Graphical model inference with external network data

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

[1] *Department of Business and Economics, Universitat Pompeu Fabra, Barcelona, Spain*
[2] *Data Science Center, Barcelona School of Economics, Spain*
[3] *School of Economics, Sichuan University, China*
[4] *Department of Statistics, University of Oxford, UK*
[5] *Department of Economics, University College London, UK*
[6] *Department of Statistical Sciences, University of Toronto, Canada*
[*] *Correspondence address jack.jewson@upf.edu*

November 2023

Abstract

We consider two applications where we study how dependence structure between many variables is linked to external network data. We first study the interplay between social media connectedness and the co-evolution of the COVID-19 pandemic across USA counties. We next study study how the dependence between stock market returns across firms relates to similarities in economic and policy indicators from text regulatory filings. Both applications are modelled via Gaussian graphical models where one has external network data. We develop spike-and-slab and graphical LASSO frameworks to integrate the network data, both facilitating the interpretation of the graphical model and improving inference. The goal is to detect when the network data relates to the graphical model and, if so, explain how. We found that counties strongly connected on Facebook are more likely to have similar COVID-19 evolution (positive partial correlations), accounting for various factors driving the mean. We also found that the association in stock market returns depends in a stronger fashion on economic than on policy indicators. The examples show that data integration can improve interpretation, statistical accuracy, and out-of-sample prediction, in some instances using significantly sparser graphical models.

Keywords: Graphical model, Network data, Spike-and-slab, COVID-19, Stock market, Social media

1 Introduction

We consider two motivating applications where one seeks to learn the dependence structure (partial correlations) across many variables, and is specifically interested in assessing whether said dependence is associated to multiple external network datasets. First, we study the dependence between COVID-19 infection rates across USA counties, and whether said dependence is linked to network data measuring Facebook connections between counties. This is an important question because individuals who are connected in social networks tend to have similar backgrounds and to be exposed to similar information. Such a shared background may lead to similar attitudes towards health prevention, and hence similar infection risks. For example, Allcott et al. (2020) found that political beliefs were strongly tied to behaviour during the COVID-19 pandemic, more specifically that Republicans practised less social distancing. It is hence important to study the association between social media and health outcomes. As described in more detail below, a study by Kuchler et al. (2021) found a link between *marginal correlations* in infection rates between counties and the Facebook index. We propose a probability model that can describe whether and how *partial correlations* depend on said index, as well as two other networks related to geographical distance and flights passenger traffic. The latter two are meant to help disentangle the effect of two counties being connected on Facebook and their being geographically close or their being major travel between them, i.e. more direct contacts. As a preview of our findings, Figure 1 (top) shows estimated (residual) partial correlations between each county pair vs. their geographical closeness and the Facebook index, Figure B.6 contains the corresponding plot for the flights network. 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. Geographically close counties also tend to have positive partial correlations while there does not appear to be a strong relationship between the estimated partial correlations and the flight passenger network. The bottom panels show our spike-and-slab model relating the network data to the probability that a partial correlation is non-zero, and to the mean and variance of the non-zero partial correlations.

As a second application, we study the dependence of stock market excess returns across companies, incorporating external data on similarities between companies in their exposure to economic and policy risks (as defined by Baker et al. (2019)). Said risks were extracted from text data in mandatory regulatory filings where companies must disclose potential risks, and the idea is that if two companies disclosed similar economy- or policy-related risks then it may be more likely that they

2

have similar stock market returns. The applied relevance of the problem arises from the longstanding insight in finance that the dependence among assets informs optimal portfolios (Markowitz, 1952). In particular, the precision matrix determines the weights across assets that minimise a portfolio's standard deviation. Bringing optimal portfolio theory to data requires estimating high-dimensional covariance/precision matrices, which is an important barrier to its practical application (Elton and Gruber, 1973). A variety of approaches have been used to tackle the problem including, recently, GLASSO (Goto and Xu, 2015), see also Senneret et al. (2016) for an empirical review. A critical observation is that we seek not only to estimate the partial correlations featuring in the precision matrix, but also to portray how they may depend on the text-based economic and policy risks, to shed light onto the joint behavior of stock market returns.

We use Gaussian graphical models (GGMs) and extensions discussed later as a convenient framework that describes the dependence among random variables in an interpretable manner, providing a suitable basis for our applications. There are however certain challenges that led us to develop a methodological framework that is another main contribution of this paper, and can be applied to numerous applications other than those considered here.

A first applied challenge is that the ease with which one can interpret the output of a graphical model deteriorates as the number of variables p gets large, i.e. there are simply too many edges to read them one by one. Our proposed model provides a way to regress 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, as illustrated in Figure 1. A second challenge is that in our applications the sample size n is moderate relative to the $p(p+1)/2$ covariance parameters. By integrating external network data one hopes to improve the accuracy of the inference, provided said data carries useful information regarding the graphical model. Our framework provides natural novel strategies to assess whether the network data is indeed useful.

To our knowledge, there are no model-based methods to incorporate multiple network-valued external data in undirected graphical models. There has been, however, active research on incorporating external data in regression. For example, Stingo et al. (2010) proposed a multivariate regression of gene expression on micro-RNA, where the prior probabilities that micro-RNAs have a non-zero coefficient depend on an external biological and structural similarity score. Similarly, Stingo et al. (2011) incorporated pathway information into regression models for gene expression, Quintana and Conti (2013) proposed a Bayesian variable selection framework where prior inclusion probabilities depend
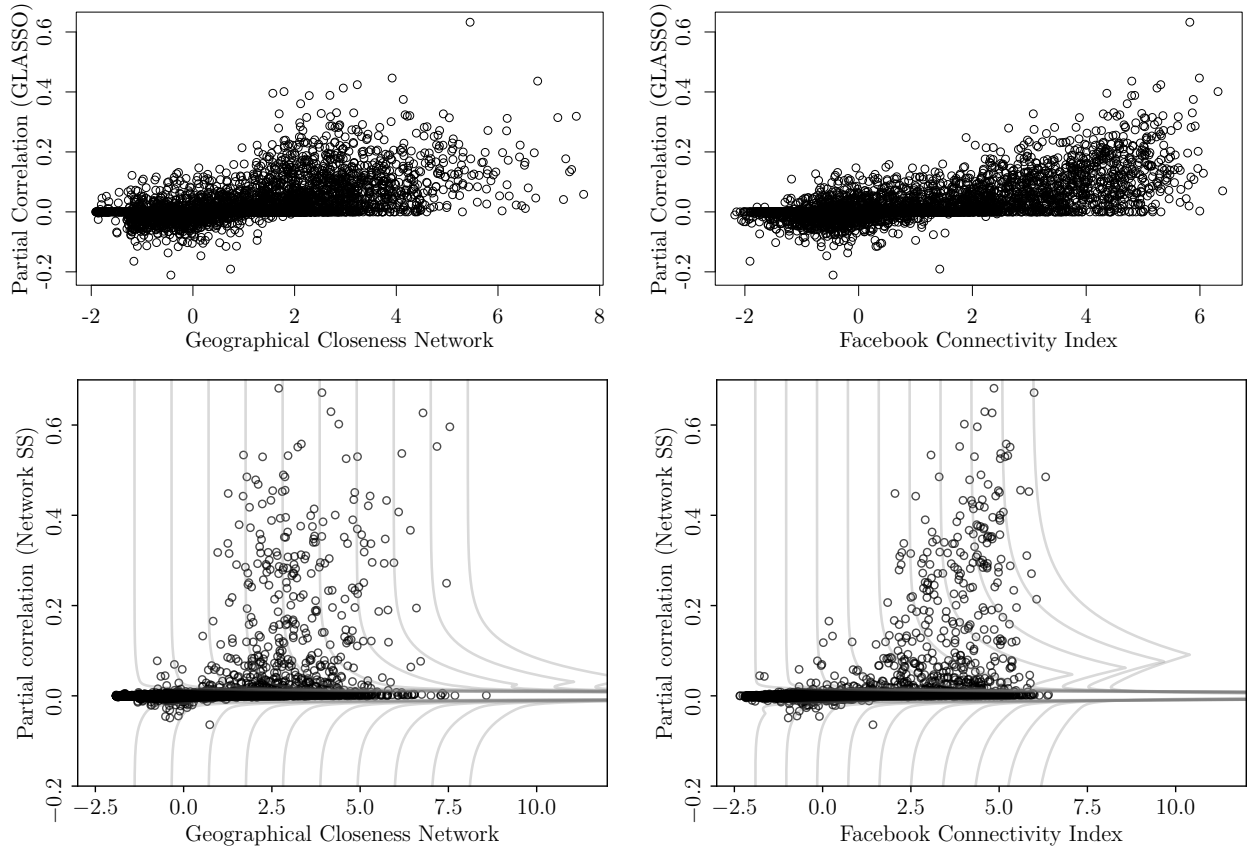
Figure 1: Residual partial correlations in COVID-19 infections (adjusted for covariates) across counties vs Geographical Closeness Network defined as $1/\log(Geodistance)$ (left) and log-Facebook Connectivity Index (right). Top panel: partial correlations estimated with graphical LASSO, with penalization parameter set via BIC. Bottom panel: fitted spike-and-slab distributions and fitted partial correlations estimated with network graphical spike-and-slab LASSO.

on meta-covariates, Cassese et al. (2014) a multivariate regression of gene expression versus copy number variations that incorporates their physical distance in the genome, Peterson et al. (2016) a regression framework using a network for covariate penalisation, and Chiang et al. (2017) a brain activity vector auto-regression that incorporates external brain information. Chen et al. (2021) predicted disease outcomes given single nucleotide polymorphisms, where the LASSO regularisation parameter depends on functional annotation categories.

There has also been work incorporating network data in graphical models, primarily in neuro-

science. Ng et al. (2012); Pineda-Pardo et al. (2014); Higgins et al. (2018) considered penalised likelihood GGMs to understand co-activation across brain regions, where one has strong grounds to believe that external network data extracted from known brain structure provides useful information. In a similar vein, Bu and Lederer (2021) use distances between brain regions to drive the regularisation of a GGM that is fit via multiple univariate regressions, and provide theoretical conditions for asymptotic learning of the GGM's structure. The main applied difference with our setting is that we wish to assess whether the network data are informative and, if so, depict how. Another difference is that we consider multiple network datasets (e.g. Facebook, distance, flights), rather than only one. In simulations we illustrate how assessing whether the network data are useful or not can lead to significant practical improvements. As discussed, the main methodological difference is that we develop a probabilistic spike-and-slab model to regress the GGM on the network data that helps interpret the presence of edges and the sign and magnitude of partial correlations. This is important in our applications, e.g. the bottom panels in Figure 1 depict that large Facebook connectivity is associated with positive partial correlations.

We develop two frameworks to integrate network data into GGM selection and parameter estimation. The first framework is a hierarchical extension of the graphical LASSO (GLASSO) (Friedman et al. (2008); Yuan and Lin (2007), see also Wang (2012) for a discussion of Bayesian counterparts). The framework largely follows that in Ng et al. (2012), except that we learn critical hyper-parameters from data and assess whether each network data is actually useful or not. We also develop tailored optimisation algorithms that build on the GOLAZO algorithm of Lauritzen and Zwiernik (2020) so that the computational cost is similar to a standard GLASSO problem, and we apply Bayesopt algorithms to speed up the search over hyper-parameter values. Our second framework is the main contribution and uses a spike-and-slab prior, with the novel feature that the slab's probability, location and variance are regressed on the network data. To ensure its practical applicability we developed a software implementation in the probabilistic programming languages `Stan` (Carpenter et al., 2017) and `NumPyro` (Bingham et al., 2019; Phan et al., 2019). The latter capitalises on efficient automatic differentiation and GPUs to help boost the computational speed. Similarly, the first framework is implemented in R.

The paper proceeds as follows. Section 2 discusses our motivating applications in more detail. Section 3 reviews the GLASSO, introduces our network-adjusted extension and its Bayesian analogue. Section 4 discusses our computational strategy for learning the graphical model and hyper-parameters that depict its association with the external network data. Section 5 uses simulations to shed light on a natural practical question: what if the network data are useless, i.e. uninformative regarding

5

the graphical model we seek to learn? We illustrate that one should assess whether the network data have useful information about the GGM and, if not, discard them to avoid deteriorating inference. Section 6 shows our main results for the COVID-19 and stock market applications, and Section 7 concludes. Code to implement all of our experiments and data pre-processing is available at https://github.com/llaurabat91/graphical-models-external-networks.

2   Motivating applications

2.1   Dependence in COVID-19 infections versus Facebook, geographical and flight networks

Studying the evolution of pandemics such as COVID-19 is of great importance for health, economic and societal reasons. There are many studies to forecast infections or to understand how they are related to various factors (e.g. health measures, temperature). We consider a further important aspect that received less attention: understanding how the disease co-evolves across (possibly distant) geographical units, and what factors are associated to such co-evolution. For example, if several counties were expected to simultaneously exhibit higher-than-expected infection rates, health authorities might need to plan resources accordingly. Further, identifying factors that are related to the co-evolution (e.g. the Facebook index) may suggest strategies to limit such coordinated growth (e.g. targeted information campaigns).

To study COVID-19 co-evolution across USA counties, we downloaded weekly infection rates from CSSE (2020a) for the period 22 January 2020 to 30 November 2021 (97 weeks total) for all USA counties ($> 3,000$ in total). We then iteratively clustered neighbouring counties with small population until all aggregated counties had at least 500,000 inhabitants, obtaining 332 aggregated counties in total. Full details of our clustering procedure are presented in Section B.3. For simplicity onwards we refer to aggregated counties simply as counties. The reason for clustering counties was two-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 $> 3,000$ counties results in a GGM with $> 4,500,000$ parameters, which imposes serious computational bottlenecks.

We also obtained data on covariates that are thought to be associated with the disease's evolution, such as temperature, population density, vaccination rates and an index measuring the stringency of pandemic measures (CSSE (2020c); Bureau (2020); CSSE (2020d); CSSE (2020b)). We defined the outcome of interest as the county log-infection rates, i.e. log infections relative to the county's

6

population. Our interest is in studying the disease co-evolution *after* accounting for factors driving the mean structure. To this end, we fitted a linear regression model that included temperature, vaccination rates, the stringency of pandemic measures, 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. See Section B 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.

Although the mean model explained most of the variance in infection rates (adjusted R^2 coefficient 0.942), certain county pairs were systematically both above or below the model predictions. Specifically, we estimated partial correlations in the regression residuals for each county pair via graphical LASSO, and obtained numerous non-zero estimates (Figure 1, top). Said partial correlations indicate that certain county pairs tend to behave better or worse than expected (given the week's overall pandemic status and other covariates) in a coordinated fashion. Our primary goal is to assess whether this coordinated behavior occurs more frequently across counties that are strongly connected via social media, given by the Facebook index. Said index defines a network of counties, measuring the strength of the connection between every pair of counties. We also consider two further networks, one based on geographical closeness (see Section 6.1) and a second measuring flow of passengers between two counties by plane (see Section B).

We see partial correlations as an appealing measure of disease co-evolution. For example, suppose that infections in County A drive those of County B, which in turn drive those of County C, then all three counties would have non-zero marginal correlation. In contrast, the partial correlation between counties A and C would be zero, suggesting there is no direct link between them. An important observation stemming from Figure 1 is that counties that are highly connected on Facebook have a higher proportion of non-zero, and positive, partial correlations. A similar observation applies to geographical distances. Hence one wishes not only to regularise to a lesser extent county pairs with a strong Facebook connection but also to describe how the average non-zero partial correlation depends on Facebook (or geographical, or flight) connectivity. This desideratum led us to develop a network-regularised spike-and-slab framework, where the slab's mean, variance and probability are regressed on the network, see Section 3.

7

2.2 Dependence in stock market returns versus text data

Our goal is to study whether and how covariation in stock market excess returns (i.e. returns above/below those that were expected, see below) across firms is associated with firms' sharing similar risks. To measure to what extent they do so, we downloaded text of the *Risk Factors* (RF) section of publicly traded firms' annual 10-K filings to the USA Securities and Exchange Commission. For each firm, we combine all filings made between 2015 and 2019, inclusive. Said filings describe exhaustively future earnings risks faced by the firms, and there is an incentive for full disclosure because investors can take legal action when firms withhold information that if disclosed would have prevented financial losses. Firms that face similar risks may have more dependent stock returns, e.g. two firms mentioning risks to oil price rises may co-move when oil prices change. Indeed, Hanley and Hoberg (2019) regressed the covariance of excess returns between pairs of financial firms on a measure of RF text overlap and showed a positive relationship in the lead-up to the global financial crisis in 2008. More recently, Davis et al. (2020) shows that firms with similar RF texts reacted similarly to the arrival of COVID-19. Our analysis goes beyond these studies by modelling partial rather than marginal correlations. We also allow distance in RF-text space to influence both the probability of a connection between firms and the mean (and variance) of the partial correlations on the network.

We consider $p = 366$ firms traded on US markets that satisfy the following conditions: i) membership in S&P500 at the end of 2019; ii) closing stock price adjusted for stock splits and dividends available in the COMPUSTAT database for every trading day between 2 January 2019 to 31 December 2019 (252 trading days in total); iii) at least one 10-K filing available in 2014-2019. For each trading day in 2019 we construct daily excess returns using the Fama-French three-factor model. Specifically, we individually regress each firm's daily log-returns on the variables contained in the daily, three Fama/French factors file downloaded from Kenneth French's Data Library website. The residual is the excess return.

To measure textual similarity between companies, we first construct a bag-of-words representation of each firm's 10-K filings during 2014-2019. We follow Baker et al. (2019), and compute firms' exposure to 16 separate *economic* risks and 20 separate *policy* risks. For each risk r, Baker et al. (2019) define a term set T_r containing terms that reflect the exposure. For example, the policy risk 'food and drug policy' is captured by the term set {prescription drug act, drug policy, food and drug administration, fda}.[1] Baker et al. (2019) show that intertemporal variation in economic and policy

---

[1]In common with the text-as-data literature, we refer here to terms even when a 'term' is a multi-word expression. See Appendix B of Baker et al. (2019) for a complete description of the term sets associated with each risk.

risk terms in newspaper articles closely tracks aggregate market volatility. This motivates the idea of using variation in these terms across individual firms to better measure their co-movement across trading days.

Let $x_{i,v}$ be the count of term v in firm i's 10-K filings during 2014-2019 and let $C_i \equiv \sum_v x_{i,v}$ be the total number of terms. We measure each firm's exposure to risk r as $\log \left(1 + \sum_{v \in T_r} x_{i,v}/C_i\right)$, i.e. logarithm of 1 plus the proportion of words referring to risk r out of the total C_i words. We use the logarithm to account for the fact that a risk term not being mentioned at all versus being mentioned once is likely to be more informative than being mentioned many times compared with slightly more times. For each pair of firms, we then measure its similarity in exposure to economic risks by computing the correlation between the vector of economy-related risks. This defines a network between companies such that the network connection between companies (j, k) is given by said correlation. We proceeded analogously to define a policy risk network by computing correlations between policy-related terms.

