscr{G}$ is denoted by $\mathrm{pa}_{\mathscr{G}}(u)$. A graph is called *directed* (*undirected*) if it contains only directed (undirected) edges. A directed graph is called Directed Acyclic Graph (DAG for short, denoted by $\mathscr{D}$) if it does not contains cycles, that is a sequence of edges $v_1 \to v_2 \to \ldots \to v_k$ such that $v_1 \equiv v_k$. A particular class of undirected graphs is represented by *decomposable* graphs, also called *chordal* or *triangulated*. An undirected graph is decomposable if every path of length $l \geq 4$ contains a chord, that is two non-consecutive adjacent vertices [6]; see for instance graph $\mathscr{G}$ in Figure 1. A graph encode a set of (marginal and) conditional independencies which determines its Markov property and can be read off from the graph itself using the notion of *d-separation* [7]. Moreover, we say that two graphs are *Markov equivalent* if and only if they encode the same conditional independencies. Markov equivalent graphs are not distinguishable in the presence of observational data only (in other terms they are "score equivalent"). Most importantly, for each decomposable undirected graph $\mathscr{G}$ we can find a perfect directed version, $\mathscr{G}^{<}$ (a DAG), which is Markov equivalent to $\mathscr{G}$ [6]; see also Figure 1.

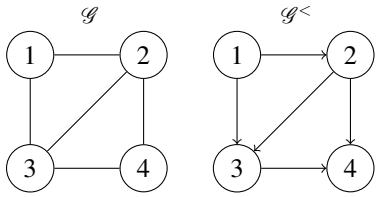

Fig. 1: A decomposable undirected graph $\mathscr{G}$ and its perfect directed version $\mathscr{G}^{<}$.

Let now Y be a response variable, $X_1,\ldots,X_q$ a collection of covariates. Each variable (both the response and covariates) can be associated to a node in a graph, whose structure will constrain the sampling distribution of the data. We are interested in selecting which covariates *directly* affect the response. In addition we allow for the presence of a dependence structure among covariates that we model by means of an undirected decomposable graph $\mathscr{G}_x = (V_x, E_x)$ (within covariates graphical structure) where $V_x = \{x_1,\ldots,x_q\}$ and $E_x \subseteq V_x \times V_x$ is the set of (undirected) edges between covariates. We also denote with $\mathscr{G}_x^<$ the perfect directed version of $\mathscr{G}_x$. Dependence relationships between covariates and response are instead represented by a directed graph $\mathscr{D}_{y|x} = (V_{xy}, E_{y|x})$ where $V_{xy} = \{x_1,\ldots,x_q,y\}$ and $E_{y|x} \subseteq \{x_1,\ldots,x_q\} \times y$. Consequently, in $\mathscr{D}_{y|x}$ we allow for the presence of directed edges *from* the covariates *to* the response only. The entire graphical structure, that we call *regression DAG*, is finally determined by the union of $\mathscr{G}_x^<$ and $\mathscr{D}_{y|x}$ and is denoted by $\mathscr{D}_{xy}$ (or simply $\mathscr{D}$ as in the sequel); see for instance Figure 2. For a given DAG $\mathscr{D}$ we can write the factorization

$$f(x_1,\ldots,x_q,y) = \prod_{j=1}^{q} f\left(x_j \,|\, \boldsymbol{x}_{\mathrm{pa}_{\mathscr{D}}(j)}\right) \cdot f\left(y \,|\, \boldsymbol{x}_{\mathrm{pa}_{\mathscr{D}}(y)}\right), \tag{1}$$

where $\mathrm{pa}_{\mathscr{D}}(j)$ is the set of parents of node j in $\mathscr{D}$.

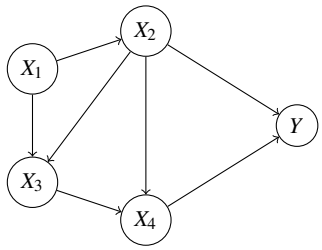

Fig. 2: A regression DAG with $q = 4$ covariates $X_1,\ldots,X_4$; the within covariates graphical structure corresponds to the perfect directed version of the decomposable graph $\mathscr{G}$ in Figure 1.

