The grass is not always greener in the neighbor's yard

Dittrich, D.

Publication date:
2018

Document Version
Publisher's PDF, also known as Version of record

Link to publication in Tilburg University Research Portal

Citation for published version (APA):
Dittrich, D. (2018). *The grass is not always greener in the neighbor's yard: Bayesian and frequentist inference methods for network autocorrelated data*. Proefschriftmaken.

The grass is not always greener in the neighbor's yard:

Bayesian and frequentist inference methods for network autocorrelated data

Dino
Dittrich

The grass is not always greener in the neighbor's yard: Bayesian and frequentist inference methods for network autocorrelated data

DINO DITTRICH

Tilburg University
Tilburg School of Social and Behavioral Sciences
Department Methodology and Statistics

ISBN: 978-94-6380-126-3
Printed by: ProefschriftMaken || `proefschriftmaken.nl`
Cover design by: Remco Wetzels || `remcowetzels.nl`

Printing was financially supported by Tilburg University.

The grass is not always greener in the neighbor's yard: Bayesian and frequentist inference methods for network autocorrelated data

Proefschrift ter verkrijging van de graad van doctor aan Tilburg University op gezag van de rector magnificus, prof. dr. E.H.L. Aarts, in het openbaar te verdedigen ten overstaan van een door het college voor promoties aangewezen commissie in de Aula van de Universiteit op vrijdag 30 november 2018 om 14.00 uur

door

Dino Dittrich,

geboren te München, Duitsland.

Promotor:	prof. dr. R.Th.A.J. Leenders
Promotor:	prof. dr. J. K. Vermunt
Copromotor:	dr. ir. J. Mulder
Overige leden van de promotiecommissie:	prof. dr. M.A.L.M. van Assen
	prof. dr. J.P. Elhorst
	prof. dr. H.A.J. Hoijtink
	dr. P. Leifeld

In memory of *Majka*.

Contents

Chapter 1

Introduction

On a quiet morning in April 2018, a few weeks before completing this thesis, I was traveling on country roads through north-eastern Bosnia. The sun was shining brightly, traffic was low, and I had time to observe the family homes, the restaurants, and the commercial areas that string together along the road between the towns of Orašje and Tuzla. Some of these looked abandoned, some were run-down, others appeared neat and tidy. Interestingly, or expectedly, the run-down as well as the neat properties seemed to cluster, and there were hardly any two of a different kind to be found next to each other. Certainly, knowing a household's income or a restaurant's revenue would have been a good indicator for the state of a property; at the same time, manpower and time are no scarce resources in Bosnia.[1] Hence, equally certainly, low financial means could not have explained alone why some properties were run-down, while others seemed well-maintained. Instead, in actively or passively determining their degree of engagement, households and estate owners are likely to also take into account their neighbors' behavior. This theoretical reasoning leads to the notion of *network autocorrelation* (Leenders, 1995; White et al., 1981).

Network autocorrelation refers to the correlation of observations for a variable of interest among related actors in a network. In general, a network is characterized by a set of actors together with a set of ties, where a tie indicates that two actors are related to each other. For example, we could define a network as the set of households in a village with ties between two households based on property adjacency. Likewise, we could also define ties between households based on social similarity rather than property adjacency.

In this thesis, we develop statistical methods for quantifying and testing the strength of the network autocorrelation of a variable of interest, as induced by a given network, while controlling for a set of explanatory variables, or covariates. This includes methods for continuous as well as count variables of interest and extends to settings in which multiple distinct networks give rise to multiple network autocorrelations.

[1] The unemployment rate in Bosnia and Herzegovina is one of the highest in Europe and reported to be 25.4% in 2016 by the International Labour Organization (ILO) at `http://ilo.org/gateway/faces/home/ctryHome?locale=EN&countryCode=BIH&_adf.ctrl-state=2oss1sbtj_9`.

1.1 The network autocorrelation model

Throughout the larger part of this thesis, we model network autocorrelation using the *network autocorrelation model* (Ord, 1975). Ever since its introduction, the network autocorrelation model has been indispensable for modeling network influence on individual behavior and has been applied in many different fields, such as criminology (Tita & Radil, 2011), ecology (McPherson & Nieswiadomy, 2005), economics (Dall'erba et al., 2009), geography (Mur et al., 2008), political science (Kowal, 2018), psychology (Barnett et al., 2014), public health (Myneni et al., 2015), and sociology (Mizruchi & Stearns, 2006).

In the network autocorrelation model, actor observations, or responses, for a variable of interest are allowed to be correlated and a *network autocorrelation parameter* ρ is estimated, quantifying the network influence on the variable of interest. Hence, an actor's response is assumed to be a function not only of a set of explanatory variables but also of the responses for the actor's neighbors, i.e., other actors in the network this actor is tied to. More precisely, the network autocorrelation model expands a standard linear regression model by including an additional term that contains a weighted sum of the actors' neighbors' responses. The corresponding weights are given in a pre-defined *connectivity matrix* whose entries stand for the extent to which two actors influence each other based on a particular influence mechanism, e.g., geographic adjacency. Then, the dependence of an actor's response on its neighbors' responses is modeled using a variance-covariance matrix that is a function of the chosen connectivity matrix.

Let $\boldsymbol{y} \in \mathbb{R}^g$ be the vector of responses for g actors in a network, and let $X \in \mathbb{R}^{g \times k}$ denote a matrix comprising values for the g actors on k covariates (possibly including a column of ones for an intercept term). Furthermore, let $W \in \mathbb{R}^{g \times g}$ be a connectivity matrix with zero diagonal, where the elements W_{ij} represent the influence of actor j on actor i; the larger W_{ij}, the larger this influence. Given the observed quantities $\boldsymbol{y}, X$, and W, the standard, or *first-order*, network autocorrelation model takes the form

$$\boldsymbol{y} = \rho W \boldsymbol{y} + X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1.1}$$

where ρ is a scalar and called the network autocorrelation parameter, $\boldsymbol{\beta} \in \mathbb{R}^k$ is a vector of regression coefficients, and $\boldsymbol{\varepsilon} \in \mathbb{R}^g$ is a vector of independent and normally distributed error terms with zero mean and variance of σ^2. The network autocorrelation parameter ρ is the model's key parameter and quantifies the effect of network ties on a variable of interest. When $\rho = 0$, the network autocorrelation model in (1.1) reduces to a standard linear regression model with only the vector of regression coefficients $\boldsymbol{\beta}$ and the error variance σ^2 to be estimated.

1.2 Inferential limitations in the network autocorrelation model

While the network autocorrelation model has yielded valuable insights into the structure of influence processes in numerous networks from a variety of fields, there was a lack of adequate statistical tools for applying the model in situations often encountered in empirical practice when starting to work on this thesis four years ago.

First, the commonly used maximum likelihood estimator for the network autocorrelation parameter ρ has been shown to be biased for high levels of network density and small network sizes (Mizruchi & Neuman, 2008; Neuman & Mizruchi, 2010; Smith, 2009). Moreover, in such scenarios the maximum likelihood estimator's sampling distribution has also been found to be strongly asymmetric (La Rocca et al., 2018), leading to distorted asymptotic standard errors and confidence intervals. Thus, relying on maximum likelihood-based inference when analyzing small and dense networks can result in incorrect conclusions about the magnitude as well as the statistical significance of the network influence on a variable of interest.

Second, classical null hypothesis significance testing procedures for the network autocorrelation parameter can only be used to falsify the precise null hypothesis $H_0 : \rho = 0$ of no network influence. In practice though, researchers are often interested in testing multiple competing hypotheses on the network autocorrelation parameter against each other and in determining which out of these hypotheses is most supported by the data. For example, when interested in testing if the network influence is zero, positive, or negative, this could be done by testing $H_0 : \rho = 0$ versus $H_1 : \rho > 0$ versus $H_2 : \rho < 0$ against one another. Classical null hypothesis significance testing procedures, however, can merely test hypothesis H_1 and hypothesis H_2 separately against the null, while they cannot quantify the amount of evidence in favor of any of the hypotheses tested (Wetzels & Wagenmakers, 2012).

Third, in many network studies, different types of network influence are likely to be present simultaneously. For example, two households can be tied to each other based on geographic adjacency and/or social similarity, where both associated networks might exert network influence. These multiple influence mechanisms can be modeled by adding as many connectivity matrices, representing the different influence mechanisms, and network autocorrelation parameters to the first-order network autocorrelation model in (1.1) as relevant to one's theory. This leads to so-called *higher-order* network autocorrelation models, which generally allow for a richer and more realistic modeling of network dependence. Most often, researchers then have expectations about the order of strength of the different influence mechanisms that they would like to test explicitly. Such expectations can be formulated as hypotheses on the network autocorrelation parameters, e.g., as $H_1 : \rho_1 > \rho_2 > 0$, $H_2 : \rho_1 = \rho_2 > 0$, or $H_3 : 0 < \rho_1 < \rho_2$, where ρ_1 and ρ_2 correspond to the strength of two influence mechanisms, respectively. However, null hypothesis significance testing procedures for the network autocorrelation model cannot be applied to test hypotheses on the relative strength of different influence mechanisms.

Fourth, the network autocorrelation model cannot be directly used to model count data, i.e., responses that can take only non-negative integer values. Nevertheless doing so when dealing with count data would lead to predicted non-integer and potentially negative responses. In some cases, it is possible to employ appropriate data transformation techniques to convert the original counts into approximately normally distributed data and fit the network autocorrelation model to the transformed data. In many cases though, these transformations are not suitable, e.g., when modeling rare events such as homicide, and the network autocorrelation model then also cannot be indirectly sensibly applied to count data.

1.3 Addressing the inferential limitations in the network autocorrelation model

In this thesis, we will develop a Bayesian framework for first- and higher-order network autocorrelation models to address the first three inferential limitations in the network autocorrelation model outlined above. Furthermore, we will propose a discrete exponential family model to analyze network autocorrelated count data. We introduce the main ideas of the Bayesian and the discrete exponential family framework, and how these frameworks can help in overcoming the inferential limitations, next.

1.3.1 Bayesian estimation

Bayesian estimation is fundamentally different from classical frequentist estimation. In Bayesian estimation, all parameters are modeled as random variables, where the information contained in the observed data is used to update the knowledge about the parameters. This prior, i.e., before observing the data, as well as posterior, i.e., after observing the data, knowledge is expressed in terms of probability distributions for the model parameters. We denote the vector of model parameters by $\boldsymbol{\theta}$ and the joint *prior distribution* for $\boldsymbol{\theta}$ by $p(\boldsymbol{\theta})$. Sometimes, genuine prior knowledge is available, e.g., based on substantive theory or from previous empirical studies, which can be employed to specify so-called *informative prior* distributions. On the other hand, often such knowledge is absent and so-called *non-informative prior* distributions, representing prior ignorance, are relied upon instead.

In a next step, the data $\boldsymbol{y}$ are taken to update the prior distribution for the model parameters and to obtain their *posterior distribution.* Applying elementary rules of probability theory (Jeffreys, 1961), the posterior distribution for $\boldsymbol{\theta}$ given the data $\boldsymbol{y}$, $p(\boldsymbol{\theta}|\boldsymbol{y})$, can be written as

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{\theta}) f(\boldsymbol{y}|\boldsymbol{\theta})}{\int p(\boldsymbol{\theta}) f(\boldsymbol{y}|\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}}, \tag{1.2}$$

where $f(\boldsymbol{y}|\boldsymbol{\theta})$ denotes the likelihood function of the data, and the denominator of (1.2) is known as the *marginal likelihood* that ensures that the posterior distribution integrates to unity (Kass & Raftery, 1995). In Bayesian estimation, the marginal likelihood can be ignored if it is finite, whereas it plays a central role in Bayesian hypothesis testing (Lynch, 2007).

The posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$ is used for all inference in the model. For example, it can be used to compute Bayesian point estimates of a model parameter, to construct so-called *credible intervals*, i.e., intervals in the domain of the posterior that contain an unknown model parameter with a specific probability, or to quantify any other statistic of interest, such as $p(\boldsymbol{\theta}|\boldsymbol{y}) > 0$.

Apart from axiomatic considerations, estimating the network autocorrelation model taking the Bayesian route is advantageous for at least two reasons. First, including empirical prior information can potentially mitigate the negative bias in the estimation of the network autocorrelation parameter ρ for high levels of network density. Second, Bayesian inference is solely based on the posterior distribution and, in contrast to maximum likelihood-based inference, does not rely on asymptotic theory for computing standard errors or credible intervals. Consequently, uncertainty about the model parameters is appropriately accounted for even in small networks.

1.3.2 Bayesian hypothesis testing

Bayesian hypothesis testing adheres to the same inherently Bayesian principle along the lines of the previous section: all hypotheses under consideration are assigned prior probabilities before the information in the observed data is used to update the initial probabilities and to obtain posterior probabilities for the hypotheses. Assume that we are interested in testing $T \geq 2$ competing hypotheses, $H_0, ..., H_{T-1}$, against each other and that one out of the T hypotheses under consideration is the true hypothesis. We denote the prior probability for a hypothesis H_t by $p(H_t)$, $t \in \{0, ..., T-1\}$, where $\sum_{t=0}^{T-1} p(H_t) = 1$. In the absence of prior preferences for the hypotheses, these prior hypotheses probabilities are typically chosen uniformly, i.e., $p(H_0) = ... = p(H_{T-1}) = 1/T$ (Berger & Sellke, 1987; Berger et al., 1997; Raftery, 1995). On the other hand, if relevant previous empirical evidence is available, it is also possible to formulate specific prior hypotheses probabilities based on such evidence.

Subsequently, the data $\boldsymbol{y}$ are used to update the prior probabilities for the hypotheses, and the posterior hypotheses probabilities are given by Bayes' theorem as

$$p(H_t|\boldsymbol{y}) = \frac{p(H_t)\, p(\boldsymbol{y}|H_t)}{\sum\limits_{t'=0}^{T-1} p(H_{t'})\, p(\boldsymbol{y}|H_{t'})} = \frac{p(H_t) \int p_t(\boldsymbol{\theta}_t)\, f_t(\boldsymbol{y}|\boldsymbol{\theta}_t)\, \mathrm{d}\boldsymbol{\theta}_t}{\sum\limits_{t'=0}^{T-1} p(H_{t'}) \int p_{t'}(\boldsymbol{\theta}_{t'})\, f_{t'}(\boldsymbol{y}|\boldsymbol{\theta}_{t'})\, \mathrm{d}\boldsymbol{\theta}_{t'}}, \tag{1.3}$$

where $p_t(\boldsymbol{\theta}_t)$ is the prior distribution for the model parameters $\boldsymbol{\theta}_t$ under hypothesis H_t and $f_t(\boldsymbol{y}|\boldsymbol{\theta}_t)$ is the likelihood function of the data under hypothesis H_t. The marginal likelihood under hypothesis H_t in (1.3), $p(\boldsymbol{y}|H_t)$, can be seen as the average likelihood under hypothesis H_t weighted by the corresponding prior $p_t(\boldsymbol{\theta}_t)$ and represents the plausibility that the data $\boldsymbol{y}$ were observed under hypothesis H_t; the larger $p(\boldsymbol{y}|H_t)$, the more plausible that the data were observed under hypothesis H_t. The posterior hypotheses probabilities in (1.3) can then be used to quantify the plausibility of any hypothesis under consideration, including (order) hypotheses on one or multiple network autocorrelation parameters.

When testing two hypotheses H_t and $H_{t'}$, $t, t' \in \{0, ..., T-1\}$, against each other, their *posterior odds*, i.e., the ratio of their posterior probabilities, measures to what extent hypothesis H_t is favored over hypothesis $H_{t'}$. By (1.3), the posterior odds can be written as

$$\frac{p\left(H_t|\boldsymbol{y}\right)}{p\left(H_{t'}|\boldsymbol{y}\right)}=\frac{p\left(H_t\right)}{p\left(H_{t'}\right)}\frac{p\left(\boldsymbol{y}|H_t\right)}{p\left(\boldsymbol{y}|H_{t'}\right)}. \tag{1.4}$$

The first fraction on the right-hand side of (1.4), $p\left(H_t\right)/p\left(H_{t'}\right)$, is the *prior odds* of the two hypotheses that indicate how much more, or less, likely one hypothesis is compared to the other prior to observing the data. The second fraction on the right-hand side of (1.4), $p\left(\boldsymbol{y}|H_t\right)/p\left(\boldsymbol{y}|H_{t'}\right)$, is the ratio of the marginal likelihoods under the two competing hypotheses and is called the *Bayes factor* (Jeffreys, 1961). We denote the Bayes factor of hypothesis H_t against hypothesis $H_{t'}$ by $BF_{tt'}$. As such, the Bayes factor measures to what extent the data change the prior odds to the posterior odds and can be interpreted as the amount of evidence in the data in support of hypothesis H_t against hypothesis $H_{t'}$. For example, when $B_{tt'}=3$, the data are three times more likely to have occurred under hypothesis H_t rather than hypothesis $H_{t'}$. Vice versa, when $B_{tt'}=1/3$, the data are three times more likely to have occurred under hypothesis $H_{t'}$ rather than hypothesis H_t.

The Bayes factor does not depend on and can be computed without specifying prior hypotheses probabilities. Hence, when testing two competing hypotheses against each other, the Bayes factor does not assume one of the hypotheses to be true but provides relative support for the two hypotheses based on the evidence in the data. However, the Bayes factor does depend on and is sensitive to the prior distribution for the model parameters under each hypothesis (Kass & Raftery, 1995). In fact, if *improper priors*, i.e., priors that do not integrate to a finite value, on the tested model parameters are imposed, the Bayes factor depends on unspecified constants and is not well-defined (O'Hagan, 1995). Such improper priors are typically used when trying to represent the absence of prior information about the model parameters. One way to resolve this issue is to use part of the information in the observed data to obtain a proper prior distribution and subsequently compute a *pseudo-Bayes factor* with the remaining information in the data (C. Han & Carlin, 2001; Robert, 2001). In Chapters 3 and 4, we will investigate the sensitivity of Bayes factor tests for the network autocorrelation parameter(s) to various priors for the latter.

1.3.3 The discrete exponential family

The discrete exponential family is a widely used general formalism for modeling data with complex dependence structures (Butts, 2007; Robins, Pattison, et al., 2007; Strauss, 1986). In the context of this thesis, we rely on the formalism to specify the joint distribution for a random count variable $\boldsymbol{Y}$ that can take only values in a finite set of count configurations $\mathcal{Y}$, i.e., we assume the count value each actor can have to be bounded. Under the formalism, the distribution for $\boldsymbol{Y}$ is written as

$$p\left(\boldsymbol{Y}=\boldsymbol{y}|\boldsymbol{\theta}\right)=\frac{\exp\left(\boldsymbol{\theta}^T\boldsymbol{t}\left(\boldsymbol{y}\right)\right)}{\sum\limits_{\boldsymbol{y}'\in\mathcal{Y}}\exp\left(\boldsymbol{\theta}^T\boldsymbol{t}\left(\boldsymbol{y}'\right)\right)}, \tag{1.5}$$

where $\boldsymbol{y}\in\mathcal{Y}$ is an attainable count configuration, $\boldsymbol{\theta}$ is a vector of real-valued parameters, and $\boldsymbol{t}(\boldsymbol{y})$ denotes a vector of sufficient statistics. The denominator of (1.5) is a normalizing constant that ensures that the defined probability distribution sums to unity.

This probability distribution is formulated in terms of its sufficient statistics that serve as summary measures of structural properties of a joint count configuration. While in principle, there are no constraints on the choice of the sufficient statistics, this choice is guided by including those sufficient statistics that describe structural properties that are linked to mechanisms giving rise to an observed count configuration. In other words, and slightly abusing notation, the sufficient statistics are specified such to represent structural properties that are well-known to be enhanced, when $\boldsymbol{\theta} > 0$, or suppressed, when $\boldsymbol{\theta} < 0$, under certain mechanisms.

Hence, the discrete exponential family formalism can be used to model network autocorrelated count data by constructing sufficient statistics that appropriately capture the net tendency of tied actors to show similar (or dissimilar) counts, commonly understood as positive (or negative) network autocorrelation. Likewise, additional sufficient statistics incorporating standard covariate or any other effects on the counts can be designed.

1.4 Outline of the thesis

The core of this thesis consists of the ensuing four chapters that have been written as journal articles and can be read comprehensibly independently from each other. Chapters 2, 3, and 4 focus on developing Bayesian methodology for the network autocorrelation model. Chapter 5 introduces a discrete exponential family model for analyzing network autocorrelated count data and is slightly less expository in style than the other chapters.

In Chapter 2, we provide new Bayesian estimation methods for the network autocorrelation model that address the issues inherent to maximum likelihood estimation of the model. For any Bayesian estimator, the prior distribution for the model parameters impacts the properties and the performance of the estimator. Thus, we motivate and derive several priors for the parameters in the network autocorrelation model in this chapter. We first derive two versions of *Jeffreys prior* (Jeffreys, 1961), *Jeffreys rule prior* and *Independence Jeffreys prior*, which have not yet been established for the network autocorrelation model. Jeffreys prior construes the concept of a non-informative prior in a formal way, and these priors can be used for a Bayesian estimation of the network autocorrelation model when prior information is unavailable. Second, we propose an informative prior for the network autocorrelation parameter ρ based on a meta-analysis of empirical applications of the network autocorrelation model. This is the first empirically justified informative prior for ρ to be found in the literature and systematically shows that positive network autocorrelation is much more prevalent in empirical practice than negative network autocorrelation. All of the resulting posterior distributions do not belong to a family of known probability distributions and hence, summary measures for the posteriors, such as the mean or quantiles, are analytically not available. We present new and more efficient procedures than currently to be found in the literature for sampling from the corresponding posterior distributions. Lastly, we conduct a simulation study to evaluate the performance of the different Bayesian estimators as well as the maximum likelihood estimator.

In Chapter 3, we introduce Bayesian hypothesis tests for the network autocorrelation parameter ρ, including precise hypotheses, e.g., $H_0 : \rho = 0$, as well as interval hypotheses, e.g., $H_1 : 0 < \rho < 1$. When testing such hypotheses on the network autocorrelation parameter using the Bayes factor, the Bayes factor can be sensitive to the choice of the prior distribution for ρ under interval hypotheses. We present three Bayes factors for testing the aforementioned hypotheses that are not yet available for the network autocorrelation model: first, we consider a Bayes factor using the empirical informative prior from Chapter 2; second, a Bayes factor based on the standard uniform prior for ρ typically used in the literature (Hepple, 1995a; Holloway et al., 2002; LeSage, 1997a); third, we develop a so-called *fractional Bayes factor* (O'Hagan, 1995) where a default prior for ρ is automatically constructed by updating an improper prior with a fraction of the information contained in the observed data, while the fractional Bayes factor itself is computed using the remaining information in the data. Furthermore, we employ the empirical informative prior also for specifying prior probabilities for interval hypotheses, which is new to the literature. We show how the presented methodology can be straightforwardly adapted to test more than two hypotheses against each other, which is particularly relevant in the network autocorrelation model, as the literature suggests that the question is not if network autocorrelation occurs but to what extent. We carry out a simulation study to investigate numerical properties of these Bayes factors, and we demonstrate their use by re-analyzing three empirical data sets from the literature.

In Chapter 4, we extend the developed methods for the first-order network autocorrelation model in Chapters 2 and 3 to higher-order network autocorrelation models involving multiple connectivity matrices and network autocorrelation parameters. As to that, we propose computationally efficient Bayesian estimation techniques for higher-order network autocorrelation models based on a general multivariate normal prior for the network autocorrelation paramaters. Moreover, we introduce adeptly implemented Bayes factors for simultaneously testing multiple order hypotheses on the network autocorrelation parameters against one another. Our Bayes factors are based on automatically constructed multivariate normal default priors for the network autocorrelation paramaters, which eliminates the need for difficult prior elicitation under each hypothesis. As such, the proposed Bayes factors provide means for quantifying the relative evidence in the data in favor of any hypothesis on the network autocorrelations parameters, including equality constraints, e.g., $H_0 : \rho_1 = \rho_2 = 0$, inequality constraints, e.g., $H_1 : \rho_1 > \rho_2 > 0$, or a combination of equality and inequality constraints, e.g., $H_2 : \rho_1 > \rho_2 = 0$. Subsequently, we investigate to what extent the Bayes factors provide evidence for a true data-generating hypothesis when tested against several competing hypotheses by means of a simulation study. Finally, we illustrate our methods on a data set from the economic growth theory, where we explore the structure of spatial autocorrelation of growth rates of labor productivity in service industry across 188 territorial units in the European Union.

In Chapter 5, we present a discrete exponential family model for analyzing network autocorrelated count data. In our approach, we use the discrete exponential family to specify the joint count distribution in terms of its sufficient statistics. We propose various sets of sufficient statistics representing network autocorrelation, standard covariate effects, and basic structural properties of the counts, such as the central tendency, the dispersion, and the sparsity of the counts, that are deemed to have generated an observed count configuration. Furthermore, we provide algorithms to simulate count configurations and to carry out maximum likelihood-based inference in the model. In addition, we introduce tailored goodness-of-fit measures based on the predictive distribution for the counts. Lastly, we illustrate the capability and the practical implementation of our model by re-investigating the causes of homicide in 343 neighborhoods in Chicago, Illinois.

In Chapter 6, the final chapter of this thesis, we summarize our main results and reflect upon their usefulness in resolving the inferential limitations in the network autocorrelation model that inspired this thesis. We conclude by pointing out remaining issues and by discussing future research topics related to network autocorrelation modeling.

Chapter 2

Bayesian estimation of the network autcorrelation model

Abstract

The network autocorrelation model has been extensively used by researchers interested modeling social influence effects in social networks. The most common inferential method in the model is classical maximum likelihood estimation. This approach, however, has known problems such as negative bias of the network autocorrelation parameter and poor coverage of confidence intervals. In this chapter, we develop new Bayesian estimation techniques for the network autocorrelation model that address the issues inherent to maximum likelihood estimation. A key ingredient of the Bayesian approach is the choice of the prior distribution. We derive two versions of Jeffreys prior, Jeffreys rule prior and Independence Jeffreys prior, which have not yet been developed for the network autocorrelation model. These priors can be used for Bayesian estimation of the model when prior information is unavailable. Moreover, we propose an informative as well as a weakly informative prior for the network autocorrelation parameter that are both based on an extensive literature review of empirical applications of the network autocorrelation model across many fields. Finally, we provide new and efficient Markov Chain Monte Carlo algorithms to sample from the resulting posterior distributions. Simulation results suggest that the considered Bayesian estimators outperform the maximum likelihood estimator with respect to bias of and frequentist coverage of credible and confidence intervals for the network autocorrelation parameter.

This chapter is based on: Dittrich, D., Leenders, R.Th.A.J., and Mulder, J. (2017). Bayesian estimation of the network autocorrelation model. *Social Networks*, 48(1): 213-236. `http://doi.org/10.1016/j.socnet.2016.09.002`.

2.1 Introduction

Identifying and estimating network influence on individual behavior is a common and important challenge encountered in social network analysis. Throughout the last decades, a number of different models studying network influence effects have emerged, out of which the network autocorrelation model is probably the most popular one (Leenders, 2002; Marsden & Friedkin, 1993; Plümper & Neumayer, 2010; W. Wang et al., 2014).

A traditional and widely used technique for parameter estimation in the network autocorrelation model is maximum likelihood estimation (Doreian, 1981; Ord, 1975), which has also been implemented in common statistical software packages such as R (Bivand & Piras, 2015; Butts, 2008; Leifeld et al., 2015; McMillen, 2013; Wilhelm & Godinho de Matos, 2015), MATLAB (LeSage, 1999), Python (Rey & Anselin, 2007), and Stata (Pisati, 2001). Despite the popularity and usefulness of maximum likelihood estimation, there are also important issues related to this estimation technique of the model. First, several simulation studies have suggested that the maximum likelihood estimator for the *network autocorrelation parameter* ρ is negatively biased under many different scenarios, that the underestimation of ρ becomes more severe for increasing network density, and that it occurs regardless of the network structure and the network size (Mizruchi & Neuman, 2008; Neuman & Mizruchi, 2010; Smith, 2009). Second, maximum likelihood-based precision estimates, such as confidence intervals, rely heavily on asymptotic theory. Consequently, the coverage of the associated confidence intervals may be distorted for small to medium samples that are often encountered in social science research, such as school classes, care teams, or members of an executive board. Notwithstanding the tremendous capability of the network autocorrelation model and the theoretical advances it has yielded for understanding the structure of social influence in social networks, the concerns regarding the maximum likelihood estimation approach may ultimately discourage researchers from utilizing the model at all.

In this chapter, we develop Bayesian statistical estimation methods for the network autocorrelation model that may attenuate the issues which have been encountered with maximum likelihood estimation. The Bayesian approach has at least two attractive features that are not shared by classical methods. First, it allows researchers to incorporate external information about the model parameters via a *prior distribution.* For example, if previous research has suggested that people in a certain network are positively influenced by each other, as is often the case in social networks, one could specify a prior distribution that assumes positive values for the network autocorrelation ρ to be more likely than negative ones. Indeed, as we will show in Section 2.4, the vast empirical literature on the model suggests that network effects are much more likely, a priori, to be in certain intervals than in others. Second, Bayesian analysis provides “exact” inference without the need for asymptotic approximations (De Oliveira & Song, 2008). This characteristic is especially appealing for small- to moderate-sized groups and can be seen as a distinct advantage of the Bayesian approach over classical frequentist methods. In other words, when networks are small, Bayesian estimation of the network autocorrelation model is statistically preferable over frequentist estimation.

Bayesian statistics is a fundamentally different approach than classical statistics. In brief, a Bayesian data analysis is carried out as follows. First, a prior distribution, or simply *prior*, for the model parameters is needed, where the prior distribution reflects the prior knowledge about the model parameters before observing the data. If prior information is available, e.g., based on published literature, an *informative prior* can be specified. On the other hand, if such information is absent, a so-called *non-informative prior* can be employed. After observing the data, Bayes' theorem is used to update the prior expectations with the information contained in the data to arrive at the *posterior distribution*, or *posterior*, for the model parameters. All inference is based on the posterior, and it is used to obtain Bayesian point estimates and *credible intervals*, the Bayesian equivalent to classical confidence intervals.

Although the specification of the prior distribution is one of the most important steps in any Bayesian analysis, it has not received much attention in the literature on Bayesian estimation of the network autocorrelation model (X. Han & Lee, 2013; Hepple, 1995b; Holloway et al., 2002; LeSage, 2000; LeSage & Pace, 2009), with the exception of LeSage (1997a) and LeSage & Parent (2007). In some cases, it is in fact difficult to elicit a prior, e.g., when prior information is absent, or when a researcher would like to add as little prior information as possible to the analysis. For these situations, non-informative priors are typically used to carry out a Bayesian analysis. In this chapter, we are the first to derive two versions of *Jeffreys prior* (Jeffreys, 1961), called *Jeffreys rule prior* and *Independence Jeffreys prior*, for the network autocorrelation model and to establish results on the propriety of the resulting posterior distributions. Jeffreys rule prior construes the concept of a non-informative prior in a formal way and is the most commonly used non-informative prior (De Oliveira & Song, 2008). Moreover, in several simulation studies of related autoregressive models, Independence Jeffreys prior has been shown to result in superior inferences compared to those based on maximum likelihood estimation (De Oliveira, 2012; De Oliveira & Song, 2008). These findings serve as another motivation to consider the two versions of Jeffreys prior for the network autocorrelation model as well.

Furthermore, we provide a novel informative prior for the network effect ρ based on an extensive literature review of empirical applications of the network autocorrelation model. To the best of our knowledge, this is the first empirically justified informative prior for ρ to be found in the literature. Because of the empirical justification of this prior, it is a reasonable "entry point" for a Bayesian analysis of the network autocorrelation model, as it summarizes the currently available evidence about observed network autocorrelations from many different sources. Moreover, we introduce a related *weakly informative prior* for ρ that can be used by a researcher who agrees that past findings should not be dismissed but who is at the same time reluctant and deliberately refrains from including all available prior information.

In addition, we present efficient *Markov Chain Monte Carlo* (MCMC) algorithms for sampling from the resulting posterior distributions, which we find to be computationally superior compared to existing schemes (LeSage, 2000; LeSage & Pace, 2009). We conduct a simulation study to investigate numerical properties of Bayesian inferences about the network effect ρ and the error variance σ^2 based on the proposed priors and to compare them to inferences coming from maximum likelihood estimation. As will be shown, the Bayesian estimator based on the informative prior performs overall the best when network effects are positive, while using the weakly informative prior eliminates virtually all the negative bias in the estimation of ρ in case of no or marginal network effects.

We proceed as follows: in Section 2.2, we discuss the network autocorrelation model in more detail. We continue with a short introduction to the Bayesian approach in regard to the model in Section 2.3. In Section 2.4, we derive two versions of Jeffreys prior and propose an informative as well as a weakly informative prior for the network autocorrelation parameter ρ based on reported network effects from the literature. Moreover, we state properties of these priors and their corresponding posteriors and provide comparisons between the priors. In Section 2.5, we present efficient MCMC implementations for Bayesian estimation of the model. We assess the numerical performance of the Bayesian estimators and the maximum likelihood estimator in a simulation study in Section 2.6. Section 2.7 concludes.

2.2 The network autocorrelation model

Originally developed by geographers (Ord, 1975), the network autocorrelation model has been used to address the problem of structured dependence ever since. In contrast to a standard linear regression model, the network autocorrelation model does not assume observations for a variable of interest to be independent from each other but allows for dependence among them. In a social network context, this has the interpretation that ego's opinion may not solely depend on exogenous variables; instead, ego's opinion might be influenced by the opinions of other actors in the network as well. Thus, in the network autocorrelation model, ego's opinion is viewed as a combination of interaction and exogenous variables, formally expressed as

$$\boldsymbol{y} = \rho W \boldsymbol{y} + X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N\left(\mathbf{0}_g, \sigma^2 I_g\right), \tag{2.1}$$

where, as in standard linear regression, $\boldsymbol{y}$ is a vector of length g consisting of values for a dependent variable for the g network actors, X is a $(g \times k)$ matrix of values for the g actors on k covariates (possibly including a vector of ones in the first column for an intercept term), $\boldsymbol{\beta}$ is a vector of regression coefficients of length k, $\mathbf{0}_g$ is a vector of zeros of length g, I_g symbolizes the $(g \times g)$ identity matrix, and $\boldsymbol{\varepsilon}$ is a vector of length g containing independent and identically normally distributed error terms with zero mean and variance of σ^2. Furthermore, W denotes a given $(g \times g)$ *connectivity matrix* representing social ties in a network, where each entry W_{ij} stands for the degree of influence of actor j (alter) on actor i (ego). By convention, we exclude loops, i.e., relationships from an actor to himself, so $W_{ii} = 0$ for all $i \in \{1, ..., g\}$. Finally, ρ is a scalar termed the network autocorrelation parameter. It is the key parameter of the model and measures the level of network influence on a variable of interest for given $\boldsymbol{y}$, W, and X. Note that when $\rho = 0$, the model reduces to a standard linear regression model.

The likelihood of the model in (2.1) is given by

$$f\left(\boldsymbol{y}|\rho,\sigma^2,\boldsymbol{\beta}\right) = \left|\det\left(A_\rho\right)\right|\left(2\pi\sigma^2\right)^{-\frac{g}{2}}\exp\left(-\frac{1}{2\sigma^2}\left(A_\rho\boldsymbol{y}-X\boldsymbol{\beta}\right)^T\left(A_\rho\boldsymbol{y}-X\boldsymbol{\beta}\right)\right), \tag{2.2}$$

where $A_\rho := I_g - \rho W$ (see e.g., Doreian, 1980). To ensure that $\left|\det\left(A_\rho\right)\right|$ is non-zero and the model's likelihood function in (2.2) is well-defined, there are restrictions on the feasible values for ρ. In practice, the admissible range of ρ is usually chosen as the interval containing $\rho = 0$ for which A_ρ is non-singular (Hepple, 1995a; Holloway et al., 2002; Lee, 2004; LeSage, 1997a, 2000; LeSage & Pace, 2009; Smith, 2009). This interval is given by $\left(\lambda_g^{-1}, \lambda_1^{-1}\right)$, where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_g$ are the ordered eigenvalues of W (Hepple, 1995a), and we follow this convention in the remainder of the chapter.[1,2] We denote the resulting set of model parameters by $\boldsymbol{\theta} := \left(\rho,\sigma^2,\boldsymbol{\beta}\right)$ and the associated parameter space by $\Theta := \Theta_\rho \times \Theta_{\sigma^2} \times \Theta_{\boldsymbol{\beta}} = \left(\lambda_g^{-1}, \lambda_1^{-1}\right) \times (0,\infty) \times \mathbb{R}^k$. Hence, the model's parameter space has the remarkable property that it depends on properties, i.e., the eigenvalues, of the connectivity matrix W.

Throughout the literature, the model is also referred to as mixed regressive-autoregressive model (Ord, 1975), spatial effects model (Doreian, 1980), network effects model (Dow et al., 1982), or spatial lag model (Anselin, 2002), and it has been applied in many different fields, such as criminology (Baller et al., 2001; Fornango, 2010; Tita & Radil, 2011), geography (Fingleton, 2001; McMillen, 2010; Seldadyo et al., 2010), political science (Beck et al., 2006; Shin & Ward, 1999; Tam Cho, 2003), and sociology (Kirk & Papachristos, 2011; Land et al., 1991; Ruggles, 2007).

2.3 Bayesian network autocorrelation modeling

The starting point of every Bayesian analysis is the formulation of prior expectations about the parameters in a statistical model. Formally, these prior expectations are expressed in terms of probability distributions, where the resulting prior distributions represent the available knowledge about the model parameters before observing data. We denote the joint prior distribution for all model parameters by $p\left(\boldsymbol{\theta}\right)$. In general, prior expectations can come from a researcher's beliefs or from accumulated empirical evidence from previous

[1] To avoid unnecessary complications, we restrict ourselves to connectivity matrices with real eigenvalues. These include all W that are either symmetric or *row standardizations*, i.e., where each row sums to one, of symmetric matrices (Smith, 2009). Furthermore, we assume that $\lambda_1 > 0$, which includes all non-zero symmetric connectivity matrices (Smith, 2009), so $\lambda_g < 0 < \lambda_1$ since $\mathrm{tr}\left(W\right) = 0$. In the common case of row-standardized connectivity matrices, it follows that $\lambda_1 = 1$ (Anselin, 1982).

[2] It is mathematically not necessary to constrain the parameter space of ρ to $\left(\lambda_g^{-1}, \lambda_1^{-1}\right)$. It suffices to exclude the reciprocals of the eigenvalues of W from the domain of ρ, as A_ρ is singular only for those values (Leenders, 1995). Some authors prefer to restrict ρ to $(-1,1)$, as $\forall\rho \in (-1,1) : A^{-1} = \sum_{k=0}^{\infty}\rho^k W^k$, which implies an underlying stationary process (Griffith, 1979). We choose the interval $\left(\lambda_g^{-1}, \lambda_1^{-1}\right)$ rather than $(-1,1)$ as admissible range of ρ, as the latter choice might yield estimates of ρ at the lower boundary of the interval and considering the whole parameter space $\mathbb{R}\setminus\left\{{\lambda_1}^{-1}, {\lambda_2}^{-1}, ..., {\lambda_g}^{-1}\right\}$ typically results in *improper* posterior distributions (see the remark in the proof of Corollary 2.1). Lastly, $\forall\rho \in \left(\lambda_g^{-1}, \lambda_1^{-1}\right) : \det\left(A_\rho\right) > 0$, so we write $\left|A_\rho\right|$ for $\left|\det\left(A_\rho\right)\right|$ in the remainder of the chapter.

studies in a field. Alternatively, one might also (purposely) stay vague and opt for a non-informative prior distribution. The idea of a non-informative prior is that it is completely dominated by the data and different methods have been proposed how to construct such priors (Bernardo, 1979; Box & Tiao, 1973; **?**; Kass & Wasserman, 1996).

After having specified a prior distribution, the data $\boldsymbol{y}$ are observed. Since the data contain information about the unknown parameters, they can be used to update the initial expectations. The information about the model parameters in the data is summarized by the likelihood function, $f\left(\boldsymbol{y}|\boldsymbol{\theta}\right)$. Linking information from the prior distribution and the data leads to the posterior distribution for the model parameters, which we denote by $p\left(\boldsymbol{\theta}|\boldsymbol{y}\right)$. Applying elementary rules of probability theory, the posterior can be written by Bayes' theorem as

$$p\left(\boldsymbol{\theta}|\boldsymbol{y}\right) = \frac{f\left(\boldsymbol{y}|\boldsymbol{\theta}\right)p\left(\boldsymbol{\theta}\right)}{p\left(\boldsymbol{y}\right)}. \tag{2.3}$$

The denominator of (2.3) is called the *marginal likelihood* and serves as normalizing constant to ensure that the posterior integrates to unity. However, as the normalizing constant does not depend on the model parameters and does not affect parameter estimation, the expression in (2.3) can be simplified to

$$p\left(\boldsymbol{\theta}|\boldsymbol{y}\right) \propto f\left(\boldsymbol{y}|\boldsymbol{\theta}\right)p\left(\boldsymbol{\theta}\right). \tag{2.4}$$

Hence, (2.4) means that the posterior distribution is proportional, with respect to the model parameters $\boldsymbol{\theta}$, to the prior distribution multiplied by the likelihood function. Formally, the normalizing constant can only be dropped if it is finite, i.e., if the posterior is integrable and thus a proper probability distribution. For the network autocorrelation model, this is the case when the network size, compared to the number of covariates, is large enough. We will come back to this in the following section.

The posterior distribution can then be used to derive point estimates of the model parameters (e.g., the posterior mean or the posterior median), credible intervals (i.e., intervals in the domain of the posterior), or to get other statistics of interest, such as the probability that the network autocorrelation is positive for given data, $p\left(\rho > 0|\boldsymbol{y}\right)$. The latter statistic is quite useful for quantifying a researcher's belief that people in a network positively influence each other with respect to some variable of interest. However, such a probability cannot be obtained when using classical frequentist methods but only when taking the Bayesian route.

2.4 Prior choices in the network autocorrelation model

The specification of the prior distribution is one of the most important steps in a Bayesian analysis. Despite its importance, prior specification in the network autocorrelation model has been largely neglected. Most of the previous work on Bayesian estimation of the model has been based on using uniform priors for ρ, $\boldsymbol{\beta}$, and $\log\left(\sigma^2\right)$ (Hepple, 1995b; Holloway et al., 2002; LeSage, 1997a, 2000). Only recently, X. Han & Lee (2013) and LeSage & Pace

(2009) considered the standard normal inverse gamma priors for $\boldsymbol{\beta}$ and σ^2 from linear regression, resulting in an inverse gamma prior for σ^2 and a normal prior for $\boldsymbol{\beta}$ conditional on σ^2, together with the standard uniform prior for ρ.

In this section, we briefly review the standard uniform prior for ρ first, before deriving two versions of Jeffreys prior and proposing two novel informative priors for the network effect ρ.

2.4.1 Flat prior

Using *flat*, or *uniform*, *priors* is the simplest and most intuitive way to quantify prior ignorance about model parameters. A uniform prior assigns equal, or uniform, probability to all possible values a parameter can attain, resulting in a flat prior density function. Applying this rationale to the network autocorrelation model means that all possible network effects ρ and regression coefficients $\boldsymbol{\beta}$ are considered as equally likely before observing the data. In mathematical notation, we denote the flat prior distributions for ρ and $\boldsymbol{\beta}$ by $p_{\mathrm{F}}(\rho) \propto 1$ and $p_{\mathrm{F}}(\boldsymbol{\beta}) \propto 1$, respectively. As noted in the previous section, for estimation purposes it suffices to give the prior distributions in these unnormalized forms. Furthermore, the error variance σ^2 is constrained to the positive axis, and it is customary to consider its logarithm and assign a flat prior to this transformed variable (Fernández et al., 2001; Kass & Wasserman, 1996). Retransforming the flat prior for $\log\left(\sigma^2\right)$ back in terms of σ^2 yields $p_{\mathrm{F}}\left(\sigma^2\right) \propto 1/\sigma^2$. Finally, under the flat prior all parameters are assumed to be a priori independent. The flat prior for $\boldsymbol{\theta} = \left(\rho, \sigma^2, \boldsymbol{\beta}\right)$ is then written as

$$p_{\mathrm{F}}(\boldsymbol{\theta}) = p_{\mathrm{F}}(\rho) \times p_{\mathrm{F}}\left(\sigma^2\right) \times p_{\mathrm{F}}(\boldsymbol{\beta}) \propto 1/\sigma^2.$$

This prior is sometimes also referred to as the diffuse prior in the literature (Hepple, 1979; LeSage, 1997a, 2000). While it is obvious that the flat prior itself is improper, i.e., the integral of $p_{\mathrm{F}}(\boldsymbol{\theta})$ on Θ is not finite, it is easy to verify that the resulting posterior distribution is proper under the very weak conditions stated in Corollary 2.1.

Corollary 2.1. *Consider the network autocorrelation model in* (2.1)*. Then,*

(i) The flat prior $p_F(\boldsymbol{\theta})$ *is unbounded and not integrable on* $\Theta = \left(\lambda_g^{-1}, \lambda_1^{-1}\right) \times (0, \infty) \times \mathbb{R}^k$.

(ii) The corresponding posterior $p_F(\boldsymbol{\theta}|\boldsymbol{y})$ *is proper on* $\Theta = \left(\lambda_g^{-1}, \lambda_1^{-1}\right) \times (0, \infty) \times \mathbb{R}^k$ *when* $g > k$, $\left(X^T X\right)^{-1}$ *exists, and* $\left(\boldsymbol{y}^T M W \boldsymbol{y}\right)^2 \neq \boldsymbol{y}^T W^T M W \boldsymbol{y}\boldsymbol{y}^T M \boldsymbol{y}$*, where* $M := I_g - X\left(X^T X\right)^{-1} X^T$.

Proof. See Appendix 2.B. ■

Thus, given the two mild regularity conditions in Corollary 2.1 (ii) hold, the flat prior yields a proper posterior when the number of actors in a network is larger than the number of external covariates. While the first regularity condition can be easily controlled for by avoiding perfect collinearity, the second one is of technical nature and needs to be checked for each data set.

2.4.2 Jeffreys rule prior

Flat priors are only one possible way to state prior ignorance; they are driven mainly by what intuitively seems to represent non-informativeness, rather than being based on a set of formal rules that defines non-informativeness mathematically. The first formal rule for specifying non-informative prior distributions was introduced by Sir Harold Jeffreys, and much of subsequent related work is based on modifications of Jeffreys' scheme (Jeffreys, 1961; Kass & Wasserman, 1996). The main motivation for Jeffreys rule prior is that statistical inference should not depend on any specific parametrization of the model, which could often be rather arbitrary. For example, if the network autocorrelation model is rewritten in terms of a precision parameter $\omega := 1/\sigma^2$, rather than σ^2, applying Jeffreys rule prior to the model formulated with respect to ω or σ^2 results in the same posterior conclusions about the network effect. Hence, when using Jeffreys rule prior, there is no need to determine a privileged parametrization as the prior is parametrization-invariant. Formally, Jeffreys rule prior is defined as

$$p_{\mathrm{J}}(\boldsymbol{\theta}) \propto \sqrt{\det\left(I\left(\boldsymbol{\theta}\right)\right)},$$

where $I(\boldsymbol{\theta})$ denotes the model's Fisher information matrix. The exact analytical form of the prior is given in Theorem 2.1. Since Jeffreys rule prior for the network autocorrelation model is improper, the propriety of the resulting posterior needs to be checked and is verified in Corollary 2.2.

Theorem 2.1. *Consider the network autocorrelation model in* (2.1) *and assume that* $\left(X^TX\right)^{-1}$ *exists. Then, Jeffreys rule prior for* $\boldsymbol{\theta} = \left(\rho, \sigma^2, \boldsymbol{\beta}\right)$*, denoted by* $p_J(\boldsymbol{\theta})$*, is*

$$p_J(\boldsymbol{\theta}) \propto \left(\sigma^2\right)^{-\frac{k+2}{2}} \left\{\operatorname{tr}\left(B_\rho^T B_\rho\right) + \operatorname{tr}\left(B_\rho^2\right) + \frac{1}{\sigma^2}\boldsymbol{\beta}^T X^T B_\rho^T M B_\rho X \boldsymbol{\beta} - \frac{2}{g}\operatorname{tr}^2\left(B_\rho\right)\right\}^{\frac{1}{2}}, \quad (2.5)$$

where $B_\rho := W A_\rho^{-1}$.

Proof. See Appendix 2.B. ■

Corollary 2.2. *Consider the network autocorrelation model in* (2.1) *and assume that* $\left(X^TX\right)^{-1}$ *exists. Then,*

(i) Jeffreys rule prior $p_J(\boldsymbol{\theta})$ *is unbounded and not integrable on* $\Theta = \left(\lambda_g^{-1}, \lambda_1^{-1}\right) \times (0, \infty) \times \mathbb{R}^k$.

(ii) Jeffreys rule posterior $p_J(\boldsymbol{\theta}|\boldsymbol{y})$ *is proper on* $\Theta = \left(\lambda_g^{-1}, \lambda_1^{-1}\right) \times (0, \infty) \times \mathbb{R}^k$ *when* $\left(\boldsymbol{y}^T M W \boldsymbol{y}\right)^2 \neq \boldsymbol{y}^T W^T M W \boldsymbol{y} \boldsymbol{y}^T M \boldsymbol{y}$.

Proof. See Appendix 2.B. ■

2.4.3 Independence Jeffreys prior

Jeffreys rule prior has the desirable property to be invariant under one-to-one parameter transformations and most often results in reasonable priors in one-dimensional models. However, applying the rule in multi-parameter models may result in poor frequentist properties of Bayesian inferences (Berger et al., 2001; De Oliveira, 2010; De Oliveira & Song, 2008), or even improper posteriors (Berger et al., 2001; Bolstad, 2009; Rubio & Steel, 2014). Thus, already Jeffreys himself argued that it is often better to consider certain blocks of parameters as a priori "independent" from each other and to compute the marginal prior for each parameter block using Jeffreys rule, assuming the other parameters to be known (De Oliveira & Song, 2008). The resulting product of the marginal priors is then called Independence Jeffreys prior. Following Bayesian analyses of related autoregressive models (Berger et al., 2001; De Oliveira, 2012; De Oliveira & Song, 2008), we split the network autocorrelation model parameters into the two blocks $\left(\rho, \sigma^2\right)$ and $\boldsymbol{\beta}$ and derive Independence Jeffreys prior based on this partitioning of the model parameters. We give the prior's analytical form in Theorem 2.2 and provide its main theoretical properties in Corollary 2.3.

Theorem 2.2. *Consider the network autocorrelation model in* (2.1). *Then, Independence Jeffreys prior for* $\boldsymbol{\theta} = \left(\rho, \sigma^2, \boldsymbol{\beta}\right)$, *denoted by* $p_{IJ}(\boldsymbol{\theta})$, *is*

$$p_{IJ}(\boldsymbol{\theta}) \propto \frac{1}{\sigma^2}\left\{\operatorname{tr}\left(B_\rho^T B_\rho\right) + \operatorname{tr}\left(B_\rho^2\right) + \frac{1}{\sigma^2}\boldsymbol{\beta}^T X^T B_\rho^T B_\rho X \boldsymbol{\beta} - \frac{2}{g}\operatorname{tr}^2\left(B_\rho\right)\right\}^{\frac{1}{2}}. \tag{2.6}$$

Proof. See Appendix 2.B. ■

Corollary 2.3. *Consider the network autocorrelation model in* (2.1). *Then,*

(i) Independence Jeffreys prior $p_{IJ}(\boldsymbol{\theta})$ *is unbounded and not integrable on* $\Theta = \left(\lambda_g^{-1}, \lambda_1^{-1}\right) \times (0, \infty) \times \mathbb{R}^k$.

(ii) Independence Jeffreys posterior $p_{IJ}(\boldsymbol{\theta}|\boldsymbol{y})$ *is proper on* $\Theta = \left(\lambda_g^{-1}, \lambda_1^{-1}\right) \times (0, \infty) \times \mathbb{R}^k$ *when* $g > k$, $\left(X^T X\right)^{-1}$ *exists, and* $\left(\boldsymbol{y}^T M W \boldsymbol{y}\right)^2 \neq \boldsymbol{y}^T W^T M W \boldsymbol{y}\boldsymbol{y}^T M \boldsymbol{y}$.

Proof. See Appendix 2.B. ■

The analytical expression of Independence Jeffreys prior in (2.6) is similar, but slightly simpler, to the one of Jeffreys rule prior in (2.5). The major difference between the two is that for Jeffreys rule prior the exponent of the error variance depends on the number of covariates, k, while it does not for Independence Jeffreys prior. For related models (Berger et al., 2001; De Oliveira, 2012; De Oliveira & Song, 2008), it has been shown that having k in the exponent of σ^2, as in Jeffreys rule prior, could result in an underestimation of the error variance. We will therefore investigate whether this is also the case in the network autocorrelation model, and if this underestimation occurs, whether it can be circumvented by using Independence Jeffreys prior. Hence, while Jeffreys rule prior is based on an invariance principle, Independence Jeffreys prior is a heuristic modification of Jeffreys rule prior that can result in better inferences.

2.4.4 An informative prior for ρ

Having discussed three prominent non-informative priors above, in this subsection, we derive a population distribution for ρ based on an extensive literature review of empirical applications of the network autocorrelation model. Subsequently, this population distribution is used as an informative prior for ρ.

In our literature search, we considered 81 peer-reviewed publications and a total of 183 estimated network autocorrelation parameters. The most important characteristics of this sample are summarized in Table 2.1.[3] As network effects in one publication are usually from the same field and closely related, this suggests that these network autocorrelations are more similar than network effects coming from different studies. To take this into account, we relied on a hierarchical approach and used the following multilevel model (Gelman et al., 2003) to estimate the population distribution of the network effects:

$$\begin{aligned} &\text{Level 1: } \rho_{ij} \sim N\left(\rho_j, \sigma_\rho^2\right), \\ &\text{Level 2: } \rho_j \sim N\left(\mu_\rho, \tau_\rho^2\right), \end{aligned} \tag{2.7}$$

where $i \in \{1, ..., n_j\}$, $j \in \{1, ..., 81\}$, ρ_{ij} is the observed i-th network effect from field j, and $\{\rho_j\}_j, \mu_\rho, \sigma_\rho^2$, and τ_ρ^2 are model parameters that have to be estimated. The distribution in Level 1 corresponds to the empirical distribution of a network effect in a specific field. The distribution in Level 2 denotes the overall population distribution in which we are ultimately interested in. We fitted the model in R (R Core Team, 2017) relying on a Bayesian framework and using standard non-informative uniform priors for μ_ρ, τ_ρ, and $\log\left(\sigma_\rho^2\right)$ (Gelman, 2006). This resulted in posterior mean estimates of $\mu_\rho = .36$ and $\tau_\rho = .19$.[4] The resulting informative prior for ρ, $p_{\text{EI}}(\rho) \sim N\left(\mu_\rho, \tau_\rho^2\right)$, and the histogram of the average network effects from each field are plotted in Figure 2.1. As can be seen, the multilevel model in (2.7) provides a reasonably good fit to the empirical data.

Figure 2.1 also shows that there are substantially more reported positive network effects than negative ones in the literature. This finding conflicts with a flat prior for ρ on $\left(\lambda_g^{-1}, \lambda_1^{-1}\right)$, which typically implies that negative network effects are a priori more likely than positive network effects and is clearly unrealistic.[5]

We combine this empirical informative prior for ρ with the standard non-informative prior for $\left(\sigma^2, \boldsymbol{\beta}\right)$ from Section 2.4.1, assuming all parameters to be a priori independent.[6]

[3] We did not attempt to be fully comprehensive here and do not claim to have included all available literature on empirical applications of the network autocorrelation model. Our selection features work that (i) used row-standardized connectivity matrices, (ii) specified the network size and the type of connectivity matrix, and (iii) employed appropriate estimation techniques for the given type of data.

[4] The associated 95% credible intervals for μ_ρ and τ_ρ were $(.33, .39)$ and $(.16, .22)$, respectively.

[5] For row-standardized connectivity matrices, it holds that $\lambda_g^{-1} \leq -1$ (Stewart, 2009, Property 10.1.2), and for most of the simulated data sets we considered, we observed that $\lambda_g^{-1} < -1$. Thus, as $\lambda_1^{-1} = 1$, in these cases the flat prior assigns more probability mass to negative network effects than to positive ones.

[6] Propriety of the resulting posterior distribution, under the conditions given in Corollary 2.1, follows immediately from the corollary's proof. Our informative prior for ρ can be easily combined with informative priors for σ^2 and $\boldsymbol{\beta}$ as well. We use non-informative improper priors for the latter parameters because our main focus lies on estimating the network effect ρ.

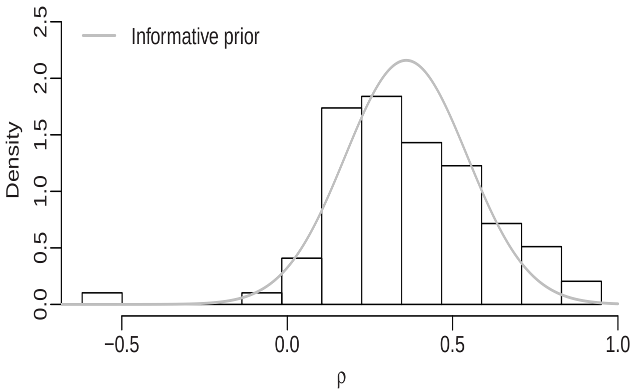

Figure 2.1 Histogram of the average network effects from each field $\{\overline{\rho}_j\}_j$, $\overline{\rho}_j := \sum_{i=1}^{n_j} \rho_{ij}/n_j$, and probability density function of the fitted normal population distribution for ρ.

Hence, the resulting empirical informative joint prior for $\boldsymbol{\theta}$ is

$$p_{\mathrm{EI}}(\boldsymbol{\theta}) = p_{\mathrm{EI}}(\rho) \times p_{\mathrm{F}}\left(\sigma^2\right) \times p_{\mathrm{F}}(\boldsymbol{\beta}) \propto N_\rho\left(.36, .19^2\right) \times 1/\sigma^2.$$

2.4.5 A weakly informative prior for ρ

There may be cases in which a researcher does not expect a network effect to be present in the data, or it may be that the researcher does not (want to) entertain the prior belief that the level of network autocorrelation in a data set is likely to fit with the empirical literature at large. In these cases, a researcher might purposely prefer to use less prior knowledge than actually available in the literature and rely on a so-called weakly informative prior distribution (Gelman et al., 2003). We construct such a weakly informative prior for ρ by imposing a normal prior that is centered around .36, as is the empirical informative prior, but with a deliberately much larger standard deviation, accounting for the uncertainty in one's prior beliefs. We set the weakly informative prior's standard deviation to .7, compared to .19 for the empirical informative prior, yielding a broad and fairly flat prior that still results in at least 62% of prior probability mass being contained in the unit interval $(0, 1)$. As for the empirical informative prior, we impose standard non-informative priors for $\left(\sigma^2, \boldsymbol{\beta}\right)$, assuming all parameters to be a priori independent. Thus,

$$p_{\mathrm{WI}}(\boldsymbol{\theta}) = p_{\mathrm{WI}}(\rho) \times p_{\mathrm{F}}\left(\sigma^2\right) \times p_{\mathrm{F}}(\boldsymbol{\beta}) \propto N_\rho\left(.36, .7^2\right) \times 1/\sigma^2.$$

Table 2.1 Characteristics of the studies used for the specification of the empirical informative prior for ρ.

	Study	Field	g	Type of W	Method	ρ
1	Andersson et al. (2010)	Property prices	1,034	Inverse distance	ML	.52
2	Anselin (1984)	House values	49	First-order contiguity	ML	.28
3	Anselin (1990)	Wage rates	25	First-order contiguity	ML	-.62
4	Anselin & Le Gallo (2006)	House prices	115,732	First-order contiguity	ML	.44
5	Anselin & Lozano-Gracia (2008)	House prices	103,867	First-order contiguity	2SLS	.33
6	Anselin et al. (2010)	House rents	1,671	First-order contiguity	HAC	.24
7	Anselin et al. (2000)	Innovation transfer	89	Distance-based contiguity	2SLS	.23
8.1	Arbia & Basile (2005)	GDP growth rates	92	First-order contiguity	ML	.33
8.2						.18
8.3						.34
9	Armstrong & Rodríguez (2006)	Property values	1,860	Inverse distance	ML	.36
10.1	Baller et al. (2001)	Homicide rates	1,412	Nearest neighbors	IV	.71
10.2						.65
10.3						.18
10.4						.23
11.1	Bernat Jr. (1996)	Economic growth	49	Squared inverse distance	ML	.35
11.2						.42
11.3						.70
12.1	Bivand & Szymanski (2000)	Garbage collection	324	First-order contiguity	ML	.15
12.2						.10
13	Bordignon et al. (2003)	Tax rates	143	First-order contiguity	ML	.16
14.1	Brueckner & Saavedra (2001)	Tax rates	70	Population weights	ML	.16
14.2						.04
14.3						.26
15.1	Buonanno et al. (2009)	Crime patterns	103	Inverse traveling distance	2SLS	-.54
15.2						.19
15.3						.21
16.1	Burt & Doreian (1982)	Scientific publishing	52	Structural equivalence	ML	.26
16.2						.21
16.3						.25
16.4						.45
16.5						.29
16.6						.31
16.7						.26
16.8						.54
17	Can (1992)	House prices	563	Squared inverse distance	ML	.41
18	Carruthers & Clark (2010)	House prices	28,165	Nearest neighbors	2SLS	.17
19.1	Chang (2008)	Water quality	94	First-order contiguity	ML	.19
19.2						.14
19.3						.49
19.4						.48
19.5						.56
19.6						.15
19.7						.42
19.8						.43
19.9						.37
19.10						.56
19.11						.44
19.12						.41
19.13						.55
19.14						.47
19.15						.36
19.16						.24
19.17						.35
19.18						.29
19.19						.25
19.20						.28
19.21						.24
19.22						.50
19.23						.42
19.24						.51
19.25						.47

Table 2.1 (continued).