In summary, our data processing produced two networks between firms that measure their similarity in risk exposures based on a particular representation of RF texts. We remark that one could use alternative text analysis tools, however our goal is to establish that text-based relational data can be useful to estimate dependence in stock returns. The optimal representation of text for this task is left as an open question. Still, as we show below, separately controlling for economic and policy risks yields important insights regarding whether government policy generates return co-movement above and beyond that generated by firm fundamentals.

See Section C and the supplementary code for the data collection, pre-processing, linear model fit, and residual checks assessing our model assumptions.

3   Model

We describe two model-fitting strategies to regress an undirected GGM on p variables onto multiple external network datasets. Section 3.1 discusses network GLASSO, which we mainly use as a computationally-convenient framework to assess whether one should add/remove each network dataset. We also discuss a Bayesian interpretation useful to check that the assumed model fits the observed data. Section 3.2 is our main contribution, a spike-and-slab model to regress partial correlations on network data. Section 3.3 discusses how to extend our framework beyond Gaussian data, as needed for the stock market application.

We set notation. Let $y_i \in \mathbb{R}^p$ be the outcome vector for individuals $i = 1, \ldots, n$ (e.g. log-infection

rates in p counties at week i, or stock excess returns for p companies at day i) and $x_i \in \mathbb{R}^d$ covariates (week indicator, temperature, percentage of fully vaccinated individuals in week i, etc.). We assume that $y_i \sim \mathcal{N}_p\left(Bx_i, \Theta^{-1}\right)$ independently across $i = 1, \ldots, n$, where B is a $p \times d$ regression coefficients matrix and Θ a $p \times p$ positive-definite precision (or inverse covariance) matrix. To ensure that the independence assumption across i is tenable, we include lagged versions of y_i into the covariates x_i, as described in Section 2 and B. For simplicity, in our applications we start by subtracting the estimated mean $\hat{B}x_i$ from y_i, where $\hat{B}$ is the least-squares estimator, and subsequently assume the outcomes to have zero mean, i.e. $y_i \sim \mathcal{N}_p(0, \Theta^{-1})$.

A convenient property of modelling $y_i \sim \mathcal{N}_p\left(0, \Theta^{-1}\right)$ is that conditional independence statements can be drawn from the graph defined by the non-zero elements of Θ. Specifically, (y_{ij}, y_{ik}) are independent given the remaining elements in y_i if and only if $\Theta_{jk} = 0$. As argued earlier, in our applications we use partial correlations as a measure of association. We denote partial correlations by

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

Importantly, in our framework, one also observes external data in the form of $Q \geq 1$ 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 the COVID-19 application $a_{jk}^{(1)}$ is the geographical closeness between counties (j, k), $a_{jk}^{(2)}$ their Facebook connection index, and $a_{jk}^{(3)}$ their flight connectivity. In the stock application, $a_{jk}^{(1)}$ is the similarity between firms (j, k) in their exposure to economic risks, and analogously $a_{jk}^{(2)}$ for policy risks.

3.1 Network graphical LASSO

Network graphical LASSO is a penalised likelihood framework to estimate $\Theta \in \mathcal{S}_+^p$ by maximising a Gaussian log-likelihood plus a graphical LASSO (GLASSO) penalty (Friedman et al., 2008; Yuan and Lin, 2007), where the magnitude of said penalty is regressed onto the network datasets. Specifically, we consider

$$\hat{\Theta} = \underset{\Theta \in \mathcal{S}_+^p}{\arg\max} \quad \log\det(\Theta) - \mathrm{tr}(S\Theta) - \sum_{j \neq k} \lambda_{jk}|\Theta_{jk}|, \tag{2}$$

where $\mathcal{S}_+^p$ is the set of non-negative definite matrices, $\mathrm{tr}(\cdot)$ the matrix trace, S the empirical covariance matrix of $(y_1, \ldots, y_n)$,

$$\lambda_{jk} = \lambda_{jk}(A^{(1)}, \ldots, A^{(Q)}) = \exp\left\{\beta_0 + \sum_{q=1}^{Q} \beta_q a_{jk}^{(q)}\right\} \tag{3}$$

are regularisation parameters, and $\beta = (\beta_0, \ldots, \beta_Q) \in \mathbb{R}^{Q+1}$ are regularisation hyperparameters that play a critical role in determining the level of sparsity in $\hat{\Theta}$. That is, each Θ_{jk} gets a potentially different penalty parameter λ_{jk}, which is a function of the network data $A^{(1)}, \ldots, A^{(Q)}$. To simplify notation, we omit the dependence on $A^{(1)}, \ldots, A^{(Q)}$ and simply use λ_{jk}, and let $A = (A^{(1)}, \ldots, A^{(Q)})$. For convenience we parameterise the penalties in terms of a scaled version of $A^{(q)}$ that is centered to have zero sample mean and unit sample variance, and which we denote by $\bar{A}^{(q)}$. GLASSO is the particular case where λ_{jk} are constant across (j, k).

Ng et al. (2012) proposed the penalty in (2)-(3), the main difference being that we consider multiple networks ($Q > 1$) and that we learn hyper-parameters β from data, including the exclusion of some networks. Two popular strategies to set hyper-parameters are using cross-validation (Friedman et al., 2008) and information criteria such as the Bayesian information criterion (BIC) (Schwarz, 1978). The former is more suitable for predictive tasks than when seeking models that help explain the data-generating truth, e.g. cross-validation does not lead to consistent model selection even in simpler linear regression where the BIC and related information criteria are consistent, see Foygel and Drton (2010); Zhang et al. (2010); Wang and Zhu (2011); Fan and Tang (2013). We hence use the BIC to learn β. Specifically, viewing $\hat{\Theta}(\beta)$ as a function of β, we choose β minimising

$$\hat{\beta}_{\mathrm{BIC}} := \underset{\beta \in \mathbb{R}^{Q+1}}{\arg\min} \mathrm{BIC}(\beta) \; = \; -2\ell_n(\hat{\Theta}(\beta)) + \left|\mathbf{E}(\hat{\Theta}(\beta))\right| \cdot \log n, \tag{4}$$

where $\ell_n(\hat{\Theta})$ is the Gaussian log-likelihood function and $|\mathbf{E}(\hat{\Theta}(\beta))|$ counts the number of edges in the graph associated with $\hat{\Theta}(\beta)$. Importantly, note that when $\beta_q = 0$ then the qth network dataset is effectively excluded. The idea is that if a network dataset does not provide useful information about Θ, then one may set $\beta_q = 0$ to avoid adding unnecessary noise to $\hat{\Theta}$, see Section 5 for an illustration. An alternative to the BIC is the Extended BIC (EBIC) (Chen and Chen, 2008). As a sensitivity check, we provide results using the EBIC to select β in Sections A.5, B.8 and C.6. In our examples the EBIC was overly conservative in selecting edges, which resulted in high mean-squared-error. Finally, we note that there are alternative approaches to choosing β, see Kuismin and Sillanpää (2021), but they require more extensive computations that become prohibitive in our setting. We also note that alternatives to (2) include the adaptive graphical LASSO, SCAD and MCP (Fan et al., 2009; Wang et al., 2016), which were proposed to reduce bias in the estimation of large entries in Θ. We focus on (2) however due to its practical appeal of defining a concave problem for which one may establish efficient optimisation methods.

One could of course consider alternative parameterisations to (3), e.g. let λ_{jk} depend non-

11

parametrically on the network data. However, (3) requires fewer hyper-parameters than a non-parametric treatment and is easy to interpret: the log-regularisation depends linearly on the networks. Further, a model-checking exercise suggested that (3) is a reasonable parameterisation for our two motivating applications. Said model-checking is best understood by adopting a Bayesian interpretation. The penalised estimator associated to (3) is equivalent to the posterior mode under independent Laplace priors (Wang, 2012) with scale parameter $1/\lambda_{jk}$, that is

$$\pi(\Theta \mid A, \beta) \;\propto\; \prod_{j>k} \frac{\lambda_{jk}}{2} \exp\left\{-\lambda_{jk}|\theta_{jk}|\right\} \mathrm{I}(\Theta \succ 0), \tag{5}$$

where $\mathrm{I}(\Theta \succ 0)$ is an indicator for Θ being positive-definite, λ_{jk} is as in (3) and β are now prior parameters. The Bayesian interpretation is that the λ_{jk}'s arise from a Laplace random effects distribution with parameter β. The *a priori* expected value of θ_{jk} is 0, which induces sparsity, and the prior variance is

$$\mathrm{Var}\left[\theta_{jk} \mid \beta, A\right] = \mathbb{E}\left[\theta_{jk}^2 \mid \beta, A\right] = \frac{2}{\lambda_{jk}^2}. \tag{6}$$

Therefore (3) assumes that the log-variance of the partial covariances θ_{jk} depends linearly on the network data

$$\log \mathbb{E}\left[\theta_{jk}^2 \mid \beta, A\right] = \log(2) - 2\left(\beta_0 + \beta_1 \bar{a}_{jk}^{(1)} + \ldots + \beta_Q \bar{a}_{jk}^{(Q)}\right). \tag{7}$$

Provided one has an initial estimate of the left-hand side of (7), which in our examples we derived from standard GLASSO estimates of θ_{jk}, one may check whether its relation to the network data is roughly linear. Such a check motivated taking the logarithm of the raw distances, Facebook connectivities and flight passenger flow to define our networks for the COVID-19 data, while the stock market risk indicator networks required no transformations. See Supplementary Sections B.6 and C.4 for further details.

3.2   Network graphical spike-and-slab LASSO

The network graphical LASSO in (2) provides sparse point estimates of partial correlations and, via its Bayesian interpretation, describes how their variance depends on the network data. In our applications, however, we also seek to describe how the proportion of non-zero partial correlations and their mean depend on the network. For example, in the COVID-19 data both the probability that two counties are conditionally dependent and the mean partial correlation grow as their Facebook connection grows (Figure 1), and similarly for the stock market data (Figure C.5). To address this

issue, we developed a spike-and-slab framework that builds on the regression setting of Rockova and George (2014) and the graphical spike-and-slab of Gan et al. (2018). 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 zero, a feature that is novel—to our knowledge—and has some independent interest.

We parameterise Θ in terms of partial correlations ρ_{jk} in (1), which facilitates interpretation and ensures that the posterior mode is invariant to scale transformations. By scale invariance we refer to the property that the estimated ρ_{jk} remain the same regardless of whether one applies a scale transformation to the input data or not, see Carter et al. (2021) for a detailed discussion. We set a prior density $\pi(\mathrm{diag}(\Theta), \rho) = \pi(\mathrm{diag}(\Theta))\pi(\rho)$, where $\sqrt{\Theta_{ii}} \sim \mathcal{IG}(a, b)$ with $a = 0.01$ and $b = 0.01$ reflecting an uninformative prior on the diagonal elements of Θ, and

$$\pi(\rho \mid \eta) = C_\eta \mathrm{I}(\rho \succ 0) \prod_{j > k} (1 - w_{jk})\mathrm{DE}(\rho_{jk}; 0, s_0) + w_{jk}\mathrm{DE}\left(\rho_{jk}; \eta_0^T a_{jk}, s_{jk}\right) \tag{8}$$

$$w_{jk} = \left(1 + e^{-\eta_2^T a_{jk}}\right)^{-1}, \quad s_{jk} = s_0(1 + \exp\left\{\eta_1^T a_{jk}\right\}),$$

where C_η is the normalisation constant and $\mathrm{I}(\rho \succ 0)$ a positive-definiteness indicator. The spike is a double-exponential with zero mean and small scale s_0 meant to capture partial correlations that are practically zero, whereas the slab has larger variance s_{jk} and may not be centered at zero. The slab prior probability w_{jk} follows a logistic regression on the network data $a_{jk} = (1, a_{jk}^{(1)}, \ldots, a_{jk}^{(Q)})$, 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}. Specifically, 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 connections a_{jk}, and similarly positive η_2 indicates a higher probability of a non-zero partial correlation for large a_{jk}.

We remark that because of the constraint $\mathrm{I}(\rho \succ 0)$ the marginal prior $\pi(\rho_{jk})$ could be fairly different from the unconstrained density inside the product in (8), 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 issue we elicit prior parameters such that the indicator $\mathrm{I}(\rho \succ 0)$ is satisfied with high prior probability, see below.

Above $\eta = (\eta_0, \eta_1, \eta_2) \in \mathbb{R}^{3(Q+1)}$ are hyper-parameters driving the regression model of the partial correlations ρ_{jk} onto the network data a_{jk}, and are a main quantity of interest in our applications. A standard strategy to set prior hyper-parameters such as η in (8) is an empirical Bayes framework where one maximises the marginal likelihood. Such a framework allows us to do inference on the η's themselves through the marginal posterior $\pi(\eta \mid y)$ and inference for Θ through the empirical Bayes

13

posterior

$$\pi(\Theta|y,\hat{\eta}) = f(y|\Theta)\pi(\Theta|\hat{\eta}),$$

where $\hat{\eta}$ maximises the marginal posterior of η given the data

$$\hat{\eta} := \arg\max_{\eta} \pi(\eta \mid y) = \arg\max_{\eta} \int \pi(\Theta, \eta|y)d\Theta = \arg\max_{\eta} \int f(y|\Theta)\pi(\Theta|\eta)\pi(\eta)d\Theta.$$

One could consider using the joint posterior $\pi(\Theta, \eta|y)$ for inference on Θ and η, but we found empirical Bayes to perform better in our experiments. See Giannone et al. (2021) for a related discussion on the desirability to learn the appropriate degree of sparsity from data in social science applications, and a related spike-and-slab proposal in a regression setting.

We next discuss our default elicitation for the prior $\pi(\rho \mid \eta)$. The guiding principle was to set a minimally-informative prior, so that data may suitably update prior beliefs, while encouraging sparse solutions and preserving the interpretability of (8). Briefly, we set $\pi(\eta)$ to be proportional to C_η^{-1} times independent Gaussian prior densities on (η_0, η_1, η_2). Adding the term C_η^{-1} is a trick to simplify computation, since then C_η drops from the posterior density $\pi(\Theta, \eta \mid y)$. Wang (2015) argued that such cancellation of prior normalisation 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 was set such that the prior mean number of edges is proportional to p, which induces sparsity, and 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, i.e. the slab captures partial correlations of a larger magnitude than the slab. Finally, the prior on η_0 was set such that the slab has zero prior mean and such that sampling entries of ρ independently from the double-exponential priors in (8) 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.

Figure A.1 plots the implied prior marginal distribution on the ρ_{jk}'s for both the COVID-19 and stock market applications showing that the prior concentrates at 0 but also features thick tails to capture true non-zero ρ_{jk}'s. The corresponding posteriors (Figure A.1, bottom panels) set significant mass away from zero, suggesting that the prior shrinkage towards 0 was not excessive. Section A.3 provides further details and lists the hyper-parameter values used in our examples. Our code contains an implementation of our prior elicitation method.

3.3 Beyond Gaussian data

In certain applications such as our stock market example, data exhibit non-Gaussian behavior such as thick tails and asymmetries, even after taking logarithmic or similar transforms (see the normality checks in Section C.3). To address this issue in this application we used a non-paranormal model, which can accommodate said departures from normality. The distribution of $y_i = (y_{i1}, \ldots, y_{ip})$ is non-paranormal if there exist strictly increasing functions $f_j : \mathbb{R} \to \mathbb{R}$ for $j = 1, \ldots, p$ such that the vector $f(y_i) := (f_1(y_{i1}), \ldots, f_p(y_{ip}))$ is Gaussian. Such a non-paranormal model may be estimated by first obtaining an estimate $\hat{f}$ from the data, for which we used the R package `huge` (Zhao et al., 2012), and subsequently applying our methodology to the transformed data $\hat{f}(y_i)$.

An interesting property of the non-paranormal family is that the graphical model can be interpreted as in the Gaussian case. The partial correlation between the transformed $f_j(y_{ij})$ and $f_k(y_{ik})$ is zero if and only if (y_{ij}, y_{ik}) are conditionally independent. Partial correlations retain an interesting interpretation in the trans-elliptical family: zero partial correlation $\rho_{jk} = 0$ indicates that y_{ij} is linearly independent with any transformation of y_{ik} (Rossell and Zwiernik, 2021).

4 Computation and inference

4.1 Network GLASSO

We first describe how to optimise (2) for a fixed β, and subsequently how to estimate $\hat{\beta}$. The main idea is that, since $\lambda_{jk} = \lambda_{jk}(A, \beta)$ are fixed for a fixed β, the network GLASSO objective in (2) is a special case of the GOLAZO class of models in Lauritzen and Zwiernik (2020). Motivated by the desire to penalise positive and negative partial correlations differently, GOLAZO algorithms consider Gaussian graphical models with likelihood penalties of the form

$$\sum_{j=1}^{p} \sum_{k \neq j} \max \left\{ L_{jk} \rho_{jk}, U_{jk} \rho_{jk} \right\}, \tag{9}$$

where $-\infty \leq L_{jk} \leq 0 \leq U_{jk} \leq \infty$ are fixed. Noting that $\lambda|x| = \max\{-\lambda x, \lambda x\}$ for positive λ, we see that the penalty in (2) is in the form of (9) with $L_{jk} = -\lambda_{jk}$ and $U_{jk} = \lambda_{jk}$. (2) is a convex problem that can be efficiently solved using a block-coordinate ascent algorithm similar to that proposed for GLASSO in (Banerjee et al., 2008). An R package is provided for GOLAZO at https://github.com/pzwiernik/golazo.

Obtaining $\hat{\beta}_{\mathrm{BIC}}$ requires maximising $\mathrm{BIC}(\beta)$ in (4). As usual when using information criteria to set

regularisation parameters, this is a non-concave function of β that exhibits discontinuities. We propose two optimisation approaches. In cases where only one or two external networks are available and p is moderate ($p \leq 200$, say) we propose a grid-search akin to that used to set the regularisation parameter in standard GLASSO. Section A.1 contains several analytic upper bounds to facilitate such a search. However, the dimension of the hyper-parameter β grows with the number K of external networks, hence grid searches are very costly when $K \geq 3$ and p is large. In these settings, we propose using Bayesian optimisation. Briefly, Bayesian optimisation first evaluates the objective function $\mathrm{BIC}(\beta)$ at a few values of the hyper-parameter β and uses a Gaussian process to estimate $\mathrm{BIC}(\beta)$ for all β. Next, an acquisition function to propose new β values at which to evaluate $\mathrm{BIC}(\beta)$, which are then used to update the Gaussian process estimate. In particular we use the R package $\texttt{rBayesianoptimisation}$ (Yan, 2016), with the 'ucb' acquisition function and maximum function evaluations as $15 + 5Q$, where Q is the number of considered networks. In our examples Bayes optimisation returned virtually identical results to a grid search, but incurred a significantly lower computational cost when $\mathrm{BIC}(\beta)$ is hard to evaluate by requiring many fewer evaluations compared with the grid search alternative.

