Title

Flexible Bayesian Methods for Inference in Psychological Science

Author

Rodriguez, Josue E.

Publication Date

2023

Peer reviewed|Thesis/dissertation

Flexible Bayesian Methods for Inference in Psychological Science

By

JOSUE E. RODRIGUEZ
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Psychology

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

PHILIPPE RAST, Chair

MIJKE RHEMTULLA

EMILIO FERRER

Committee in Charge

2023

Para mí madre y mí padre, quienes abandonaron sus vidas para que sus hijos tuvieran una mejor. Sí se pudo. To my sister, who I want to feel proud when she sees her older brother. To Shea, I strove to be the person you think I am — I hope this helps.

Contents

Flexible Bayesian Methods for Inference in Psychological Science

Abstract

This dissertation is a collection of three papers I wrote during my time in graduate school. Each proposes a novel way in which a Bayesian statistical technique may be applied or conceptualized for the purpose of better aligning statistical hypotheses and research aims, or improving upon the status quo with respect to the application of statistical methods in psychological science. On a personal note, these articles represent endeavors that pushed my intellectual limits and challenged my grit and mettle. The articles are presented as chapters, and in the chronological order in which they were written. I hope this reflects my thought process and growth throughout my time in graduate school. The first chapter presents a framework for integrating exploratory and confirmatory analyses in psychological network research. It is argued that while network analysis has been proposed as a tool for hypothesis generation, there is untapped potential for confirmatory hypothesis testing. We suggest using Bayesian Gaussian graphical models to first generate and then test ordered hypotheses based on the conditional (in)dependence structure of psychological networks. The second chapter proposes the use of the Bayesian bootstrap method to estimate various correlation coefficients commonly used in the social-behavioral sciences. We demonstrate how the Bayesian bootstrap can be used to estimate Pearson's, Spearman's, Gaussian rank, Kendall's τ, and polychoric correlations. We also describe a method for comparing correlations and evaluating null associations among the estimated correlations. Finally, in an effort to provide a more nuanced understanding of individual differences than standard approaches, the third chapter explores the spike-and-slab prior distribution for random effect selection in mixed-effects models. Simulation studies were conducted to evaluate the spike-and-slab prior in accurately distinguishing "average" and "non-average" individuals. The results highlight the spike-and-slab prior's ability to identify individual differences, even in situations with low between-person variance. This dissertation concludes by offering some discussion on why Bayesian analyses are more flexible than standard approaches, and how this flexibility can lead to higher-quality inferences in psychological science.

This dissertation is a collection of three papers I wrote during my time in graduate school. Each proposes a novel way in which a Bayesian statistical technique may be applied or conceptualized

for the purpose of better aligning statistical hypotheses and research aims, or improving upon the status quo with respect to the application of statistical methods in psychological science. On a personal note, these articles represent endeavors that pushed my intellectual limits and challenged my grit and mettle. The articles are presented as chapters, and in the chronological order in which they were written. I hope this reflects my thought process and growth throughout my time in graduate school. The first chapter presents a framework for integrating exploratory and confirmatory analyses in psychological network research. It is argued that while network analysis has been proposed as a tool for hypothesis generation, there is untapped potential for confirmatory hypothesis testing. We suggest using Bayesian Gaussian graphical models to first generate and then test ordered hypotheses based on the conditional (in)dependence structure of psychological networks. The second chapter proposes the use of the Bayesian bootstrap method to estimate various correlation coefficients commonly used in the social-behavioral sciences. We demonstrate how the Bayesian bootstrap can be used to estimate Pearson's, Spearman's, Gaussian rank, Kendall's τ, and polychoric correlations. We also describe a method for comparing correlations and evaluating null associations among the estimated correlations. Finally, in an effort to provide a more nuanced understanding of individual differences than standard approaches, the third chapter explores the spike-and-slab prior distribution for random effect selection in mixed-effects models. Simulation studies were conducted to evaluate the spike-and-slab prior in accurately distinguishing "average" and "non-average" individuals. The results highlight the spike-and-slab prior's ability to identify individual differences, even in situations with low between-person variance. This dissertation concludes by offering some discussion on why Bayesian analyses are more flexible than standard approaches, and how this flexibility can lead to higher-quality inferences in psychological science.

Acknowledgments

A huge thanks to Philippe Rast for taking me on as a graduate student and allowing me the space to freely pursue my interests. Your advice has been invaluable, in academia and in life. I'd like to thank Emilio Ferrer for instilling in me the courage in my voice when I speak, and Mijke Rhemtulla who was a constant source of inspiration to become a better academic.

In particular, I'd like to thank Donny Williams who spent countless hours simultaneously playing the roles of coach and closest friend.

Of course, I'd like to thank all of the friends I made along this journey, without whom I may have never unglued my eyes from a computer. Beverly, Lee, Anna, Madison, Toby, Simran, Diego, and MJ — thank you.

CHAPTER 1

On Formalizing Expectations: Bayesian Testing of Central Structures in Psychological Networks

1.1. Introduction

Network theory has emerged as a popular framework in the social-behavioral sciences for analyzing psychological constructs (Cramer et al., 2012; Dalege et al., 2019; Epskamp, Maris, et al., 2018; McNally, 2016). The underlying rationale is that a group of observed variables, say, self-reported symptoms, form a dynamic system wherein they mutually influence and interact with one another (Borsboom, 2017; McNally et al., 2015). In networks, observed variables are called "nodes" and the featured connections between them are called "edges". This work will focus on psychological networks in which the edges are undirected and represent conditional dependence between nodes representing symptoms of mental disorders, that is, pairwise relations between symptoms after controlling for all other symptoms. This approach has led to powerful new insights into a range of mental disorders including obsessive compulsive disorder (OCD; McNally et al., 2017), depression (Boschloo et al., 2016; Fried et al., 2016; Hoorelbeke et al., 2016), anxiety (Beard et al., 2016), and posttraumatic stress disorder (PTSD; Afzali et al., 2017; Armour et al., 2017; Fried et al., 2018; McNally et al., 2015).

