Biometrics, 2024, 80(4), ujae151
https://doi.org/10.1093/biomtc/ujae151
Biometric Methodology

Graphical model inference with external network data

Jack Jewson[1,*], Li Li[2], Laura Battaglia[3], Stephen Hansen[4], David Rossell[5,6], Piotr Zwiernik[5,6,7]

[1]Department of Econometrics and Business Statistics, Monash University, Wellington Road, Clayton, Victoria 3800, Australia, [2]School of Economics, Sichuan University, No. 24, South Section 1, Yihuan Road and No. 29, Jiuyanqiao Wangjiang Road, Chengdu, Sichuan 610065, China, [3]Department of Statistics, University of Oxford, 24-29 St Giles', Oxford OX1 3LB, United Kingdom, [4]Department of Economics, University College London, Drayton House, 30 Gordon St, London WC1H 0AN, United Kingdom, [5]Department of Business and Economics, Universitat Pompeu Fabra, Ramon Trias Fargas 25-27, Barcelona 08005, Spain, [6]Data Science Center, Barcelona School of Economics, Ramon Trias Fargas 25-27, Barcelona 08005, Spain, [7]Department of Statistical Sciences, University of Toronto, 700 University Ave 9th Floor, Toronto ON M5G 1X6, Canada

[*]Corresponding author: Jack Jewson, Department of Econometrics and Business Statistics, Monash University, Wellington Road, Clayton, Victoria 3800, Australia (jack.jewson@monash.edu).

ABSTRACT

A frequent challenge when using graphical models in practice is that the sample size is limited relative to the number of parameters. They also become hard to interpret when the number of variables p gets large. We consider applications where one has external data, in the form of networks between variables, that can improve inference and help interpret the fitted model. An example of interest regards the interplay between social media and the co-evolution of the COVID-19 pandemic across USA counties. We develop a spike-and-slab prior framework that depicts how partial correlations depend on the networks, by regressing the edge probabilities, average partial correlations, and their variance on the networks. The goal is to detect when the network data relates to the graphical model and, if so, explain how. We develop computational schemes and software in R and probabilistic programming languages. Our applications show that incorporating network data can improve interpretation, statistical accuracy, and out-of-sample prediction.

KEYWORDS: Bayesian inference; data integration; graphical model; network data; spike-and-slab.

1 INTRODUCTION

Gaussian graphical models (GGMs) are a convenient framework to describe the dependence among p random variables. As practical limitations, GGMs require estimating an inherently large number of parameters and are harder to interpret when p is large. We propose a Bayesian framework designed for situations where external data can help increase the accuracy and interpretability of GGM inference. A motivating application is learning the dependence structure between COVID-19 infection rates across USA counties, and whether said dependence is linked to network data measuring Facebook connections between counties. Kuchler et al. (2022) found a link between *marginal correlations* in said infection rates and the Facebook index. We propose a probability model to describe whether and how *partial correlations* depend on said index, and on 2 other networks measuring geographical distance and flight passenger traffic. As a preview, Figure 1 shows estimated (residual) partial correlations between each county pair vs their geographical closeness and the Facebook index. Counties that are highly connected on Facebook have a higher proportion of positive partial correlations, whereas for those lowly connected most non-zero partial correlations are negative. See Section 6 for further details.

The motivation for our methodology is 2-fold. Firstly, the ease with which one can interpret a GGM deteriorates as p grows, that is, there are simply too many edges to read them one by one. Our model regresses the probability of an edge being present, as well as the mean and variance of the associated (non-zero) partial correlation, on external network data. Said regression helps understand when one can expect an edge to be present, and to have a certain sign and magnitude. A second challenge is when the sample size n is moderate relative to the $p(p+1)/2$ covariance parameters. By integrating external network data one hopes to improve inferential accuracy, provided said data carries useful information regarding the graphical model. We discuss strategies to assess whether the network data is indeed useful.

To our knowledge, there are no model-based methods to regress the partial correlations of an undirected graphical model on multiple network-valued datasets. There is, however, work on incorporating external data in regression. Stingo et al. (2010) proposed a regression of gene expression on micro-RNA, where the prior probabilities for non-zero coefficients depend on a certain similarity score, whereas Stingo et al. (2011) incorporated pathway information. As another example, Quintana and Conti (2013) proposed a Bayesian variable selection framework where prior inclusion probabilities depend on meta-covariates.

Closer to our work, Bu and Lederer, 2021; Higgins et al., 2018; Ng et al., 2012; Pineda-Pardo et al., 2014 propose graphical LASSO (GLASSO) frameworks where penalization parameter depends on 1 network. Peterson et al. (2015) considered multiple graphical models where prior edge inclusion probabilities can be encouraged to resemble an external network. The main difference with our work is that these authors considered that

Received: January 17, 2024; Revised: October 1, 2024; Accepted: November 19, 2024

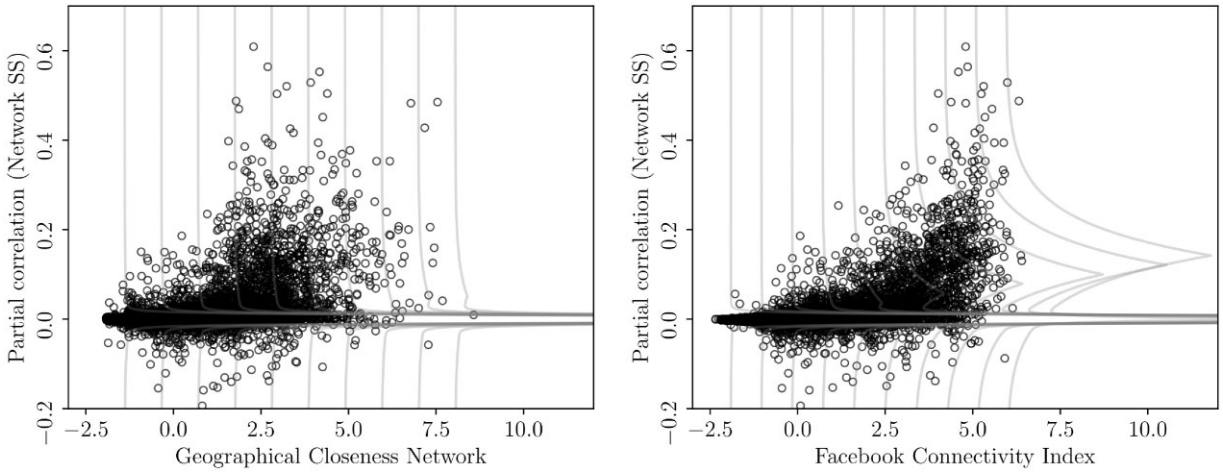


FIGURE 1 Estimated residual partial correlations in COVID-19 infections (adjusted for covariates) for all pairs of counties (y-axis) vs the counties' connection in 2 network datasets (x-axis). Left: Geographical closeness network defined as $1/\log(Geodistance)$. Right: log-Facebook connectivity index. The gray lines show the spike-and-slab distributions fitted to the estimated partial correlations, as a function of the network data.

one has strong grounds to believe that the single external dataset provides useful information. Hence, there is no need to learn the relation between the GGM and the external data. In contrast, we regress the GGM on multiple external datasets, estimate and interpret the corresponding parameters. We use a spike-and-slab model to assess whether the external data is indeed associated to the GGM's structure: the edge probability, the mean and variance of non-zero partial correlations. As a secondary contribution, we extend the GLASSO framework of Ng et al. (2012) to allow for multiple network datasets, and to assess whether the network data are informative via information criteria. Our examples show that if the external data were non-informative (independent of the GGM's structure), then inference can suffer unless one effectively removes the external data from the model. See Section 3.2 for a more extensive literature discussion.