4.2 Spike-and-slab

The full parameter of interest is $(\mathrm{diag}(\Theta), \rho, \eta)$, where $\eta = (\eta_0, \eta_1, \eta_2)$ are the hyper-parameters in (8). To approximate their posterior distribution $\pi(\mathrm{diag}(\Theta), \rho, \eta \mid y)$ given the data y we used Hamiltonian Monte Carlo (see Neal (2011) for a review). Specifically, we developed an $\texttt{R}$ implementation using the $\texttt{Stan}$ software (Carpenter et al., 2017), as well as a Python implementation using the $\texttt{NumPyro}$ package (Phan et al., 2019). Sections A.2 and A.4 describe further implementation details and our code provides both implementations. The purpose of the $\texttt{R}$ version is to make our methods available to the ample $\texttt{R}$ community, whereas $\texttt{NumPyro}$ provides significant computational savings by using improvements in automatic differentiation and enabling the use of GPUs. The savings were substantial, Section D demonstrates that greater than an order of magnitude speed up was possible even in simple experimental settings.

The output of both implementations are N posterior samples $(\mathrm{diag}(\Theta^{(i)}), \rho^{(i)}, \eta^{(i)})$ for $i = 1, \ldots, N$ that can be used to approximate the posterior distribution or suitable summaries such as the marginal posterior mean and standard deviation of any parameter. Of particular interest to us is to estimate the posterior probability for the presence of an edge between any two nodes (j, k), i.e. that the partial correlation ρ_{jk} was generated by the slab in (8). We next discuss how to estimate said posterior probability using the posterior samples.

To ease notation re-write the prior as

$$\pi(\rho_{jk} \mid \eta) = (1 - w_{jk}(\eta))\pi_0(\rho_{jk} \mid \eta) + w_{jk}(\eta)\pi_1(\rho_{jk} \mid \eta) \tag{10}$$

where $\pi_0(\rho_{jk} \mid \eta)$ is the spike prior density, $\pi_1(\rho_{jk} \mid \eta)$ the slab prior density, and $w_{jk}(\eta)$ the slab prior probability. The idea is that any ρ_{jk} generated by the spike takes a near-zero value, i.e. the partial correlation is either truly zero or small enough to be practically irrelevant. Let $z_{jk} = 1$ indicate that ρ_{jk} was generated from the slab and $z_{jk} = 0$ that it was generated from the spike, i.e. $P(z_{jk} = 1 \mid \eta) = w_{jk}$. A measure of evidence in favor of the presence of the edge is the posterior probability

$$P(z_{jk} = 1 \mid y) = \int P(z_{jk} = 1 \mid \rho_{jk}, \eta)\pi(\rho_{jk}, \eta \mid y)d\rho_{jk}d\eta, \tag{11}$$

where from Bayes rule

$$P(z_{jk} = 1 \mid \rho_{jk}, \eta) = \frac{w_{jk}(\eta)\pi_1(\rho_{jk} \mid \eta)}{(1 - w_{jk}(\eta))\pi_0(\rho_{jk} \mid \eta) + w_{jk}(\eta)\pi_1(\rho_{jk} \mid \eta)}. \tag{12}$$

Given B posterior samples from $\pi(\rho, \eta \mid y)$, (11) may be easily estimated by

$$\hat{P}(z_{jk} = 1 \mid y) = \frac{1}{N}\sum_{I=1}^{N} P(z_{jk} = 1 \mid \rho_{jk}^{(I)}, \eta^{(I)}) \tag{13}$$

The description above applies in a full Bayesian treatment where η has a posterior distribution, in our empirical Bayes framework we simply replaced η by $\hat{\eta}$ in (10)-(13).

Our decision rule is to include edge (j, k) whenever $\hat{P}(z_{jk} = 1 \mid y) \geq t$ for some threshold $t \in [0, 1]$. We used $t = 0.95$. In problems where the goal is to estimate Θ it is customary to use $t = 0.5$, see Barbieri and Berger (2004). In contrast, in structural learning where one seeks to control the posterior expected false discovery proportion below some given level α, Müller et al. (2004) showed that the optimal threshold maximising statistical power is to set the largest t such that

$$\frac{1}{|D|} \sum_{(j,k)\in D} \hat{P}(z_{jk} = 0 \mid y) \leq \alpha$$

where D is the set of included edges. In particular, setting $t = 1 - \alpha$ ensures that the posterior expected false discovery proportion is below α.

4.3 Empirical Bayes

The empirical Bayes estimate $\hat{\eta}$ discussed in Section 3.2 requires marginalizing the joint posterior $\pi(\Theta, \eta \mid y)$. This is possible given N posterior samples $(\Theta^{(i)}, \eta^{(i)})$ for $i = 1, \ldots, N$ from the latter,

17

since then by definition $\eta^{(i)}$ are samples from $\pi(\eta \mid y)$. Then one may obtain $\hat{\eta}$ by maximising a kernel density estimate of $\pi(\eta \mid y)$, for example. Given that the accuracy of kernel density estimators degrades as dimensionality grows, in our examples when $\dim(\eta) > 2$ we instead obtain marginal mode estimators $\hat{\eta}_j = \arg\max_{\eta_j} \pi(\eta_j \mid y)$.

5   Simulation study

We conducted a simulation study to illustrate two important practical points. First, that when the network data are informative regarding the structure of Θ, incorporating said data improves inference. Second as just as important, that when the network data are useless inference does not suffer too much. To this end, we compared standard GLASSO with the network GLASSO of Section 3.1 and the network graphical spike-and-slab of Section 3.2 in several settings. We also considered the siGGM method Higgins et al. (2018), which is analogous to the network GLASSO in (2) but hyper-parameters are set to enforces the assumption that the network data are related to Θ, rather than learning from data whether this is the case or not. As discussed in Section 4, the network GLASSO hyper-parameters β are set via the BIC using grid-search optimisation, and the spike-and-slab hyper-parameters η using empirical Bayes. We considered a setting where there is a single binary network A with entries $a_{jk} \in \{0, 1\}$ and considered $p \in \{10, 50\}$ and sample sizes $n \in \{100, 200\}$ (results for $n = 500$ are in Table A.2). We then generated 50 independent datasets where $y_i \sim \mathcal{N}(0, \Theta^{-1})$ independently across $i = 1, \ldots, n$. 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 (i.e. Θ_{jk} where $|j - k| > 1$), a proportion of $0.5/p$ were set to non-zero values uniformly spaced in $[-0.1, 0.1]$ (i.e. the number of edges in the graphical model grows linearly with p).

We consider a setting where the network data are useless (independent network), and two settings where they are increasingly informative. To measure the degree to which the network data $a_{jk} \in \{0, 1\}$ is informative we count the proportion of overlaps where $a_{jk} = \mathrm{I}(\Theta_{jk} \neq 0)$, i.e. 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 $a_{jk} = 1$ 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 $a_{jk} = 1$ 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.

Code to reproduce our simulations is available in the GitHub repository. For each setting, we report the mean squared estimation error (MSE), the false discovery rate (FDR), and the false negative rate (Benjamini and Hochberg, 1995). The FDR is the expected proportion of false positive edges among the edges estimated to be present, a measure of type I error, whereas the FNR is the expected proportion of false negative edges among those not reported to be present, which measures statistical power. Under the GLASSO methods, an edge is declared if the corresponding estimate of ρ_{jk} was non-zero (rounded to 5 decimal places). For the spike-and-slab model an edge is declared when the posterior probability that ρ_{jk} arises from slab (12), conditional on empriical Bayes estimates $\hat{\eta}$ is above 0.95.

Table 1 summarises the results. For all sample sizes, the network GLASSO significantly reduced the MSE when the network data were mildly or strongly informative ($A_{0.75}$ and $A_{0.85}$), whereas it attained a similar MSE to standard GLASSO in the uninformative network setting (A_{ind}). The FDR was significantly above the usually accepted level of 0.05. Regarding the spike-and-slab formulations, they consistently achieved an FDR below 0.05 and a small FNR, and in large p situations a further improvement of the MSE compared with the network-GLASSO methods. Adding network data improved the spike-and-slab MSE and FNR, particularly when p was large relative to n. The FDR did not noticeably improve, but it was already near-zero when not using the network data. 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 minutes to run, whereas the `NumPyro` spike-and-slab implementation took close to 20 minutes (and `Stan` nearly 2 hours), see Section D for further details.

We stress that when the network data are useless (A_{ind}) the performance of Network GLASSO

remained similar to GLASSO, and that of Network SS to that of a standard spike-and-slab. In contrast the performance of SIGGM was poor in this setting, illustrating the practical value of assessing whether the network data is useful for inference, as done in our two 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 significantly above 0.05.

6   Results

6.1   COVID-19 infection rates

Recall that the outcomes are log-infection rates for USA counties during $n = 97$ weeks and that a regression model was fit to account for various factors driving the mean infection rates. These included week and county indicators, temperature and vaccination rate and serial correlation terms, see Section 2. 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 three 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 Section B for more details). Pearson's correlation between A_1 and A_2 is 0.746, i.e. there is a large overlap in the information given by both networks and it is hence desirable to use a principled model to disentangle their effects.

As a first exercise, we used network GLASSO to determine what network datasets are informative with respect to the target partial correlations. As $p = 332$ is large and the hyper-parameter dimension is $\dim(\beta) = 4$, we estimated $\hat{\beta}_{\mathrm{BIC}}$ using Bayesian optimisation, as described in Section 4. Table 2 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. The estimated coefficients for both networks $(\hat{\beta}_1, \hat{\beta}_2)$ were negative, i.e. counties that are close geographically or highly-connected at Facebook are regularised less. The larger coefficient $\hat{\beta}_2$ in the joint model suggests that the effect of the Facebook network is greater. Interestingly, the three network-regularised solutions were significantly sparser relative to the 628 edges detected by GLASSO.

Despite these solutions being sparser, they included some edges that were not included by GLASSO. Figure 2a shows edges that were only selected when adding the geographical network A_1, which largely

Table 1: Simulation results under non, mildly and strongly informative networks A_{ind}, $A_{0.75}$ and $A_{0.85}$. For SS and network SS models edges declared when posterior probability > 0.95.

	n	$p = 10$			$p = 50$		
		MSE	FDR	FNR	MSE	FDR	FNR
GLASSO	100	0.350	0.370	0.098	3.505	0.442	0.292
Network GLASSO, $A_{ind.}$	100	0.354	0.340	0.122	3.623	0.392	0.306
Network GLASSO, $A_{0.75}$	100	0.291	0.258	0.093	2.847	0.421	0.251
Network GLASSO, $A_{0.85}$	100	**0.170**	0.174	0.120	2.246	0.426	0.223
SS	100	0.222	**0.000**	0.086	1.611	**0.000**	0.023
Network SS, $A_{ind.}$	100	0.237	0.003	0.082	1.631	0.004	0.025
Network SS, $A_{0.75}$	100	0.234	0.007	0.073	1.462	0.005	0.023
Network SS, $A_{0.85}$	100	0.189	0.047	0.060	**1.280**	0.002	0.022
SIGGM, A_{ind}	100	0.534	0.683	0.047	4.815	0.866	0.017
SIGGM, $A_{0.75}$	100	0.304	0.492	0.019	3.203	0.837	0.010
SIGGM, $A_{0.85}$	100	0.197	0.385	**0.028**	2.749	0.794	**0.009**
GLASSO	200	0.184	0.416	0.022	1.794	0.476	0.181
Network GLASSO, $A_{ind.}$	200	0.201	0.378	0.040	1.871	0.439	0.189
Network GLASSO, $A_{0.75}$	200	0.161	0.309	0.022	1.515	0.412	0.181
Network GLASSO, $A_{0.85}$	200	0.096	0.204	0.098	1.241	0.388	0.173
SS	200	0.109	**0.000**	0.056	0.672	0.002	0.017
Network SS, $A_{ind.}$	200	0.127	0.007	0.053	0.671	**0.002**	0.017
Network SS, $A_{0.75}$	200	0.114	0.007	0.048	0.597	0.003	0.015
Network SS, $A_{0.85}$	200	**0.091**	0.023	0.041	**0.527**	0.002	0.015
SIGGM, A_{ind}	200	0.273	0.666	**0.015**	2.108	0.839	0.009
SIGGM, $A_{0.75}$	200	0.181	0.487	**0.015**	1.470	0.797	0.009
SIGGM, $A_{0.85}$	200	0.105	0.381	0.026	1.138	0.751	**0.008**

correspond to counties that are close to each other. Figure 2b shows an analogous plot when using the Facebook network A_2, interestingly there are connections between faraway counties in the west, north-east and south-east. Figure B.8 further portrays the estimated graphical model when using both networks.
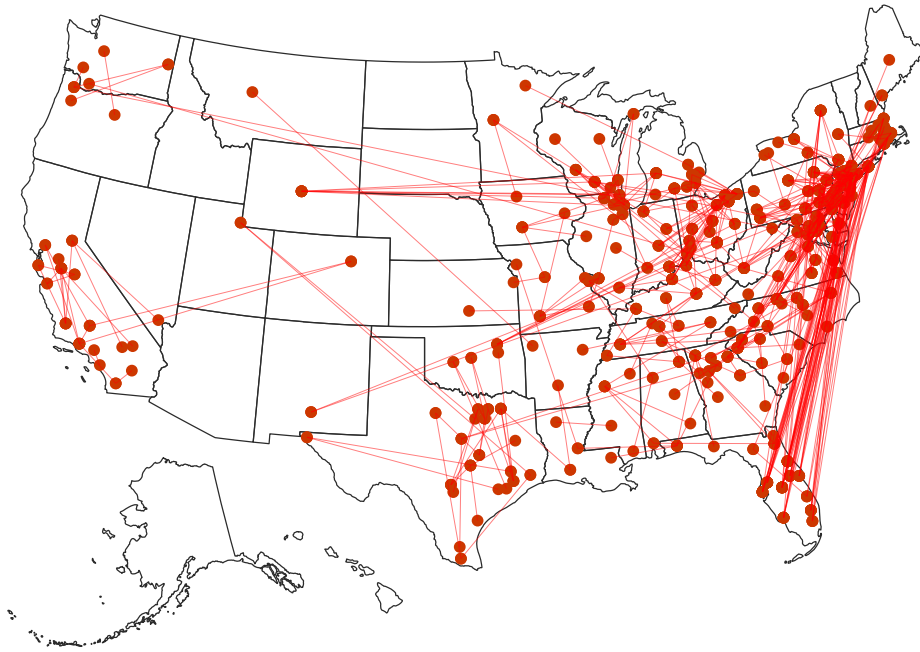
Table 2: Eight models for the COVID-19 data. A_1, A_2 and A_3: networks defined by $1/\log(Geodist)$, $\log(Facebook)$ and $A_3 = \log(1 + Flights)$. BIC values account for the extra hyper-parameters in the network GLASSO models. 10-fold: 10-fold cross-validated log-likelihood

Method	BIC	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	Edges	10-fold
GLASSO	23158.558	-1.376				2637	113.208
Network GLASSO- A_1	16237.865	0.230	-1.053			1427	122.692
Network GLASSO- A_2	15207.178	0.738		-1.301		1430	122.516
Network GLASSO- A_3	24227.657	-1.295			-0.085	2602	102.794
Network GLASSO- A_1 & A_2	**15064.079**	1.500	0.355	-1.695		1197	**124.372**
Network GLASSO- A_1 & A_3	16057.853	0.527	-1.193		0.531	1377	119.583
Network GLASSO- A_2 & A_3	15217.319	0.493		-1.131	0.377	1339	122.396
Network GLASSO- A_1, A_2 & A_3	15448.091	0.212	-0.063	-1.093	-0.103	1598	121.276

To further assess the relative performance of the eight 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 remarkably sparser (1,197 vs. 2,637 edges).

We next applied our spike-and-slab framework to obtain further insights on how the proportion of edge connections, as well as the mean partial correlation, depend on the two networks. We initially ran 20,000 MCMC iterations, thinning to 1 in every 10, to sample from $\pi(\text{diag}(\Theta), \rho, \eta \mid y)$. The resulting chains for network hyperparameters η had average effective sample size (ESS) of 473.8 and average R-hat value of 1.004, providing us with sufficient confidence to use these chains to do inference on η and produce empirical Bayes estimates. We then ran a second MCMC, fixing $\hat{\eta}$, for 4,000 iterations thinning to 1 in every 10. The resulting chains for partial correlations ρ had an average ESS of 372.8 and an average R-hat value of 1.003, suggesting that the chain converged.

As discussed earlier, the bottom panels in Figure 1 display 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 Figure B.6. Table 3 presents the corresponding (empirical Bayes) hyper-parameter estimates, and Figure B.7 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

(a) Edges identified by Network GLASSO - A_1 (geographical network) but not by GLASSO

(b) Edges identified by Network GLASSO - A_2 (Facebook network) but not by GLASSO

Figure 2: Edges identified by Network GLASSO but not by standard GLASSO.

variance (respectively) of the non-zero ρ_{jk}, i.e. the slab location and variance parameters, increase for counties that are strongly connected in the network. Similarly, positive entries in η_2 indicates a higher probability of there being a non-zero partial correlation between such counties. Table 3 hence shows that counties strongly connected in the Facebook and geographic networks had more non-zero partial correlations (relative to less connected counties), and that both the mean and variance of the partial correlations were also larger. The flight passenger network was estimated to have no effect on there being a non-zero partial correlation, nor on their mean, and a mild effect on the variance of non-zero partial correlation (in agreement with the BIC and cross-validation results in Table 2). The coefficients for the Facebook network are larger in absolute value than those of the geographical network indicating that the Facebook network has the stronger association with the dependence in COVID-19 rates. This is further illustrated in Figure B.7. These results illustrate the greater flexibility provided by the network spike-and-slab models to portray the relation between the network data and the partial correlations. For completeness, Table B.1 summarises the selected graphical model under a 0.5 and 0.95 posterior probability threshold for declaring an edge.

Altogether, our results support that there is a fairly strong association between social media connections and the co-evolution of the pandemic, even when accounting for geographical closeness and a number of factors driving the mean structure, and that said association is not driven by airplane travel.