	Study	Field	g	Type of W	Method	ρ
20	Cohen & Coughlin (2008)	House prices	508	Inverse distance	ML	.26
21	Conway et al. (2010)	House prices	260	First-order contiguity	ML	.11
22	Dall'erba (2005)	GDP growth rates	48	Most accessible neighbors	ML	.40
23	Doreian (1980)	Huk rebellion	57	First-order contiguity	ML	.47
24.1	Doreian (1980)	Voting behavior	64	First-order contiguity	ML	.61
24.2						.26
24.3	Doreian (1981)					.12
24.4						.29
24.5	Leenders (2002)					.31
25	Dow (2007)	Income contributions	158	Lexical distance	2SLS	.76
26	Easterly & Levine (1998)	GDP growth rates	234	Economic size	2SLS	.55
27	Elhorst (2014)	Crime rates	49	First-order contiguity	ML	.43
28	Ertur et al. (2007)	GDP growth rates	138	Nearest neighbors	ML	.75
29.1	Fingleton (2001)	Economic growth	178	Economic size&distance	3SLS	-.19
29.2						.56
29.3						.73
30	Fingleton et al. (2005)	Change in employment	408	Squared inverse distance	2SLS	.41
31	Fingleton & Le Gallo (2008)	House prices	353	Economic distance	ML	.72
32	Florax et al. (2002)	Agricultural yields	100	First-order contiguity	ML	.50
33	Ford & Rork (2010)	Patent rates	186	First-order contiguity	ML	.08
34	Fornango (2010)	Homicide rates	110	First-order contiguity	ML	.30
35	Gimpel & Schuknecht (2003)	Voting turnout	363	Distance-based contiguity	ML	.67
36.1	Gould (1991)	Battalion enlistment	20	Cross-district enlistment	ML	.29
36.2						.49
36.3						.49
37	Greenbaum (2002)	Teacher salaries	483	Inverse income difference	ML	.66
38	Heikkila & Kantiotou (1992)	Police expenditures	57	First-order contiguity	ML	.43
39.1	Heyndels & Vuchelen (1998)	Tax rates	589	First-order contiguity	3SLS	.67
39.2						.70
40	Holloway et al. (2002)	Crop adoption	406	First-order contiguity	Bayes	.54
41	Hunt et al. (2005)	Fishing trip prices	770	Inverse distance-based	ML	.80
42.1	Joines et al. (2003)	Hospitalization rates	100	First-order contiguity	ML	.53
42.2						.51
43	Kalenkoski & Lacombe (2008)	Youth employment	3,065	First-order contiguity	ML	.49
44.1	Kalnins (2003)	Fast food prices	1,385	Distance&contiguity-based	ML	.11
44.2						.21
45.1	Kim & Goldsmith (2009)	Property values	262	Nearest neighbors	2SLS	.22
45.2			523			.19
45.3			730			.14
46	Kim & Zhang (2005)	Land values	731	Nearest neighbors	ML	.39
47.1	Kirk & Papachristos (2011)	Homicide rates	342	First-order contiguity	ML	.43
47.2						.33
48.1	Land et al. (1991)	Church adherence	731	Inverse distance	2SLS	.33
48.2			697			.29
48.3			663			.28
49	Lauridsen et al. (2010)	Medical expenditures	400	Inverse distance	ML	.87
50	LeSage (1997b)	House values	88	First-order contiguity	ML	.45
51	Levine et al. (1995)	Road accidents	362	Squared inverse distance	ML	.22
52.1	Lin (2010)	GPA scores	68,131	Friendship	ML	.30
52.2			49,559			.29
52.3			79,067			.30
53	Lu & Zhang (2011)	Tree heights	3,982	Variogram	ML	.59
54	McMillen (2010)	Land ratios	1,322	First-order contiguity	ML	.71
55.1	McMillen et al. (2007)	Tuition fees	929	Distance&contiguity-based	ML	.22
55.2						.34
56.1	McPherson & Nieswiadomy (2005)	Species threat	113	Shared border length	ML	.23
56.2						.16
57	Moreno & Trehan (1997)	Worker output growth	89	Inverse distance	ML	.51
58.1	Morenoff (2003)	Birth weights	342	First-order contiguity	2SLS	.53
58.2						.69
59.1	Mur et al. (2008)	Purchasing power	1,274	Distance&contiguity-based	ML	.60
59.2						.61
60	Niebuhr (2010)	R&D spillovers	95	First-order contiguity	ML	.16

Table 2.1 (continued).

	Study	Field	g	Type of W	Method	ρ
61.1	Osland (2010)	Voting patterns	1,691	Nearest neighbors	ML	.07
61.2			766			.06
62	Patton & McErlean (2003)	Land prices	197	Squared inverse distance	IV	.66
63	Plümper & Neumayer (2010)	Tax rates	581	First-order contiguity	ML	.12
64.1	Pons-Novell & Viladecans-Marsal (1999)	GDP growth rates	74	First-order contiguity	ML	.23
64.2						.20
64.3						.17
65	Revelli (2003)	Expenditure levels	238	Contiguity-based	ML	.11
66	Ruggles (2007)	Co-residence behavior	276	Shared border length	ML	.15
67	Rupasingha et al. (2002)	Income growth	2,995	First-order contiguity	ML	.49
68.1	Saavedra (2000)	Welfare competition	47	First-order contiguity	ML	.28
68.2						.30
68.3						.32
69	Seldadyo et al. (2010)	Governance patterns	188	Nearest neighbors	ML	.28
70	Shin & Ward (1999)	Military spendings	95	Distance&contiguity-based	ML	.08
71.1	Tam Cho (2003)	Campaign donations	671	Inverse distance	2SLS	.06
71.2			455		ML	.04
71.3			657		ML	.03
71.4			1,183		ML	.03
71.5			1,420		2SLS	.03
71.6			2,072		2SLS	.03
71.7			1,821		2SLS	.03
71.8			2,288		2SLS	.02
71.9			2,206		2SLS	.03
71.10			291		ML	.07
71.11			229		ML	.06
71.12			249		ML	.06
71.13			273		ML	.05
71.14			458		2SLS	.05
71.15			502		2SLS	.05
71.16			698		2SLS	.05
71.17			606		2SLS	.04
71.18			660		2SLS	.05
71.19			752		2SLS	.03
71.20			401		2SLS	.00
71.21			613		2SLS	.02
71.22			581		2SLS	.02
71.23			324		ML	.05
71.24			918		ML	.01
71.25			760		2SLS	.03
71.26			701		ML	.06
71.27			980		2SLS	.05
71.28			874		ML	.07
72	Tita & Greenbaum (2009)	Gun violence	244	Gang rivalry	ML	.22
73	Varga (2000)	Technology innovation	125	Distance-based contiguity	IV	.14
74	Halleck Vega & Elhorst (2015)	Cigarette sales	1,380	First-order contiguity	ML	.20
75	Vitale et al. (2016)	Student performance	66	Personal advice	ML	.31
76	Voss & Chi (2006)	Population change	1,837	Nearest neighbors	ML	.27
77.1	Voss et al. (2006)	Child poverty	3,136	First-order contiguity	ML	.31
77.2						.27
78	Wilhelmsson (2002)	House prices	1,377	Inverse distance	ML	.95
79.1	Whitt (2010)	Crime rates	85	First-order contiguity	ML	.37
79.2						.58
79.3						.50
79.4						.54
80	Won Kim et al. (2003)	House prices	609	Distance&contiguity-based	2SLS	.55
81.1	N. Yang et al. (2012)	Wine prices	79	Nearest neighbors	ML	.33
81.2			876	Nearest neighbors		.34

Note: g = network size; 2SLS = two-stage least squares; 3SLS = three-stage least squares; HAC = Kelejian-Prucha heteroskedasticity and autocorrelation consistent estimator; IV = instrumental variables; ML = maximum likelihood.

2.4.6 Graphical prior comparison

In order to get more insight into the differences between the discussed priors, we inspected them graphically. We base our visualization on a randomly generated network of 20 actors with four covariates, including an intercept term. The shape of these priors is essentially the same for other data sets that are generated under different specifications of W and X (not shown).

Figure 2.2 shows the flat prior, the conditional Jeffreys rule prior, the conditional Independence Jeffreys prior, the empirical informative prior, and the weakly informative prior for ρ for the simulated data set. We fixed σ^2 to 1 and $\boldsymbol{\beta}$ to (1,1,1,1) for both versions of Jeffreys prior as the corresponding marginal priors for ρ are analytically not available. The graphs of the two versions of Jeffreys prior are "bathtub-shaped", contrary to the flat prior and the informative priors for ρ. In particular, $p_{\text{IJ}}\left(\rho|\sigma^2, \boldsymbol{\beta}\right)$ assigns substantial weight to values for ρ close to the boundaries of the admissible interval for ρ, while $p_{\text{J}}\left(\rho|\sigma^2, \boldsymbol{\beta}\right)$ does essentially the same but with slightly more weight for values for ρ close to the left boundary and less prior mass for values for ρ close to the right boundary.[7]

As the main analytical difference between Jeffreys rule prior and Independence Jeffreys prior is that for the latter the exponent of the error variance does not depend on the number of covariates, we also considered the bivariate conditional prior for $\left(\rho, \sigma^2|\boldsymbol{\beta} = (1,1,1,1)\right)$. In contrast to the conditional prior for ρ, $p_{\text{J}}\left(\rho, \sigma^2|\boldsymbol{\beta}\right)$ places more prior mass at boundary values of the two-dimensional parameter space $\left(\lambda_g^{-1}, \lambda_1^{-1}\right) \times (0, \infty)$, compared to $p_{\text{IJ}}\left(\rho, \sigma^2|\boldsymbol{\beta}\right)$ (not shown). Thus, we expect the Bayesian posterior estimates of ρ and σ^2 based on Jeffreys rule prior to tend more towards their respective boundary values compared to the estimates based on Independence Jeffreys prior.

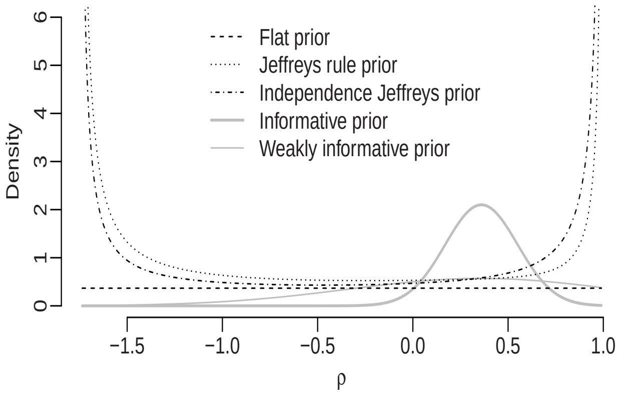

Figure 2.2 Conditional prior distributions for $\left(\rho|\sigma^2 = 1, \boldsymbol{\beta} = (1,1,1,1)\right)$ for simulated data.

[7] The (inverse of the) eigenvalues of the simulated network yield $(-1.75, 1)$ as the admissible interval for ρ as defined in Section 2.2.

2.5 Bayesian computation

In this section, we present an efficient algorithm for a Bayesian estimation of the network autocorrelation model. The methodology can be used to sample from the various arising posterior distributions based on the priors discussed in Section 2.4. As is common in Bayesian computation, the goal is to obtain a sample from the joint posterior for the unknown model parameters by sequentially drawing from the conditional posteriors, i.e., given the remaining parameters and the data (Gelfand & Smith, 1990; Geman & Geman, 1984). This is repeated until a sufficiently large sample is obtained.[8] We propose to sample the parameters according to the following blocks: $(\rho, \beta_1), \sigma^2$, and $\widetilde{\boldsymbol{\beta}}$, where β_1 denotes the model's intercept and $\widetilde{\boldsymbol{\beta}} = (\beta_2, ..., \beta_k)$ contains all the other regression coefficients. The reason for simultaneously sampling ρ and β_1 in one block is the high posterior correlation between these parameters.[9] Sampling these parameters separately would result in *slow mixing*, i.e., more draws would be needed to get both a good approximation of the posterior distribution and small estimation errors (Brooks, 1998; Gelman et al., 2003; Raftery & Lewis, 1996).

We illustrate the sampling algorithm when using the flat prior and the informative prior firsts, before discussing sampling from the more complex posteriors based on Jeffreys rule prior and Independence Jeffreys prior. For the former, the conditional posteriors are given by (see e.g., LeSage, 1997a)

$$p\left((\rho, \beta_1)\,|\sigma^2, \widetilde{\boldsymbol{\beta}}, \boldsymbol{y}\right) \propto |A_\rho| \exp\left(-\frac{1}{2\sigma^2}\left(A_\rho \boldsymbol{y} - X\boldsymbol{\beta}\right)^T\left(A_\rho \boldsymbol{y} - X\boldsymbol{\beta}\right)\right) p(\rho), \tag{2.8}$$

$$p\left(\sigma^2|\,(\rho, \beta_1), \widetilde{\boldsymbol{\beta}}, \boldsymbol{y}\right) \sim IG\left(\frac{g}{2}, \frac{\left(A_\rho \boldsymbol{y} - X\boldsymbol{\beta}\right)^T\left(A_\rho \boldsymbol{y} - X\boldsymbol{\beta}\right)}{2}\right), \tag{2.9}$$

$$p\left(\widetilde{\boldsymbol{\beta}}|\,(\rho, \beta_1), \sigma^2, \boldsymbol{y}\right) \sim N\left(\boldsymbol{\mu}_{\widetilde{\boldsymbol{\beta}}}, \Sigma_{\widetilde{\boldsymbol{\beta}}}\right), \tag{2.10}$$

where $IG(\cdot, \cdot)$ denotes the inverse gamma distribution and $\boldsymbol{\mu}_{\widetilde{\boldsymbol{\beta}}}$ and $\Sigma_{\widetilde{\boldsymbol{\beta}}}$ are given in Appendix 2.C.

Sampling from the inverse gamma distribution in (2.9) and the normal distribution in (2.10) is straightforward, whereas due to the appearance of the determinant in (2.8), the conditional posterior for (ρ, β_1) does not have a well-known form. In order to efficiently sample from this distribution, we rely on the *Metropolis-Hastings algorithm* (Hastings, 1970; Metropolis et al., 1953). In the algorithm, a candidate-generating distribution is chosen from which candidate values for the target distribution, here: the conditional posterior, are drawn. The specification of the candidate-generating distribution is crucial for the algorithm's efficiency, where we aim to construct a distribution that closely ap-

[8] This approach is needed as for none of the priors previously discussed the corresponding posterior belongs to a family of known probability distributions. Geman & Geman (1984) showed that sampling from the sequence of conditional posteriors for all parameters indeed produces estimates that converge in the limit to the true marginal posteriors for the parameters.

[9] This correlation is particularly pronounced for high levels of network density and we have not found this issue being discussed in the literature before. Only Hepple (1995b) provided a plot of the bivariate marginal posterior density $p_{\mathrm{F}}\left((\rho, \beta_1)\,|\boldsymbol{y}\right)$ for an empirical data set that clearly shows this dependence.

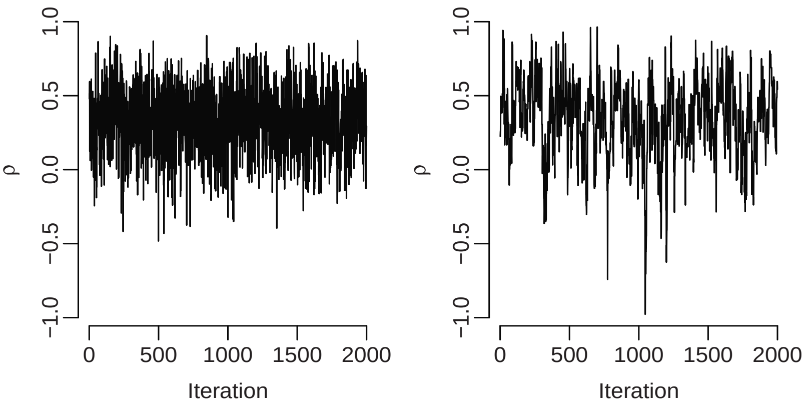

Figure 2.3 Trace plots of posterior draws for ρ for novel (left) and random walk scheme (right) for simulated data.

proximates the actual conditional posterior target distribution, which typically results in efficient solutions (Chib & Greenberg, 1995).

In (2.8), we approximate $\log(|A_\rho|)$ by a second-order Taylor polynomial at $\rho = 0$, which results in a normal approximation of the first factor, $|A_\rho|$. The second factor, $\exp(\cdot)$, if considered as a function of (ρ, β_1), has a bivariate normal kernel. The third factor, i.e., the marginal prior for ρ, is ignored in the candidate-generating distribution when using the flat prior and is normal for the informative priors. Hence, the overall product of these normal densities results in a bivariate normal candidate-generating distribution for (ρ, β_1) that incorporates the dependence between the two parameters and is tailored to the conditional posterior for (ρ, β_1).

Due to the complex prior expression for both Jeffreys rule and Independence Jeffreys prior, a Metropolis-Hastings step is needed to sample from all three conditional posteriors when employing these priors. For the first parameter block, (ρ, β_1), we use the same candidate-generating distribution as for the flat prior, as the prior information for (ρ, β_1) is quite vague compared to the likelihood. For the conditional posteriors for σ^2, we propose inverse gamma distributions as candidate-generating distributions but with different shape parameters than those used in (2.9), accounting for the different exponents of σ^2 in the two priors. Finally, we rely on the normal distribution in (2.10) as the candidate-generating distribution for the conditional posterior for $\widetilde{\boldsymbol{\beta}}$. All details and the full sampling schemes for all of the discussed priors can be found in Appendix 2.C.

We implemented our approach and compared its performance to existing sampling schemes that do not block (ρ, β_1) but build on a one-dimensional random walk algorithm to generate draws for ρ instead (Holloway et al., 2002; LeSage, 2000; LeSage & Pace, 2009). We found that our method produces well-mixed Markov chains with very low autocorrelations. Figure 2.3 displays sample trace plots of posterior draws for ρ based on our algorithm (left panel) and based on existing schemes (right panel) when using the flat prior. As can be seen, our algorithm generates Markov chains that are moving quicker and explore the parameter space much faster compared to traditional methods.[10]

2.6 Simulation study

We performed a thorough simulation study to examine properties of the Bayesian estimators based on the flat prior, Jeffreys rule prior, Independence Jeffreys prior, the empirical informative prior, and the weakly informative prior, and compared them to those based on maximum likelihood estimation. The main focus in this study was to evaluate the bias of ρ and the frequentist coverage of credible and confidence intervals for ρ for the different estimators, i.e., the extent to which a true data-generating network effect is contained in those intervals. Furthermore, as the most likely outcome of the negative bias in the estimation of ρ is a Type II error, we considered the average of Type I and Type II errors as well.[11] Such average error rates are increasingly used as optimal decision criteria instead of the prevailing paradigm, which is fixing Type I error probability and then minimizing Type II error probability (Chance & Rossman, 2006; DeGroot & Schervish, 2010; Pericchi & Pereira, 2016). Lastly, we also investigated the bias in the estimation of σ^2, as it is known that Jeffreys rule prior can result in poor estimates of the error variance in multi-parameter models (De Oliveira, 2012; De Oliveira & Song, 2008).

2.6.1 Study design

In our study design, we largely followed setups from previous simulation studies of the network autocorrelation model (Mizruchi & Neuman, 2008; Neuman & Mizruchi, 2010; W. Wang et al., 2014). Hence, we generated data $\boldsymbol{y}$ by using random networks and varying the size of the network, the density of the network, the number of covariates, and the magnitude of ρ. We did so by $\boldsymbol{y} = (I_g - \rho W)^{-1}(X\boldsymbol{\beta} + \boldsymbol{\varepsilon})$.[12] We considered three network sizes ($g \in \{10, 20, 50\}$), three levels of network density ($d \in \{.1, .3, .5\}$), two sets of covariates plus an intercept term ($k \in \{4, 7\}$), and three fixed network effect sizes ($\rho \in \{0, .2, .5\}$).[13] We obtained random binary symmetric connectivity matrices with zeros in the diagonal entries by relying on the rgraph() function from the sna package in R (Butts, 2008) and subsequently row-normalized the raw connectivity matrices. Moreover, we drew independent values from a standard normal distribution for the elements of X (excluding the first column of X which is a vector of ones), $\boldsymbol{\beta}$, and $\boldsymbol{\varepsilon}$, so $\sigma^2 = 1$. In addition to simulating data using a fixed network effect ρ, we also allowed for fluctuations in the underlying network effects by sampling them from the estimated population distribu-

[10] Also note that there are no parameters to be tuned in the Metropolis-Hastings steps in our approach, such as the variances of candidate-generating distributions. This stands in stark contrast to existing schemes, where this is commonly done in order to achieve specific acceptance rates.

[11] We thank an anonymous reviewer for this suggestion.

[12] For all simulated data sets we looked at, none of the regularity conditions needed for posterior propriety was violated, and it seems highly unlikely to encounter such a situation in empirical practice.

[13] Simulation results for negative values for ρ and different specifications of W are available from the authors upon request. We do not present them here, as the analyses provide no additional, i.e., different, insights.

tion from Section 2.4.4. As the true network autocorrelation is unknown in practice, this appears to be a much more realistic simulation setup compared to setting ρ to a specific value a priori.[14] In total, we considered 60 scenarios (1 network size × 3 network densities × 1 set of covariates × 4 sampling schemes for ρ and 2 network sizes × 3 network densities × 2 sets of covariates × 4 sampling schemes for ρ) and simulated 500 data sets for each scenario.

For the Bayesian estimators, we drew 10,000 samples from the corresponding posteriors, applying the sampling schemes described in Section 2.5. We used the marginal posterior median as point estimator and 95% equal-tailed credible intervals by discarding the 2.5% smallest and largest draws, respectively, for coverage analyses of ρ and σ^2. In contrast to that, we employed asymptotic standard errors based on the model's observed Fisher information matrix to obtain maximum likelihood-based confidence intervals for ρ and σ^2.[15]

2.6.2 Simulation results

Table 2.2 shows the average bias of ρ for the different estimators. Overall, the Bayesian estimators based on the non-informative priors yield similar results to those based on maximum likelihood estimation. In particular, there is still some negative bias present, which is a well-known issue in the network autocorrelation model. On the other hand, if the true underlying ρ equals zero, the Bayesian estimator based on the weakly informative prior eliminates virtually all the negative bias in the estimation of ρ. Furthermore, when the data-generating network effect is positive, using the empirical informative prior generally results in the least absolute bias of ρ. Given our review of empirically observed network autocorrelations in Section 2.4.4, this is clearly the most common situation to be encountered in practice. Lastly, we also observe a much smaller increase in bias for higher levels of network density for this estimator, compared to the non-informative Bayesian ones and the maximum likelihood estimator.

Table 2.3 shows the empirical frequentist coverage of equal-tailed 95% credible intervals for ρ for the Bayesian estimators and the coverage of asymptotic 95% confidence intervals for ρ. The coverage of credible intervals based on the flat prior and Independence Jeffreys prior is very close to the nominal .95. In contrast to that, the coverage of credible intervals corresponding to Jeffreys rule prior and the coverage of maximum likelihood-based confidence intervals are below nominal for all considered scenarios. The problem of subpar coverage of maximum likelihood-based confidence intervals for ρ is completely resolved

[14] In fact, we sampled ρ from the estimated population distribution truncated to $(-1, 1)$ to ensure that the generated network effects always lied in the chosen admissible interval $\left(\lambda_g^{-1}, 1\right)$. Note that less than 0.1% of probability mass of the estimated population distribution actually falls outside $(-1, 1)$. For each draw for ρ from this estimated population distribution, we recorded the drawn value for ρ (which was the true value for ρ for that particular draw) and base our simulation results on those recorded underlying network effects.

[15] All computation was performed in R using self-written routines. We used maximum likelihood estimates as starting values for the MCMC procedures and discarded the first 1,000 iterations as so-called *burn-in* values (Gelman et al., 2003). As most of the marginal posterior distributions for ρ and σ^2 were skewed, we opted for the posterior median as Bayesian point estimator, which was a less extreme estimator than the posterior mean or the posterior mode.

when using Bayesian estimators based on the flat prior or Independence Jeffreys prior.

Table 2.4 reports the average empirical Type I and Type II error rates of ρ for the different estimators. In general, the average error rates increase with the network density due to the negative bias in the estimation of ρ, and they decrease for higher network autocorrelations as a result of higher power. For all considered scenarios, the Bayesian estimator based on the empirical informative prior clearly yields the smallest average Type I and Type II error rates across the board. The other estimators perform relatively similar to each other, with the maximum likelihood estimator having slightly smaller average error rates than the remaining Bayesian ones but still considerably higher than the estimator based on the empirical informative prior. The greater power of the maximum likelihood estimator, compared to the Bayesian estimators based on the non-informative priors, comes at the price of underestimating the standard error of ρ. In turn, this results in narrower confidence intervals for ρ, leading to lower coverage but slightly higher power. Regardless, estimating ρ using the empirical informative prior yields the lowest average Type I and Type II error rates.

Table 2.5 displays the average bias of σ^2 for the Bayesian estimators and the maximum likelihood estimator. The estimates of σ^2 corresponding to the use of the flat prior, Independence Jeffreys prior, and the informative priors are nearly unbiased, while the results based on Jeffreys rule prior and maximum likelihood estimation exhibit a large negative bias. This bias is particularly pronounced for a higher number of covariates. We also investigated the associated coverage of Bayesian equal-tailed 95% credible intervals and asymptotic 95% confidence intervals for σ^2. In line with the results for the average bias of σ^2, we found that the coverage of credible intervals based on the flat prior, Independence Jeffreys prior, and the informative priors is very close to the nominal .95. On the other hand, the coverage of credible intervals corresponding to Jeffreys rule prior and the coverage of maximum likelihood-based confidence intervals are well below the nominal rate for all considered scenarios. These results are not shown here but are available from the authors upon request.

Based on our simulation output, we suggest the following: first, if a researcher is willing to expect that his, or her, study might have a network effect along the lines of the overall distribution of network autocorrelation effects across the literature at large, using the empirical informative prior is highly recommended as it leads to a dramatic decrease of the bias in the estimation of ρ. Furthermore, the corresponding estimator exhibits by far the smallest average Type I and Type II error rates of ρ and accurately estimates σ^2. At the same time, applying the empirical informative prior can result in a mild overestimation of ρ for small positive network effects. However, we believe this to be less of a concern than falsely dismissing positive network effects and stress that overall, this estimator performs clearly the best.

Second, if a researcher does not expect a network effect to be present in the data, or if the researcher does not (want to) entertain the prior belief that the level of network autocorrelation in a data set is likely to fit with the extant empirical empirical literature, relying on the weakly informative prior yields nearly adequate point estimates of the network effect in these cases. This does, however, require the researcher to sacrifice the Type I and Type II error rate reducing benefits of the empirical informative prior.

Third, if a researcher prefers to refrain from employing any empirical-based prior information, we recommend using the non-informative Independence Jeffreys prior. While this does not attenuate the negative bias in the estimation of ρ, the issue of poor coverage of confidence intervals, associated with maximum likelihood estimation of the model, can be completely eluded at least. We wish to emphasize that there is never a case in which maximum likelihood estimation can be recommended.

Lastly, when analyzing a real data set, we advise researchers to estimate the model using all three recommended priors. If the resulting estimates of ρ are close to each other, this implies that the data contain sufficient information and the estimates are most likely highly reliable; else, this strongly points at (negative) bias in the estimation of the network effect.

2.7 Conclusions

In this chapter, we derived two versions of Jeffreys prior for the network autocorrelation model that provide default Bayesian analyses of the model. Moreover, we specified an empirical informative prior and a weakly informative prior for the network effect ρ based on reported network effects from the literature.

We evaluated the Bayesian estimators by means of a simulation study and compared their performance to the performance of the maximum likelihood estimator. We found that the Bayesian estimator based on the empirical informative prior performs superior and that the estimator based on the weakly informative prior can be a useful alternative. Concomitantly, we also provided a very efficient MCMC implementation of the Bayesian approach that is preferable to existing sampling schemes and ensures a fast and accurate Bayesian estimation of the network autocorrelation model.

In order to allow researchers and practitioners to easily use the newly developed methods in this chapter, it is essential to make them accessible in a statistical software package. In addition, as we primarily focused on Bayesian point estimation in this work, further work needs to be done in studying Bayesian model selection procedures for the discussed priors. Finally, despite the improved numerical properties of the Bayesian estimators, the negative bias of ρ in the model is not entirely resolved. We did resolve much of the bias for data sets that are typical in the empirical literature at large, but more research is needed to untangle it completely. It remains a major challenge to investigate what causes this negative bias and why it becomes increasingly salient at high levels of network density.

Table 2.2 Average bias of ρ based on using the flat prior (F), Jeffreys rule prior (J), Independence Jeffreys prior (IJ), the empirical informative prior (EI), the weakly informative prior (WI), and the maximum likelihood estimator (ML) for 500 simulated data sets. For each scenario, the smallest absolute bias is printed in bold face.

$g = 10, k = 4$	$\rho = 0$						$\rho = .2$						$\rho = .5$						$\rho \sim N\left(.36, .19^2\right)(-1,1)$					
d	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML
.1	-.028	-.046	-.034	.240	**.027**	-.041	-.103	-.074	-.085	.076	**-.064**	-.079	-.151	**-.068**	-.104	-.109	-.136	-.082	-.124	-.075	-.095	**-.020**	-.099	-.083
.3	-.088	-.115	-.093	.247	**-.008**	-.110	-.158	-.137	-.141	**.087**	-.095	-.140	-.242	-.145	-.197	**-.139**	-.210	-.161	-.223	-.173	-.195	**-.056**	-.175	-.180
.5	-.172	-.203	-.158	.282	**-.015**	-.204	-.253	-.220	-.216	**.106**	-.125	-.235	-.453	-.367	-.405	**-.169**	-.343	-.391	-.364	-.296	-.319	**-.051**	-.248	-.318
$g = 20, k = 4$	$\rho = 0$						$\rho = .2$						$\rho = .5$						$\rho \sim N\left(.36, .19^2\right)(-1,1)$					
d	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML
.1	-.023	-.026	-.024	.143	**-.001**	-.025	-.055	**-.039**	-.047	.046	-.040	-.042	-.085	**-.047**	-.067	-.088	-.085	-.054	-.059	-.032	-.046	**-.014**	-.052	-.038
.3	-.082	-.080	-.072	.234	**-.009**	-.085	-.151	-.125	-.131	**.076**	-.096	-.136	-.228	-.156	-.192	**-.141**	-.202	-.181	-.198	-.144	-.173	**-.052**	-.159	-.164
.5	-.227	-.205	-.196	.268	**-.054**	-.229	-.307	-.254	-.265	**.093**	-.163	-.287	-.404	-.288	-.353	**-.162**	-.301	-.340	-.370	-.274	-.321	**-.066**	-.257	-.320
$g = 20, k = 7$	$\rho = 0$						$\rho = .2$						$\rho = .5$						$\rho \sim N\left(.36, .19^2\right)(-1,1)$					
d	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML
.1	-.018	-.021	-.018	.132	**.001**	-.020	-.048	**-.031**	-.041	.034	-.037	-.033	-.077	**-.038**	-.063	-.081	-.078	-.042	-.062	-.032	-.051	**-.026**	-.057	-.035
.3	-.057	-.057	-.052	.213	**-.002**	-.059	-.094	-.067	-.080	.078	**-.056**	-.074	-.172	**-.099**	-.143	-.126	-.158	-.116	-.152	-.101	-.129	**-.033**	-.122	-.113
.5	-.121	-.098	-.096	.272	**.003**	-.115	-.227	-.181	-.191	**.094**	-.121	-.202	-.321	-.203	-.269	**-.155**	-.261	-.242	-.274	-.185	-.232	**-.044**	-.195	-.216
$g = 50, k = 4$	$\rho = 0$						$\rho = .2$						$\rho = .5$						$\rho \sim N\left(.36, .19^2\right)(-1,1)$					
d	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML
.1	-.011	-.011	-.010	.125	**.006**	-.011	-.043	-.035	-.039	.034	**-.033**	-.037	-.051	**-.030**	-.043	-.068	-.052	-.036	-.047	-.032	-.041	**-.019**	-.043	-.037
.3	-.093	-.078	-.082	.225	**-.020**	-.089	-.140	-.106	-.121	**.077**	-.088	-.125	-.215	-.145	-.191	**-.140**	-.193	-.181	-.175	-.117	-.152	**-.043**	-.140	-.147
.5	-.274	-.224	-.246	.262	**-.080**	-.261	-.297	-.216	-.262	**.095**	-.154	-.265	-.437	-.319	-.398	**-.168**	-.332	-.383	-.343	-.236	-.303	**-.051**	-.233	-.294
$g = 50, k = 7$	$\rho = 0$						$\rho = .2$						$\rho = .5$						$\rho \sim N\left(.36, .19^2\right)(-1,1)$					
d	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML
.1	-.010	-.011	-.033	.097	**.001**	-.011	-.023	**-.016**	-.020	.032	-.017	-.017	-.047	**-.029**	-.042	-.063	-.049	-.033	-.043	-.031	-.039	**-.024**	-.041	-.033
.3	-.067	-.060	-.061	.197	**-.015**	-.065	-.084	-.060	-.072	.074	**-.052**	-.071	-.144	**-.093**	-.124	-.122	-.138	-.115	-.117	-.075	-.100	**-.033**	-.099	-.094
.5	-.165	-.137	-.147	.248	**-.044**	-.157	-.195	-.137	-.164	**.089**	-.110	-.169	-.260	-.157	-.225	**-.145**	-.220	-.206	-.270	-.188	-.240	**-.059**	-.204	-.229

Table 2.3 Empirical frequentist coverage of 95% credible and confidence intervals for ρ based on using the flat prior (F), Jeffreys rule prior (J), Independence Jeffreys prior (IJ), the empirical informative prior (EI), the weakly informative prior (WI), and the maximum likelihood estimator (ML) for 500 simulated data sets. For each scenario, the most accurate coverage is printed in bold face.

$g = 10, k = 4$	$\rho = 0$						$\rho = .2$						$\rho = .5$						$\rho \sim N\left(.36, .19^2\right)(-1, 1)$					
d	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML
.1	.976	.860	**.966**	.842	.984	.800	.978	.878	**.964**	.992	.988	.800	.942	.836	.940	.966	**.946**	.776	.946	.868	.946	.922	**.950**	.820
.3	.976	.838	**.956**	.880	.980	.784	**.956**	.832	.942	.994	.978	.750	**.948**	.870	**.948**	.988	.960	.802	.936	.834	.930	.934	**.952**	.776
.5	.976	.854	**.954**	.902	.992	.800	.974	.878	**.952**	.996	.988	.816	.926	.856	.928	.990	**.956**	.806	.936	.872	.940	.914	**.958**	.796
$g = 20, k = 4$	$\rho = 0$						$\rho = .2$						$\rho = .5$						$\rho \sim N\left(.36, .19^2\right)(-1, 1)$					
d	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML
.1	.962	.918	**.952**	.876	.970	.900	**.944**	.906	.934	.984	**.956**	.896	.920	.906	.926	**.944**	.924	.896	.946	.904	.938	.932	**.950**	.864
.3	.962	.914	**.950**	.862	.976	.892	.964	.918	**.956**	1	.980	.894	.926	.906	.930	.972	**.942**	.876	.934	.918	.940	.904	**.952**	.892
.5	.970	.920	**.964**	.930	.994	.886	**.942**	.916	**.942**	.998	.986	.874	.922	.926	.932	1	**.964**	.888	.926	.928	.934	.934	**.956**	.896
$g = 20, k = 7$	$\rho = 0$						$\rho = .2$						$\rho = .5$						$\rho \sim N\left(.36, .19^2\right)(-1, 1)$					
d	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML
.1	.964	.874	**.954**	.900	.978	.852	**.952**	.874	.946	.986	.968	.844	.932	.876	**.942**	.932	.932	.850	.960	.872	**.954**	**.954**	.960	.850
.3	.962	.874	**.954**	.870	.980	.842	.978	.914	**.970**	.994	.984	.890	.946	.876	**.948**	.986	.954	.890	.928	.880	.932	.924	**.946**	.850
.5	.978	.886	**.972**	.870	.998	.856	.954	.882	**.948**	.998	.978	.848	.932	.906	**.942**	.992	.962	.848	**.950**	.904	.948	.932	.964	.858
$g = 50, k = 4$	$\rho = 0$						$\rho = .2$						$\rho = .5$						$\rho \sim N\left(.36, .19^2\right)(-1, 1)$					
d	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML
.1	.960	**.948**	.958	.896	.968	.944	.940	.930	.934	.986	**.946**	.928	.956	**.946**	.956	.956	.956	.936	.944	.932	.928	.944	**.946**	.924
.3	**.948**	.930	.942	.856	.970	.924	**.948**	.942	.944	.998	.984	.932	.926	.942	.936	.982	**.946**	.928	.944	**.946**	.938	.940	.966	.924
.5	**.930**	.916	.928	.916	.988	.898	.946	.940	**.950**	1	.978	.922	.892	.922	.904	.998	**.932**	.904	.916	.942	.924	.922	**.946**	.932
$g = 50, k = 7$	$\rho = 0$						$\rho = .2$						$\rho = .5$						$\rho \sim N\left(.36, .19^2\right)(-1, 1)$					
d	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML
.1	.934	.910	.928	.882	**.946**	.902	.962	.932	**.956**	.972	.962	.918	.954	.930	**.948**	.944	.954	.926	.946	.938	**.948**	.940	.942	.924
.3	.958	.922	**.948**	.856	.976	.906	.954	.936	**.952**	.996	.974	.922	.948	.938	**.950**	.982	.954	.926	**.950**	.930	**.950**	.954	.958	.916
.5	.954	.932	**.952**	.886	.990	.918	**.964**	.904	.966	1	.984	.928	**.954**	.962	.958	.988	.972	.942	.918	.928	.928	.916	**.938**	.916

Table 2.4 Average of empirical Type I and Type II error rates of ρ resulting from 95% credible and confidence intervals for ρ based on using the flat prior (F), Jeffreys rule prior (J), Independence Jeffreys prior (IJ), the empirical informative prior (EI), the weakly informative prior (WI), and the maximum likelihood estimator (ML) for 500 simulated data sets. For the scenarios where the data-generating value for ρ is zero, the empirical Type I error rate equals the proportion of times in which zero was not contained in the credible or confidence interval. For the scenarios where the data-generating value for ρ is not zero, the Type II error rate equals the proportion of times in which zero was contained in the credible or confidence interval. For each scenario, the smallest average error rate is printed in bold face.

$g = 10, k = 4$	$\rho = .2$						$\rho = .5$						$\rho \sim N\left(.36, .19^2\right)(-1, 1)$					
d	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML
.1	.488	.482	.482	**.413**	.478	.486	.364	.273	.323	**.204**	.342	.269	.400	.344	.379	**.284**	.389	.343
.3	.494	.497	.497	**.422**	.497	.477	.454	.392	.427	**.252**	.442	.374	.461	.430	.449	**.320**	.450	.404
.5	.505	.522	.510	**.452**	.500	.522	.502	.505	.505	**.396**	.496	.495	.499	.502	.500	**.402**	.491	.474
$g = 20, k = 4$	$\rho = .2$						$\rho = .5$						$\rho \sim N\left(.36, .19^2\right)(-1, 1)$					
d	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML
.1	.452	.431	.446	**.321**	.440	.417	.164	.147	.152	**.092**	.149	.141	.290	.263	.278	**.200**	.280	.258
.3	.501	.506	.504	**.421**	.498	.501	.426	.384	.412	**.252**	.411	.374	.490	.426	.450	**.318**	.443	.419
.5	.509	.503	.506	**.452**	.499	.507	.498	.481	.486	**.346**	.482	.470	.501	.500	.495	**.387**	.491	.494
$g = 20, k = 7$	$\rho = .2$						$\rho = .5$						$\rho \sim N\left(.36, .19^2\right)(-1, 1)$					
d	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML
.1	.427	.390	.420	**.293**	.410	.385	.141	.115	.124	**.073**	.120	.122	.271	.223	.262	**.173**	.256	.221
.3	.497	.487	.492	**.380**	.481	.489	.383	.316	.358	**.185**	.354	.302	.445	.391	.431	**.306**	.430	.385
.5	.504	.501	.501	**.448**	.499	.499	.486	.447	.478	**.323**	.472	.434	.493	.470	.483	**.372**	.479	.459
$g = 50, k = 4$	$\rho = .2$						$\rho = .5$						$\rho \sim N\left(.36, .19^2\right)(-1, 1)$					
d	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML
.1	.412	.393	.403	**.295**	.396	.388	.098	.093	.095	**.065**	.088	.087	.237	.223	.236	**.172**	.229	.219
.3	.506	.497	.498	**.423**	.494	.496	.416	.406	.413	**.248**	.401	.400	.452	.436	.447	**.299**	.437	.421
.5	.516	.516	.515	**.451**	.496	.516	.512	.498	.506	**.353**	.482	.494	.515	.496	.510	**.381**	.485	.484
$g = 50, k = 7$	$\rho = .2$						$\rho = .5$						$\rho \sim N\left(.36, .19^2\right)(-1, 1)$					
d	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML
.1	.351	.335	.351	**.242**	.338	.326	.073	.066	.071	.064	**.058**	.069	.193	.187	.190	**.160**	.180	.188
.3	.486	.472	.479	**.360**	.471	.470	.350	.312	.338	**.167**	.323	.307	.404	.380	.398	**.265**	.384	.381
.5	.509	.501	.502	**.431**	.490	.504	.481	.435	.464	**.269**	.451	.425	.490	.469	.490	**.361**	.475	.468

Table 2.5 Average bias of σ^2 based on using the flat prior (F), Jeffreys rule prior (J), Independence Jeffreys prior (IJ), the empirical informative prior (EI), the weakly informative prior (WI), and the maximum likelihood estimator (ML) for 500 simulated data sets. For each scenario, the smallest absolute bias is printed in bold face.

$g = 10, k = 4$	$\rho = 0$						$\rho = .2$						$\rho = .5$						$\rho \sim N\left(.36, .19^2\right)(-1,1)$					
d	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML
.1	.072	-.440	**-.008**	.220	.078	-.492	.122	-.416	**.034**	.139	.112	-.470	.192	-.413	**.054**	.187	.183	-.467	.099	-.452	**-.018**	.139	.095	-.501
.3	**.021**	-.466	-.064	.211	.036	-.516	.067	-.446	**-.032**	.126	.065	-.497	.147	-.423	**.010**	.132	.134	-.476	.083	-.453	**-.044**	.130	.081	-.504
.5	**.007**	-.469	-.079	.189	.034	-.518	**.033**	-.453	-.062	.107	.040	-.503	.098	-.423	**-.028**	.111	.092	-.475	.081	-.436	**-.037**	.126	.080	-.488
$g = 20, k = 4$	$\rho = 0$						$\rho = .2$						$\rho = .5$						$\rho \sim N\left(.36, .19^2\right)(-1,1)$					
d	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML
.1	-.006	-.238	-.043	.041	**-.005**	-.274	.017	-.223	-.023	.016	**.014**	-.259	.040	-.218	**-.011**	.044	.039	-.253	.039	-.214	**-.008**	.049	.038	-.250
.3	.021	-.213	**-.018**	.089	.026	-.250	.021	-.213	**-.019**	.041	.020	-.251	.037	-.208	**-.012**	.038	.034	-.244	.033	-.209	**-.014**	.047	.032	-.245
.5	-.018	-.240	-.055	.064	**-.004**	-.276	**.014**	-.216	-.028	.055	.020	-.253	.047	-.194	**-.006**	.051	.044	-.231	.016	-.216	-.031	.034	**.015**	-.252
$g = 20, k = 7$	$\rho = 0$						$\rho = .2$						$\rho = .5$						$\rho \sim N\left(.36, .19^2\right)(-1,1)$					
d	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML
.1	.037	-.372	**-.016**	.094	.037	-.400	.043	-.371	**-.012**	.037	.040	-.399	.049	-.380	**-.018**	.050	.049	-.407	.088	-.354	**.022**	.105	.088	-.382
.3	**.021**	-.379	-.032	.097	.023	-.408	.056	-.356	**.001**	.063	.052	-.386	.029	-.382	-.033	.025	**.024**	-.409	.041	-.371	**-.019**	.049	.038	-.399
.5	**.005**	-.385	-.047	.075	.011	-.413	**.014**	-.379	-.039	.049	.015	-.407	.054	-.360	**-.009**	.047	.047	-.388	.047	-.362	**-.014**	.064	.043	-.391
$g = 50, k = 4$	$\rho = 0$						$\rho = .2$						$\rho = .5$						$\rho \sim N\left(.36, .19^2\right)(-1,1)$					
d	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML
.1	**.002**	-.090	-.013	.016	**.002**	-.107	.002	-.090	-.014	**.000**	-.001	-.107	-.008	-.103	-.026	**-.004**	-.008	-.118	**.003**	-.092	-.014	.009	**.003**	-.108
.3	-.004	-.093	-.019	.021	**-.002**	-.110	.011	-.080	-.006	.019	**-.002**	-.097	.022	-.072	**.003**	.021	.021	-.088	**.009**	-.084	**-.009**	.015	**.009**	-.100
.5	-.012	-.099	-.028	.022	**-.006**	-.117	**-.005**	-.093	-.022	.009	-.007	-.110	.015	-.076	**-.004**	.020	.015	-.093	**.002**	-.088	-.015	.010	.003	-.104
$g = 50, k = 7$	$\rho = 0$						$\rho = .2$						$\rho = .5$						$\rho \sim N\left(.36, .19^2\right)(-1,1)$					
d	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML	F	J	IJ	EI	WI	ML
.1	.017	-.141	**-.010**	.032	.018	-.156	**.000**	-.156	-.018	-.003	-.001	-.170	.016	-.145	**-.004**	.019	.016	-.159	**.007**	-.151	-.012	.011	**.007**	-.166
.3	**.004**	-.150	-.014	.028	.005	-.165	**-.001**	-.155	-.020	.004	-.002	-.170	.028	-.132	**.008**	.027	.027	-.147	.012	-.144	**-.007**	.015	.011	-.159
.5	-.007	-.158	-.024	.021	**-.005**	-.173	-.006	-.157	-.024	**.003**	-.007	-.172	.022	-.135	**.001**	.020	.020	-.150	.024	-.132	**.005**	.028	.023	-.148

Appendix 2.A Auxiliary facts

(1) Let $A_\rho = I_g - \rho W$. If $\left(X^T X\right)^{-1}$ exists, then

$$\begin{aligned}
&\left(A_\rho \boldsymbol{y} - X\boldsymbol{\beta}\right)^T \left(A_\rho \boldsymbol{y} - X\boldsymbol{\beta}\right) \\
&= \left(A_\rho \boldsymbol{y} - X\hat{\boldsymbol{\beta}}\right)^T \left(A_\rho \boldsymbol{y} - X\hat{\boldsymbol{\beta}}\right) + \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)^T \left(X^T X\right) \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right) \\
&= \boldsymbol{y}^T A_\rho^T M A_\rho \boldsymbol{y} + \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)^T \left(X^T X\right) \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right),
\end{aligned}$$

where $\hat{\boldsymbol{\beta}} := \left(X^T X\right)^{-1} X^T A_\rho \boldsymbol{y}$ and $M = I_g - X\left(X^T X\right)^{-1} X^T$. Note that M is symmetric and idempotent.

(2) Let A be an invertible matrix, and let $\boldsymbol{u}$ and $\boldsymbol{v}$ be two vectors. Then, $A + \boldsymbol{u}\boldsymbol{v}^T$ is invertible and (see e.g., Ding & Zhou, 2007, Lemma 1.1)

$$\det\left(A + \boldsymbol{u}\boldsymbol{v}^T\right) = \left(1 + \boldsymbol{v}^T A^{-1} \boldsymbol{u}\right) \det\left(A\right).$$

(3) Let $\boldsymbol{Z} \sim N\left(\boldsymbol{\mu}, \Sigma\right)$ and let A be a symmetric matrix. Then (see e.g., Mathai & Provost, 1992, Corollary 3.2b.1),

$$\mathbb{E}\left[\boldsymbol{Z}^T A \boldsymbol{Z}\right] = \operatorname{tr}\left(A\Sigma\right) + \boldsymbol{\mu}^T A \boldsymbol{\mu}.$$

(4) Let A and B be symmetric and positive semi-definite matrices. Then (see e.g., X. Yang, 2000, Lemma 1),

$$0 \leq \operatorname{tr}\left(AB\right) \leq \operatorname{tr}\left(A\right) \operatorname{tr}\left(B\right).$$

(5) Let A and B be matrices. Then,

$$\operatorname{tr}\left(AB\right) \leq \frac{1}{2}\left(\operatorname{tr}\left(A^2\right) + \operatorname{tr}\left(B^2\right)\right).$$

(6) Let A be an idempotent matrix. Then, the eigenvalues of A are either 0 or 1 (see e.g., Harville, 1997, Theorem 21.8.2).

Appendix 2.B Proofs

Proof of Corollary 2.1

(i) Follows immediately from the prior's definition and $\Theta = \left(\lambda_g^{-1}, \lambda_1^{-1}\right) \times (0, \infty) \times \mathbb{R}^k$.

(ii) Using auxiliary fact (1), Hepple (1995a) showed that if $\left(X^T X\right)^{-1}$ exists and $g > k$, the corresponding marginal posterior for ρ is

$$p_{\mathrm{F}}\left(\rho|\boldsymbol{y}\right) \propto |A_\rho| \boldsymbol{y}^T A_\rho^T M A_\rho \boldsymbol{y}^{-\frac{g-k}{2}}.$$

As $|A_\rho| \leq 1$ and the assumption that $\left(\boldsymbol{y}^T MW\boldsymbol{y}\right)^2 \neq \boldsymbol{y}^T W^T MW\boldsymbol{y}\boldsymbol{y}^T M\boldsymbol{y}$ ensures that $\boldsymbol{y}^T A_\rho^T MA_\rho\boldsymbol{y} > 0$, $p_{\mathrm{F}}\left(\rho|\boldsymbol{y}\right)$ is bounded on $\left(\lambda_g^{-1}, \lambda_1^{-1}\right)$, which proves the statement.

Remark: As for $\rho \to \infty$, $|A_\rho| = \mathcal{O}\left(\rho^{g-m_0}\right)$ (where $m_0 \geq 0$ denotes the algebraic multiplicity of a potential zero eigenvalue of W) and $\boldsymbol{y}^T A_\rho^T MA_\rho\boldsymbol{y}^{-(g-k)/2} = \mathcal{O}\left(\rho^{k-g}\right)$, it follows that $p_{\mathrm{F}}\left(\rho|\boldsymbol{y}\right) = \mathcal{O}\left(\rho^{k-m_0}\right)$. Hence, the marginal posterior for ρ is integrable on $\mathbb{R} \setminus \left\{{\lambda_1}^{-1}, {\lambda_2}^{-1}, ..., {\lambda_g}^{-1}\right\}$ only if $k < m_0 - 1$, which is typically not the case.

Proof of Theorem 2.1

The model's Fisher information Matrix for $\boldsymbol{\theta} = \left(\rho, \sigma^2, \boldsymbol{\beta}\right)$ is (see e.g., Doreian, 1981)

$$\begin{aligned}
I\left(\boldsymbol{\theta}\right) &= \frac{1}{\sigma^2} I^*\left(\boldsymbol{\theta}\right) \\
&= \frac{1}{\sigma^2}\begin{bmatrix} \sigma^2\left(\mathrm{tr}\left(B_\rho^T B_\rho\right) + \mathrm{tr}\left(B_\rho^2\right)\right) + \boldsymbol{\beta}^T X^T B_\rho^T B_\rho X\boldsymbol{\beta} & \mathrm{tr}\left(B_\rho\right) & \left(X^T B_\rho X\boldsymbol{\beta}\right)^T \\ \mathrm{tr}\left(B_\rho\right) & \frac{g}{2\sigma^2} & \mathbf{0}_k^T \\ X^T B_\rho X\boldsymbol{\beta} & \mathbf{0}_k & X^T X \end{bmatrix} \qquad (2.11) \\
&= \frac{1}{\sigma^2}\begin{bmatrix} I^*_{\rho,\rho} & I^*_{\rho,\sigma^2} & I^*_{\rho,\boldsymbol{\beta}} \\ I^*_{\sigma^2,\rho} & I^*_{\sigma^2,\sigma^2} & I^*_{\sigma^2,\boldsymbol{\beta}} \\ I^*_{\boldsymbol{\beta},\rho} & I^*_{\boldsymbol{\beta},\sigma^2} & I^*_{\boldsymbol{\beta},\boldsymbol{\beta}} \end{bmatrix}.
\end{aligned}$$

Using cofactor expansion and determinant properties of block matrices (see e.g., Harville, 1997), we can write

$$\begin{aligned}
\det\left(I^*\left(\boldsymbol{\theta}\right)\right) &= -I^*_{\sigma^2,\rho}\det\begin{pmatrix} I^*_{\rho,\sigma^2} & I^*_{\rho,\boldsymbol{\beta}} \\ I^*_{\boldsymbol{\beta},\sigma^2} & I^*_{\boldsymbol{\beta},\boldsymbol{\beta}} \end{pmatrix} + I^*_{\sigma^2,\sigma^2}\det\begin{pmatrix} I^*_{\rho,\rho} & I^*_{\rho,\boldsymbol{\beta}} \\ I^*_{\boldsymbol{\beta},\rho} & I^*_{\boldsymbol{\beta},\boldsymbol{\beta}} \end{pmatrix} \\
&= -I^*_{\sigma^2,\rho}\det\left(I^*_{\rho,\sigma^2}\right)\det\left(I^*_{\boldsymbol{\beta},\boldsymbol{\beta}}\right) + I^*_{\sigma^2,\sigma^2}\det\begin{pmatrix} I^*_{\rho,\rho} & I^*_{\rho,\boldsymbol{\beta}} \\ I^*_{\boldsymbol{\beta},\rho} & I^*_{\boldsymbol{\beta},\boldsymbol{\beta}} \end{pmatrix} \\
&= -{I^*_{\sigma^2,\rho}}^2\det\left(I^*_{\boldsymbol{\beta},\boldsymbol{\beta}}\right) + I^*_{\sigma^2,\sigma^2}\det\left(I^*_{\rho,\rho}\right)\det\left(I^*_{\boldsymbol{\beta},\boldsymbol{\beta}} - I^*_{\boldsymbol{\beta},\rho}{I^*_{\rho,\rho}}^{-1}I^*_{\rho,\boldsymbol{\beta}}\right) \\
&= -{I^*_{\sigma^2,\rho}}^2\det\left(I^*_{\boldsymbol{\beta},\boldsymbol{\beta}}\right) + I^*_{\sigma^2,\sigma^2}I^*_{\rho,\rho}\det\left(I^*_{\boldsymbol{\beta},\boldsymbol{\beta}} - I^*_{\boldsymbol{\beta},\rho}{I^*_{\rho,\rho}}^{-1}I^*_{\rho,\boldsymbol{\beta}}\right). \qquad (2.12)
\end{aligned}$$

By auxiliary fact (2), we can further write (2.12) as

$$\begin{aligned}
\det\left(I^*\left(\boldsymbol{\theta}\right)\right) &= -{I^*_{\sigma^2,\rho}}^2\det\left(I^*_{\boldsymbol{\beta},\boldsymbol{\beta}}\right) + I^*_{\sigma^2,\sigma^2}I^*_{\rho,\rho}\left(1 - I^*_{\rho,\boldsymbol{\beta}}{I^*_{\boldsymbol{\beta},\boldsymbol{\beta}}}^{-1}I^*_{\boldsymbol{\beta},\rho}{I^*_{\rho,\rho}}^{-1}\right)\det\left(I^*_{\boldsymbol{\beta},\boldsymbol{\beta}}\right) \\
&= \det\left(I^*_{\boldsymbol{\beta},\boldsymbol{\beta}}\right)\left(I^*_{\sigma^2,\sigma^2}I_{\rho,\rho} - I^*_{\sigma^2,\sigma^2}I^*_{\rho,\boldsymbol{\beta}}{I^*_{\boldsymbol{\beta},\boldsymbol{\beta}}}^{-1}I^*_{\boldsymbol{\beta},\rho} - {I^*_{\sigma^2,\rho}}^2\right).
\end{aligned}$$

Plugging in the actual entries for the respective blocks yields

$$
\begin{aligned}
&I^*_{\sigma^2,\sigma^2} I^*_{\rho,\rho} - I^*_{\sigma^2,\sigma^2} I^*_{\rho,\boldsymbol{\beta}} {I^*_{\boldsymbol{\beta},\boldsymbol{\beta}}}^{-1} I^*_{\boldsymbol{\beta},\rho} - {I^*_{\sigma^2,\rho}}^2 \\
&= \frac{g}{2\sigma^2}\left(\sigma^2\left(\operatorname{tr}\left(B_\rho^T B_\rho\right) + \operatorname{tr}\left(B_\rho^2\right)\right) + \boldsymbol{\beta}^T X^T B_\rho^T B_\rho X \boldsymbol{\beta}\right) \\
&\quad - \frac{g}{2\sigma^2}\left(X^T B_\rho X \boldsymbol{\beta}\right)^T \left(X^T X\right)^{-1} X^T B_\rho X \boldsymbol{\beta} - \operatorname{tr}^2\left(B_\rho\right) \\
&= \frac{g}{2}\left(\operatorname{tr}\left(B_\rho^T B_\rho\right) + \operatorname{tr}\left(B_\rho^2\right) + \frac{1}{\sigma^2}\boldsymbol{\beta}^T X^T B_\rho^T \left(I_g - X\left(X^T X\right)^{-1} X^T\right) B_\rho X \boldsymbol{\beta} - \frac{2}{g}\operatorname{tr}^2\left(B_\rho\right)\right) \\
&= \frac{g}{2}\left(\operatorname{tr}\left(B_\rho^T B_\rho\right) + \operatorname{tr}\left(B_\rho^2\right) + \frac{1}{\sigma^2}\boldsymbol{\beta}^T X^T B_\rho^T M B_\rho X \boldsymbol{\beta} - \frac{2}{g}\operatorname{tr}^2\left(B_\rho\right)\right).
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\det\left(I\left(\boldsymbol{\theta}\right)\right) = \det\left(\frac{1}{\sigma^2} I^*\left(\boldsymbol{\theta}\right)\right) = \left(\sigma^2\right)^{-k-2} \det\left(I^*\left(\boldsymbol{\theta}\right)\right) \\
&= \left(\sigma^2\right)^{-k-2} \det\left(X^T X\right) \frac{g}{2}\left(\operatorname{tr}\left(B_\rho^T B_\rho\right) + \operatorname{tr}\left(B_\rho^2\right) + \frac{1}{\sigma^2}\boldsymbol{\beta}^T X^T B_\rho^T M B_\rho X \boldsymbol{\beta} - \frac{2}{g}\operatorname{tr}^2\left(B_\rho\right)\right) \\
&\propto \left(\sigma^2\right)^{-k-2}\left(\operatorname{tr}\left(B_\rho^T B_\rho\right) + \operatorname{tr}\left(B_\rho^2\right) + \frac{1}{\sigma^2}\boldsymbol{\beta}^T X^T B_\rho^T M B_\rho X \boldsymbol{\beta} - \frac{2}{g}\operatorname{tr}^2\left(B_\rho\right)\right),
\end{aligned}
$$

from which, together with the definition of Jeffreys rule prior, the result follows.