The paper proceeds as follows. Section 2 reviews the standard GLASSO and the graphical spike-and-slab. Section 3 introduces our network-informed spike-and-slab (NI-SS) framework and the network-informed GLASSO (NI-GLASSO). Section 4 discusses our computational strategy. Section 5 uses simulations to shed light on a natural question: what if the network data are uninformative regarding the graphical model we seek to learn? Section 6 shows our results for the COVID-19 application, and Section 7 concludes. Code to reproduce our results is available in the supplementary material.

2 BACKGROUND AND NOTATION

Let $y_i \in \mathbb{R}^p$ be the outcome for individuals $i = 1, \ldots, n$ (eg, log-infection rates in p counties at week i) and $X_i \in \mathbb{R}^{p\times d}$ covariates (eg, temperature or percentage of fully vaccinated individuals in week i across the p counties). We assume that $y_i \sim \mathcal{N}_p\left(X_i b, \Theta^{-1}\right)$ independently across i, where $b \in \mathbb{R}^d$ are regression coefficients and Θ a $p \times p$ positive-definite precision (or inverse covariance) matrix. See supplementary Section A for extensions to non-Gaussian data. To ensure that the independence assumption across i is tenable, one may include lagged versions of y_i into X_i. Keeping in mind our target COVID-19 application where covariate effects were similar across the p counties, this model assumes that covariates have the same effects on all p outcomes and hence $b \in \mathbb{R}^d$, relative to the pd parameters needed in a general multivariate regression. Although, we view b as a nuisance parameter, we include it in our Bayesian framework to account for the uncertainty in its estimation. This comes at a cost: evaluating the likelihood requires $\min\{d^2p^2, np^2 + npd\}$ operations, relative to the $O(p^2)$ operations when one assumes a zero-mean outcome (see supplemenatry Section B.7). A faster alternative is to first estimate $\hat{b}$ and then define $y_i - X_i\hat{b} \sim \mathcal{N}_p(0, \Theta^{-1})$. This alternative is not reported here, but provided similar results in our examples.

The key novelty is that one observes $Q \geq 1$ networks between variables. These are $p \times p$ symmetric matrices $A^{(1)}, \ldots, A^{(Q)}$, where $a_{jk}^{(q)}$ measures strength of the connection between variables (j, k). In our COVID-19 example, $a_{jk}^{(1)}$ is the geographical closeness between counties (j, k), $a_{jk}^{(2)}$ their Facebook connection index, and $a_{jk}^{(3)}$ their flight connectivity. We assume that the network data are fixed, that is, we do not fit any model to said data. Each network may contain binary $(a_{jk}^{(q)} \in \{0, 1\})$, count $(a_{jk}^{(q)} \in \mathbb{N})$ or real $(a_{jk}^{(q)} \in \mathbb{R})$ entries.

Assuming $y_i \sim \mathcal{N}_p\left(X_i b, \Theta^{-1}\right)$, then (y_{ij}, y_{ik}) are independent given the remaining elements in y_i (and X_i) if and only if $\Theta_{jk} = 0$. We denote partial correlations by

$$\rho_{jk} := \operatorname{corr}(y_{ij}, y_{ik} \mid y_{i\{1,\ldots,p\}\setminus\{j,k\}}) = -\frac{\Theta_{jk}}{\sqrt{\Theta_{jj}\Theta_{kk}}}. \quad (1)$$

GLASSO (Yuan and Lin, 2007; Friedman et al., 2008) is a popular approach to sparsely estimate Θ by maximizing the Gaussian log-likelihood plus a LASSO penalty. While alternatives exist (eg, Fan et al., 2009) a practical appeal of GLASSO is that it defines a concave problem allowing for fast optimization. A Bayesian analogue is to obtain the posterior mode under

independent Double Exponential priors (Wang, 2012). While such a prior encourages values of Θ_{jk} that are shrunken toward 0, it does not quantify the probability that $\Theta_{jk} = 0$ and thus conduct edge selection. This issue can be addressed using a spike-and-slab framework (Gan et al., 2019). Carter et al. () proposed parameterising the prior on Θ in terms of the partial correlations, both facilitate the prior's interpretation and to ensure that the posterior mode is invariant to scale transformations, that is, the estimated ρ_{jk}'s remain the same if one applies a scale transformation to Y.

3 MODEL

3.1 Spike-and-slab framework

We propose a framework to regress partial correlations ρ_{jk} on multiple network datasets. We seek to describe how the proportion of non-zero partial correlations, as well as their mean and variance, depend on the networks. To address this we extend Gan et al. (2019). The main novelty is that both the slab prior probability and its parameters depend on network data. In particular, the slab need not be centered at 0, a feature that is novel —to our knowledge—even in simpler regression settings and may have some independent interest. We set a prior density

$$\begin{aligned}\pi(\rho \mid \eta) &= C_\eta \mathrm{I}(\rho \succ 0) \prod_{j>k} (1 - w_{jk}) \mathrm{DE}(\rho_{jk}; 0, s_0) \\ &\quad + w_{jk} \mathrm{DE}\left(\rho_{jk}; \eta_0^T a_{jk}, s_{jk}\right) \\ w_{jk} &= \left(1 + e^{-\eta_2^T a_{jk}}\right)^{-1}, \; s_{jk} = s_0(1 + \exp\left\{\eta_1^T a_{jk}\right\}),\end{aligned} \tag{2}$$

where C_η is the normalizing constant, which depends on $\eta = (\eta_0, \eta_1, \eta_2) \in \mathbb{R}^{3(Q+1)}$. The spike is a double-exponential with 0 mean and small scale s_0 meant to capture near-zero partial correlations ρ_{jk}. The slab has larger variance and captures non-zero ρ_{jk}. The slab prior probability w_{jk} follows a logistic regression on the network data $a_{jk} = (1, a_{jk}^{(1)}, \ldots, a_{jk}^{(Q)})^T$, its mean $\eta_0^T a_{jk}$ depends linearly on a_{jk} and its variance s_{jk} is larger than s_0 by a factor that also depends on a_{jk}. Our default spike prior variance is $s_0 = 0.003$, considering that $|\rho_{jk}| < 0.01$ are practically irrelevant, see supplementary Section B.2 for details and supplementary Section C.8.3 for a sensitivity analysis to s_0. We set independent uninformative $\sqrt{\Theta_{ii}} \sim \mathcal{IG}(0.01, 0.01)$, and $\pi(\Theta \mid \eta) = \pi(\mathrm{diag}(\Theta))\pi(\rho \mid \eta)$. For the regression coefficients, we set a minimally informative prior $b \sim \mathcal{N}_d(0, 5^2)$.

Importantly, positive entries in η_0 and η_1 indicate that the mean and variance (respectively) of the non-zero partial correlations increase for larger network values a_{jk} and, similarly, positive η_2 indicates a higher probability of a non-zero partial correlation for large a_{jk}. Zero entries in η indicate a lack of association.