Table 3: Network spike-and-slab empirical Bayes (marginal MAP) estimates and 95% posterior intervals for COVID-19 data. 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 includes 0.

	Intercept	A_1	A_2	A_3
η_0 (slab location)	-0.008	0.006	0.017	**0.0**
95% interval	(-0.009, -0.005)	(0.003,0.008)	(0.014,0.018)	(-0.002,0.002)
η_1 (slab dispersion)	2.285	0.054	0.178	-0.071
95% interval	(2.110, 2.507)	(0.003, 0.105)	(0.108, 0.240)	(-0.144, -0.002)
η_2 (slab probability)	-2.694	0.336	0.771	**-0.1**
95% interval	(-3.088,-2.397)	(0.154, 0.513)	(0.608, 0.949)	(-0.247, 0.047)

6.2 Stock market excess returns

Recall that the outcomes are log-daily excess returns of $p = 366$ US companies. The first network is an economic risks network A_1 where $a_{jk}^{(1)}$ is the Pearson's correlation between vectors of economic risks faced by firms j and k. The rth element of these vectors is $\log(1 + \text{prop}_r)$ where prop_r is proportion of 10-K terms that reflect the rth economic risk according to the dictionaries of Baker et al. (2019). A_2 is the equivalent but for vectors of policy risks. Pearson's correlation between the two networks was 0.301, suggesting that they provide largely different information. See Section C for a description of the data pre-processing.

We firstly run GLASSO using no network data and then network GLASSO using only the Economic network, only the Policy network, and finally using both networks. Table 4 compares these four models. The model including both networks attained the best BIC value and their estimated parameters $(\hat{\beta}_1, \hat{\beta}_2)$ are both negative. That is, partial correlations between companies that have large connections in the network are more likely to be non-zero, and are hence less regularised. The estimated graphical model when using both networks is sparser than under standard GLASSO.

Table 4: Four models for the stock market data. A_1 is the Economic network, A_2 the Policy network. BIC values account for the extra hyper-parameters in the network GLASSO models. 10-fold is the 10-fold cross-validation log-likelihood

Method	BIC	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	Edges	10-fold
GLASSO	74078.33	-1.639			2623	-467.128
Network GLASSO A_1	72459.55	-1.137	-0.677		2770	**-463.683**
Network GLASSO A_2	73857.75	-1.107		-0.776	2211	-467.128
Network GLASSO A_1 & A_2	**72392.50**	-0.176	-0.932	-0.671	2058	-467.42

To further assess the four models we evaluated their out-of-sample log-likelihood using 10-fold cross-validation. The models incorporating only the economic risks network performed best in this prediction exercise.

To gain further insights into the relation between partial correlations and the network data, we applied our spike-and-slab framework. We initially run 10,000 MCMC iterations, thinning to 1 in every 10, to sample from $\pi(\text{diag}(\Theta), \rho, \eta \mid y)$. The resulting chains for network hyper-parameters η had an average effective sample size (ESS) of 453.9 and an average R-hat value of 1.002, suggesting MCMC convergence. From these samples we obtained empirical Bayes estimate $\hat{\eta}$ and then ran a
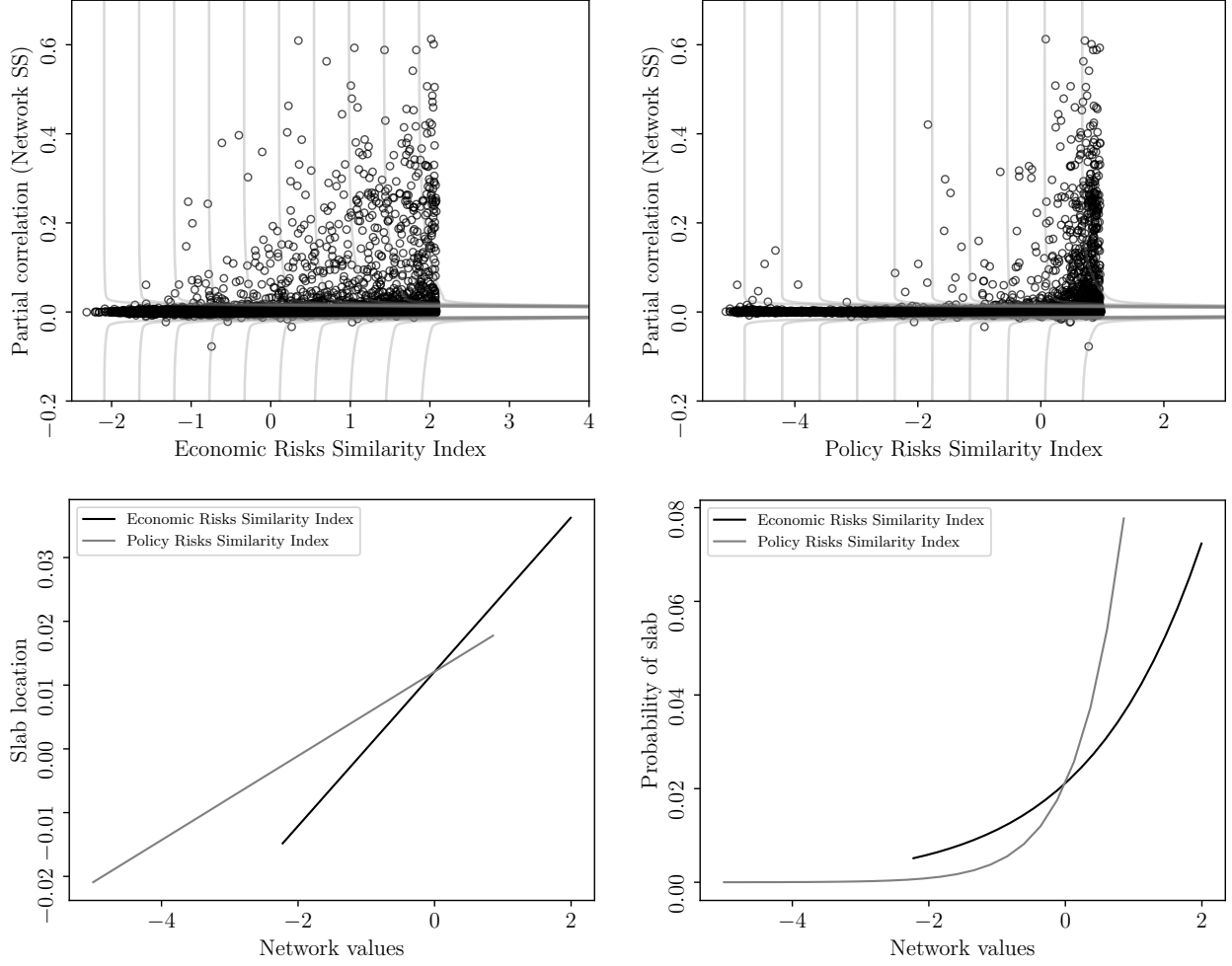
Figure 3: Residual partial correlations of the stock market excess returns across firms vs Economy risk (left) and Policy risk (right). Top panel: fitted spike-and-slab distributions and fitted partial correlations estimated with network spike-and-slab model. Bottom panel: Stock-market data: Slab location (**left**) and slab probability (**right**) as a function of both networks estimated by empirical Bayes.

second MCMC, fixing $\hat{\eta}$, for 4,000 iterations thinning to 1 in every 10. The resulting chains for partial correlations ρ had an average ESS of 370.0 and an average R-hat of 1.003.

Figure 3 shows the estimated spike-and-slab distributions for the partial correlations as a function of both networks, and Table 5 the corresponding hyper-parameter estimates. Recall that the network values were standardised so although they are correlations they do not lie in $[-1, 1]$. Companies with strong connections in either network are estimated to have both a larger probability of a non-zero partial correlation (positive η_2) and a larger mean non-zero partial correlation (positive η_0). Interestingly, the policy network had the larger positive effect on the probability of a non-zero partial correlation, but its effect on their mean is smaller than the economic network's.

In short, both the economic and policy risk networks appear to contain independent information about the partial correlations among firms' stock returns. The GLASSO model suggests economic risks are more associated with such correlations: when both networks are included, the estimated impact of the economic risk network on the regularisation is stronger than that of the policy risk network and the out-of-sample predictive exercises prefers only the model with the economic network. At the same time, the additional structure of the spike-and-slab model reveals a more subtle pattern. The relationship between the strength of the connection in the economic risks networks and the partial correlation evolves more smoothly than for the strength of the connection in the policy network. Only when firms are strongly connected in the latter is there an inferred impact on their partial correlations.

Since the pioneering work of Hassan et al. (2019), economists have used word-count-based approaches to measure and evaluate firm-level exposure to political and policy risks. But exposure to policy risks is in part a function of economic risks: for example, firms exposed to air travel via their business model will also be exposed to regulation of the Federal Aviation Administration. Our findings suggest that, after accounting for shared economic risks, shared policy risks only matter for excess return co-movement once firms are strongly connected.

For completeness, Table C.1 summarises the selected graphical model under a 0.5 and 0.95 posterior probability threshold for declaring an edge.

7 Discussion

We believe that our two frameworks to regress a graphical model on network data should have interest beyond our motivating COVID-19 and stock market applications. Specifically, the Bayesian framework provides a rich model to depict the probability that parameters are non-zero as well as the distribution

Table 5: Network spike-and-slab empirical Bayes (marginal MAP) estimates and 95% posterior credible intervals for the stock market data. A_1 is the Economic network, A_2 the Policy network.

	intercept	A_1	A_2
η_0 (slab location)	0.012	0.012	0.007
95% interval	(0.01, 0.014)	(0.009,0.014)	(0.003,0.009)
η_1 (slab dispersion)	2.943	0.284	-0.353
95% interval	(2.827,3.063)	(0.194, 0.369)	(-0.447,-0.249)
η_2 (slab probability)	-3.834	0.644	1.59
95% interval	(-4.122,-3.553)	(0.519, 0.808)	(1.267, 1.93)

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 hyperparameters ameliorated the consequences in a worst-case scenario where one introduces uninformative external data.

Our COVID-19 application found that geographical closeness and a Facebook connectivity network were both informative about the dynamics of COVID-19 cases, allowing for the estimation of a sparser graphical model that predicted better out-of-sample. The Facebook network had a greater association with COVID-19, suggesting for example to consider social media campaigns to help improve disease outcomes. We stress that our findings should be understood as associations between social media and disease progression, rather than causal connections. For example, although Nyhan et al. (2023) found that Facebook feeds are skewed towards politically like-minded sources, there was little evidence that increasing exposure to more diverse sources reduced polarization. It is therefore possible that the Facebook index serves as a proxy for like-minded attitudes, rather than shared social media causally driving people to have similar attitudes. In the stock market application we found similarity in firm risk exposures to economic and policy risks were both informative of the firms co-evolution in excess returns. While the policy network appeared to have a stronger relationship to whether two firms were connected, the economic network was more predictive of the behaviour of connected observations. These findings suggest that by understanding better the role played by risks declared by firms on their stock market behavior, it may be possible to design better portfolio strategies.

Further methodological work could consider richer relationships for how the graphical models can depend on the networks, e.g. by considering non-parametric models. Another interesting avenue would be developing computational methods to scale our algorithms to even higher dimensions.

Acknowledgements

References

Hunt Allcott, Levi Boxell, Jacob Conway, Matthew Gentzkow, Michael Thaler, and David Yang. Polarization and public health: Partisan differences in social distancing during the coronavirus pandemic. *Journal of public economics*, 191:104254, 2020.

Scott R. Baker, Nicholas Bloom, Steven J. Davis, and Kyle J. Kost. Policy News and Stock Market Volatility. *National Bureau of Economic Research Working Paper Series*, (w25720), 2019.

Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research*, 9:485–516, 2008.

M.M. Barbieri and J.O. Berger. Optimal predictive model selection. *The Annals of Statistics*, 32(3): 870–897, 2004.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57(1):289–300, 1995.

Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul A. Szerlip, Paul Horsfall, and Noah D. Goodman. Pyro: Deep universal probabilistic programming. *Journal of Machine Learning Research*, 20:28:1–28:6, 2019.

Yunqi Bu and Johannes Lederer. Integrating additional knowledge into the estimation of graphical models. *The international journal of biostatistics*, 18(1):1–17, 2021.

U.S. Census Bureau. Us 2019 population data, 2020. Available from github: https://www2.census.gov/programs-surveys/popest/tables/2010-2019/counties/totals/.

Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1):1–32, 2017.

Jack Storror Carter, David Rossell, and Jim Q Smith. Partial correlation graphical lasso. *arXiv preprint arXiv:2104.10099*, 2021.

Alberto Cassese, Michele Guindani, and Marina Vannucci. A Bayesian integrative model for genetical genomics with spatially informed variable selection. *Cancer informatics*, 13:S13784, 2014.

J. Chen and Z. Chen. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008.

Ting-Huei Chen, Nilanjan Chatterjee, Maria Teresa Landi, and Jianxin Shi. A penalized regression framework for building polygenic risk models based on summary statistics from genome-wide association studies and incorporating external information. *Journal of the American Statistical Association*, 116(533):133–143, 2021.

Sharon Chiang, Michele Guindani, Hsiang J Yeh, Zulfi Haneef, John M Stern, and Marina Vannucci. Bayesian vector autoregressive model for multi-subject effective connectivity inference using multimodal neuroimaging data. *Human brain mapping*, 38(3):1311–1332, 2017.

CSSE. Covid19 infection rates, 2020a. Available from github: https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_US.csv.

CSSE. Government coronavirus policies, 2020b. Available from github: https://github.com/CSSEGISandData/COVID-19_Unified-Dataset.

CSSE. Daily average near-surface temperature, 2020c. Available from github: https://github.com/CSSEGISandData/COVID-19_Unified-Dataset/tree/master/Hydromet.

CSSE. U.s. vaccination data, 2020d. Available from github: https://github.com/govex/COVID-19/tree/master/data_tables/vaccine_data/us_data/time_series.

Steven J. Davis, Stephen Hansen, and Cristhian Seminario-Amez. Firm-Level Risk Exposures and Stock Returns in the Wake of COVID-19. Working Paper 27867, National Bureau of Economic Research, 2020.

Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222, 1987.

Edwin J. Elton and Martin J. Gruber. Estimating the Dependence Structure of Share Prices–Implications for Portfolio Selection. *The Journal of Finance*, 28(5):1203–1232, 1973. ISSN 0022-1082. doi: 10.2307/2978758.

Eugene F. Fama and Kenneth R. French. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56, 1993. ISSN 0304-405X. doi: https://doi.org/10.1016/0304-405X(93)90023-5. URL https://www.sciencedirect.com/science/article/pii/0304405X93900235.

Jianqing Fan, Yang Feng, and Yichao Wu. Network exploration via the adaptive LASSO and SCAD penalties. *Annals of Applied Statistics*, 3(2):521–541, 2009.

Yingying Fan and Cheng Yong Tang. Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society B*, 75(3):531–552, 2013.

Rina Foygel and Mathias Drton. Extended bayesian information criteria for gaussian graphical models. *Advances in Neural Information Processing Systems*, 23:604–612, 2010.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

L. Gan, N.N. Narisetty, and F. Liang. Bayesian regularization for graphical models with unequal shrinkage. *Journal of the American Statistical Association*, just-accepted:1–14, 2018.

Domenico Giannone, Michele Lenza, and Giorgio E. Primiceri. Economic Predictions With Big Data: The Illusion of Sparsity. *Econometrica*, 89(5):2409–2437, 2021. ISSN 0012-9682. doi: 10.3982/ECTA17842.

Shingo Goto and Yan Xu. Improving Mean Variance Optimization through Sparse Hedging Restrictions. *The Journal of Financial and Quantitative Analysis*, 50(6):1415–1441, 2015. ISSN 0022-1090.

Kathleen Weiss Hanley and Gerard Hoberg. Dynamic Interpretation of Emerging Risks in the Financial Sector. *The Review of Financial Studies*, 32(12):4543–4603, 2019. ISSN 0893-9454. doi: 10.1093/rfs/hhz023.

Tarek A Hassan, Stephan Hollander, Laurence van Lent, and Ahmed Tahoun. Firm-Level Political Risk: Measurement and Effects. *The Quarterly Journal of Economics*, 134(4):2135–2202, 2019. ISSN 0033-5533. doi: 10.1093/qje/qjz021.

Ixavier A Higgins, Suprateek Kundu, and Ying Guo. Integrative bayesian analysis of brain functional networks incorporating anatomical knowledge. *NeuroImage*, 181:263–278, 2018.

Matthew D Hoffman and Andrew Gelman. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

Theresa Kuchler, Dominic Russel, and Johannes Stroebel. The geographic spread of covid-19 correlates with the structure of social networks as measured by facebook. *Journal of Urban Economics*, page 103314, 2021.

Markku Kuismin and Mikko J Sillanpää. Mcpese: Monte carlo penalty selection for graphical lasso. *Bioinformatics*, 37(5):726–727, 2021.

Steffen Lauritzen and Piotr Zwiernik. Locally associated graphical models and mixed convex exponential families. *arXiv*, 2008.04688:1–34, 2020. to appear in Annals of Statistics.

Harry Markowitz. Portfolio Selection. *The Journal of Finance*, 7(1):77–91, 1952. ISSN 0022-1082. doi: 10.2307/2975974.

P. Müller, G. Parmigiani, C. Robert, and J. Rousseau. Optimal sample size for multiple testing: the case of gene expression microarrays. *Journal of the American Statistical Association*, 99(468):990–1001, 2004.

Radford Neal. *MCMC using Hamiltonian dynamics*, pages 113–162. Chapman and Hall/CRC, 2011.

Bernard Ng, Gaël Varoquaux, Jean-Baptiste Poline, and Bertrand Thirion. A novel sparse graphical approach for multimodal brain connectivity inference. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 707–714. Springer, 2012.

Brendan Nyhan, Jaime Settle, Emily Thorson, Magdalena Wojcieszak, Pablo Barberá, Annie Y Chen,

Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, et al. Like-minded sources on facebook are prevalent but not polarizing. *Nature*, 620(7972):137–144, 2023.

Christine B Peterson, Francesco C Stingo, and Marina Vannucci. Joint bayesian variable and graph selection for regression models with network-structured predictors. *Statistics in medicine*, 35(7): 1017–1031, 2016.

Du Phan, Neeraj Pradhan, and Martin Jankowiak. Composable effects for flexible and accelerated probabilistic programming in numpyro. *arXiv*, 1912.11554:1–10, 2019.

José Angel Pineda-Pardo, Ricardo Bruña, Mark Woolrich, Alberto Marcos, Anna C Nobre, Fernando Maestú, and Diego Vidaurre. Guiding functional connectivity estimation by structural connectivity in meg: an application to discrimination of conditions of mild cognitive impairment. *Neuroimage*, 101:765–777, 2014.