Proof of Corollary 2.2

(i) From the definition of Jeffreys rule prior, it follows that

$$
\begin{aligned}
&\int_0^\infty p_{\mathrm{J}}\left(\boldsymbol{\theta}\right) \mathrm{d}\sigma^2 \\
&> \int_0^1 p_{\mathrm{J}}\left(\boldsymbol{\theta}\right) \mathrm{d}\sigma^2 \\
&\propto \int_0^1 \left(\sigma^2\right)^{-\frac{k+2}{2}} \left\{\operatorname{tr}\left(B_\rho^T B_\rho\right) + \operatorname{tr}\left(B_\rho^2\right) + \frac{1}{\sigma^2}\boldsymbol{\beta}^T X^T B_\rho^T M B_\rho X \boldsymbol{\beta} - \frac{2}{g}\operatorname{tr}^2\left(B_\rho\right)\right\}^{1/2} \mathrm{d}\sigma^2 \\
&> \int_0^1 \left(\sigma^2\right)^{-\frac{k+2}{2}} \left\{\operatorname{tr}\left(B_\rho^T B_\rho\right) + \operatorname{tr}\left(B_\rho^2\right) + \boldsymbol{\beta}^T X^T B_\rho^T M B_\rho X \boldsymbol{\beta} - \frac{2}{g}\operatorname{tr}^2\left(B_\rho\right)\right\}^{1/2} \mathrm{d}\sigma^2 \\
&= \left\{\operatorname{tr}\left(B_\rho^T B_\rho\right) + \operatorname{tr}\left(B_\rho^2\right) + \boldsymbol{\beta}^T X^T B_\rho^T M B_\rho X \boldsymbol{\beta} - \frac{2}{g}\operatorname{tr}^2\left(B_\rho\right)\right\}^{1/2} \int_0^1 \left(\sigma^2\right)^{-\frac{k+2}{2}} \mathrm{d}\sigma^2 \\
&= \infty.
\end{aligned}
$$

(ii) Defining $h_1\left(\rho\right) := \operatorname{tr}\left(B_\rho^T B_\rho\right) + \operatorname{tr}\left(B_\rho^2\right) \geq 0$ and $h_2\left(\rho, \boldsymbol{\beta}\right) := \boldsymbol{\beta}^T X^T B_\rho^T M B_\rho X \boldsymbol{\beta} \geq 0$, we can write for Jeffreys rule posterior

$$\begin{aligned} p_{\mathrm{J}}\left(\boldsymbol{\theta}|\boldsymbol{y}\right) \propto \left(\sigma^2\right)^{-\frac{k+2}{2}} & \sqrt{\operatorname{tr}\left(B_\rho^T B_\rho\right) + \operatorname{tr}\left(B_\rho^2\right) + \frac{1}{\sigma^2}\boldsymbol{\beta}^T X^T B_\rho^T M B_\rho X \boldsymbol{\beta} - \frac{2}{g}\operatorname{tr}^2\left(B_\rho\right)} \\ & |A_\rho| \left(\sigma^2\right)^{-\frac{g}{2}} \exp\left(-\frac{1}{2\sigma^2}\left(A_\rho \boldsymbol{y} - X\boldsymbol{\beta}\right)^T \left(A_\rho \boldsymbol{y} - X\boldsymbol{\beta}\right)\right) \\ \le \left(\sigma^2\right)^{-\frac{g+k}{2}-1} & \sqrt{h_1\left(\rho\right) + \frac{1}{\sigma^2} h_2\left(\rho, \boldsymbol{\beta}\right)} \\ & |A_\rho| \exp\left(-\frac{1}{2\sigma^2}\left(A_\rho \boldsymbol{y} - X\boldsymbol{\beta}\right)^T \left(A_\rho \boldsymbol{y} - X\boldsymbol{\beta}\right)\right). \end{aligned} \tag{2.13}$$

Using auxilliary fact (1) and integrating (2.13) over $\boldsymbol{\beta}$ yields

$$\begin{aligned} & \int_{\mathbb{R}^k} \left(\sigma^2\right)^{-\frac{g+k}{2}-1} \sqrt{h_1\left(\rho\right) + \frac{1}{\sigma^2} h_2\left(\rho, \boldsymbol{\beta}\right)} |A_\rho| \\ & \quad \exp\left(-\frac{1}{2\sigma^2}\left(A_\rho \boldsymbol{y} - X\boldsymbol{\beta}\right)^T \left(A_\rho \boldsymbol{y} - X\boldsymbol{\beta}\right)\right) \mathrm{d}\boldsymbol{\beta} \\ & \le \left(\sigma^2\right)^{-\frac{g+k}{2}-1} |A_\rho| \\ & \quad \int_{\mathbb{R}^k} \left(\sqrt{h_1\left(\rho\right)} + \sqrt{\frac{1}{\sigma^2} h_2\left(\rho, \boldsymbol{\beta}\right)}\right) \exp\left(-\frac{1}{2\sigma^2}\left(A_\rho \boldsymbol{y} - X\boldsymbol{\beta}\right)^T \left(A_\rho \boldsymbol{y} - X\boldsymbol{\beta}\right)\right) \mathrm{d}\boldsymbol{\beta} \\ & = \left(\sigma^2\right)^{-\frac{g+k}{2}-1} |A_\rho| \exp\left(-\frac{1}{2\sigma^2} \boldsymbol{y}^T A_\rho^T M A_\rho \boldsymbol{y}\right) \\ & \quad \int_{\mathbb{R}^k} \left(\sqrt{h_1\left(\rho\right)} + \sqrt{\frac{1}{\sigma^2} h_2\left(\rho, \boldsymbol{\beta}\right)}\right) \exp\left(-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)^T \left(X^T X\right) \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)\right) \mathrm{d}\boldsymbol{\beta} \\ & = \left(\sigma^2\right)^{-\frac{g+k}{2}-1} |A_\rho| \exp\left(-\frac{1}{2\sigma^2} \boldsymbol{y}^T A_\rho^T M A_\rho \boldsymbol{y}\right) \sqrt{h_1\left(\rho\right)} \\ & \quad \int_{\mathbb{R}^k} \exp\left(-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)^T \left(X^T X\right) \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)\right) \mathrm{d}\boldsymbol{\beta} \qquad (2.14) \\ & \quad + \left(\sigma^2\right)^{-\frac{g+k}{2}-1} |A_\rho| \exp\left(-\frac{1}{2\sigma^2} \boldsymbol{y}^T A_\rho^T M A_\rho \boldsymbol{y}\right) \left(\sigma^2\right)^{-\frac{1}{2}} \\ & \qquad \int_{\mathbb{R}^k} \sqrt{h_2\left(\rho, \boldsymbol{\beta}\right)} \exp\left(-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)^T \left(X^T X\right) \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)\right) \mathrm{d}\boldsymbol{\beta}. \qquad (2.15) \end{aligned}$$

The integrand in (2.14) is the kernel of the probability density function of a multivariate normal random variable $\boldsymbol{Z} \sim N\left(\hat{\boldsymbol{\beta}}, \sigma^2\left(X^T X\right)^{-1}\right)$, so when $\left(X^T X\right)^{-1}$ exists, it follows that

$$\int_{\mathbb{R}^k} \exp\left(-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)^T \left(X^T X\right) \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)\right) \mathrm{d}\boldsymbol{\beta} \propto \left(\sigma^2\right)^{\frac{k}{2}}.$$

Similarly, the integrand in (2.15) can be expressed as the expected value for the square root of a quadratic form involving $\boldsymbol{Z}$. By using auxiliary fact (3) and Jensen's inequality, we can write

$$\begin{aligned}
&\int_{\mathbb{R}^k} \sqrt{h_2\left(\rho, \boldsymbol{\beta}\right)} \exp\left(-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)^T \left(X^T X\right)\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)\right) \mathrm{d}\boldsymbol{\beta} \\
&\propto \left(\sigma^2\right)^{\frac{k}{2}} \int_{\mathbb{R}^k} \sqrt{\boldsymbol{z}^T X^T B_\rho^T M B_\rho X \boldsymbol{z}} \left((2\pi)^k \left|\sigma^2 \left(X^T X\right)^{-1}\right|\right)^{-\frac{1}{2}} \\
&\qquad\qquad \exp\left(-\frac{1}{2\sigma^2}\left(\boldsymbol{z}-\hat{\boldsymbol{\beta}}\right)^T \left(X^T X\right)\left(\boldsymbol{z}-\hat{\boldsymbol{\beta}}\right)\right) \mathrm{d}\boldsymbol{z} \\
&= \left(\sigma^2\right)^{\frac{k}{2}} \mathbb{E}_{\boldsymbol{Z}}\left[\sqrt{\boldsymbol{Z}^T X^T B_\rho^T M B_\rho X \boldsymbol{Z}}\right] \\
&\leq \left(\sigma^2\right)^{k/2} \sqrt{\mathbb{E}_{\boldsymbol{Z}}\left[\boldsymbol{Z}^T X^T B_\rho^T M B_\rho X \boldsymbol{Z}\right]} \\
&= \left(\sigma^2\right)^{\frac{k}{2}} \sqrt{\operatorname{tr}\left(X^T B_\rho^T M B_\rho X \sigma^2 \left(X^T X\right)^{-1}\right) + \hat{\boldsymbol{\beta}}^T X^T B_\rho^T M B_\rho X \hat{\boldsymbol{\beta}}} \\
&= \left(\sigma^2\right)^{\frac{k}{2}} \sqrt{\sigma^2 \operatorname{tr}\left(B_\rho^T M B_\rho P\right) + \hat{\boldsymbol{\beta}}^T X^T B_\rho^T M B_\rho X \hat{\boldsymbol{\beta}}} \\
&\leq \left(\sigma^2\right)^{\frac{k}{2}} \left(\left(\sigma^2\right)^{\frac{1}{2}} \sqrt{\operatorname{tr}\left(B_\rho^T M B_\rho P\right)} + \sqrt{\hat{\boldsymbol{\beta}}^T X^T B_\rho^T M B_\rho X \hat{\boldsymbol{\beta}}}\right),
\end{aligned}$$

where $P := X\left(X^T X\right)^{-1} X^T$, which is symmetric, idempotent, and positive semi-definite, so $\operatorname{tr}\left(B_\rho^T M B_\rho P\right) \geq 0$ by auxiliary fact (4). Combining these observations with (2.14) and (2.15), it follows that

$$\begin{aligned}
&\left(\sigma^2\right)^{-\frac{g+k}{2}-1} \left|A_\rho\right| \exp\left(-\frac{1}{2\sigma^2} \boldsymbol{y}^T A_\rho^T M A_\rho \boldsymbol{y}\right) \\
&\int_{\mathbb{R}^k} \left(\sqrt{h_1\left(\rho\right)} + \sqrt{\frac{1}{\sigma^2} h_2\left(\rho, \boldsymbol{\beta}\right)}\right) \exp\left(-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)^T \left(X^T X\right)\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)\right) \mathrm{d}\boldsymbol{\beta} \\
&\leq \left(\sigma^2\right)^{-\frac{g+k}{2}-1} \left|A_\rho\right| \exp\left(-\frac{1}{2\sigma^2} \boldsymbol{y}^T A_\rho^T M A_\rho \boldsymbol{y}\right) \\
&\quad \left(\sqrt{h_1\left(\rho\right)} \left(\sigma^2\right)^{\frac{k}{2}} + \left(\sigma^2\right)^{\frac{k-1}{2}} \left(\left(\sigma^2\right)^{\frac{1}{2}} \sqrt{\operatorname{tr}\left(B_\rho^T M B_\rho P\right)} + \sqrt{\hat{\boldsymbol{\beta}}^T X^T B_\rho^T M B_\rho X \hat{\boldsymbol{\beta}}}\right)\right)
\end{aligned}$$

$$= \left(\sigma^2\right)^{-\frac{g}{2}-1} \left|A_\rho\right| \exp\left(-\frac{1}{2\sigma^2} \boldsymbol{y}^T A_\rho^T M A_\rho \boldsymbol{y}\right) \left(\sqrt{h_1\left(\rho\right)} + \sqrt{\operatorname{tr}\left(B_\rho^T M B_\rho P\right)}\right) \tag{2.16}$$

$$+ \left(\sigma^2\right)^{-\frac{g+1}{2}-1} \left|A_\rho\right| \exp\left(-\frac{1}{2\sigma^2} \boldsymbol{y}^T A_\rho^T M A_\rho \boldsymbol{y}\right) \sqrt{\hat{\boldsymbol{\beta}}^T X^T B_\rho^T M B_\rho X \hat{\boldsymbol{\beta}}}. \tag{2.17}$$

Next, observe that the terms involving σ^2 in (2.16) and (2.17) correspond to kernels of probability density functions of inverse gamma distributed random variables, so integrating over σ^2 yields

$$\int_0^\infty \left(\sigma^2\right)^{-\frac{g}{2}-1} \exp\left(-\frac{1}{2\sigma^2} \boldsymbol{y}^T A_\rho^T M A_\rho \boldsymbol{y}\right) \mathrm{d}\sigma^2 \propto \boldsymbol{y}^T A_\rho^T M A_\rho \boldsymbol{y}^{-\frac{g}{2}}, \tag{2.18}$$

$$\int_0^\infty \left(\sigma^2\right)^{-\frac{g+1}{2}-1} \exp\left(-\frac{1}{2\sigma^2}\boldsymbol{y}^T A_\rho^T M A_\rho \boldsymbol{y}\right) \mathrm{d}\sigma^2 \propto \boldsymbol{y}^T A_\rho^T M A_\rho \boldsymbol{y}^{-\frac{g+1}{2}}, \tag{2.19}$$

where the terms on the right hand side of (2.18) and (2.19) are bounded for $\rho \in \left(\lambda_g^{-1}, \lambda_1^{-1}\right)$ by assumption. Hence, it only remains to show that the term

$$|A_\rho| \left(\sqrt{h_1(\rho)} + \sqrt{\operatorname{tr}\left(B_\rho^T M B_\rho P\right)} + \sqrt{\hat{\boldsymbol{\beta}}^T X^T B_\rho^T M B_\rho X \hat{\boldsymbol{\beta}}}\right)$$

is bounded for $\rho \in \left(\lambda_g^{-1}, \lambda_1^{-1}\right)$. First, applying auxiliary fact (5) yields

$$\begin{aligned}
\sqrt{h_1(\rho)} &= \sqrt{\operatorname{tr}\left(B_\rho^T B_\rho\right) + \operatorname{tr}\left(B_\rho^2\right)} \leq \sqrt{\operatorname{tr}\left(B_\rho^T B_\rho\right)} + \sqrt{\operatorname{tr}\left(B_\rho^2\right)} \\
&\leq \sqrt{\frac{1}{2}\left(\operatorname{tr}\left(B_\rho^T B_\rho^T\right) + \operatorname{tr}\left(B_\rho^2\right)\right)} + \sqrt{\operatorname{tr}\left(B_\rho^2\right)} \\
&\propto \sqrt{\operatorname{tr}\left(B_\rho^2\right)} = \sqrt{\sum_{i=1}^{g} \frac{\lambda_i^2}{\left(1-\rho\lambda_i\right)^2}} \\
&\leq \sum_{i=1}^{g} \sqrt{\frac{\lambda_i^2}{\left(1-\rho\lambda_i\right)^2}} \leq \sum_{i=1}^{g} \frac{|\lambda_i|}{1-\rho\lambda_i}.
\end{aligned}$$

Second, using auxiliary facts (4) and (6), it holds that

$$\begin{aligned}
\sqrt{\operatorname{tr}\left(B_\rho^T M B_\rho P\right)} &\leq \sqrt{\operatorname{tr}\left(B_\rho^T M B_\rho\right)\operatorname{tr}(P)} \propto \sqrt{\operatorname{tr}\left(B_\rho^T M B_\rho\right)} = \sqrt{\operatorname{tr}\left(B_\rho B_\rho^T M\right)} \\
&\leq \sqrt{\operatorname{tr}\left(B_\rho B_\rho^T\right)\operatorname{tr}(M)} \propto \sqrt{\operatorname{tr}\left(B_\rho^T B_\rho\right)} \leq \sum_{i=1}^{g} \frac{|\lambda_i|}{1-\rho\lambda_i},
\end{aligned} \tag{2.20}$$

where (2.20) follows from the considerations above and the idempotence of M and P, respectively. Finally, with auxiliary facts (4)-(6) and after some algebraic manipulations, we can write

$$\begin{aligned}
&\sqrt{\hat{\boldsymbol{\beta}}^T X^T B_\rho^T M B_\rho X \hat{\boldsymbol{\beta}}} = \sqrt{\left(\left(X^T X\right)^{-1} X^T A_\rho \boldsymbol{y}\right)^T X^T B_\rho^T M B_\rho X \left(X^T X\right)^{-1} X^T A_\rho \boldsymbol{y}} \\
&= \sqrt{\operatorname{tr}\left(B_\rho^T M B_\rho P A_\rho \boldsymbol{y}\boldsymbol{y}^T A_\rho^T P\right)} = \sqrt{\operatorname{tr}\left(B_\rho^T M B_\rho P A_\rho \boldsymbol{y}\boldsymbol{y}^T A_\rho^T P^T\right)} \\
&\leq \sqrt{\operatorname{tr}\left(B_\rho^T M B_\rho\right)\operatorname{tr}\left(\left(\boldsymbol{y}^T A_\rho^T P^T\right)^T \boldsymbol{y}^T A_\rho^T P^T\right)} = \sqrt{\operatorname{tr}\left(B_\rho B_\rho^T M\right)\boldsymbol{y}^T A_\rho^T P A_\rho \boldsymbol{y}} \\
&\leq \sqrt{\operatorname{tr}\left(B_\rho B_\rho^T\right)\operatorname{tr}(M)\,\boldsymbol{y}^T A_\rho^T P A_\rho \boldsymbol{y}} \propto \sqrt{\operatorname{tr}\left(B_\rho^T B_\rho\right)\boldsymbol{y}^T A_\rho^T P A_\rho \boldsymbol{y}} \\
&= \sqrt{\operatorname{tr}\left(B_\rho^T B_\rho\right)}\sqrt{\left(\boldsymbol{y}^T A_\rho^T P A_\rho \boldsymbol{y}\right)}.
\end{aligned} \tag{2.21}$$

As $\boldsymbol{y}^T A_\rho^T P A_\rho \boldsymbol{y}$ is bounded for $\rho \in \left(\lambda_g^{-1}, \lambda_1^{-1}\right)$, the expression in (2.21) can be bounded itself by a multiple of the sum term in (2.20). Furthermore, if m_1 and m_g denote the algebraic multiplicity of λ_1 and λ_g, respectively, then

$$\begin{aligned}
&|A_\rho| \sum_{i=1}^{g} \frac{|\lambda_i|}{1-\rho\lambda_i} \\
&= \left(\prod_{i=1}^{g}(1-\rho\lambda_i)\right)\left(\sum_{i=1}^{g} \frac{|\lambda_i|}{1-\rho\lambda_i}\right) \\
&= \left(\prod_{i=1}^{g}(1-\rho\lambda_i)\right)\frac{|\lambda_1|}{1-\rho\lambda_1} + \ldots + \left(\prod_{i=1}^{g}(1-\rho\lambda_i)\right)\frac{|\lambda_g|}{1-\rho\lambda_g} \\
&= |\lambda_1|\left(\prod_{i=m_1+1}^{g}(1-\rho\lambda_i)\right) + \ldots + |\lambda_g|\left(\prod_{i=g-m_g}^{g}(1-\rho\lambda_i)\right) \qquad (2.22) \\
&< \infty,
\end{aligned}$$

as every summand in (2.22) is bounded for $\rho \in \left(\lambda_g^{-1}, \lambda_1^{-1}\right)$. This completes the proof.

Proof of Theorem 2.2

The model's Fisher information Matrix in (2.11) gives

$$\begin{aligned}
&\det\left(I_{(\rho,\sigma^2),(\rho,\sigma^2)}(\boldsymbol{\theta})\right) = \left(\sigma^2\right)^{-2}\left(I^*_{\rho,\rho}I^*_{\sigma^2,\sigma^2} - {I^*_{\rho,\sigma^2}}^2\right) \\
&= \left(\sigma^2\right)^{-2}\left(\frac{g}{2\sigma^2}\left(\sigma^2\left(\mathrm{tr}\left(B_\rho^T B_\rho\right) + \mathrm{tr}\left(B_\rho^2\right)\right) + \boldsymbol{\beta}^T X^T B_\rho^T B_\rho X\boldsymbol{\beta}\right) - \mathrm{tr}^2\left(B_\rho\right)\right) \\
&\propto \left(\sigma^2\right)^{-2}\left(\mathrm{tr}\left(B_\rho^T B_\rho\right) + \mathrm{tr}\left(B_\rho^2\right) + \frac{1}{\sigma^2}\boldsymbol{\beta}^T X^T B_\rho^T B_\rho X\boldsymbol{\beta} - \frac{2}{g}\,\mathrm{tr}^2\left(B_\rho\right)\right),
\end{aligned}$$

and $\det\left(I_{(\boldsymbol{\beta},\boldsymbol{\beta})}(\boldsymbol{\theta})\right) \propto 1$. The result follows from these observations and by the definition of Independence Jeffreys prior.

Proof of Corollary 2.3

These results are proved in an identical way as the ones in Corollary 2.2 and follow almost immediately.

Appendix 2.C Posterior sampling

We outlined the sampling procedure and gave the conditional posteriors based on the flat prior and the informative priors in Section 2.5. However, it remains to specify the exact forms of the candidate-generating distributions for the conditional posteriors for the parameter blocks (ρ, β_1) and $\widetilde{\boldsymbol{\beta}}$. As to the conditional posterior for (ρ, β_1) based on the flat prior, we first approximate $\log\left(|A_\rho|\right)$ by a second-order Taylor polynomial at $\rho = 0$, so $|A_\rho| \approx \exp\left(-\rho^2 \sum_{i=1}^{g} \lambda_i^2/2\right)$. Using this approximation, we can write

$$
\begin{aligned}
& p_{\mathrm{F}}\left(\left(\rho, \beta_1\right)|\sigma^2, \widetilde{\boldsymbol{\beta}}, \boldsymbol{y}\right) \\
& \propto \left|A_\rho\right| \exp\left(-\frac{1}{2\sigma^2}\left(A_\rho \boldsymbol{y} - X\boldsymbol{\beta}\right)^T\left(A_\rho \boldsymbol{y} - X\boldsymbol{\beta}\right)\right) \\
& \approx \exp\left(-\frac{\rho^2}{2}\sum_{i=1}^{g}\lambda_i^2\right)\exp\left(-\frac{1}{2\sigma^2}\left(A_\rho \boldsymbol{y} - X\boldsymbol{\beta}\right)^T\left(A_\rho \boldsymbol{y} - X\boldsymbol{\beta}\right)\right) \\
& \propto \exp\left(-\frac{\rho^2}{2}\sum_{i=1}^{g}\lambda_i^2 - \frac{1}{2\sigma^2}\left(\rho^2 \boldsymbol{y}^T W^T W \boldsymbol{y} - 2\rho \boldsymbol{y}^T W^T\left(\boldsymbol{y} - \widetilde{X}\widetilde{\boldsymbol{\beta}}\right) + 2\rho\beta_1 \boldsymbol{y}^T W^T \mathbf{1}_g\right.\right. \\
& \qquad\qquad \left.\left. -2\beta_1 \mathbf{1}_g^T\left(\boldsymbol{y} - \widetilde{X}\widetilde{\boldsymbol{\beta}}\right) + \beta_1^2 g\right)\right), \qquad (2.23)
\end{aligned}
$$

where the proportionality holds with respect to (ρ, β_1), $\widetilde{X}$ denotes the matrix X with its first column removed, $\widetilde{\boldsymbol{\beta}} = (\beta_2, ..., \beta_g)$, and $\mathbf{1}_g$ is the vector of ones of length g. The expression in (2.23) corresponds to the kernel of a bivariate normal density $q_{\mathrm{F}}(\rho, \beta_1)$ with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. By equating coefficients and after some algebraic manipulation, it follows that

$$
\begin{aligned}
\boldsymbol{\mu} &= \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \begin{pmatrix} \frac{\mathbf{1}_g^T\left(\left(\boldsymbol{y}-\widetilde{X}\widetilde{\boldsymbol{\beta}}\right)\boldsymbol{y}^T W^T \mathbf{1}_g - \mathbf{1}_g \boldsymbol{y}^T W^T\left(\boldsymbol{y}-\widetilde{X}\widetilde{\boldsymbol{\beta}}\right)\right)}{\left(\boldsymbol{y}^T W^T \mathbf{1}_g\right)^2 - g\left(\sigma^2 \sum_{i=1}^{g}\lambda_i^2 + \boldsymbol{y}^T W^T W \boldsymbol{y}\right)} \\ \frac{\boldsymbol{y}^T W^T\left(\boldsymbol{y}-\widetilde{X}\widetilde{\boldsymbol{\beta}}\right) - \mu_1\left(\sigma^2 \sum_{i=1}^{g}\lambda_i^2 + \boldsymbol{y}^T W^T W \boldsymbol{y}\right)}{\boldsymbol{y}^T W^T \mathbf{1}_g} \end{pmatrix}, \\
\Sigma &= \sigma^2 \begin{pmatrix} \sigma^2 \sum_{i=1}^{g}\lambda_i^2 + \boldsymbol{y}^T W^T W \boldsymbol{y} & \boldsymbol{y}^T W^T \mathbf{1}_g \\ \boldsymbol{y}^T W^T \mathbf{1}_g & g \end{pmatrix}^{-1}.
\end{aligned}
$$

We use this candidate-generating normal distribution also for the conditional posterior for (ρ, β_1) based on Jeffreys rule prior and Independence Jeffreys prior, as for these priors the prior information for (ρ, β_1) is quite vague compared to the likelihood. Note that due to the chosen parameter space of ρ, $q_{\mathrm{F}}(\rho, \beta_1)$ is in fact truncated to $\left(\lambda_g^{-1}, \lambda_1^{-1}\right) \times \mathbb{R}$. In the simulation study, we relied on the rtmvnorm() function from the tmvtnorm package in R to sample from this truncated distribution (Wilhelm & Manjunath, 2015). Similarly, we can obtain the corresponding mean vector and covariance matrix of the candidate-generating bivariate normal distribution for (ρ, β_1) when using a normal prior for ρ.

The conditional posterior for $\widetilde{\boldsymbol{\beta}}$ based on the flat prior and the informative priors is a multivariate normal distribution and can be directly sampled from, for which we used the rmvnorm() function from the mvtnorm package in R (Genz et al., 2014). Its mean vector and covariance matrix are given by

$$\boldsymbol{\mu}_{\widetilde{\boldsymbol{\beta}}} = \boldsymbol{\mu}_{\boldsymbol{\beta}_2} + \Sigma_{\boldsymbol{\beta}_{21}}\Sigma_{\beta_{11}}^{-1}\left(\beta_1 - \mu_{\beta_1}\right), \qquad (2.24)$$

$$\Sigma_{\widetilde{\boldsymbol{\beta}}} = \Sigma_{\boldsymbol{\beta}_{22}} - \Sigma_{\boldsymbol{\beta}_{21}}\Sigma_{\beta_{11}}^{-1}\Sigma_{\boldsymbol{\beta}_{12}}, \qquad (2.25)$$

where

$$\boldsymbol{\mu}_{\boldsymbol{\beta}} = \left(X^TX\right)^{-1}X^TA_\rho\boldsymbol{y} = \begin{pmatrix}\mu_{\beta_1}\\ \boldsymbol{\mu}_{\boldsymbol{\beta}_2}\end{pmatrix} \quad \text{sized} \begin{pmatrix}1\times 1\\ (k-1)\times 1\end{pmatrix},$$

$$\Sigma_{\boldsymbol{\beta}} = \sigma^2\left(X^TX\right)^{-1} = \begin{pmatrix}\Sigma_{\beta_{11}} & \Sigma_{\boldsymbol{\beta}_{12}}\\ \Sigma_{\boldsymbol{\beta}_{21}} & \Sigma_{\boldsymbol{\beta}_{22}}\end{pmatrix} \text{sized} \begin{pmatrix}1\times 1 & 1\times(k-1)\\ (k-1)\times 1 & (k-1)\times(k-1)\end{pmatrix}.$$

For the same reasons as before, we use this candidate-generating distribution also for the conditional posterior for $\widetilde{\boldsymbol{\beta}}$ based on Jeffreys rule prior and Independence Jeffreys prior.

Combining (2.2) and (2.5), the full conditionals based on Jeffreys rule prior can be written as

$$\begin{aligned}
p_{\mathrm{J}}\left((\rho,\beta_1)\,|\sigma^2,\widetilde{\boldsymbol{\beta}},\boldsymbol{y}\right) &\propto |A_\rho|\exp\left(-\frac{1}{2\sigma^2}\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}\right)\\
&\qquad\left(\operatorname{tr}\left(B_\rho^TB_\rho\right)+\operatorname{tr}\left(B_\rho^2\right)+\frac{1}{\sigma^2}\boldsymbol{\beta}^TX^TB_\rho^TMB_\rho X\boldsymbol{\beta}-\frac{2}{g}\operatorname{tr}^2\left(B_\rho\right)\right)^{\frac{1}{2}},\\
p_{\mathrm{J}}\left(\sigma^2|\,(\rho,\beta_1)\,,\widetilde{\boldsymbol{\beta}},\boldsymbol{y}\right) &\propto \left(\sigma^2\right)^{-\frac{g+k}{2}-1}\exp\left(-\frac{1}{2\sigma^2}\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}\right)\\
&\qquad\left(\operatorname{tr}\left(B_\rho^TB_\rho\right)+\operatorname{tr}\left(B_\rho^2\right)+\frac{1}{\sigma^2}\boldsymbol{\beta}^TX^TB_\rho^TMB_\rho X\boldsymbol{\beta}-\frac{2}{g}\operatorname{tr}^2\left(B_\rho\right)\right)^{\frac{1}{2}},\\
p_{\mathrm{J}}\left(\widetilde{\boldsymbol{\beta}}|\,(\rho,\beta_1)\,,\sigma^2,\boldsymbol{y}\right) &\propto \exp\left(-\frac{1}{2\sigma^2}\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}\right)\\
&\qquad\left(\operatorname{tr}\left(B_\rho^TB_\rho\right)+\operatorname{tr}\left(B_\rho^2\right)+\frac{1}{\sigma^2}\boldsymbol{\beta}^TX^TB_\rho^TMB_\rho X\boldsymbol{\beta}-\frac{2}{g}\operatorname{tr}^2\left(B_\rho\right)\right)^{\frac{1}{2}},
\end{aligned}$$

where $\boldsymbol{\varepsilon} = A_\rho\boldsymbol{y} - X\boldsymbol{\beta}$. As none of these full conditionals is of known analytical form, a Metropolis-Hastings step for each parameter (block) is needed. The candidate-generating distributions for the conditional posteriors for (ρ,β_1) and $\widetilde{\boldsymbol{\beta}}$ have already been given above, while we propose

$$q_{\mathrm{J}}\left(\sigma^2|\,(\rho,\beta_1)\,,\widetilde{\boldsymbol{\beta}},\boldsymbol{y}\right) \sim IG\left(\frac{g+k+1}{2},\frac{\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}}{2}\right)$$

as candidate-generating distribution for $p_{\mathrm{J}}\left(\sigma^2|\,(\rho,\beta_1)\,,\widetilde{\boldsymbol{\beta}},\boldsymbol{y}\right)$, which resulted in well-mixed Markov chains. Equivalently, we can also easily formulate the conditional posteriors based on Independence Jeffreys prior, where we suggest

$$q_{\mathrm{IJ}}\left(\sigma^2|\,(\rho,\beta_1)\,,\widetilde{\boldsymbol{\beta}},\boldsymbol{y}\right) \sim IG\left(\frac{g+1}{2},\frac{\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}}{2}\right)$$

as corresponding candidate-generating distribution for $p_{\mathrm{IJ}}\left(\sigma^2|\,(\rho,\beta_1)\,,\widetilde{\boldsymbol{\beta}},\boldsymbol{y}\right)$.

We outline the full sampling algorithm based on using the flat prior in the following:

(1) Set starting values $\left(\rho^0, \beta_1^0\right), \left(\sigma^2\right)^0$, and $\widetilde{\boldsymbol{\beta}}^0$, e.g., to their maximum likelihood estimates, and the number of draws N.

(2) Repeat steps (3) - (5) for $i = 1 : N$.

(3) Perform a Metropolis-Hastings step for (ρ, β_1) with the target density $p_{\mathrm{F}}\left((\rho, \beta_1) | \sigma^2, \widetilde{\boldsymbol{\beta}}, \boldsymbol{y}\right)$ and the candidate-generating density $q_{\mathrm{F}}(\rho, \beta_1)$, i.e.,

- Draw from $q_{\mathrm{F}}(\rho, \beta_1)$ until a draw $\left(\hat{\rho}, \hat{\beta}_1\right)$ satisfies $\left(\hat{\rho}, \hat{\beta}_1\right) \in \left(\lambda_g^{-1}, \lambda_1^{-1}\right) \times \mathbb{R}$. Draw u from the uniform distribution $U(0,1)$.
- Calculate the acceptance probability $\alpha\left[\left(\rho^{i-1}, \beta_1^{i-1}\right), \left(\hat{\rho}, \hat{\beta}_1\right)\right]$, defined as

$$\alpha\left[\left(\rho^{i-1}, \beta_1^{i-1}\right), \left(\hat{\rho}, \hat{\beta}_1\right)\right] :=$$
$$\min\left(\frac{p_{\mathrm{F}}\left(\hat{\rho}, \hat{\beta}_1 | \left(\sigma^2\right)^{i-1}, \widetilde{\boldsymbol{\beta}}^{i-1}, \boldsymbol{y}\right) q_{\mathrm{F}}\left(\rho^{i-1}, \beta_1^{i-1}\right)}{p_{\mathrm{F}}\left(\rho^{i-1}, \beta_1^{i-1} | \left(\sigma^2\right)^{i-1}, \widetilde{\boldsymbol{\beta}}^{i-1}, \boldsymbol{y}\right) q_{\mathrm{F}}\left(\hat{\rho}, \hat{\beta}_1\right)}, 1\right).$$

- If $u \leq \alpha\left[\left(\rho^{i-1}, \beta_1^{i-1}\right), \left(\hat{\rho}, \hat{\beta}_1\right)\right]$, set $\left(\rho^i, \beta_1^i\right) = \left(\hat{\rho}, \hat{\beta}_1\right)$.
- Else, set $\left(\rho^i, \beta_1^i\right) = \left(\rho^{i-1}, \beta_1^{i-1}\right)$.

(4) Draw $\left(\sigma^2\right)^i$, given $\left(\rho^i, \beta_1^i\right)$ and $\widetilde{\boldsymbol{\beta}}^{i-1}$, from the inverse gamma distribution in (2.9).

(5) Draw $\widetilde{\boldsymbol{\beta}}^i$, given $\left(\rho^i, \beta_1^i\right)$ and $\left(\sigma^2\right)^i$, from the $(k-1)$-variate normal distribution with mean $\boldsymbol{\mu}_{\widetilde{\boldsymbol{\beta}}}$ and covariance matrix $\Sigma_{\widetilde{\boldsymbol{\beta}}}$ as in (2.24), (2.25).

Note that when using Jeffreys rule prior or Independence Jeffreys prior, the direct sampling procedures in (4) and (5) are replaced by Metropolis-Hastings steps based on the corresponding candidate-generating densities.

Chapter 3

Bayesian hypothesis testing in the network autocorrelation model

Abstract

Currently available (classical) testing procedures for the network autocorrelation parameter can only be used to falsify a precise null hypothesis of no network effect. Classical methods can be neither used to quantify evidence for the null nor to test multiple hypotheses on the network autocorrelation parameter simultaneously. This article presents flexible Bayes factor testing procedures that do not have these limitations. We propose Bayes factors based on an empirical and a uniform prior for the network effect, respectively, first. Next, we develop a fractional Bayes factor where a default prior is automatically constructed. Simulation results suggest that the first two Bayes factors show superior performance and are the Bayes factors we recommend. We apply the recommended Bayes factors to three data sets from the literature and compare the results to those coming from classical analyses using p-values. R code for efficient computation of the Bayes factors is provided.

This chapter is based on: Dittrich, D., Leenders, R.Th.A.J., and Mulder, J. (in press). Network Autocorrelation Modeling: A Bayes Factor Approach for Testing (Multiple) Precise and Interval Hypotheses. *Sociological Methods & Research*. `http://doi.org/10.1177/0049124117729712`.

3.1 Introduction

The network autocorrelation model (Ord, 1975) has been extensively used to represent theories of social influence throughout recent decades. It allows researchers to quantify the strength of a peer effect in a network for a given theory of interpersonal influence while controlling for sociological and other covariates. The identification and magnitude of the peer, or network, effect ρ, also known as the *network autocorrelation*, is often the focus of interest in model applications. Typically, a researcher aims to identify if there is social influence present in the network, resulting in an inferential test of $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$. Subsequently, if the null hypothesis is rejected, the researcher then concludes that there is evidence for some degree of social influence.

Even though the network autocorrelation model and this statistical approach have yielded many interesting and theoretically useful findings, more intricate hypothesis tests are often more informative. For example, when a researcher is interested in testing whether the degree of social influence is either zero, small, medium, or large, a more informative test would be $H_0 : \rho = 0$ versus $H_1 : 0 < \rho \leq .25$ versus $H_2 : .25 < \rho \leq .5$ versus $H_3 : .5 < \rho < 1$. This not only allows the researcher to conclude if there is evidence for a non-zero amount of network autocorrelation in the network, but it grants the researcher the opportunity to simultaneously test several strengths of social influence against each other as well. This chapter focuses on Bayesian hypothesis testing procedures for such multiple precise and interval hypotheses on the network autocorrelation.

The standard approach to testing a network effect is null hypothesis significance testing. Classical null hypothesis significance tests such as the Wald test, the likelihood ratio test, or the Lagrange multiplier test are based on different test statistics, summary values constructed from the sample, which have asymptotically known distributions under the null hypothesis (Leenders, 1995). Then, assuming the null hypothesis to be true, the probability of observing a test statistic at least as extreme as the observed test statistic is calculated. This probability is called the p-value. Subsequently, the p-value is compared to a pre-specified significance level α, which is usually set to .05 (Weakliem, 2004). If the p-value is smaller than α, the null hypothesis is rejected. In this case, one would conclude that there is enough evidence in the data to reject the null hypothesis of no network effect. If the p-value is larger than α, there is not enough evidence in the data to reject the null.

Classical null hypothesis significance testing in the network autocorrelation model has a number of disadvantages. First, the procedure cannot be used to provide evidence in favor of the null hypothesis (Wetzels & Wagenmakers, 2012); it can only falsify the null hypothesis. If the p-value is larger than α, this ultimately implies a state of ignorance where the null can be neither rejected nor supported by the data. For example, if the estimated autocorrelation parameter is $\hat{\rho} = .16$ with a two-sided p-value of .08 (so the null hypothesis that $\rho = 0$ would not be rejected based on $\alpha = .05$), this does not mean that the null hypothesis is “accepted”; it merely means that judgment regarding the rejection of this particular hypothesis is suspended and that no degree of belief in the hypothesis has been determined. On the other hand, if the p-value is smaller than α, it is still not possible to say how much evidence there is against the null (and certainly not how much evidence there is in favor of ρ being .16 indeed), only that there is enough evidence to reject the null that $\rho = 0$ based on a chosen significance level. Second, classical null hypothesis significance tests are not consistent. If the null is true and the sample size grows to infinity, there is still a probability of α (typically .05) of drawing the incorrect conclusion that the null is false. This is undesirable, as one should be able to draw more accurate conclusions with growing sample size. Third, p-values in the network autocorrelation model depend heavily on asymptotic theory and consequently, the Type I error rate is not controlled for in an accurate manner in the case of small networks (Dittrich et al., 2017). A final important issue in the context of this chapter is that p-values cannot be adequately used when testing multiple competing hypotheses against each other (Shaffer, 1995). Instead, one can only test each hypothesis against the null that does not answer the question which hypothesis, out of a set of precise and interval hypotheses, is most supported by the data.

In this chapter, we propose *Bayes factor* tests (Jeffreys, 1961; Kass & Raftery, 1995; Mulder & Wagenmakers, 2016) as an alternative approach to classical null hypothesis significance testing in the network autocorrelation model. The Bayes factor is a Bayesian hypothesis testing criterion that circumvents the aforementioned issues with null hypothesis significance testing. First, in contrast to classical null hypothesis significance testing, it allows the researcher to evaluate and quantify the relative evidence in the data in favor of the null, or any other, hypothesis against another hypothesis (Kuha, 2004). These hypotheses can be precise hypotheses, e.g., $H_0 : \rho = 0$, or interval hypotheses, e.g., $H_1 : 0 < \rho < 1$. For example, a Bayes factor of $B_{01} = 5$ implies that it is five times more likely to observe the data under the null hypothesis than under a specific alternative hypothesis H_1. Second, Bayes factors are consistent, i.e., if the null hypothesis is true, the Bayes factor B_{01} tends to infinity as the sample size goes to infinity (Casella et al., 2009). In other words, the larger the sample size, the more do the data support one hypothesis over another. Third, the Bayes factor provides “exact” inference without the need for asymptotic approximations (De Oliveira & Song, 2008). Lastly, the Bayes factor can be straightforwardly extended to test more than two hypotheses against each other, e.g., $H_0 : \rho = 0$ versus $H_1 : -.25 < \rho < 0$ versus $H_2 : 0 < \rho \leq .25$ versus $H_3 : .25 < \rho \leq .5$ versus $H_4 : .5 < \rho < 1$ (Raftery et al., 1997). This feature is of particular relevance in the network autocorrelation model, as in many network studies, researchers do not doubt that social influence occurs but are interested in testing competing theories about its strength. In summary, these advantageous properties explain the increasing usage of Bayes factors in social science research, such as in ANOVA (Klugkist et al., 2005), linear regression models (Braeken et al., 2015), repeated measures (Mulder et al., 2009), or structural equation modeling (Gu et al., 2014).

So far, only two Bayes factors in the network autocorrelation model have been developed in the literature. Hepple (1995a) proposed a Bayes factor for testing competing connectivity matrices against each other, while LeSage & Parent (2007) provided Bayes factors for testing different explanatory variables. We have neither found any Bayes factor for the standard one-sided test $H_0 : \rho = 0$ versus $H_1 : 0 < \rho < 1$ nor for any multiple hypothesis tests. This is surprising, as the network autocorrelation parameter is at the heart of the model and testing for network effects is of crucial importance for network scientists when testing for and understanding theories of social influence. In sum, the objective of this chapter is to provide methodology that

- makes it possible to test multiple competing hypotheses on ρ against each other and precisely quantify the amount of evidence in favor of any of the hypotheses tested (including a null hypothesis),
- works for any combination of precise and/or interval hypotheses, and
- overcomes the problems with classical null hypothesis significance testing of ρ.

We also provide ready-to-use R code (R Core Team, 2017) to make the methodology easily applicable for applied researchers.

In order to compute the Bayes factor, so-called *prior distributions*, or simply *priors*, for the unknown model parameters have to be specified under each hypothesis. These priors quantify which values for the parameters are most likely before observing the data. For the testing problems considered in this chapter, the prior for the network autocorrelation parameter ρ under the alternative(s) is most important. We develop and explore several Bayes factors for testing the network effect: first, a Bayes factor based on an empirical *informative prior* that stems from an extensive literature review of empirical applications of the network autocorrelation model; second, a Bayes factor based on a uniform prior that assumes every value for ρ to be equally likely a priori; third, a so-called *fractional Bayes factor* (O'Hagan, 1995) that can be computed without needing to formulate a *proper*, i.e., integrable, *prior* distribution for ρ based on one's prior beliefs. Subsequently, we conduct a simulation study to investigate the numerical properties of and differences between the proposed Bayes factors and then use the Bayes factors to re-analyze three data sets from the literature. Finally, we give R code for the computation of the Bayes factors.

The chapter is organized as follows: in Section 3.2, we discuss the network autocorrelation model in more detail and continue with a short introduction to Bayesian hypothesis testing in Section 3.3. In Section 3.4, we motivate several prior choices for the network autocorrelation parameter ρ. We assess the numerical performance of the Bayes factors in a simulation study in Section 3.5 and highlight their practical use with three examples in Section 3.6. Section 3.7 concludes.

3.2 The network autocorrelation model

Most social phenomena are embedded within networks of interdependencies. Building from a standard linear regression model, the network autocorrelation model effectively incorporates such interdependencies between individuals. In the model, the network structure is explicitly used to account for network influence on a variable of interest and to estimate the magnitude of this influence that is considered to be a model parameter, the network autocorrelation ρ. Formally, the network autocorrelation model is expressed as

$$\boldsymbol{y} = \rho W \boldsymbol{y} + X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N\left(\boldsymbol{0}_g, \sigma^2 I_g\right), \tag{3.1}$$

where $\boldsymbol{y}$ is a $(g \times 1)$ vector of values for a dependent variable for the g network actors, X is a $(g \times k)$ matrix of values for the g actors on k covariates (possibly including a column of ones for an intercept term), $\boldsymbol{\beta}$ is a $(k \times 1)$ vector of regression coefficients, $\boldsymbol{0}_g$ is a $(g \times 1)$ vector of zeros, I_g denotes the $(g \times g)$ identity matrix, and $\boldsymbol{\varepsilon}$ is a $(g \times 1)$ vector containing independent and identically normally distributed error terms with zero mean and variance of σ^2. Furthermore, W is a given $(g \times g)$ *connectivity matrix* representing social ties in a network, with W_{ij} denoting the degree of influence of actor j on actor i.[1] Finally, the network autocorrelation ρ is the key parameter of the model and quantifies the social influence for given $\boldsymbol{y}$, W, and X. We denote the resulting set of model parameters as $\boldsymbol{\theta} := \left(\rho, \sigma^2, \boldsymbol{\beta}\right)$.

Subsequently, we will repeatedly rely on the model's likelihood, given by

$$f\left(\boldsymbol{y}|\rho, \sigma^2, \boldsymbol{\beta}\right) = \left|\det\left(A_\rho\right)\right| \left(2\pi\sigma^2\right)^{-\frac{g}{2}} \exp\left(-\frac{1}{2\sigma^2}\left(A_\rho \boldsymbol{y} - X\boldsymbol{\beta}\right)^T \left(A_\rho \boldsymbol{y} - X\boldsymbol{\beta}\right)\right),$$

where $A_\rho := I_g - \rho W$ (see e.g., Doreian, 1980). To ensure that the model is well-defined, there are restrictions on the region of support for ρ. Typically, this region is chosen as the interval containing $\rho = 0$ for which A_ρ is non-singular (Hepple, 1995a; LeSage & Parent, 2007; Smith, 2009). In this case, the corresponding admissible interval for ρ is given by $\left(\lambda_g^{-1}, \lambda_1^{-1}\right)$, where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_g$ are the ordered eigenvalues of W (Hepple, 1995a). For *row-standardized* connectivity matrices W, i.e., where each row sum is one, it holds that $\lambda_1 = 1$ (Anselin, 1982). Without loss of generality, we restrict ourselves to such commonly used row-standardized connectivity matrices in the remainder of this chapter. Hence, the model's parameter space becomes $\Theta := \Theta_\rho \times \Theta_{\sigma^2} \times \Theta_{\boldsymbol{\beta}} = \left(\lambda_g^{-1}, 1\right) \times (0, \infty) \times \mathbb{R}^k$.

Throughout the literature, the model has also been named as mixed regressive-autoregressive model (Ord, 1975), spatial effects model (Doreian, 1980), network effects model (Marsden & Friedkin, 1993), or spatial lag model (Anselin, 2002), and it has been applied in many different fields, such as criminology (Baller et al., 2001; Tita & Radil, 2011), geography (McMillen, 2010; Mur et al., 2008), political science (Beck et al., 2006; Gimpel & Schuknecht, 2003), or sociology (Duke, 1993; Kirk & Papachristos, 2011; Mizruchi & Stearns, 2006).

3.3 Bayesian hypothesis testing

In many network studies, researchers have expectations about the magnitude of the network effect. An interesting research question is whether the network effect can be classified as zero, small, medium, or large. These expectations can be translated to a set of multiple

[1] By convention, we exclude loops, i.e., relationships from an actor to himself, so $W_{ii} = 0$ for all $i \in \{1, ..., g\}$.

hypotheses on the network autocorrelation by setting $H_0 : \rho = 0$, $H_1 : 0 < \rho \leq .25$, $H_2 : .25 < \rho \leq .5$, and $H_3 : .5 < \rho < 1$, and the question to be answered is which of these hypotheses is most plausible. In general, such a test is much more insightful than the standard test of no network effect versus "some" (positive) network effect, $H_0 : \rho = 0$ versus $H_1 : 0 < \rho < 1$. In order to illustrate this, consider a situation in which the estimated network autocorrelation parameter is $\hat{\rho} = .16$, with a 95% confidence interval for ρ of $(-.06, .37)$, and a one-sided p-value of .08. Using the standard significance level of $\alpha = .05$, we would not reject the null hypothesis that $\rho = 0$ and conclude that there is no statistically significant network effect present in the data. Based on the confidence interval, however, we do have quite a lot of confidence that the network effect may be positive. Hence, based on these classical outcomes, it is very difficult to state how plausible it is that the true network effect is zero, small, medium, or large, which was the initial research question.

The Bayes factor is a Bayesian hypothesis testing criterion that resolves this issue by providing a means to directly quantify how plausible each hypothesis is after observing the data. Suppose that we are interested in testing $T \geq 2$ hypotheses, H_0, H_1, H_2, ..., H_{T-1}. First, in Bayesian hypothesis testing, prior probabilities have to be assigned to both the model parameters under each hypothesis and to the hypotheses themselves. We denote these latter *prior hypotheses probabilities* by $p\left(H_0\right)$, $p\left(H_1\right)$, ..., $p\left(H_{T-1}\right)$, with $\sum_{t=0}^{T-1} p\left(H_t\right) = 1$, which reflect how plausible we believe each hypothesis to be (relative to each other) before observing the data. There are multiple ways to assign prior hypotheses probabilities, e.g., by assuming equal prior probabilities (reflecting prior ignorance), i.e., $p\left(H_0\right) = ... = p\left(H_{T-1}\right) = 1/T$ (Hepple, 1995a; LeSage & Parent, 2007), or by formulating specific prior probabilities for the various hypotheses. We will discuss procedures for eliciting prior probabilities for interval hypotheses in Section 3.4.

Next, after observing the data $\boldsymbol{y}$, Bayes' theorem is applied to update the prior expectations with the information contained in the data. The resulting *posterior hypotheses probabilities*, $p\left(H_t|\boldsymbol{y}\right)$, can then be written as

$$p\left(H_t|\boldsymbol{y}\right) = \frac{p\left(\boldsymbol{y}|H_t\right) p\left(H_t\right)}{p\left(\boldsymbol{y}\right)} = \frac{p\left(\boldsymbol{y}|H_t\right) p\left(H_t\right)}{\sum\limits_{t'=0}^{T-1} p\left(\boldsymbol{y}|H_{t'}\right) p\left(H_{t'}\right)}, t \in \{0, 1, ..., T-1\}. \tag{3.2}$$

These posterior probabilities quantify how probable each hypothesis is after observing the data, the quantity that researchers are typically interested in.

The term $p\left(\boldsymbol{y}|H_t\right)$ in (3.2) is called the *marginal likelihood* under hypothesis H_t and denotes the probability that the data were observed under hypothesis H_t. It is computed by integrating the product of the model's likelihood function and the prior distribution for the model parameters under hypothesis H_t. Hence, the marginal likelihood can be seen as a weighted likelihood over the parameter space under hypothesis H_t, with the prior under hypothesis H_t acting as a weight function. In formal notation,

$$p\left(\boldsymbol{y}|H_t\right) = \int_{\Theta_t} f\left(\boldsymbol{y}|\boldsymbol{\theta}_t\right) p_t\left(\boldsymbol{\theta}_t|H_t\right) \mathrm{d}\boldsymbol{\theta}_t, \tag{3.3}$$

where $\boldsymbol{\theta}_t$ are the model parameters under hypothesis H_t, $p_t\left(\boldsymbol{\theta}_t|H_t\right)$ their prior density, and Θ_t the corresponding parameter space. For example, for $H_0 : \rho = 0$ and $H_1 : 0 < \rho < 1$ in the network autocorrelation model (3.1), it follows that $\boldsymbol{\theta}_0 = \left(\sigma^2, \boldsymbol{\beta}\right)$, $\Theta_0 = \mathbb{R}^+ \times \mathbb{R}^k$, and $\boldsymbol{\theta}_1 = \left(\rho, \sigma^2, \boldsymbol{\beta}\right)$, $\Theta_1 = (0,1) \times \mathbb{R}^+ \times \mathbb{R}^k$.

When performing pairwise model comparisons between two hypotheses H_t and $H_{t'}$, $t, t' \in \{0, ..., T-1\}$, we can consider the ratio of the corresponding posterior hypotheses probabilities. In this case, the normalizing constant $p\left(\boldsymbol{y}\right)$ in (3.2) cancels out and we can write

$$\frac{p\left(H_t|\boldsymbol{y}\right)}{p\left(H_{t'}|\boldsymbol{y}\right)} = \frac{p\left(\boldsymbol{y}|H_t\right)}{p\left(\boldsymbol{y}|H_{t'}\right)} \times \frac{p\left(H_t\right)}{p\left(H_{t'}\right)} := BF_{tt'} \times \frac{p\left(H_t\right)}{p\left(H_{t'}\right)}. \tag{3.4}$$

The term $p\left(H_t\right)/p\left(H_{t'}\right)$ in (3.4) is called the *prior odds* of the two hypotheses and quantifies how much more, or less, likely a researcher expects hypothesis H_t to be compared to hypothesis $H_{t'}$ before observing the data. When a researcher does not a priori believe one to be more likely than the other, the prior odds can be set equal to one. The term $p\left(H_t|\boldsymbol{y}\right)/p\left(H_{t'}|\boldsymbol{y}\right)$ in (3.4) is known as the *posterior odds* and reflects how much more (if larger than one), or less (if smaller than one), likely hypothesis H_t is than hypothesis $H_{t'}$ after taking the observed data into account. For example, if the posterior odds is five, this means that hypothesis H_t is five times more likely than hypothesis $H_{t'}$ for this data set. From (3.4) we can see that the posterior odds can be written as the prior odds multiplied by the Bayes factor, $BF_{tt'}$, which is defined as the ratio of two marginal likelihoods. Hence, the Bayes factor indicates to what extent the data change the prior odds to the posterior odds. Note that the Bayes factor can be used to quantify the relative evidence in the data in favor of the hypotheses without needing to specify how plausible they are before observing the data. If both models are considered as equally likely a priori, i.e., $p\left(H_t\right) = p\left(H_{t'}\right)$, the Bayes factor equals the posterior odds.[2]

The Bayes factor is a measure of relative evidence; it quantifies the amount of evidence in the data in favor of one hypothesis *relative* to another hypothesis. Jeffreys (1961) proposed a classification scheme to group Bayes factors into different categories, see Table 3.1. For example, there is "substantial" (relative) evidence in the data for hypothesis H_t when the Bayes factor $BF_{tt'}$ exceeds three and, equivalently, "substantial" evidence for hypothesis $H_{t'}$ when the Bayes factor is less than 1/3. These labels provide some rough guidelines when speaking of relative evidence in favor of a hypothesis but the interpretation should ultimately depend on the context of the research question (Kass & Raftery, 1995).

[2] We use the terms hypothesis and model interchangeably throughout the chapter.

Table 3.1 Evidence categories for the Bayes factor $BF_{tt'}$ as given by Jeffreys (1961).

$BF_{tt'}$			$\log(BF_{tt'})$			Interpretation
	$>$	100		$>$	4.61	Decisive evidence for hypothesis H_t
30	-	100	3.40	-	4.61	Very strong evidence for hypothesis H_t
10	-	30	2.30	-	3.40	Strong evidence for hypothesis H_t
3	-	10	1.10	-	2.30	Substantial evidence for hypothesis H_t
1	-	3	0	-	1.10	Not worth more than a bare mention
1/3	-	1	-1.10	-	0	Not worth more than a bare mention
1/10	-	1/3	-2.30	-	-1.10	Substantial evidence for hypothesis $H_{t'}$
1/30	-	1/10	-3.40	-	-2.30	Strong evidence for hypothesis $H_{t'}$
1/100	-	1/30	-4.61	-	-3.40	Very strong evidence for hypothesis $H_{t'}$
	$<$	1/100		$<$	-4.61	Decisive evidence for hypothesis $H_{t'}$

3.4 Bayes factor tests for the network autocorrelation parameter

In this section, we propose three Bayes factor tests when testing precise hypotheses, $H_{\text{precise}} : \rho = c$, and interval hypotheses, $H_{\text{interval}} : a_1 < \rho < a_2$, on the network autocorrelation parameter. One of the most important steps in Bayesian hypothesis testing is the prior specification of the model parameters. In the network autocorrelation model (3.1), a prior for ρ must be specified under an interval hypothesis, while no prior for ρ needs to be formulated under a precise hypothesis, as ρ is not a free parameter in this case. Despite its importance, prior specification of the network effect ρ has been largely neglected in the scarce literature on Bayesian hypothesis testing in the network autocorrelation model. The previous works of Hepple (1995a), X. Han & Lee (2013), and LeSage (2014a) are exclusively based on a uniform prior for ρ when testing the plausibility of different connectivity matrices (Hepple, 1995a) and spatial model specifications (X. Han & Lee, 2013; LeSage, 2014a), respectively. LeSage & Parent (2007) additionally employed a beta prior for ρ in a variable selection problem. As these authors did not consider testing ρ in particular, the prior choice was also less important in those contexts. In our setting, however, the prior for the network effect ρ under the alternative(s) should be carefully chosen, as the Bayes factor can be sensitive to the prior for the tested parameter (Kass & Raftery, 1995; Liu & Aitkin, 2008; Sinharay & Stern, 2002).

Prior expectations about the network autocorrelation parameter can be formulated based on a researcher's beliefs or stem from previous empirical evidence from the literature. On the other hand, if the available prior information is weak or a researcher prefers to avoid adding prior information to an analysis, so-called *non-informative priors* are often used. Such non-informative priors are typically improper, i.e., they do not integrate to a finite value, and are supposed to be completely dominated by the data (Gelman et al., 2013). In the following, we first present an empirical informative prior for ρ, second, a vague proper prior for ρ, and third, an improper prior for the network effect. We combine these different marginal prior distributions for ρ with the standard non-informative prior for the nuisance parameters $(\sigma^2, \boldsymbol{\beta})$, $p(\sigma^2, \boldsymbol{\beta}) \propto 1/\sigma^2$, assuming all parameters to be a priori independent (Hepple, 1995a; Holloway et al., 2002; LeSage, 1997a). However, note that the exact choice of the prior for the nuisance parameters hardly has an effect on the Bayes factor as long as this prior is relatively vague (Kass & Raftery, 1995).[3]

3.4.1 The Bayes factor based on an empirical prior

In our review of published empirical applications of the network autocorrelation model in Chapter 2, we showed that medium network effects, e.g., $\rho \approx .3$, are much more likely to be found in real-world networks than larger effects, e.g., $\rho \approx .8$, or negative effects, e.g., $\rho \approx -.2$. We also showed that the distribution of empirically observed network effects is well-approximated by a normal distribution centered around .36 with a standard deviation of .19. Unless a new study considers a case that is fundamentally different from the networks studied in the literature at large to date, a network autocorrelation in a new study is likely to come from a population distribution for ρ that resembles this normal distribution. This yields the empirically motivated prior

$$p_{\mathrm{E}}\left(\rho|H_{\text{interval}}\right) \sim N\left(.36, .19^2\right)(a_1, a_2), \tag{3.5}$$

which is the aforementioned normal distribution with a mean of .36 and a standard deviation of .19, truncated to the corresponding parameter space of ρ under an interval hypothesis. Based on this empirical prior, the marginal likelihoods under precise and under interval hypotheses on ρ, respectively, are given by

$$p_{\mathrm{E}}\left(\boldsymbol{y}|H_{\text{precise}}\right) = \pi^{-\frac{g-k}{2}}\Gamma\left(\frac{g-k}{2}\right)\sqrt{\left|\left(X^TX\right)^{-1}\right|}\,|A_c|\,\boldsymbol{y}^TA_c^TMA_c\boldsymbol{y}^{-\frac{g-k}{2}}, \tag{3.6}$$

$$\begin{aligned} &p_{\mathrm{E}}\left(\boldsymbol{y}|H_{\text{interval}}\right) \\ &= \pi^{-\frac{g-k}{2}}\Gamma\left(\frac{g-k}{2}\right)\sqrt{\left|\left(X^TX\right)^{-1}\right|}\int p_{\mathrm{E}}\left(\rho|H_{\text{interval}}\right)|A_\rho|\,\boldsymbol{y}^TA_\rho^TMA_\rho\boldsymbol{y}^{-\frac{g-k}{2}}\mathrm{d}\rho \qquad (3.7) \\ &= \frac{\pi^{-\frac{g-k}{2}}\Gamma\left(\frac{g-k}{2}\right)\sqrt{\left|\left(X^TX\right)^{-1}\right|}}{\sqrt{2\pi\times .19^2}\left(\Phi\left(\frac{a_2-.36}{.19}\right)-\Phi\left(\frac{a_1-.36}{.19}\right)\right)}\int\limits_{a_1}^{a_2}\exp\left(-\frac{\left(\rho-.36\right)^2}{.19^2}\right)|A_\rho|\,\boldsymbol{y}^TA_\rho^TMA_\rho\boldsymbol{y}^{-\frac{g-k}{2}}\mathrm{d}\rho, \end{aligned}$$

where $\Gamma(\cdot)$ represents the gamma function, $M := I_g - X\left(X^TX\right)^{-1}X^T$, and $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution (see e.g., Hepple, 1995a).

The uni-dimensional integral in (3.7) does not have a closed-form solution and has to be evaluated numerically. This can be done by relying on standard numerical methods, e.g., Simpson's rule (Atkinson, 1989), and we present R code therefor in Appendix 3.A, allowing researchers to use the Bayes factor without having to deal with the formulae themselves. Subsequently, the desired Bayes factors are obtained through (3.4) by using the marginal likelihoods under the precise and interval hypotheses under consideration.

[3] Improper priors for nuisance parameters appearing in both the null and the alternative model(s) are routinely used in Bayesian hypothesis testing (Hepple, 1995a; Jeffreys, 1961).

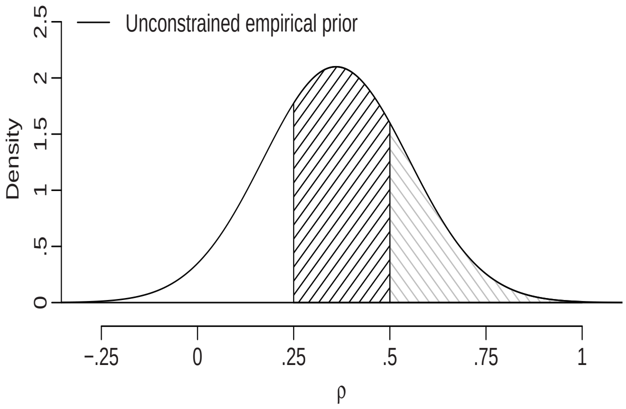

Figure 3.1 Probability density function of the unconstrained empirical prior for ρ, $p_{\mathrm{E}}(\rho) \sim N\left(.36, .19^2\right)$. The shaded areas under the probability density function correspond to the probabilities of ρ falling in the intervals $(.25, .5]$ (black) and $(.5, 1)$ (gray), which are equal to .49 and .23, respectively.

The unconstrained version of the empirical prior in (3.5), i.e., $p_{\mathrm{E}}(\rho) \sim N\left(.36, .19^2\right)$, can also be employed to determine prior probabilities for interval hypotheses. In this approach, the prior hypotheses probabilities are based on the probabilities of the tested model parameter falling in the respective intervals under a proper unconstrained prior. As an example, consider the two interval hypotheses $H_1 : .25 < \rho \leq .5$ and $H_2 : .5 < \rho < 1$. The probabilities of ρ falling in the intervals $(.25, .5]$ and $(.5, 1)$ under the unconstrained empirical prior are equal to .49 and .23, respectively, see Figure 3.1.[4] These probabilities give a quantification of the plausibility of the hypotheses under the unconstrained empirical prior. Hence, the corresponding prior odds, $p(H_1)/p(H_2)$, in this example is 2.12 ($\approx .49/.23$). In other words, under the empirical prior and before considering the data, a value for ρ inside the interval $(.25, .5]$ is 2.12 times as likely as ρ being inside $(.5, 1)$. Thus, if we assume that either hypothesis H_1 or hypothesis H_2 is true and that their prior odds corresponds to 2.12, the prior probabilities for the hypotheses are $p(H_1) = .68$ ($\approx .49/(.49 + .23)$) and $p(H_2) = .32$ ($\approx .23/(.49 + .23)$). This seems reasonable, as medium effects (hypothesis H_1) are generally more plausible than large effects (hypothesis H_2) in social network research (Dittrich et al., 2017). To the best of our knowledge, using an unconstrained prior to specify prior odds of interval hypotheses is novel in the literature.[5]

In the remainder of this chapter, we rely on the following method to assign prior model probabilities when testing one precise null hypothesis and $T - 1$ interval hypotheses. As

[4] In R, these probabilities are calculated as $pnorm(.5, mean = .36, sd = .19) - pnorm(.25, mean = .36, sd = .19) = .49$ and $pnorm(1, mean = .36, sd = .19) - pnorm(.5, mean = .36, sd = .19) = .23$. We rounded all probabilities to two decimal places in this chapter.

[5] Only Mulder (2014a) discussed a similar method for assigning prior probabilities to hypotheses with order constraints on the tested parameters.

there are T hypotheses in total, we set a prior probability of $1/T$ to the precise null hypothesis. Subsequently, the remaining probability of $(T-1)/T$ is divided upon the interval hypotheses, $H_1, ..., H_{T-1}$, using the prior probabilities of the intervals under a proper unconstrained prior. For example, consider $H_0 : \rho = 0$, $H_1 : -.25 < \rho < 0$, $H_2 : 0 < \rho \leq .25$, $H_3 : .25 < \rho \leq .5$, and $H_4 : .5 < \rho < 1$. Here, we test five hypotheses in total, so the null hypothesis is assigned a prior probability of $1/5 = .2$. The remaining probability of $4/5 = .8$ is split between the four interval hypotheses based on the probability mass contained in the four intervals under the unconstrained empirical prior. For the hypotheses considered above, the probabilities of ρ falling in these intervals are .03, .25, .49, and .23. As we have already set $p(H_0) = .2$, there is a total probability of .8 left for the remaining hypotheses. Rescaling and making them add up to .8 results in the prior model probabilities $p(H_1) = .02$, $p(H_2) = .20$, $p(H_3) = .39$, and $p(H_4) = .19$.[6] Finally, when combining the prior hypotheses probabilities and the marginal likelihoods in (3.6) and (3.7), the corresponding posterior model probabilities can be calculated via (3.2).