There may be cases where the network covariates mainly affect the edge inclusion probabilities, or the mean/variance of non-zero partial correlations, but not all 3 components. One can inspect the posterior distribution of (η_0, η_1, η_2) to decide what components are needed, for example, if the 95% posterior interval of η_{0k} contains 0 then one may set it to 0 (see Section 6 and supplementary Section C.8.2). As discussed in Section 4, we set Gaussian priors on η. This is for simplicity, given that our COVID-19 example only has $Q = 3$ networks, and that most earlier literature used $Q = 1$. In settings with larger Q, a shrinkage or a selection prior (eg, spike-and-slab) on η may be more sensible. Because of the constraint $\mathrm{I}(\rho \succ 0)$ the marginal prior $\pi(\rho_{jk} \mid \eta)$ could be fairly different from the unconstrained density inside the product in (2), then w_{jk} could not be interpreted as the prior probability of an edge, and similarly for $\eta_0^T a_{jk}$ and s_{jk}. To address this, we elicit a prior $\pi(\eta)$ such that the indicator $\mathrm{I}(\rho \succ 0)$ is satisfied with high prior probability, see Section 4.1.

The linear predictors $(\eta_0^T a_{jk}, \eta_1^T a_{jk}, \eta_2^T a_{jk})$ in (2) can be generalized to a semi-parametric additive model, by simply replacing $a_{j,k}$ by a suitable basis (eg, splines), and applying our methodology as presented. Supplementary Figures C.6–C.7 show that, while not perfect, the linearity assumptions were reasonable for the COVID-19 example, hence we focus on these for simplicity.

In (2), $\eta = (\eta_0, \eta_1, \eta_2) \in \mathbb{R}^{3(Q+1)}$ drive the regression of the partial correlations onto the network data, a main quantity of interest in our framework. A standard strategy to learn such hyper-parameters is empirical Bayes, where one maximizes the marginal posterior

$$\begin{aligned}\hat{\eta} &:= \arg\max_\eta \pi(\eta \mid Y) \\ &= \arg\max_\eta \int e^{\ell(Y; b, \Theta)} \pi(b, \Theta \mid \eta) \pi(\eta) d\Theta db\end{aligned}$$

where $\ell(Y; b, \Theta)$ is the log-likelihood of $y_i \sim \mathcal{N}_p\left(X_i b, \Theta^{-1}\right)$, and then inference on (b, Θ) is based on the empirical Bayes posterior

$$\pi(b, \Theta | Y, \hat{\eta}) = e^{\ell(Y; b, \Theta)} \pi(b, \Theta | \hat{\eta}).$$

One could use the joint posterior $\pi(b, \Theta, \eta | Y)$ for inference on Θ and η, but empirical Bayes performed better in our experiments. See Giannone et al. (2021) for a related discussion on the use of empirical Bayes with spike-and-slab regression in social science applications.

3.2 Comparison to NI-GLASSO

As discussed, Ng et al. (2012) and Pineda-Pardo et al. (2014) allowed the GLASSO regularization parameter to depend on network data. In such penalized likelihood frameworks, it is customary to subtract the estimated mean $X_i \hat{b}$ from y_i, and subsequently assume $y_i - X_i \hat{b} \sim \mathcal{N}_p(0, \Theta^{-1})$. The authors proposed estimating

$$\hat{\Theta} = \arg\max_{\Theta \in \mathcal{S}_+^p} \quad \log\det(\Theta) - \mathrm{tr}(S\Theta) - \sum_{j \neq k} \lambda_{jk} |\Theta_{jk}|. \tag{3}$$

Each Θ_{jk} gets a different penalty λ_{jk}, which is a pre-specified function of the network data. Both authors allow λ_{jk} to depend on only 1 network a_{jk} and hard-code that larger a_{jk} results in smaller penalization λ_{jk}. In contrast, our framework learns whether ρ_{jk} should depend on a_{jk} and, if so, how. That is, we regress the GGM on the network. We also consider multiple networks, as is necessary to disentangle their effects, for example, having more Facebook connections between counties vs being closer geographically.

We also consider a small but practically important extension of the GLASSO methods, as a secondary contribution of our paper. We specify

$$\lambda_{jk} = \lambda_{jk}(A^{(1)}, \ldots, A^{(Q)}) = \exp\left\{\beta_0 + \sum_{q=1}^{Q} \beta_q a_{jk}^{(q)}\right\}, \quad (4)$$

where $\beta = (\beta_0, \ldots, \beta_Q) \in \mathbb{R}^{Q+1}$ are hyper-parameters. Unlike in Ng et al. (2012) and Pineda-Pardo et al. (2014), we learn if the network affects the graphical model via positive or negative β's, and consider the possibility of excluding some networks. If a network dataset does not provide useful information about Θ, then one may set $\beta_q = 0$ to avoid adding unnecessary noise to $\hat{\Theta}$. We call this method NI-GLASSO.

Two standard strategies to estimate β are cross-validation and the Bayesian information criterion (BIC). The former is more suitable for prediction than when seeking to explain the data-generating truth, for example, it does not lead to consistent model selection even in simpler linear regression (Foygel and Drton, 2010; Zhang et al., 2010). We consider

$$\hat{\beta}_{\text{BIC}} := \underset{\beta \in \mathbb{R}^{Q+1}}{\arg\min}\, \text{BIC}(\beta) = -2\ell(Y; 0, \hat{\Theta}(\beta)) + \left|\mathbf{E}(\hat{\Theta}(\beta))\right| \cdot \log n, \quad (5)$$

where $\ell(Y; 0, \hat{\Theta})$ is the log-likelihood for centered data (3) and $|\mathbf{E}(\hat{\Theta}(\beta))|$ is the number of edges associated with $\hat{\Theta}(\beta)$. Supplementary Section B.6.2 contains a comparison with the extended BIC.

For a fixed β, the NI-GLASSO in (3) is a special case of the GOLAZO of Lauritzen and Zwiernik (2022). It is a convex problem where $\hat{\Theta}(\beta)$ can be efficiently found using a block-coordinate ascent algorithm. In cases where there are only $Q = 1$ or $Q = 2$ external networks and p is moderate ($p \leq 200$, say), $\hat{\beta}_{\text{BIC}}$ can be found with a grid search. However, grid searches are very costly when $Q \geq 3$ and p is large. In such settings, we propose using Bayesian optimization, in particular the *R* package `rBayesianoptimisation` (Yan, 2016).

Finally, it is interesting to compare NI-GLASSO in (3) with our NI-SS in (2). The Bayesian interpretation of (3) is that Θ_{jk}'s arise from a Laplace random effects distribution with 0 mean and prior variance $\text{Var}\left[\Theta_{jk} \mid \beta, A\right] = 2/\lambda_{jk}^2$. From (4),

$$\log \text{Var}\left[\Theta_{jk} \mid \beta, A\right] = \log(2) - 2\left(\beta_0 + \beta_1 \tilde{a}_{jk}^{(1)} + \ldots + \beta_Q \tilde{a}_{jk}^{(Q)}\right). \quad (6)$$

This log-linear model for the variance of the precision matrix entries is similar to that for the slab variance in (2). However, (2) is more flexible in that also the probability of a non-zero partial correlation and their mean value (arguably, more interesting than their variance) are regressed on the networks. For example, Figure 1 suggests that the mean and proportion of non-zero partial correlations increase as counties get more connected both geographically and on Facebook.