MA Quintana and DV Conti. Integrative variable selection via Bayesian model uncertainty. *Statistics in medicine*, 32(28):4938–4953, 2013.

V. Rockova and E.I. George. EMVS: The EM approach to Bayesian variable selection. *Journal of the American Statistical Association*, 109(506):828–846, 2014.

David Rossell and Piotr Zwiernik. Dependence in elliptical partial correlation graphs. *Electronic Journal of Statistics*, 15(2):4236–4263, 2021.

G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

J.G. Scott and J.O Berger. Bayes and empirical Bayes multiplicity adjustment in the variable selection problem. *The Annals of Statistics*, 38(5):2587–2619, 2010.

Marc Senneret, Yannick Malevergne, Patrice Abry, Gerald Perrin, and Laurent Jaffrès. Covariance Versus Precision Matrix Estimation for Efficient Asset Allocation. *IEEE Journal of Selected Topics in Signal Processing*, 10(6):982–993, 2016. ISSN 1941-0484. doi: 10.1109/JSTSP.2016.2577546.

Roger W Sinnott. Virtues of the haversine. *Sky and telescope*, 68(2):158, 1984.

Francesco C Stingo, Yian A Chen, Marina Vannucci, Marianne Barrier, and Philip E Mirkes. A Bayesian graphical modeling approach to microRNA regulatory network inference. *The annals of applied statistics*, 4(4):2024–2048, 2010.

Francesco C Stingo, Yian A Chen, Mahlet G Tadesse, and Marina Vannucci. Incorporating biological information into linear models: A Bayesian approach to the selection of pathways and genes. *The annals of applied statistics*, 5(3):1–24, 2011.

Hao Wang. Bayesian graphical LASSO models and efficient posterior computation. *Bayesian Analysis*, 7(4):867–886, 2012.

Hao Wang. Scaling it up: Stochastic search structure learning in graphical models. *Bayesian Analysis*, 10(2):351–377, 2015.

Lingxiao Wang, Xiang Ren, and Quanquan Gu. Precision matrix estimation in high dimensional Gaussian graphical models with faster rates. *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS 2016*, 51:177, 2016.

Tao Wang and Lixing Zhu. Consistent tuning parameter selection in high dimensional sparse linear regression. *Journal of Multivariate Analysis*, 102(7):1141–1151, 2011.

Yachen Yan. rbayesianoptimization: Bayesian optimization of hyperparameters. *R package version*, 1(0), 2016.

Ming Yuan and Yi Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

Giacomo Zanella and Gareth Roberts. Multilevel linear models, gibbs samplers and multigrid decompositions (with discussion). *Bayesian Analysis*, 16(4):1309–1391, 2021.

Yiyun Zhang, Runze Li, and Chih-Ling Tsai. Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association*, 105(489):312–323, 2010.

Tuo Zhao, Han Liu, Kathryn Roeder, John Lafferty, and Larry Wasserman. The huge package for high-dimensional undirected graph estimation in r. *The Journal of Machine Learning Research*, 13 (1):1059–1062, 2012.

Supplementary Material

Section A provides further details for the implementation of our network GLASSO and network spike-and-slab models. Section B contains further information related to our COVID-19 data application, including the data collection, preprocessing, linear model estimation and diagnostic checks, network specification and linearity check, as well as further results and figures. Section C provides analogous information for the stock market data. Lastly, Section D provides a performance comparison of the network GLASSO frequentist model with the network spike-and-slab Bayesian model using `Stan` and `NumPyro`. Code to implement all of our experiments and data pre-processing is available at https://github.com/llaurabat91/graphical-models-external-networks.

A Implementation details for network GLASSO and network spike-and-slab

A.1 Bounding the region for optimal GOLAZO hyperparameters β.

The GOLAZO algorithm (Section 8.1 in Lauritzen and Zwiernik (2020)) is a block coordinate descent algorithm where the j-th row is optimised with other entries of Σ fixed by solving a quadratic program

$$\min_d \ d^T(\Sigma_{\backslash j})^{-1}d \qquad \text{subject to } |\Sigma_{ij} - S_{ij}| \leq \lambda_{ij} \text{ for all } i < j \text{ and } \Sigma_{ii} = S_{ii} \text{ for all } i, \qquad \text{(A.1)}$$

where d contains the off-diagonal entries of the j-th row of Σ (the diagonal entry always satisfies $\Sigma_{jj} = S_{jj}$).

The following lemma guarantees that for large enough λ_{ij} the solution is to set all parameter estimates to zero.

Lemma A.1. If $\lambda_{jk} \geq |S_{jk}|$ for all $k \neq j$ then $d = 0$ optimises (A.1).

Proof. Under the given condition $d = 0$ is always feasible. Since $d = 0$ is also the global minimum, the result follows. □

We can therefore assume that $\lambda_{jk} < |S_{jk}|$ for at least one pair (j, k). That is, we may restrict attention to β satisfying

$$\max_{j \neq k} \log(\lambda_{jk}) \ = \ \max_{j \neq k} \beta_0 + \sum_{q=1}^{Q} \beta_q \bar{a}_{jk}^{(q)} \ \leq \ \max_{j \neq k} \log|S_{jk}|.$$

Note that this expression bounds the range of possible optima for each β_q given the rest, and in particular for β_0 we obtain

$$\beta_0 \leq \max_{j \neq k} \{\log |S_{jk}| - \sum_{q=1}^{Q} \beta_q \bar{a}_{jk}^{(q)}\},$$

which is $\leq \max_{j \neq k} \log |S_{jk}|$ at the initialisation step where $\beta_1 = \ldots = \beta_Q = 0$.

In particular, we propose the following procedure. First initialise $\hat{\beta}_0$ (the first entry in $\hat{\beta}$), such that $\hat{\lambda} = \exp(\hat{\beta}_0)$, where $\hat{\lambda}$ maximises the BIC in (4) over a univariate grid. Assuming that all variables in y_i are standardised to unit sample variance, the grid search is facilitated by Lemma A.1 and the analytic upper bound that

$$\hat{\beta}_0 \leq \log \left(\max_{k \neq j} \{|R_{jk}|\} \right), \tag{A.2}$$

where R is the empirical correlation matrix.

Second, we conduct a grid search on the whole vector β, with the first entry being centered around $\hat{\beta}_0$. The grid search is again facilitated by the Lemma A.1 which shows that one may restrict attention to β such that

$$\max_{j \neq k} \lambda_{jk} = \max_{j \neq k} e^{\beta_0 + \sum_{q=1}^{Q} \beta_q \bar{a}_{jk}^{(q)}} \leq 1 - |R_{jk}|$$

since increasing λ_{jk} beyond this bound will not change $\hat{\Theta}$. Within the grid search, we also use the solution obtained for a particular β as a warm start for subsequent values of β.

Further, the fact that Σ in (A.1) must be positive definite allows for the construction of further simple bounds. For every $i \neq j$ we necessarily have $\Sigma_{jk}^2 \leq \Sigma_{jj} \Sigma_{kk} = S_{jj} S_{kk}$, or equivalently, $\Sigma_{jk} \in [-\sqrt{S_{jj} S_{kk}}, \sqrt{S_{jj} S_{kk}}]$. It follows that, without loss of generality, we can restrict attention to that $\lambda_{jk} \leq \sqrt{S_{jj} S_{kk}} - |S_{jk}|$ giving

$$\beta_0 + \sum_{q=1}^{Q} \beta_q \bar{a}_{jk}^{(q)} \leq \log(\sqrt{S_{jj} S_{kk}} - |S_{jk}|) \qquad \text{for all } j \neq k.$$

A.2 Implementation of spike-and-slab

If the spike has a very small variance, or the slab has too bigger variance it can be difficult for an MCMC sampler to efficiently explore both spaces. We use a rescaling trick to facilitate efficient MCMC inference for the network spike-and-slab model. Rather than sample directly from $\pi(\rho)$ as defined by (8), for each ρ_{jk} we define latent variables $\tilde{\rho}_{jk}^{spike}$, $\tilde{\rho}_{jk}^{slab}$ and u_{jk}. We then sample

$$\tilde{\rho}_{jk}^{spike} \sim \text{DE}(0, 1), \quad \tilde{\rho}_{jk}^{slab} \sim \text{DE}(0, 1) \text{ and } u_{jk} \sim \text{Unif}[0, 1],$$

and set

$$\rho_{jk} = \text{I}(u_{jk} > w_{jk}) \left(s_0 \times \tilde{\rho}_{jk}^{spike} \right) + \text{I}(u_{jk} \le w_{jk}) \left(\eta_0^T a_{jk} + s_{jk} \times \tilde{\rho}_{jk}^{slab} \right).$$

It is straightforward to see that the marginal distribution of ρ_{jk} matches that defined in (8). Lastly, to make such an implementation suitable for MCMC samplers that require differentiability, we approximate the indicator $\text{I}(u_{jk} > w_{jk})$ with a sigmoid function

$$\text{I}(x \ge 0) \approx \sigma_k(x) = \frac{1}{1 + \exp(-kx)} \text{ for large } k,$$

taking $k = 100$.

A.3 Prior elicitation

We elicit spike-and-slab prior parameters (η_0, η_1, η_2) that encourage sparse solutions, avoid pathological values, and maintain their specified intuition whilst being minimally informative. We finish this section with a table of the values used in the simulations and in our applications. For interpretability, we treat the spike's scale parameter s_0 as a constant. Recall that the spike captures partial correlations ρ_{jk} that are considered to be 0 for all practical purposes, which here we consider to be $|\rho_{jk}| < 0.01$. We hence set s_0 such that the spike has most of its density below this threshold, i.e. $\Pi(\rho_{ij} \in (-\tau, \tau); s_0) = 0.95$, where $\tau = 0.01$. This gave the value $s_0 = 0.003$

Consider first the hyperparameters $(\eta_{00}, \eta_{10}, \eta_{20})$ defining the intercept of the regression of the slab's mean, variance, and prior probability on the network data. We set the priors

$$\eta_{00} \sim \mathcal{N}\left(0, g_0^2\right)$$
$$\eta_{10} \sim \mathcal{N}\left(m_1, g_1^2\right)$$
$$\eta_{20} \sim \mathcal{N}\left(m_2, g_2^2\right).$$

For the hyperparameters that capture the effect of each network $A^{(q)}$, where $q = 1, \ldots, Q$, we set

$$\eta_{0q} \sim \mathcal{N}\left(0, g_0^2\right)$$
$$\eta_{1q} \sim \mathcal{N}\left(0, g_1^2\right)$$
$$\eta_{2q} \sim \mathcal{N}\left(0, g_2^2\right).$$

Centering the prior of η_{00} at 0 encodes the absence of information about whether partial correlations are positive or negative on average. Similarly, centering the priors of $(\eta_{0q}, \eta_{1q}, \eta_{2q})$ at zero reflects no

prior knowledge on whether the network data are predictive of ρ and in which direction. To set the remaining hyperparameters we assume the networks have been standardised and conduct the prior elicitation for the average value of the networks (i.e. $\bar{a}_{jk}^{(q)} = 0$ for all networks q). As a result, our prior elicitation is invariant to the network(s) considered.

The prior on η_2 was set based on sparsity and minimal informativeness considerations. Specifically, we set the prior expected number of edges (non-zero partial correlations) to scale linearly with p, so that each node is expected to have a constant degree as p grows. When all networks are at their average value the slab prior probability is $w = 1/(1 + e^{-\eta_{20}})$. A standard non-informative prior on slab prior probabilities is a $\text{Beta}(m_w v_w, m_w(1 - v_w))$ distribution (Scott and Berger, 2010), where m_w is the prior mean and v_w is often interpreted as the prior 'sample size'. We take the minimally informative choice $v_w = 1$. Regarding m_w, we set it such that the prior expected number of edges is p. Since the prior expected number of edges is

$$\mathbb{E}\left[\sum_{j=1}^{p}\sum_{k<j}\mathbb{I}(\rho_{jk} \in \text{slab})\right] = \frac{p(p-1)}{2}w,$$

for $m_w = \frac{2}{(p-1)}$ the expected number of edges is p. Based on these considerations, we set the (m_2, g_2^2) featuring in the prior of η_{20} and η_{2q} so that the implied prior on w has the same mean and variance as the Beta prior described above.

Regarding the prior on η_1, we considered that for the slab to capture non-zero partial correlations its prior scale parameter at the average value of the networks $s_{jk} = s_0(1 + \exp\{-\eta_{10}\})$ should be significantly larger than that of the spike s_0. We hence set m_1 and g_1 such that the prior mode of s_1 is $10 \times s_0$, as well as $s_1 > 3 \times s_0$ with prior probability 0.99.

Finally, the prior on η_0 was set based on prior positive-definiteness considerations. Specifically, the positive-definiteness indicator $I(\rho \succ 0)$ induces dependence in the spike-and-slab prior density, i.e. it can produce a joint prior that is vastly different from the product of independent priors on each ρ_{jk}. Such a discrepancy is undesirable for prior interpretation, particularly in our setting where the priors and their hyperparameters are objects of interest that describe how ρ_{jk} depends on network data. To address this issue we set prior parameters such that the prior probability of ρ being positive definite when independently sampling its elements is at least 0.95. Conditional on the priors specified for (η_1, η_2), g_0 was set to the largest value (i.e. least informative) that guarantees at least 0.95 probability that ρ is positive-definite under independent sampling from the unconstrained spike-and-slab prior components.

Table A.1: Network spike-and-slab prior hyperparameters

	$p = 10$	$p = 50$	COVID-19 data ($p = 332$)	Stock data ($p = 366$)
s_0	0.003	0.003	0.003	0.003
g_0	0.145	0.152	0.002	0.002
m_1	-2.197	-2.197	-2.197	-2.197
g_1	0.661	0.661	0.3	0.35
m_2	-2.722	-6.737	-7.789	-10.16
g_2	3.278	3.395	1.02	1.81

A.3.1 Elicited values

Table A.1 presents the elicited values used in our simulations and real data examples. Code to elicit priors following the specification above for further examples is available in the GitHub repository. As the dimension of the data increases, only the prior for η_2 changes greatly. This is a result of the assumption that the number of edges grows linearly with p, and therefore Θ is *a priori* assumed more sparse for larger p.

To assess the impact of these default prior choices, it is useful to display the implied prior marginal distribution on the ρ_{jk}'s. Figure A.1 shows that in both the COVID-19 and stock market applications most of the prior probability is contained in $\rho_{jk} \in (-0.5, 0.5)$, which seems a sensible prior interval. The prior concentrates significant mass around 0, which induces shrinkage, but also features thick tails, which favors capturing truly non-zero ρ_{jk}'s. Indeed, the corresponding posteriors (Fig. A.1, bottom panels) set significant mass away from zero, suggesting that the prior shrinkage towards 0 was not excessive.

A.4 Reparametrisation of the network hyperparameters

An advantage of the Bayesian network spike-and-slab approach is that it allows us to do inference for the network hyperparameter as was done in Tables 3 and 5. Such inferences, however, require that the effective sample size (ESS) of the sampled hyperparameters is sufficiently high. We observed empirically that hyperparameters attain lower ESS. Although this phenomenon has not been studied in our graphical model settings, in hierarchical models it is well understood that parameters associated to higher levels have strictly slower MCMC mixing, and that said mixing can be improved by reparameterising the problem (Zanella and Roberts, 2021). We applied the following transformation
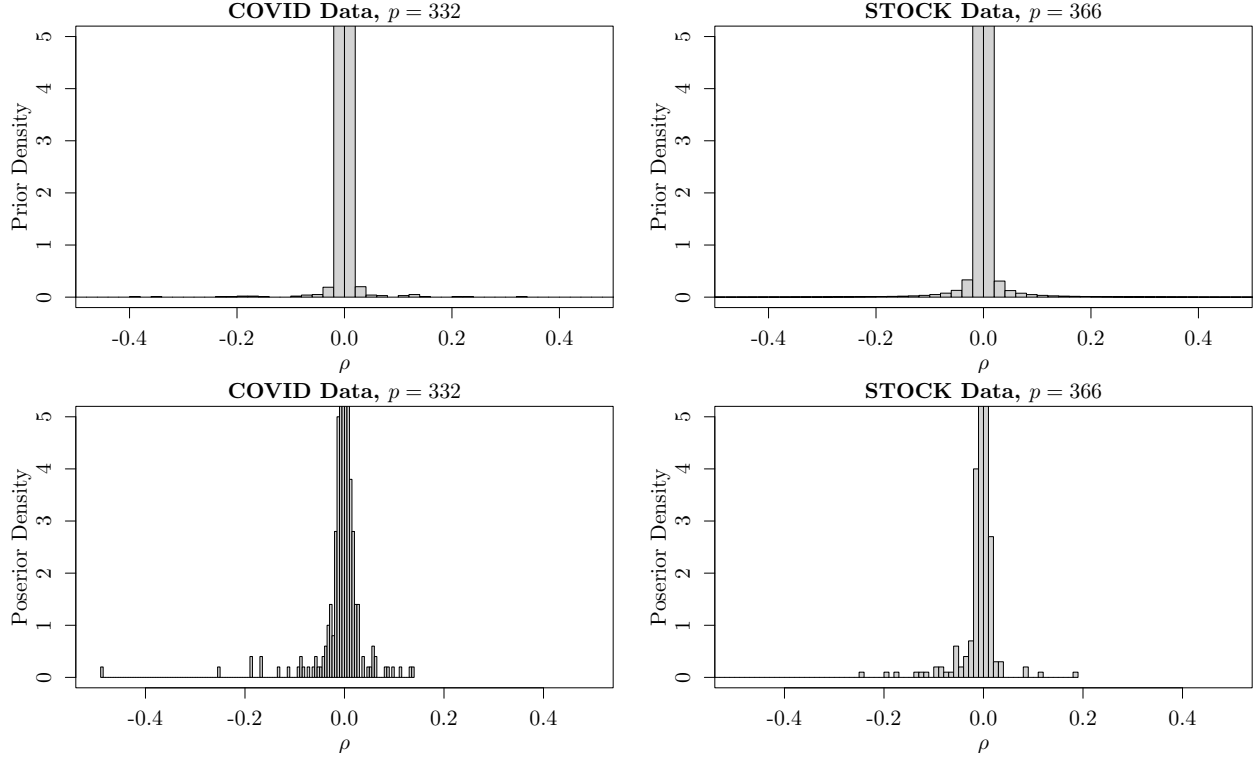
Figure A.1: Elicited prior distribution and posterior distribution for ρ_{jk}, $j = 1, \ldots, p$, $k < j$ for the COVID-19 data $p = 332$ and stock market data $p = 366$.

of the hyperparameters to facilitate their sampling.