In a Gaussian framework we now assume $X_1,\ldots,X_q,Y \,|\, \Omega_{\mathscr{D}} \sim \mathscr{N}(\mathbf{0}, \Omega_{\mathscr{D}}^{-1})$, where $\Omega_{\mathscr{D}}^{-1}$ is the precision matrix (inverse of the covariance matrix $\Sigma_{\mathscr{D}}$) Markov w.r.t. DAG $\mathscr{D}$ and hence *constrained* by $\mathscr{D}$. Let also $\boldsymbol{y}$ be a $(n,1)$ vector of observations from the response Y, $\boldsymbol{X}$ a (n,q) data matrix collecting the observations from the q covariates. We denote with $\mathscr{S}$ the set of all regression DAGs on $q+1$ nodes which will represent our model space. The objective is then to perform model selection within the space of regression DAG models given the observed data $(\boldsymbol{X},\boldsymbol{y})$; we approach such problem adopting a Bayesian methodology. Specifically, let $f(\boldsymbol{y},\boldsymbol{X} \,|\, \Omega_{\mathscr{D}})$ be the likelihood function, $p(\Omega_{\mathscr{D}})$ a prior assigned to the DAG model parameter $\Omega_{\mathscr{D}}$. We are interested in evaluating the marginal likelihood $m(\boldsymbol{y},\boldsymbol{X} \,|\, \mathscr{D})$ of a generic $\mathscr{D} \in \mathscr{S}$ which from a Bayesian perspective represents the

score assigned to model $\mathscr{D}$,

$$m(\boldsymbol{y},\boldsymbol{X}\,|\,\mathscr{D}) = \int f(\boldsymbol{y},\boldsymbol{X}\,|\,\Omega_{\mathscr{D}})p(\Omega_{\mathscr{D}})\,d\Omega_{\mathscr{D}}. \tag{2}$$

To this end we rely on the *objective Bayes* method of [3] who derive a closed formula for $m(\boldsymbol{y},\boldsymbol{X}\,|\,\mathscr{D})$. Let now $p(\mathscr{D})$ be a prior assigned to $\mathscr{D}$. Bayesian prior-to-posterior analysis amounts to evaluate the posterior probability of $\mathscr{D}$ given the data,

$$p(\mathscr{D}\,|\,\boldsymbol{y},\boldsymbol{X}) = \frac{m(\boldsymbol{y},\boldsymbol{X}\,|\,\mathscr{D})p(\mathscr{D})}{\sum_{\mathscr{D}\in\mathscr{S}} m(\boldsymbol{y},\boldsymbol{X}\,|\,\mathscr{D})p(\mathscr{D})}, \tag{3}$$

for each $\mathscr{D}\in\mathscr{S}$; see also [2]. Since an exhaustive enumeration of *all* the regression DAGs on $q+1$ nodes is not feasible, we construct a Markov chain Monte Carlo (MCMC) algorithm to traverse the model space and approximate the posterior distribution in Eq. 3. Our MCMC method is based on a Markov chain on the model space which performs moves between graphs through additions and removals of edges provided that each proposed graph *falls inside* the model space (equivalently it must be a regression DAG); see also [2] for a general theoretical framework. The output of our MCMC consists in a collection of regression DAGs visited by the Markov chain at each time. Accordingly, the posterior distribution in (3) is approximated by the number of visits of each model. In addition we can compute the posterior probability of inclusion of each edge and obtain a single model estimate, if required, by selecting those edges whose posterior probability is greater than some threshold (e.g. 0.5); see also [2] fo details and the output of Figure 3.

3 Application

We consider Italy as destination country and inflows from 171 different origin countries at 3 different time spans, i.e. 2000, 2010 and 2016. In this first analysis, we consider each year separately. The response variable Y is then the logarithm of the annual number of migrants from origin country i to Italy in a given year t[1]. The set of covariates includes the following characteristics of the origin country:

- X_1: total population,
- X_2: percentage of urban population,
- X_3: Potential Support Ratio (PSR), defined as the ratio of people younger than 15 to the working-age population (those aged 15-64),
- X_4: Infant Mortality Rate (IMR), defined as the probability of a live birth to die before one year of age,
- X_5: distance between the capital of origin country and Italy.

[1] Data Sources: Y, OECD Stat, `https://stats.oecd.org`; $X_1 - X_4$, The Woldbank Database, `https://data.worldbank.org/`; X_5, Centre d'Etudes Prospectives et d'Informations Internationales (CEPII), `http://www.cepii.fr/`

The dataset contains no missing data and all variables were zero-centred. The normality assumption is reasonably satisfied after log-transformations.