3.4.2 The Bayes factor based on a uniform prior

The uniform prior treats all possible network effects under the alternative(s) as equally likely before observing the data, resulting in a uniform prior distribution for ρ. As the region of support for ρ is bounded, the uniform prior for the network autocorrelation is a vague proper prior. Hence, it is less informative than the empirical prior but also does not represent complete prior ignorance that is typically expressed by using improper priors. The uniform prior under an interval hypothesis is written as

$$p_{\text{U}}\left(\rho|H_{\text{interval}}\right) \sim U\left(a_1, a_2\right),$$

where $U(a_1, a_2)$ denotes the uniform distribution on (a_1, a_2). The marginal likelihood under a precise hypothesis H_{precise} remains the same as in (3.6), while the marginal likelihood under an interval hypothesis H_{interval} in combination with a uniform prior for ρ is given by

$$\begin{aligned} & p_{\text{U}}\left(\boldsymbol{y}|H_{\text{interval}}\right) \\ & = \pi^{-\frac{g-k}{2}}\Gamma\left(\frac{g-k}{2}\right)\sqrt{\left|\left(X^TX\right)^{-1}\right|}\int p_{\text{U}}\left(\rho|H_{\text{interval}}\right)\left|A_\rho\right|\boldsymbol{y}^TA_\rho^TMA_\rho\boldsymbol{y}^{-\frac{g-k}{2}}\mathrm{d}\rho \\ & = \pi^{-\frac{g-k}{2}}\Gamma\left(\frac{g-k}{2}\right)\sqrt{\left|\left(X^TX\right)^{-1}\right|}\frac{1}{a_2-a_1}\int_{a_1}^{a_2}\left|A_\rho\right|\boldsymbol{y}^TA_\rho^TMA_\rho\boldsymbol{y}^{-\frac{g-k}{2}}\mathrm{d}\rho. \end{aligned} \tag{3.8}$$

Again, the Bayes factor is computed as the ratio of two marginal likelihoods based on the uniform prior for ρ. Similarly as for the empirical prior, an unconstrained uniform prior

[6] Note that the prior odds of any two interval hypotheses does not depend on the choice of the prior probability for the precise hypothesis.

can also be used to specify prior model probabilities. In the uniform prior setting, this implies that the prior odds of two interval hypotheses is equal to the ratio of their interval lengths. Posterior model probabilities can then be obtained via (3.2) by plugging in the marginal likelihoods under consideration and the prior probabilities for the hypotheses based on the uniform prior.

3.4.3 The fractional Bayes factor

Finally, there may be situations in which a researcher does not have any prior beliefs about possible network effects if the null were false, or in which a researcher prefers not to specify a proper prior for ρ based on external knowledge. Such prior ignorance is usually reflected by employing improper priors. However, if improper priors for the tested parameters are imposed, the Bayes factor depends on unknown normalizing constants and is not well-defined (O'Hagan, 1995). To that end, the fractional Bayes factor methodology was originally proposed by O'Hagan (1995) as a way to bypass this issue. In the fractional Bayes factor, the main idea is to split the information in the data into two different fractions that sum up to one. The first (small) fraction, denoted by b, is used to update the initial improper prior into a proper default prior; the second fraction, $1-b$, is taken to evaluate the hypotheses under investigation based on this proper default prior. In mathematical notation, this comes down to rewriting the marginal likelihood in (3.3) as

$$\begin{aligned} p\left(\boldsymbol{y}|H_t, b\right) &= \frac{\int_{\Theta_t} f\left(\boldsymbol{y}|\boldsymbol{\theta}_t\right) p_{\mathrm{NI}}\left(\boldsymbol{\theta}_t|H_t\right) \mathrm{d}\boldsymbol{\theta}_t}{\int_{\Theta_t} f\left(\boldsymbol{y}|\boldsymbol{\theta}_t\right)^b p_{\mathrm{NI}}\left(\boldsymbol{\theta}_t|H_t\right) \mathrm{d}\boldsymbol{\theta}_t} \\ &= \int_{\Theta_t} f\left(\boldsymbol{y}|\boldsymbol{\theta}_t\right)^{1-b} p\left(\boldsymbol{\theta}_t|H_t, \boldsymbol{y}^b\right) \mathrm{d}\boldsymbol{\theta}_t, \end{aligned} \tag{3.9}$$

where $p_{\mathrm{NI}}\left(\boldsymbol{\theta}_t|H_t\right)$ denotes a non-informative improper prior density for $\boldsymbol{\theta}_t$ under hypothesis H_t and $p\left(\boldsymbol{\theta}_t|H_t, \boldsymbol{y}^b\right) := f\left(\boldsymbol{y}|\boldsymbol{\theta}_t\right)^b p_{\mathrm{NI}}\left(\boldsymbol{\theta}_t|H_t\right) / \int_{\Theta_t} f\left(\boldsymbol{y}|\boldsymbol{\theta}_t\right)^b p_{\mathrm{NI}}\left(\boldsymbol{\theta}_t|H_t\right) \mathrm{d}\boldsymbol{\theta}_t$ is the updated proper default prior. Thus, in order to compute the marginal likelihood in the network autocorrelation model in the fractional Bayes factor approach, one needs to choose a non-informative improper prior for ρ and to specify the fraction b. As improper prior for ρ we use

$$p_{\mathrm{NI}}\left(\rho|H_{\mathrm{interval}}\right) = \left(1-\rho\right)^{-1} \mathbb{1}_{(0,1)}\left(\rho\right). \tag{3.10}$$

This prior approximates the model's Independence Jeffreys prior very well (see Appendix 3.C) that has been shown to outperform the standard uniform prior for ρ in Bayesian estimation of the model in Chapter 2.[7] At the same time, Independence Jeffreys prior itself also imposes an a priori dependence structure between the model parameters (Dittrich et al., 2017), which makes it difficult to compare its inferences to those based

[7] The prior in (3.10) has the same asymptotic behavior as the model's conditional Independence Jeffreys prior for ρ for $\rho \to 1$, see Appendix 3.C for a proof.

on the previously proposed normal and uniform marginal priors for ρ. For this reason, we rely on the marginal improper prior for ρ in (3.10) as it relaxes this a priori dependence. Note that in the fractional Bayes factor approach, we consider $H_{\text{interval}} : 0 < \rho < 1$ as the only alternative interval hypothesis. Else, if a researcher had more precise expectations about the magnitude of the network effect, e.g., $H_{\text{interval}} : .25 < \rho < .5$, it seems much more sensible to use a proper prior for ρ instead.

Typically, the fraction b in the fractional Bayes factor is chosen as the smallest value for which the updated default prior in (3.9) is proper (Berger & Mortera, 1999; Mulder, 2014b; O'Hagan, 1995). This choice results in maximal possible use of the information in the data for hypothesis testing. If the proposed improper prior in (3.10) is combined with the standard non-informative prior for $\left(\sigma^2, \boldsymbol{\beta}\right)$, $p\left(\sigma^2, \boldsymbol{\beta}\right) \propto 1/\sigma^2$, the corresponding updated prior in (3.9) is proper if $b > k/g$ (see Appendix 3.D for a proof). The proof shows that this also holds for the updated prior in (3.9) under a precise hypothesis H_{precise}. We denote the resulting choice for b as $b_1 = (k+1)/g$. On the other hand, if misspecification of the improper prior is a concern, larger values for b may be preferred, as they can reduce the sensitivity of the fractional Bayes factor to prior misspecification (Conigliani & O'Hagan, 2000; O'Hagan, 1995). Since empirical network autocorrelations are more likely to come from the estimated unconstrained population distribution for ρ in (3.5) than from a distribution that resembles the improper prior in (3.10), prior misspecification is indeed a valid concern here. In this case, O'Hagan (1995) suggested to use $b_2 = \max\left(k+1; \ln(g)\right)/g$, which makes the fractional Bayes factor more robust but increases slowly with g, or $b_3 = \max\left(k+1; \sqrt{g}\right)/g$, when sensitivity to misspecification of the prior is a serious concern. Finally, the marginal likelihoods under a precise hypothesis H_{precise} and under an interval hypothesis H_{interval} in the fractional Bayes factor approach are given by

$$p\left(\boldsymbol{y}|H_{\text{precise}}, b\right) = b^{\frac{gb}{2}} \pi^{\frac{g(b-1)}{2}} \frac{\Gamma\left(\frac{g-k}{2}\right)}{\Gamma\left(\frac{gb-k}{2}\right)} |A_c|^{1-b} \boldsymbol{y}^T A_c^T M A_c \boldsymbol{y}^{\frac{g(b-1)}{2}}, \tag{3.11}$$

$$\begin{aligned} p\left(\boldsymbol{y}|H_{\text{interval}}, b\right) &= b^{\frac{gb}{2}} \pi^{\frac{g(b-1)}{2}} \frac{\Gamma\left(\frac{g-k}{2}\right) \int p_{\text{NI}}\left(\rho|H_{\text{interval}}\right) |A_\rho| \boldsymbol{y}^T A_\rho^T M A_\rho \boldsymbol{y}^{-\frac{g-k}{2}} \mathrm{d}\rho}{\Gamma\left(\frac{gb-k}{2}\right) \int p_{\text{NI}}\left(\rho|H_{\text{interval}}\right) |A_\rho|^b \boldsymbol{y}^T A_\rho^T M A_\rho \boldsymbol{y}^{-\frac{gb-k}{2}} \mathrm{d}\rho} \\ &= b^{\frac{gb}{2}} \pi^{\frac{g(b-1)}{2}} \frac{\Gamma\left(\frac{g-k}{2}\right) \int_0^1 (1-\rho)^{-1} |A_\rho| \boldsymbol{y}^T A_\rho^T M A_\rho \boldsymbol{y}^{-\frac{g-k}{2}} \mathrm{d}\rho}{\Gamma\left(\frac{gb-k}{2}\right) \int_0^1 (1-\rho)^{-1} |A_\rho|^b \boldsymbol{y}^T A_\rho^T M A_\rho \boldsymbol{y}^{-\frac{gb-k}{2}} \mathrm{d}\rho}. \end{aligned} \tag{3.12}$$

For more technical details about the computation of (3.12), we refer the reader to Appendix 3.E. As before, the marginal likelihoods in (3.11) and (3.12) are used to calculate the fractional Bayes factor itself, which is again the ratio of two marginal likelihoods. If a researcher is also interested in obtaining posterior model probabilities, a default choice is to impose equal prior model probabilities, i.e., $p\left(H_0\right) = p\left(H_1\right) = .5$, while subjective prior model probabilities could be assigned if a researcher has clear beliefs about the plausibility of the null hypothesis of a zero network effect.

3.5 Simulation study

In this section, we present results of a simulation study that investigated the performance of and the differences between the Bayes factors discussed in Section 3.4. To that end, we first looked at the one-sided test A: $H_0 : \rho = 0$ versus $H_1 : 0 < \rho < 1$. In particular, we explored which Bayes factor converges fastest to a true data-generating hypothesis and assessed the sensitivity of the fractional Bayes factor to the choice of b. Next, we considered a multiple hypothesis test B with a precise hypothesis and four interval hypotheses: $H_0 : \rho = 0$ versus $H_1 : -.25 < \rho < 0$ versus $H_2 : 0 < \rho \leq .25$ versus $H_3 : .25 < \rho \leq .5$ versus $H_4 : .5 < \rho < 1$. Our primary goal here was to study how fast the posterior model probabilities converge to the true model for different data-generating hypotheses and to check how robust these findings are to different marginal priors for ρ as well as different prior model probabilities.

3.5.1 Study design

In order to mimic realistic networks, we considered two different approaches to obtain connectivity matrices W in test A. First, we used the well-known *small-world* structure (Watts & Strogatz, 1998) to generate simulated networks. Such networks are highly clustered, i.e., there is a tendency that network actors create very dense subnetworks but at the same time have small path lengths, i.e., there is a high probability that any two actors in the network are connected by short paths of acquaintances (Watts & Strogatz, 1998). Typically, this results in networks in which most actors are linked to only a few others, while some actors, also known as *hubs*, have a lot of ties. These hubs function as connectors between different subnetworks and shorten the path lengths between two actors in the entire network. Small-world structures can be observed in online social networks (Fu et al., 2007), scientific collaboration networks (Newman, 2001), or corporate elite networks (Davis et al., 2003). We obtained simulated small-world networks by relying on the watts.strogatz.game() function from the igraph package in R (Csárdi & Nepusz, 2006). In the underlying algorithm, a ring lattice of g actors, each connected to its d nearest neighbors by undirected ties, is constructed first. Next, with probability r, each tie in the network is randomly rewired. Following Neuman & Mizruchi (2010) and W. Wang et al. (2014), we set $r = .1$, which lead to highly clustered networks with low average path lengths. In our simulation study, we considered 14 network sizes ($g \in \{25, 50, 75, 100, 150, 200, 300, ..., 1000\}$) and two average degrees ($d \in \{4, 8\}$). The simulated connectivity matrices were binary, i.e., $W_{ij} = 1$ if there was a tie between actor i and j and zero otherwise. Subsequently, we row-normalized the generated symmetric raw connectivity matrices.[8]

Second, we also ran simulations using two prominent contiguity-based spatial networks; the 49 neighborhoods in Columbus, Ohio, analyzed in e.g., Anselin (1988), Elhorst (2014), and Hepple (1995a), and the 64 Louisiana parishes studied in Doreian (1980), Howard (1971), and Leenders (2002), among others. In general, networks based on spatial conti-

[8] According to Watts & Strogatz (1998), small-world networks are usually large and sparse with $g \gg d \gg \ln(g)$. Obviously, this relationship does not hold for all of our generated networks as e.g., $\ln(100) = 4.61$. However, we aimed to construct networks with a partially realistic, non-random configuration and simulated networks that at least resemble small-world structures.

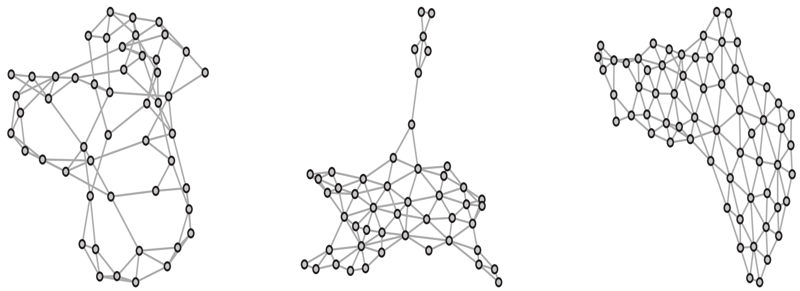

Figure 3.2 Simulated small-world network for $g = 50$, $d = 4$, $r = .1$ (left), the network of the 49 Columbus neighborhoods (middle), and the network of the 64 Louisiana parishes (right).

guity do not exhibit small-world properties, as there may be no short path between two distant nodes. Thus, we relied on these two real-world networks to gain insights into the behavior of the Bayes factors for network configurations that are different from typical small-world structures. We set W_{ij} to one if area i is adjacent to area j, to zero otherwise, and row-standardized the raw adjacency matrices. Graphical representations of the two spatial networks and an example of a simulated small-world network appear in Figure 3.2.[9]

For each of the network types, we included three covariates plus an intercept term (so $k = 4$) and used three fixed network effect sizes ($\rho \in \{0, .25, .5\}$) to generate $\boldsymbol{y}$ via $\boldsymbol{y} = (I_g - \rho W)^{-1} (X\boldsymbol{\beta} + \boldsymbol{\varepsilon})$ for test A. Furthermore, we also sampled network effects from the estimated empirical population distribution from Section 3.4.1, truncated to $(0, 1)$, rather than fixing ρ to a specific value. As the true network autocorrelation is unknown in practice, this is a more realistic setup than choosing specific values for ρ a priori. Finally, we drew independent values from a standard normal distribution for the elements of X (excluding the first column which is a vector of ones), $\boldsymbol{\beta}$, and $\boldsymbol{\varepsilon}$. Hence, we considered 120 scenarios for test A in total (14 small-world networks × 2 average degrees × 4 sampling schemes for ρ and 2 spatial networks × 4 sampling schemes for ρ) and simulated 1,000 data sets for each scenario.

For test B, we generated network effects using the empirical prior for ρ, truncated to the corresponding parameter space under each hypothesis. We assigned prior probabilities to the hypotheses based on both the unconstrained empirical the uniform prior for ρ. We drew values for the elements of X, $\boldsymbol{\beta}$, and $\boldsymbol{\varepsilon}$ as described in the previous paragraph, so for test B we examined 140 scenarios in total (14 small-world networks × 2 average degrees × 5 data-generating hypotheses) and generated 1,000 data sets for each scenario.

[9] We created the network plots using the plot.igraph() function from the igraph package in R.

3.5.2 Simulation results

Figure 3.3 and Table 3.2 show the average weight of evidence, i.e., the natural logarithm of the Bayes factor (Good, 1985), for the different Bayes factors and network structures for test A: $H_0 : \rho = 0$ versus $H_1 : 0 < \rho < 1$.[10]

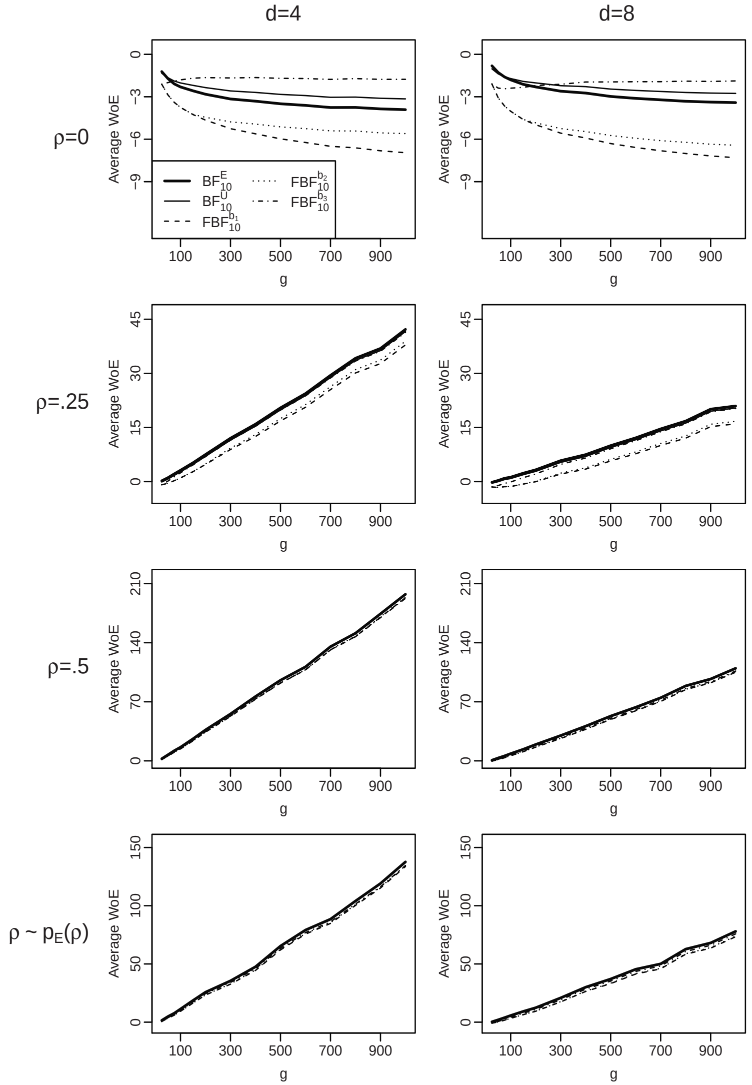

Figure 3.3 Average weight of evidence (WoE) for BF^E_{10} (thick solid line), BF^U_{10} (thin solid line), $FBF^{b_1}_{10}$ (dashed line), $FBF^{b_2}_{10}$ (dotted line), and $FBF^{b_3}_{10}$ (dot-dashed line) for test A: $H_0 : \rho = 0$ versus $H_1 : 0 < \rho < 1$ for 1,000 simulated data sets using generated small-world networks.

[10] In Table 3.2, we do not present the results for the fractional Bayes factor based on b_2, as $b_1 = b_2$ for both spatial networks when $k = 4$.

Table 3.2 Average weight of evidence for BF_{10}^{E}, BF_{10}^{U}, $FBF_{10}^{b_1}$, and $FBF_{10}^{b_3}$ for test A: $H_0 : \rho = 0$ versus $H_1 : 0 < \rho < 1$ for 1,000 simulated data sets using the network of the 49 Columbus neighborhoods and the network of the 64 Louisiana parishes.

	Columbus neighborhoods				Louisiana parishes			
	$\rho = 0$	$\rho = .25$	$\rho = .5$	$\rho \sim p_{\mathrm{E}}(\rho)$	$\rho = 0$	$\rho = .25$	$\rho = .5$	$\rho \sim p_{\mathrm{E}}(\rho)$
$\ln\left(BF_{10}^{E}\right)$	-1.7	1.0	7.4	4.8	-1.9	1.4	9.0	6.0
$\ln\left(BF_{10}^{U}\right)$	-1.7	.7	7.0	4.5	-1.7	1.0	8.6	5.7
$\ln\left(FBF_{10}^{b_1}\right)$	-2.8	-.5	6.0	3.4	-3.2	-.4	7.3	4.4
$\ln\left(FBF_{10}^{b_3}\right)$	-2.0	.2	6.2	3.8	-2.0	.5	7.7	5.0

We can conclude the following from these results. First, the Bayes factor based on the empirical prior, BF_{10}^{E} (thick solid line), and the Bayes factor based on the uniform prior, BF_{10}^{U} (thin solid line), show consistent behavior, i.e., the evidence for a true data-generating hypothesis is increasing with the network size. In addition, they almost always provide most evidence for the true hypothesis, except for $\rho = .25$ and some smaller network sizes in the small-world networks. Second, the evidence for a true alternative hypothesis grows with the network size at a faster rate than the evidence for a true null, as is common in other statistical models (Johnson & Rossell, 2010) and for precise hypotheses in general. Third, the Bayes factor based on the empirical prior results in slightly even more evidence for a true hypothesis than the Bayes factor based on the uniform prior. Fourth, the smaller the average degree, the bigger the evidence for a true alternative hypothesis provided by both BF_{10}^{E} and BF_{10}^{U} for fixed g and ρ. This behavior is due to the negative bias of ρ for increasing network densities in the model (Mizruchi & Neuman, 2008; Neuman & Mizruchi, 2010; Smith, 2009). Fifth, the fractional Bayes factors based on b_1 (dashed line) and b_2 (dotted line) yield very similar results. Overall, they provide less evidence for the alternative hypothesis compared to BF_{10}^{E}, or BF_{10}^{U}, and appear to be biased towards the null. For example, this bias is manifest from the fact that networks of approximately 300 nodes are needed before the fractional Bayes factors based on b_1 and b_2 result in evidence for the true alternative hypothesis when ρ equals .25 and $d = 8$, see Figure 3.3. By way of comparison, the Bayes factors based on the empirical prior and the uniform prior already point towards evidence for the alternative for small networks in this scenario. Sixth, the fractional Bayes factor based on b_3 (dot-dashed line) does not show consistent behavior when the null is true. In particular, the evidence for a true null hypothesis does not increase with the network size but remains constant, or in some cases even decreases.

In order to provide more insights into the behavior of the fractional Bayes factor, we investigated its behavior more thoroughly. To that end, we observe that for small values for b, the updated marginal prior for ρ in the fractional Bayes factor approach in (3.9) is just proper, with most of its probability mass still at values close to one for which the likelihood function is vanishing, see Figure 3.4. As the marginal likelihood under hypothesis H_1 in the fractional Bayes factor approach is essentially an average weighted likelihood over $(0, 1)$, with the updated prior acting as weight function, the fractional Bayes factors based on b_1 and b_2 tend to favor the null by construction. On the other

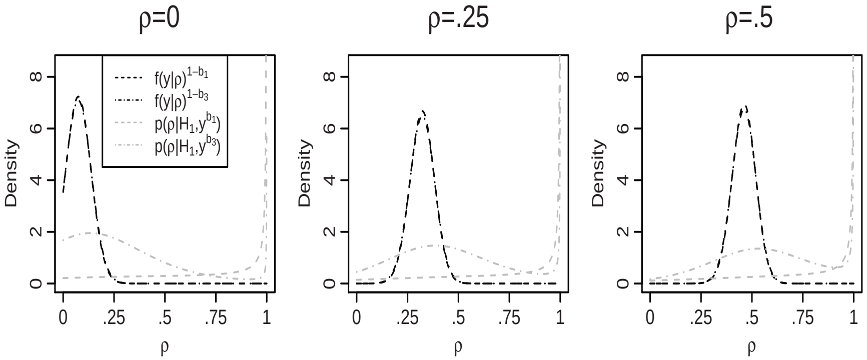

Figure 3.4 Normalized integrated likelihood components $f\left(\boldsymbol{y}|\rho\right)^{1-b_1}$ (black dashed line), $f\left(\boldsymbol{y}|\rho\right)^{1-b_3}$ (black dot-dashed line), and the updated marginal priors $p\left(\rho|H_1,\boldsymbol{y}^{b_1}\right)$ (gray dashed line) and $p\left(\rho|H_1,\boldsymbol{y}^{b_3}\right)$ (gray dot-dashed line) under $H_1 : 0 < \rho < 1$ based on $p_{\mathrm{NI}}\left(\rho|H_1\right) = \left(1-\rho\right)^{-1}$ for simulated data using generated small-world networks ($g = 100, d = 8$). The integrated likelihood component and the updated marginal prior based on b_2 are not plotted, as they are graphically indistinguishable from the curves based on b_1.

hand, for large values for b, the updated marginal prior for ρ exhibits a local maximum near the maximum likelihood estimate of ρ, see Figure 3.4. When the data are generated under the null, this results in a relatively larger average weighted likelihood over $(0, 1)$ and considerable support for the alternative, which explains the inconsistent behavior of the fractional Bayes factor based on the fraction b_3 when the null is true.

Given the results from our simulation study, we recommend using the Bayes factor based on the empirical prior for ρ when testing $H_0 : \rho = 0$ versus $H_1 : 0 < \rho < 1$, or the Bayes factor based on the uniform prior as a reasonable alternative. We do not recommend any of the fractional Bayes factors for this test, as they are either biased towards the null when the data are generated under the alternative or they show inconsistent behavior when the null is true. As a result, the fractional Bayes factors provide little improvement over classical null hypothesis significance testing, whereas the Bayes factors based on the empirical and the uniform prior for ρ clearly do perform considerably well.

Figure 3.5 reports the average posterior model probabilities for test B based on the empirical prior (left panel) and the ones based on the uniform prior (right panel) for different network sizes and data-generating hypotheses.[11] In general, the evidence for a true data-generating hypothesis is increasing with the sample size in all scenarios. Furthermore, the data-generating hypothesis always receives the most support, except for some very small network sizes and when the data are based on negative network effects in combination with using the empirical prior for ρ. In the latter case, the prior probability for hypothesis H_1, $p_{\mathrm{E}}\left(H_1\right) = .02$, is very small compared to the one for the null, $p_{\mathrm{E}}\left(H_0\right) = .20$. This means that the data need to support hypothesis H_1 approximately 10 times more than

[11] Simulation results for $d = 8$ are available from the authors upon request. We do not present them here, as they do no provide any additional, i.e., different, insights.

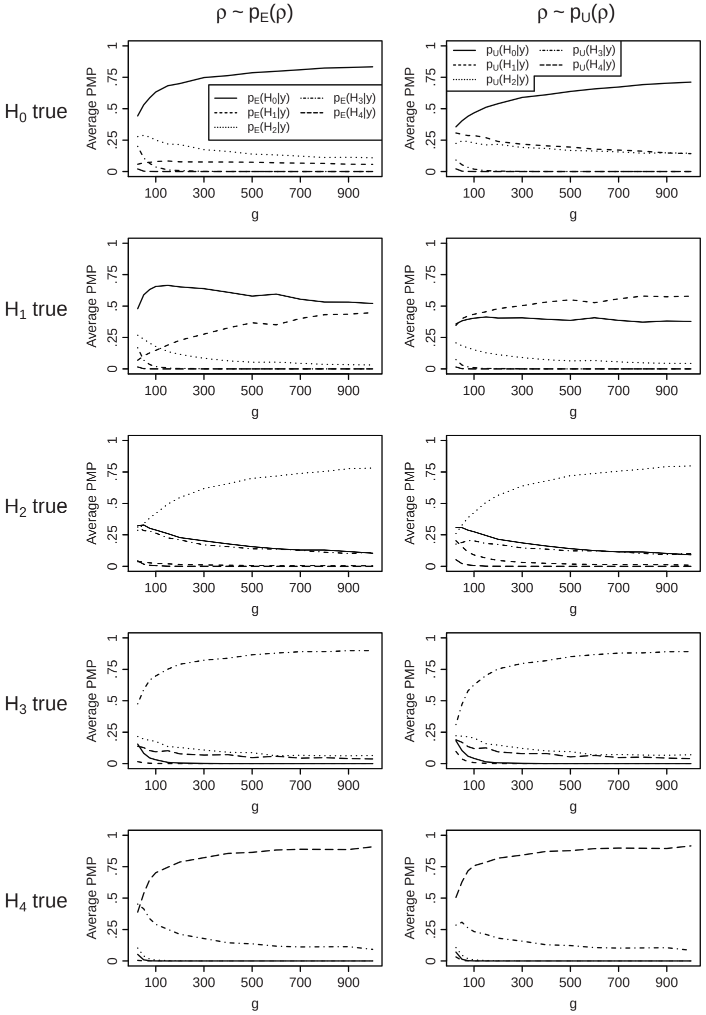

Figure 3.5 Average posterior model probabilities (PMP) for the hypotheses H_0 (solid line), H_1 (dashed line), H_2 (dotted line), H_3 (dot-dashed line), and H_4 (long-dashed line) for test B: $H_0 : \rho = 0$ versus $H_1 : -.25 < \rho < 0$ versus $H_2 : 0 < \rho \leq .25$ versus $H_3 : .25 < \rho \leq .5$ versus $H_4 : .5 < \rho < 1$ based on the empirical prior for ρ, $p_{\mathrm{E}}\left(H_t|\boldsymbol{y}\right), t \in \{0,1,2,3,4\}$ (left panel), and based on the uniform prior for ρ, $p_{\mathrm{U}}\left(H_t|\boldsymbol{y}\right)$ (right panel), for 1,000 simulated data sets using generated small-world networks.

the null such that the hypotheses receive at least equal posterior probability. However, we believe this behavior not to be a real concern, as the empirical literature suggests that it is highly unlikely to observe negative values for ρ in practice (Dittrich et al., 2017). Else, if a researcher deems such negative network autocorrelations to be more plausible than implied by $p_{\mathrm{E}}\left(H_1\right) = .02$, hypothesis H_1 should accordingly be given a higher prior probability.

3.6 Empirical examples

In the following, we apply the Bayes factor based on the empirical prior for ρ to three data sets from the literature to quantify the relative evidence in the data for competing hypotheses of interest. We tested the five hypotheses $H_0 : \rho = 0$, $H_1 : -.25 < \rho < 0$, $H_2 : 0 < \rho \leq .25$, $H_3 : .25 < \rho \leq .5$, and $H_4 : .5 < \rho < 1$ against each other, corresponding to the notion of "no network effect", a "minor negative network effect", a "minor (positive) network effect", a "medium network effect", and a "large network effect", respectively. We assigned prior probabilities to these hypotheses using the unconstrained empirical prior from Section 3.4, which yields $p_{\mathrm{E}}(H_0) = .20$, $p_{\mathrm{E}}(H_1) = .02$, $p_{\mathrm{E}}(H_2) = .20$, $p_{\mathrm{E}}(H_3) = .39$, and $p_{\mathrm{E}}(H_4) = .19$. In order to check for robustness to the choice of the prior for ρ and the prior model probabilities, we performed our analyses also using a uniform prior for the network autocorrelation parameter as well as for specifying prior model probabilities via the uniform unconstrained prior, i.e., $p_{\mathrm{U}}(H_0) = .20$, $p_{\mathrm{U}}(H_1) = p_{\mathrm{U}}(H_2) = p_{\mathrm{U}}(H_3) = .16$, and $p_{\mathrm{U}}(H_4) = .32$. Finally, we compared the results to those coming from classical tests using p-values.

3.6.1 Crime data

In the cross-sectional data set for 49 neighborhoods in Columbus, Ohio, first analyzed by Anselin (1988), the network autocorrelation model was used to explain the 1980 neighborhood crime rates, operationalized as the combined total of residential burglaries and vehicle thefts per 1,000 households. This data set is openly accessible as part of the columbus data from the R package spdep (Bivand & Piras, 2015) and Figure 3.6 (left) shows the spatial distribution of the crime rates. Anselin (1988) modeled these crime rates as a function of household income and housing value (in 1,000USD$), plus an intercept term. In his study, both explanatory variables had a negative impact on the crime rate, while the maximum likelihood estimate of the network effect was $\hat{\rho}_{\mathrm{ML}} = .40$, with a 95% confidence interval for ρ of $(.17, .64)$, and $p = .0008$.[12]

Using the empirical prior for ρ, the posterior hypotheses probabilities are $p_{\mathrm{E}}(H_0|\boldsymbol{y}) = .01$, $p_{\mathrm{E}}(H_1|\boldsymbol{y}) = .00$, $p_{\mathrm{E}}(H_2|\boldsymbol{y}) = .11$, $p_{\mathrm{E}}(H_3|\boldsymbol{y}) = .74$, and $p_{\mathrm{E}}(H_4|\boldsymbol{y}) = .14$. In other words, hypothesis H_3 is by far the most likely hypothesis and approximately 83 ($\approx .74/.01$), 1744, 7, and 6 times more plausible than hypothesis H_0, H_1, H_2, and H_4, respectively. Furthermore, the Bayes factor of hypothesis H_3 against hypothesis H_0 is 42.3, which is considered as very strong evidence in the data for hypothesis H_3, and the Bayes factor of hypothesis H_3 against hypothesis H_4 (the second most supported hypothesis) is 2.6, which implies minor evidence for a medium effect relative to a large effect. R code for computing these posterior model probabilities and corresponding Bayes factors is provided in Appendix 3.A. When relying on the uniform prior for ρ, the posterior model probabilities are $p_{\mathrm{U}}(H_0|\boldsymbol{y}) = .02$, $p_{\mathrm{U}}(H_1|\boldsymbol{y}) = .00$, $p_{\mathrm{U}}(H_2|\boldsymbol{y}) = .14$, $p_{\mathrm{U}}(H_3|\boldsymbol{y}) = .64$, and $p_{\mathrm{U}}(H_4|\boldsymbol{y}) = .20$. Again, hypothesis H_3 is by far the most likely hypothesis out of the five, followed by

[12] All reported p-values are for the one-sided test $H_0 : \rho = 0$ versus $H_1 : 0 < \rho < 1$ and based on the Wald test statistic using the expected Fisher information matrix for computing standard errors.

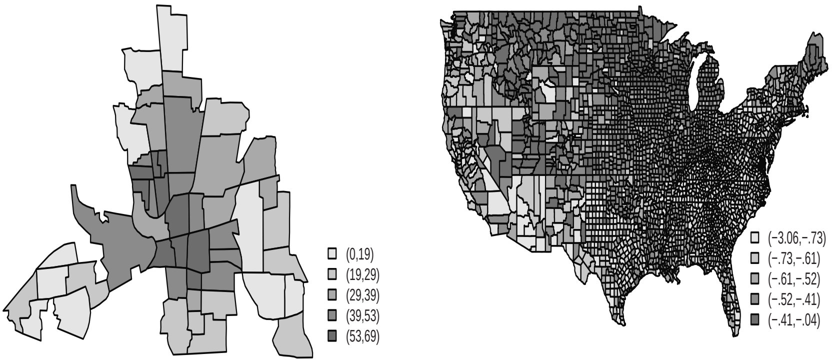

Figure 3.6 Number of residential burglaries and vehicle thefts per thousand households across 49 neighborhoods in Columbus, Ohio, in 1980 (left) and logarithm of voter turnout in the 1980 U.S. presidential election across 3,076 US counties (right). The shading color of an entity indicates to which quintile of the sample it belongs.

hypothesis H_4, which is approximately three times as unlikely as hypothesis H_3. Thus, in line with the results from classical maximum likelihood-based inference, there is very strong evidence for a positive network effect. Contrary to maximum likelihood-based inference, however, the Bayesian approach also allows one to draw conclusions as to how much support there is in the data for particular values for ρ. In this data set, evidence is strong that the network effect resides between .25 and .5, a conclusion that is neither affected by the choice of the prior for ρ nor by the choice of the prior model probabilities. Ultimately, among these Columbus neighborhoods, there is the most evidence for medium autocorrelation (between .25 and .5) with respect to crime rates. Table 3.3 summarizes the findings.

3.6.2 Threatened birds data

In the following example, McPherson & Nieswiadomy (2005) studied the percentage of threatened birds in 113 countries around the globe in the year 2000 via the network autocorrelation model, considering that "threats to a species in one country may spill over to neighboring countries' species" (McPherson & Nieswiadomy, 2005, p.401). In this data set, the spatial connectivity matrix was based on the shared border length between two neighboring countries, which made several island countries isolates in the network.[13] In addition, the authors included 10, mainly socio-economic, explanatory variables in the analysis, plus an intercept term. The resulting maximum likelihood estimate of the network autocorrelation was $\hat{\rho}_{\text{ML}} = .16$, with a 95% confidence interval for ρ of $(-.06, .37)$, and $p = .08$, so "in other words, threats to birds spill over into adjoining countries" (McPherson & Nieswiadomy, 2005, p.405).

[13] All raw connectivity matrices in the examples were subsequently row-normalized by the authors.

Table 3.3 Bayes factors BF_{t0}^{E} and BF_{t0}^{U}, $t \in \{0,1,2,3,4\}$, posterior model probabilities $p_{\mathrm{E}}\left(H_t|\boldsymbol{y}\right)$ and $p_{\mathrm{U}}\left(H_t|\boldsymbol{y}\right)$ for the hypotheses $H_0 : \rho = 0$, $H_1 : -.25 < \rho < 0$, $H_2 : 0 < \rho \leq .25$, $H_3 : .25 < \rho \leq .5$, and $H_4 : .5 < \rho < 1$, and maximum likelihood estimates $\hat{\rho}_{\mathrm{ML}}$ of ρ and corresponding p-values for the crime data set, the threatened birds data set, and the voting data set.

	Crime data					Threatened birds data					Voting data				
	H_0	H_1	H_2	H_3	H_4	H_0	H_1	H_2	H_3	H_4	H_0	H_1	H_2	H_3	H_4
BF_{t0}^{E}	1	.4	12.5	42.3	16.3	1	.5	2.0	.5	0.0	1	0.0	$> 10^6$	$> 10^6$	$> 10^6$
BF_{t0}^{U}	1	.2	9.4	42.0	6.4	1	.3	1.9	.5	0.0	1	0.0	$> 10^6$	$> 10^6$	$> 10^6$
$p_{\mathrm{E}}\left(H_t\|\boldsymbol{y}\right)$	.01	.00	.11	.74	.14	.26	.01	.50	.23	.00	.00	.00	.00	.00	1
$p_{\mathrm{U}}\left(H_t\|\boldsymbol{y}\right)$	.02	.00	.14	.64	.20	.32	.07	.49	.12	.00	.00	.00	.00	.00	1
ML-based	$\hat{\rho}_{\mathrm{ML}} = .40, p = .0008$					$\hat{\rho}_{\mathrm{ML}} = .16, p = .08$					$\hat{\rho}_{\mathrm{ML}} = .61, p < 10^{-6}$				

Based on the empirical prior for ρ, the data yield posterior model probabilities of $p_{\mathrm{E}}\left(H_0|\boldsymbol{y}\right) = .26$, $p_{\mathrm{E}}\left(H_1|\boldsymbol{y}\right) = .01$, $p_{\mathrm{E}}\left(H_2|\boldsymbol{y}\right) = .50$, $p_{\mathrm{E}}\left(H_3|\boldsymbol{y}\right) = .23$, and $p_{\mathrm{E}}\left(H_4|\boldsymbol{y}\right) = .00$. Hence, the hypothesis that there is a minor spillover effect of threats to birds across adjoining countries, i.e., $H_2 : 0 < \rho \leq .25$, is most supported by the data and consequently results in the highest Bayes factor, see Table 3.3. However, the Bayes factor of a minor spillover effect against no spillover effect is only 2.0, so the support in the data for hypothesis H_2 is far from decisive. In case of considering a uniform for ρ, the resulting posterior probabilities for the hypotheses are similar and given by $p_{\mathrm{U}}\left(H_0|\boldsymbol{y}\right) = .32$, $p_{\mathrm{U}}\left(H_1|\boldsymbol{y}\right) = .07$, $p_{\mathrm{U}}\left(H_2|\boldsymbol{y}\right) = .49$, $p_{\mathrm{U}}\left(H_3|\boldsymbol{y}\right) = .12$, and $p_{\mathrm{U}}\left(H_4|\boldsymbol{y}\right) = .00$. Again, none of the alternative hypotheses receives convincing evidence to outweigh the null. Overall, the data provide most evidence for ρ being between 0 and .25 but the Bayes factors also show that the strength of this evidence is rather small. These findings vividly illustrate the well-known issue that p-values tend to overestimate the evidence against the null (Berger & Sellke, 1987; Jeffreys, 1961; Rouder et al., 2009; Sellke et al., 2001; Wagenmakers, 2007).

3.6.3 Voting data

The voting data set contains voter turnout in the 1980 U.S. presidential election from 3,107 U.S. counties. Pace & Barry (1997) employed a Spatial Durbin model, a variant of the network autocorrelation model given by $\boldsymbol{y} = \rho W\boldsymbol{y} + \alpha\mathbf{1} + X\boldsymbol{\beta} + WX\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$ (where α denotes the model's intercept term, $\mathbf{1}$ a vector of ones, and $\boldsymbol{\gamma}$ another vector of regression coefficients), to analyze the logarithm of the voter turnout (TURNOUT) across these U.S. counties.[14] The spatial distribution of the logarithm of the voter turnout is shown in Figure 3.6 (right).[15] As explanatory variables, the authors used the logarithm of, first, the population of 18 years of age or older eligible to vote in each county (POP); second, the population of 25 years of age or older with a 12-th grade or higher education in each county (EDUCATION); third, the number of owner-occupied housing units in each county (HOUSES); fourth, the aggregate income of each county (INCOME), as well as the

[14] Note that a Spatial Durbin model can be represented as a network autocorrelation (3.1) by rewriting $\alpha\mathbf{1} + X\boldsymbol{\beta} + WX\boldsymbol{\gamma}$ as $\tilde{X}\tilde{\boldsymbol{\beta}}$, where $\tilde{X} := (\mathbf{1}, X, WX)$ and $\tilde{\boldsymbol{\beta}} := (\alpha, \boldsymbol{\beta}, \boldsymbol{\gamma})$.

[15] We created the U.S. county map by using the map() function from the maps package in R (Becker et al., 2016). The depicted county map does not include several very small counties in Virginia, which is why there are data for only 3,076, instead of the full 3,107, counties displayed.

four corresponding spatially lagged variables. Furthermore, the connectivity matrix W in this example was constructed on the basis of the four nearest neighbors of each county.[16] Except for POP and the spatially lagged HOUSES and INCOME variables, all other predictors were found to have a positive impact on TURNOUT, with a maximum likelihood estimate of $\hat{\rho}_{\mathrm{ML}} = .62$ (Pace & Barry, 1997, p.242), an associated 95% confidence interval for ρ of $(.58, .65)$, and $p < 10^{-6}$.

Computing the corresponding posterior model probabilities underpins the decisive evidence for a large network effect, as $p_{\mathrm{E}}\left(H_0|\boldsymbol{y}\right) = p_{\mathrm{E}}\left(H_1|\boldsymbol{y}\right) = p_{\mathrm{E}}\left(H_2|\boldsymbol{y}\right) = p_{\mathrm{E}}\left(H_3|\boldsymbol{y}\right) = .00$ and $p_{\mathrm{E}}\left(H_4|\boldsymbol{y}\right) = 1$. Accordingly, the concomitant Bayes factor of hypothesis H_4 against any other considered hypothesis exceeds 10^6, which provides "decisive" evidence in the data in favor of hypothesis H_4 compared to hypothesis H_0, H_1, H_2, and H_3, respectively. When using the uniform prior, the posterior model probabilities remain essentially unchanged, as the evidence in the data for a large network effect is conclusive in this example. Although the implications of the Bayesian approach seem in line with the traditional approach, this is not the case completely: the only thing we can deduce from classical null hypothesis significance testing here is that we can reject the null hypothesis that $\rho = 0$. On the other hand, the Bayesian approach gives us much more detail about which values for ρ are most supported by the data and quantifies to what extent.

3.7 Conclusions

In this chapter, we developed three Bayes factors for testing precise and interval hypotheses on the network effect in the network autocorrelation model. The Bayesian approach to these tests comes with several practical advantages compared to classical null hypothesis significance testing. For example, the Bayes factors and the resulting posterior model probabilities allow us to quantify the amount of evidence for a precise null hypothesis, or any other hypothesis, and they allow us to test multiple precise and interval hypotheses simultaneously without any of the drawbacks of classical null hypothesis significance testing.

We ran an extensive simulation study to evaluate the numerical behavior of the presented Bayes factors for a wide range of network configurations. We found that the Bayes factor based on an empirical prior for the network effect, relying on a summary of published network autocorrelations from many different sources, is always consistent, displays superior performance, and is the Bayes factor we recommend. At the same time, using a uniform prior for ρ instead yields properties that are almost as good as those based on the empirical prior. Finally, we do not recommend employing improper priors for ρ in combination with the fractional Bayes factor methodology. We illustrated the practical use of the recommended Bayes factors with three examples and provided computer code in R, making the proposed Bayes factors easily available to researchers interested in testing for the existence and magnitude of a network effect in the network autocorrelation model.

[16] Data on the dependent and the independent variables were taken from the 1980 U.S. Census and are available along with a sparse matrix representation of W from the Spatial Econometrics Toolbox for Matlab at `http://spatial-econometrics.com/html/jplv7.zip`, files *elect.data* and *elect.ford*.

Given the importance of the network autocorrelation model in a variety of fields, we believe there is much value to having available an approach that makes it possible for researchers to test hypotheses that go beyond the standard significance test $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$, which is only useful for falsifying the null. The new Bayesian tests provide means for quantifying evidence in favor of any hypothesis and enable researchers to test multiple hypotheses against one another in a single analysis including, but not restricted to, any combination of precise and interval hypotheses. Overall, we hope that these tools will enrich the toolkit of researchers studying network effects through the network autocorrelation model.

Acknowledgment

We thank Michael A. McPherson and Michael L. Nieswiadomy for sharing their data.

Appendix 3.A Calculating Bayes factors using R

```
## Function to compute the logarithm of the marginal likelihood
## under a precise hypothesis rho=c and data y, X, W
lnmarglik.p <- function(y, X, W, c) {
 g <- length(y)
 k <- ncol(X)
 gminusk2 <- .5*(g - k)
 EW <- eigen(W, only.values=TRUE)$values
 XtX <- t(X) %*% X
 M <- diag(g) - X %*% solve(XtX) %*% t(X)
 yMy <- sum(c(M %*% y)**2)
 Wy <- W %*% y
 MWy <- c(M %*% Wy)
 yMWy <- sum(y*MWy)
 yWMWy <- sum(MWy**2)
 yAMAy <- c(yMy - 2*c*yMWy + c**2*yWMWy)
 lnml <- (- gminusk2*log(pi) + lgamma(gminusk2) - .5*log(det(XtX))
          + Re(sum(log(1 - c*EW))) - gminusk2*log(yAMAy))
 return(lnml)
}

## Function to compute the logarithm of the marginal likelihood under an
## interval hypothesis a1<rho<a2 for a normal prior (mean, sd)
## for rho, data y, X, W, and N grid points
lnmarglik.n <- function(N=1e3, y, X, W, mean, sd, a1, a2) {
 g <- length(y)
 k <- ncol(X)
 gminusk2 <- .5*(g - k)
 EW <- eigen(W, only.values=TRUE)$values
 XtX <- t(X) %*% X
 M <- diag(g) - X %*% solve(XtX) %*% t(X)
 yMy <- sum(c(M %*% y)**2)
 Wy <- W %*% y
 MWy <- c(M %*% Wy)
 yMWy <- sum(y*MWy)
 yWMWy <- sum(MWy**2)
 scalefac <- .999
 rhoseqh <- seq(scalefac*a1, scalefac*a2, length=100)
 yAMAyh <- c(yMy - 2*rhoseqh*yMWy + rhoseqh**2*yWMWy)
 lognormdensh <- dnorm(rhoseqh, mean=mean, sd=sd, log=T)
 inth <- NULL
 for (r in 1:100) {
  inth[r] <- Re(sum(log(1 - rhoseqh[r]*EW))) - gminusk2*log(yAMAyh[r])
             + lognormdensh[r]}
 d <- 650 - max(inth)
 rhoseq <- seq(scalefac*a1, scalefac*a2, length=2*N - 1)
 yAMAy <- c(yMy - 2*rhoseq*yMWy + rhoseq**2*yWMWy)
 spread <- (rhoseq[3] - rhoseq[1])/6
 weights <- c(1, rep(c(4, 2), .5*(2*N - 4)), c(4, 1))
 normc <- pnorm(a2, mean=mean, sd=sd) - pnorm(a1, mean=mean, sd=sd)
 lognormdens <- dnorm(rhoseq, mean=mean, sd=sd, log=T)
 int <- NULL
 for (r in 1:(2*N - 1)) {
  int[r] <- weights[r]*exp(Re(sum(log(1 - rhoseq[r]*EW)))
            - gminusk2*log(yAMAy[r]) + lognormdens[r] + d)}
 lnml <- (- gminusk2*log(pi) + lgamma(gminusk2) - .5*log(det(XtX))
          - log(normc) - d + log(spread) + log(sum(int)))
 return(lnml)
}

## Function to compute the logarithm of the marginal likelihood under an
## interval hypothesis a1<rho<a2 for a uniform prior for rho,
## data y, X, W, and N grid points
lnmarglik.u <- function(N=1e3, y, X, W, a1, a2) {
 g <- length(y)
 k <- ncol(X)
 gminusk2 <- .5*(g - k)
 EW <- eigen(W, only.values=TRUE)$values
 XtX <- t(X) %*% X
 M <- diag(g) - X %*% solve(XtX) %*% t(X)
 yMy <- sum(c(M %*% y)**2)
 Wy <- W %*% y
 MWy <- c(M %*% Wy)
 yMWy <- sum(y*MWy)
 yWMWy <- sum(MWy**2)
 scalefac <- .999
 rhoseqh <- seq(scalefac*a1, scalefac*a2, length=100)
 yAMAyh <- c(yMy - 2*rhoseqh*yMWy + rhoseqh**2*yWMWy)
 inth <- NULL
 for (r in 1:100) {
  inth[r] <- Re(sum(log(1 - rhoseqh[r]*EW))) - gminusk2*log(yAMAyh[r])}
 d <- 650 - max(inth)
 rhoseq <- seq(scalefac*a1, scalefac*a2, length=2*N - 1)
 yAMAy <- c(yMy - 2*rhoseq*yMWy + rhoseq**2*yWMWy)
 spread <- (rhoseq[3] - rhoseq[1])/6
 weights <- c(1, rep(c(4, 2), .5*(2*N - 4)), c(4, 1))
 int <- NULL
 for (r in 1:(2*N - 1)) {
  int[r] <- weights[r]*exp(Re(sum(1 - rhoseq[r]*EW))
            - gminusk2*log(yAMAy[r]) + d)}
 lnml <- (- gminusk2*log(pi) + lgamma(gminusk2) - .5*log(det(XtX))
          - log(a2 - a1) - d + log(spread) + log(sum(int)))
 return(lnml)
}
```

For all of the scenarios considered in this chapter, using $N = 1,000$ equally spaced grid points for ρ proved to be more than sufficient to obtain reliable results for the marginal likelihoods given in (3.7) and (3.8). In addition, we also compared our numerical integration scheme to an importance sampling procedure (A. Owen & Zhou, 2000). As to that, we approximated $\log\left(|A_\rho|\right)$ and $\log\left(\boldsymbol{y}^T A_\rho^T M A_\rho \boldsymbol{y}\right)$ by second-order Taylor polynomials at their maximum values, $\rho = 0$ and $\rho = \boldsymbol{y}^T MW\boldsymbol{y}/\boldsymbol{y}W^T MW\boldsymbol{y}$, respectively. This results in normal approximations of $|A_\rho|$ and $\boldsymbol{y}^T A_\rho^T M A_\rho \boldsymbol{y}^{-(g-k)/2}$. Hence, the overall expressions in (3.7) and (3.8) can be approximated by normal distributions, which we used as importance sampling distributions. R code therefor is available from the authors upon request. The two methods give virtually identical results and we thank an anonymous reviewer for this suggestion.

We provide code for computing the logarithms of the marginal likelihoods only, as calculating the marginal likelihoods themselves directly might result in underflow in R. Obtaining the Bayes factor from two logarithms of the marginal likelihoods is straightforward, while posterior probabilities for the hypotheses can be calculated via

$$
\begin{aligned}
p\left(H_t|\boldsymbol{y}\right) &= \frac{p\left(\boldsymbol{y}|H_t\right)p\left(H_t\right)}{\sum\limits_{t'=0}^{T-1} p\left(\boldsymbol{y}|H_{t'}\right)p\left(H_{t'}\right)} \\
&= \frac{\exp\left(\log\left(p\left(\boldsymbol{y}|H_t\right)\right) + \log\left(p\left(H_t\right)\right)\right)}{\sum\limits_{t'=0}^{T-1} \exp\left(\log\left(p\left(\boldsymbol{y}|H_{t'}\right)\right) + \log\left(p\left(H_{t'}\right)\right)\right)} \\
&= \frac{\exp\left(\log\left(p\left(\boldsymbol{y}|H_t\right)\right) + \log\left(p\left(H_t\right)\right) + d\right)}{\sum\limits_{t'=0}^{T-1} \exp\left(\log\left(p\left(\boldsymbol{y}|H_{t'}\right)\right) + \log\left(p\left(H_{t'}\right)\right) + d\right)},
\end{aligned}
$$

where an auxiliary constant d, e.g., $d = 650 - \max\limits_{t\in\{0,1,\ldots,T-1\}} \log\left(p\left(\boldsymbol{y}|H_t\right)\right)$, might be added in case that the marginal likelihoods are too small to be distinguished from zero in R.

```
## Run the script below to compute posterior model probabilities and
## Bayes factors based on the empirical prior for rho
## for the crime data set (Anselin, 1988)
install.packages("spdep"); library(spdep); data(columbus)
priormean <- .36; priorsd <- .19
nc <- pnorm(1, mean=priormean, sd=priorsd)
      - pnorm(-.25, mean=priormean, sd=priorsd)
prior.H0 <- .2
prior.H1 <- .8*(pnorm(0, mean=priormean, sd=priorsd)
                - pnorm(-.25, mean=priormean, sd=priorsd))/nc
prior.H2 <- .8*(pnorm(.25, mean=priormean, sd=priorsd)
                - pnorm(0, mean=priormean, sd=priorsd))/nc
prior.H3 <- .8*(pnorm(.5, mean=priormean, sd=priorsd)
                - pnorm(.25, mean=priormean, sd=priorsd))/nc
prior.H4 <- .8*(pnorm(1, mean=priormean, sd=priorsd)
                - pnorm(.5, mean=priormean, sd=priorsd))/nc
prior.H <- c(prior.H0, prior.H1, prior.H2, prior.H3, prior.H4)
W.crime.list <- nb2listw(col.gal.nb)
crime.ml <- lagsarlm(CRIME~INC + HOVAL, data=columbus, listw=W.crime.list)
summary(crime.ml) # maximum likelihood estimates
W.crime <- nb2mat(col.gal.nb)
X.crime <- cbind(rep(1, nrow(W.crime)), columbus$INC, columbus$HOVAL)
y.crime <- columbus$CRIME
lnmarglik.H0.crime <- lnmarglik.p(y=y.crime, X=X.crime, W=W.crime, c=0)
lnmarglik.H1.crime <- lnmarglik.n(N=1e3, y=y.crime, X=X.crime, W=W.crime,
                                mean=priormean, sd=priorsd, a1=-.25, a2=0)
lnmarglik.H2.crime <- lnmarglik.n(N=1e3, y=y.crime, X=X.crime, W=W.crime,
                                mean=priormean, sd=priorsd, a1=0, a2=.25)
lnmarglik.H3.crime <- lnmarglik.n(N=1e3, y=y.crime, X=X.crime, W=W.crime,
                                mean=priormean, sd=priorsd, a1=.25, a2=.5)
lnmarglik.H4.crime <- lnmarglik.n(N=1e3, y=y.crime, X=X.crime, W=W.crime,
                                mean=priormean, sd=priorsd, a1=.5, a2=1)
ln.marglik.crime <- c(lnmarglik.H0.crime, lnmarglik.H1.crime,
               lnmarglik.H2.crime, lnmarglik.H3.crime, lnmarglik.H4.crime)
exp(ln.marglik.crime-lnmarglik.H0.crime) # Bayes factors
(exp(ln.marglik.crime + log(prior.H))
/sum(exp(ln.marglik.crime + log(prior.H)))) #posterior model probabilities
```

Appendix 3.B Auxiliary facts

(1) Let A and B be symmetric and positive semi-definite matrices. Then (see e.g., X. Yang, 2000, Lemma 1),

$$0 \leq \operatorname{tr}(AB) \leq \operatorname{tr}(A)\operatorname{tr}(B).$$

(2) Let A and B be matrices. Then,

$$\operatorname{tr}(AB) \leq \frac{1}{2}\left(\operatorname{tr}\left(A^2\right) + \operatorname{tr}\left(B^2\right)\right).$$

(3) Let $A_\rho = I_g - \rho W$. If $\left(X^T X\right)^{-1}$ exists, then

$$\begin{aligned}
&\left(A_\rho \boldsymbol{y} - X\boldsymbol{\beta}\right)^T \left(A_\rho \boldsymbol{y} - X\boldsymbol{\beta}\right) \\
&= \left(A_\rho \boldsymbol{y} - X\hat{\boldsymbol{\beta}}\right)^T \left(A_\rho \boldsymbol{y} - X\hat{\boldsymbol{\beta}}\right) + \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)^T \left(X^T X\right)\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right) \\
&= \boldsymbol{y}^T A_\rho^T M A_\rho \boldsymbol{y} + \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)^T \left(X^T X\right)\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right),
\end{aligned}$$

where $\hat{\boldsymbol{\beta}} := \left(X^T X\right)^{-1} X^T A_\rho \boldsymbol{y}$ and $M = I_g - X\left(X^T X\right)^{-1} X^T$.

Appendix 3.C Asymptotic prior behavior

We need to show that the proposed improper prior for ρ, $p_{\text{NI}}\left(\rho | H_{\text{interval}}\right) = (1-\rho)^{-1}\,\mathbb{1}_{(0,1)}(\rho)$, is asymptotically of the same order as the conditional Independence Jeffreys prior for ρ for $\rho \to 1$. In particular, it is to show that

(i) $p_{\text{IJ}}\left(\rho | \sigma^2 = \sigma_1^2, \boldsymbol{\beta} = \boldsymbol{\beta}_1\right) = \mathcal{O}\left((1-\rho)^{-1}\right)$,

(ii) $(1-\rho)^{-1} = \mathcal{O}\left(p_{\text{IJ}}\left(\rho | \sigma^2 = \sigma_1^2, \boldsymbol{\beta} = \boldsymbol{\beta}_1\right)\right)$,

where $p_{\text{IJ}}\left(\rho | \sigma^2 = \sigma_1^2, \boldsymbol{\beta} = \boldsymbol{\beta}_1\right)$ denotes the model's conditional Independence Jeffreys prior for ρ and $\sigma_1^2 \in \mathbb{R}^+$ and $\boldsymbol{\beta}_1 \in \mathbb{R}^k$ are constants.

Proof.