Rather than sample directly from the priors for the hyperparameters as outlined in Section A.3, we reparameterised and sampled

$$\tilde{\eta}_{iq} \sim \mathcal{N}\left(0, \frac{p(p-1)}{2n}\right), \quad i = 1, 2, 3, \quad q = 0, 1, \ldots, Q.$$

The original η hyperparameters can then be recovered as

$$\eta_{i0} = m_i + \tilde{\eta}_{i0} \times g_i / \sqrt{p(p-1)/2n},$$
$$\eta_{iq} = 0 + \tilde{\eta}_{iq} \times g_i / \sqrt{p(p-1)/2n}, \quad i = 1, 2, 3, \quad q = 1, \ldots, Q,$$

where $m_0 := 0$. The idea behind this is to first standardise the η's to all have mean 0 and variance 1, before adjusting the variance of the $\tilde{\eta}$'s by the square-root of the ratio of the number of ρ's $(p(p-1)/2)$ from which the η's are learned, to the number of observations Y (n) from which the ρ's themselves are learned. Such a reparametrisation leaves the model completely unchanged, but we found this improved the ESS of the $\eta's$.

A.5 Additional simulation results

A.5.1 Simulations with $n = 500$

Table A.2 presents simulation results from Section 4 in the additional case where the sample size $n = 500$. These show that when n is large relative to p, network information helps to a lesser extent.

Table A.2: Simulation results for $n = 500$ under non, mildly and strongly informative networks A_{ind}, $A_{0.75}$ and $A_{0.85}$. For SS and network SS models edges declared when posterior probability > 0.95.

	n	$p = 10$			$p = 50$		
		MSE	FDR	FNR	MSE	FDR	FNR
GLASSO	500	0.082	0.367	0.002	0.825	0.410	0.032
Network GLASSO, $A_{ind.}$	500	0.085	0.315	0.007	0.766	0.443	0.035
Network GLASSO, $A_{0.75}$	500	0.066	0.270	**0.000**	0.604	0.419	0.031
Network GLASSO, $A_{0.85}$	500	0.045	0.195	0.008	0.512	0.386	0.027
SS	500	**0.030**	0.000	0.023	0.198	0.002	0.009
Network SS, $A_{ind.}$	500	0.034	**0.000**	0.023	0.201	**0.001**	0.010
Network SS, $A_{0.75}$	500	0.032	0.002	**0.018**	0.193	0.001	0.009
Network SS, $A_{0.85}$	500	0.033	0.008	0.022	0.183	0.001	0.009
SIGGM, A_{ind}	500	0.104	0.658	0.000	0.968	0.775	**0.007**
SIGGM, $A_{0.75}$	500	0.068	0.478	0.001	0.606	0.712	0.008
SIGGM, $A_{0.85}$	500	0.047	0.375	0.021	0.524	0.683	0.008

A.5.2 The EBIC to learn the network hyperparameters

As a sensitivity check, we also consider using the EBIC (Chen and Chen, 2008) to select hyperparameters for the GLASSO and Network GLASSO models

$$\text{EBIC}(\lambda) = -2\ell_n(\hat{\Theta}(\lambda)) + |\mathbf{E}(\hat{\Theta}(\lambda))| \log n + 4|\mathbf{E}(\hat{\Theta}(\lambda))|\gamma_{\text{EBIC}} \log p \qquad (\text{A.3})$$

Compared with the BIC, (A.3) has an additional complexity penalty, controlled by hyperparameter γ. Foygel and Drton (2010) recommend $\gamma_{\text{EBIC}} \in [0, 0.5]$ where $\gamma_{\text{EBIC}} = 0$ recovers the BIC. Table A.3 presents the results of the experiments introduced in Section 4 when using the EBIC with $\gamma_{\text{EBIC}} = 0.5$ to select hyperparameters. Comparing these results with Table 1 shows that using the EBIC reduced the FDR relative to the BIC, however, this generally results in much more conservative edge selection which damaged the MSE.

Table A.3: GLASSO and network GLASSO simulation results under non, mildly and strongly informative networks A_{ind}, $A_{0.75}$ and $A_{0.85}$ with EBIC rule ($\gamma_{\text{EBIC}} = 0.5$) for learning the β hyperparameters.

		$p = 10$			$p = 50$		
	n	MSE	FDR	FNR	MSE	FDR	FNR
GLASSO	100	0.474	0.243	0.176	6.628	0.163	0.566
Network GLASSO, $A_{ind.}$	100	0.556	0.163	0.253	7.008	0.128	0.632
Network GLASSO, $A_{0.75}$	100	0.383	0.138	0.162	5.691	0.112	0.504
Network GLASSO, $A_{0.85}$	100	0.195	0.103	0.153	4.566	0.098	0.414
GLASSO	200	0.254	0.283	0.060	2.726	0.224	0.241
Network GLASSO, $A_{ind.}$	200	0.265	0.223	0.082	2.678	0.227	0.248
Network GLASSO, $A_{0.75}$	200	0.200	0.176	0.058	2.155	0.206	0.216
Network GLASSO, $A_{0.85}$	200	0.108	0.118	0.120	1.837	0.188	0.207
GLASSO	500	0.101	0.281	0.004	0.958	0.327	0.138
Network GLASSO, $A_{ind.}$	500	0.099	0.235	0.011	1.002	0.286	0.142
Network GLASSO, $A_{0.75}$	500	0.074	0.185	0.000	0.781	0.272	0.153
Network GLASSO, $A_{0.85}$	500	0.051	0.116	0.096	0.698	0.214	0.158

B COVID-19 data analysis

This section provides additional details for the analysis of the COVID-19 infection rate data.

B.1 Data sources

To undertake our analysis, we collected and combined the following datasets.

1. U.S. population data

U.S. population data for 2019 were sourced from https://www2.census.gov/programs-surveys/popest/tables/2010-2019/counties/totals/.

2. FIPS code data

To allow for a better match between different datasets, we also extracted the "FIPS code" that uniquely identifies counties within the U.S. from the U.S. Bureau of Labor Statistics https://www.bls.gov/cew/classifications/areas/sic-area-titles.htm.

3. COVID-19 infection data

Time series data of confirmed COVID-19 infections in each U.S. county was obtained from https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_U.S..csv. Figure B.1 plots of the weekly aggregated confirmed COVID-19 infections

4. COVID-19 vaccination data

State-level vaccination data was obtained from https://github.com/govex/COVID-19/tree/master/data_tables/vaccine_data/us_data/time_series.

5. Policy data

The Oxford COVID-19 Government Response Tracker https://github.com/CSSEGISandData/COVID-19_Unified-Dataset tracks individual policy measures across 20 indicators. They also calculate several indices to give an overall impression of government activity. We used their Containment and Health indices to summarise the policy variables.

6. Temperature data

43

Figure B.1: Weekly SMALLCAPS:COVID-19 Cases per county for the 99 biggest counties in the U.S.

We extracted the daily average near-surface air temperature from the 'Hydromet' folder of the above repository https://github.com/CSSEGISandData/COVID-19_Unified-Dataset/tree/master/Hydromet.

7. U.S. area data

Population densities were obtained by dividing the county population by the area of the region. Area data were obtained from the U.S. Census Bureau https://tigerweb.geo.census.gov/tigerwebmain/TIGERweb_main.html.

8. Geocloseness data

To measure the Geographical distance between two counties we use the Haversine distance (Sinnott, 1984) which assumes the earth is spherical. The latitude and longitude of each county were downloaded from the U.S. Census Bureau https://tigerweb.geo.census.gov/tigerwebmain/TIGERweb_main.html.

9. Facebook connectivity data

The Facebook Social Connectedness Index (SCI), obtained from https://data.humdata.org/dataset/

`social-connectedness-index`, uses an anonymised snapshot of all active Facebook users and their friendship networks to measure the intensity of connectedness between locations. Specifically, it measures the relative probability that two individuals across two locations are friends with each other on Facebook.

10. Flight connectivity data

Flight data between airports in the US was downloaded from `https://essd.copernicus.org/articles/13/357/2021/essd-13-357-2021.html`

11. Airport Information

Name, ICAO code, and Geographical location of US airports were extracted from `https://www.flightradar24.com/52.52,13.39/4`. This data allowed us to assign airports to the county(s) that they were part of.

12. county-to-MSA crosswalk

When counties are part of large urban areas known as metropolitan statistical area (MSA) we allocate the flights proportionally to all counties in the MSA. For example, a flight from JFK in New York to LAX in Los Angeles is not recorded between just those two counties, but rather allocated between all county pairs that form the NY and LA MSAs proportionally to the populations of these counties in the MSAs. The membership of counties to MSAs was downloaded from `https://www.census.gov/geographies/reference-files/time-series/demo/metro-micro/delineation-files.html`.

13. Flight capacities

The capacity of certain plane models was extracted from the folowing links

- `https://www.seatguru.com/airlines/American_Airlines/fleetinfo.php`

- `https://www.seatguru.com/airlines/Delta_Airlines/fleetinfo.php`

- `https://www.seatguru.com/airlines/JetBlue_Airways/fleetinfo.php`

- `https://www.seatguru.com/airlines/Southwest_Airlines/fleetinfo.php`

- `https://www.seatguru.com/airlines/Spirit_Airlines/fleetinfo.php`,

allowing for the estimation of the number of passengers on each flight.

Producing our flight connectivity network required the following steps

- Assign each airport to the county in which it is located and distribute flights between counties that make up MSA's to the other counties in the MSA proportionally to their population

- Use airline capacity data to estimate the number of passengers on each flight and therefore the number of passengers flowing between two counties

- Standardise this by the population of each county to estimate population flow

B.2   Data processing

Once the data was collected, some minimal data preprocessing was required to prepare the data for our analysis. This consisted mainly of variable transformation and imputing of missing values.

B.2.1   Variables transformation

Natural logarithms were taken of the variables '*confirmed case*', '*population density*' and '*number of vaccinations*'.

B.2.2   Missing values

In addition, there were missing values in covariates Containment and Health Index data (CHI) as well as the Temperature (Temp) data and the Vaccination data. We imputed these missing values as follows

1. The CHI values were calculated as a function of different policy measures (https://github.com/OxCGRT/covid-policy-tracker/blob/master/documentation/index_methodology.md). On several occasions either these policy measures or their flags were missing. We imputed these as follows

   - Missing flags were imputed as 0's, i.e. no flags
   - Missing values before the first recorded value were imputed as 0, i.e. assuming no measures were in place before the first recorded measure

- Missing values in between two recorded values were imputed as an average of the before and after measures

- Missing values after the last recorded value were imputed as the last seen measure, i.e. assuming a continuation

2. For the Temp data, the temperatures for 18 counties were not recorded at all. We imputed these using the nearest county geographically whose temperature data was available.

3. The vaccination data was only recorded from the 14th of December 2020 and therefore all vaccination counts before this date were imputed as 0's.

B.3 Meta-County Clustering

Before Clustering the data we removed some counties whose data were not available. From the FIPS data we downloaded there were 3144 counties. District of Columbia did not have COVID-19 policy variable available and we could not compute the population density for Valdez-Cordova Census Area in Alaska so we removed these. Five counties were removed as they were not available in the SCI index and 8 counties in Connecticut were removed because they did not have any flight connection data. This left 3129 counties.

Starting with 3129 counties, we hierarchically clustered small counties together such that the resulting meta-counties all have population greater than 500,000. The clustering procedure is described in the following steps and is implemented by our code.

Step 1: Remove the 'big' counties. Any county whose population was greater than 500,000 is extracted and left unchanged. There were 136 'big' counties leaving 2993 'small' counties.

Step 2: Cluster small counties with each state.

i) Within each state find the smallest 'small' county and combine this with the 'small' county within that state whose centroid is closest to create a 'meta-county'

ii) Update the county centroid as the average of the latitude and longitude of the two combined counties

This procedure is repeated with each state until either all of the 'meta-counties' have population greater than 500,000, or there is only one meta-county left for that state. This resulted in 196 meta counties

This clustering procedure resulted in $p = 332$ counties. Once the meta-counties have been created the number of cases were summed, populations are combined, areas combined, temperatures averaged and the vaccination and CHI variables inputted for that state. This allowed the model described below to be estimated together on the large counties and the meta-counties made up of smaller counties.

B.4   Model description

Our final response variable is the log of the weekly COVID-19 infections per 10,000 members of the population (i.e. cases / population $\times$ 10,000). This results in data $y_1, \ldots, y_n$ where $y_i = (y_{i1}, ..., y_{ip})$ is the log of the standardised weekly COVID-19 infections at week i in the $p = 332$ counties and meta-counties. The sample interval is from 22 January 2020 to 30 November 2021 resulting in $n = 97$ weeks of data.

Our graphical model posits $y_i \sim \mathcal{N}_p(\mu_i, \Theta^{-1})$ where $\mu_i = (\mu_{i1}, ..., \mu_{ip})$. For convenience, we decouple the estimate of μ_i from Θ. We pose a regression model for μ_{ij} and then estimate Θ using the residuals of this model assuming zero mean as in Section 2. Our generalised additive regression model for y_{ij} can be summarised as follows

$$
\begin{aligned}
log(confirmed)_{ij} = b_0 &+ b_1 \times log(Lag_{confirmed})_{ij} + b_2 \times log(popdensi)_j \\
&+ b_3 \times Cum\_vaccinated_{i,state_j} + b_4 \times CHI_{i,state_j} \\
&+ s(Temp)_{ij} + \gamma_2 \times Time_2 + ... + \gamma_T \times Time_T + \epsilon_{ij}
\end{aligned}
$$

where

(1) $log(confirmed)_{ij}$ represents the natural logarithm of weekly per 10,000 people confirmed case in county j at time i.

(2) $log(Lag_{confirmed})_{ij}$ a first-order auto-regressive term measuring the infection rate at the previous time point $i - 1$ for each county j

(3) $log(popdensi)_j$ is the population density for county j

(4) $Cum\_vaccinated_{i,state_j}$ is the cumulative number of vaccinated individuals in the state to which county j belongs by time i

(5) $CHI_{i,state_j}$ represented the Containment and Health Index summarising COVID-19 policies/measures put in place in the state to which county j belongs and time i (wearing masks, closing schools, etc.)

(6) $s(Temp)_{ij}$ is a non-parametric smooth of the average temperature for county j at time i implemented in `mgcv` package in R

(7) $Time_i$ is an indicator for week i and provides a weekly fixed effect term estimating the mean infections across all counties at time i

(8) ϵ_{ij} are the residuals of county j at time i

With such a model we aim to remove the effect of the most relevant covariates that drive the mean number of infections, allowing Θ^{-1} to capture dependencies unexplained by these covariates.

B.5 Checking model goodness-of-fit

The main assumptions behind our assumed model require that the residuals ϵ_{ij} are Gaussian distributed and independent across $i = 1, \ldots, n$ time points. We provide diagnostic plots to check these assumptions.

Figure B.2 plots the fitted values $\hat{y}_{ij}$ and each of the predictors against the residuals ϵ_{ij}. This demonstrates that the assumption that the covariates are linearly related to the response is satisfactory and that the residuals appear reasonably homoskedastic. Figure B.3 shows a histogram of the standardised residuals and Q-Q-normal plots for ϵ_{ij}. The Gaussian assumption is tenable here.

The raw COVID-19 data exhibited strong serial correlation. To address this issue we added a first-order auto-regressive term. Figure B.4 plots the autocorrelation functions and partial autocorrelation functions for further lags after incorporating the AR1 term. These indicate that higher-order terms are unnecessary. After adding an AR1 term the interpretation of the errors (and their covariance) changes: they measure the infection rate relative to the covariates and to the infection rate of the previous week, i.e. they capture whether certain counties are growing faster/slower than expected (relative to the next week). So the model is investigating the growth rates, rather than absolute infection numbers.

B.6 The network predictors

A further assumption of our proposed network GLASSO models, as discussed in Section 2.1, is that there is a linear relation between $\log \mathbb{E}[\rho_{jk}^2 | A]$ and the network entries $a_{jk}^{(q)}$. To achieve linearity we defined our two network predictors as

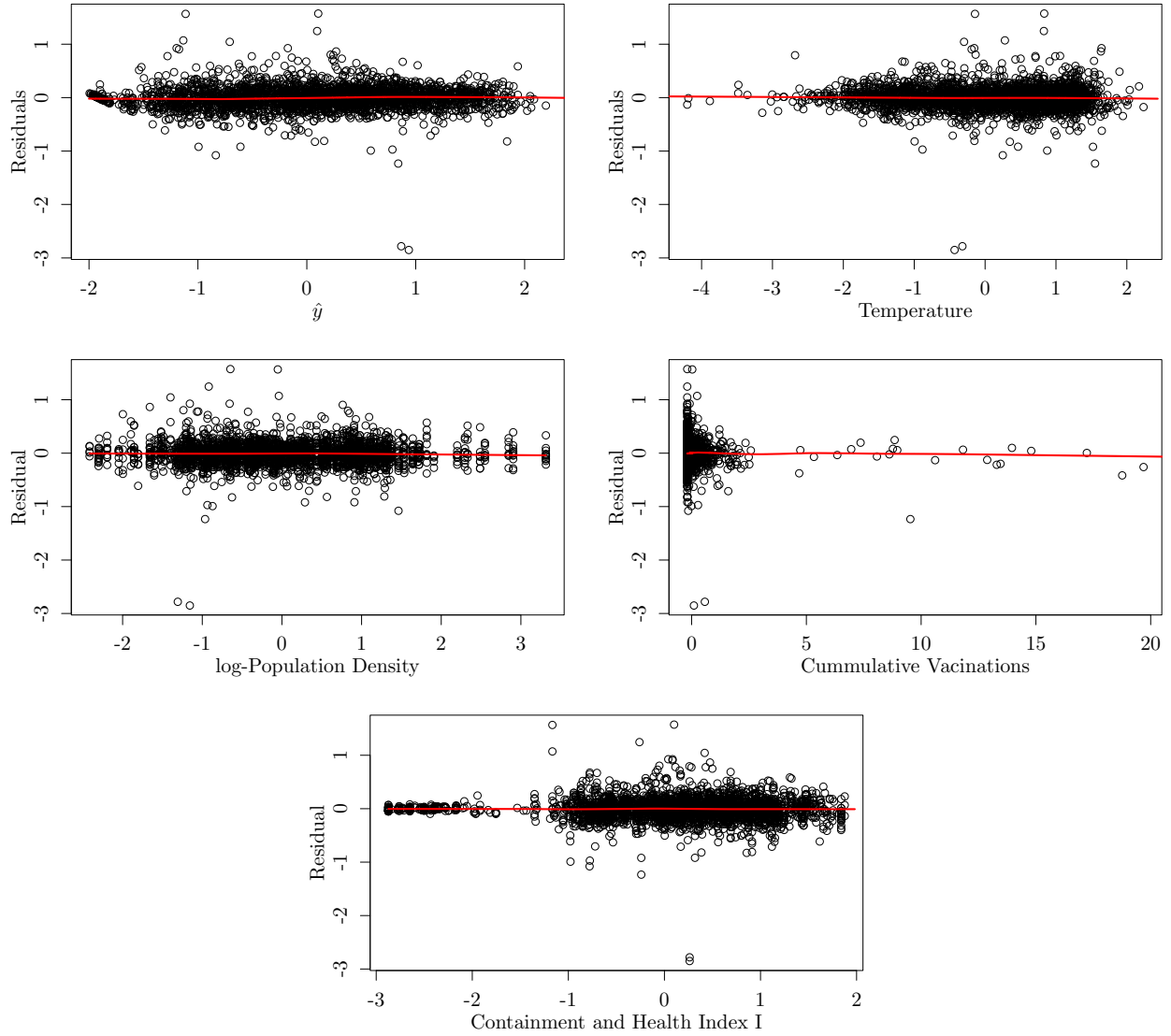$$A_1 := 1/\log(Geodist), \quad A_2 := \log(Facebook), \quad A_3 := \log(1 + Flights).$$