4 PRIOR ELICITATION AND INFERENCE

4.1 Hyper-parameter prior elicitation

To complete our model, we specify a prior $\pi(\eta)$ on the hyper-parameters. Our guiding principle is to set a minimally informative prior, so that data may suitably update prior beliefs, while encouraging sparse solutions and preserving the interpretability of (2). We set $\pi(\eta)$ to be proportional to C_η^{-1} times independent Gaussian priors on (η_0, η_1, η_2). Adding the term C_η^{-1} helps simplify computations, since then C_η drops from the posterior density $\pi(\Theta, \eta \mid y)$. Wang (2015) argued that such cancellation of prior normalization constants does not adversely affect spike-and-slab priors in graphical model settings (as long as the constant affects hyper-parameters η but not parameters Θ, as in our case).

The prior on η_2, which drives the prior probability of an edge, was set such that the prior mean number of edges is proportional to p and hence induces sparsity. The prior parameters were also set such that the prior sample size can be thought of as 1, in analogy to the standard default $Beta(0.5, 0.5)$ prior in a Binomial experiment. The prior on η_1 was set such that the prior mode of the slab's scale is $10s_0$ and greater than $3s_0$ with probability 0.99, that is, the slab captures partial correlations of a larger magnitude than the spike. Finally, the prior on η_0 was set such that the slab has 0 prior mean and such that sampling entries of ρ independently from the double-exponential priors in (2) returns a positive-definite matrix with 0.95 prior probability. This ensures that $\pi(\rho \mid \eta)$ is similar to its unconstrained version where one drops the positive-definiteness indicator, as otherwise w_{jk} cannot be interpreted as the marginal slab probability.

Supplementary Figure B.1 plots the implied prior marginal distribution on the ρ_{jk}'s for our applications. The prior concentrates at 0 but also features thick tails to capture true non-zero ρ_{jk}'s. The corresponding posteriors (bottom panels) set significant mass away from 0, suggesting that the prior shrinkage toward 0 was not excessive. Supplementary Section B.2 provides further details.

4.2 Posterior inference

The full parameter of interest is $(b, \text{diag}(\Theta), \rho, \eta)$, where $\eta = (\eta_0, \eta_1, \eta_2)$ are the hyper-parameters in (2). In our framework, η is particularly interesting as it describes whether and how the GGM depends on the network data. To approximate the posterior distribution $\pi(b, \text{diag}(\Theta), \rho, \eta \mid y)$, we used Hamiltonian Monte Carlo. We developed an `R` implementation in `Stan` (Carpenter et al., 2017), and a Python implementation using the `NumPyro` package (Phan et al., 2019), see supplementary Sections B.1 and B.3 for details. `NumPyro` provides significant computational savings via faster automatic differentiation and the use of GPUs. Supplementary S ection E shows an order of magnitude speed ups even in simple settings.

The output of both implementations is L posterior samples $(b^{(l)}, \text{diag}(\Theta^{(l)}), \rho^{(l)}, \eta^{(l)})$ for $l = 1, \ldots, L$. Of particular interest to us is estimating the posterior probability for the presence of an edge between any 2 nodes (j, k), that is, that the partial correlation ρ_{jk} was generated by the slab in (2). To ease

notation re-write the prior (2) as

$$\pi(\rho_{jk} \mid \eta) = (1 - w_{jk}(\eta))\pi_0(\rho_{jk} \mid \eta) + w_{jk}(\eta)\pi_1(\rho_{jk} \mid \eta), \quad (7)$$

where $\pi_0(\rho_{jk} \mid \eta)$ is the spike density, $\pi_1(\rho_{jk} \mid \eta)$ the slab density, and $w_{jk}(\eta)$ the slab probability. Let $z_{jk} = 1$ indicate that ρ_{jk} was generated from the slab and $z_{jk} = 0$ that it arose from the spike, that is, $P(z_{jk} = 1 \mid \eta) = w_{jk}(\eta)$. Consider the marginal posterior probability

$$P(z_{jk} = 1 \mid Y) = \int P(z_{jk} = 1 \mid \rho_{jk}, \eta)\pi(\rho_{jk}, \eta \mid Y)d\rho_{jk}d\eta, \quad (8)$$

where from Bayes rule

$$P(z_{jk} = 1 \mid \rho_{jk}, \eta) = \frac{w_{jk}(\eta)\pi_1(\rho_{jk} \mid \eta)}{(1 - w_{jk}(\eta))\pi_0(\rho_{jk} \mid \eta) + w_{jk}(\eta)\pi_1(\rho_{jk} \mid \eta)}. \quad (9)$$

Given L posterior samples $(\rho^{(l)}, \eta^{(l)})$ from $\pi(\rho, \eta \mid Y)$, (8) may be estimated by

$$\hat{P}(z_{jk} = 1 \mid y) = \frac{1}{L}\sum_{l=1}^{L} P(z_{jk} = 1 \mid \rho_{jk}^{(l)}, \eta^{(l)}). \quad (10)$$

The description above applies in a full Bayesian treatment where η has a posterior distribution. In our empirical Bayes framework, we simply replaced η and $\eta^{(l)}$ by $\hat{\eta}$ in (7)–(10).

Our decision rule is to claim a non-zero partial correlation $\rho_{jk} \neq 0$ when $\hat{P}(z_{jk} = 1 \mid y) \geq 0.95$. The 0.95 threshold guarantees that the posterior expected false discovery proportion is below 0.05 (Müller et al., 2004), and is a common default to control false positives.

4.3 Empirical Bayes

The empirical Bayes estimate $\hat{\eta}$ (Section 3.1) requires marginalizing the joint posterior $\pi(b, \Theta, \eta \mid y)$. Given L posterior samples $(b^{(l)}, \Theta^{(l)}, \eta^{(l)})$ for $l = 1, \ldots, L$ from the latter, by definition $\eta^{(l)}$ are samples from $\pi(\eta \mid y)$. Then, for example, one may obtain $\hat{\eta}$ by maximizing a kernel density estimate of $\pi(\eta \mid y)$. Given that the accuracy of density estimators degrades with dimension, in our examples when $\dim(\eta) > 2$, we instead obtained marginal mode estimators $\hat{\eta}_j = \arg\max_{\eta_j} \pi(\eta_j \mid y)$.