(i) Applying auxiliary facts (1), (2), and using the functional form of the model's Independence Jeffreys prior given in Dittrich et al. (2017) yields

$$\begin{aligned}
&p_{\text{IJ}}\left(\rho | \sigma^2 = \sigma_1^2, \boldsymbol{\beta} = \boldsymbol{\beta}_1\right) \\
&\propto \frac{1}{\sigma_1^2}\left\{\operatorname{tr}(B_\rho^T B_\rho) + \operatorname{tr}\left(B_\rho^2\right) + \frac{1}{\sigma_1^2}\boldsymbol{\beta}_1^T X^T B_\rho^T B_\rho X \boldsymbol{\beta}_1 - \frac{2}{g}\operatorname{tr}^2\left(B_\rho\right)\right\}^{\frac{1}{2}} \\
&\propto \left\{\operatorname{tr}(B_\rho^T B_\rho) + \operatorname{tr}\left(B_\rho^2\right) + \frac{1}{\sigma_1^2}\boldsymbol{\beta}_1^T X^T B_\rho^T B_\rho X \boldsymbol{\beta}_1 - \frac{2}{g}\operatorname{tr}^2\left(B_\rho\right)\right\}^{\frac{1}{2}} \\
&\leq \left\{\operatorname{tr}(B_\rho^T B_\rho) + \operatorname{tr}\left(B_\rho^2\right) + \frac{1}{\sigma_1^2}\boldsymbol{\beta}_1^T X^T B_\rho^T B_\rho X \boldsymbol{\beta}_1\right\}^{\frac{1}{2}}
\end{aligned}$$

$$
\begin{aligned}
&= \left\{\operatorname{tr}(B_\rho^T B_\rho) + \operatorname{tr}\left(B_\rho^2\right) + \frac{1}{\sigma_1^2}\operatorname{tr}\left(X\boldsymbol{\beta}_1\boldsymbol{\beta}_1^T X^T B_\rho^T B_\rho\right)\right\}^{\frac{1}{2}} \\
&\leq \left\{\frac{1}{2}\left(\operatorname{tr}(B_\rho^T B_\rho^T) + \operatorname{tr}\left(B_\rho^2\right)\right) + \operatorname{tr}\left(B_\rho^2\right)\right. \\
&\qquad \left. + \frac{1}{\sigma_1^2}\operatorname{tr}\left(X\boldsymbol{\beta}_1\boldsymbol{\beta}_1^T X^T\right)\frac{1}{2}\left(\operatorname{tr}(B_\rho^T B_\rho^T) + \operatorname{tr}\left(B_\rho^2\right)\right)\right\}^{\frac{1}{2}} \\
&\propto \operatorname{tr}\left(B_\rho^2\right)^{\frac{1}{2}},
\end{aligned}
$$

where $B_\rho := WA_\rho^{-1}$. Without loss of generality, we assume throughout the remainder of the proof that the multiplicity of the largest eigenvalue of W, $\lambda_1 = 1$, is one. Then, there exists $\rho^{'}$ such that $(1-\rho)^{-2} > \sum_{i=2}^{g} \frac{\lambda_i^2}{(1-\rho\lambda_i)^2}$ for all $\rho \in \left(\rho^{'}, 1\right)$. As the eigenvalues of B_ρ^2 are given by $\left(\lambda_i^2/\left(1-\rho\lambda_i\right)^2\right)_i, i \in \{1, ..., g\}$, it holds for all $\rho \in \left(\rho^{'}, 1\right)$ that

$$
\operatorname{tr}\left(B_\rho^2\right)^{\frac{1}{2}} = \left\{\sum_{i=1}^{g} \frac{\lambda_i^2}{\left(1-\rho\lambda_i\right)^2}\right\}^{\frac{1}{2}} < \left\{\frac{2}{\left(1-\rho\right)^2}\right\}^{\frac{1}{2}} \propto \frac{1}{1-\rho}.
$$

(ii) Note that

$$
\operatorname{tr}^2\left(B_\rho\right) = \left\{\sum_{i=1}^{g} \frac{\lambda_i}{1-\rho\lambda_i}\right\}^2 = \sum_{i=1}^{g} \frac{\lambda_i^2}{\left(1-\rho\lambda_i\right)^2} + \sum_{i,j=1, i\neq j}^{g} \frac{\lambda_i\lambda_j}{\left(1-\rho\lambda_i\right)\left(1-\rho\lambda_j\right)}.
$$

Consequently, there exists $\rho^{''}$ such that $\operatorname{tr}^2\left(B_\rho\right) \leq 2\left(1-\rho\right)^{-2}$ for all $\rho \in \left(\rho^{''}, 1\right)$. Furthermore, it holds that $\operatorname{tr}\left(B_\rho^2\right) = \sum_{i=1}^{g} \frac{\lambda_i^2}{(1-\rho\lambda_i)^2} \geq (1-\rho)^{-2}$. Thus, we can write for all $\rho \in \left(\rho^{''}, 1\right)$

$$
\begin{aligned}
& p_{\mathrm{IJ}}\left(\rho | \sigma^2 = \sigma_1^2, \boldsymbol{\beta} = \boldsymbol{\beta}_1\right) \\
&\propto \frac{1}{\sigma_1^2}\left\{\operatorname{tr}(B_\rho^T B_\rho) + \operatorname{tr}\left(B_\rho^2\right) + \frac{1}{\sigma_1^2}\boldsymbol{\beta}_1^T X^T B_\rho^T B_\rho X\boldsymbol{\beta}_1 - \frac{2}{g}\operatorname{tr}^2\left(B_\rho\right)\right\}^{\frac{1}{2}} \\
&\propto \left\{\operatorname{tr}(B_\rho^T B_\rho) + \operatorname{tr}\left(B_\rho^2\right) + \frac{1}{\sigma_1^2}\boldsymbol{\beta}_1^T X^T B_\rho^T B_\rho X\boldsymbol{\beta}_1 - \frac{2}{g}\operatorname{tr}^2\left(B_\rho\right)\right\}^{\frac{1}{2}} \\
&\geq \left\{\operatorname{tr}\left(B_\rho^2\right) - \frac{2}{g}\operatorname{tr}^2\left(B_\rho\right)\right\}^{\frac{1}{2}} \\
&\geq \left\{\frac{1}{\left(1-\rho\right)^2} - \frac{2}{g}\frac{2}{\left(1-\rho\right)^2}\right\}^{\frac{1}{2}} \\
&\propto \frac{1}{1-\rho},
\end{aligned}
$$

which completes the proof. ■

Appendix 3.D Minimum bound for b in the fractional Bayes factor approach

When using the proposed improper prior for ρ, $p_{\mathrm{NI}}\left(\rho|H_{\mathrm{interval}}\right) = (1-\rho)^{-1}\,\mathbb{1}_{(0,1)}\left(\rho\right)$, in combination with the standard non-informative prior for the remaining nuisance parameters, $p\left(\sigma^2, \boldsymbol{\beta}\right) \propto 1/\sigma^2$, we can write for the the resulting updated prior in (3.9) under H_{interval}

$$
\begin{aligned}
& p\left(\rho, \sigma^2, \boldsymbol{\beta}|H_{\mathrm{interval}}, \boldsymbol{y}^b\right) \propto f\left(\boldsymbol{y}|\rho, \sigma^2, \boldsymbol{\beta}\right)^b p_{\mathrm{NI}}\left(\rho, \sigma^2, \boldsymbol{\beta}|H_{\mathrm{interval}}\right) \\
& \propto (1-\rho)^{-1}\left|A_\rho\right|^b\left(\sigma^2\right)^{-\frac{gb}{2}-1} \exp\left(-\frac{b}{2\sigma^2}\left(A_\rho \boldsymbol{y} - X\boldsymbol{\beta}\right)^T\left(A_\rho \boldsymbol{y} - X\boldsymbol{\beta}\right)\right).
\end{aligned}
$$

Integrating over $\boldsymbol{\beta}$ and using auxiliary fact (3) yields

$$
\begin{aligned}
& \int_{\mathbb{R}^k} p\left(\rho, \sigma^2, \boldsymbol{\beta}|H_{\mathrm{interval}}, \boldsymbol{y}^b\right) \mathrm{d}\boldsymbol{\beta} \\
& \propto \int_{\mathbb{R}^k} (1-\rho)^{-1}\left|A_\rho\right|^b\left(\sigma^2\right)^{-\frac{gb}{2}-1} \exp\left(-\frac{b}{2\sigma^2}\left(A_\rho \boldsymbol{y} - X\boldsymbol{\beta}\right)^T\left(A_\rho \boldsymbol{y} - X\boldsymbol{\beta}\right)\right) \mathrm{d}\boldsymbol{\beta} \\
& = (1-\rho)^{-1}\left|A_\rho\right|^b\left(\sigma^2\right)^{-\frac{gb}{2}-1} \int_{\mathbb{R}^k} \exp\left(-\frac{b}{2\sigma^2}\left(A_\rho \boldsymbol{y} - X\boldsymbol{\beta}\right)^T\left(A_\rho \boldsymbol{y} - X\boldsymbol{\beta}\right)\right) \mathrm{d}\boldsymbol{\beta} \\
& = (1-\rho)^{-1}\left|A_\rho\right|^b\left(\sigma^2\right)^{-\frac{gb}{2}-1} \exp\left(-\frac{b}{2\sigma^2}\boldsymbol{y}^T A_\rho^T M A_\rho \boldsymbol{y}\right) \\
& \quad \int_{\mathbb{R}^k} \exp\left(-\frac{b}{2\sigma^2}\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)^T\left(X^T X\right)\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)\right) \mathrm{d}\boldsymbol{\beta}.
\end{aligned} \tag{3.13}
$$

If $\left(X^T X\right)^{-1}$ exists, the integrand in (3.13) is the kernel of the probability density function of a multivariate normal random variable $\boldsymbol{Z} \sim N\left(\hat{\boldsymbol{\beta}}, \frac{\sigma^2}{b}\left(X^T X\right)^{-1}\right)$. Thus, it follows that

$$
\int_{\mathbb{R}^k} \exp\left(-\frac{b}{2\sigma^2}\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)^T\left(X^T X\right)\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)\right) \mathrm{d}\boldsymbol{\beta} \propto \left(\sigma^2\right)^{\frac{k}{2}}.
$$

Hence, overall

$$
\begin{aligned}
& \int_{\mathbb{R}^k} p\left(\rho, \sigma^2, \boldsymbol{\beta}|H_{\mathrm{interval}}, \boldsymbol{y}^b\right) \mathrm{d}\boldsymbol{\beta} \\
& \propto (1-\rho)^{-1}\left|A_\rho\right|^b\left(\sigma^2\right)^{-\frac{gb}{2}-1} \exp\left(-\frac{b}{2\sigma^2}\boldsymbol{y}^T A_\rho^T M A_\rho \boldsymbol{y}\right) \\
& \quad \int_{\mathbb{R}^k} \exp\left(-\frac{b}{2\sigma^2}\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)^T\left(X^T X\right)\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)\right) \mathrm{d}\boldsymbol{\beta} \\
& \propto (1-\rho)^{-1}\left|A_\rho\right|^b\left(\sigma^2\right)^{-\frac{gb-k}{2}-1} \exp\left(-\frac{b}{2\sigma^2}\boldsymbol{y}^T A_\rho^T M A_\rho \boldsymbol{y}\right).
\end{aligned} \tag{3.14}
$$

Next, observe that the terms in (3.14) involving σ^2 correspond to the kernel of the probability density function of an inverse gamma distributed random variable when $gb > k$, i.e., when $b > k/g$. In this case, integrating over σ^2 gives

$$\begin{aligned}&\int_0^\infty (1-\rho)^{-1} |A_\rho|^b \left(\sigma^2\right)^{-\frac{gb-k}{2}-1} \exp\left(-\frac{b}{2\sigma^2}\boldsymbol{y}^T A_\rho^T M A_\rho \boldsymbol{y}\right) \mathrm{d}\sigma^2 \\ &\propto (1-\rho)^{-1} |A_\rho|^b \boldsymbol{y}^T A_\rho^T M A_\rho \boldsymbol{y}^{-\frac{gb-k}{2}}. \end{aligned} \tag{3.15}$$

As the last term in (3.15) is a quadratic polynomial in ρ, it is bounded for $\rho \in (0,1)$. Thus, it only remains to show that

$$\int_0^1 (1-\rho)^{-1} |A_\rho|^b \mathrm{d}\rho = \int_0^1 (1-\rho)^{-1} (1-\rho)^b \prod_{i=2}^g (1-\rho\lambda_i)^b \,\mathrm{d}\rho < \infty. \tag{3.16}$$

Similarly to before, the last product in (3.16) is bounded for $\rho \in (0,1)$, so it suffices to check that

$$\int_0^1 (1-\rho)^{-1} (1-\rho)^b \,\mathrm{d}\rho = \int_0^1 (1-\rho)^{b-1} \,\mathrm{d}\rho = \frac{1}{b}\left[(1-\rho)^b\right]_0^1 < \infty,$$

which proves the statement.

Appendix 3.E Fractional Bayes factor computation

Computing the integral in the numerator of (3.12) can be done using standard numerical techniques. On the other hand, evaluating the integral in the denominator of (3.12) directly is numerically unstable, as the integrand approaches infinity at the upper bound. Integration by parts provides a simple solution that results in a new smooth integrand and is presented in the following. Without loss of generality, we assume that the multiplicity of the largest eigenvalue of W, $\lambda_1 = 1$, is one. Then, the integral in the denominator of (3.12) can be written as

$$\begin{aligned}&\int_0^1 (1-\rho)^{-1} |A_\rho|^b \boldsymbol{y}^T A_\rho^T M A_\rho \boldsymbol{y}^{-\frac{gb-k}{2}} \mathrm{d}\rho \\ &= \int_0^1 (1-\rho)^{-1} (1-\rho)^b \prod_{i=2}^g (1-\rho\lambda_i)^b \,\boldsymbol{y}^T A_\rho^T M A_\rho \boldsymbol{y}^{-\frac{gb-k}{2}} \mathrm{d}\rho \\ &= \int_0^1 (1-\rho)^{b-1} \prod_{i=2}^g (1-\rho\lambda_i)^b \,\boldsymbol{y}^T A_\rho^T M A_\rho \boldsymbol{y}^{-\frac{gb-k}{2}} \mathrm{d}\rho \\ &= \int_0^1 (1-\rho)^{b-1} H(\rho)\,\mathrm{d}\rho \\ &= \left[-\frac{1}{b}(1-\rho)^b H(\rho)\right]_0^1 - \int_0^1 -\frac{1}{b}(1-\rho)^b H'(\rho)\,\mathrm{d}\rho \end{aligned} \tag{3.17}$$

$$= \frac{1}{b}\left(\boldsymbol{y}^T M \boldsymbol{y}^{-\frac{gb-k}{2}} + \int_0^1 (1-\rho)^b\, H'(\rho)\, \mathrm{d}\rho\right),$$

where (3.17) follows from integration by parts and $H(\rho) := \prod_{i=2}^{g} (1-\rho\lambda_i)^b\, \boldsymbol{y}^T A_\rho^T M A_\rho \boldsymbol{y}^{-\frac{gb-k}{2}}$. After some algebraic manipulation, we can write

$$\begin{aligned} H'(\rho) &= -H(\rho)\left(b\sum_{i=2}^{g}\frac{\lambda_i}{1-\rho\lambda_i} + \frac{gb-k}{\boldsymbol{y}^T A_\rho^T M A_\rho \boldsymbol{y}}\left(\rho\boldsymbol{y}^T W^T M W\boldsymbol{y} - \boldsymbol{y}^T M W\boldsymbol{y}\right)\right) \\ &= H(\rho)\, h(\rho), \end{aligned}$$

with $h(\rho) := -\left(b\sum_{i=2}^{g}\frac{\lambda_i}{1-\rho\lambda_i} + \frac{gb-k}{\boldsymbol{y}^T A_\rho^T M A_\rho \boldsymbol{y}}\left(\rho\boldsymbol{y}^T W^T M W\boldsymbol{y} - \boldsymbol{y}^T M W\boldsymbol{y}\right)\right)$. Hence, overall it holds that

$$\begin{aligned} &\int_0^1 (1-\rho)^{-1}\, |A_\rho|^b \boldsymbol{y}^T A_\rho^T M A_\rho \boldsymbol{y}^{-\frac{gb-k}{2}}\,\mathrm{d}\rho \\ &= \frac{1}{b}\left(\boldsymbol{y}^T M \boldsymbol{y}^{-\frac{gb-k}{2}} + \int_0^1 (1-\rho)^b\, H'(\rho)\,\mathrm{d}\rho\right) \\ &= \frac{1}{b}\left(\boldsymbol{y}^T M \boldsymbol{y}^{-\frac{gb-k}{2}} + \int_0^1 (1-\rho)^b\, H(\rho)\, h(\rho)\,\mathrm{d}\rho\right) \\ &= \frac{1}{b}\left(\boldsymbol{y}^T M \boldsymbol{y}^{-\frac{gb-k}{2}} + \int_0^1 |A_\rho|^b \boldsymbol{y}^T A_\rho^T M A_\rho \boldsymbol{y}^{-\frac{gb-k}{2}} h(\rho)\,\mathrm{d}\rho\right), \end{aligned} \tag{3.18}$$

where the remaining integral in (3.18) is smooth and can be well-approximated by standard schemes such as Simpson's rule.

Chapter 4

Bayesian analysis of higher-order network autocorrelation models

Abstract

The network autocorrelation model has been the workhorse for estimating and testing the strength of theories of social influence in a network. In many network studies, different types of social influence are present simultaneously and can be modeled using various connectivity matrices. Often, researchers have expectations about the order of strength of these different influence mechanisms. However, currently available methods cannot be applied to test a specific order of social influence in a network. In this chapter, we first present flexible Bayesian techniques for estimating network autocorrelation models with multiple network autocorrelation parameters. Second, we develop new Bayes factors that allow researchers to test hypotheses with order constraints on the network autocorrelation parameters in a direct manner. Concomitantly, we give efficient algorithms for sampling from the posterior distributions and for computing the Bayes factors. Simulation results suggest that frequentist properties of Bayesian estimators based on non-informative priors for the network autocorrelation parameters are overall slightly superior to those based on maximum likelihood estimation. Furthermore, when testing statistical hypotheses, the Bayes factors show consistent behavior with evidence for a true data-generating hypothesis increasing with the sample size. Finally, we illustrate our methods using a data set from the economic growth theory.

This chapter is under review at *Sociological Methodology* (revise and resubmit) as: Dittrich, D., Leenders, R.Th.A.J., and Mulder, J. Network autocorrelation modeling: Bayesian techniques for estimating and testing multiple network autocorrelations.

4.1 Introduction

Social network research plays an important role in understanding how persons, organizations, or countries influence each other's behavior, decision-making, or well-being. The network autocorrelation model (Ord, 1975; Doreian, 1981) has been the workhorse for estimating and testing the strength of social influence in a given network (Fujimoto et al., 2011). In the network autocorrelation model, actors' behavior, opinions, or well-beings are assumed to be correlated and a *network autocorrelation parameter* ρ is estimated, representing the strength of a social influence mechanism in the network. The network autocorrelation model has been used to analyze network influence on individual behavior across many different fields, such as criminology (Tita & Radil, 2011), ecology (McPherson & Nieswiadomy, 2005), economics (Kalenkoski & Lacombe, 2008), geography (Mur et al., 2008), organization studies (Mizruchi & Stearns, 2006), political science (Gimpel & Schuknecht, 2003), and sociology (Burt & Doreian, 1982).

While the network autocorrelation model has yielded many useful findings, the standard, or *first-order*, specification of the model implicitly assumes the presence of a single network influence mechanism on the outcome of interest only. However, this may be too restrictive in many cases, as different types of social influence are likely to be present simultaneously. For example, an actor is often a member of multiple distinct but potentially overlapping networks, such as a friendship network, a collaboration network, or an information-sharing network. Similarly, ties need not only be defined by social interaction but can also refer to geographical proximity, money flows, or joint memberships. Each of these networks may have some connection to the outcome of interest; hence, a model that ignores multiple social influence mechanisms might be overly simplistic.

Besides the fact that individuals are often members of multiple, potentially overlapping, networks, conversely it is also the case that many networks are characterized by subgroups. For example, children in school classes may belong to separate social classes and we might ask if, with respect to school performance and petty crime, children of socially disadvantaged backgrounds influence each other based on the same influence mechanism, say friendship, stronger than those of a more privileged background? Another example of grouping can be found in economic growth theory, where, with respect to economic growth, central nations are expected to be subject to other processes than peripheral developing nations (Dall'erba et al., 2009; Leenders, 1995).

The network autocorrelation model can be straightforwardly extended to include multiple influence mechanisms and different subgroups within a network (McMillen et al., 2007). Badinger & Egger (2011), Elhorst et al. (2012), Hepple (1995a), and Lee & Liu (2010) provided theoretical discussions of and estimation procedures for these so-called *higher-order* network autocorrelation models, while empirical applications can be found in e.g., Beck et al. (2006), Dall'erba et al. (2009), Lacombe (2004), McMillen et al. (2007), and Tita & Radil (2011).

In this chapter, we develop a fully Bayesian framework for estimating higher-order network autocorrelation models and for simultaneously testing multiple constraints on the relative order of network effects, such as $H_0 : \rho_1 = \rho_2 = 0$, $H_1 : \rho_1 > \rho_2 = 0$, $H_2 : \rho_1 > \rho_2 > 0$, or $H_3 : \rho_1 = \rho_2 > 0$, where ρ_1 and ρ_2 quantify the strength of different influence mechanisms, respectively. Using a Bayesian approach for estimating and testing higher-order network autocorrelation models has several advantages compared to classical methods such as maximum likelihood estimation and null hypothesis significance testing. First, in contrast to maximum likelihood estimation of higher-order models, relying on Bayesian estimation eliminates the need to perform an optimization procedure over a constrained parameter space, the latter not always resulting in the optimal parameter estimates (LeSage & Pace, 2011). Second, opposed to null hypothesis significance testing, so-called *Bayes factors* allow researchers to quantify relative evidence in the data in favor of the null, or any other, hypothesis against another hypothesis (Kass & Raftery, 1995) and can also be easily extended to test more than two hypotheses against each other simultaneously (Raftery et al., 1997). Hence, this enables researchers to precisely test multiple network operationalizations against each other. Third, Bayes factors have been proven to be very effective for testing hypotheses with order constraints on the parameters of interest (Braeken et al., 2015; Klugkist et al., 2005; Mulder, 2016; Mulder & Wagenmakers, 2016). For example, this makes it possible to precisely test whether social influence is larger among actors with low socio-economic status (SES) than from actors of high SES to those with low SES (or more complicated combinations of equality and inequality expectations). This cannot be done using classical tests and is of particular importance in higher-order network autocorrelation models, as in this setting, researchers often have expectations about the order of strength of different network effects. Whereas such expectations have tended to remain implicit in most research, Bayes factors permit researchers to state them as actual hypotheses and then test them in a precise and straightforward manner.

Thus, we propose Bayes factors for testing multiple hypotheses on the relative importance of social influence in a given network. The presented methodology not only allows a researcher to conclude if there is evidence in the data for, or against, non-zero network autocorrelations in the network, but it grants the researcher the opportunity to simultaneously test any number of competing hypotheses on the relative strength of the network effects against each other as well. Subsequently, we conduct an extensive simulation study to investigate and show the desirable numerical properties of the new procedures, which we then use to re-analyze a data set from the economic growth literature.

We proceed as follows. In the next section, we present higher-order network autocorrelation models in detail before introducing Bayesian estimation and hypothesis testing techniques for the model in Sections 4.3 and 4.4. Concomitantly, we provide efficient implementations for estimating higher-order network autocorrelation models and for computing Bayes factors involving order hypotheses on the network autocorrelation parameters. We assess the numerical behavior of the proposed methods in Section 4.5. In Section 4.6, we illustrate our approaches with an empirical example and Section 4.7 concludes.

4.2 The network autocorrelation model

4.2.1 The first-order network autocorrelation model

Building on a standard linear regression model, the network autocorrelation model relaxes the assumption of independence of observations and allows for correlation between them by explicitly using the underlying network structure. More precisely, an actor's response is modeled as the weighted sum of the actor's neighbor responses and a linear combination of actor attributes. In mathematical notation, the first-order network autocorrelation model is given by

$$\boldsymbol{y} = \rho W \boldsymbol{y} + X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N\left(\mathbf{0}_g, \sigma^2 I_g\right), \tag{4.1}$$

where $\boldsymbol{y}$ is a vector of length g containing the observations for a variable of interest for the g actors in a network, $X \in \mathbb{R}^{g \times k}$ is a standard design matrix (possibly including a vector of ones in the first column for an intercept term), $\boldsymbol{\beta} \in \mathbb{R}^k$ is a vector of k regression coefficients as in standard linear regression, $\boldsymbol{\varepsilon} \in \mathbb{R}^g$ comprises the error terms that are assumed to be independent and identically normally distributed with zero mean and variance of σ^2, $\mathbf{0}_g$ is a vector of zeros of length g, and I_g denotes the $(g \times g)$ identity matrix. Furthermore, W is a $(g \times g)$ *connectivity matrix*, where a non-zero entry W_{ij} amounts to the influence of actor j on actor i and $W_{ii} = 0$ for all $i \in \{1, ..., g\}$. Typically, W is *row-standardized*, i.e., all rows sum up to one, which in this case means that the term $W\boldsymbol{y}$ represents the vector of the actors' neighbors' average responses. Finally, ρ is called the network autocorrelation parameter and quantifies the magnitude of social influence on a variable of interest in a given network as induced by W. For a substantive interpretation of the model, see Leenders (1995, 2002).

The model's likelihood is multivariate normal and can be written as

$$f\left(\boldsymbol{y}|\rho, \sigma^2, \boldsymbol{\beta}\right) = \left|\det\left(A_\rho\right)\right| \left(2\pi\sigma^2\right)^{-\frac{g}{2}} \exp\left(-\frac{1}{2\sigma^2}\left(A_\rho \boldsymbol{y} - X\boldsymbol{\beta}\right)^T \left(A_\rho \boldsymbol{y} - X\boldsymbol{\beta}\right)\right), \tag{4.2}$$

where $A_\rho := I_g - \rho W$ (see e.g., Doreian, 1981). Usually, the parameter space of ρ is chosen as the interval around $\rho = 0$ for which A_ρ is non-singular (Hepple, 1995a; LeSage & Parent, 2007; Smith, 2009). The bounds of this feasible range of ρ are determined by the eigenvalues of W with the smallest and largest real part, respectively, which means that ρ has to be contained in $\left(1/Re\left(\lambda_g\left[W\right]\right), 1/Re\left(\lambda_1\left[W\right]\right)\right)$, where $\lambda_1\left[W\right], ..., \lambda_g\left[W\right]$ denote the eigenvalues of W with $Re\left(\lambda_1\left[W\right]\right) \geq ... \geq Re\left(\lambda_g\left[W\right]\right)$ (Hepple, 1995a). The model's overall parameter space of $\boldsymbol{\theta} := \left(\rho, \sigma^2, \boldsymbol{\beta}\right)$ is then given by $\Theta := \Theta_\rho \times \Theta_{\sigma^2} \times \Theta_{\boldsymbol{\beta}} = \left(1/Re\left(\lambda_g\left[W\right]\right), 1/Re\left(\lambda_1\left[W\right]\right)\right) \times (0, \infty) \times \mathbb{R}^k$.[1]

[1] Except for Leenders (1995), the literature on the network autocorrelation model ignores the occurrence of potentially complex eigenvalues that can arise when considering non-symmetric connectivity matrices. If W is row-standardized, it follows that $Re\left(\lambda_1\left[W\right]\right) = 1$. Lastly, as $\det\left(A_\rho\right) > 0$ for all $\rho \in \Theta_\rho$, we simply write $|A_\rho|$ for $|\det\left(A_\rho\right)|$ in the following.

4.2.2 Higher-order network autocorrelation models

The standard, or first-order, network autocorrelation model in (4.1) is limited to a single network autocorrelation parameter ρ and a single connectivity matrix W. Hence, in this model the social influence is assumed to be homogeneously distributed in the network based on a single influence mechanism. Extending the first-order model to higher-order network autocorrelation models allows for a richer dependence structure by including multiple connectivity matrices, representing different influence mechanisms, e.g., geographic adjacency and social similarity. This amounts to the functional form

$$\boldsymbol{y} = \sum_{r=1}^{R} \rho_r W_r \boldsymbol{y} + X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N\left(\mathbf{0}_g, \sigma^2 I_g\right), \tag{4.3}$$

where $\{W_r\}_r$ are distinct connectivity matrices and the corresponding network autocorrelation parameters $\{\rho_r\}_r$ denote the strength of the different influence mechanisms.

In practice, there can be overlap between connectivity matrices, i.e., different connectivity matrices may share common ties. While partially overlapping connectivity matrices do not pose identification problems as long as there is no complete overlap (Elhorst et al., 2012), overlap does make interpretability of the network autocorrelation parameters more difficult (Elhorst et al., 2012; LeSage & Pace, 2011). In particular, partial overlap may result in empirically unlikely negative autocorrelations (Dittrich et al., 2017; Elhorst et al., 2012). We will analyze the numerical effect of overlapping connectivity matrices on the estimation of and hypothesis tests on $\boldsymbol{\rho} := (\rho_1, ..., \rho_R)$ in more detail in a simulation study in Section 4.5.

Higher-order network autocorrelation models do not only allow to consider multiple influence mechanisms but also to partition a network into several subgroups. In the latter case, we include possible heterogeneity in social influence strengths by allowing for different levels of network autocorrelation within and between subgroups for a given influence mechanism, e.g., geographic adjacency. Dividing the actors in a network into S subgroups, with sizes $g_1, ..., g_S$ and $\sum_{s=1}^{S} g_s = g$, we can express a model with multiple subgroups using the representation in (4.3) by writing

$$\begin{aligned}\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \cdots \\ \boldsymbol{y}_S \end{bmatrix} &= \begin{bmatrix} \rho_{11}W_{11} & \cdots & \rho_{1S}W_{1S} \\ \cdots & \cdots & \cdots \\ \rho_{S1}W_{S1} & \cdots & \rho_{SS}W_{SS} \end{bmatrix} \begin{bmatrix} \boldsymbol{y}_1 \\ \cdots \\ \boldsymbol{y}_S \end{bmatrix} + X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N\left(\mathbf{0}_g, \sigma^2 I_g\right) \\ &= \left(\rho_{11} \begin{bmatrix} W_{11} & 0 & \cdots \\ 0 & \cdots & 0 \end{bmatrix} + \cdots + \rho_{SS} \begin{bmatrix} 0 & \cdots & 0 \\ 0 & \cdots & W_{SS} \end{bmatrix}\right) \begin{bmatrix} \boldsymbol{y}_1 \\ \cdots \\ \boldsymbol{y}_S \end{bmatrix} + X\boldsymbol{\beta} + \boldsymbol{\varepsilon},\end{aligned}$$

where $\boldsymbol{y}_s$ is a vector of length g_s containing the observations for the g_s actors in the s-th subgroup of the network, $W_{ss'}$ is a $(g_s \times g_{s'})$ connectivity matrix defining the influence relationships between members of subgroup s' and members of subgroup s, and $\rho_{ss'}$ is the network autocorrelation parameter representing the strength of the social influence of the actors in subgroup s' on the actors in subgroup s. As the sizes of the S subgroups potentially differ, each $W_{ss'}$ is typically row-standardized separately, which removes scale effects and eases direct comparison between the network autocorrelation parameters (McMillen et al., 2007).

The structure of the likelihood function of higher-order network autocorrelation models remains the same as the one of the first-order network autocorrelation model in (4.2), with A_ρ being replaced by $A_{\boldsymbol{\rho}} := I_g - \sum_{r=1}^{R} \rho_r W_r$. As in the first-order model, we define the R-dimensional parameter space of $\boldsymbol{\rho} = (\rho_1, ..., \rho_R)$ as the space containing the origin for which $A_{\boldsymbol{\rho}}$ is non-singular. Elhorst et al. (2012) provided a simple general procedure for checking if a point $\boldsymbol{\rho}^* \in \mathbb{R}^R$, given $W_1, ..., W_R$, lies in the corresponding feasible parameter space $\Theta_{\boldsymbol{\rho}}$.[2]

4.2.3 Application of a higher-order network autocorrelation model: Economic growth of labor productivity

In this subsection, we introduce a data set from the economic growth literature that prompts questions readily answered using Bayes factors. Here, we merely describe the data set and the research questions, while we will come back and provide solutions to them in Section 4.6.

Dall'erba et al. (2009) employed a second-order network autocorrelation model to explain the growth rates of labor productivity in service industry across 188 European regions in 12 countries from 1980 to 2003. In order to adequately deal with interregional spillovers, the authors introduced two different spatial weight matrices, W_1 and W_2, "under the assumption that economic interactions decrease very substantially when a national border is passed" (Dall'erba et al., 2009, p.337). Hence, W_1 was constructed using the three nearest neighbors of a region within the same country, while W_2 was based on the three nearest neighbors in the bordering countries. These raw binary connectivity matrices were subsequently row-normalized by the authors. In addition to an intercept term, Dall'erba et al. (2009) considered four more explanatory variables: the growth rate of market service output in a region, the initial labor productivity gap between the region and the leading region, a measure of urbanization of the region, and a measure of the accessibility of the region. Thus, their model is given by

$$\boldsymbol{y} = \rho_1 W_1 \boldsymbol{y} + \rho_2 W_2 \boldsymbol{y} + \beta_1 \boldsymbol{X}_{\cdot 1} + \beta_2 \boldsymbol{X}_{\cdot 2} + \beta_3 \boldsymbol{X}_{\cdot 3} + \beta_4 \boldsymbol{X}_{\cdot 4} + \beta_5 \boldsymbol{X}_{\cdot 5} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N\left(\boldsymbol{0}, \sigma^2 I\right), \tag{4.4}$$

where $\boldsymbol{y} \in \mathbb{R}^{188}$ is the vector of growth rates of labor productivity in service industry across the 188 regions, $\boldsymbol{\beta} \in \mathbb{R}^5$ represents the vector of the four regression coefficients plus an intercept term, $X \in \mathbb{R}^{188 \times 5}$ contains the values for the explanatory variables for the 188 regions, where $\boldsymbol{X}_{\cdot i}, i \in \{1, ..., 5\}$, denotes the i-th column of X, $\boldsymbol{0} \in \mathbb{R}^{188}$ is a vector of zeros, and $I \in \mathbb{R}^{188 \times 188}$ represents the corresponding identity matrix.

[2] In Elhorst et al. (2012), the term $\tan(\alpha) / r_{\max}[W^*]$ needs to be replaced by $\tan(\alpha) / r_{\min}[W^*]$ in Equation (15) on page 213.

The authors found that the estimate of ρ_1, reflecting interactions within the same country, was positive and statistically significant, indicating the presence of positive spatial within-country spillover effects. On the other hand, the estimate of ρ_2 was very close to zero and statistically not significant. Dall'erba et al. (2009) concluded by saying that "the results obtained also confirm the hypothesis that economic interactions decrease very substantially when a national border is passed (indeed, the coefficient reflecting external spillovers is not statistically significant)" (Dall'erba et al., 2009, p.342). However, in order to draw this conclusion, one needs to directly test a corresponding hypothesis, e.g., $H_1 : \rho_1 > \rho_2 = 0$, against a (set of) competing hypothesis (hypotheses), such as $H_0 : \rho_1 = \rho_2 = 0$, $H_2 : \rho_1 > \rho_2 > 0$, or (and) $H_3 : \rho_1 = \rho_2 > 0$. These four hypotheses correspond to the notion of "no network effects" (hypothesis H_0), "a positive within-country network effect only" (hypothesis H_1), "positive but decreasing network effects after a national border is passed" (hypothesis H_2), and "positive and equally strong within-country and between-country network effects" (hypothesis H_3). Currently, no formal statistical method is available to directly test such hypotheses on multiple network autocorrelations. In the remainder of this chapter, we develop a Bayesian framework for testing and quantifying the evidence in the data for such hypotheses involving equality and order constraints on the network effects. We will come back to this empirical example and test these hypotheses against each other using Bayes factors in Section 4.6. Lastly, the authors stated that "there is evidence that the coefficients in a growth model are potentially varying for different subsets of the total sample" (Dall'erba et al., 2009, p.342). In Section 4.6, we will also investigate if there is such evidence in this data set by considering a network autocorrelation model with two subgroups, allowing for differing levels of network autocorrelation within and between the two subgroups.

4.3 Bayesian estimation of higher-order network autocorrelation models

4.3.1 Prior specification

Bayesian estimation starts with formulating prior expectations about the parameters in a model that is done in terms of so-called *prior distributions*, or *priors*. These priors summarize the (lack of) information about the model parameters before observing the data. If such prior information is available, e.g., based on previous literature, *informative priors* for the parameters of interest can be formulated. In Chapter 2, we performed a literature study for the first-order network autocorrelation model, where we looked at the distribution of reported network autocorrelations across many different fields. Our results showed that most of the analyzed data in the literature exhibit positive network autocorrelation between 0 and .5, while it seems highly unlikely to observe negative network autocorrelation, as previously also noted by e.g., Neuman & Mizruchi (2010). This information could then be used to formulate an informative prior for ρ in a first-order network autocorrelation model (see e.g., Dittrich et al., in press).

On the other hand, if such prior information is missing, or a researcher deliberately refrains from adding additional information to the model through the prior, usually so-called *non-informative priors* are used (Gelman et al., 2003). In the network autocorrelation model, σ^2 and $\boldsymbol{\beta}$ are commonly assigned the standard non-informative priors $p\left(\sigma^2\right) \propto 1/\sigma^2$ and $p\left(\boldsymbol{\beta}\right) \propto 1$, respectively (Hepple, 1995a; Holloway et al., 2002; LeSage, 1997b, 2000). These priors assume that all possible values for $\log\left(\sigma^2\right)$ and $\boldsymbol{\beta}$ are equally likely a priori. We also do so throughout this chapter. Note that these priors are not *proper* in the sense that they do not integrate to a finite value, but this does not affect estimation of the model.

We use a general R-variate normal prior for $\boldsymbol{\rho}$, $p\left(\boldsymbol{\rho}\right) = \phi_{\boldsymbol{\mu},\Sigma}\left(\boldsymbol{\rho}\right) \mathbb{1}_{\Theta_{\boldsymbol{\rho}}}\left(\boldsymbol{\rho}\right) c^{-1}$, where $\phi_{\boldsymbol{\mu},\Sigma}\left(\cdot\right)$ denotes the probability density function of a multivariate normal distribution with prior mean $\boldsymbol{\mu}$ and prior covariance matrix Σ, $\mathbb{1}_{\cdot}\left(\cdot\right)$ is the standard indicator function, and $c := \int_{\Theta_{\boldsymbol{\rho}}} \phi_{\boldsymbol{\mu},\Sigma}\left(\boldsymbol{\rho}\right) \mathrm{d}\boldsymbol{\rho}$ is a normalizing constant representing the probability mass of $\phi_{\boldsymbol{\mu},\Sigma}\left(\cdot\right)$ contained in the network autocorrelation parameters' space $\Theta_{\boldsymbol{\rho}}$. If researchers have sufficient prior information about the network autocorrelations, they can specify $\boldsymbol{\mu}$ and Σ directly. Alternatively, when specifying Σ vaguely enough, i.e., with very large diagonal elements, the prior becomes essentially identical to a proper uniform distribution for $\boldsymbol{\rho}$ on the bounded parameter space $\Theta_{\boldsymbol{\rho}}$.

In summary, we use the following priors for the model parameters, which we assume to be a priori independent from each other:

$$
\begin{aligned}
p\left(\boldsymbol{\rho}\right) &= \phi_{\boldsymbol{\mu},\Sigma}\left(\boldsymbol{\rho}\right) \mathbb{1}_{\Theta_{\boldsymbol{\rho}}}\left(\boldsymbol{\rho}\right) c^{-1}, && (4.5)\\
p\left(\sigma^2\right) &\propto 1/\sigma^2, && (4.6)\\
p\left(\boldsymbol{\beta}\right) &\propto 1, && (4.7)\\
p\left(\boldsymbol{\rho}, \sigma^2, \boldsymbol{\beta}\right) &= p\left(\boldsymbol{\rho}\right) \times p\left(\sigma^2\right) \times p\left(\boldsymbol{\beta}\right).
\end{aligned}
$$

4.3.2 Posterior computation

After having specified a prior distribution for the model parameters, the information contained in the observed data $\boldsymbol{y}$ is used to update the prior distribution and to arrive at the *posterior distribution*, or simply *posterior*. The posterior is used for all Bayesian inference in the model, e.g., to obtain point estimates of model parameters (the posterior mean or the posterior median), to construct Bayesian credible intervals (i.e., intervals in the domain of the posterior), or to determine other statistics of interest, such as the probability that one network effect is stronger than another one for given data, $p\left(\rho_1 > \rho_2 | \boldsymbol{y}\right)$. In this subsection, we specify the posterior for higher-order network autocorrelation models based on the priors from Section 4.3.1 and provide an automatic and efficient scheme to sample from this posterior.

First, Bayes' theorem gives that the posterior is proportional to the prior multiplied by the likelihood, more precisely

$$
\begin{aligned}
p\left(\boldsymbol{\rho}, \sigma^2, \boldsymbol{\beta} | \boldsymbol{y}\right) &= \frac{p\left(\boldsymbol{\rho}, \sigma^2, \boldsymbol{\beta}\right) f\left(\boldsymbol{y} | \boldsymbol{\rho}, \sigma^2, \boldsymbol{\beta}\right)}{\int_{\Theta_{\boldsymbol{\rho}}} \int_{\Theta_{\sigma^2}} \int_{\Theta_{\boldsymbol{\beta}}} p\left(\boldsymbol{\rho}, \sigma^2, \boldsymbol{\beta}\right) f\left(\boldsymbol{y} | \boldsymbol{\rho}, \sigma^2, \boldsymbol{\beta}\right) \mathrm{d}\boldsymbol{\beta} \mathrm{d}\sigma^2 \mathrm{d}\boldsymbol{\rho}} \\
&\propto p\left(\boldsymbol{\rho}, \sigma^2, \boldsymbol{\beta}\right) f\left(\boldsymbol{y} | \boldsymbol{\rho}, \sigma^2, \boldsymbol{\beta}\right).
\end{aligned} \tag{4.8}
$$

The denominator of (4.8) is called the *marginal likelihood* and ensures that the posterior integrates to unity. The marginal likelihood does not depend on any model parameters and can be ignored in Bayesian estimation. On the other hand, when testing hypotheses, the marginal likelihood does play a central role as it quantifies how plausible the data are under a specific hypothesis, which we will discuss in the following section.

Next, using the priors in (4.5), (4.6), (4.7), and the likelihood function in (4.2), we can express the posterior $p\left(\boldsymbol{\rho}, \sigma^2, \boldsymbol{\beta} | \boldsymbol{y}\right)$ for higher-order network autocorrelation models as

$$
\begin{aligned}
p\left(\boldsymbol{\rho}, \sigma^2, \boldsymbol{\beta} | \boldsymbol{y}\right) \propto{} & |A_{\boldsymbol{\rho}}| \left(\sigma^2\right)^{-\frac{g}{2}-1} \\
& \exp\left(-\frac{1}{2}\left(\boldsymbol{\rho}-\boldsymbol{\mu}\right)^T \Sigma^{-1} \left(\boldsymbol{\rho}-\boldsymbol{\mu}\right) - \frac{1}{2\sigma^2}\left(A_{\boldsymbol{\rho}}\boldsymbol{y} - X\boldsymbol{\beta}\right)^T \left(A_{\boldsymbol{\rho}}\boldsymbol{y} - X\boldsymbol{\beta}\right)\right).
\end{aligned} \tag{4.9}
$$

However, the posterior distribution in (4.9) does not belong to a family of known probability distributions, so we cannot directly infer its posterior mean, its quantiles, or other quantities of interest.[3] In this case, it is common to sample random draws from the posterior distribution and to use these posterior draws to approximate any desired statistic. An efficient method is to sequentially draw from the conditional posterior distributions, i.e., the posterior distribution of one parameter (block) given the remaining parameters and the data (Geman & Geman, 1984; Gelfand & Smith, 1990).[4] Extending the proposed method for the first-order network autocorrelation model in Chapter 2 to higher-order models, we sample the model parameters according to the following blocks: $(\boldsymbol{\rho}, \beta_1), \sigma^2$, and $\widetilde{\boldsymbol{\beta}}$, where β_1 denotes the model's intercept and $\widetilde{\boldsymbol{\beta}} = (\beta_2, ..., \beta_k)$ contains the remaining regression coefficients. By simultaneously sampling $\boldsymbol{\rho}$ and β_1, we can better capture potential posterior correlation between the network effects as well as potential correlation between the network effects and the intercept (Dittrich et al., 2017). The conditional posteriors for the proposed blocks are then given by (see e.g., LeSage, 1997a)

[3] The posterior $p\left(\boldsymbol{\rho}, \sigma^2, \boldsymbol{\beta} | \boldsymbol{y}\right)$ in (4.9) is proper given very mild regularity conditions. The proof for higher-order models is quasi-identical to and an adaptation of the one for the first-order model in Chapter 2.

[4] This is iteratively repeated N times for a large number N. Geman & Geman (1984) showed that as $N \to \infty$, the draws based on the sequence of conditional posteriors can be seen as samples from the actual marginal posteriors, i.e., the posteriors for a parameter (block) given the data, e.g., $p\left(\boldsymbol{\rho} | \boldsymbol{y}\right)$.

$$p\left(\boldsymbol{\rho}, \beta_1 | \sigma^2, \widetilde{\boldsymbol{\beta}}, \boldsymbol{y}\right)$$
$$\propto |A_{\boldsymbol{\rho}}| \exp\left(-\frac{1}{2}\left(\boldsymbol{\rho}-\boldsymbol{\mu}\right)^T \Sigma^{-1}\left(\boldsymbol{\rho}-\boldsymbol{\mu}\right) - \frac{1}{2\sigma^2}\left(A_{\boldsymbol{\rho}}\boldsymbol{y} - X\boldsymbol{\beta}\right)^T\left(A_{\boldsymbol{\rho}}\boldsymbol{y} - X\boldsymbol{\beta}\right)\right), \quad (4.10)$$

$$p\left(\sigma^2 | \boldsymbol{\rho}, \beta_1, \widetilde{\boldsymbol{\beta}}, \boldsymbol{y}\right) \sim IG\left(\frac{g}{2}, \frac{\left(A_{\boldsymbol{\rho}}\boldsymbol{y} - X\boldsymbol{\beta}\right)^T\left(A_{\boldsymbol{\rho}}\boldsymbol{y} - X\boldsymbol{\beta}\right)}{2}\right), \quad (4.11)$$

$$p\left(\widetilde{\boldsymbol{\beta}} | \boldsymbol{\rho}, \beta_1, \sigma^2, \boldsymbol{y}\right) \sim N\left(\boldsymbol{\mu}_{\widetilde{\boldsymbol{\beta}}}, \Sigma_{\widetilde{\boldsymbol{\beta}}}\right), \quad (4.12)$$

where $IG(\cdot,\cdot)$ denotes the inverse gamma distribution and $\boldsymbol{\mu}_{\widetilde{\boldsymbol{\beta}}}$ and $\Sigma_{\widetilde{\boldsymbol{\beta}}}$ are given in Appendix 4.A.

Drawing from the conditional posteriors in (4.11) and (4.12) can be done using standard statistical software. In contrast, the conditional posterior in (4.10) does not have a well-known form and cannot be directly sampled from. Instead, we use the so-called *Metropolis-Hastings algorithm* (Hastings, 1970; Metropolis et al., 1953) to generate draws from the conditional posterior for $(\boldsymbol{\rho}, \beta_1)$. In short, the algorithm generates candidate values for the conditional posterior from a candidate-generating distribution that can be easily sampled from and subsequently accepts, or rejects, the draws with a certain probability. The algorithm's efficiency mainly depends on the shape of the proposed candidate-generating distribution; if possible, exploiting the form of the conditional posterior and specifying a candidate-generating distribution that closely approximates it results in efficient solutions (Chib & Greenberg, 1995).

As to that, we first approximate $\log(|A_{\boldsymbol{\rho}}|)$ by a quadratic polynomial in $\boldsymbol{\rho}$ by virtue of Jacobi's formula and the Mercator series, see Appendix 4.A. Next, we observe that the logarithm of the exponential in (4.10) can also be written as a quadratic polynomial in $(\boldsymbol{\rho}, \beta_1)$. Hence, the logarithm of the conditional posterior itself can be approximated by a quadratic polynomial in $(\boldsymbol{\rho}, \beta_1)$. Finally, by equating coefficients of this quadratic polynomial with the log-kernel of the probability density function of a $(R+1)$-variate normal distribution, the density in (4.10) can be approximated by a $(R+1)$-variate normal candidate-generating density for $(\boldsymbol{\rho}, \beta_1)$ that is tailored to the conditional posterior for $(\boldsymbol{\rho}, \beta_1)$.[5] All details and the full sampling scheme can be found in Appendix 4.A.

We implemented our proposed approach in R (R Core Team, 2017) and compared its performance to a sampling scheme that does not block the network autocorrelation parameters and the intercept but uses one-dimensional random walk algorithms to generate draws for each network effect sequentially, as in Zhang et al. (2013). Figure 4.1 shows exemplary trace plots of posterior draws for ρ_1 and ρ_2 based on the two sampling schemes and the data in Dall'erba et al. (2009) and model (4.4). We can observe that

[5] It can (rarely) happen that, after equating coefficients, the obtained covariance matrix of the normal candidate-generating distribution is not positive definite, as the Hessian in the second-order approximation of $\log(|A_{\boldsymbol{\rho}}|)$ itself is not always positive definite, e.g., for $W_1 = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$ and $W_2 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$. Thus, if the initially obtained matrix is not positive definite, we instead use its nearest positive definite one as the normal candidate-generating distribution's covariance matrix. This can be done using the nearPD() function from the Matrix package in R (Bates & Maechler, 2017).

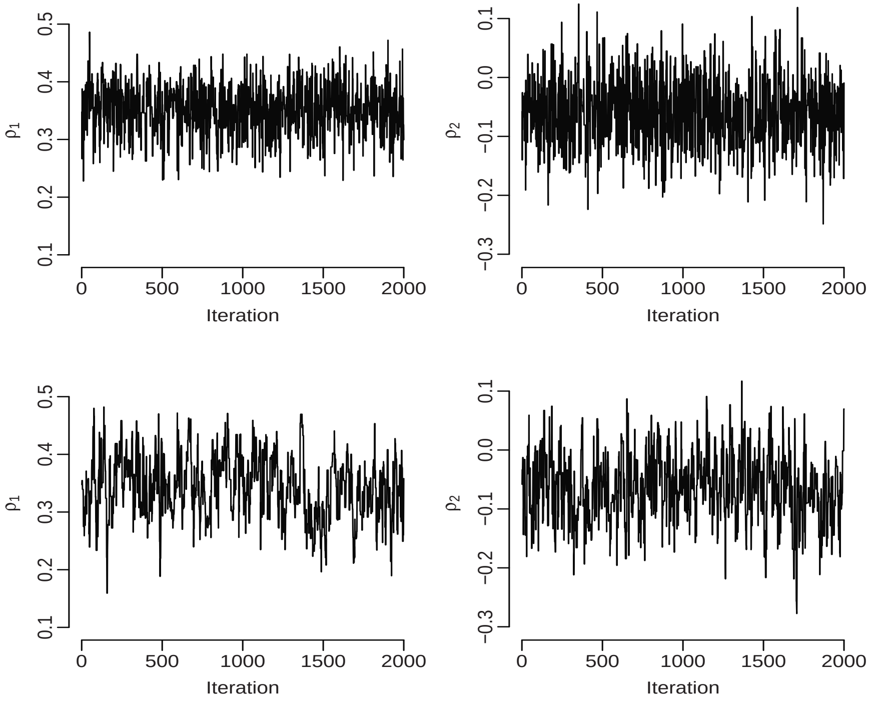

Figure 4.1 Trace plots of posterior draws for ρ_1 and ρ_2 based on our proposed scheme (top row) and a random walk algorithm (bottom row) for the data in Dall'erba et al. (2009) and model (4.4).

our method results in a more efficient implementation than drawing each network effect separately, as it generates Markov chains that explore the corresponding parameter space of (ρ_1, ρ_2) much faster. Lastly, our approach is fully automatic in the sense that there are no parameters to be tuned in the Metropolis-Hastings algorithm, such as the variances of candidate-generating distributions.

To conclude, the presented sampling algorithm allows researchers to automatically and efficiently draw from the posterior based on a general multivariate normal prior for the network autocorrelation parameters, including informative as well as non-informative specifications. Such efficient sampling is essential for performing any Bayesian estimation of the model, which solely relies upon the generated posterior draws.

4.4 Bayesian hypothesis testing in higher-order network autocorrelation models

In many network studies, researchers have competing theories about the specific order of different network effect strengths. These theories can be formulated as hypotheses on the network autocorrelation parameters, e.g., as $H_1 : \rho_1 > \rho_2 = 0$, $H_2 : \rho_1 > \rho_2 > 0$, or $H_3 : \rho_1 = \rho_2 > 0$, and can include as many network autocorrelation parameters as relevant to one's theory. The focus of interest then lies on which substantive theory, or hypothesis, is most plausible and most supported by the data and how strongly. In this chapter, we consider $T \geq 2$ constrained hypotheses on the network effects, where a hypothesis H_t, $t \in \{0, ..., T-1\}$, contains q_t^I inequality and q_t^E equality constraints on $\boldsymbol{\rho}$, i.e.,

$$H_t := \begin{cases} R_t^I \boldsymbol{\rho} > \boldsymbol{r}_t^I \\ R_t^E \boldsymbol{\rho} = \boldsymbol{r}_t^E, \end{cases} \tag{4.13}$$

where R_t^I and $\boldsymbol{r}_t^I$ are a $\left(q_t^I \times R\right)$ matrix and a vector of length R, respectively, containing the coefficients of the q_t^I inequality constraints under hypothesis H_t. Equivalently, the $\left(q_t^E \times R\right)$ matrix R_t^E and the vector $\boldsymbol{r}_t^E$ contain the coefficients of the q_t^E equality constraints. For example, the constraints induced by the three hypotheses $H_1 : \rho_1 > \rho_2 = 0$, $H_2 : \rho_1 > \rho_2 > 0$, and $H_3 : \rho_1 = \rho_2 > 0$ can be represented by (4.13) according to[6]

$$\begin{aligned}
&H_1: \quad R_1^I = (1,0)\,, && r_1^I = 0, && R_1^E = (0,1)\,, && r_1^E = 0,\\
&H_2: \quad R_2^I = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}, && \boldsymbol{r}_2^I = (0,0)\,, && && \\
&H_3: \quad R_3^I = (1,0)\,, && r_3^I = 0, && R_3^E = (1,-1)\,, && r_3^E = 0.
\end{aligned}$$

4.4.1 The Bayes factor

The Bayes factor is a comparative Bayesian hypothesis testing criterion that directly quantifies the relative evidence for a hypothesis in the data. The Bayes factor of hypothesis H_t against hypothesis $H_{t'}$, $t, t' \in \{0, ..., T-1\}$, is defined as the ratio of the marginal likelihoods under the two hypotheses, i.e., in the network autocorrelation model as

$$\begin{aligned} B_{tt'} &= \frac{m_t(\boldsymbol{y})}{m_{t'}(\boldsymbol{y})} \qquad\qquad (4.14) \\ &= \frac{\int_{\Theta_{\boldsymbol{\rho}_t}} \int_0^\infty \int_{\mathbb{R}^k} p_t(\boldsymbol{\rho}_t)\, p\left(\sigma^2\right) p(\boldsymbol{\beta})\, f\left(\boldsymbol{y}|\boldsymbol{\rho}_t, \sigma^2, \boldsymbol{\beta}\right) \mathrm{d}\boldsymbol{\beta} \mathrm{d}\sigma^2 \mathrm{d}\boldsymbol{\rho}_t}{\int_{\Theta_{\boldsymbol{\rho}_{t'}}} \int_0^\infty \int_{\mathbb{R}^k} p_{t'}(\boldsymbol{\rho}_{t'})\, p(\sigma^2)\, p(\boldsymbol{\beta})\, f(\boldsymbol{y}|\boldsymbol{\rho}_{t'}, \sigma^2, \boldsymbol{\beta})\, \mathrm{d}\boldsymbol{\beta} \mathrm{d}\sigma^2 \mathrm{d}\boldsymbol{\rho}_{t'}}, \end{aligned}$$

where $\boldsymbol{\rho}_t$ are the network autocorrelation parameters under hypothesis H_t, $p_t(\boldsymbol{\rho}_t)$ denotes their prior density, and $\Theta_{\boldsymbol{\rho}_t}$ the corresponding parameter space (Kass & Raftery, 1995). We assume common priors for σ^2 and $\boldsymbol{\beta}$ under both hypothesis H_t and hypothesis $H_{t'}$ as they are seen as nuisance parameters in the presented framework. The exact form of the priors for these nuisance parameters typically does not alter the magnitude of the Bayes factor (Kass & Raftery, 1995).

[6] For hypothesis H_1, R_1^I and r_1^I imply $1 \times \rho_1 + 0 \times \rho_2 > 0$, while R_1^E and r_1^E lead to $0 \times \rho_1 + 1 \times \rho_2 = 0$. Together, this constitutes hypothesis H_1. For hypothesis H_2, we have $1 \times \rho_1 - 1 \times \rho_2 > 0$ and $0 \times \rho_1 + 1 \times \rho_2 > 0$. Analogously, for hypothesis H_3, $1 \times \rho_1 + 0 \times \rho_2 > 0$ and $1 \times \rho_1 - 1 \times \rho_2 = 0$.

Table 4.1 Evidence categories for the Bayes factor $BF_{tt'}$ as given by Jeffreys (1961).

$BF_{tt'}$			$\log\left(BF_{tt'}\right)$			Interpretation
	$>$	100		$>$	4.61	Decisive evidence for hypothesis H_t
30	-	100	3.40	-	4.61	Very strong evidence for hypothesis H_t
10	-	30	2.30	-	3.40	Strong evidence for hypothesis H_t
3	-	10	1.10	-	2.30	Substantial evidence for hypothesis H_t
1	-	3	0	-	1.10	Not worth more than a bare mention
1/3	-	1	-1.10	-	0	Not worth more than a bare mention
1/10	-	1/3	-2.30	-	-1.10	Substantial evidence for hypothesis $H_{t'}$
1/30	-	1/10	-3.40	-	-2.30	Strong evidence for hypothesis $H_{t'}$
1/100	-	1/30	-4.61	-	-3.40	Very strong evidence for hypothesis $H_{t'}$
	$<$	1/100		$<$	-4.61	Decisive evidence for hypothesis $H_{t'}$

The marginal likelihood under hypothesis H_t, $m_t\left(\boldsymbol{y}\right)$, is a weighted average likelihood over the parameter space under hypothesis H_t, with the prior $p_t\left(\boldsymbol{\rho}_t\right)$ under hypothesis H_t acting as a weight function. As such, it can be interpreted as the probability that the data were observed under hypothesis H_t. Hence, the Bayes factor, as the ratio of two marginal likelihoods, quantifies the relative evidence that the data were observed under hypothesis H_t rather than hypothesis $H_{t'}$. For example, when $B_{tt'} = 5$, this indicates that the data are five times more likely to have occurred under hypothesis H_t compared to hypothesis $H_{t'}$. Conversely, when $B_{tt'} = 1/5$, it is five times more likely to have observed the data under hypothesis $H_{t'}$ than under hypothesis H_t.

In order to facilitate interpretation of the Bayes factor, Jeffreys (1961) proposed a classification scheme that groups Bayes factors into different categories, see Table 4.1. For example, there is "strong" evidence in the data for hypothesis H_t, relative to hypothesis $H_{t'}$, when $B_{tt'} > 10$ and, equivalently, "strong" relative evidence in the data for hypothesis $H_{t'}$ when $B_{tt'} < 1/10$. This grouping provides verbal descriptions and rules of thumb when speaking of relative evidence in the data in favor of a hypothesis but is still somewhat arbitrary. Ultimately, the interpretation of the magnitude of a Bayes factor should hinge upon the context of the research question (Kass & Raftery, 1995). For some introductory texts on Bayes factor testing in social science research, we refer the interested reader to Raftery (1995), van de Schoot et al. (2011), or Wagenmakers (2007).

4.4.2 Bayes factor computation

In this section, we present efficient methods to compute marginal likelihoods and Bayes factors in higher-order network autocorrelation models. Using a multivariate normal prior for $\boldsymbol{\rho}_t$ under hypothesis H_t, $p_t\left(\boldsymbol{\rho}_t\right) = \phi_{\boldsymbol{\mu}_t,\Sigma_t}\left(\boldsymbol{\rho}_t\right)\mathbb{1}_{\Theta_{\boldsymbol{\rho}_t}}\left(\boldsymbol{\rho}_t\right)c_t^{-1}$, $c_t := \int_{\Theta_{\boldsymbol{\rho}_t}}\phi_{\boldsymbol{\mu}_t,\Sigma_t}\left(\boldsymbol{\rho}_t\right)\mathrm{d}\boldsymbol{\rho}_t$, the non-informative prior $p\left(\sigma^2,\boldsymbol{\beta}\right) \propto 1/\sigma^2$ for the nuisance parameters σ^2 and $\boldsymbol{\beta}$, and after analytically integrating out σ^2 and $\boldsymbol{\beta}$, the Bayes factor of hypothesis H_t against hypothesis $H_{t'}$ in (4.14) reduces to

$$\begin{aligned}B_{tt'} &= \frac{m_t(\boldsymbol{y})}{m_{t'}(\boldsymbol{y})}\\ &= \frac{c_{t'}\left|2\pi\Sigma_t\right|^{-\frac{1}{2}}}{c_t\left|2\pi\Sigma_{t'}\right|^{-\frac{1}{2}}}\\ &\quad\frac{\int_{\Theta_{\boldsymbol{\rho}_t}}\left|A_{\boldsymbol{\rho}_t}\right|\exp\left(-\frac{1}{2}\left(\boldsymbol{\rho}_t-\boldsymbol{\mu}_t\right)^T\Sigma_t^{-1}\left(\boldsymbol{\rho}_t-\boldsymbol{\mu}_t\right)\right)\boldsymbol{y}^TA_{\boldsymbol{\rho}_t}^TMA_{\boldsymbol{\rho}_t}\boldsymbol{y}^{-\frac{g-k}{2}}\mathrm{d}\boldsymbol{\rho}_t}{\int_{\Theta_{\boldsymbol{\rho}_{t'}}}\left|A_{\boldsymbol{\rho}_{t'}}\right|\exp\left(-\frac{1}{2}\left(\boldsymbol{\rho}_{t'}-\boldsymbol{\mu}_{t'}\right)^T\Sigma_{t'}^{-1}\left(\boldsymbol{\rho}_{t'}-\boldsymbol{\mu}_{t'}\right)\right)\boldsymbol{y}^TA_{\boldsymbol{\rho}_{t'}}^TMA_{\boldsymbol{\rho}_{t'}}\boldsymbol{y}^{-\frac{g-k}{2}}\mathrm{d}\boldsymbol{\rho}_{t'}},\end{aligned} \quad (4.15)$$

where $M := I_g - X\left(X^TX\right)^{-1}X^T$.[7]

The normalizing constants c_t and $c_{t'}$ in (4.15) correspond to the prior probabilities that the unconstrained priors for $\boldsymbol{\rho}_t$ under hypothesis H_t and for $\boldsymbol{\rho}_{t'}$ under hypothesis $H_{t'}$, $N\left(\boldsymbol{\mu_t}, \Sigma_t\right)$ and $N\left(\boldsymbol{\mu}_{t'}, \Sigma_{t'}\right)$, are in agreement with the constraints imposed under the two hypotheses. They can be approximated by simple rejection sampling, i.e., by sampling draws from the unconstrained priors and recording the proportions of draws that are in agreement with the constraints. The remaining integrals in the numerator and the denominator of (4.15) do not have closed-form solutions and have to be evaluated numerically. For this purpose, we rely on an importance sampling procedure (A. Owen & Zhou, 2000) that is explained next.