Figure B.2: Plots of the fitted values and each covariate against the residuals for the COVID-19 data. The **red** line corresponds to the LOWESS smooth.
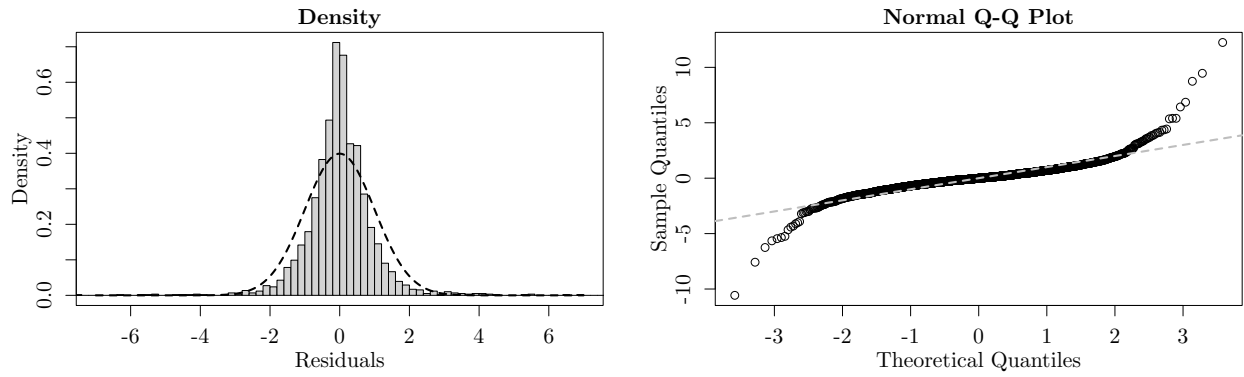
Figure B.3: COVID-19 data. **Left** Histogram of the standardised residuals compared with the standard Gaussian density. **Right** Q-Q Normal plot of the standardised residuals.
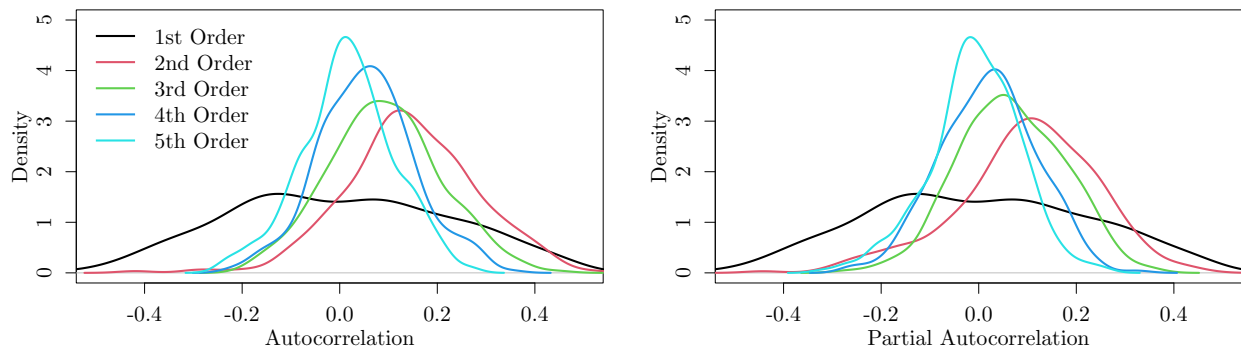

Figure B.4: Residual autocorrelation functions and partial autocorrelation functions after incorporating the AR1 term for the COVID-19 data.

Figure B.5: Assessing the linear relation between $\log \mathbb{E}[\hat{\rho}_{jk}^2 | A]$ and the network matrices, where $\hat{\rho}_{jk}$ is the GLASSO estimate. The points represent the log-mean values of $\hat{\rho}_{jk}^2$ within 10 equispaced bins defined for each network.

Figure B.5 illustrates that after such transformations, the assumption of linearity is reasonably satisfied.

B.7   Supplementary figures

The top of Figure B.6 the flight connectivity network against partial correlations estimated by GLASSO. It appears that as the flight connectivity goes up, the variance in the partial correlations decreases slightly. However the dependence between the network and the partial correlations is much smaller than was observed between the geographical or Facebook networks and the partial correlations in Figure 1. The fitted spike-and-slab distributions in the bottom of Figure B.6 further demonstrate this.

Table B.1 summarises the estimated graphical model under the network spike-and-slab model using a posterior slab probability threshold of > 0.5 and > 0.95. The number of edges estimated under both
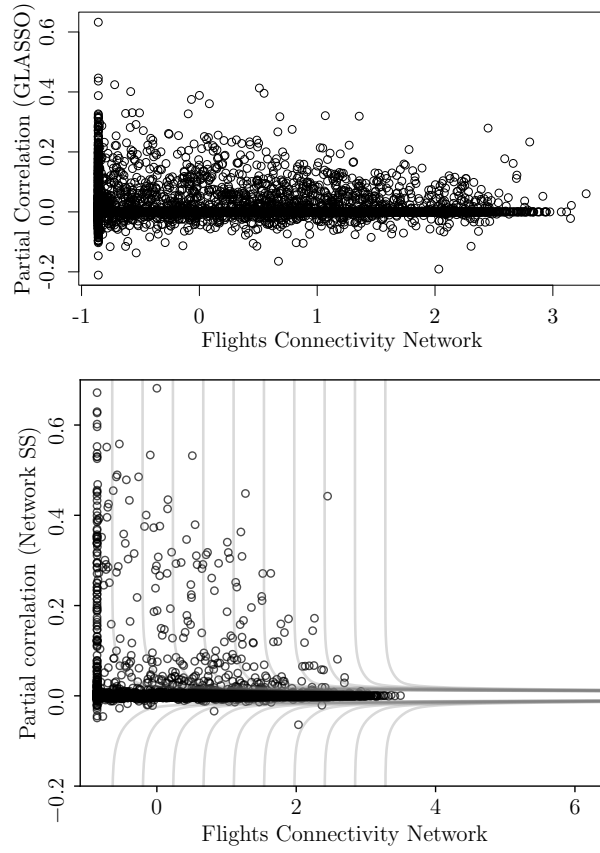
Figure B.6: Residual partial correlations in COVID-19 infections (adjusted for covariates) across counties vs flight connectivity network defined as $\log(1+Flight)$. Top panel: partial correlations estimated with graphical LASSO, with penalization parameter set via BIC. Bottom panel: fitted spike-and-slab distributions and fitted partial correlations estimated with network graphical spike-and-slab LASSO.

the 0.5 and 0.95 slab probability threshold is considerably smaller than the number of edges estimated under the network GLASSO models. Under the 0.95 slab probability threshold, the estimated number of edges is more conservative.

Figure B.7 shows how the estimated network hyperparameters of Table 3 affect the location of the slab and the probability of being in the slab marginally for each network when fixing the other two networks to their means. We see that while as both the geographical closeness and Facebook networks increase the location of the slab and the probability of being in the slab increases, the Facebook network has the larger effect.

53

Table B.1: COVID-19 data: Edge counts of the network spike-and-slab model when declaring an edge for posterior slab probability > 0.5 and > 0.95

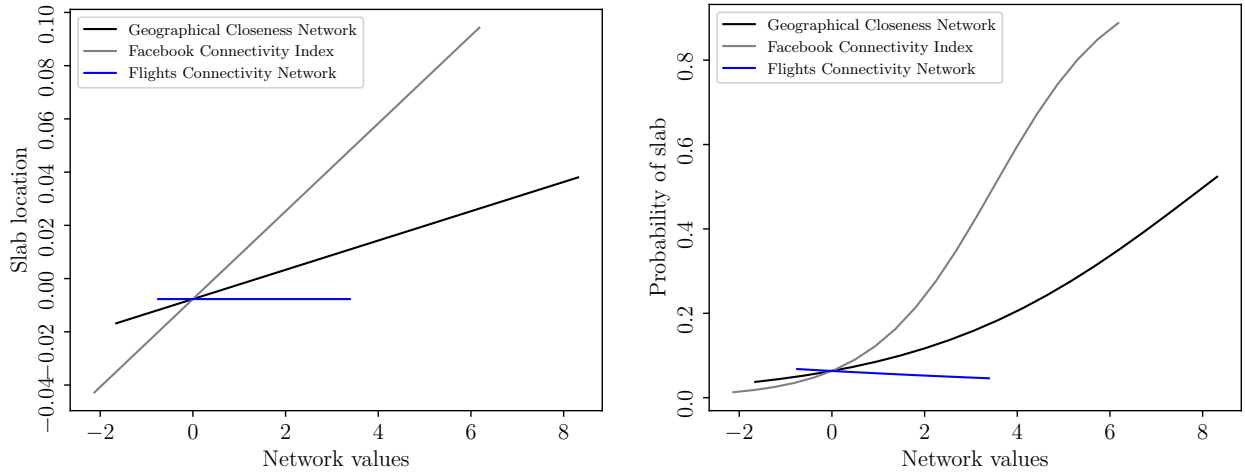|	Edges (> 0.5)	Non-Edges (> 0.5)	Edges (> 0.95)	Non-Edges (> 0.95)
Network SS	249	54697	102	54844

Figure B.7: COVID-19 data: Slab location (**left**) and slab probability (**right**) as a function of the three networks estimated by empirical Bayes.
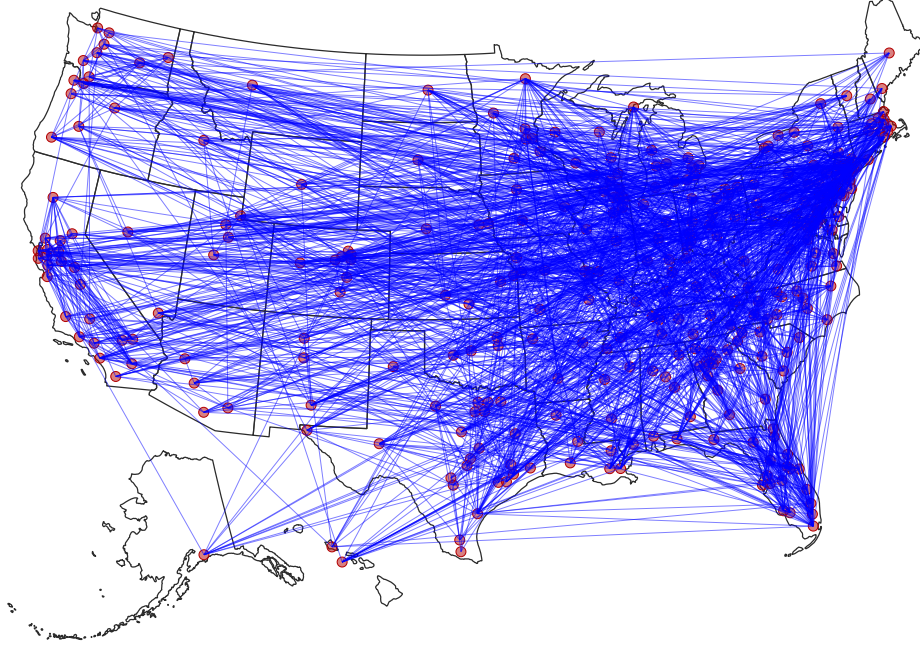
B.7.1 U.S. map plots

Figure B.8 visualises the network given by non-zero elements of the GLASSO estimated Θ with no network information (top) and the network GLASSO estimate of Θ obtained when using both A_1 and A_2 (bottom), the model achieving the smallest BIC, on top of a U.S. map. The network GLASSO estimates a much sparser network, but we see there are still edges present between counties that are geographically close as well as those that are farther away.
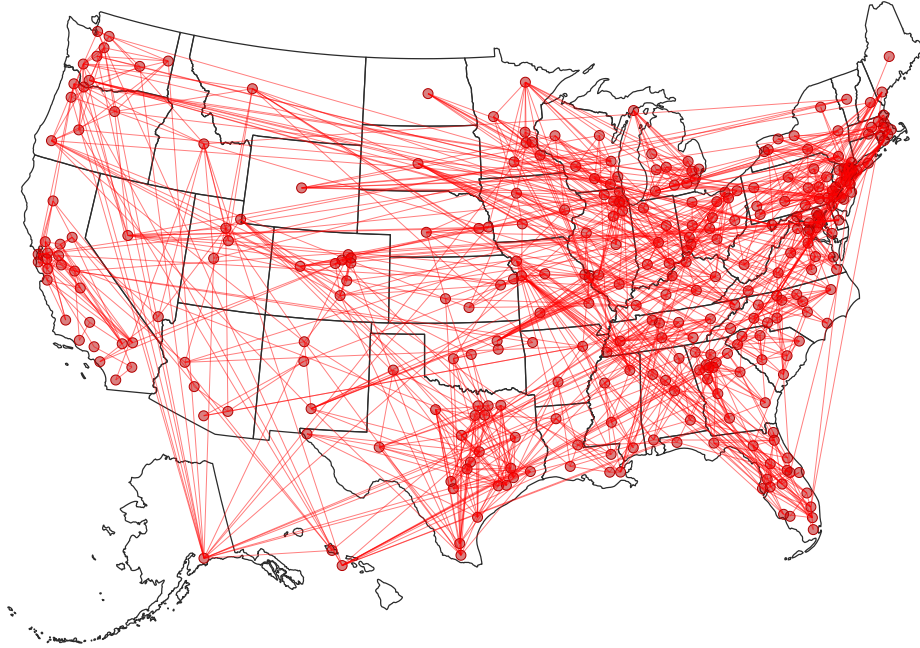
B.8 Results using the EBIC

Similarly to the simulations, we also investigate the sensitivity of our COVID-19 data results by considering selecting hyperparameters using the EBIC with $\gamma_{\text{EBIC}} = 0.5$. Table B.2 presents these results. From the number of edges, we can see that using the EBIC estimates sparser networks than under the BIC, but the out-of-sample test set estimate suggests these estimates may be too sparse. Importantly, we see that the improvement of the network GLASSO methods over standard GLASSO is still apparent when using the EBIC selection criteria.

Table B.2: Eight models for the COVID-19 data when using the EBIC ($\gamma_{\text{EBIC}} = 0.5$) to learn the network hyperparameters. A_1, A_2 and A_3: networks defined by $1/\log(Geodist)$, $\log(Facebook)$ and $\log(1 + Flight)$. EBIC values account for the extra hyper-parameters in the network GLASSO models. 10-fold: 10-fold cross-validated log-likelihood

Method	EBIC	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	Edges	10-fold
GLASSO	32204.000	-0.062				0	24.317
Network GLASSO- A_1	27150.357	1.622	-1.120			766	81.846
Network GLASSO- A_2	24461.011	2.903		-1.147		617	113.533
Network GLASSO- A_3	32220.185	0.959			-0.165	0	24.317
Network GLASSO- A_1 & A_2	24303.227	5.200	-1.038	-1.162		730	112.771
Network GLASSO- A_1 & A_3	26453.480	2.407	-1.414		1.005	766	90.710
Network GLASSO- A_2 & A_3	**23931.443**	4.063		-1.457	-0.255	589	**114.372**
Network GLASSO- A_1, A_2 & A_3	24927.80	2.845	-0.274	-1.159	0.246	796	113.764

(a) Edges identified in GLASSO with no network

(b) Edges identified in network GLASSO with networks A_1 and A_2, the model achieveing the smallest BIC

Figure B.8: Edges identified by GLASSO and network GLASSO with the geographical closeness and Facebook networks.

C   Stock market data preparation

This section provides additional details for the analysis of the stock market excess returns data.

C.1   Data sources

To undertake our analysis, we collected and combined the following datasets.

1. Stock price data

We extracted the daily closing stock price for $p = 366$ firms satisfying the following criteria: closing stock prices adjusted for stock splits and dividends were available in the COMPUSTAT database for every trading day between 2 January 2019 to 31 December 2019 (leaving $n = 252$ time points), the stocks were associated to a member of the S&P500 at the end of 2019, and we could retrieve their 10-K filings for at least one of the years 2014-2019. The data was downloaded from the Center for Research in Security Prices (CRSP) database accessed via Wharton Research Data Services (WRDS).

2. S&P 500 firms

The list of S&P 500 firms was downloaded from https://web.archive.org/web/20190912150512/ https://en.wikipedia.org/wiki/List_of_S%26P_500_companies which corresponds to the wikipedia page listing the S&P500 retrieved on 12/09/2019, its last archived data in 2019.

3. Fama/French Three-Factor Model

We constructed excess returns using the Fama-French three-factor model (Fama and French, 1993). The three factors are the 1) overall market return, 2) a measure of firm size, and 3) a measure of book-to-market ratio. The daily Fama/French factors were downloaded from https://mba.tuck. dartmouth.edu/pages/faculty/ken.french/data_library.html. For each stock, we regress 2019 daily returns on the three factors (plus a constant) and extract the residual as the excess return.

4. Risk measures

Our network data measures the similarity of two companies' risk exposures stratified into Economic and Policy risks. The 10-K risk exposure data counts for each risk category the number of sentences within a company's 10-K filings that contained any member of a dictionary associated with that risk category. We manually construct these using the dictionary terms listed in Baker et al. (2019). From this data, we can construct a $p \times p$ network matrix for firms, where each entry X_{ij} represents the

degree of "closeness" between firm X_i and firm X_j.

C.2 Model description

Our final response variable is the log daily returns for $p = 366$ U.S. firms throughout 2019, resulting in $n = 251$ observations. We are, however, interested in the graphical model, $\mathcal{N}_p(0, \Theta^{-1})$, of the 'excess returns', defined as the residuals of a linear model regressing the log-returns on the Fama-French factors.