5 SIMULATION STUDY

We conducted simulations to illustrate two important practical points, for simplicity in the case where the outcome Y has 0 mean. First, that when the network data are informative regarding the structure of the GGM, incorporating said data improves inference. Second and just as important, that when the network data are useless (ie, our model is misspecified) inference does not suffer too much. To this end, we compared standard GLASSO with the NI-SS of Section 3.1 and to the NI-GLASSO extension of Ng et al. (2012) and Pineda-Pardo et al. (2014) of Section 3.2 in several settings. We also considered the sIGGM method of Higgins et al. (2018) which is analogous to the NI-GLASSO in (3) but hyper-parameters enforce the assumption that the network data are related to Θ, rather than learning from data whether this is the case or not. We considered a setting with a single binary network A with entries $a_{jk} \in \{0, 1\}$ and considered $p = 50$ and sample sizes $n \in \{100, 200\}$ (results for $p = 10$ and $n = 500$ are in supplementary Figures B.2 and B.3). We then generated 50 independent datasets where $y_i \sim \mathcal{N}(0, \Theta^{-1})$. We set the data-generating Θ to have unit diagonal and most non-zero entries along the main tri-diagonal (Θ_{jk} where $|j - k| = 1$). Specifically, a proportion of 0.95 of the tri-diagonal entries were set to non-zero values uniformly spaced in $[0.2, 0.5]$. Regarding entries outside the main tri-diagonal (ie, Θ_{jk} where $|j - k| > 1$), a proportion of $0.5/p$ were set to be uniformly spaced in $[-0.1, 0.1]$ (ie, the number of edges grows linearly with p).

We consider a setting where the network data are useless (independent network), and 2 settings where they are increasingly informative. In the former setting, our prior model (2) is fully misspecified, since neither the proportion of non-zero Θ_{jk}, their mean or variance depend on the network data.

To measure the degree to which the network data $a_{jk} \in \{0, 1\}$ are informative, we count the proportion of overlaps where $a_{jk} = \mathrm{I}(\Theta_{jk} \neq 0)$, that is, the presence/absence of an edge in the network A matches that of an edge in Θ. We considered the following settings:

(1) Independent network: The tri-diagonal elements of A are set such that half of them are 1 and half of them 0, equally for the elements outside the main tri-diagonal, half of these are 1 and half of these are 0. This led to a 0.533 and 0.502 proportion of edges that agree between A and $\mathrm{I}(\Theta \neq 0)$ for $p = 10$ and 50, respectively.
(2) Mildly informative network: The tri-diagonal elements of A are set such that the proportion of $a_{jk} = 1$ is 0.75, alternatively for the elements outside the main tri-diagonal the proportion of $a_{jk} = 1$ is 0.25. This led to a 0.778 and 0.747 proportion of edges that agree between A and $\mathrm{I}(\Theta \neq 0)$ for $p = 10$ and 50, respectively.
(3) Strongly informative network: The tri-diagonal elements of A are set such that the proportion of $a_{jk} = 1$ is 0.85, alternatively for the elements outside the main tri-diagonal, the proportion of $a_{jk} = 1$ is 0.15. This led to a 0.867 and 0.844 proportion of edges that agree between A and $\mathrm{I}(\Theta \neq 0)$ for $p = 10$ and 50, respectively.

Supplementary Section B.6.3 shows an additional simulation where our prior is partially misspecified and the network is only informative about the mean of non-zero Θ_{jk}, but not about their proportion / variance.

For each setting, we report the mean squared estimation error (MSE), the false discovery rate (FDR), and the false negative rate (FNR) (Benjamini and Hochberg, 1995). The FDR is the expected proportion of false positive edges among the edges reported to be present, a measure of type I error, whereas the FNR is the expected proportion of false negatives among those not reported, which measures type II error. For the GLASSO

methods, an edge is declared if the estimate of ρ_{jk} was non-zero (rounded to 5 decimal places).

Figure 2 presents the results. Adding network data improved the spike-and-slab MSE and FNR when the network data were mildly or strongly informative ($A_{0.75}$ and $A_{0.85}$), whereas it attained similar performance to the standard spike-and-slab in the uninformative network setting (A_{ind}). The FDR did not noticeably improve, but it was always consistently below the usually accepted level of 0.05. The MSE of the NI-GLASSO behaved similarly, improving when the networks were informative and not deteriorating drastically when the network was uninformative. However, while the FDR improved when the networks were informative it was significantly above the 0.05 level. For larger p, the NI-SS also improved the MSE compared with the network-GLASSO methods. These findings suggest that the spike-and-slab formulations tend to attain better inference than the GLASSO counterparts. However, the latter may be more appealing in settings with pressing computational demands. For example, in the $p = 50, n = 100, A_{.85}$ setting GOLAZO took just over 5 min, whereas the `NumPyro` NI-SS implementation took close to 20 min (and `Stan` nearly 2 h), see supplementary Section E.

In contrast to NI-SS and NI-GLASSO, the performance of sIGGM was poor when the network data were useless (A_{ind}), illustrating the practical value of assessing whether the network data is useful for inference, as done in our 2 frameworks. In the informative network data settings, the performance of sIGGM improved, although its MSE was higher than for our methodology, and the FDR levels were significantly above 0.05.

6 COVID-19 INFECTION RATES

We downloaded weekly COVID-19 infection rates from CSSE (2020) for the period January 22, 2020 to November 30, 2021 ($n = 97$ weeks) for all USA counties ($> 3,000$). We then iteratively clustered neighboring counties with small populations until all aggregated counties had at least 500 000 inhabitants, obtaining 332 aggregated counties in total (see supplementary Section C.3 for full details). For simplicity onward, we refer to aggregated counties simply as counties. The reason for clustering counties was 2-fold. First, the weekly infection rates for smaller counties are subject to high variance, and hence less reliable than when grouping counties. Second, working with $>$ 3000 counties results in a GGM with >4 500 000 parameters, which imposes serious computational bottlenecks.

We defined the outcome of interest as the county log-infection rates, that is, log infections relative to the county's population. Our goal is to study the disease co-evolution *after* accounting for factors driving the mean structure. To this end, we included covariates temperature, vaccination rates, an index measuring the stringency of pandemic measures (CSSE, 2020), a weekly fixed effect term estimating the mean infections across all counties in that particular week, and a first-order auto-regressive term measuring the infection rate in the previous week into the model. See supplementary Section C and the supplementary code for the data collection, pre-processing, and residual checks assessing the linearity and normality assumptions, and that higher-order auto-regressive terms are not needed.

The goal is to regress the residual partial correlations between counties, which measure the extent to which COVID-19 co-evolved in these counties, on 3 network datasets. These are a geographical closeness network A_1 where $a_{jk}^{(1)}$ is the reciprocal of the log-geographic distance between counties (j, k) (hence larger values indicate smaller distance), a Facebook network A_2 where $a_{jk}^{(2)}$ is the log-Facebook connection index between (j, k), and a flight network A_3 where $a_{jk}^{(3)}$ is the logarithm of 1 + the flight passenger flow between (j, k) (see supplementary Sections C.2 and C.6 for more details). Pearson's correlation between A_1 and A_2 is 0.746, that is, the networks provide overlapping information that we wish to disentangle. We standardized the networks to all have off-diagonal entries with mean 0 and variance 1.

As a first exercise, we used NI-GLASSO to determine what network datasets are informative with respect to the target partial correlations. Supplementary Table C.1 shows a summary comparing the 8 models defined by the inclusion/exclusion of each network data. The model attaining the best BIC value includes the geographical and Facebook networks, suggesting that they both carry relevant information to help learn the graphical model, but not the flight network. To further assess the relative performance of the 8 models, we undertook a 10-fold cross-validation exercise where we assessed the log-likelihood (as a measure of predictive accuracy) in an out-of-sample fashion. The models incorporating the Facebook and geographical network also performed much better than standard GLASSO according to this predictive criterion, despite being significantly sparser (1197 vs 2637 edges).