Let $h_t\left(\boldsymbol{\rho}_t\right) := \left|A_{\boldsymbol{\rho}_t}\right|\exp\left(-\frac{1}{2}\left(\boldsymbol{\rho}_t-\boldsymbol{\mu}_t\right)^T\Sigma_t^{-1}\left(\boldsymbol{\rho}_t-\boldsymbol{\mu}_t\right)\right)\boldsymbol{y}^TA_{\boldsymbol{\rho}_t}^TMA_{\boldsymbol{\rho}_t}\boldsymbol{y}^{-\frac{g-k}{2}}$ denote the integrand in the numerator of (4.15) (all steps equivalently apply to $h_{t'}\left(\boldsymbol{\rho}_{t'}\right)$). Then, we can write for the numerator of (4.15)

$$\begin{aligned}I_t &:= \int_{\Theta_{\boldsymbol{\rho}_t}} h_t\left(\boldsymbol{\rho}_t\right)\mathrm{d}\boldsymbol{\rho}_t = \int_{\Theta_{\boldsymbol{\rho}_t}} q_t\left(\boldsymbol{\rho}_t\right)\frac{h_t\left(\boldsymbol{\rho}_t\right)}{q_t\left(\boldsymbol{\rho}_t\right)}\mathrm{d}\boldsymbol{\rho}_t = \mathbb{E}\left[\frac{h_t\left(\boldsymbol{P}\right)}{q_t\left(\boldsymbol{P}\right)}\right] \quad (4.16)\\ &\approx N^{-1}\sum_{i=1}^{N}\frac{h_t\left(\boldsymbol{\rho}_i\right)}{q_t\left(\boldsymbol{\rho}_i\right)} := \widehat{I}_t,\end{aligned}$$

where $\boldsymbol{P}$ is a random variable with probability density function $q_t\left(\cdot\right)$ known as the *importance density*, $\mathbb{E}\left[h_t\left(\boldsymbol{P}\right)/q_t\left(\boldsymbol{P}\right)\right]$ denotes the expected value for $h_t\left(\boldsymbol{P}\right)/q_t\left(\boldsymbol{P}\right)$, and $\boldsymbol{\rho}_i$ are draws from $q_t\left(\cdot\right)$, forming realizations of $\boldsymbol{P}$. The specification of the importance density is crucial for the algorithm's efficiency, where we aim to construct a density that closely follows the actual integrand but has heavier tails than the latter and is easy to sample from (A. Owen & Zhou, 2000).

As in Section 4.3.2, we approximate $\log\left(\left|A_{\boldsymbol{\rho}_t}\right|\right)$ by a second-order polynomial in $\boldsymbol{\rho}_t$ at its maximum, the origin. This results in a normal approximation of $\left|A_{\boldsymbol{\rho}_t}\right|$. We apply the same rationale to the third term in $h_t\left(\boldsymbol{\rho}_t\right)$, $\boldsymbol{y}^TA_{\boldsymbol{\rho}_t}^TMA_{\boldsymbol{\rho}_t}\boldsymbol{y}^{-(g-k)/2}$. Hence, $h_t\left(\boldsymbol{\rho}_t\right)$ can be approximated by the product of three multivariate normal densities that itself is a multivariate normal density, which we use as importance density in (4.16).[8] Finally, as $\boldsymbol{\rho}_t$

[7] We can use improper priors for the nuisance parameters σ^2 and $\boldsymbol{\beta}$, $p\left(\sigma^2, \boldsymbol{\beta}\right) \propto 1/\sigma^2$, as they appear in both hypothesis H_t and hypothesis $H_{t'}$ and the corresponding normalizing constants cancel out after integrating out σ^2 and $\boldsymbol{\beta}$ (Hepple, 1995a). Note that depending on the specifications of hypothesis H_t and hypothesis $H_{t'}$, the dimensions of Σ_t and $\Sigma_{t'}$ may differ, e.g., for $H_1 : \rho_1 > \rho_2 = 0$ and $H_2 : \rho_1 > \rho_2 > 0$.

[8] As in Section 4.3.2, if the resulting covariance matrix of the normal distribution approximating $\left|A_{\boldsymbol{\rho}_t}\right|$ is not positive definite, we use its nearest positive definite matrix as covariance matrix instead.

approaches the boundary of $\Theta_{\boldsymbol{\rho}_t}$, the proposed normal importance density has heavier tails than $h_t(\boldsymbol{\rho}_t)$, since in this case $\left|A_{\boldsymbol{\rho}_t}\right|$ decreases toward zero, while the normal importance density does not. This ensures a finite variance of the importance sampling estimate $\widehat{I_t}$ and reliable estimation of the associated Bayes factors. All details hereto can be found in Appendix 4.B.

4.4.3 A default prior for ρ

When testing multiple hypotheses against each other, a prior for the tested model parameters has to be specified under each hypothesis. Arguably, eliciting a prior under each hypothesis directly can become difficult and cumbersome, especially with a large number of hypotheses at hand. As an alternative, we propose an automatic *empirical Bayes* procedure (Carlin & Louis, 2000) for constructing a default prior $p_t(\boldsymbol{\rho}_t)$ under each hypothesis H_t such that the marginal likelihood under every hypothesis H_t is maximized.

First, we center the multivariate normal default prior $p_t(\boldsymbol{\rho}_t)$ under hypothesis H_t around the origin. The motivation for this choice is that the origin is located at the boundary of typical (in)equality constrained hypotheses in the network autocorrelation model, such as $H_1 : \rho_1 > \rho_2 = 0$, $H_2 : \rho_1 > \rho_2 > 0$, or $H_3 : \rho_1 = \rho_2 > 0$, and previous literature on order constrained hypothesis testing has suggested that "there is a gain of evidence for the inequality constrained hypothesis that is supported by the data when the unconstrained prior is located on the boundary" (Mulder, 2014b, p.452). Second, in contrast to Bayesian estimation, assigning very large values to the diagonal elements of the prior's covariance matrix Σ_t is not feasible in hypothesis testing. In hypothesis testing, we need to explicitly calculate the normalizing constant $c_t = \int_{\Theta_{\boldsymbol{\rho}_t}} \phi_{\boldsymbol{\mu}_t,\Sigma_t}(\boldsymbol{\rho}_t)\,\mathrm{d}\boldsymbol{\rho}_t$ and a vague formulation of Σ_t makes this computation either unstable or tremendously time consuming due to the fairly small parameter space $\Theta_{\boldsymbol{\rho}_t}$.[9] Instead, we set the prior covariance matrix Σ_t of the free network autocorrelation parameter(s) under a hypothesis, e.g., ρ_1 under hypothesis $H_1 : \rho_1 > \rho_2 = 0$, to the product of the corresponding asymptotic variance-covariance matrix of the maximum likelihood estimate of $\boldsymbol{\rho}_t$ and a hypothesis-specific scaling factor σ_t^2, similarly as in Zellner's g-prior (Zellner, 1986). In mathematical notation, $\Sigma_t = \sigma_t^2 I(\boldsymbol{\rho}_t)^{-1}$, where $I(\boldsymbol{\rho}_t)$ denotes the submatrix of the network autocorrelation model's Fisher information matrix $I\left(\boldsymbol{\rho}_t, \sigma^2, \boldsymbol{\beta}\right)$. Hence, there is only one free parameter in the prior specification of Σ_t left, σ_t^2. Following Hansen & Yu (2001) and Liang et al. (2008), we employ a local empirical Bayes approach and choose σ_t^2 such that the associated marginal likelihood $m_t(\boldsymbol{y})$ is maximized, avoiding arbitrary prior specification. As there is no analytical solution to this maximization problem, one way to approximate the maximum of $m_t(\boldsymbol{y})$ is to compute the marginal likelihood on a grid of increasing values for σ_t^2 until a stopping rule is reached, e.g., until the marginal likelihood is not increasing anymore, or until it is not increasing by more than some tolerance factor.[10]

[9] As $\Theta_{\boldsymbol{\rho}_t}$ is bounded and small, Bartlett's paradox (Bartlett, 1957) is not an issue in the considered tests on the network autocorrelations.

[10] The marginal likelihood $m_t(\boldsymbol{y})$ can be strictly increasing with σ_t^2, which is why we cannot use more efficient optimization techniques, such as Newton's method or the BFGS algorithm (Nocedal & Wright, 2006).

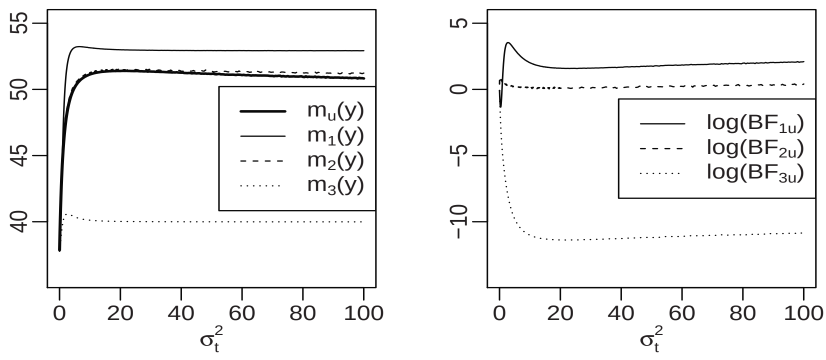

Figure 4.2 Marginal likelihoods $m_t(\boldsymbol{y})$, $t \in \{1, 2, 3, u\}$, under the hypotheses $H_1 : \rho_1 > \rho_2 = 0$, $H_2 : \rho_1 > \rho_2 > 0$, $H_3 : \rho_1 = \rho_2 = 0$, and $H_u : (\rho_1, \rho_2) \in \Theta_{(\rho_1,\rho_2)}$ as a function of σ_t^2 (left) and the logarithm of the Bayes factors $\log(BF_{1u})$, $\log(BF_{2u})$, and $\log(BF_{3u})$ as a function of σ_t^2 (right) for the data in Dall'erba et al. (2009) and model (4.4).

Figure 4.2 shows the marginal likelihoods under the three constrained hypotheses $H_1 : \rho_1 > \rho_2 = 0$, $H_2 : \rho_1 > \rho_2 > 0$, and $H_3 : \rho_1 = \rho_2 > 0$, the marginal likelihood under an unconstrained hypothesis $H_u : (\rho_1, \rho_2) \in \Theta_{(\rho_1,\rho_2)}$, and the logarithm of the Bayes factors of the three constrained hypotheses against the unconstrained hypothesis H_u as a function of σ_t^2, $t \in \{1, 2, 3, u\}$, for the data in Dall'erba et al. (2009) and model (4.4). As can be seen, all of the marginal likelihoods sharply increase for smaller values for σ_t^2 before they gradually decrease after having reached their respective maxima. At the same time, the associated Bayes factors, in which we are ultimately interested, appear fairly robust to the choice of σ_t^2, except for extremely small values for σ_t^2. For the vast majority of data sets we looked at, we observed essentially the same pattern with almost all of the optimal values for σ_t^2 laying between 2 and 10.

In summary, in this section we showed how researchers can use Bayes factors to test and quantify the evidence in the data for hypotheses with order constraints on the network autocorrelation parameters. Parallel hereto, we provided methodology to efficiently compute such Bayes factors without any need to subjectively elicit priors for the network effects. Altogether, this ultimately allows network scholars to test and verify any kind of expectations they have about the strength of different network effects.

4.5 Simulation study

We performed a simulation study to investigate the performance of the proposed Bayesian estimator and the proposed Bayes factors in a second-order network autocorrelation model. First, we compared the Bayesian estimator from Section 4.3.2 to the maximum likelihood estimator in terms of bias of the network effects and frequentist coverage of the corresponding credible and confidence intervals. Here, we use the term coverage to indicate the proportion of times in which the true, i.e., data-generating, network effects were contained in the credible and confidence intervals, respectively. Second, as researchers are generally interested in testing whether (some) network effects are zero or whether one network effect is larger than another one, we considered a multiple hypothesis test with the following five hypotheses: $H_1 : \rho_1 > \rho_2 = 0$, $H_2 : \rho_1 > \rho_2 > 0$, $H_3 : \rho_1 = \rho_2 > 0$, $H_4 : 0 < \rho_1 < \rho_2$, and $H_5 : 0 = \rho_1 < \rho_2$. We investigated if and how fast the different Bayes factors converge to a true data-generating hypothesis and how robust these findings are to various degrees of overlap between two connectivity matrices.

4.5.1 Study design

In our simulation study, we generated data $\boldsymbol{y}$ via $\boldsymbol{y} = A^{-1}_{(\rho_1,\rho_2)} \left(X\boldsymbol{\beta} + \boldsymbol{\varepsilon}\right), \boldsymbol{\varepsilon} \sim N\left(\mathbf{0}_g, I_g\right)$, for four network sizes g ($g \in \{50, 100, 200, 400\}$), three levels of overlap between W_1 and W_2 (0%, 20%, 40%), and both W_1 and W_2 having an average degree of four. We simulated random non-symmetric binary connectivity matrices using the rgraph() function from the sna package in R (Butts, 2008), randomly rearranged ties when accounting for overlap, and subsequently row-standardized the raw connectivity matrices. Furthermore, we drew independent values from a standard normal distribution for the elements of $X \in \mathbb{R}^{g\times 4}$ (excluding the first column which is a vector of ones), $\boldsymbol{\beta} \in \mathbb{R}^4$, and $\boldsymbol{\varepsilon} \in \mathbb{R}^g$.

In our first experiment, we set the two network effects to $(\rho_1, \rho_2) = (.2, .2)$ and simulated 1,000 data sets for each of the 12 scenarios (4 network sizes × 3 levels of overlap × 1 network effects size).[11] For the Bayesian estimator, we employed the standard improper prior $p\left(\sigma^2, \boldsymbol{\beta}\right) \propto 1/\sigma^2$ for the nuisance parameters and a non-informative bivariate normal prior for (ρ_1, ρ_2), $p\left(\rho_1, \rho_2\right) \propto N\left(\mathbf{0}_2, 100 \times I_2\right)$, which essentially corresponds to a uniform prior for $(\rho_1, \rho_2) \in \Theta_{(\rho_1,\rho_2)}$. We drew 1,000 realizations from the resulting posteriors by relying on the methods described in Section 4.3.2, taking the maximum likelihood estimate of $\left(\left(\rho_1, \rho_2\right), \sigma^2, \boldsymbol{\beta}\right)$ as starting value in the sampling algorithm, see Appendix 4.A. We used the marginal posterior median as point estimator and the 95% equal-tailed credible interval for coverage analysis. We obtained the maximum likelihood estimates as well as their standard errors and associated asymptotic confidence intervals applying the lnam() function from the sna package in R.

In our second experiment, we considered 41 network effects sizes (ρ_1, ρ_2) ($(\rho_1, \rho_2) \in \{(.4, 0), (.39, .01) ..., (0, .4)\}$) and simulated 100 data sets for each of the 492 scenarios (4 network sizes × 3 levels of overlap × 41 network effects sizes). Figure 4.3 shows the trajectory of the network effects and a graphical representation of the five tested hypotheses $H_1 : \rho_1 > \rho_2 = 0$, $H_2 : \rho_1 > \rho_2 > 0$, $H_3 : \rho_1 = \rho_2 > 0$, $H_4 : 0 < \rho_1 < \rho_2$, and $H_5 : 0 = \rho_1 < \rho_2$. We specified the prior under each of the five hypotheses based on the proposed empirical Bayes procedure in Section 4.4.3.[12] In order to compute the normalizing constants c_2 and c_4, we generated draws from the unconstrained bivariate normal prior for (ρ_1, ρ_2) until we obtained 1,000 draws in agreement with the constraints imposed

[11] Simulation results for different combinations of (ρ_1, ρ_2) are available from the authors upon request. We do not present them here as they do not provide any additional insights.

[12] We used the lnam() function from the sna package in R to obtain the asymptotic variance-covariance matrices of the maximum likelihood estimates of the free network autocorrelation parameters. Furthermore, we started with $\sigma_t^2 = 1$, $t \in \{1, 2, 3, 4, 5\}$, and kept increasing σ_t^2 by 1 until either the increment in the marginal likelihood $m_t\left(\boldsymbol{y}\right)$ was less than .01 or σ_t^2 reached the cut-off value 20.

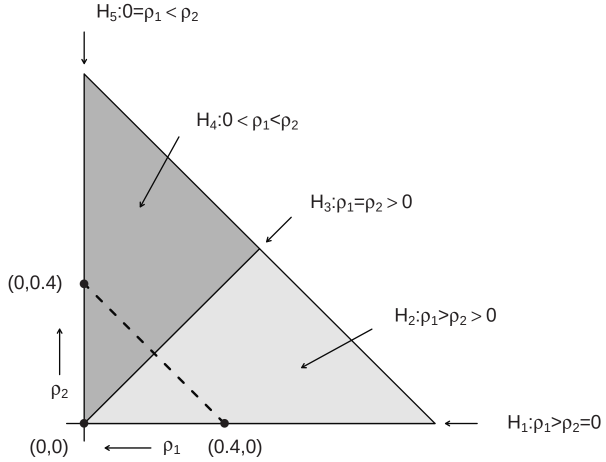

Figure 4.3 Graphical representation of the admissible subspaces of (ρ_1, ρ_2) under the five constrained hypotheses and the trajectory of the data-generating network effects $(\rho_1, \rho_2) = (.4, 0)$ (dashed line).

under hypothesis H_2 and hypothesis H_4, respectively. Then, we approximated the normalizing constants by the reciprocals of the proportion of the total number of draws in agreement with the constraints. For the hypotheses with only one free network autocorrelation parameter, i.e., $H_1 : \rho_1 > \rho_2 = 0$, $H_3 : \rho_1 = \rho_2 > 0$, and $H_5 : 0 = \rho_1 < \rho_2$, we directly obtained the corresponding normalizing constants by using the pnorm() function in R, as the bounds of the feasible range of a single free network autocorrelation parameter are known exactly, see Section 4.2.1. Finally, for all hypotheses we drew 1,000 realizations from their (unconstrained) importance densities and computed the logarithm of the Bayes factor of each constrained hypothesis against an unconstrained reference hypothesis $H_u : (\rho_1, \rho_2) \in \Theta_{(\rho_1,\rho_2)}$.[13]

4.5.2 Simulation results

Table 4.2 shows the average estimates and root mean squared errors of ρ_1 and ρ_2 for the Bayesian as well as the maximum likelihood estimator. Overall, the two estimators yield nearly identical results for all considered scenarios. As expected, the (negative) bias in the estimation of the network effects and the associated root mean squared errors are decreasing with the network size, the bias being virtually non-existent for $g = 400$. Introducing 20% and 40% of overlap between two connectivity matrices does not appear to impact the estimation results, even if there is mild negative correlation between the estimated network effects in these cases, see Figure 4.4.

[13] For the hypotheses with only one free network autocorrelation parameter, i.e., $H_1 : \rho_1 > \rho_2 = 0$, $H_3 : \rho_1 = \rho_2 > 0$, and $H_5 : 0 = \rho_1 < \rho_2$, we directly generated 1,000 draws in agreement with the respective constraints by using the rtruncnorm() function from the truncnorm package in R (Trautmann et al., 2015). As the unconstrained hypothesis H_u only serves as reference hypothesis to which all other hypotheses are compared to, we did not maximize over σ_u^2 when computing $m_u(\boldsymbol{y})$ but fixed σ_u^2 to 5.

Table 4.2 Average posterior median and maximum likelihood estimates of $(\rho_1, \rho_2) = (.2, .2)$ and corresponding average root mean squared errors (RMSE) for 1,000 simulated data sets.

	0% overlap				20% overlap				40% overlap			
	Estimate		RMSE		Estimate		RMSE		Estimate		RMSE	
	ρ_1	ρ_2	ρ_1	ρ_2	ρ_1	ρ_2	ρ_1	ρ_2	ρ_1	ρ_2	ρ_1	ρ_2
$g = 50$												
Bayes	.149	.152	.155	.157	.160	.148	.163	.151	.159	.166	.168	.168
MLE	.157	.160	.156	.157	.167	.154	.164	.151	.164	.172	.168	.169
$g = 100$												
Bayes	.180	.182	.107	.104	.178	.184	.111	.107	.188	.183	.107	.107
MLE	.182	.184	.108	.104	.179	.186	.111	.107	.189	.185	.108	.108
$g = 200$												
Bayes	.189	.187	.074	.074	.198	.190	.076	.075	.194	.195	.077	.077
MLE	.190	.188	.074	.074	.198	.190	.076	.075	.194	.195	.078	.077
$g = 400$												
Bayes	.196	.196	.052	.051	.197	.197	.050	.051	.198	.196	.054	.053
MLE	.196	.196	.052	.051	.197	.197	.050	.051	.198	.196	.053	.053

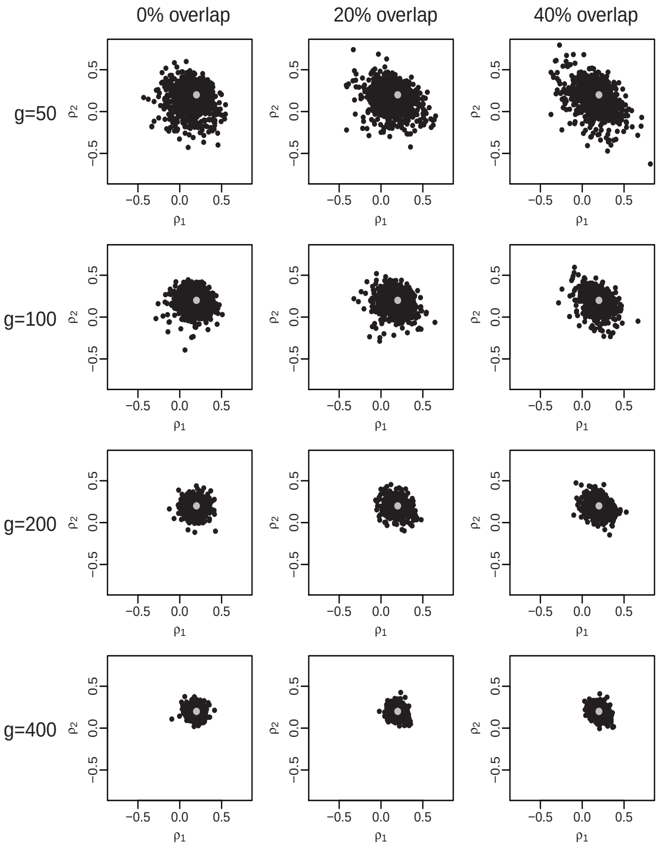

Figure 4.4 Posterior median estimates (black) of $(\rho_1, \rho_2) = (.2, .2)$ (gray) for 1,000 simulated data sets.

Table 4.3 Empirical frequentist coverage of 95% credible and confidence intervals for ρ_1 and ρ_2 for 1,000 simulated data sets.

	0% overlap		20% overlap		40% overlap	
	ρ_1	ρ_2	ρ_1	ρ_2	ρ_1	ρ_2
$g = 50$						
Bayes	.953	.949	.936	.946	.948	.943
MLE	.928	.928	.912	.937	.923	.930
$g = 100$						
Bayes	.949	.958	.949	.946	.957	.961
MLE	.936	.950	.942	.936	.951	.955
$g = 200$						
Bayes	.943	.948	.947	.953	.954	.937
MLE	.934	.943	.950	.951	.951	.932
$g = 400$						
Bayes	.954	.948	.964	.940	.953	.943
MLE	.955	.949	.964	.946	.953	.943

Table 4.3 reports the empirical frequentist coverage of Bayesian equal-tailed 95% credible intervals and asymptotic 95% maximum likelihood-based confidence intervals for ρ_1 and ρ_2. The coverage of Bayesian credible intervals is very close to the nominal .95 for all considered scenarios, while the coverage of confidence intervals is below nominal for network sizes of 50. These observations are in line with the subpar coverage of maximum likelihood-based confidence intervals for small samples in the first-order network autocorrelation model in Chapter 2.

Based on the results from our first simulation experiment, we draw two main conclusions. First, we recommend using the non-informative Bayesian estimator over the maximum likelihood estimator, as both estimators yield nearly identical network effect estimates but the coverage of Bayesian credible intervals appears accurate, whereas for smaller network sizes the coverage of maximum likelihood-based confidence intervals is not. Second, estimating second-order network autocorrelation models with moderately overlapping connectivity matrices, i.e., with up to 40% shared ties, does not alter the estimation of the network effects. This second finding is of particular importance to social network researchers who often encounter distinct but partially overlapping networks in empirical practice.

Figure 4.5 displays the average logarithm of the Bayes factors of the hypotheses H_1 (thick solid line), H_2 (thick dashed line), H_3 (dotted line), H_4 (dashed line), and H_5 (solid line) against an unconstrained reference hypothesis H_u as a function of network effects (ρ_1, ρ_2) from $(.4, 0)$ to $(, 0.4)$. Overall, the results indicate that the Bayes factors show consistent behavior, i.e., there is the most evidence for the data-generating hypothesis if the network size is large enough. This evidence is monotonically increasing with the network size. In particular, there is little discrimination between the five hypotheses for $g = 50$, while there is clear support for the data-generating hypothesis for network sizes of 200 and 400. Two of the lines in Figure 4.5 are discontinued due to numerical reasons: when computing the Bayes factor, we need to calculate the probability mass of the unconstrained importance density contained in the parameter space imposed by

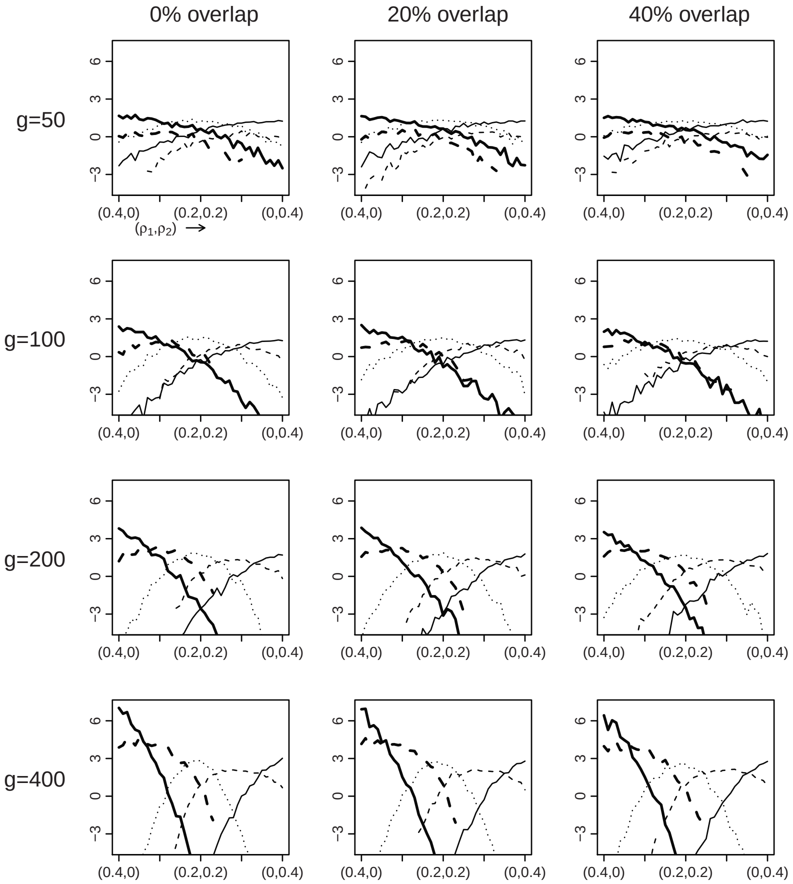

Figure 4.5 Average logarithm of the Bayes factors $\log(B_{tu})$, $t \in \{1, 2, 3, 4, 5\}$, of the hypotheses $H_1 : \rho_1 > \rho_2 = 0$ (thick solid line), $H_2 : \rho_1 > \rho_2 > 0$ (thick dashed line), $H_3 : \rho_1 = \rho_2 > 0$ (dotted line), $H_4 : 0 < \rho_1 < \rho_2$ (dashed line), and $H_5 : 0 = \rho_1 < \rho_2$ (solid line) against $H_u : (\rho_1, \rho_2) \in \Theta_{(\rho_1,\rho_2)}$ as function of network effects (ρ_1, ρ_2) from $(.40, 0)$ to $(0, .40)$ for 100 simulated data sets.

the constraints under a hypothesis, see Appendix 4.B. For the Bayes factors involving the purely inequality constrained hypotheses $H_2 : \rho_1 > \rho_2 > 0$ and $H_4 : 0 < \rho_1 < \rho_2$, we approximated these probabilities numerically by the proportion of 1,000 draws from the unconstrained importance densities that were in agreement with hypothesis H_2 and hypothesis H_4, respectively. For some data sets, however, none of the draws were in agreement with hypothesis H_2, or hypothesis H_4, in which case we set the corresponding marginal likelihood to $-\infty$. If this happened for at least one of the 100 simulated data sets, then the average logarithm of the Bayes factor was $-\infty$ as well. Finally, as in our first simulation experiment, these findings are robust to moderate degrees of overlap between two connectivity matrices.

4.6 Application revisited

In this section, we re-analyze a data set from the economic growth literature initially studied by Dall'erba et al. (2009) and address the questions raised in Section 4.2.3. First, we re-estimated the second-order network autocorrelation model in (4.4) based on non-informative priors for all model parameters and compared the results to those coming from maximum likelihood estimation. Second, we used Bayes factors to quantify the relative evidence in the data for different competing hypotheses of interest with respect to this data set. Finally, we considered a network autocorrelation model with two subgroups, assuming only one dominant common influence mechanism within and between the two subgroups.

4.6.1 Bayesian estimation of a second-order network autocorrelation model

Table 4.4 displays the results of a Bayesian estimation of the second-order model in (4.4), along with the corresponding maximum likelihood estimates.[14] The Bayesian and the maximum likelihood estimates of all parameters are similar to each other, in line with the results from our simulation study in Section 4.5.2. In particular, the (Bayesian) estimate of ρ_1, reflecting interactions within the same country, is of large positive magnitude (.350), while the (Bayesian) estimate of ρ_2, reflecting spillovers from regions in neighboring countries, is much smaller and close to zero (-.058). Dall'erba et al. (2009) concluded by saying that "the results obtained also confirm the hypothesis that economic interactions decrease very substantially when a national border is passed (indeed, the coefficient reflecting external spillovers is not statistically significant)" (Dall'erba et al., 2009, p.342).

4.6.2 Bayesian hypothesis testing in a second-order network autocorrelation model

Using Bayes factors, we quantified the evidence in the data for two hypotheses representing the notion of decreasing economic interactions once a national border is passed, $H_1 : \rho_1 > \rho_2 = 0$ and $H_2 : \rho_1 > \rho_2 > 0$, and tested them against two competing hypotheses, $H_0 : \rho_1 = \rho_2 = 0$ and $H_3 : \rho_1 = \rho_2 > 0$.[15] Furthermore, we also included a hypothesis $H_c : \neg\left(H_0 \vee H_1 \vee H_2 \vee H_3\right)$ in our test that consists of the complement of all other possible hypotheses on (ρ_1, ρ_2) except hypotheses H_0, H_1, H_2, and H_3, i.e., which contains all the orders of network effects we did not hypothesize.

Table 4.5 provides the Bayes factors for every pair out of the set of the five considered hypotheses above using the prior specifications from Sections 4.4.2 and 4.4.3. Notably,

[14] As in our simulation study in Section 4.5, we used the prior $p\left(\rho_1, \rho_2, \sigma^2, \boldsymbol{\beta}\right) \propto N\left(\mathbf{0}_2, 100 \times I_2\right) \times 1/\sigma^2$ and relied on the lnam() function from the sna package in R to compute the maximum likelihood estimates and the corresponding confidence intervals.

[15] We considered the hypothesis $H_2 : \rho_1 > \rho_2 > 0$ rather than a hypothesis $H_{2'} : \rho_1 > \rho_2$ here, as the hypotheses H_1 and $H_{2'}$ would not be unambiguous. This does not mean that hypothesis $H_{2'}$ in itself cannot be tested against other hypotheses, only the combination of hypotheses in a given test has to be considered carefully.

Table 4.4 Posterior median and maximum likelihood estimates and associated 95% Bayesian credible and confidence intervals (in brackets) for the data in Dall'erba et al. (2009) and model (4.4).

Parameter	Bayes	MLE
ρ_1	.350	.348
	$(.238, .464)$	$(.235, .460)$
ρ_2	-.058	-.058
	$(-.169, .052)$	$(-.168, .052)$
Intercept	-.682	-.696
	$(-.887, -.451)$	$(-.924, -.469)$
Market Service Growth	.484	.483
	$(.384, .585)$	$(.385, .580)$
Productivity Gap	.212	.218
	$(.122, .296)$	$(.127, .309)$
Urbanization	3.094×10^{-5}	3.119×10^{-5}
	$\left(4.728 \times 10^{-6}, 5.703 \times 10^{-5}\right)$	$\left(5.894 \times 10^{-6}, 5.648 \times 10^{-5}\right)$
Accessibility	1.052×10^{-5}	1.164×10^{-5}
	$\left(-3.937 \times 10^{-6}, 2.516 \times 10^{-5}\right)$	$\left(-3.132 \times 10^{-6}, 2.642 \times 10^{-5}\right)$

Table 4.5 Bayes factors $B_{tt'}$, $t, t' \in \{0, 1, 2, 3, c\}$, for the hypotheses $H_0 : \rho_1 = \rho_2 = 0$, $H_1 : \rho_1 > \rho_2 = 0$, $H_2 : \rho_1 > \rho_2 > 0$, $H_3 : \rho_1 = \rho_2 > 0$, and $H_c : \neg (H_0 \vee H_1 \vee H_2 \vee H_3)$ for the data in Dall'erba et al. (2009) and model (4.4).

Hypothesis	H_0	H_1	H_2	H_3	H_c
H_0	-	1.893×10^{-7}	3.043×10^{-5}	.059	2.966×10^{-6}
H_1	5.283×10^{6}	-	160.746	3.117×10^{5}	14.085
H_2	3.286×10^{4}	6.220×10^{-3}	-	1.939×10^{3}	.089
H_3	16.945	3.208×10^{-6}	5.157×10^{-4}	-	4.539×10^{-5}
H_c	3.732×10^{5}	.071	11.359	2.203×10^{4}	-

$H_1 : \rho_1 > \rho_2 = 0$ is the hypothesis most supported by the data and approximately 5.283×10^6, 160.746, 3.117×10^5, and 14.085 times more supported than hypothesis H_0, H_2, H_3, and H_c, respectively. Moreover, there is the least evidence for the null in the data. Consequently, regardless the specification of alternative expectations about ρ_1 and ρ_2, the hypothesis that both network effects are zero has to be strongly rejected. Although these implications seem in line with the authors' claim that network effects are decreasing after a national border is passed, using Bayes factors provides us with much more extensive conclusions about the characteristic evidence in the data. Hence, we can now quantify how much more likely these conclusions are than competing conclusions (hypotheses) and how (un)likely it is that an entirely different mechanism H_c generated the data. Ultimately, in this data set there is the most and very strong evidence for a positive within-country network effect only.

4.6.3 Bayesian hypothesis testing in a fourth-order network autocorrelation model

Dall'erba et al. (2009) also pointed out potentially asymmetric growth rates across the regions, depending on the initial productivity level of a region. Thus, the authors pro-

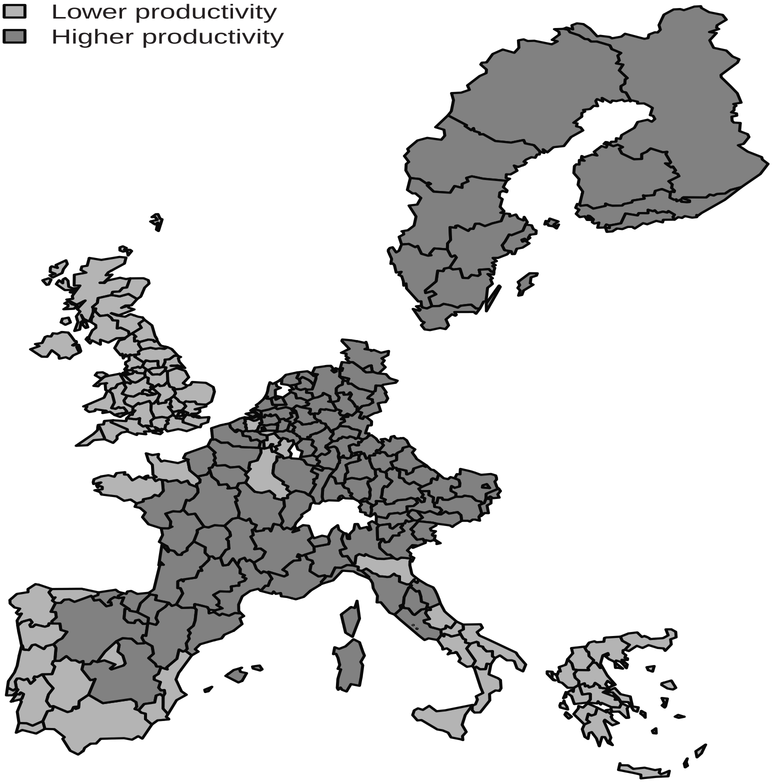

Figure 4.6 Spatial distribution of productivity levels in 1980 across the 188 regions.

ceeded by dividing the sample into two clusters; 111 initially more productive regions and 77 initially less productive regions, implying a core-periphery pattern, see Figure 4.6.[16] Next, they separately estimated two second-order network autocorrelation models for the two clusters. Here, for illustrative purposes, we allow for varying levels of network autocorrelation within and between the two clusters and consider a model with two subgroups instead. For example, we could expect the network effects among regions of the same subgroup to be larger than the network effects between regions of different subgroups, or we could expect the initially more productive regions to influence the initially less productive ones more strongly than the other way around.

Our previous analyses in Section 4.6.2 suggested that there is very strong evidence for a positive within-country network effect only, i.e., $\rho_1 > 0$ and $\rho_2 = 0$. Thus, we merely

[16] We created the map of the European NUTS-2 regions by using Eurostat data from `http://ec.europa.eu/eurostat/cache/GISCO/geodatafiles/NUTS_2010_60M_SH.zip`, file *NUTS_RG_60M_2010.shp*, and the readOGR() function from the rgdal package in R (Bivand et al., 2017). The depicted map shows 189 instead of the original 188 regions, as the Trentino-Alto Adige region in Italy has been divided into two NUTS-2 regions.

consider spillover effects within the same country, in other words, we assume that only W_1 plays a role, not W_2. We denote by $\rho_{hh}, \rho_{hl}, \rho_{lh}$, and ρ_{ll} the network effect within regions with initially higher productivity levels, the network effect of the initially less productive regions on the initially more productive regions, the network effect of the initially more productive regions on the initially less productive ones, and the network effect within regions with initially lower productivity levels, respectively. Accordingly, $\boldsymbol{y}_h \in \mathbb{R}^{111}$ and $\boldsymbol{y}_l \in \mathbb{R}^{77}$ contain the growth rates of labor productivity of the initially more and the initially less productive regions, respectively, and we partitioned W_1, the unstandardized connectivity matrix using the three nearest neighbors of a region within the same state, into the four submatrices $W_{hh} \in \mathbb{R}^{111\times111}$, $W_{hl} \in \mathbb{R}^{111\times77}$, $W_{lh} \in \mathbb{R}^{77\times111}$, and $W_{ll} \in \mathbb{R}^{77\times77}$, representing ties within and between the two subgroups.[17] This resulted in the following fourth-order network autocorrelation model

$$\begin{aligned}\boldsymbol{y} = \begin{bmatrix}\boldsymbol{y}_h\\ \boldsymbol{y}_l\end{bmatrix} &= \begin{bmatrix}\rho_{hh}W_{hh} & \rho_{hl}W_{hl}\\ \rho_{lh}W_{lh} & \rho_{ll}W_{ll}\end{bmatrix}\begin{bmatrix}\boldsymbol{y}_h\\ \boldsymbol{y}_l\end{bmatrix} + \beta_1\boldsymbol{X}_{\cdot1} + \beta_2\boldsymbol{X}_{\cdot2} + \beta_3\boldsymbol{X}_{\cdot3} + \beta_4\boldsymbol{X}_{\cdot4} + \beta_5\boldsymbol{X}_{\cdot5} + \boldsymbol{\varepsilon}\\ &= \left(\rho_{hh}\begin{bmatrix}W_{hh} & 0\\ 0 & 0\end{bmatrix} + \rho_{hl}\begin{bmatrix}0 & W_{hl}\\ 0 & 0\end{bmatrix} + \rho_{lh}\begin{bmatrix}0 & 0\\ W_{lh} & 0\end{bmatrix} + \rho_{ll}\begin{bmatrix}0 & 0\\ 0 & W_{ll}\end{bmatrix}\right)\begin{bmatrix}\boldsymbol{y}_h\\ \boldsymbol{y}_l\end{bmatrix} \qquad (4.17)\\ &\quad + \beta_1\boldsymbol{X}_{\cdot1} + \beta_2\boldsymbol{X}_{\cdot2} + \beta_3\boldsymbol{X}_{\cdot3} + \beta_4\boldsymbol{X}_{\cdot4} + \beta_5\boldsymbol{X}_{\cdot5} + \boldsymbol{\varepsilon}.\end{aligned}$$

We generally expect the network effects within the two subgroups to be larger than the network effects between subgroups, i.e., $\{\rho_{hh}, \rho_{ll}\} > \{\rho_{hl}, \rho_{lh}\}$, where the ">" sign holds pairwise for any two elements of the first and second set, respectively. Furthermore, hypotheses of substantial interest might be based on expectations of positive network effects within both subgroups but with potentially differing magnitudes. We translated these expectations to the hypotheses $H_1 : \{\rho_{hh} > \rho_{ll} > 0\} \wedge \{\{\rho_{hh}, \rho_{ll}\} > \{\rho_{hl}, \rho_{lh}\}\}$, $H_2 : \{\rho_{hh} = \rho_{ll} > 0\} \wedge \{\{\rho_{hh}, \rho_{ll}\} > \{\rho_{hl}, \rho_{lh}\}\}$, and $H_3 : \{0 < \rho_{hh} < \rho_{ll}\} \wedge \{\{\rho_{hh}, \rho_{ll}\} > \{\rho_{hl}, \rho_{lh}\}\}$.[18] We supplemented these three hypotheses with the hypothesis of no network effects, $H_0 : \rho_{hh} = \rho_{hl} = \rho_{lh} = \rho_{ll} = 0$, and the complement of all the orders of network effects we did not have hypotheses for, $H_c : \neg\,(H_0 \vee H_1 \vee H_2 \vee H_3)$.

Table 4.6 shows the Bayes factors for every pair out of the set of the five considered hypotheses. As can be seen, $H_2 : \{\rho_{hh} = \rho_{ll} > 0\} \wedge \{\{\rho_{hh}, \rho_{ll}\} > \{\rho_{hl}, \rho_{lh}\}\}$ is the hypothesis that is most supported by the data and receives approximately 3.149×10^5, 3.222, 3.704, and 55.556 more support than hypothesis H_0, H_1, H_3, and H_c, respectively. Hence, there is no evidence in the data for differing network effects within the initially more and the initially less productive regions, while there is very strong evidence for the network effects within the two subgroups to be larger than the network effects between the subgroups.

[17] We row-standardized W_{hh}, W_{hl}, W_{lh}, and W_{ll} separately.

[18] Another hypothesis of interest could be that (with respect to the growth of labor productivity) the initially more productive regions influence the initially less productive ones stronger than the other way around, i.e., $\rho_{lh} > \rho_{hl}$. We did not include this hypothesis, as it would overlap with the other considered hypotheses and because of the undesirable behavior of such overlapping-hypotheses Bayes factors (Morey & Rouder, 2011). Separate analyses showed, however, that there is actually no evidence for a hypothesis involving $\rho_{lh} > \rho_{hl}$.

Table 4.6 Bayes factors $B_{tt'}$, $t, t' \in \{0, 1, 2, 3, c\}$, for the hypotheses $H_0 : \rho_{hh} = \rho_{hl} = \rho_{lh} = \rho_{ll} = 0$, $H_1 : \{\rho_{hh} > \rho_{ll} > 0\} \wedge \{\rho_{hh}, \rho_{ll}\} > \{\rho_{hl}, \rho_{lh}\}$, $H_2 : \{\rho_{hh} = \rho_{ll} > 0\} \wedge \{\rho_{hh}, \rho_{ll}\} > \{\rho_{hl}, \rho_{lh}\}$, $\{H_3 : 0 < \rho_{hh} < \rho_{ll}\} \wedge \{\rho_{hh}, \rho_{ll}\} > \{\rho_{hl}, \rho_{lh}\}$, and $H_c : \neg(H_0 \vee H_1 \vee H_2 \vee H_3)$ for the data in Dall'erba et al. (2009) and model (4.17).

Hypothesis	H_0	H_1	H_2	H_3	H_c
H_0	-	1.023×10^{-5}	3.176×10^{-6}	1.177×10^{-5}	1.800×10^{-4}
H_1	9.773×10^4	-	.310	1.151	17.241
H_2	3.149×10^5	3.222	-	3.704	55.556
H_3	8.497×10^4	.869	.270	-	14.925
H_c	5.563×10^3	.058	.018	.067	-

4.7 Conclusions

In this chapter, we developed Bayesian techniques for estimating and testing higher-order network autocorrelation models with multiple network autocorrelations. In particular, we provided default Bayes factors that enable researchers to test hypotheses with order constraints on the network effects in a direct manner. Thus, the proposed methods allow researchers to simultaneously test any number of competing hypotheses on the relative strength of network effects against each other and to quantify the amount of evidence in the data for any of these hypotheses. This has not yet been possible using the currently available statistical techniques in the literature on network autocorrelation models.

We ran an extensive simulation study to evaluate the numerical behavior of the presented Bayesian procedures for a number of different network specifications, including varying network sizes and network overlap. First, we found that the Bayesian estimator based on a non-informative prior and the maximum likelihood estimator have comparable frequentist properties under most scenarios, except for smaller network sizes. For smaller network sizes, only the Bayesian estimator exhibits accurate coverage of credible intervals and overall shows slightly superior performance. Second, we observed that the introduced Bayes factors always result in the largest evidence for a true data-generating hypothesis, with this evidence increasing with the network size. Furthermore, we also provided efficient algorithms for sampling from the posterior distributions and for computing the Bayes factors. We illustrated the practical utility of the Bayes factors by applying them to a data set on economic growth in 188 European regions. This resulted in additional and more precise insights into how the various network effects are related to each other in comparison to classical null hypothesis significance testing.

Given the many, often implicit, expectations researchers have about the relative importance of different network effects, we hope that by enabling researchers to test these expectations directly and explicitly, higher-order network autocorrelation models will bring for a more thorough understanding of social contagion processes that goes beyond the current state of the art.

Acknowledgment

We thank Sandy Dall'erba, Marco Percoco, and Gianfranco Piras for sharing their data.

Appendix 4.A Posterior sampling

We outlined the procedure for sampling from the full posterior $p\left(\boldsymbol{\rho}, \sigma^2, \boldsymbol{\beta}|\boldsymbol{y}\right)$ for higher-order network autocorrelation models in Section 4.3.2. However, it remains to specify the exact form of the candidate-generating distribution for the conditional posterior $p\left(\boldsymbol{\rho}, \beta_1|\sigma^2, \widetilde{\boldsymbol{\beta}}, \boldsymbol{y}\right)$ and the expressions $\boldsymbol{\mu}_{\widetilde{\boldsymbol{\beta}}}$ and $\Sigma_{\widetilde{\boldsymbol{\beta}}}$ in (4.12).

Approximating the conditional posterior for $(\boldsymbol{\rho}, \beta_1)$

In the following, we show how to approximate $p\left(\boldsymbol{\rho}, \beta_1|\sigma^2, \widetilde{\boldsymbol{\beta}}, \boldsymbol{y}\right)$ by a $(R+1)$-variate normal distribution. First, by Jacobi's formula (see e.g., Hall, 2003, Theorem 2.11), for any complex matrix X it holds that $\det\left(\exp\left(X\right)\right) = \left|\exp\left(X\right)\right| = \exp\left(\operatorname{tr}\left(X\right)\right)$. As we set the R-dimensional parameter space of $\boldsymbol{\rho}$ as the space containing the origin for which $A_{\boldsymbol{\rho}}$ is non-singular, we know that $|A_{\boldsymbol{\rho}}| > 0$, $A_{\boldsymbol{\rho}}$ is invertible, and that $\log\left(A_{\boldsymbol{\rho}}\right)$ exists (see e.g., Higham, 2008, Theorem 1.27). Thus, we can write for $X := A_{\boldsymbol{\rho}} = I_g - \sum_{r=1}^{R} \rho_r W_r$

$$\begin{aligned} &\left|\exp\left(A_{\boldsymbol{\rho}}\right)\right| = \exp\left(\operatorname{tr}\left(A_{\boldsymbol{\rho}}\right)\right) \\ \Leftrightarrow &\left|\exp\left(\log\left(A_{\boldsymbol{\rho}}\right)\right)\right| = \exp\left(\operatorname{tr}\left(\log\left(A_{\boldsymbol{\rho}}\right)\right)\right) \\ \Leftrightarrow &\log\left(\left|A_{\boldsymbol{\rho}}\right|\right) = \operatorname{tr}\left(\log\left(A_{\boldsymbol{\rho}}\right)\right). \end{aligned} \tag{4.18}$$

Using the Mercator series for the matrix logarithm (see e.g., Hall, 2003, Theorem 2.7), we can rewrite (4.18) as

$$\begin{aligned} \log\left(\left|A_{\boldsymbol{\rho}}\right|\right) = \operatorname{tr}\left(\log\left(A_{\boldsymbol{\rho}}\right)\right) &= \operatorname{tr}\left(\sum_{m=1}^{\infty} (-1)^{m+1} \frac{\left(A_{\boldsymbol{\rho}} - I_g\right)^m}{m}\right) \\ &= \sum_{m=1}^{\infty} (-1)^{m+1} \frac{1}{m} \operatorname{tr}\left(\left(-\sum_{r=1}^{R} \rho_r W_r\right)^m\right). \end{aligned} \tag{4.19}$$

The first two sum terms in (4.19) are given by

$$\begin{aligned} m = 1: \quad & (-1)^2 \operatorname{tr}\left(-\sum_{r=1}^{R} \rho_r W_r\right) = -\sum_{r=1}^{R} \rho_r \operatorname{tr}\left(W_r\right) = 0, \\ m = 2: \quad & (-1)^3 \frac{1}{2} \operatorname{tr}\left(\left(-\sum_{r=1}^{R} \rho_r W_r\right)^2\right) = -\frac{1}{2} \sum_{r,r'=1}^{R} \rho_r \rho_{r'} \operatorname{tr}\left(W_r W_{r'}\right). \end{aligned}$$

Hence, $\log\left(\left|A_{\boldsymbol{\rho}}\right|\right)$ can be approximated by a quadratic polynomial in $\boldsymbol{\rho}$ as

$$\log\left(|A_{\boldsymbol{\rho}}|\right) \approx -\frac{1}{2}\sum_{r,r'=1}^{R} \rho_r\rho_{r'}\operatorname{tr}\left(W_rW_{r'}\right),$$
$$\Rightarrow |A_{\boldsymbol{\rho}}| \approx \exp\left(-\frac{1}{2}\sum_{r,r'=1}^{R} \rho_r\rho_{r'}\operatorname{tr}\left(W_rW_{r'}\right)\right). \tag{4.20}$$

Second, after some algebraic manipulation,

$$\exp\left(-\frac{1}{2}\left(\boldsymbol{\rho}-\boldsymbol{\mu}\right)^T\Sigma^{-1}\left(\boldsymbol{\rho}-\boldsymbol{\mu}\right)\right) \propto \exp\left(-\frac{1}{2}\left(\sum_{r,r'=1}^{R}\rho_r\rho_{r'}\Sigma^{-1}_{rr'} - 2\sum_{r,r'=1}^{R}\rho_r\Sigma^{-1}_{rr'}\mu_{r'}\right)\right), \tag{4.21}$$

$$\exp\left(-\frac{1}{2\sigma^2}\left(A_{\boldsymbol{\rho}}\boldsymbol{y}-X\boldsymbol{\beta}\right)^T\left(A_{\boldsymbol{\rho}}\boldsymbol{y}-X\boldsymbol{\beta}\right)\right)$$
$$\propto \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{r,r'=1}^{R}\rho_r\rho_{r'}\boldsymbol{y}^TW_r^TW_{r'}\boldsymbol{y} - 2\sum_{r=1}^{R}\rho_r\boldsymbol{y}^TW_r^T\left(\boldsymbol{y}-\widetilde{X}\widetilde{\boldsymbol{\beta}}\right)\right.\right.$$
$$\left.\left.+2\sum_{r=1}^{R}\rho_r\beta_1\boldsymbol{y}^TW_r^T\mathbf{1}_g - 2\beta_1\mathbf{1}_g^T\left(\boldsymbol{y}-\widetilde{X}\widetilde{\boldsymbol{\beta}}\right)+\beta_1^2g\right)\right), \tag{4.22}$$

where the proportionality in (4.21) holds with respect to $\boldsymbol{\rho}$, in (4.22) with respect to $(\boldsymbol{\rho}, \beta_1)$, $\widetilde{X}$ denotes the matrix X with its first column removed, $\widetilde{\boldsymbol{\beta}} = (\beta_2, ..., \beta_k)$, and $\mathbf{1}_g$ is the vector of ones of length g. By equating the coefficients of the product of (4.20), (4.21), and (4.22), i.e., the approximated conditional posterior for $(\boldsymbol{\rho}, \beta_1)$, to the kernel of the probability density function of a multivariate normal distribution $N\left(\boldsymbol{\mu}_{MH}, \Sigma_{MH}\right)$, we obtain for the coefficients of Σ^{-1}_{MH}

$$\Sigma^{-1}_{MH\,rr'} = \operatorname{tr}\left(W_rW_{r'}\right) + \Sigma^{-1}_{rr'} + \boldsymbol{y}^TW_r^TW_{r'}\boldsymbol{y}/\sigma^2, \quad r, r' \in \{1, ..., R\},$$
$$\Sigma^{-1}_{MH\,(R+1)r'} = \Sigma^{-1}_{MH\,r'(R+1)} = \boldsymbol{y}^TW_{r'}^T\mathbf{1}_g/\sigma^2,$$
$$\Sigma^{-1}_{MH\,(R+1)(R+1)} = g/\sigma^2,$$

and $\boldsymbol{\mu}_{MH} = \Sigma_{MH}\boldsymbol{z}_{MH}$, where $\boldsymbol{z}_{MH}$ is a vector of length $R+1$ with

$$z_{MH_r} = \boldsymbol{y}^TW_r^T\left(\boldsymbol{y}-\widetilde{X}\widetilde{\boldsymbol{\beta}}\right)/\sigma^2 + \sum_{r'=1}^{R}\Sigma^{-1}_{rr'}\boldsymbol{\mu},$$
$$z_{MH_{R+1}} = \mathbf{1}_g^T\left(\boldsymbol{y}-\widetilde{X}\widetilde{\boldsymbol{\beta}}\right)/\sigma^2.$$

If the $(R \times R)$ matrix $T, T_{rr'} := \operatorname{tr}\left(W_rW_{r'}\right)$, is not positive definite, Σ^{-1}_{MH}, and consequently Σ_{MH}, may not be positive definite either. In this case, we consider the nearest positive definite matrix to Σ^{-1}_{MH} as Σ^{-1}_{MH} instead. Subsequently, we use $q_{MH}\left(\boldsymbol{\rho}, \beta_1\right) = \phi_{\boldsymbol{\mu}_{MH},\Sigma_{MH}}\left(\boldsymbol{\rho}, \beta_1\right)$ as candidate-generating density in the Metropolis-Hastings algorithm.

Sampling from the conditional posterior for $\widetilde{\boldsymbol{\beta}}$

The $(k-1)$-variate normal conditional posterior for $\widetilde{\boldsymbol{\beta}}$ in (4.12) can be directly sampled from, with its mean vector and covariance matrix given by $\boldsymbol{\mu}_{\widetilde{\boldsymbol{\beta}}} = \boldsymbol{\mu}_{\boldsymbol{\beta}_2} + \Sigma_{\boldsymbol{\beta}_{21}} \Sigma_{\beta_{11}}^{-1} (\beta_1 - \mu_{\beta_1})$ and $\Sigma_{\widetilde{\boldsymbol{\beta}}} = \Sigma_{\boldsymbol{\beta}_{22}} - \Sigma_{\boldsymbol{\beta}_{21}} \Sigma_{\beta_{11}}^{-1} \Sigma_{\boldsymbol{\beta}_{12}}$, where

$$\boldsymbol{\mu}_{\boldsymbol{\beta}} = \left(X^T X\right)^{-1} X^T A_{\boldsymbol{\rho}} \boldsymbol{y} = \begin{pmatrix} \mu_{\beta_1} \\ \boldsymbol{\mu}_{\boldsymbol{\beta}_2} \end{pmatrix} \quad \text{sized} \quad \begin{pmatrix} 1 \times 1 \\ (k-1) \times 1 \end{pmatrix}, \tag{4.23}$$

$$\Sigma_{\boldsymbol{\beta}} = \sigma^2 \left(X^T X\right)^{-1} = \begin{pmatrix} \Sigma_{\beta_{11}} & \Sigma_{\boldsymbol{\beta}_{12}} \\ \Sigma_{\boldsymbol{\beta}_{21}} & \Sigma_{\boldsymbol{\beta}_{22}} \end{pmatrix} \quad \text{sized} \quad \begin{pmatrix} 1 \times 1 & 1 \times (k-1) \\ (k-1) \times 1 & (k-1) \times (k-1) \end{pmatrix}. \tag{4.24}$$

Sampling algorithm

The full sampling algorithm for drawing from the posterior $p\left(\boldsymbol{\rho}, \sigma^2, \boldsymbol{\beta} | \boldsymbol{y}\right)$ in higher-order network autocorrelation models can be written as follows:

(1) Set starting values $\left(\boldsymbol{\rho}^0, \beta_1^0\right), \left(\sigma^2\right)^0$, and $\widetilde{\boldsymbol{\beta}}^0$, e.g., to their maximum likelihood estimates, and the number of draws N.

(2) Repeat steps (3) - (5) for i=1:N.

(3) Perform a Metropolis-Hastings step for $(\boldsymbol{\rho}, \beta_1)$ with the target density $p\left(\boldsymbol{\rho}, \beta_1 | \sigma^2, \widetilde{\boldsymbol{\beta}}, \boldsymbol{y}\right)$ and the candidate-generating density $q_{MH}(\boldsymbol{\rho}, \beta_1)$, i.e.,

- Draw from $q_{MH}(\boldsymbol{\rho}, \beta_1)$ until a draw $\left(\hat{\boldsymbol{\rho}}, \hat{\beta}_1\right)$ satisfies $\left(\hat{\boldsymbol{\rho}}, \hat{\beta}_1\right) \in \Theta_{\boldsymbol{\rho}} \times \mathbb{R}$. Draw u from the uniform distribution $U(0,1)$.
- Calculate the acceptance probability $\alpha\left[\left(\boldsymbol{\rho}^{i-1}, \beta_1^{i-1}\right), \left(\hat{\boldsymbol{\rho}}, \hat{\beta}_1\right)\right]$, defined as

$$\begin{aligned} &\alpha\left[\left(\boldsymbol{\rho}^{i-1}, \beta_1^{i-1}\right), \left(\hat{\boldsymbol{\rho}}, \hat{\beta}_1\right)\right] := \\ &\min\left(\frac{p\left(\hat{\boldsymbol{\rho}}, \hat{\beta}_1 | \left(\sigma^2\right)^{i-1}, \widetilde{\boldsymbol{\beta}}^{i-1}, \boldsymbol{y}\right) q_{MH}\left(\boldsymbol{\rho}^{i-1}, \beta_1^{i-1}\right)}{p\left(\boldsymbol{\rho}^{i-1}, \beta_1^{i-1} | \left(\sigma^2\right)^{i-1}, \widetilde{\boldsymbol{\beta}}^{i-1}, \boldsymbol{y}\right) q_{MH}\left(\hat{\boldsymbol{\rho}}, \hat{\beta}_1\right)}, 1\right). \end{aligned}$$

- If $u \leq \alpha\left[\left(\boldsymbol{\rho}^{i-1}, \beta_1^{i-1}\right), \left(\hat{\boldsymbol{\rho}}, \hat{\beta}_1\right)\right]$, set $\left(\boldsymbol{\rho}^i, \beta_1^i\right) = \left(\hat{\boldsymbol{\rho}}, \hat{\beta}_1\right)$.
- Else, set $\left(\boldsymbol{\rho}^i, \beta_1^i\right) = \left(\boldsymbol{\rho}^{i-1}, \beta_1^{i-1}\right)$.

(4) Draw $\left(\sigma^2\right)^i$, given $\left(\boldsymbol{\rho}^i, \beta_1^i\right)$ and $\widetilde{\boldsymbol{\beta}}^{i-1}$, from the inverse gamma distribution in (4.11).

(5) Draw $\widetilde{\boldsymbol{\beta}}^i$, given $\left(\boldsymbol{\rho}^i, \beta_1^i\right)$ and $\left(\sigma^2\right)^i$, from the $(k-1)$-variate normal distribution with mean $\boldsymbol{\mu}_{\widetilde{\boldsymbol{\beta}}}$ and covariance matrix $\Sigma_{\widetilde{\boldsymbol{\beta}}}$ as in (4.23), (4.24).

Appendix 4.B Bayes factor computation

In the following, we show how the integral $I_t = \int_{\Theta_{\boldsymbol{\rho}_t}} h_t(\boldsymbol{\rho}_t)\,\mathrm{d}\boldsymbol{\rho}_t$ in (4.16) can be effectively approximated by its importance sampling estimate $\widehat{I}_t$,

$$\begin{aligned}\widehat{I}_t &= N^{-1}\sum_{i=1}^{N}\frac{h_t(\boldsymbol{\rho}_i)}{q_t(\boldsymbol{\rho}_i)}\\ &= N^{-1}\sum_{i=1}^{N}\frac{\left|A_{\boldsymbol{\rho}_i}\right|\exp\left(-\frac{1}{2}(\boldsymbol{\rho}_i-\boldsymbol{\mu}_t)^T\Sigma_t^{-1}(\boldsymbol{\rho}_i-\boldsymbol{\mu}_t)\right)\boldsymbol{y}^TA_{\boldsymbol{\rho}_i}^TMA_{\boldsymbol{\rho}_i}\boldsymbol{y}^{-\frac{g-k}{2}}}{q_t(\boldsymbol{\rho}_i)},\end{aligned} \tag{4.25}$$

where $\boldsymbol{\rho}_i$ are draws from a suitable importance density $q_t(\cdot)$. We specify $q_t(\cdot)$ such that it closely follows the integrand $h_t(\boldsymbol{\rho}_t)$ but has heavier tails than the latter, which ensures a reliable estimation of I_t.