The 'excess returns' for stock j are therefore estimated, separately for each firm, using the following model

$$r_{ij} - Rf_i = b_{0j} + b_{1j} \times SMB_i + b_{2j} \times HML_i + b_{3j} \times (Rm - Rf)_i + \epsilon_{ij} \tag{C.1}$$

where

(1) r_{ij} is the log daily return of stock j at time i defined as $r_{ij} = \log p_{ij} - \log p_{i-1j}$, where p_{ij} is the closing price of firm j on day i

(2) Rf_i is the risk free rate at time i.

(3) SMB_i (Small Minus Big) is the average return on the three small portfolios minus the average return on the three big portfolios at time i.

(4) HML_i (High Minus Low) is the average return on the two value portfolios minus the average return on the two growth portfolios at time i.

(5) $(Rm - Rf)_i$, the excess return on the market at time i, value-weighted return of all CRSP firms incorporated in the U.S. and listed on the NYSE, AMEX, or NASDAQ that have a CRSP share code of 10 or 11 at the beginning of i's month, good shares and price data at the beginning of i's month, and good return data for i minus the one-month Treasury bill rate.

(6) Coefficients b_{0j}, b_{1j}, b_{2j} and b_{3j} are estimated for firm j using ordinary least squares.

C.3 Checking model goodness-of-fit

Similarly to Section B.5, we produce diagnostic plots to confirm the validity of the linear-model and the Gaussianity and independence of its residuals.
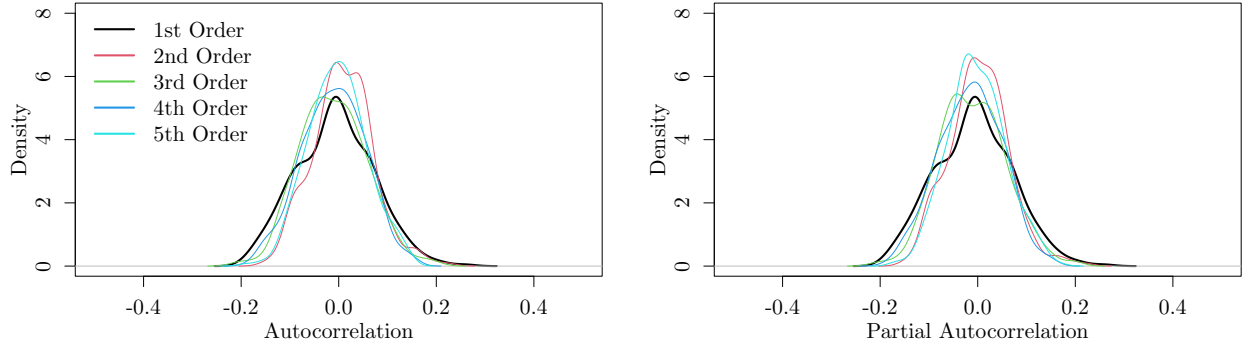
Figure C.1: Residual autocorrleation function and partial autocorrelation function for the stock market data.

Figure C.1 plots autocorrelation functions and partial autocorrelation functions, demonstrating that the observations can be considered independent and that there is no need to consider autoregressive terms. Figure C.2 plots the fitted values $\hat{y}_{ij}$ and each of the predictors against the residuals ϵ_{ij}, demonstrating that the assumption that the covariates are linearly related to the response is satisfactory and that the residuals appear reasonably homoskedastic.

While the Gaussian assumption was tenable for the COVID-19 data, Figure C.3 shows that this is not the case for the stock market data. There is evidence of considerably heavier tails than Gaussianity. To address this issue we fit a non-paranormal model based on transforming the data into $f(\epsilon_i) := (f_1(\epsilon_{i1}), \ldots, f_p(\epsilon_{ip}))$, where $\hat{f}$ was estimated using the R package `huge` (Zhao et al., 2012). Figure C.4 shows a histogram and Q-Q-normal plot for $f(\epsilon_i)$, where the Gaussian assumption is more tenable.

C.4   The network predictors

The price data from CRSP is arranged by TIC, a unique stock identifier, while the risk measures are arranged by CIK, a unique company identifier. Any two stocks associated with the same compnay had the same risk scores.

Based on the construction of Baker et al. (2019), we divided the 37 risk factors into two categories: the economic risks (containing 17 risks) and the policy risks (containing 20 risks) and standardised the sentence coutsn by the total number of sentences in the 10-K fillings. Then, for each risk type, we centered the $\log(1 + counts)$ and evaluated the Pearson's correlation between all pairs of companies to obtain two network matrices E_{pears} and P_{pears}.

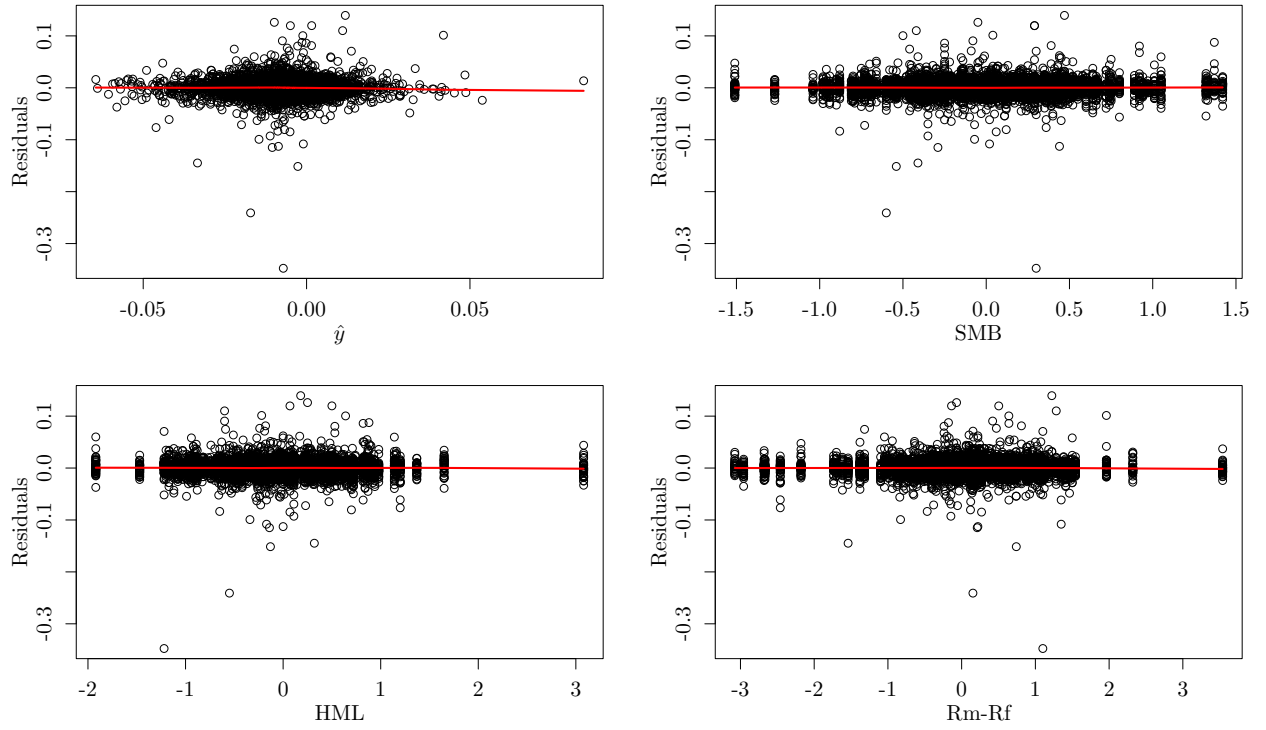Figure C.5 demonstrates that for both networks there appears to be an increased chance of having

Figure C.2: Plots of the fitted values and each covariate against the residuals for the stock market data. The red line corresponds to the LOWESS smooth.
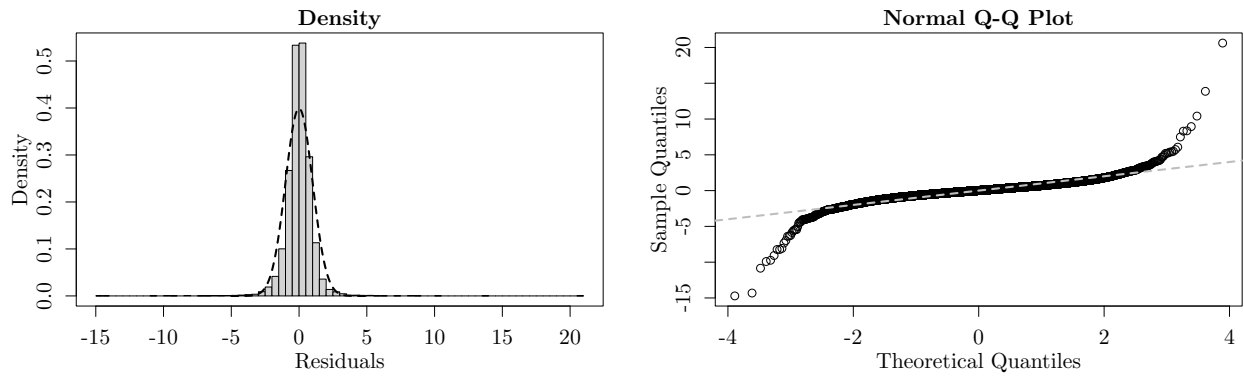

Figure C.3: Stock market data. **Left** Histogram of the standardised residuals compared with the standard Gaussian density. **Right** Q-Q Normal plot of the standardised residuals.
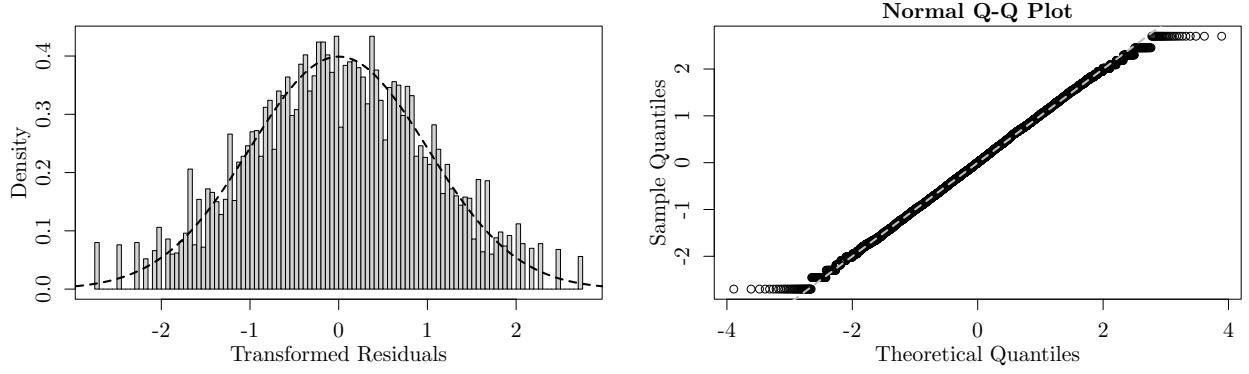
Figure C.4: Stock market data. **Left** Histogram of the transformed residuals compared with the standard Gaussian density. **Right** Q-Q Normal plot of the transformed residuals.
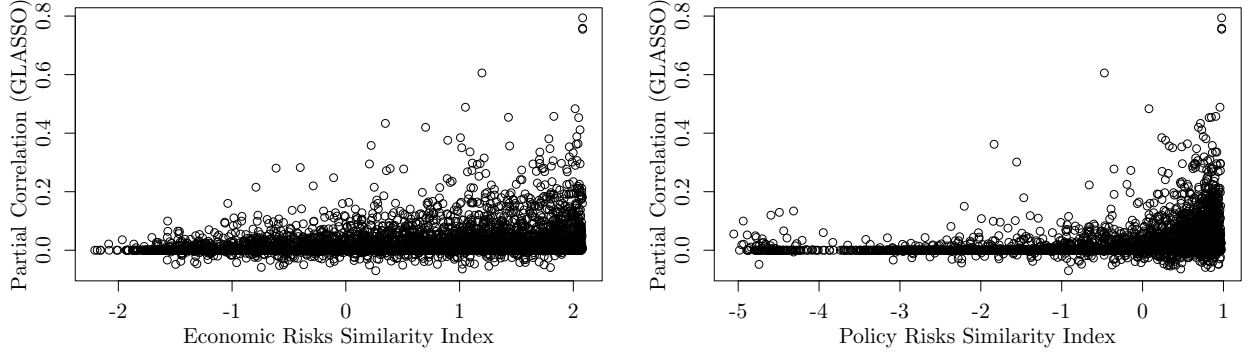

Figure C.5: Residual partial correlations of the stock market excess returns across firms vs Economy risk (left) and Policy risk (right). Partial correlations were estimated with GLASSO, with penalization parameter set via BIC.

positive partial correlation if the two firms have highly correlated risk factors. Figure C.6 demonstrates that no further transformation of the networks is required to satisfy the network GLASSO assumption of linearity.

C.5 Supplementary figures

Table C.1 summarises the estimated graphical model under the network spike-and-slab model using a posterior slab probability threshold of > 0.5 and > 0.95. The number of edges estimated under both slab probability threshold is smaller than the number of edges estimated under the network GLASSO models. Under the 0.95 slab probability threshold, the estimated number of edges is more
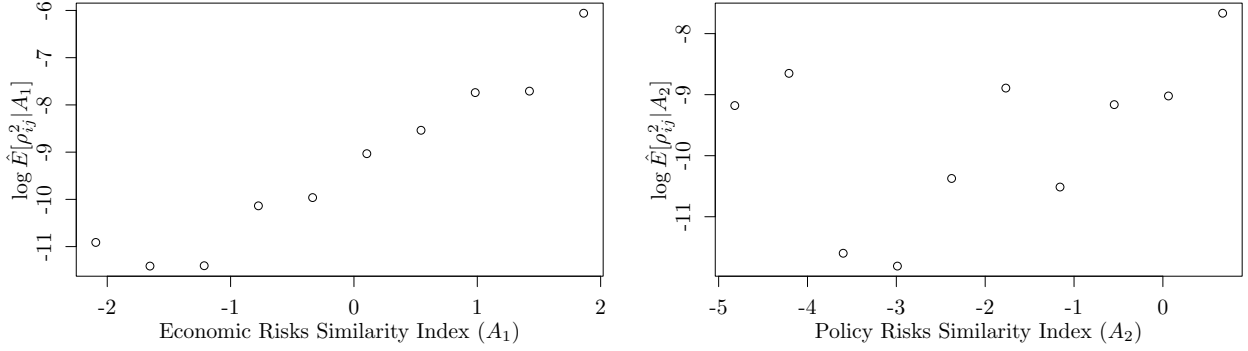
Figure C.6: Assessing the linear relation between $\log \mathbb{E}[\hat{\rho}_{jk}^2 | A]$ and the network matrices, where $\hat{\rho}_{jk}$ is the GLASSO estimate. The points represent the log-mean values of $\hat{\rho}_{jk}^2$ within 10 equispaced bins defined for each network.

Table C.1: Stock market data: Edge counts of the network spike-and-slab model when declaring an edge for posterior slab probability > 0.5 and > 0.95

	Edges (> 0.5)	Non-Edges (> 0.5)	Edges (> 0.95)	Non-Edges (> 0.95)
Network SS	377	66418	189	66606

conservative.

C.6   Results using the EBIC

Table B.2 presents results investigating the stability of our stock market data analysis to selecting hyperparameters using the EBIC with $\gamma_{\text{EBIC}} = 0.5$ rather than the BIC. The EBIC continues to estimate sparser networks than the BIC, but to the detriment of the out-of-sample test set score. Importantly, we see that the improvement of the network GLASSO methods over standard GLASSO is still apparent when using the EBIC selection criteria.

D   Stan vs NumPyro

We estimated our network spike-and-slab models using the No-U-Turn Sampler (NUTS) (Hoffman and Gelman, 2014), an extension of Hamiltonian Monte Carlo (HMC, Duane et al. 1987) that automates the setting of the step-size in the Hamiltonian discretisation. Two probabilistic programming imple-

Table C.2: Four models for the stock market data when using the EBIC ($\gamma_{\text{EBIC}} = 0.5$) to learn the network hyperparameters. A_1 is the Economic network, A_2 the Policy network. EBIC values account for the extra hyper-parameters in the network GLASSO models. 10-fold is the 10-fold cross-validation log-likelihood.

Method	EBIC	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	Edges	10-fold
GLASSO	88588.75	-0.9106			616	-494.781
Network GLASSO- A_1	86140.75	3.350	-2.604		572	-493.508
Network GLASSO- A_2	87675.87	0.531		-2.081	732	-494.090
Network GLASSO- A_1 & A_2	**84150**	10.289	-4.020	-5.718	468	**-492.872**

mentations of NUTS are `Stan` (Carpenter et al., 2017) and `NumPyro` (Bingham et al., 2019; Phan et al., 2019). We provide implementations of our algorithm in both languages, but for our experiments, we found `NumPyro`'s ability to take advantage of parallel computing for automatic differentiation provided a considerable speed up.

We illustrate this using one of our simulated examples from Section 4. We consider network matrix $A_{0.85}$, $n = 100$ and $p = 10$ and $p = 50$. We ran both `Stan` and `NumPyro` for 2000 warm-up iterations and 2000 sampling iterations. Table D.1 compares the time taken to sample and the effective sample size (ESS) of the resulting sample averaged across 10 repeat datasets. We present the ESS as separately averaged across the ρ model parameters and the network hyperparameter η. We see that both methods produce similar ESS but that `NumPyro` does so over six times faster.

We also take this opportunity to demonstrate how efficient the network GLASSO is when implemented as a special case of the GOLAZO algorithm (Lauritzen and Zwiernik, 2020). For the same datasets considered above, we implement the network GLASSO using 50×50 grid search to estimate the network hyperparameters. We see that the GOLAZO algorithm takes a fraction of the time to run as the Bayesian implementation even when using a rudimentary grid-search optimisation scheme.

Lastly, above we limited `NumPyro`'s access to only 6 cores on one machine. Using more cores, for example on a GPU, provides the potential for `NumPyro` to achieve even greater speed-ups for higher dimensional problems beyond the simple one considered here.

Table D.1: Comparison of time taken for network GLASSO implemented using the GOLAZO algorithm and the network spike-and-slab sampling algorithms in `Stan` and `NumPyro`.

$p = 10$	Time (s)	ESS ρ's	ESS η's	$p = 50$	Time (s)	ESS ρ's	ESS η's
GOLAZO	14.94	-	-	GOLAZO	312.78	-	-
Stan	184.93	855	373	Stan	7162.11	1663	268
NumPyro	28.99	977	522	NumPyro	1133.057	1604	293