We then applied our NI-SS framework to obtain further insights into how the proportion of edge connections, as well as the mean partial correlation, depend on the 3 networks. Supplementary Section C.8 summarizes our sampling procedure. Figure 1 displays the fitted spike-and-slab distribution as a function of both the geographical closeness and Facebook networks. The corresponding plot for the flight network is in supplementary Figure C.5. Table 1 presents the corresponding (empirical Bayes) hyper-parameter estimates, and Figure 3 displays the estimated prior slab mean and prior slab probability as functions of the networks. Recall that positive entries in η_0 and η_1 indicate that the mean and variance (respectively) of the non-zero ρ_{jk}, that is, the slab location and variance parameters, increase for counties that are strongly connected in the network. Similarly, positive entries in η_2 indicate a higher probability of there being a non-zero partial correlation between such counties. Table 1 hence shows that counties strongly connected in the Facebook had more non-zero partial correlations (relative to less connected counties), and that counties strongly connected in the Facebook and geographic networks had larger non-zero partial correlations. The flight passenger network was estimated to have no effect on there being a non-zero partial correlation, nor on their mean, as both of their credibility intervals contain 0, and a mild effect on the variance of non-zero partial correlation (in agreement with the BIC and cross-validation results in supplementary Table C.1). The coefficients for the Facebook network are larger in absolute value than those of the geographical network indicating that the Facebook network has a stronger association with the dependence on

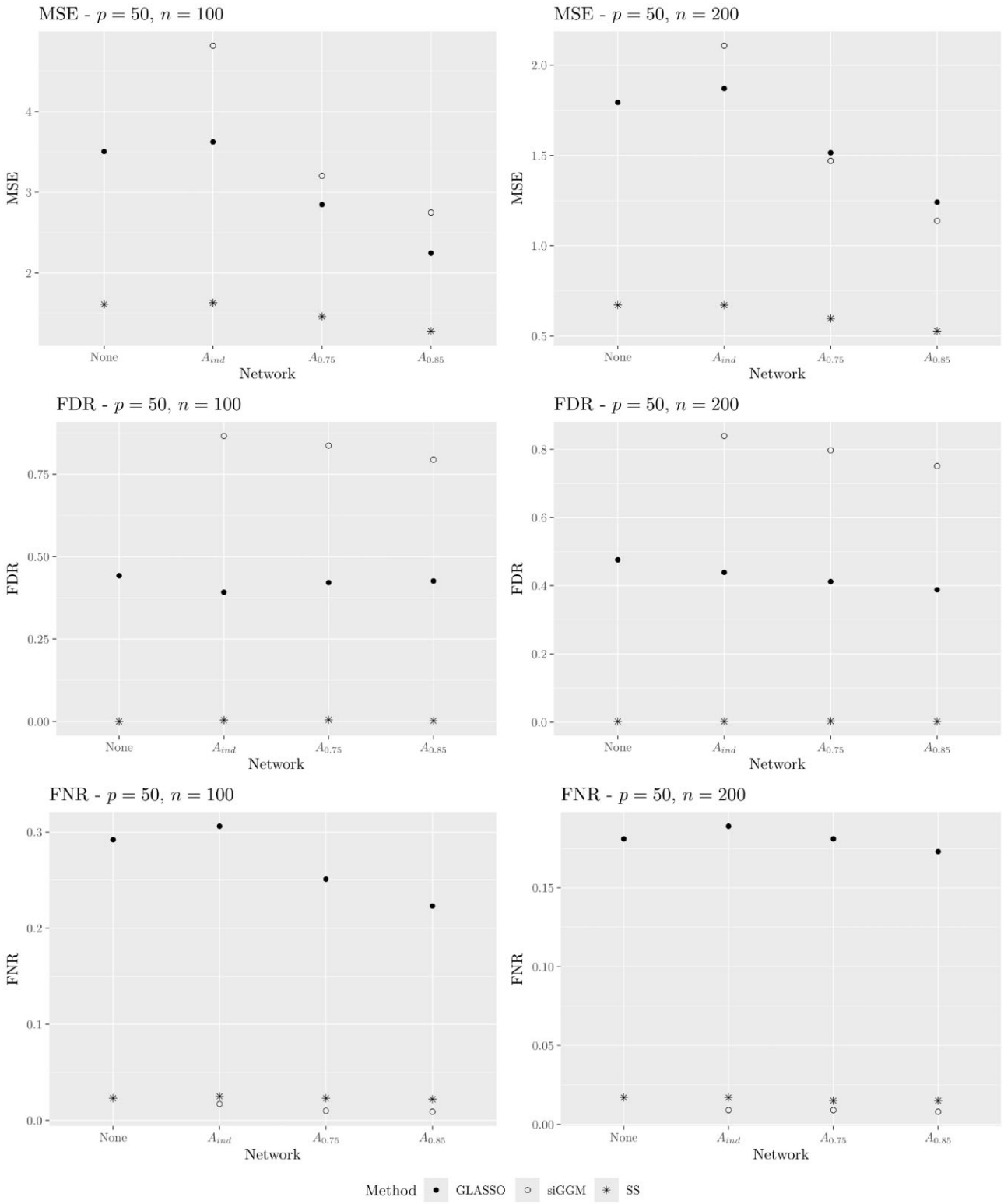


FIGURE 2 Simulation results with $p = 50$ under non, mildly, and strongly informative networks A_{ind}, $A_{0.75}$, and $A_{0.85}$. For SS and NI-SS models, edges declared when posterior probability > 0.95.

TABLE 1 NI-SS empirical Bayes (marginal MAP) estimates and 95% posterior intervals for COVID-19 data.

	Intercept	A_1	A_2	A_3
η_0 (slab location)	−0.002	**0.007**	**0.024**	0.006
95% interval	(−0.006, 0.003)	(0.002, 0.011)	(0.021, 0.028)	(−0.001, 0.009)
η_1 (slab dispersion)	**2.874**	0.025	0.043	**−0.162**
95% interval	(2.661, 3.092)	(−0.033, 0.084)	(−0.026, 0.113)	(−0.247, −0.082)
η_2 (slab probability)	**−3.797**	0.111	**1.066**	0.069
95% interval	(−4.291, −3.423)	(−0.033, 0.258)	(0.899, 1.286)	(−0.069, 0.221)

A_1, A_2, and A_3: networks defined by $1/\log(Geodist)$, $\log(Facebook)$, and $\log(1 + Flights)$. Bold values where the credibility interval does not includes 0.