For the sake of brevity, only results for the latest year considered (2016) are presented. Similar results, available upon request from the Authors, were obtained for years 2000 and 2010. Results are summarized in Figure 3. The left panel contains the $(5,5)$ heatmap with marginal posterior probabilities of edge inclusion for the decomposable within covariates graphical structure and the $(5,1)$ heatmap with probabilities of inclusion for the directed edges between covariates and response. In the right panel we instead report the median probability graph model, which is obtained by selecting those edges whose posterior probability of inclusion is greater than 0.5. With regards to the within covariates structure it appears that PSR (X_3) is clearly related to the percentage of urban population (X_2) and IMR (X_4). Such result is very reasonable as all these variables concern development and economic conditions. Moreover, among the covariates, only X_1, X_3 and X_4 *directly* affect the response. Consequently, we would say that the effect of the other covariates on the migration flow is "filtered" by them. For instance, using a more technical terminology, Y is conditionally independent of X_2 given $\{X_3, X_4\}$, $Y \perp\!\!\!\perp X_2 \,|\, \{X_3, X_4\}$.

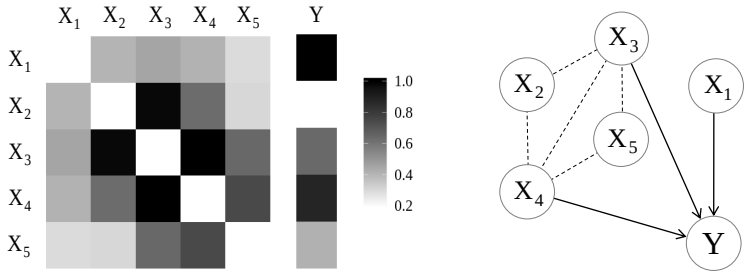

Fig. 3: Heatmaps with marginal posterior probabilities of edge inclusion for the graphical structure within covariates and between covariates and response (left panel). Median probability graph model (right panel).

4 Discussion

Graphical models represent an effective and powerful tool to study dependencies between variables and provide results that are easy and straightforward to interpret. In addition our methodology, being fully Bayesian, returns a posterior distribution over the space of *all* possible regression DAG models. This in turn provides a coherent quantification of any measure of uncertainty around the *strength* of the dependence relationship between variables. For simplicity we considered in our study only few

independent variables, but many other demographic characteristics can be included. Our results are coherent with the literature; in addition we explored the underlining structure between the determinants of international migrations.

Alternative techniques to investigate dependency relationships between variables are of course present in the literature. Among these, Structural Equation Modelling (SEM) and path analysis are the most used. However, while path analysis typically requires causal assumptions underlying the dependencies between variables and aim at estimating the size of such causal relationships under a given path diagram, our approach is more targeted to discover conditional independencies between variables and hence more general. Moreover, differently from SEM, the proposed method implicitly assumes that there are no latent variables in the system.

This contribution clearly presents some limitations which give room for future improvements. First, we analysed the data for each year separately without accounting for the effect of time over migration flows. Furthermore, we based our study on a Bayesian methodology for model selection of Gaussian graphical models which cannot be easily adapted to include different types of variables (e.g. categorical), such as the presence of colonial links or a common language between countries which have also proven to be important determinants of international migrations.

References

1. Azose, J.J., Raftery, A.E.: Estimating large correlation matrices for international migration. The Annals of Applied Statistics **12** (2), 940–970 (2018)
2. Castelletti, F.: Bayesian model selection of Gaussian DAG models. Under review (2019)
3. Consonni, G., La Rocca, L.: Objective Bayes Factors for Gaussian Directed Acyclic Graphical Models. Scandinavian Journal of Statistics. **39** (4), 321–354 (2012)
4. Fertig, M., Schmidt, C. M.: Aggregate-Level Migration Studies as a Tool for Forecasting Future Migration Streams. IZA Discussion Paper **183** (2000)
5. Kim, K., Cohen, J.E.: Determinants of International Migration Flows to and from Industrialized Countries: A Panel Data Approach Beyond Gravity. International Migration Review. **44** (4), 899–932 (2010)
6. Lauritzen, S. L.: Graphical Models. Oxford University Press, Oxford (1996)
7. Pearl, J.: Causality: Models, Reasoning, and Inference. Cambridge University Press, Cambridge (2000)