As in Appendix 4.A, we approximate $\log\left(\left|A_{\boldsymbol{\rho}_t}\right|\right)$ by a quadratic polynomial in $\boldsymbol{\rho}_t$ at its maximum value, the origin. This results in a normal approximation of $\left|A_{\boldsymbol{\rho}_t}\right|$, i.e., $\left|A_{\boldsymbol{\rho}_t}\right| \approx N\left(\mathbf{0}_R, T^{-1}\right)$, where $T_{rr'} := \operatorname{tr}(W_rW_{r'})$, $r, r' \in \{1, ..., R\}$. In the case that T is not positive definite, we use the nearest positive definite matrix to T instead. The second term in the denominator of (4.25) already equals the kernel of the probability density function of the normal distribution $N(\boldsymbol{\mu}_t, \Sigma_t)$. Lastly, we also approximate the logarithm of the third term in $h_t(\boldsymbol{\rho}_t)$ by a second-order Taylor polynomial at its maximum. It follows that $\boldsymbol{y}^TA_{\boldsymbol{\rho}_t}^TMA_{\boldsymbol{\rho}_t}\boldsymbol{y}^{-(g-k)/2} \approx N(\boldsymbol{\mu}_3, \Sigma_3)$, where $\boldsymbol{\mu}_3 = \left(\boldsymbol{y}^TW_.^TMW_.\boldsymbol{y}\right)^{-1}\boldsymbol{y}^TMW_.\boldsymbol{y}$, $\Sigma_3 = \left(\boldsymbol{y}^TW_.^TMW_.\boldsymbol{y}\right)^{-1}\left(\boldsymbol{y}^TA_{\boldsymbol{\mu}_3}^TMA_{\boldsymbol{\mu}_3}\boldsymbol{y}\right)/(g-k)$ and $\left(\boldsymbol{y}^TW_.^TMW_.\boldsymbol{y}\right)_{rr'} := \boldsymbol{y}^TW_r^TMW_{r'}\boldsymbol{y}$, $\left(\boldsymbol{y}^TMW_.\boldsymbol{y}\right)_r := \boldsymbol{y}^TMW_r\boldsymbol{y}$. Thus, $h_t(\boldsymbol{\rho}_t)$ can be approximated by a product of three multivariate normal densities that is multivariate normal itself, so $q_t(\boldsymbol{\rho}_t) := \phi_{\boldsymbol{\mu}_{IS_t},\Sigma_{IS_t}}(\boldsymbol{\rho}_t)\,\mathbb{1}_{\Theta_{\boldsymbol{\rho}_t}}(\boldsymbol{\rho}_t)\,c_{IS_t}^{-1}$, $c_{IS_t} := \int_{\Theta_{\boldsymbol{\rho}_t}}\phi_{\boldsymbol{\mu}_{IS_t},\Sigma_{IS_t}}(\boldsymbol{\rho}_t)\,\mathrm{d}\boldsymbol{\rho}_t$, with $\Sigma_{IS_t} = \left(T + \Sigma_t^{-1} + \Sigma_3^{-1}\right)^{-1}$, $\boldsymbol{\mu}_{IS_t} = \Sigma_{IS_t}\left(\Sigma_t^{-1}\boldsymbol{\mu}_t + \Sigma_3^{-1}\boldsymbol{\mu}_3\right)$.

Calculating $\widehat{I}_t$ directly might result in underflow in R, which is why we show how to compute its logarithm only, next. We can write

$$\begin{aligned}\log\left(\widehat{I}_t\right) &= \log\left(N^{-1}\sum_{i=1}^{N}\frac{h_t(\boldsymbol{\rho}_i)}{q_t(\boldsymbol{\rho}_i)}\right) = -\log(N) + \log\left(\sum_{i=1}^{N}\frac{h_t(\boldsymbol{\rho}_i)}{q_t(\boldsymbol{\rho}_i)}\right)\\ &= -\log(N) + \log\left(\sum_{i=1}^{N}\exp\left(\log\left(\frac{h_t(\boldsymbol{\rho}_i)}{q_t(\boldsymbol{\rho}_i)}\right) + d - d\right)\right)\\ &= -\log(N) - d + \log\left(\sum_{i=1}^{N}\exp\left(\log\left(\frac{h_t(\boldsymbol{\rho}_i)}{q_t(\boldsymbol{\rho}_i)}\right) + d\right)\right)\\ &= -\log(N) - d + \log\left(\sum_{i=1}^{N}\exp\left(\log\left(\left|A_{\boldsymbol{\rho}_i}\right|\right) - \frac{g-k}{2}\log\left(\boldsymbol{y}^TA_{\boldsymbol{\rho}_i}^TMA_{\boldsymbol{\rho}_i}\boldsymbol{y}\right)\right.\right.\\ &\qquad -\frac{1}{2}(\boldsymbol{\rho}_i-\boldsymbol{\mu}_t)^T\Sigma_t^{-1}(\boldsymbol{\rho}_i-\boldsymbol{\mu}_t)\\ &\qquad \left.\left. -\log\left(\frac{\phi_{\boldsymbol{\mu}_{IS_t},\Sigma_{IS_t}}(\boldsymbol{\rho}_i)\,\mathbb{1}_{\Theta_{\boldsymbol{\rho}_t}}(\boldsymbol{\rho}_i)}{c_{IS_t}}\right) + d\right)\right)\end{aligned}$$

$$
\begin{aligned}
&\approx -\log(N) - d + \log\Bigg(\sum_{i=1}^{N} \exp\bigg(\log(|A_{\boldsymbol{\rho}_i}|) - \frac{g-k}{2}\log\left(\boldsymbol{y}^T A_{\boldsymbol{\rho}_i}^T M A_{\boldsymbol{\rho}_i}\boldsymbol{y}\right) \\
&\qquad -\frac{1}{2}(\boldsymbol{\rho}_i - \boldsymbol{\mu}_t)^T \Sigma_t^{-1}(\boldsymbol{\rho}_i - \boldsymbol{\mu}_t) \\
&\qquad - \log\left(\frac{\phi_{\boldsymbol{\mu}_{IS_t},\Sigma_{IS_t}}(\boldsymbol{\rho}_i)\,\mathbb{1}_{\Theta_{\boldsymbol{\rho}_t}}(\boldsymbol{\rho}_i)}{\frac{\sum_{i=1}^{N}\mathbb{1}_{\Theta_{\boldsymbol{\rho}_t}}(\boldsymbol{\rho}_i)}{N}}\right) + d\bigg)\Bigg) \\
&= -2\log(N) + \log\left(\sum_{i=1}^{N}\mathbb{1}_{\Theta_{\boldsymbol{\rho}_t}}(\boldsymbol{\rho}_i)\right) - d \\
&\quad + \log\Bigg(\sum_{i=1}^{N}\exp\bigg(\log(|A_{\boldsymbol{\rho}_i}|) - \frac{g-k}{2}\log\left(\boldsymbol{y}^T A_{\boldsymbol{\rho}_i}^T M A_{\boldsymbol{\rho}_i}\boldsymbol{y}\right) \\
&\qquad -\frac{1}{2}(\boldsymbol{\rho}_i - \boldsymbol{\mu}_t)^T \Sigma_t^{-1}(\boldsymbol{\rho}_i - \boldsymbol{\mu}_t) - \log\left(\phi_{\boldsymbol{\mu}_{IS_t},\Sigma_{IS_t}}(\boldsymbol{\rho}_i)\,\mathbb{1}_{\Theta_{\boldsymbol{\rho}_t}}(\boldsymbol{\rho}_i)\right) + d\bigg)\Bigg),
\end{aligned}
$$

where $\boldsymbol{\rho}_i$ are draws from the unconstrained importance density $N\left(\boldsymbol{\mu}_{IS_t}, \Sigma_{IS_t}\right)$ and d is an auxiliary constant, e.g., $d = -\frac{g-k}{2}\min_{i\in\{1,\ldots,N\}}\left(\boldsymbol{y}^T A_{\boldsymbol{\rho}_i}^T M A_{\boldsymbol{\rho}_i}\boldsymbol{y}\right)$, which is added to prevent the marginal likelihood to become too small to be distinguished from zero in R. The auxiliary constant d is set in advance after generating the N draws from the unconstrained importance density $N\left(\boldsymbol{\mu}_{IS_t}, \Sigma_{IS_t}\right)$ first.

Chapter 5

A discrete exponential family model for network autocorrelated count data

Abstract

We introduce a discrete exponential family model for analyzing network autocorrelated count data. In our approach, we model the joint distribution for the counts using a discrete exponential family specified in terms of sufficient statistics of a count configuration. We propose several sufficient statistics representing key structural properties of a count configuration, such as lower-order moments, zero inflation, and network autocorrelation. As such, the approach does not rely on any distributional assumptions on the marginal or conditional counts and is flexible enough to model a wide range of count patterns. We provide algorithms to simulate count configurations from the model and to perform maximum likelihood-based inference, along with goodness-of-fit measures assessing the model fit to observed data based on simulated count configurations. Finally, we illustrate our model by re-analyzing the sources of reported homicide counts in 343 neighborhoods in Chicago, Illinois.

This chapter will be submitted for publication as: Dittrich, D., Butts, C.T., Leenders, R.Th.A.J., and Mulder, J. A discrete exponential family model for network autocorrelated count data.

5.1 Introduction

Individual behavior, corporate decisions, or entries intro armed conflict do not happen in vacuum. Instead, individuals, firms, and countries interact and thereby influence as well as are influenced by each other. Out of the numerous models that address effects of such social interaction on a variable of interest, the network autocorrelation model (Ord, 1975) has been the flagship model for incorporating global network autocorrelation in cross-sectional data. It is also known as spatial effects model (Doreian, 1980), network effects model (Dow et al., 1982), mixed regressive-spatial autoregressive model (Anselin, 1988), or spatial lag model (Anselin, 2002), and has been widely applied in a variety of fields, such as criminology (Tita & Radil, 2011), ecology (McPherson & Nieswiadomy, 2005), economics (Conway et al., 2010), geography (Dall'erba, 2005), and sociology (Mizruchi & Stearns, 2006).

For the moment, assume that we observed values for a variable of interest for g actors in a network, who may be tied to each other based on a given influence mechanism, e.g., friendship. The network autocorrelation model expands a standard linear regression model and accommodates potential *network autocorrelation*, i.e., interdependence of the observations for the actors across the network, by including an additional endogenous covariate. For each actor in the network, the additional covariate consists of a weighted sum of the values for the variable of interest for this actor's neighbors, i.e., other actors in the network this actor is tied to. The associated regression coefficient is known as the network autocorrelation parameter ρ. Thus, the network autocorrelation model is given by

$$\boldsymbol{y} = \rho W \boldsymbol{y} + X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{5.1}$$

$$\Leftrightarrow \boldsymbol{y} = \left(I_g - \rho W\right)^{-1} \left(X\boldsymbol{\beta} + \boldsymbol{\varepsilon}\right), \tag{5.2}$$

where $\boldsymbol{y} \in \mathbb{R}^g$ contains the values for a dependent variable of interest for the g actors, $X \in \mathbb{R}^{g\times k}$ comprises values for the g actors on k covariates, $\boldsymbol{\beta} \in \mathbb{R}^k$ is a vector of regression coefficients, $\boldsymbol{\varepsilon} \in \mathbb{R}^g$ is a vector of error terms, $I_g \in \mathbb{R}^{g\times g}$ denotes the identity matrix, and $W \in \mathbb{R}^{g\times g}$ is an a priori defined *connectivity matrix*, specifying the influence relationships between the actors; the larger the entry W_{ij}, the larger the influence of actor j on actor i, where we exclude relationships from an actor to himself, i.e., $W_{ii} = 0$ for all $i \in \{1, ..., g\}$. The scalar network autocorrelation parameter ρ quantifies the magnitude of the network autocorrelation on the variable of interest in the network defined by W. Under regularity conditions, the matrix inverse in the reduced form of the model in (5.2) can be rewritten using the so-called "Leontief expansion" as $\left(I_g - \rho W\right)^{-1} = I_g + \rho W + \rho^2 W^2 + ...$ (see e.g., Griffith, 1979). Consequently, in the network autocorrelation model, a change in one actor's covariate value does not only affect this actor's neighbors' values for a variable of interest but potentially, depending on the structure of W, also those for all other actors in the network, with the impact on the variable of interest decreasing with the network distance from the actor. Such a spillover pattern is also called a *global* spillover (Anselin, 2003; Elhorst, 2010; LeSage, 2014b).

Despite the usefulness of the network autocorrelation model, an important limitation of the model is that it cannot be directly applied to count data, which would, among other issues, lead to predicted non-integer and potentially negative values for a variable of interest. As a potential ad-hoc remedy, some authors suggested to transform the counts into an approximately normal variable and then fit a network autocorrelation model to the transformed variable (LeSage & Pace, 2008; Quddus, 2008). Often though, count data cannot be appropriately transformed to become approximately normal, e.g., when modeling rare counts such as violent crimes or the number of sexual partners; merely transforming counts is of limited applicability in such complex real-life settings and also does not substantially address the problem of how to adequately deal with network autocorrelated counts. Hence, any prudent model for network autocorrelated count data has to forfeit a direct functional relationship between the dependent variable of interest and the covariates while retaining the network autocorrelation model's global spillover pattern.

To be sure, a number of models for dealing with count data building on global spillover patterns have been proposed in the literature. McMillen (1992) was the first to provide an extension of the network autocorrelation model for binary data and LeSage (2000) further refined the method to allow modeling of heteroskedastic data.[1] More recently, Lambert et al. (2010) and Glaser (2017) developed spatial autoregressive count models for Poisson and negative binomial data. In both approaches, the conditional expected count for each actor is modeled as a function of the actor's neighbors' counts and actor covariates. Castro et al. (2012) and Bhat et al. (2014) considered models that are characterized by the counts being driven by a Gaussian latent variable that is assumed to follow the network autocorrelation model in (5.1). Subsequently, in all of the four latter approaches the authors constructed a pseudo-likelihood function by multiplying suitable conditional probability mass functions and maximized the resulting product to obtain a maximum pseudo-likelihood estimate of the model parameters. However, these maximum pseudo-likelihood estimates ignore dependencies between the conditional probability mass functions and cannot adequately capture the dependence in the data when network autocorrelation is strong. Liesenfeld et al. (2016) presented a model for spatially correlated Poisson and negative binomial data similar to the ones in Castro et al. (2012) and Bhat et al. (2014) but additionally provided a numerically accurate, albeit computationally very expensive, algorithm for full maximum likelihood estimation of their model. In contrast, Bhati (2008) took a rather different approach when modeling spatially correlated counts by introducing a semi-parametric estimator based on a generalized cross-entropy formulation.

In this chapter, we propose a discrete exponential family model for network autocorrelated count data that unifies several ideas from the literature. In our proposed model, we directly model the joint distribution for all actor counts using a discrete exponential family specified by its sufficient statistics. These sufficient statistics are specified such to represent structural properties of a joint count configuration and that are characteristic of underlying processes believed to have generated an observed count configuration. For

[1] Martinetti & Geniaux (2016), McMillen (2013), and Wilhelm & Godinho de Matos (2015) implemented estimation procedures for the probit network autocorrelation model in R, while LeSage (1999) did so in MATLAB.

example, under network autocorrelation, we would expect neighboring actors to exhibit fairly similar counts, which would be accordingly captured by a sufficient statistic embodying network autocorrelation. There are several advantages to using a discrete exponential family for the joint count distribution for the actors. First, our model naturally incorporates global spillover patterns across the network. Second, we do not (have to) make any restrictive and potentially limiting distributional assumptions on the conditional counts. Third, by choosing appropriate sufficient statistics, our model is flexible enough to accommodate highly skewed raw count distributions as well as count configurations exhibiting excess zeros, making it particularly appealing when analyzing rare counts or events. Fourth, we are able to easily simulate joint count configurations from the model and to compute accurate simulation-based maximum likelihood estimates, providing the basis for inference in the model. Hence, our model permits for principled inference in the presence of both strong network autocorrelation and complex real-life data structures going well beyond standard Poisson and negative binomial specifications.

We proceed as follows. In the next section, we motivate and describe our discrete exponential family model for network autocorrelated count data. In Section 5.3, we introduce several sufficient statistics representing a range of structural count properties and show how to interpret them. In Section 5.4, we present simulation and maximum likelihood estimation procedures for the model as well as useful measures for assessing model fit to observed data. We apply our model to re-analyze the drivers of homicide counts registered in 343 neighborhoods in Chicago, Illinois, in Section 5.5. We conclude this chapter with a short discussion of our main findings and highlight directions for fruitful future research within our proposed framework in Section 5.6.

5.2 Model definition

Let us consider g actors and an associated g-variate random count variable of interest $\boldsymbol{Y}$ that can take values in $\{0, ..., m-1\}^g$. We denote this set of attainable count configurations by $\mathcal{Y} := \{0, ..., m-1\}^g$ and define an attainable count as an integer between 0 and $m-1$. Here, the upper bound $m-1$ for the maximum count each actor can have needs to be chosen in advance, such that the set of attainable count configurations is large enough to contain all likely count configurations.

Under this general framework, we are interested in modeling a joint actor count configuration while accounting for exogenous and endogenous mechanisms that have supposedly generated the count configuration. For this, we use a discrete exponential family that specifies the joint distribution for the counts precisely in terms of these potentially count-generating mechanisms. In the literature, the discrete exponential family has been extensively employed, e.g., to model various forms of relational data through exponential random graph models (Holland & Leinhardt, 1981; Krivitsky, 2012; Krivitsky & Butts, 2017), to define probabilistic graphical models (Lauritzen, 1996), or to control the arrangement of "objects" into "locations" in generalized location systems (Butts, 2007). As we can reformulate the modeling task in this chapter as assigning actors to attainable counts, our approach is closely related to the one in generalized location systems and we will repeatedly draw parallels between the two.

Using notation similar to that of Butts (2007) and Hummel et al. (2012), we define the probability of the random count variable $\boldsymbol{Y}$ occupying a particular count configuration $\boldsymbol{y}$ as

$$p\left(\boldsymbol{Y}=\boldsymbol{y}|\boldsymbol{\theta}\right):=\frac{\exp\left(\boldsymbol{\theta}^T\boldsymbol{t}\left(\boldsymbol{y}\right)\right)}{\kappa\left(\boldsymbol{\theta}\right)},\boldsymbol{y}\in\mathcal{Y}, \tag{5.3}$$

where $\boldsymbol{\theta}$ is a vector of real-valued parameters, $\boldsymbol{t}\left(\boldsymbol{y}\right)$ is a vector of sufficient statistics, and $\kappa\left(\boldsymbol{\theta}\right):=\sum_{\boldsymbol{y}'\in\mathcal{Y}}\exp\left(\boldsymbol{\theta}^T\boldsymbol{t}\left(\boldsymbol{y}'\right)\right)$ denotes a normalizing constant, which ensures that the probabilities of all attainable count configurations in (5.3) sum to unity. The specification in (5.3) is a fairly general one, where the sufficient statistics can be seen as summary measures of structural properties of a joint count configuration. In practice, the sufficient statistics are chosen such to represent mechanisms that are hypothesized to have generated and can explain an observed count configuration, e.g., why some actors have zero counts, why others exhibit high counts, or why certain actor counts tend to cluster. The parameter θ_i then weights the relative importance of the corresponding sufficient statistic $t_i\left(\boldsymbol{y}\right)$.

5.3 Model specification and interpretation

In this section, we first motivate the choice for specific sufficient statistics embodying a variety of common structural properties of a count configuration that govern the count distribution, before we show how to properly interpret the corresponding model parameters.

5.3.1 Specification of sufficient statistics

Network autocorrelation

We include network autocorrelation to our model by proposing a sufficient statistic, originally introduced slightly differently in Butts (2007), that captures the tendency of connected actors to show similar (or different) counts. Let $W\in\mathbb{R}^{r\times g\times g}$ be an array of r a priori defined connectivity matrices specifying the influence relationships between the actors, where W_{ijk} amounts to the influence of actor k on actor j based on influence mechanism i. Moreover, let $R\in\mathbb{R}^{m\times r}$ be a matrix containing values for the m attainable counts on r count attributes, i.e., values that are functions of the counts. For example, R could simply contain the attainable counts $(0,...,m-1)$ in each of its r columns. Alternatively, when trying to account for the shape of the raw count distribution, R could also contain a reference count distribution's percentiles. We set the elements of the vector of sufficient statistics $\boldsymbol{t}^{\mathrm{NAC}}\left(\boldsymbol{y}\right):=\left(t_1^{\mathrm{NAC}}\left(\boldsymbol{y}\right),...,t_r^{\mathrm{NAC}}\left(\boldsymbol{y}\right)\right)$ representing network autocorrelation to

$$t_i^{\mathrm{NAC}}\left(\boldsymbol{y}\right):=-\sum_{j=1}^{g}\sum_{k=1}^{g}W_{ijk}\left|R_{y_ji}-R_{y_ki}\right|,i\in\left\{1,...,r\right\}, \tag{5.4}$$

where y_j is the count for actor j and $R_{y_j i}$ is the value for this count's i-th attribute. We denote the corresponding parameter vector by $\boldsymbol{\rho} \in \mathbb{R}^r$ and negate the sum in (5.4) such that a positive parameter value leads to positive network autocorrelation.[2] In order to illustrate the behavior of the sufficient statistics in (5.4), consider a network of urban neighborhoods equipped with a single binary adjacency matrix W, i.e., $W_{jk} = 1$ if neighborhood j and k are adjacent and zero otherwise, and $R = (0, ..., m-1)$. In this case, similar counts between adjacent neighborhoods result in relatively larger (less negative) values for the (scalar) sufficient statistic, which would imply positive network autocorrelation across the neighborhood network. Equivalently, more divergent counts between adjacent neighborhoods would lead to smaller (more negative) values for the sufficient statistic and reveal negative network autocorrelation. Accordingly, the magnitude of ρ indicates the importance of network autocorrelation as underlying mechanism in explaining an observed count configuration. Lastly, the specification in (5.4) can include several connectivity matrices, e.g., based on geographic adjacency and social similarity between neighborhoods, resulting in multiple network autocorrelations and adding to the model's flexibility.

Covariate modeling

In addition to network autocorrelation, our model also needs to incorporate count homo- or heterogeneity based on exogenous actor attributes. We model such effects by sufficient statistics that are based on products of actor and count attributes.[3] Let $X \in \mathbb{R}^{g \times q}$ contain values for the g actors on q actor attributes (possible including columns of ones) and let $Q \in \mathbb{R}^{m \times q}$ carry values for the m attainable counts on q count attributes. We define the elements of the vector of sufficient statistics $\boldsymbol{t}^{\mathrm{CHG}}(\boldsymbol{y}) := \left(t_1^{\mathrm{CHG}}(\boldsymbol{y}), ..., t_q^{\mathrm{CHG}}(\boldsymbol{y})\right)$ reflecting count homo- or heterogeneity based on exogenous actor attributes as

$$t_{i'}^{\mathrm{CHG}}(\boldsymbol{y}) := \sum_{j=1}^{g} Q_{y_j i'} X_{j i'}, i' \in \{1, ..., q\}, \tag{5.5}$$

and we denote the corresponding parameter vector by $\boldsymbol{\beta} \in \mathbb{R}^q$. For example, consider again a network of urban neighborhoods with $\boldsymbol{Q}_{\cdot i'} = (0, ..., m-1)$, where $\boldsymbol{Q}_{\cdot i'}$ denotes the i'-th column of Q, and let $\boldsymbol{X}_{\cdot i'}$ comprise values for a resource deprivation index of the neighborhoods. Then, ceteris paribus, $\beta_{i'} > 0$ implies that more resource deprived neighborhoods are more likely to exhibit higher counts for a variable of interest, e.g., homicide counts, and less resource deprived ones are accordingly less likely so.

If one or several columns of X do not vary across actors but are constant, i.e., equal to a vector of ones, the sufficient statistics in (5.5) can also be used to capture basic endogenous forces that gave rise to an observed count configuration. These basic endogenous forces can be seen as control variables that govern the baseline shape of the count

[2] The sufficient statistics in (5.4) also bear resemblance to Geary's c (Geary, 1954), a well-known measure of spatial autocorrelation, and more remotely to a sufficient statistic in Krivitsky (2012) that models mutuality in directed valued exponential random graph models.

[3] Since the probability mass function in (5.3) is invariant up to additive constants, any sufficient statistic that is invariant over $\mathcal{Y}$, e.g., solely based on actor attributes such as age, gender, etc., can be left out (Butts, 2007). Instead, all relevant sufficient statistics need to link actor and count attributes.

distribution notwithstanding other, more substantive, effects based on network autocorrelation or exogenous actor attributes, which may not always be available. We show how to attain some control over basic properties of the count configuration, such as the mean, the variance, and the sparsity of the counts, next. First, we can perfectly model the mean count of a count configuration by setting a column of Q to $(0, ..., m-1)$, leading to a sufficient statistic that is the sum of the actor counts, $t^{\text{AVG}}(\boldsymbol{y}) := \sum_{j=1}^{g} y_j$. Second, we can reach partial control over the variance of the counts by taking a column of Q to be $(0^a, ..., (m-1)^a), a \neq 1$, resulting in a sufficient statistic of the form $t^{\text{VAR}}(\boldsymbol{y}) = \sum_{j=1}^{g} y_j^a$. We have found experimentally that choosing $a = 2$ typically appropriately controls the variance of the counts. Third, we can capture count sparsity, i.e., the tendency of count configurations to have excess zeros, by specifying a column of Q as the binary vector $(1, 0, ..., 0)$. This produces a sufficient statistic $t^{\text{ZIF}}(\boldsymbol{y}) := \sum_{j=1}^{g} \mathbb{1}(y_j = 0)$, which simply equals the number of zero counts in a count configuration.

Taken together, the sufficient statistics in (5.4) and (5.5) can be substituted into (5.3) to fully determine the probability of observing a count configuration $\boldsymbol{y}$, given by

$$\begin{aligned} p(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{\theta}) &= \frac{\exp\left(\boldsymbol{\theta}^T \boldsymbol{t}(\boldsymbol{y})\right)}{\kappa(\boldsymbol{\theta})}, \boldsymbol{y} \in \mathcal{Y}, \\ &= \frac{\exp\left((\boldsymbol{\rho}, \boldsymbol{\beta})^T \left(\boldsymbol{t}^{\text{NAC}}(\boldsymbol{y}), \boldsymbol{t}^{\text{CHG}}(\boldsymbol{y})\right)\right)}{\kappa(\boldsymbol{\theta})} \\ &= \frac{\exp\left(-\sum_{i=1}^{r} \rho_i \sum_{j=1}^{g} \sum_{k=1}^{g} W_{ijk} \left|R_{y_j i} - R_{y_k i}\right| + \sum_{i'=1}^{q} \beta_{i'} \sum_{j=1}^{g} Q_{y_j i'} X_{j i'}\right)}{\kappa(\boldsymbol{\theta})}, \end{aligned} \tag{5.6}$$

where $\boldsymbol{\theta} := (\boldsymbol{\rho}, \boldsymbol{\beta})$ and $\boldsymbol{t}(\boldsymbol{y}) := \left(\boldsymbol{t}^{\text{NAC}}(\boldsymbol{y}), \boldsymbol{t}^{\text{CHG}}(\boldsymbol{y})\right)$.

5.3.2 Interpretation of model parameters

In contrast to the standard network autocorrelation model, a parameter θ_i in the discrete exponential family model in (5.6) cannot be interpreted as the effect that a one-unit change in an actor attribute has on the expected actor counts in the network. Instead, the most direct way to interpret parameters in the model is in terms of the probability ratio of two count configurations $\boldsymbol{y}, \boldsymbol{y}' \in \mathcal{Y}$. In this case,

$$\frac{p(\boldsymbol{Y} = \boldsymbol{y}'|\boldsymbol{\theta})}{p(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{\theta})} = \frac{\exp\left(\boldsymbol{\theta}^T \boldsymbol{t}(\boldsymbol{y}')\right)}{\kappa(\boldsymbol{\theta})} \frac{\kappa(\boldsymbol{\theta})}{\exp\left(\boldsymbol{\theta}^T \boldsymbol{t}(\boldsymbol{y})\right)} = \exp\left(\boldsymbol{\theta}^T \left(\boldsymbol{t}(\boldsymbol{y}') - \boldsymbol{t}(\boldsymbol{y})\right)\right). \tag{5.7}$$

Hence, the log-odds of observing count configuration $\boldsymbol{y}'$ rather than count configuration $\boldsymbol{y}$ increase by θ_i for a one-unit change in the associated sufficient statistic.[4] While (5.7) gives a clear quantitative interpretation of the model parameters in terms of relative

[4] As the normalizing constant $\kappa(\boldsymbol{\theta})$ implicitly depends on X, the probability ratio in (5.7) involving a change in X would depend on the ratio of the corresponding intractable normalizing constants and θ_i could not be interpreted in the same way.

count configuration probabilities and changes in the sufficient statistics, the changes in the sufficient statistics are ultimately a result of changes in the counts themselves. Thus, it is more intuitive and insightful to consider the effect that a model parameter has on the probability ratio in (5.7) based on increments, or reductions, in the counts. Let $\boldsymbol{y} \in \mathcal{Y}$ be a count configuration in which actor j has count k. Using (5.7) and conditional on the counts for the remaining actors, we can express the conditional log-odds of actor j having an attainable count l instead as

$$\log\left(\frac{p\left(Y_j = l|\boldsymbol{Y}_{-j} = \boldsymbol{y}_{-j};\boldsymbol{\theta}\right)}{p\left(Y_j = k|\boldsymbol{Y}_{-j} = \boldsymbol{y}_{-j};\boldsymbol{\theta}\right)}\right) = \log\left(\frac{p\left(Y_j = l,\boldsymbol{Y}_{-j} = \boldsymbol{y}_{-j}|\boldsymbol{\theta}\right)}{p\left(Y_j = k,\boldsymbol{Y}_{-j} = \boldsymbol{y}_{-j}|\boldsymbol{\theta}\right)}\frac{p\left(\boldsymbol{Y}_{-j} = \boldsymbol{y}_{-j}|\boldsymbol{\theta}\right)}{p\left(\boldsymbol{Y}_{-j} = \boldsymbol{y}_{-j}|\boldsymbol{\theta}\right)}\right)$$
$$= -\sum_{i=1}^{r}\rho_i\sum_{j'=1}^{g}\left(W_{ijj'}+W_{ij'j}\right)\left(\left|R_{li}-R_{y_{j'}i}\right|-\left|R_{ki}-R_{y_{j'}i}\right|\right)+\sum_{i'=1}^{q}\beta_{i'}X_{ji'}\left(Q_{li'}-Q_{ki'}\right), \quad (5.8)$$

where $\boldsymbol{Y}_{-j}, \boldsymbol{y}_{-j}$ denote all elements of $\boldsymbol{Y}$ and $\boldsymbol{y}$, respectively, other than element j.

5.4 Model inference

5.4.1 Simulation

In order to perform maximum likelihood-based inference in the model, it is essential to be able to simulate count configurations from the model given a particular parameter value $\boldsymbol{\theta}$. Opposed to the standard network autocorrelation model, it is not possible to simulate count configurations from the model directly; however, count configurations can be straightforwardly simulated using the *Metropolis algorithm* (Metropolis et al., 1953). In the Metropolis algorithm, starting from an initial count configuration $\boldsymbol{y}^0 \in \mathcal{Y}$, a candidate count configuration $\hat{\boldsymbol{y}} \in \mathcal{Y}$ for the target distribution $p\left(\boldsymbol{Y}|\boldsymbol{\theta}\right)$ is proposed, first. Next, the candidate count configuration $\hat{\boldsymbol{y}}$ is accepted with probability $\alpha := \min\left(1, p\left(\boldsymbol{Y} = \hat{\boldsymbol{y}}|\boldsymbol{\theta}\right)/p\left(\boldsymbol{Y} = \boldsymbol{y}^0|\boldsymbol{\theta}\right)\right)$. If the candidate count configuration is accepted, $\hat{\boldsymbol{y}}$ becomes the next element in the sequence of simulated count configurations, i.e., $\boldsymbol{y}^1 = \hat{\boldsymbol{y}}$, else $\boldsymbol{y}^1 = \boldsymbol{y}^0$. This so-called Metropolis step is repeated a large number of times until the desired number of draws of count configurations has been obtained. These draws can then be, possibly after an appropriate number of initial draws, regarded as realizations from the target distribution $p\left(\boldsymbol{Y}|\boldsymbol{\theta}\right)$ itself.

Following similar simulation procedures for related discrete exponential family models (Butts, 2007; Snijders, 2002), we form a candidate count configuration $\hat{\boldsymbol{y}} \in \mathcal{Y}$, given count configuration $\boldsymbol{y} \in \mathcal{Y}$, by assigning one randomly chosen actor j a random attainable count k, i.e., $\hat{y}_j = k, \hat{\boldsymbol{y}}_{-j} = \boldsymbol{y}_{-j}$. In each Metropolis step, the corresponding acceptance probability can then be easily calculated using (5.8). However, as the described Metropolis algorithm often starts with a random count configuration and at most one actor's count is changed in each step of the algorithm, the first simulated count configurations are typically highly dependent on the initial count configuration and not representative of the target distribution $p\left(\boldsymbol{Y}|\boldsymbol{\theta}\right)$. Thus, an initial number of draws, known as the *burn-in* (Gelman et al., 2003), is usually discarded.[5] Moreover, the sequence of simulated count configurations may be sub-sampled, or *thinned*, i.e., every $k-1$ out of k draws may be discarded, due to limitations in computer memory and storage, or due to intending to reduce the autocorrelation in the draws, e.g., when aiming to obtain (nearly) independent realizations. In general though, thinning does not improve statistical efficiency (Geyer, 1992; Link & Eaton, 2012; A. B. Owen, 2017).

5.4.2 Estimation

In this section, we present techniques for maximum (pseudo-)likelihood estimation of the model. While the availability of the full likelihood function in (5.3) readily permits likelihood-based inference in theory, the intractability of the normalizing constant $\kappa(\boldsymbol{\theta})$ makes direct evaluation or maximization of the likelihood function numerically infeasible. One way to bypass this issue is to construct a pseudo-likelihood function defined as the product of the conditional likelihoods and to maximize this product to obtain the maximum pseudo-likelihood estimate of the model parameters. The maximum pseudo-likelihood estimator, however, is only identical to the maximum likelihood estimator in case of no network autocorrelation and has been shown to be generally inferior to the maximum likelihood estimator in structurally similar exponential random graph models (Desmarais & Cranmer, 2012b; Robins, Pattison, et al., 2007; van Duijn et al., 2009). Nevertheless, we first establish the model's maximum pseudo-likelihood estimator, which will serve as an initial approximation to the maximum likelihood estimator. Subsequently, we describe a simulation-based stepping algorithm that, starting from the maximum pseudo-likelihood estimate, iteratively moves towards the maximum likelihood estimate.

Maximum pseudo-likelihood estimation

As explained in the previous paragraph, we first need to specify the conditional likelihood for an actor given the counts for all other actors in the network. Following Butts (2007), the conditional likelihood for actor j is given by

$$p\left(Y_j = y_j | \boldsymbol{Y}_{-j} = \boldsymbol{y}_{-j}; \boldsymbol{\theta}\right) = \left(\sum_{k=0}^{m-1} \exp\left(\boldsymbol{\theta}^T \left(\boldsymbol{t}\left({}^{j,k}\boldsymbol{y}\right) - \boldsymbol{t}\left(\boldsymbol{y}\right)\right)\right)\right)^{-1}, \qquad (5.9)$$

where ${}^{j,k}\boldsymbol{y}$ denotes a vector with ${}^{j,k}y_j = k, {}^{j,k}\boldsymbol{y}_{-j} = \boldsymbol{y}_{-j}$, and the differences in the sufficient statistics in (5.9) can be efficiently evaluated using (5.8). Hence, the resulting pseudo-likelihood function equals

$$\tilde{p}\left(\boldsymbol{Y} = \boldsymbol{y} | \boldsymbol{\theta}\right) := \prod_{j=1}^{g} p\left(Y_j = y_j | \boldsymbol{Y}_{-j} = \boldsymbol{y}_{-j}; \boldsymbol{\theta}\right),$$

[5] An informal but powerful empirical way to check for the algorithm's convergence is to inspect the trace plots of the sufficient statistics of the simulated count configurations. An overview on more formal convergence diagnostics can be found in Cowles & Carlin (1996).

and the maximum pseudo-likelihood estimate $\tilde{\boldsymbol{\theta}}$ is taken to be the maximizer of this function, which can be numerically obtained using standard optimization techniques such as Newton's method or trust region optimization (Nocedal & Wright, 2006).

Maximum likelihood estimation

Given that the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ exists, it is a well-known result that $\mathbb{E}_{\hat{\boldsymbol{\theta}}}\left[\boldsymbol{t}\left(\boldsymbol{Y}\right)\right] = \boldsymbol{t}\left(\boldsymbol{y}^{\text{obs}}\right)$, where $\boldsymbol{t}\left(\boldsymbol{y}^{\text{obs}}\right)$ refers to the vector of observed sufficient statistics (Barndorff-Nielsen, 1978). This results provides the theoretical justification for a heuristic approach for computing the maximum likelihood estimate in the model, which is based on simulating count configurations given an initial parameter value and successive refinement of the initial value by comparing the simulated sufficient statistics to the observed ones. For this, we rely on the stepping algorithm from Hummel et al. (2012) that has been originally introduced for maximum likelihood estimation of exponential random graph models.

In short, we set an arbitrary parameter value $\boldsymbol{\theta}^0$, consider the difference between the log-likelihoods of the observed count configuration $\boldsymbol{y}^{\text{obs}}$ given $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^0$, respectively, and subsequently maximize this difference over $\boldsymbol{\theta}$. However, due to the appearance of the intractable normalizing constants $\kappa\left(\boldsymbol{\theta}\right)$ and $\kappa\left(\boldsymbol{\theta}^0\right)$ in the log-likelihood difference, direct evaluation and maximization is impossible. Instead, we use an analytic law-of-large numbers approximation provided by Geyer (1992) and given by

$$\begin{aligned} &\log\left(p\left(\boldsymbol{Y} = \boldsymbol{y}^{\text{obs}}|\boldsymbol{\theta}\right)\right) - \log\left(p\left(\boldsymbol{Y} = \boldsymbol{y}^{\text{obs}}|\boldsymbol{\theta}^0\right)\right) \\ &\approx \left(\boldsymbol{\theta} - \boldsymbol{\theta}^0\right)^T \boldsymbol{t}\left(\boldsymbol{y}^{\text{obs}}\right) - \log\left(\frac{1}{n}\sum_{i=1}^{n} \exp\left(\left(\boldsymbol{\theta} - \boldsymbol{\theta}^0\right)^T \boldsymbol{t}\left(\boldsymbol{y}_i\right)\right)\right), \end{aligned} \tag{5.10}$$

where $\boldsymbol{y}^1, ..., \boldsymbol{y}^n$ are simulated count configurations given $\boldsymbol{\theta}^0$. Regrettably, the approximation in (5.10) only works well if $\boldsymbol{\theta}$ is "close" to $\boldsymbol{\theta}^0$ (Geyer, 1992; Hummel et al., 2012). Worse, if the vector of observed sufficient statistics $\boldsymbol{t}\left(\boldsymbol{y}^{\text{obs}}\right)$ is not contained in the convex hull of the vectors of simulated sufficient statistics $\boldsymbol{t}\left(\boldsymbol{y}^1\right), ..., \boldsymbol{t}\left(\boldsymbol{y}^n\right)$, the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ does not even exist (Desmarais & Cranmer, 2012a; Hunter et al., 2012).[6]

Before describing a stepping algorithm that iteratively shifts $\boldsymbol{\theta}^0$ closer to $\hat{\boldsymbol{\theta}}$, the approximation in (5.10) itself can be improved by replacing it with a log-normal approximation (Hummel et al., 2012). This log-normal approximation is based on a normal approximation of the vector of sufficient statistics $\boldsymbol{t}\left(\boldsymbol{Y}\right) \sim N\left(\boldsymbol{m}_0, \Sigma_0\right)$, where $\boldsymbol{m}_0$ and Σ_0 are the mean vector and variance-covariance matrix of $\boldsymbol{t}\left(\boldsymbol{Y}\right)$ given $\boldsymbol{\theta}^0$, respectively. Hence, (5.10) can be rewritten as

$$\begin{aligned} &\log\left(p\left(\boldsymbol{Y} = \boldsymbol{y}^{\text{obs}}|\boldsymbol{\theta}\right)\right) - \log\left(p\left(\boldsymbol{Y} = \boldsymbol{y}^{\text{obs}}|\boldsymbol{\theta}^0\right)\right) \\ &\approx \left(\boldsymbol{\theta} - \boldsymbol{\theta}^0\right)^T \left(\boldsymbol{t}\left(\boldsymbol{y}^{\text{obs}}\right) - \hat{\boldsymbol{m}}_0\right) - \frac{1}{2}\left(\boldsymbol{\theta} - \boldsymbol{\theta}^0\right)^T \hat{\Sigma}_0 \left(\boldsymbol{\theta} - \boldsymbol{\theta}^0\right), \end{aligned} \tag{5.11}$$

[6] The is.inCH() function from the ergm package in R (Handcock et al., 2017; Hunter et al., 2008) can be used to check whether a vector lies in the closure of the convex hull of a set of vectors.

where $\hat{\boldsymbol{m}}_0$ and $\hat{\Sigma}_0$ are the sample estimators of $\boldsymbol{m}_0$ and Σ_0, respectively, and (5.11) is maximized by $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^0 + \hat{\Sigma}_0^{-1}\left(\boldsymbol{t}\left(\boldsymbol{y}^{\text{obs}}\right) - \hat{\boldsymbol{m}}_0\right)$. Thus, replacing (5.10) by (5.11) not only provides an improvement in terms of accuracy of the maximum likelihood estimate but also leads to a reduction in computation time, as it eliminates the need for a numerical optimization procedure.

In the actual stepping algorithm, in each iteration t of the algorithm, the vector of observed sufficient statistics $\boldsymbol{t}\left(\boldsymbol{y}^{\text{obs}}\right)$ is moved towards the sample mean of the simulated sufficient statistics $\boldsymbol{t}\left(\boldsymbol{y}^1\right), ..., \boldsymbol{t}\left(\boldsymbol{y}^n\right)$ given $\boldsymbol{\theta}^t$ by replacing $\boldsymbol{t}\left(\boldsymbol{y}^{\text{obs}}\right)$ with a convex combination of the two in the log-likelihood difference in (5.11). Hummel et al. (2012) set a safety margin, 1.05, and chose the weight $\gamma^t \in (0, 1]$ in this convex combination adaptively as the largest value such that $1.05\gamma^t \boldsymbol{t}\left(\boldsymbol{y}^{\text{obs}}\right) + \left(1 - 1.05\gamma^t\right) n^{-1} \sum_{i=1}^n \boldsymbol{t}\left(\boldsymbol{y}^i\right)$ is in the convex hull of $\boldsymbol{t}\left(\boldsymbol{y}^1\right), ..., \boldsymbol{t}\left(\boldsymbol{y}^n\right)$.[7] Subsequently, in each iteration the adjusted log-likelihood difference is maximized over $\boldsymbol{\theta}$. Once $\gamma^t = 1$ for two consecutive iterations in the algorithm, the algorithm is terminated and the final iterated maximum $\boldsymbol{\theta}^{t+1}$ is taken as the maximum likelihood estimate. The algorithm can be summarized in six steps as follows:

(1) Set t to zero and choose $\boldsymbol{\theta}^0$, e.g., as the maximum pseudo-likelihood estimate $\tilde{\boldsymbol{\theta}}$.

(2) Simulate count configurations $\boldsymbol{y}^1, ..., \boldsymbol{y}^n$ given $\boldsymbol{\theta}^t$.

(3) Compute the sample mean $\hat{\boldsymbol{m}}_t$ of the simulated sufficient statistics,
$\hat{\boldsymbol{m}}_t = n^{-1} \sum_{i=1}^n \boldsymbol{t}\left(\boldsymbol{y}^i\right)$.

(4) Choose the largest $\gamma^t \in (0, 1]$ such that $1.05\gamma^t \boldsymbol{t}\left(\boldsymbol{y}^{\text{obs}}\right) + \left(1 - 1.05\gamma^t\right) \hat{\boldsymbol{m}}_t$ lies in the convex hull of $\boldsymbol{t}\left(\boldsymbol{y}^1\right), ..., \boldsymbol{t}\left(\boldsymbol{y}^n\right)$.

(5) Replace $\boldsymbol{t}\left(\boldsymbol{y}^{\text{obs}}\right)$ with $\gamma^t \boldsymbol{t}\left(\boldsymbol{y}^{\text{obs}}\right) + \left(1 - \gamma^t\right) \hat{\boldsymbol{m}}_t$ in (5.11) and maximize. Set $\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t + \hat{\Sigma}_t^{-1}\left(\boldsymbol{t}\left(\boldsymbol{y}^{\text{obs}}\right) - \hat{\boldsymbol{m}}_t\right)$, where $\hat{\Sigma}_t = (n-1)^{-1} \sum_{i=1}^n \left(\boldsymbol{t}\left(\boldsymbol{y}^i\right) - \hat{\boldsymbol{m}}_t\right)\left(\boldsymbol{t}\left(\boldsymbol{y}^i\right) - \hat{\boldsymbol{m}}_t\right)^T$.

(6) If $\gamma^t = \gamma^{t-1} = 1$ (for $t > 0$), terminate and return the (approximate) maximum likelihood estimate $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^{t+1}$. Else, set $t = t + 1$ and return to step (2).

After the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ has been obtained, we draw a final sample of simulated count configurations $\boldsymbol{y}^1, ..., \boldsymbol{y}^n$ given $\hat{\boldsymbol{\theta}}$ to estimate the standard error of $\hat{\boldsymbol{\theta}}$. The standard error consists of two components. First, the typical error resulting from fluctuations in the maximum likelihood estimate's sampling distribution, and second, the error introduced by approximating the log-likelihood difference in (5.11).

Relying on standard asymptotic theory, the first error component can be computed using the inverse of the model's Fisher information matrix $I(\boldsymbol{\theta})$, which itself can only be approximated by the variance-covariance matrix of the simulated count configurations $\boldsymbol{y}^1, ..., \boldsymbol{y}^n$ given $\hat{\boldsymbol{\theta}}$, $\hat{\Sigma}_{t+1}$.[8] However, as is the case for exponential random graph models, it has not yet been shown that such asymptotic arguments can be legitimately employed for approximating the first error component (Butts, 2007). Hummel et al. (2012) and Hunter

[7] More conservative, i.e., larger, values for the safety margin than 1.05 can be chosen if the iterated maximum $\boldsymbol{\theta}^t$ gets stuck, while values between 1 and 1.05 may lead to a faster convergence of the algorithm.

[8] More precisely, the approximate first error component equals the square root of the diagonal of $\hat{\Sigma}_{t+1}^{-1}$.

& Handcock (2006) provided details on the computation of the second error component, which we experimentally found to be negligible in magnitude compared to the first error component when drawing sufficiently large samples of count configurations in step (2) of the above algorithm. Alternatively, approximate standard errors can also be obtained by running a parametric bootstrap procedure, i.e., by taking the standard deviation of maximum likelihood estimates of repeatedly simulated count configurations given $\hat{\boldsymbol{\theta}}$, which naturally accommodates both error components but requires longer computation times.

5.4.3 Goodness-of-fit

Once the maximum likelihood estimate and its standard error are determined, it remains to assess how well the estimated model fits an observed count configuration. Traditional measures for assessing (relative) *goodness-of-fit* such as the AIC (Akaike, 1974), the BIC (Schwarz, 1978), or chi-squared tests (Pearson, 1900) cannot be adequately used in our proposed model for one or several of the following reasons. First, the assumptions underlying the derivation of all of these measures are violated here, as in general, the counts are not independent and identically distributed across the network. Second, network autocorrelated data do not have a clear notion of sample size (Berger et al., 2014), which can lead to ambiguous conclusions depending on the definition of effective sample size adopted. Third, evaluating the model's likelihood function directly is infeasible and hinders practical applicability of the AIC and the BIC. While several extensions of the above measures for correlated data exist, we instead turn to graphical goodness-of-fit procedures based on the predictive count distribution that are straightforward to implement and interpret, and which provide a much richer picture of goodness-of-fit than scalar summary measures.

By construction of the stepping algorithm in Section 5.4.2, simulated sufficient statistics given the maximum likelihood estimate will center around the observed sufficient statistics. At the same time, some estimated models may still lead to simulated count configurations that are vastly different from the observed count configuration and that typically exhibit a very high number of minimum and maximum attainable counts. This phenomenon is well-known in exponential random graph models and termed *degeneracy* (Handcock, 2003). In exponential random graph models, degeneracy problems largely stem from including certain sufficient statistics such as *triangle* and *k-star* counts (Krivitsky, 2012) but we have not found that using any of the sufficient statistics in Section 5.3.1 systematically results in degeneracy of our model.

If the estimated model is deemed non-degenerate, sufficient statistics that have not been added to the model are not guaranteed to be well-reproduced by the simulated count configurations. However, if these unmodeled sufficient statistics are recovered well, there is evidence that the modeled structural properties are the only ones necessary to adequately describe the count generating process (Robins, Snijders, et al., 2007). Else, this suggests refinements to the model, e.g., by including more parameters. In the following, we propose two (sets of) sufficient statistics for assessing model fit that serve as independent goodness-of-fit measures: the *raw count distribution*, and *count 2-stars*.

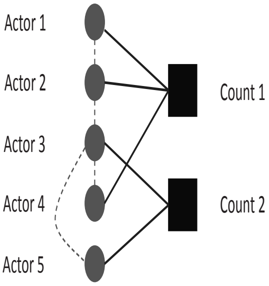

Figure 5.1 Hypothetical count configuration for five actors (A) in a network, where dashed lines indicate ties between actors. Here, the neighbor count 2-stars are $(A1, A2)$ and $(A3, A5)$.

First, we would like our estimated model to properly capture the observed raw count distribution, irrespectively of the dependence of the counts. Accordingly, we denote the number of actors in a network having count i by the sufficient statistic $t_i^{\mathrm{C}}(\boldsymbol{y}) := \sum_{j=1}^{g} \mathbb{1}(y_j = i), i \in \{0, ..., m-1\}$. Second, we wish our model to appropriately reveal potential clustering tendencies in the counts. While the sufficient statistics embodying network autocorrelation in (5.4) are designed to reflect such clustering tendencies for connected actors, we introduce a set of count 2-star sufficient statistics that check if the model can reproduce various other forms of clustering in the data. We define a *general count 2-star* as a pair of actors that have the same count, adapting the definition of a *general event 2-star* in exponential random graph models for affiliation networks in Agneessens & Roose (2008). Since the total number of general count 2-stars in a count configuration is a function of the raw count distribution, modeling the number of general count 2-stars and the raw count distribution is equivalent though.[9] Instead, we consider more specific count 2-stars that also take network properties and/or actor attributes into account. We define a *neighbor count 2-star* as a pair of tied actors, based on connectivity mechanism i, that have the same count. We set the associated sufficient statistic to the total number of neighbor count 2-stars in a count configuration, i.e., $t_i^{2*W}(\boldsymbol{y}) := \sum_{j=1}^{g} \sum_{k=1}^{g} \mathbb{1}(W_{ijk} > 0)\, \mathbb{1}(y_j = y_k), i \in \{1, ..., r\}$. As such, it captures rather rigidly count clustering tendencies of tied actors and we have found that it only mildly correlates with the sufficient statistic $t_i^{\mathrm{NAC}}(\boldsymbol{y})$ representing network autocorrelation.[10] Figure 5.1 visualizes neighbor count 2-stars for a simple hypothetical count and network configuration. Similarly, we can define *attribute count 2-stars* that include actor attributes and measure count clustering based on actor attributes, e.g., by counting the number of pairs of actors that not only have the same count but also the same covariate value.

[9] The total number of general count 2-stars in a count configuration can be written as $t^{2^*}(\boldsymbol{y}) := \sum_{i=0}^{m-1} \binom{t_i^{\mathrm{C}}(\boldsymbol{y})}{2}$.

[10] We ignored the strength of actor ties in the definition of $t_i^{2*W}(\boldsymbol{y})$, which could be easily modeled though by replacing $\mathbb{1}(W_{ijk} > 0)$ with W_{ijk}.

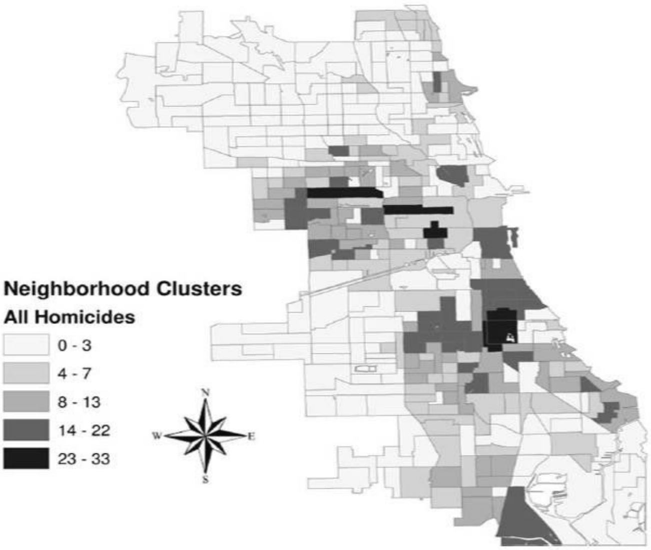

Figure 5.2 Spatial distribution of homicide counts for the years 1989 through 1991 across the 343 Chicago neighborhood clusters.

5.5 Application: Homicide in Chicago neighborhoods

In this section, we illustrate our model by reanalyzing homicide data for the city of Chicago, Illinois.[11] The data consist of aggregated homicide counts for the years 1989 through 1991 and socio-economic variables for 343 neighborhood clusters taken from the 1990 census. These neighborhood clusters are composed of the city's 865 census tracts that are geographically adjacent and socially similar, resulting in fairly demographically-homogenous neighborhood clusters (Morenoff et al., 2001). Figure 5.2 shows the spatial distribution of the homicide counts across the 343 neighborhoods, revealing that neighborhoods with similar homicide counts tend to cluster in space.[12]

Bhati (2008) employed a generalized Poisson regression model using four socio-economic variables to explain the spatial patterns of homicide across the 343 neighborhoods. These were, first, a neighborhood resource deprivation index (RDI); second, the residential stability of a neighborhood (RST); third, young men aged 15 to 25 as a proportion of a neighborhood's total population (MEN); fourth, the logarithm of a neighborhood's population (POP). As expected, all of these variables were found to be positively associated with homicide. In addition, Bhati (2008) considered more extended models that hinted at overdispersion, compared to a Poisson specification, and spatial autocorrelation in the counts.

[11] The data can be obtained from the Inter-university Consortium for Political and Social Research after submitting a data access proposal at `http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/4079`.

[12] The figure was copied from Bhati (2008) with kind permission from the author.

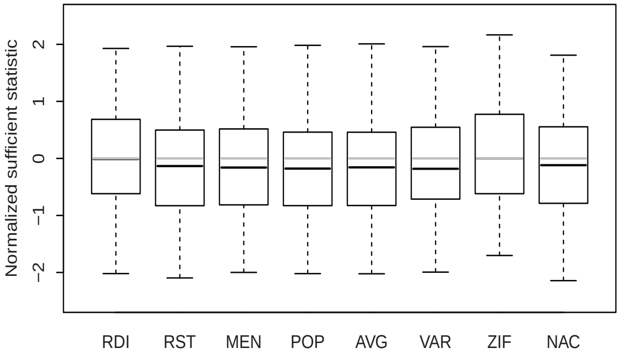

Figure 5.3 Box plots of the normalized simulated sufficient statistics for the Chicago homicide data set. The box plots show the median (thick black solid line), the interquartile range (black solid rectangle), as well as the 2.5-th and 97.5-th percentiles (short solid black lines). The thick grey lines show the normalized observed sufficient statistics that are equal zero.

We applied the proposed discrete exponential family model to explain the structural properties of the Chicago homicide counts as a function of exogenous and endogenous forces in the neighborhood network. We used the four socio-economic variables described above as actor attributes and accordingly specified sufficient statistics of the form in (5.5), representing count heterogeneity based on actor attributes, where we took the corresponding elements of Q to be the attainable counts themselves.[13] Furthermore, we included three endogenous sufficient statistics of the form in (5.5) to model the baseline shape of the count distribution: first, the sum of all counts, controlling for the mean count (AVG); second, the sum of all squared counts, controlling for the variance of the counts (VAR); third, the number of zero counts, controlling for zero-inflation (ZIF). Lastly, we added a scalar sufficient statistic of the form in (5.4) embodying network autocorrelation (NAC) to the model that is based on a binary spatial neighborhood adjacency matrix.

We implemented our model and obtained the maximum likelihood estimate using the stepping algorithm in Section 5.4.2.[14] Subsequently, we simulated 25,000,000 count configurations given the maximum likelihood estimate and sub-sampled every 25,000-th draw to compute standard errors and assess goodness-of-fit; Figure 5.3 depicts box plots of the corresponding normalized simulated sufficient statistics, i.e., after subtracting the observed sufficient statistics and dividing by the respective sample standard deviations. As is evident, their sample means closely follow the observed sufficient statistics, which suggests that the stepping algorithm has indeed converged to the maximum likelihood estimate. Moreover, the simulated marginal predictive distributions for the homicide counts in Figure 5.4 do not indicate any degeneracy issues with our model specification.

[13] We set the attainable upper bound of the counts to twice the largest count value in the data, which is 33. Our results are also robust to larger values for the upper bound that lead to longer computation times though.

[14] We set the initial value $\boldsymbol{\theta}^0$ to the maximum pseudo-likelihood estimate, the safety margin in the stepping algorithm to 10, and iteratively refined $\boldsymbol{\theta}^0$ until the stopping criterion was reached.

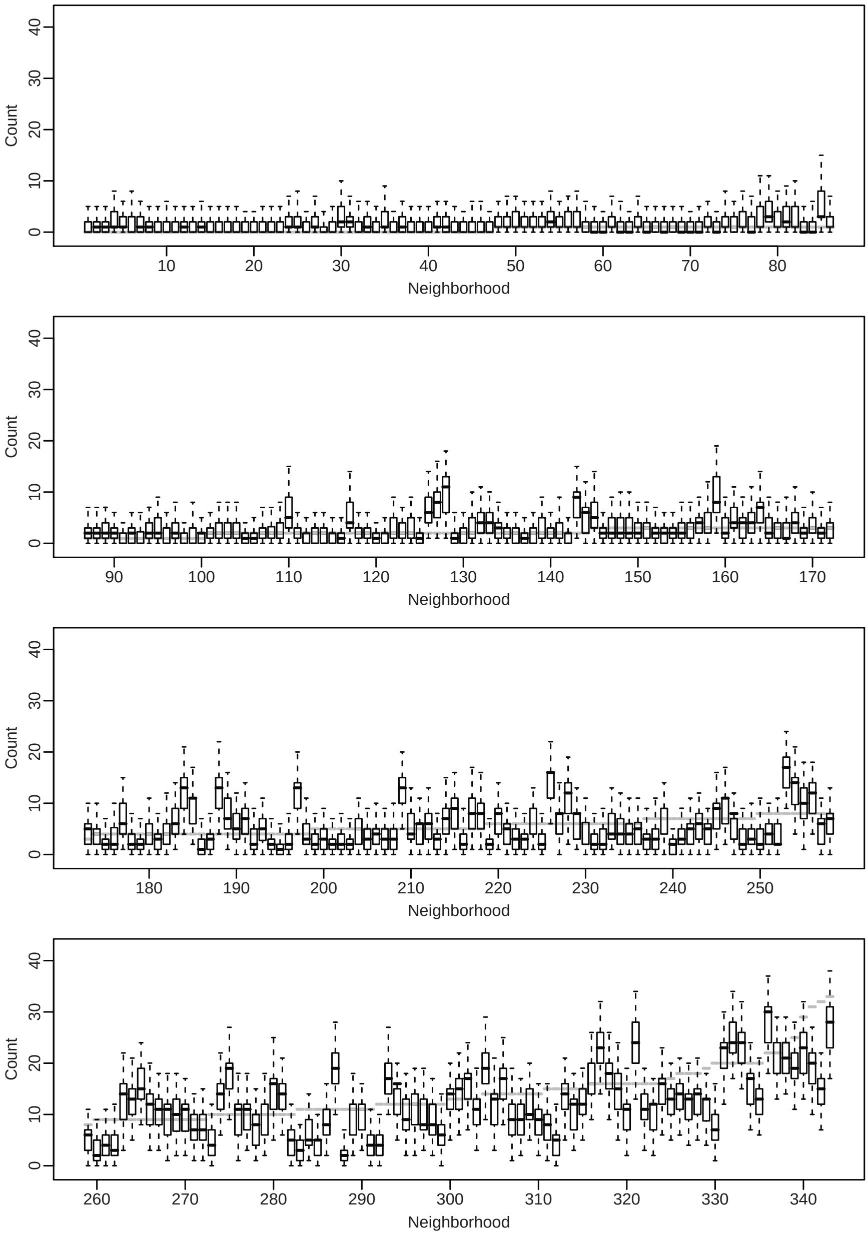

Figure 5.4 Box plots of the simulated marginal homicide counts for the 343 Chicago neighborhoods. The box plots show the mode (thick black solid line), the interquartile range (black solid rectangle), as well as the 2.5-th and 97.5-th percentiles (short solid black line). The thick grey lines show the observed homicide counts.

Table 5.1 Maximum likelihood estimates and asymptotic as well as bootstrapped standard errors (SE) under the full model and under a reduced model that ignores network autocorrelation for the Chicago homicide data set.

	Full model			Reduced model		
		Asymptotic	Bootstrapped		Asymptotic	Bootstrapped
Parameter	Estimate	SE	SE	Estimate	SE	SE
Resource deprivation	0.341	0.038	0.044	0.415	0.038	0.045
Residential stability	0.895	0.412	0.447	0.944	0.412	0.440
Young men	2.907	0.862	0.996	3.103	0.905	1.039
Population	0.530	0.067	0.074	0.536	0.066	0.075
Count mean	-5.090	0.611	0.681	-5.113	0.614	0.677
Count variance	-0.017	0.003	0.003	-0.025	0.003	0.003
Zero inflation	-0.280	0.188	0.215	-0.474	0.209	0.208
Network autocorrelation	0.021	0.004	0.005			

Table 5.1 reports the maximum likelihood estimates and their asymptotic standard errors. In addition, we computed bootstrapped standard errors based on the 1,000 drawn count configurations given the maximum likelihood estimate, which are also included in Table 5.1. While the bootstrapped standard errors are consistently somewhat larger than their asymptotic counter parts, these differences do not alter any of our substantive findings. At the same time, this underlines that researchers should be cautious when interpreting asymptotic standard errors in the model and computer-intensive bootstrapped standard errors are not available. In line with the results in Bhati (2008), the four covariates RDI, RST, MEN, and POP are also positively associated with higher levels of homicide in our model. The endogenous baseline terms are generally difficult to interpret but their negative coefficients suggest that the homicide counts are smaller, vary less, and exhibit fewer zeros than in typical count configurations given a model that would ignore these baseline terms. Finally, the positive coefficient of the network autocorrelation parameter indicates that the spatial clustering in the homicide counts cannot be accounted for by socio-economic variability in the neighborhoods and the shape of the baseline count distribution alone; instead, underlying network autocorrelation is central to explaining the non-random spatial homicide pattern in the data.

We assessed the model fit based on two independent goodness-of-fit measures from Section 5.4.2, the raw count distribution and neighbor count-2 stars. Figure 5.5 (top row) shows the simulated sampling distribution of these quantities, along with the actual raw count distribution and the number of neighbor count 2-stars in the observed data. As can be seen, the model captures the raw count distribution very well, as is the case for the neighbor count 2-stars. Hence, the proposed model specification is able to reproduce a variety of distinct structural properties of the homicide counts besides the one explicitly modeled and can be said to adequately describe the count generating process.