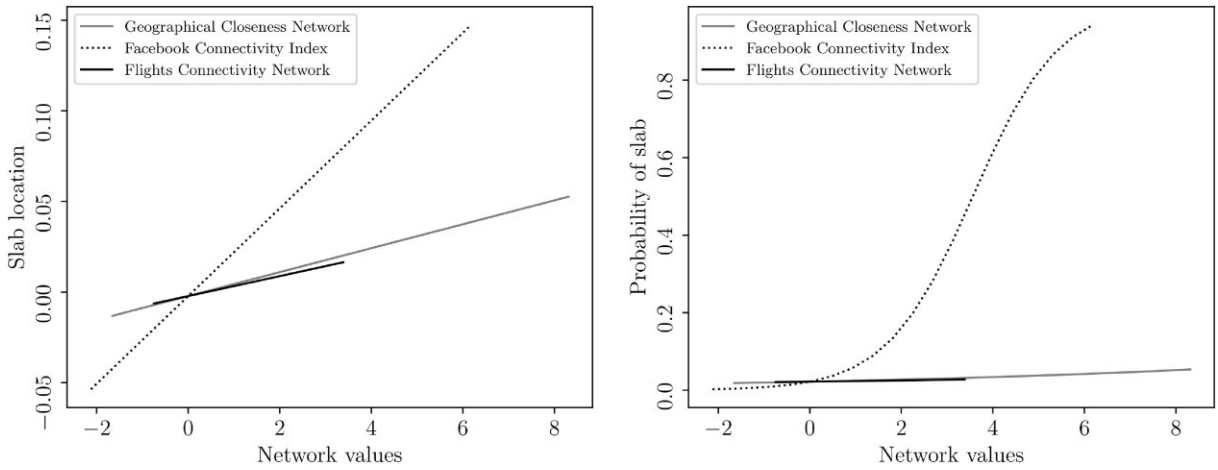


FIGURE 3 COVID-19 data: Slab location (left) and slab probability (right) as a function of the 3 networks estimated by empirical Bayes.

COVID-19 rates. This is further illustrated in Figure 3. Table 1 suggests that some hyper-parameters could be set to 0, for example, there isn't strong evidence that the slab location depends on the flights network (A_3). Supplementary Section C.8.2 provides additional results for the case where the empirical Bayes estimate for any parameter whose 95% credibility intervals contains 0 is set to 0. Overall, Figure 3 and Table 1 help interpret the GGM, in terms of when one expects an edge to be present and/or have a particular sign.

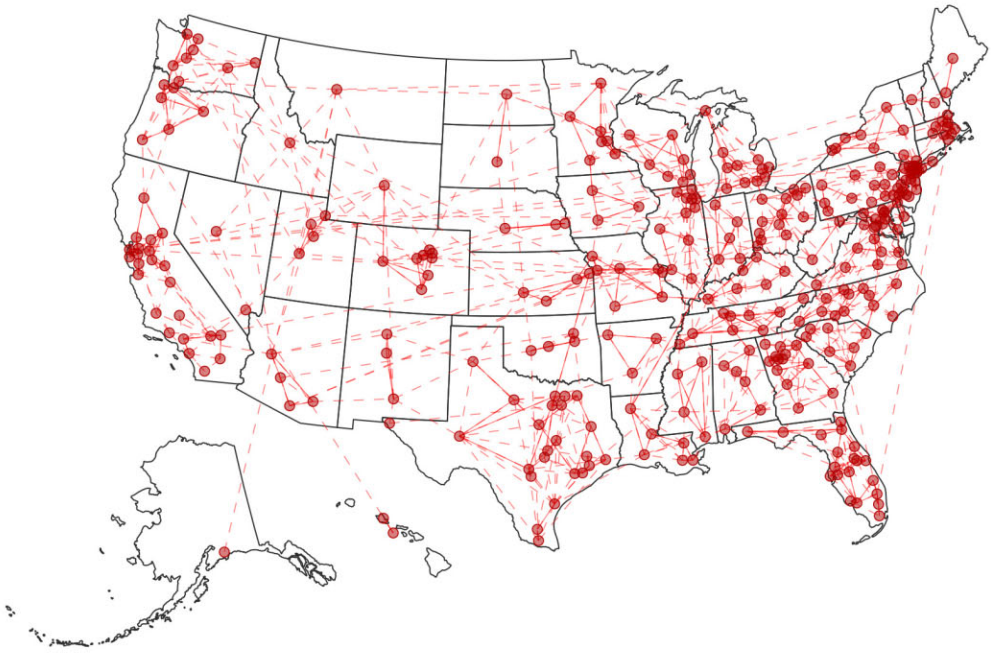
To further investigate the effects of including network data, supplementary Table C.3 compares the number of edges selected by the NI-SS to those of an SS prior using no network data. The former returns a denser graphical model, and both declare considerably fewer edges than the penalized likelihood methods. These findings agree with those of Section 5, where the more conservative nature of the spike-and-slab resulted in a good FDR control. Hence, supplementary Table C.3 suggests that NI-SS increases the power to detect edges that would be missed if the network data were not included. Such extra edges (Figure 4) occur both between geographically close and more distant counties, for example, governed by the same party.

COVID-19 diffuses locally, and it is, therefore, not surprising that the geographical distance network was informative. The greater importance assigned to the Facebook network by our model is, however, intriguing. Individuals who are connected in social networks tend to have similar backgrounds and political leanings, and to be exposed to similar information. Such a shared background may lead to similar attitudes toward health prevention, and hence similar infection risks, explaining this dependence. For example, Allcott et al. (2020) found that political beliefs were strongly tied to behavior during the COVID pandemic, more specifically that Republicans practised less social distancing. Our study reveals a similar association between social media and health outcomes.

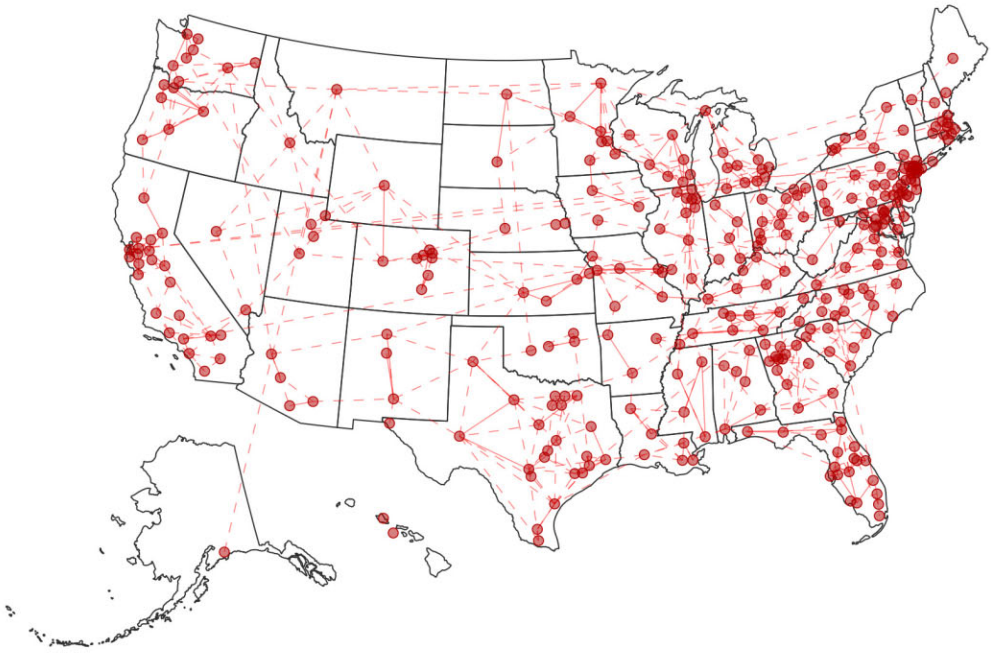
7 DISCUSSION

We hope that our framework to regress a graphical model on network data has interest beyond our motivating COVID-19 and stock market applications (supplementary Section D). The spike-and-slab provides a rich depiction for the probability that parameters are non-zero as well as the distribution of non-zero parameters. Such a framework should find applicability in many other problems, for example, high-dimensional regression or factor models. Our results showed that the external (network) data was particularly helpful in situations where the problem dimension was large relative to the sample size n, as is often the case in applications. Further, we observed that the ability to learn hyper-parameters ameliorated the consequences in a worst-case scenario where one introduces uninformative external data.