As the development of this model was primarily inspired by the desire to sensibly analyze network autocorrelated count data, we were particularly interested in understanding the effect of network autocorrelation on structural properties of the homicide counts.

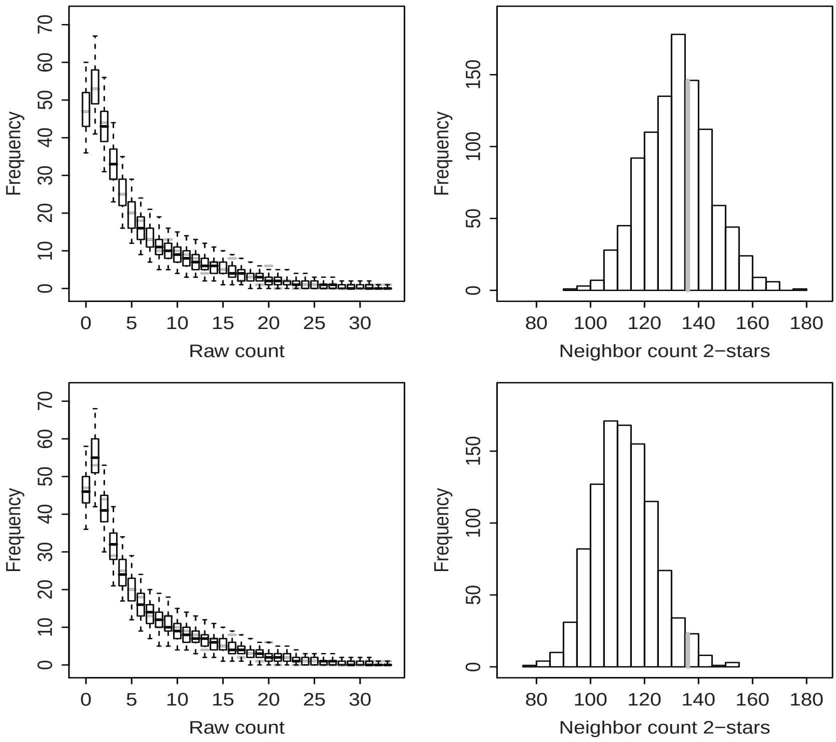

Figure 5.5 Left panel: Box plots of the simulated raw count distribution under the full model (top row) and under a reduced model ignoring network autocorrelation (bottom) for the Chicago homicide data set. The box plots show the median (thick black solid line), the interquartile range (black solid rectangle), as well as the 2.5-th and 97.5-th percentiles (short solid black lines). The thick grey lines show the observed raw counts. Right panel: Histogram of the simulated number of neighbor count 2-stars under the full model (top) and under the reduced model (bottom). The thick grey lines show the observed number of neighbor count 2-stars.

Therefore, we estimated a reduced model leaving out network autocorrelation and compared its inferences to those under the full model including network autocorrelation. The maximum likelihood estimates under the reduced model in Table 5.1 are overall qualitatively similar to the ones under the full model, with the estimates of the four covariates being slightly inflated and the estimates of the baseline terms slightly deflated compared to the full model. Furthermore, Figure 5.5 (bottom row) shows that the reduced model preserves the raw count distribution, while it does not seem to accurately capture the number of neighbor count 2-stars, i.e., the number of adjacent neighborhoods that have the same count. In sum, these results provide further evidence that network autocorrelation is an essential mechanism underlying the spatial distribution of homicide in this data set.

5.6 Conclusions

In this chapter, we introduced a discrete exponential family model for analyzing network autocorrelated count data. In the model, we used a discrete exponential family to specify the joint count distribution, where we did not rely on any distributional assumptions on the marginal or conditional counts. We showed how to specify the joint distribution in terms of sufficient statistics that capture a range of characteristic structural properties of a count configuration, such as network autocorrelation, the number of zero counts, or count homo- or heterogeneity through actor attributes variability. In addition, we provided algorithms to simulate count configurations from the model and to perform maximum likelihood-based inference. We implemented and illustrated the usefulness of our model by analyzing the drivers of homicide across 343 neighborhoods in Chicago, Illinois, where we found that network autocorrelation was fundamental to understanding the spatial clustering of the homicide counts.

The work in this chapter leaves room for several improvements and extensions of the model. First, we operationalized and modeled network autocorrelation through a statistic that is based on the sum of absolute differences between counts, or generally count attributes, of tied actors. At the same time, this is only one possible way of capturing network autocorrelation in a count configuration, and it may well be that other formulations, e.g., along the lines of the introduced neighbor count 2-stars or building upon spatial association measures, are better suited in certain situations. It remains to systematically analyze the capacity and power of different operationalizations and to formulate recommendations as to when to use which. Second, we relied upon a simple Metropolis algorithm to simulate count configurations from the model that explores the space of attainable count configurations rather slowly, in particular when the chosen upper bound for the counts is very high. Thus, developing adaptive algorithms that automatically find “good” candidate count configurations would be a valuable improvement of the model. Third, we implicitly used a discrete uniform *base measure* in the discrete exponential family that defines the joint count distribution. Generalizing the model to include additional base measures, where the base measure represents the joint count distribution in absence of any other endogenous or exogenous forces, and thus allowing for even richer modeling is left for future work.

To conclude, we hope that providing researchers with a flexible and interpretable model for analyzing network autocorrelated count data of manifold distributional shapes will contribute to a better understanding of influence patterns underlying many count processes.

Chapter 6

Epilogue

In this thesis, we developed a fully Bayesian framework to model network autocorrelation of a variable of interest using the network autocorrelation model. Furthermore, we introduced a discrete exponential family model for analyzing network autocorrelated count data for which the network autocorrelation model itself is not well-suited. In this final chapter, we summarize our most important findings, highlight the remaining limitations of our work, and outline future research directions for modeling network autocorrelation.

6.1 Bayesian analysis of the network autocorrelation model

We have provided comprehensive Bayesian inference methods for the network autocorrelation model in the first three chapters of the body of the thesis. In Chapter 2, we focused on Bayesian estimation of a first-order model and considered estimation of higher-order models in Chapter 4. Analogously, we presented Bayes factors for testing hypotheses on a single network autocorrelation parameter ρ in a first-order model in Chapter 3, which we then generalized to test equality and inequality constrained hypotheses on multiple network autocorrelation parameters in higher-order models in Chapter 4. For both Bayesian estimation of and hypothesis testing in the network autocorrelation model, the specification of the prior for the network autocorrelation parameter(s) is a major challenge. Thus, we constructed several theoretically and empirically guided priors for the network autocorrelation parameter(s) and assessed the sensitivity of the performance of our methods to different prior choices.

By means of an extensive simulation study, we found that the Bayesian estimators based on the two versions of the newly derived Jeffreys prior for the first-order model do not perform substantially differently than the maximum likelihood estimator with respect to bias of the network autocorrelation parameter; in particular, relying on these priors does not lessen the severe negative bias in the estimation of ρ for high levels of network density. At the same time, using Independence Jeffreys prior and the standard uniform prior for ρ results in accurate coverage of credible intervals for ρ even for high levels of network density and small network sizes, as opposed to the below-nominal coverage of maximum likelihood-based confidence intervals in such scenarios. We also observed that the Bayesian estimator based on the empirical informative prior for ρ dramatically decreases the bias in the estimation of ρ if the expected network autocorrelation in a study is in line with previous empirical results on the magnitude of ρ. Meanwhile, several authors (Bao, 2013; Z. Yang, 2015; Yu et al., 2015) have proposed different bias-corrected maximum likelihood estimation procedures that almost eliminate the bias in the estimation of ρ when network density is not excessively high, as is the case for most spatial networks. However, it remains to investigate if the good performance of these procedures also holds for very small and dense networks that are often encountered in social network research. Nevertheless, this stimulates the future development of similar bias-corrected Bayesian estimators in the network autocorrelation model.

We did not pursue further the two versions of Jeffreys prior in higher-order models, as relying on these priors in a first-order model merely marginally improves inferences compared to employing the uniform prior for ρ but results in more complex posteriors and considerably longer computation times.[1] Moreover, higher-order models have not been applied nearly as extensively in the literature as the first-order model, and we saw little value in constructing an empirical reference prior for the network autocorrelation parameters here. Instead, we used a general multivariate normal prior for the network autocorrelation parameters in combination with standard non-informative priors for the remaining model parameters. Based on a non-informative and essentially uniform prior specification for the network autocorrelation parameters in a second-order model, we came to qualitatively same conclusions as in the first-order model in terms of bias of the network autocorrelation parameters and coverage of the corresponding credible intervals. In addition, these conclusions are robust to modest overlap between two connectivity matrices.

Apart from enabling researchers to include various amounts of prior knowledge about the model parameters to their analyses, our advocated Bayesian framework also permits researchers to simultaneously test multiple competing hypotheses on the network autocorrelation parameter(s), which allows for much richer and more nuanced insights compared to classical null hypothesis significance testing. First, we proposed several Bayes factors for the first-order model that quantify the amount of relative evidence in the data for precise and interval hypotheses on ρ. Similar as in Bayesian estimation, the amount of evidence for interval hypotheses, such as $H_1 : 0 < \rho < 1$, is sensitive to the chosen prior for ρ though. We conducted a large simulation study and found that Bayes factors based on the empirical informative prior and the uniform prior for ρ provide the largest evidence for a true data-generating hypothesis and show consistent behavior, i.e., the evidence for a true data-generating hypothesis is increasing with the network size. On the other hand, we also noticed that using a fractional Bayes factor based on an improper prior for ρ results in sub-optimal inferences and is therefore not recommended. Second, we presented Bayes factors based on automatically constructed multivariate normal priors for the network autocorrelation parameters for testing any number of equality and inequality constrained hypotheses on them in higher-order models, such as $H_1 : \rho_1 > \rho_2 = 0$ and $H_2 : \rho_1 > \rho_2 > 0$.

[1] The derivation of Jeffreys rule prior and Independence Jeffreys prior in higher-order network autocorrelation models is a straightforward exercise and analogous to the derivation in the first-order model.

As in the first-order model, the analyzed Bayes factors appear to be consistent but require somewhat larger network sizes to provide substantial evidence for a true-data generating hypothesis. At the same time, both in first-order and higher-order models, the evidence for a true data-generating hypothesis is also decreasing with the network density, owing to the increasingly distorted (concentrated) likelihood function. Hence, future work on Bayesian bias-correction procedures would not only be of great value for reducing the negative bias in the estimation of the network autocorrelation parameter(s) but equally beneficial for improving inference about them using Bayes factors.

All of our proposed methods in the network autocorrelation model rely on the assumption that the variable of interest is normally distributed and can be modeled assuming homogeneous statistical errors across the network. Needless to say, this assumption may often not be met in empirical practice and more research is needed to examine how robust our findings are to violations of these underlying conjectures. While it may seem tempting to simply expand our methods to accommodate variables of interest following other continuous distributions, such as the beta, the gamma, or the log-normal, using generalized linear models theory, such approaches would usually build upon pseudo-likelihood inference and may not appropriately capture complex dependencies in the data. Instead, a more promising approach could be based on generalizing the discrete exponential family model from Chapter 5 to continuous data, allowing for full likelihood-based inference and retaining distributional flexibility.[2]

6.2 Network autocorrelation modeling of count data

In Chapter 5, we made use of an entirely different formalism to model network autocorrelated count data, the discrete exponential family. In contrast to most network autocorrelation count models currently available in the literature, relying on the discrete exponential formalism allows for full likelihood-based inference, and we demonstrated how the formalism can be utilized to model network autocorrelated count data. In particular, we showed how to incorporate network autocorrelation and covariate effects by defining the joint count distribution as a discrete exponential family specified in terms of suitable sufficient statistics representing these effects. Hence, even though leaving the network autocorrelation model framework and taking a rather different modeling approach might make this thesis appear somewhat less consistent at first sight, we believe this stride makes it in fact methodologically more sound upon second thought.

One considerable strength of our proposed model is that no potentially restrictive distributional assumptions on the marginal or conditional counts need to be made, but the model is able to handle a wide range of different count configurations. On the other hand, we implicitly assumed a discrete uniform base measure in the discrete exponential family formulation of the joint count distribution. The choice of this base measure determines the joint count distribution net of any other effects in the model and likewise, at least par-

[2] Interestingly, recent research on the network autocorrelation model has focused more on developing model extensions for non-continuous data such as ordinal data (Dow, 2008) and multinomial data (Y. Wang et al., 2014).

tially, governs the shape of the marginal count distributions under a fully specified model that is, however, in general analytically unavailable. The derivation of other meaningful base measures, e.g., along the lines of Krivitsky (2012) and yielding Poisson-like marginal count distributions, is an important topic left for future research. This especially applies to a more general model with no a priori upper bound for the counts, where certain base measures, in combination with particular parameter configurations, may lead to non-finite normalizing constants.

Finally, we provided maximum likelihood-based inference methods for the model in Chapter 5 only, as opposed to proposing Bayesian inferential tools for the network autocorrelation model in Chapters 2, 3, and 4. Here, practical implementation of similar Bayesian tools is greatly hindered by the intractability of the normalizing constant in the model's likelihood, resulting in so-called *doubly intractable* posteriors (Murray et al., 2006) due to an additional parameter-dependent normalizing constant. For the future, adapting existing algorithms for Bayesian analyses of related exponential random graph models (Caimo & Friel, 2011) would be a valuable addition to the inferential toolbox for the model.

6.3 Concluding thoughts

In addition to introducing new methodology, we devoted considerable effort to developing efficient implementations of these methods, which we meticulously described throughout the thesis. While, in principle, this allows researchers to straightforwardly reconstruct our implementations in their familiar software environment and apply our methods in practice, we are aware that it is rather the wish being the father to the thought here; merely providing verbal and symbolic guidance will hardly bridge the gap between presented theory and practical application. Thus, a major overarching goal for future research is to bundle our existing implementations in a freely available software package to help disseminate our methods within the network science community and facilitate obtaining more profound insights into the structure of network autocorrelation. An obvious thought that comes to mind is how these insights can be actually used to answer substantive research problems encountered in empirical practice. Admittedly, this raises questions that go beyond the scope of this thesis and primarily need to be addressed by applied network researchers and policy makers. To conclude, we nevertheless hope that the methods developed in this thesis can ultimately help policy makers in conducting targeted interventions taking into account the structure of network autocorrelation.

References

Agneessens, F., & Roose, H. (2008). Local Structural Properties and Attribute Characteristics in 2-mode Networks: p* Models to Map Choices of Theater Events. *Journal of Mathematical Sociology*, *32*(3), 204-237.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716-723.

Andersson, H., Jonsson, L., & Ögren, M. (2010). Property Prices and Exposure to Multiple Noise Sources: Hedonic Regression with Road and Railway Noise. *Environmental and Resource Economics*, *45*(1), 73-89.

Anselin, L. (1982). A Note on Small Sample Properties of Estimators in a First-Order Spatial Autoregressive Model. *Environment and Planning A*, *14*(8), 1023-1030.

Anselin, L. (1984). Specification tests on the structure of interaction in spatial econometric models. *Papers of the Regional Science Association*, *54*(1), 165-182.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models.* Dordrecht, The Netherlands: Springer.

Anselin, L. (1990). Some robust approaches to testing and estimation in spatial econometrics. *Regional Science and Urban Economics*, *20*(2), 141-163.

Anselin, L. (2002). Under the hood: Issues in the specification and interpretation of spatial regression models. *Agricultural Economics*, *27*(3), 247-267.

Anselin, L. (2003). Spatial Externalities, Spatial Multipliers, And Spatial Econometrics. *International Regional Science Review*, *26*(2), 153-166.

Anselin, L., & Le Gallo, J. (2006). Interpolation of Air Quality Measures in Hedonic House Price Models: Spatial Aspects. *Spatial Economic Analysis*, *1*(1), 31-52.

Anselin, L., & Lozano-Gracia, N. (2008). Errors in variables and spatial effects in hedonic house price models of ambient air quality. *Empirical Economics*, *34*(1), 5-34.

Anselin, L., Lozano-Gracia, N., Deichmann, U., & Lall, S. (2010). Valuing Access to Water - A Spatial Hedonic Approach, with an Application to Bangalore, India. *Spatial Economic Analysis*, *5*(2), 161-179.

Anselin, L., Varga, A., & Acs, Z. (2000). Geographical and sectoral characteristics of academic knowledge externalities. *Papers in Regional Science*, *79*(4), 435-443.

Arbia, G., & Basile, R. (2005). Spatial dependence and non-linearities in regional growth behaviour in Italy. *Statistica*, *65*(2), 145-167.

Armstrong, R. J., & Rodríguez, D. A. (2006). An evaluation of the Accessibility Benefits of Commuter Rail in Eastern Massachusetts using Spatial Hedonic Price Functions. *Transportation*, *33*(1), 21-43.

Atkinson, K. E. (1989). *An Introduction to Numerical Analysis* (Second ed.). New York, NY: John Wiley & Sons.

Badinger, H., & Egger, P. (2011). Estimation of higher-order spatial autoregressive cross-section models with heteroscedastic disturbances. *Papers in Regional Science*, *90*(1), 213-235.

Baller, R. D., Anselin, L., Messner, S. F., Deane, G., & Hawkins, D. F. (2001). Structural Covariates of U.S. County Homicide Rates: Incorporating Spatial Effects. *Criminology*, *39*(3), 561-588.

Bao, Y. (2013). Finite Sample Bias of the QMLE in Spatial Autoregressive Models. *Econometric Theory*, *29*(1), 68-88.

Barndorff-Nielsen, O. E. (1978). *Information and Exponential Families in Statistical Theory.* New York, NY: John Wiley & Sons.

Barnett, N. P., Ott, M. Q., Rogers, M. L., Loxley, M., Linkletter, C., & Clark, M. A. (2014). Peer Associations for Substance Use and Exercise in a College Student Social Network. *Health Psychology*, *33*(10), 1134-1142.

Bartlett, M. S. (1957). A comment on D. V. Lindley's statistical paradox. *Biometrika*, *44*(3-4), 533-534.

Bates, D., & Maechler, M. (2017). Matrix: Sparse and Dense Matrix Classes and Methods [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=Matrix` (R package version 1.2-8)

Beck, N., Gleditsch, K. S., & Beardsley, K. (2006). Space is More than Geography: Using Spatial Econometrics in the Study of Political Economy. *International Studies Quarterly*, *50*(1), 27-44.

Becker, R. A., Brownrigg, R., Deckmyn, A., Minka, T. P., & Wilks, A. R. (2016). maps: Draw Geographical Maps [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=maps` (R package version 3.1.1)

Berger, J. O., Bayarri, M. J., & Pericchi, L. R. (2014). The Effective Sample Size. *Econometric Reviews*, *33*(1-4), 197-217.

Berger, J. O., Boukai, B., & Wang, Y. (1997). Unified Frequentist and Bayesian Testing of a Precise Hypothesis. *Statistical Science*, *12*(3), 133-160.

Berger, J. O., de Oliveira, V., & Sansó, B. (2001). Objective Bayesian Analysis of Spatially Correlated Data. *Journal of the American Statistical Association*, *96*(456), 1361-1374.

Berger, J. O., & Mortera, J. (1999). Default Bayes Factors for Nonnested Hypothesis Testing. *Journal of the American Statistical Association*, *94*(446), 542-554.

Berger, J. O., & Sellke, T. (1987). Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence. *Journal of the American Statistical Association*, *82*(397), 112-122.

Bernardo, J. M. (1979). Reference Posterior Distributions for Bayesian Inference. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *41*(2), 113-147.

Bernat Jr., G. A. (1996). Does Manufacturing Matter? A Spatial Econometric View of Kaldor's Laws. *Journal of Regional Science*, *36*(3), 463-477.

Bhat, C. R., Paleti, R., & Singh, P. (2014). A Spatial Multivariate Count model for Firm Location Decisions. *Journal of Regional Science*, *54*(3), 462-502.

Bhati, A. S. (2008). A Generalized Cross-Entropy Approach for Modeling Spatially Correlated Counts. *Econometric Reviews*, *27*(4-6), 574-595.

Bivand, R., Keitt, T., & Rowlingson, B. (2017). rgdal: Bindings for the "Geospatial" Data Abstraction Library [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=rgdal` (R package version 1.2-8)

Bivand, R., & Piras, G. (2015). Comparing Implementations of Estimation Methods for Spatial Econometrics. *Journal of Statistical Software*, *63*(18), 1-36.

Bivand, R., & Szymanski, S. (2000). Modelling the spatial impact of the introduction of Compulsory Competitive Tendering. *Regional Science and Urban Economics*, *30*(2), 203-219.

Bolstad, W. M. (2009). *Understanding Computational Bayesian Statistics.* New York, NY: John Wiley & Sons.

Bordignon, M., Cerniglia, F., & Revelli, F. (2003). In search of yardstick competition: a spatial analysis of Italian municipality property tax setting. *Journal of Urban Economics*, *54*(2), 199-217.

Box, G. E., & Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis.* New York, NY: John Wiley & Sons.

Braeken, J., Mulder, J., & Wood, S. (2015). Relative Effects at Work: Bayes Factors for Order Hypotheses. *Journal of Management*, *41*(2), 544-573.

Brooks, S. P. (1998). Markov chain Monte carlo method and its application. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *47*(1), 69-100.

Brueckner, J. K., & Saavedra, L. A. (2001). Do Local Governments Engage in Strategic Property-Tax Competition? *National Tax Journal*, *54*(2), 203-229.

Buonanno, P., Montolio, D., & Vanin, P. (2009). Does Social Capital Reduce Crime? *Journal of Law and Economics*, *52*(1), 145-170.

Burt, R. S., & Doreian, P. (1982). Testing a structural model of perception: Conformity and deviance with respect to Journal norms in elite sociological methodology. *Quality and Quantity*, *16*(2), 109-150.

Butts, C. T. (2007). Models for Generalized Location Systems. *Sociological Methodology*, *37*(1), 283-348.

Butts, C. T. (2008). Social Network Analysis with sna. *Journal of Statistical Software*, *24*(6), 1-51.

Caimo, A., & Friel, N. (2011). Bayesian inference for exponential random graph models. *Social Networks*, *33*(1), 41-55.

Can, A. (1992). Specification and estimation of hedonic housing price models. *Regional Science and Urban Economics*, *22*(3), 453-474.

Carlin, B. C., & Louis, T. A. (2000). Empirical Bayes: Past, Present and Future. *Journal of the American Statistical Association*, *95*(452), 1286-1289.

Carruthers, J. I., & Clark, D. E. (2010). Valuing Environmental Quality: A Space-based Strategy. *Journal of Regional Science*, *50*(4), 801-832.

Casella, G., Girón, F. J., Martínez, M. L., & Moreno, E. (2009). Consistency of Bayesian procedures for variable selection. *The Annals of Statistics*, *37*(3), 1207-1228.

Castro, M., Paleti, R., & Bhat, C. R. (2012). A latent variable representation of count data models to accommodate spatial and temporal dependence: Application to predicting crash frequency at intersections. *Transportation Research Part B: Methodological*, *46*(1), 253-272.

Chance, B. L., & Rossman, A. J. (2006). *Investigating Statistical Concepts, Applications, and Methods*. Belmont, CA: Duxbury Press.

Chang, H. (2008). Spatial analysis of water quality trends in the Han River basin, South Korea. *Water Research*, *42*(13), 3285-3304.

Chib, S., & Greenberg, E. (1995). Understanding the Metropolis-Hastings Algorithm. *The American Statistician*, *49*(4), 327-335.

Cohen, J. P., & Coughlin, C. C. (2008). Spatial Hedonic Models of Airport Noise, Proximity, and Housing Prices. *Journal of Regional Science*, *48*(5), 859-878.

Conigliani, C., & O'Hagan, A. (2000). Sensitivity of the Fractional Bayes Factor to Prior Distributions. *The Canadian Journal of Statistics*, *28*(2), 343-352.

Conway, D., Li, C. Q., Wolch, J., Kahle, C., & Jerrett, M. (2010). A Spatial Autocorrelation Approach for Examining the Effects of Urban Greenspace on Residential Property Values. *The Journal of Real Estate Finance and Economics*, *41*(2), 150-169.

Cowles, M. K., & Carlin, B. P. (1996). Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. *Journal of the American Statistical Association*, *91*(434), 883-904.

Csárdi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, *Complex Systems*, 1695. Retrieved from `http://igraph.org`

Dall'erba, S. (2005). Productivity convergence and spatial dependence among Spanish regions. *Journal of Geographical Systems*, *7*(2), 207-227.

Dall'erba, S., Percoco, M., & Piras, G. (2009). Service industry and cumulative growth in the regions of Europe. *Entrepreneurship & Regional Development*, *21*(4), 333-349.

Davis, G. F., Yoo, M., & Baker, W. E. (2003). The Small World of the American Corporate Elite, 1982–2001. *Strategic Organization*, *1*(3), 301-326.

DeGroot, M. H., & Schervish, M. J. (2010). *Probability and Statistics* (Fourth ed.). Boston, MA: Addison-Wesley.

De Oliveira, V. (2010). Objective Bayesian Analysis for Gaussian Random Fields. In M.-H. Chen, P. Müller, D. Sun, K. Ye, & D. K. Dey (Eds.), *Frontiers of Statistical Decision Making and Bayesian Analysis: In Honor of James O. Berger* (p. 497-511). New York, NY: Springer.

De Oliveira, V. (2012). Bayesian analysis of conditional autoregressive models. *Annals of the Institute of Statistical Mathematics*, *64*(1), 107-133.

De Oliveira, V., & Song, J. J. (2008). Bayesian Analysis of Simultaneous Autoregressive Models. *Sankhya: The Indian Journal of Statistics, Series B (2008-)*, *70*(2), 323-350.

Desmarais, B. A., & Cranmer, S. J. (2012a). Statistical Inference for Valued-Edge Networks: The Generalized Exponential Random Graph Model. *PLOS ONE*, *7*(1), 1-12.

Desmarais, B. A., & Cranmer, S. J. (2012b). Statistical mechanics of networks: Estimation and uncertainty. *Physica A: Statistical Mechanics and its Applications*, *391*(4), 1865-1876.

Ding, J., & Zhou, A. (2007). Eigenvalues of rank-one updated matrices with some applications. *Applied Mathematics Letters*, *20*(12), 1223-1226.

Dittrich, D., Leenders, R. T., & Mulder, J. (2017). Bayesian estimation of the network autocorrelation model. *Social Networks*, *48*, 213-236.

Dittrich, D., Leenders, R. T., & Mulder, J. (in press). Network Autocorrelation Modeling: A Bayes Factor Approach for Testing (Multiple) Precise and Interval Hypotheses. *Sociological Methods & Research*.

Doreian, P. (1980). Linear Models with Spatially Distributed Data: Spatial Disturbances or Spatial Effects? *Sociological Methods & Research*, *9*(1), 29-60.

Doreian, P. (1981). Estimating Linear Models with Spatially Distributed Data. *Sociological Methodology*, *12*, 359-388.

Dow, M. M. (2007). Galton's Problem as Multiple Network Autocorrelation Effects: Cultural Trait Transmission and Ecological Constraint. *Cross-Cultural Research*, *41*(4), 336-363.

Dow, M. M. (2008). Network Autocorrelation Regression With Binary and Ordinal Dependent Variables. *Cross-Cultural Research*, *42*(4), 394-419.

Dow, M. M., White, D. R., & Burton, M. L. (1982). Multivariate Modeling with Interdependent Network Data. *Cross-Cultural Research*, *17*(3-4), 216-245.

Duke, J. B. (1993). Estimation of the Network Effects Model in a Large Data Set. *Sociological Methods & Research*, *21*(4), 465-481.

Easterly, W., & Levine, R. (1998). Troubles with the Neighbours: Africa's Problem, Africa's Opportunity. *Journal of African Economies*, *7*(1), 120-142.

Elhorst, J. P. (2010). Applied Spatial Econometrics: Raising the Bar. *Spatial Economic Analysis*, *5*(1), 9-28.

Elhorst, J. P. (2014). *Spatial Econometrics: From Cross-Sectional Data to Spatial Panels.* Heidelberg, Germany: Springer.

Elhorst, J. P., Lacombe, D. J., & Piras, G. (2012). On model specification and parameter space definitions in higher order spatial econometric models. *Regional Science and Urban Economics*, *42*(1-2), 211-220.

Ertur, C., Le Gallo, J., & LeSage, J. P. (2007). Local versus Global Convergence in Europe: A Bayesian Spatial Econometric Approach. *The Review of Regional Studies*, *37*(1), 82-108.

Fernández, C., Ley, E., & Steel, M. F. (2001). Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics*, *16*(5), 563-576.

Fingleton, B. (2001). Theoretical economic geography and spatial econometrics: dynamic perspectives. *Journal of Economic Geography*, *1*(2), 201-225.

Fingleton, B., Igliori, D., & Moore, B. (2005). Cluster Dynamics: New Evidence and Projections for Computing Services in Great Britain. *Journal of Regional Science*, *45*(2), 283-311.

Fingleton, B., & Le Gallo, J. (2008). Estimating spatial models with endogenous variables, a spatial lag and spatially dependent disturbances: Finite sample properties. *Papers in Regional Science*, *87*(3), 319-339.

Florax, R. J., Voortman, R. L., & Brouwer, J. (2002). Spatial dimensions of precision agriculture: a spatial econometric analysis of millet yield on Sahelian coversands. *Agricultural Economics*, *27*(3), 425-443.

Ford, T. C., & Rork, J. C. (2010). Why buy what you can get for free? The effect of foreign direct investment on state patent rates. *Journal of Urban Economics*, *68*(1), 72-81.

Fornango, R. J. (2010). When Space Matters: Spatial Dependence, Diagnostics, and Regression Models. *Journal of Criminal Justice Education*, *21*(2), 117-135.

Fu, F., Chen, X., Liu, L., & Wang, L. (2007). Social dilemmas in an online social network: The structure and evolution of cooperation. *Physics Letters A*, *371*(1-2), 58-64.

Fujimoto, K., Chou, C.-P., & Valente, T. W. (2011). The network autocorrelation model using two-mode data: Affiliation exposure and potential bias in the autocorrelation parameter. *Social Networks*, *33*, 231-243.

Geary, R. C. (1954). The Contiguity Ratio and Statistical Mapping. *The Incorporated Statistician*, *5*(3), 115-145.

Gelfand, A. E., & Smith, A. F. (1990). Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, *85*(410), 398-409.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, *1*(3), 515-533.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis* (Third ed.). Boca Raton, FL: Chapman & Hall/CRC Press.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian Data Analysis* (Second ed.). Boca Raton, FL: Chapman & Hall/CRC Press.

Geman, S., & Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*(6), 721-741.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., & Hothorn, T. (2014). mvtnorm: Multivariate Normal and t Distributions [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=mvtnorm` (R package version 1.0-2)

Geyer, C. J. (1992). Practical Markov Chain Monte Carlo. *Statistical Science*, *7*(4), 473-483.

Gimpel, J. G., & Schuknecht, J. E. (2003). Political participation and the accessibility of the ballot box. *Political Geography*, *22*(5), 471-488.

Glaser, S. (2017). *Modelling of Spatial Effects in Count Data* (Unpublished doctoral dissertation). Universität Hohenheim.

Good, I. (1985). Weight of Evidence: A Brief Survey. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian Statistics 2* (p. 249-269). Amsterdam, The Netherlands: North-Holland.

Gould, R. V. (1991). Multiple Networks and Mobilization in the Paris Commune, 1871. *American Sociological Review*, *56*(6), 716-729.

Greenbaum, R. T. (2002). A spatial study of teachers' salaries in Pennsylvania school districts. *Journal of Labor Research*, *23*(1), 69-86.

Griffith, D. A. (1979). Urban Dominance, Spatial Structure, and Spatial Dynamics: Some Theoretical Conjectures and Empirical Implications. *Economic Geography*, *55*(2), 95-113.

Gu, X., Mulder, J., Deković, M., & Hoijtink, H. (2014). Bayesian evaluation of inequality constrained hypotheses. *Psychological Methods*, *19*(4), 511-527.

Hall, B. C. (2003). *Lie Groups, Lie Algebras, and Representations: An Elementary Introduction*. New York, NY: Springer.

Halleck Vega, S., & Elhorst, J. P. (2015). The SLX model. *Journal of Regional Science*, *55*(3), 339-363.

Han, C., & Carlin, B. P. (2001). Markov Chain Monte Carlo Methods for Computing Bayes Factors. *Journal of the American Statistical Association*, *96*(455), 1122-1132.

Han, X., & Lee, L.-F. (2013). Bayesian estimation and model selection for spatial Durbin error model with finite distributed lags. *Regional Science and Urban Economics*, *43*(5), 816-837.

Handcock, M. S. (2003). Statistical Models for Social Networks: Inference and Degeneracy. In R. Breiger, K. Carley, & P. Pattison (Eds.), *Dynamic Social Network Modelling and Analysis: Workshop Summary and Papers* (p. 229-240). Washington, D.C.: National Academies Press.

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., Krivitsky, P. N., & Morris, M. (2017). ergm: Fit, Simulate and Diagnose Exponential-Family Models for Networks [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=ergm` (R package version 3.8.0)

Hansen, M. H., & Yu, B. (2001). Model Selection and the Principle of Minimum Description Length. *Journal of the American Statistical Association*, *96*(454), 746-774.

Harville, D. A. (1997). *Matrix Algebra From a Statistician's Perspective.* New York, NY: Springer.

Hastings, W. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, *57*(1), 97-109.

Heikkila, E. J., & Kantiotou, C. (1992). Calculating fiscal impacts where spatial effects are present. *Regional Science and Urban Economics*, *22*(3), 475-490.

Hepple, L. W. (1979). Bayesian analysis of the linear model with spatial dependence. In C. P. A. Bartels & R. H. Ketellapper (Eds.), *Exploratory and explanatory statistical analysis of spatial data* (p. 179-199). Dordrecht, The Netherlands: Springer.

Hepple, L. W. (1995a). Bayesian Techniques in Spatial and Network Econometrics: 1. Model Comparison and Posterior Odds. *Environment and Planning A: Economy and Space*, *27*(3), 447-469.

Hepple, L. W. (1995b). Bayesian Techniques in Spatial and Network Econometrics: 2. Computational Methods and Algorithms. *Environment and Planning A: Economy and Space*, *27*(4), 615-644.

Heyndels, B., & Vuchelen, J. (1998). Tax Mimicking Among Belgian Municipalities. *National Tax Journal*, *51*(1), 89-101.

Higham, N. J. (2008). *Functions of Matrices: Theory and Computation*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Holland, P. W., & Leinhardt, S. (1981). An Exponential Family of Probability Distributions for Directed Graphs. *Journal of the American Statistical Association*, *76*(373), 33-50.

Holloway, G., Shankar, B., & Rahman, S. (2002). Bayesian spatial probit estimation: a primer and an application to HYV rice adoption. *Agricultural Economics*, *27*(3), 383-402.

Howard, P. H. (1971). *Political Tendencies in Louisiana*. Baton Rouge, LA: Louisiana State University Press.

Hummel, R. M., Hunter, D. R., & Handcock, M. S. (2012). Improving Simulation-Based Algorithms for Fitting ERGMs. *Journal of Computational and Graphical Statistics*, *21*(4), 920-939.

Hunt, L. M., Boxall, P., Englin, J., & Haider, W. (2005). Remote tourism and forest management: a spatial hedonic analysis. *Ecological Economics*, *53*(1), 101-113.

Hunter, D. R., & Handcock, M. S. (2006). Inference in Curved Exponential Family Models for Networks. *Journal of Computational and Graphical Statistics*, *15*(3), 565-583.

Hunter, D. R., Handcock, M. S., Butts, C. T., Goodreau, S. M., & Morris, M. (2008). ergm: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks. *Journal of Statistical Software*, *24*(3), 1-29.

Hunter, D. R., Krivitsky, P. N., & Schweinberger, M. (2012). Computational Statistical Methods for Social Network Models. *Journal of Computational and Graphical Statistics*, *21*(4), 856-882.

Jeffreys, H. (1961). *Theory of Probability* (Third ed.). Oxford, UK: Oxford University Press.

Johnson, V. E., & Rossell, D. (2010). On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*(2), 143-170.

Joines, J. D., Hertz-Picciotto, I., Carey, T. S., Gesler, W., & Suchindran, C. (2003). A spatial analysis of county-level variation in hospitalization rates for low back problems in North Carolina. *Social Science & Medicine*, *56*(12), 2541-2553.

Kalenkoski, C. M., & Lacombe, D. J. (2008). Effects of Minimum Wages on Youth Employment: the Importance of Accounting for Spatial Correlation. *Journal of Labor Research*, *29*(4), 303-317.

Kalnins, A. (2003). Hamburger Prices and Spatial Econometrics. *Journal of Economics & Management Strategy*, *12*(4), 591-616.

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773-795.

Kass, R. E., & Wasserman, L. (1996). The Selection of Prior Distributions by Formal Rules. *Journal of the American Statistical Association*, *91*(435), 1343-1370.

Kim, J., & Goldsmith, P. (2009). A Spatial Hedonic Approach to Assess the Impact of Swine Production on Residential Property Values. *Environmental and Resource Economics*, *42*(4), 509-534.

Kim, J., & Zhang, M. (2005). Determining Transit's Impact on Seoul Commercial Land Values: An Application of Spatial Econometrics. *International Real Estate Review*, *8*(1), 1-26.

Kirk, D. S., & Papachristos, A. V. (2011). Cultural Mechanisms and the Persistence of Neighborhood Violence. *American Journal of Sociology*, *116*(4), 1190-1233.

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality Constrained Analysis of Variance: A Bayesian Approach. *Psychological Methods*, *10*(4), 477-493.

Kowal, M. S. (2018). Corporate politicking, together: trade association ties, lobbying, and campaign giving. *Business and Politics*, *20*(1), 98-131.

Krivitsky, P. N. (2012). Exponential-family random graph models for valued networks. *Electronic Journal of Statistics*, *6*, 1100-1128.

Krivitsky, P. N., & Butts, C. T. (2017). Exponential-family Random Graph Models for Rank-order Relational Data. *Sociological Methodology*, *47*(1), 68-112.

Kuha, J. (2004). AIC and BIC: Comparisons of Assumptions and Performance. *Sociological Methods & Research*, *33*(2), 188-229.

Lacombe, D. J. (2004). Does Econometric Methodology Matter? An Analysis of Public Policy Using Spatial Econometric Techniques. *Geographical Analysis*, *36*(2), 105-118.

Lambert, D. M., Brown, J. P., & Florax, R. J. (2010). A two-step estimator for a spatial lag model of counts: Theory, small sample performance and an application. *Regional Science and Urban Economics*, *40*(4), 241-252.

Land, K. C., Deane, G., & Blau, J. R. (1991). Religious Pluralism and Church Membership: A Spatial Diffusion Model. *American Sociological Review*, *56*(2), 237-249.

La Rocca, M., Porzio, G. C., Vitale, M. P., & Doreian, P. (2018). Finite Sample Behavior of MLE in Network Autocorrelation Models. In F. Mola, C. Conversano, & M. Vichi (Eds.), *Classification, (Big) Data Analysis and Statistical Learning* (p. 43-50). Cham, Switzerland: Springer.

Lauridsen, J., Maté Sánchez, M., & Bech, M. (2010). Public pharmaceutical expenditure: identification of spatial effects. *Journal of Geographical Systems*, *12*(2), 175-188.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford, UK: Clarendon Press.

Lee, L.-F. (2004). Asymptotic Distributions of Quasi-Maximum Likelihood Estimators for Spatial Autoregressive Models. *Econometrica*, *72*(6), 1899-1925.

Lee, L.-F., & Liu, X. (2010). Efficient GMM estimation of high order spatial autoregressive models with autoregressive disturbances. *Econometric Theory*, *26*(1), 187-230.

Leenders, R. T. (1995). *Structure and Influence: Statistical models for the dynamics of actor attributes, network structure and their interdependence*. Amsterdam, The Netherlands: Thela Thesis.

Leenders, R. T. (2002). Modeling social influence through network autocorrelation: constructing the weight matrix. *Social Networks*, *24*(1), 21-47.

Leifeld, P., Cranmer, S. J., & Desmarais, B. A. (2015). xergm: Extensions for Exponential Random Graph Models [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=xergm` (R package version 1.6)

LeSage, J. P. (1997a). Bayesian Estimation of Spatial Autoregressive Models. *International Regional Science Review*, *20*(1-2), 113-129.

LeSage, J. P. (1997b). Regression analysis of spatial data. *Journal of Regional Analysis and Policy*, *27*(2), 83-94.

LeSage, J. P. (1999). *Spatial Econometrics.* (`http://www.rri.wvu.edu/WebBook/LeSage/spatial/wbook.pdf`)

LeSage, J. P. (2000). Bayesian Estimation of Limited Dependent Variable Spatial Autoregressive Models. *Geographical Analysis*, *32*(1), 19-35.

LeSage, J. P. (2014a). Spatial econometric panel data model specification: A Bayesian approach. *Spatial Statistics*, *9*, 122-145.

LeSage, J. P. (2014b). What Regional Scientists Need to Know About Spatial Econometrics. *The Review of Regional Studies*, *44*(1), 13-32.

LeSage, J. P., & Pace, R. K. (2008). Spatial Econometric Modeling of Origin-Destination Flows. *Journal of Regional Science*, *48*(5), 941-967.

LeSage, J. P., & Pace, R. K. (2009). *Introduction to Spatial Econometrics.* Boca Raton, FL: Chapman & Hall/CRC Press.

LeSage, J. P., & Pace, R. K. (2011). Pitfalls in Higher Order Model Extensions of Basic Spatial Regression Methodology. *The Review of Regional Studies*, *41*(1), 13-26.

LeSage, J. P., & Parent, O. (2007). Bayesian Model Averaging for Spatial Econometric Models. *Geographical Analysis*, *39*(3), 241-267.

Levine, N., Kim, K. E., & Nitz, L. H. (1995). Spatial analysis of Honolulu motor vehicle crashes: II. Zonal generators. *Accident Analysis and Prevention*, *27*(5), 675-685.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of g Priors for Bayesian Variable Selection. *Journal of the American Statistical Association*, *103*(481), 410-423.

Liesenfeld, R., Richard, J.-F., & Vogler, J. (2016). Likelihood Evaluation of High-Dimensional Spatial Latent Gaussian Models with Non-Gaussian Response Variables. In B. H. Baltagi, J. P. LeSage, & R. K. Pace (Eds.), *Spatial Econometrics: Qualitative and Limited Dependent Variables* (p. 35-77). Bingley, UK: Emerald Group Publishing Limited.

Lin, X. (2010). Identifying Peer Effects in Student Academic Achievement by Spatial Autoregressive Models with Group Unobservables. *Journal of Labor Economics*, *28*(4), 825-860.

Link, W. A., & Eaton, M. J. (2012). On thinning of chains in MCMC. *Methods in Ecology and Evolution*, *3*(1), 112-115.

Liu, C. C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, *52*(6), 362-375.

Lu, J., & Zhang, L. (2011). Modeling and Prediction of Tree Height–Diameter Relationships Using Spatial Autoregressive Models. *Forest Science*, *57*(3), 252-264.

Lynch, S. M. (2007). *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*. New York, NY: Springer.

Marsden, P. V., & Friedkin, N. E. (1993). Network Studies of Social Influence. *Sociological Methods & Research*, *22*(1), 127-151.

Martinetti, D., & Geniaux, G. (2016). ProbitSpatial: Probit with Spatial Dependence, SAR and SEM Models [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=ProbitSpatial` (R package version 1.0)

Mathai, A. M., & Provost, S. B. (1992). *Quadratic Forms in Random Variables: Theory and Applications*. New York, NY: Marcel Dekker.

McMillen, D. P. (1992). Probit with Spatial Autocorrelation. *Journal of Regional Science*, *32*(3), 335-348.

McMillen, D. P. (2010). Issues in Spatial Data Analysis. *Journal of Regional Science*, *50*(1), 119-141.

McMillen, D. P. (2013). McSpatial: Nonparametric spatial data analysis [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=McSpatial` (R package version 2.0)

McMillen, D. P., Singell Jr., L. D., & Waddell, G. R. (2007). Spatial Competition and the Price of College. *Economic Inquiry*, *45*(4), 817-833.

McPherson, M. A., & Nieswiadomy, M. L. (2005). Environmental Kuznets curve: threatened species and spatial effects. *Ecological Economics*, *55*(3), 395-407.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equations of State Calculations by Fast Computing Machines. *Journal of Chemical Physics*, *21*(6), 1087-1092.

Mizruchi, M. S., & Neuman, E. J. (2008). The effect of density on the level of bias in the network autocorrelation model. *Social Networks*, *30*(3), 190-200.

Mizruchi, M. S., & Stearns, L. B. (2006). The Conditional Nature of Embeddedness: A Study of Borrowing by Large U.S. Firms, 1973-1994. *American Sociological Review*, *71*(2), 310-333.

Moreno, R., & Trehan, B. (1997). Location and the Growth of Nations. *Journal of Economic Growth*, *2*(4), 399-418.

Morenoff, J. D. (2003). Neighborhood Mechanisms and the Spatial Dynamics of Birth Weight. *American Journal of Sociology*, *108*(5), 976-1017.

Morenoff, J. D., Sampson, R. J., & Raudenbusch, S. W. (2001). Neighborhood Inequality, Collective Efficacy, and the Spatial Dynamics of Urban Violence. *Criminology*, *39*(3), 517-558.

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, *16*(4), 406-419.

Mulder, J. (2014a). Bayes factors for testing inequality constrained hypotheses: Issues with prior specification. *British Journal of Mathematical and Statistical Psychology*, *67*(1), 153-171.

Mulder, J. (2014b). Prior adjusted default Bayes factors for testing (in)equality constrained hypotheses. *Computational Statistics and Data Analysis*, *71*, 448-463.

Mulder, J. (2016). Bayes factors for testing order-constrained hypotheses on correlations. *Journal of Mathematical Psychology*, *72*, 104-115.

Mulder, J., Klugkist, I., van de Schoot, R., Meeus, W. H., Selfhout, M., & Hoijtink, H. (2009). Bayesian model selection of informative hypotheses for repeated measurements. *Journal of Mathematical Psychology*, *53*(6), 530-546.

Mulder, J., & Wagenmakers, E.-J. (2016). Editors' introduction to the special issue "Bayes factors for testing hypotheses in psychological research: Practical relevance and new developments". *Journal of Mathematical Psychology*, *72*, 1-5.

Mur, J., López, F., & Angulo, A. (2008). Symptoms of Instability in Models of Spatial Dependence. *Geographical Analysis*, *40*(2), 189-211.

Murray, I., Ghahramani, Z., & MacKay, D. J. C. (2006). MCMC for doubly-intractable distributions. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence* (p. 359-366). Arlington, VA: AUAI Press.

Myneni, S., Fujimoto, K., Cobb, N., & Cohen, T. (2015). Content-Driven Analysis of an Online Community for Smoking Cessation: Integration of Qualitative Techniques, Automated Text Analysis, and Affiliation Networks. *American Journal of Public Health*, *105*(6), 1206-1212.

Neuman, E. J., & Mizruchi, M. S. (2010). Structure and bias in the network autocorrelation model. *Social Networks*, *32*(4), 290-300.

Newman, M. E. J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(2), 404-409.

Niebuhr, A. (2010). Migration and innovation: Does cultural diversity matter for regional R&D activity? *Papers in Regional Science*, *89*(3), 563-585.

Nocedal, J., & Wright, S. (2006). *Numerical Optimization* (Second ed.). New York, NY: Springer.

O'Hagan, A. (1995). Fractional Bayes Factors for Model Comparison. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *57*(1), 99-138.

Ord, K. (1975). Estimation Methods for Models of Spatial Interaction. *Journal of the American Statistical Association*, *70*(349), 120-126.

Osland, L. (2010). An Application of Spatial Econometrics in Relation to Hedonic House Price Modeling. *Journal of Real Estate Research*, *32*(3), 289-320.

Owen, A., & Zhou, Y. (2000). Safe and Effective Importance Sampling. *Journal of the American Statistical Association*, *95*(449), 135-143.

Owen, A. B. (2017). Statistically Efficient Thinning of a Markov Chain Sampler. *Journal of Computational and Graphical Statistics*, *26*(3), 738-744.

Pace, R. K., & Barry, R. (1997). Quick Computation of Spatial Autoregressive Estimators. *Geographical Analysis*, *29*(3), 232-247.

Patton, M., & McErlean, S. (2003). Spatial Effects within the Agricultural Land Market in Northern Ireland. *Journal of Agricultural Economics*, *54*(1), 35-54.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, *50*(302), 157-175.

Pericchi, L., & Pereira, C. (2016). Adaptative significance levels using optimal decision rules: Balancing by weighting the error probabilities. *Brazilian Journal of Probability and Statistics*, *30*(1), 70-90.

Pisati, M. (2001). Tools for spatial data analysis. *Stata Technical Bulletin*, *60*, 21-37.

Plümper, T., & Neumayer, E. (2010). Model specification in the analysis of spatial dependence. *European Journal of Political Research*, *49*(3), 418-442.

Pons-Novell, J., & Viladecans-Marsal, E. (1999). Kaldor's Laws and Spatial Dependence: Evidence for the European Regions. *Regional Studies*, *33*(5), 443-451.

Quddus, M. A. (2008). Modelling area-wide count outcomes with spatial correlation and heterogeneity: An analysis of London crash data. *Accident Analysis and Prevention*, *40*(4), 1486-1497.

R Core Team. (2017). R: A Language and Environment for Statistical Computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Raftery, A. E. (1995). Bayesian Model Selection in Social Research. *Sociological Methodology*, *25*, 111-163.

Raftery, A. E., & Lewis, S. M. (1996). Implementing MCMC. In W. Gilks, S. Richardson, & D. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice* (p. 115-130). Boca Raton, FL: Chapman & Hall/CRC Press.

Raftery, A. E., Madigan, D., & Hoeting, J. A. (1997). Bayesian Model Averaging for Linear Regression Models. *Journal of the American Statistical Association*, *92*(437), 179-191.

Revelli, F. (2003). Reaction or interaction? Spatial process identification in multi-tired government structures. *Journal of Urban Economics*, *53*(1), 29-53.

Rey, S. J., & Anselin, L. (2007). PySAL: A Python Library of Spatial Analytical Methods. *The Review of Regional Studies*, *37*(1), 5-27.

Robert, C. (2001). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. New York, NY: Springer.

Robins, G., Pattison, P., Kalish, Y., & Lusher, D. (2007). An introduction to exponential random graph (p*) models for social networks. *Social Networks*, *29*(2), 173-191.

Robins, G., Snijders, T., Wang, P., Handcock, M., & Pattison, P. (2007). Recent developments in exponential random graph (p*) models for social networks. *Social Networks*, *29*(2), 192-215.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225-237.

Rubio, F. J., & Steel, M. F. (2014). Inference in Two-Piece Location-Scale Models with Jeffreys Priors. *Bayesian Analysis*, *9*(1), 1-22.

Ruggles, S. (2007). The Decline of Intergenerational Coresidence in the United States, 1850 to 2000. *American Sociological Review*, *72*(6), 964-989.

Rupasingha, A., Goetz, S. J., & Freshwater, D. (2002). Social and institutional factors as determinants of economic growth: Evidence from the United States counties. *Papers in Regional Science*, *81*(2), 139-155.

Saavedra, L. A. (2000). A Model of Welfare Competition with Evidence from AFDC. *Journal of Urban Economics*, *47*(2), 248-279.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2), 461-464.

Seldadyo, H., Elhorst, J. P., & De Haan, J. (2010). Geography and governance: Does space matter? *Papers in Regional Science*, *89*(3), 625-640.

Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of p Values for Testing Precise Null Hypotheses. *The American Statistician*, *55*(1), 62-71.

Shaffer, J. P. (1995). Multiple Hypothesis Testing. *Annual Review of Psychology*, *46*, 561-584.

Shin, M., & Ward, M. D. (1999). Lost in Space: Political Geography and the Defense-Growth Trade-Off. *The Journal of Conflict Resolution*, *43*(6), 793-817.

Sinharay, S., & Stern, H. S. (2002). On the Sensitivity of Bayes Factors to the Prior Distributions. *The American Statistician*, *56*(3), 196-201.

Smith, T. E. (2009). Estimation Bias in Spatial Models with Strongly Connected Weight Matrices. *Geographical Analysis*, *41*(3), 307-332.

Snijders, T. A. B. (2002). Markov Chain Monte Carlo Estimation of Exponential Random Graph Models. *Journal of Social Structure*, *3*(2).

Stewart, W. J. (2009). *Probability, Markov Chains, Queues, and Simulation: The Mathematical Basis of Performance Modeling.* Princeton, NJ: Princeton University Press.

Strauss, D. (1986). On a General Class of Models for Interaction. *SIAM Review*, *28*(4), 513-527.

Tam Cho, W. K. (2003). Contagion Effects and Ethnic Contribution Networks. *American Journal of Political Science*, *47*(2), 368-387.

Tita, G. E., & Greenbaum, R. T. (2009). Crime, Neighborhoods, and Units of Analysis: Putting Space in Its Place. In D. Weisburd, W. Bernasco, & G. J. Bruinsma (Eds.), *Putting Crime in its Place* (p. 145-170). New York, NY: Springer.

Tita, G. E., & Radil, S. M. (2011). Spatializing the Social Networks of Gangs to Explore Patterns of Violence. *Journal of Quantitative Criminology*, *27*(4), 521-545.

Trautmann, H., Steuer, D., Mersmann, O., & Bornkamp, B. (2015). truncnorm: Truncated normal distribution [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=truncnorm` (R package version 1.0-7)

van de Schoot, R., Mulder, J., Hoijtink, H., van Aken, M. A., Dubas, J. S., Orobio de Castro, B., ... Romeijn, J.-W. (2011). An introduction to Bayesian model selection for evaluating informative hypotheses. *European Journal of Developmental Psychology*, *8*(6), 713-729.

van Duijn, M. A. J., Gile, K. J., & Handcock, M. S. (2009). A framework for the comparison of maximum pseudo-likelihood and maximum likelihood estimation of exponential family random graph models. *Social Networks*, *31*(1), 52-62.

Varga, A. (2000). Local Academic Knowledge Transfers and the Concentration of Economic Activity. *Journal of Regional Science*, *40*(2), 289-309.

Vitale, M. P., Porzio, G. C., & Doreian, P. (2016). Examining the effect of social influence on student performance through network autocorrelation models. *Journal of Applied Statistics*, *43*(1), 115-127.

Voss, P. R., & Chi, G. (2006). Highways and Population Change. *Rural Sociology*, *71*(1), 33-58.

Voss, P. R., Long, D. D., Hammer, R. B., & Friedman, S. (2006). County child poverty rates in the US: a spatial regression approach. *Population Research and Policy Review*, *25*(4), 369-391.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779-804.

Wang, W., Neuman, E. J., & Newman, D. A. (2014). Statistical power of the social network autocorrelation model. *Social Networks*, *38*, 88-99.

Wang, Y., Kockelman, K. M., & Damien, P. (2014). A spatial autoregressive multinomial probit model for anticipating land-use change in Austin, Texas. *The Annals of Regional Science*, *52*(1), 251-278.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world networks'. *Nature*, *393*, 440-442.

Weakliem, D. L. (2004). Introduction to the Special Issue on Model Selection. *Sociological Methods & Research*, *33*(2), 167-187.

Wetzels, R., & Wagenmakers, E.-J. (2012). A default Bayesian hypothesis test for correlations and partial correlations. *Psychonomic Bulletin & Review*, *19*(6), 1057-1064.

White, D. R., Burton, M. L., & Dow, M. M. (1981). Sexual Division of Labor in African Agriculture: A Network Autocorrelation Analysis. *American Anthropologist*, *83*(4), 824-849.

Whitt, H. P. (2010). The Civilizing Process and Its Discontents: Suicide and Crimes against Persons in France, 1825–1830. *American Journal of Sociology*, *116*(1), 130-186.

Wilhelm, S., & Godinho de Matos, M. (2015). spatialprobit: Spatial Probit Models [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=spatialprobit` (R package version 0.9-11)

Wilhelm, S., & Manjunath, B. (2015). tmvtnorm: Truncated Multivariate Normal and Student t Distribution [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=tmvtnorm` (R package version 1.4-10)

Wilhelmsson, M. (2002). Spatial Models in Real Estate Economics. *Housing, Theory and Society*, *19*(2), 92-101.

Won Kim, C., Phipps, T. T., & Anselin, L. (2003). Measuring the benefits of air quality improvement: a spatial hedonic approach. *Journal of Environmental Economics and Management*, *45*(1), 24-39.

Yang, N., McCluskey, J. J., & Brady, M. P. (2012). The Value of Good Neighbors: A Spatial Analysis of the California and Washington State Wine Industries. *Land Economics*, *88*(4), 674-684.

Yang, X. (2000). A Matrix Trace Inequality. *Journal of Mathematical Analysis and Applications*, *250*(1), 372-374.

Yang, Z. (2015). A general method for third-order bias and variance corrections on a nonlinear estimator. *Journal of Econometrics*, *186*(1), 178-200.

Yu, D., Bai, P., & Ding, C. (2015). Adjusted quasi-maximum likelihood estimator for mixed regressive, spatial autoregressive model and its small sample bias. *Computational Statistics and Data Analysis*, *87*, 116-135.

Zellner, A. (1986). On Assessing Prior Distributions and Bayesian Regression Analysis with g-Prior Distributions. In P. K. Goel & A. Zellner (Eds.), *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti* (p. 233-243). Amsterdam, The Netherlands: North-Holland.

Zhang, B., Thomas, A. C., Doreian, P., Krackhardt, D., & Krishnan, R. (2013). Contrasting Multiple Social Network Autocorrelations for Binary Outcomes, With Applications To Technology Adoption. *ACM Transactions on Management Information Systems*, *3*(4), 18:1–18:21.

Acknowledgments

I would like to take the opportunity and thank several people who shaped the past four years and contributed to this thesis in different ways.

First, I would like to thank my promotors and copromotor. Joris, thank you for your continuous academic guidance, the confidence shown in me, and both the attention and freedom you gave me. I am happy that our relationship grew beyond the office over the years, helping me not only develop as a researcher but also as a person. Roger, thank you for bringing up football data in the research project description that probably gave way to my time in Tilburg to start with. The football data are still to be analyzed by the three of us but you steadily encouraged me to also take the perspective of a more applied researcher from time to time, which greatly improved the comprehensibility of my writing. Jeroen, your door was always open and I would like to thank you for supporting my personal development by facilitating teaching opportunities, summer school visits, and research stays.

In fall 2017, I had the chance to visit Professor George Tita and Professor Carter Butts at UC Irvine. George, thank you for inviting me to Irvine without having met me in person, for introducing me to dozens of your colleagues and asking every single one of them to share empirical data with us, and for showing me Irvine's finest gourmet places. Carter, thank you for the warm welcome, for allowing me to join your lab group, and for the many stimulating and fruitful discussions. Mario, thank you for hosting me in Irvine and the fun times in Southern California and beyond. I hope that our friendship will continue to last across seas and continents in spite of near apartment burn downs, fish poisonings, and late-night officer chats.

Furthermore, I would like to thank the members of the Bayes lab group for taking their time to read and give feedback on several draft papers over the years. It was always motivating to present and exchange thoughts with you and receive well-intentioned reviews on early stage research. I am especially grateful to Davide, Florian, Geert, Jesper, Kyle, Lianne, Reza, Sara, and Xynthia.

To all of my other (former) colleagues at the department: thank you for being such an easy-going group of guys and girls, creating an atmosphere in which we saw each other much more as friends rather than competitors. To Mattis, thank you for being a great office mate and a great friend, being there in times of need, nearby or far away. To Jaap Joris, thank you for the enthusiasm and the fresh ideas you brought to our office in the past two years as well as the always entertaining out-of-office hours spent together. To Erwin, thank you for sharing your thesis template and thereby making this thesis look somewhat tidy. To Luc, thank you for boosting my teaching self-confidence and for nudging me into

becoming a supporter of Willem II. To Paulette, thank you for answering my countless questions before moving to Tilburg and for being so hospitable. To Reza, thank you for radiating so much positive energy every time we see each other, you are an inspiration. To Robbie, thank you for all the sports talk and action, I hope to join you in more data-driven reasoning soon. To Ruslan paşa, thank you for bringing back music making to my life, I cannot overstate how much this means to me. To Zsuzsa, thank you for motivating me to keep exploring the world and trying out something new. To Anne-Marie, Liesbeth, and Marieke, thank you for all the patience, kindness, and help when I was bogged down in paper work.

Finding joy beyond work has always been very important to me, and I cherish the friendships I made in Tilburg outside the office that also gave inspiration for times spent in the office. Alem, I think we can both safely say "it is a small world", and thank you for being an anchor for me in Tilburg. Diogo, you became my first friend in Tilburg and I miss our dinners, spontaneous trips, and listening to your guitar playing. Gijs, I admire the dedication you have, which made me look forward to our practices for days. Roland, I still have to figure out if and when you sleep but that left more time to have fun on and off the tennis court. Dear Yugo band, dear Asmir, Dušan, Jovana, Marko, Nataša, Nebojša, and Stevan, thank you for the Wednesday lunches, for the nights out, for the trips together, and for all the laughs, you rock. I would also like to thank Christina, Francesca, and Gaby for your friendship.

To my ski ninjas, Mario, Moritz, Nora, Ozren, and Peter B., thank you for brightening up winters in the lowlands and for not being a complete débutant in powder anymore. To Akan, Antonio, Feli, Isai, Jason, Juan, Michael S., Miles, Olaf, and Pedro, thank you for the good times in Irvine. To Annalisa, thank you for helping me grow as a person in so many ways. To my "old" Munich friends, Alex D., Alex G., Beni, Đorđe, Fabio, Matthias, Michael C., Peter L., Robert, and Verena, thank you for the nights out back home, for the spontaneous visits, and for the many memorable travels. You always made me return in high spirits from shorter or longer home office stays.

Nataša and Barend, I am very happy that you are holding my back on defense day. Although we stayed at different departments and barely understood each other's research, we shared the ups and downs of academic life and the two of you made me feel comfortable to reach out at any time for whatever reason there was. Nataša, thank you for being the good and caring person you are. Barend, thank you for the always active and fun times spent together, from A(rmbrustschützenzelt) to Z(aventem).

Finally, I wish to express my gratitude to my family. To my sister, Irma, thank you for always believing in my academic skills and especially for your support throughout the last few months of writing. To my parents, thank you for putting things into perspective from time to time and for always putting my well-being first. I am grateful for the opportunities you created for me and my sister from early childhood on, which eventually allowed me to be able to pursue a Ph.D. The three of you mean everything to me.