Future work could consider richer models for how the GGM depends on the networks, for example, non-parametric, or situations where the graphical model and associated network data vary across time. Another interesting avenue would be developing computational methods that scale to even higher dimensions. A possible strategy is to replace our continuous spike by a point-mass at 0. In sparse settings, such a prior could lead



(a) Edges identified by NI-SS thresholding the posterior inclusion probability at 0.95 (solid) and 0.5 (dashed).



(b) Edges identified by NI-SS and not by SS thresholding the posterior inclusion probability at 0.95 (solid) and 0.5 (dashed).

FIGURE 4 Edges identified by NI-SS.

to Markov chain Monte Carlo iterations where one updates models of dimension much lower than the $p(p+1)/2$ required by a continuous spike, albeit one would have to design efficient algorithms to search over the $2^{p(p+1)/2}$ models.

ACKNOWLEDGMENTS

We thank Dennis Kristensen for helpful comments and Du Phan for advising us on `NumPyro`.

SUPPLEMENTARY MATERIALS

Supplementary material is available at *Biometrics* online.

Web Appendices, Tables, and Figures, containing additional implementation details and further simulation results, and data and code to implement the simulations in Section 5 and real data analysis in Section 6 are available with this paper at the Biometrics website on Oxford Academic. The data and code can also be found at https://github.com/llaurabat91/graphical-models-external-networks.

FUNDING

J.J., D.R., and P.Z. were funded by Government of Spain's PID2022-138268NB-I00 financed by MCIN/AEI/10.13039/501100011033 and FEDER, by PGC2018-101643-B-I00 and the Ayudas Fundación BBVA Proyectos de Investigación Científica en Matemáticas 2021. J.J. was also funded by Juan de la Cierva Formación FJC2020-046348-I, P.Z. by NSERC grant RGPIN-2023-03481, and D.R. by Europa Excelencia EUR2020-112096, the AEI/10.13039/501100011033, EU "NextGenerationEU"/PRT, and Consolidación investigadora CNS2022-135963 by the AEI. L.L. acknowledges the financial support from the China Scholarship Council (No. 202006240148). S.H. acknowledges funding from European Research Council Consolidator Grant 864863, which supported his time and the work of L.B.

CONFLICT OF INTEREST

None declared.

DATA AVAILABILITY

The data underlying this article are available in the article and in its online supplementary material.

REFERENCES

Allcott, H., Boxell, L., Conway, J., Gentzkow, M., Thaler, M. and Yang, D. (2020). Polarization and public health: Partisan differences in social distancing during the coronavirus pandemic. *Journal of Public Economics*, 191, 104254.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57, 289–300.

Bu, Y. and Lederer, J. (2021). Integrating additional knowledge into the estimation of graphical models. *The International Journal of Biostatistics*, 18, 1–17.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M. et al. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76, 1–32.

Carter, J. S., Rossell, D. and Smith, J. Q. (2024). Partial correlation graphical LASSO. *Scandinavian Journal of Statistics* 51, 32–63.

CSSE. (2020). COVID-19 Data. Available from github: https://github.com/CSSEGISandData/COVID-19 [Accessed 1 March 2022].

Fan, J., Feng, Y. and Wu, Y. (2009). Network exploration via the adaptive LASSO and SCAD penalties. *Annals of Applied Statistics*, 3, 521–541.

Foygel, R. and Drton, M. (2010). Extended Bayesian Information Criteria for Gaussian Graphical Models. *Advances in Neural Information Processing Systems*, 23, 604–612.

Friedman, J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9, 432–441.

Gan, L., Narisetty, N. and Liang, F. (2019). Bayesian Regularization for Graphical Models with Unequal Shrinkage. *Journal of the American Statistical Association*, 114, 1218–1231.

Giannone, D., Lenza, M. and Primiceri, G. E. (2021). Economic Predictions With Big Data: The Illusion of Sparsity. *Econometrica*, 89, 2409–2437.

Higgins, I. A., Kundu, S. and Guo, Y. (2018). Integrative Bayesian analysis of brain functional networks incorporating anatomical knowledge. *NeuroImage*, 181, 263–278.

Kuchler, T., Russel, D. and Stroebel, J. (2022). The geographic spread of COVID-19 correlates with the structure of social networks as measured by Facebook. *Journal of Urban Economics*, 127, 103314.

Lauritzen, S. and Zwiernik, P. (2022). Locally associated graphical models and mixed convex exponential families. *The Annals of Statistics*, 50, 3009–3038.

Müller, P., Parmigiani, G., Robert, C. and Rousseau, J. (2004). Optimal sample size for multiple testing: the case of gene expression microarrays. *Journal of the American Statistical Association*, 99, 990–1001.

Ng, B., Varoquaux, G., Poline, J.-B. and Thirion, B. (2012). A novel sparse graphical approach for multimodal brain connectivity inference. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 707–714. Berlin Heidelberg, Springer.

Peterson, C., Stingo, F. C. and Vannucci, M. (2015). Bayesian inference of multiple Gaussian graphical models. *Journal of the American Statistical Association*, 110, 159–174.

Phan, D., Pradhan, N. and Jankowiak, M. (2019). Composable effects for flexible and accelerated probabilistic programming in NumPyro. In: *Program Transformations for ML Workshop at NeurIPS 2019*.

Pineda-Pardo, J. A., Bruña, R., Woolrich, M., Marcos, A., Nobre, A. C., Maestú, F. et al. (2014). Guiding functional connectivity estimation by structural connectivity in MEG: an application to discrimination of conditions of mild cognitive impairment. *Neuroimage*, 101, 765–777.

Quintana, M. and Conti, D. (2013). Integrative variable selection via Bayesian model uncertainty. *Statistics in Medicine*, 32, 4938–4953.

Stingo, F. C., Chen, Y. A., Tadesse, M. G. and Vannucci, M. (2011). Incorporating biological information into linear models: A Bayesian approach to the selection of pathways and genes. *The Annals of Applied Statistics*, 5, 1–24.

Stingo, F. C., Chen, Y. A., Vannucci, M., Barrier, M. and Mirkes, P. E. (2010). A Bayesian graphical modeling approach to microRNA regulatory network inference. *The Annals of Applied Statistics*, 4, 2024–2048.

Wang, H. (2012). Bayesian graphical LASSO models and efficient posterior computation. *Bayesian Analysis*, 7, 867–886.

Wang, H. (2015). Scaling it up: Stochastic search structure learning in graphical models. *Bayesian Analysis*, 10, 351–377.

Yan, Y. (2016). rBayesianOptimization: Bayesian optimization of hyperparameters. *R package version 1.0.0*. https://CRAN.R-project.org/package=rBayesianOptimization

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94, 19–35.

Zhang, Y., Li, R. and Tsai, C.-L. (2010). Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association*, 105, 312–323.

Received: January 17, 2024; Revised: October 1, 2024; Accepted: November 19, 2024