This surge of research stems from a shift away from the "common cause" perspective to the "network" perspective of mental disorders (Cramer et al., 2010b; McNally, 2016). The key distinction lies in the assumptions of their respective statistical models. The latter uses network models that account for the mutual interactions between psychopathological symptoms (Borsboom, 2017; Borsboom & Cramer, 2013), whereas the former uses latent variable models that fail to capture mutual relationships between symptoms due to the assumption of "local independence" (Cramer et al., 2010a, but see Bringmann and Eronen, 2018). There are also notable differences relating to their perceived purpose. For example, undirected networks are customarily estimated with a

data-driven approach thought to be ideal for hypothesis generation (Epskamp & Fried, 2018; Epskamp, van Borkulo, et al., 2018). On the other hand, latent variable models have a long tradition of confirmatory hypothesis testing (e.g., Bentler, 1980). Although this distinction is commonplace, network modeling has untapped potential for confirmatory testing of conditional (in)dependencies.

Confirmatory testing with networks remains uncommon in part because edges are often thought to *merely* represent a causal skeleton (Borsboom, 2017; Borsboom & Cramer, 2013; Epskamp, van Borkulo, et al., 2018). Typically causality is associated with directionality, that is, say, $A \rightarrow B$, which implies that A causes B (e.g., Pearl, 2009). Estimating such a graph would require abandoning a partial correlation network. This is because the relations are inherently *undirected*. Hence, the notion of using networks to generate *causal* hypotheses perhaps implies that an alternative model is needed for confirmatory testing. This is not the case. Networks are an effective method to study pairwise relationships and can be used for confirmatory hypothesis testing (Epskamp et al., 2017; Ryan et al., 2019). In fact, approaches for estimating directed graphs of conditional dependencies (e.g., DAGs) are also inherently data-driven (e.g., Kalisch & Bühlmann, 2007). This is distinct from confirmatory testing, where the focus is on *a priori* expectations that allow for rich inference. Testing expectations is analogous to predicting the observed data — an important signature of a theory's explanatory power.

A prerequisite for using networks in a confirmatory setting is having hypotheses to test. Here there is also untapped potential. For example, although hypothesis generation is commonly proposed as an advantage of network models (Cramer et al., 2012; Epskamp & Fried, 2018; Epskamp, Waldorp, et al., 2018; Ryan et al., 2019), we are not aware of any examples in psychology that have actually formulated hypotheses to then test. The unrealized idea is to use a network estimated in an exploratory setting to generate hypotheses regarding, say, which nodes are most central in a system (Epskamp, van Borkulo, et al., 2018; Jones et al., 2019; Robinaugh et al., 2016). The main contribution of this work is bringing to fruition the idea of using networks to generate hypotheses for testing in a confirmatory setting.

In this work, we focus on testing hypotheses related to central nodes and the conditional (in)dependence structure therein. In psychopathology, special attention has been drawn to central nodes due to the idea that intervening on them would affect the rest of the network (Beard et al.,

2016; McNally et al., 2015; Robinaugh et al., 2016). This idea implies the notion of causality, but in fact, centrality measures have been critiqued as poor indicators of causal influence (Dablander & Hinne, 2019). However, because centrality scores are summary statistics that describe an exploratory analysis they can be used to formulate confirmatory hypotheses. In particular, strength-based metrics (Jones et al., 2019; Newman, 2010) are useful for developing hypotheses related to the edge weights of central nodes. This is perhaps unsurprising, given that they can be calculated using a population parameter with a known distribution (e.g., a partial correlation, Fisher, 1924; Yule, 1897). Together, centrality indices provide untapped sources of information that can be used to narrow the focus on to particular aspects of an estimated network.

To test hypotheses related to the edge weights, we use recently proposed Bayesian methodology that readily allows for exploratory and confirmatory testing in partial correlation networks, or Gaussian graphical models (GGMs; Williams & Mulder, 2020). This approach facilitates a workflow wherein central nodes can be identified in an exploratory stage and hypotheses related to these nodes can then be tested in a confirmatory setting. A particular advantage of the confirmatory aspect is that hypotheses are expressed using (in)equality constraints on the parameters of interest and tested against competing theoretical expectations. For instance, one could test $\mathcal{H}_1 : \rho_{12} > \rho_{13} > \rho_{14} > 0$ against $\mathcal{H}_2 : \rho_{12} = \rho_{13} = \rho_{14} = 0$ (Hoijtink, 2001; Hoijtink et al., 2019; Mulder, 2016). This provides a formal comparison between $\mathcal{H}_1$, which states that there is an order to the edge weights, or effect sizes, and they are all positive, versus $\mathcal{H}_2$, which expresses that they are all equal to zero. A major contribution of this work is to extend the idea of informative Bayesian testing to psychological networks, in addition to providing a comprehensive framework that can propel the field toward developing formal models (e.g. Borsboom et al., 2020; Haslbeck et al., 2019).

This work is organized as follows. We first give a concise overview of Gaussian graphical models. We then proceed to illustrate how hypotheses can be derived based on an exploratory network analysis. Here we show that the information encoded by the conditional (in)dependence structures can be used to formalize theoretical expectations. Next we provide an overview of the confirmatory strategy in this work, where the advantage of adopting a Bayesian approach for confirmatory testing is made clear. Specifically, the ability to directly compare theoretical models formulated through an exploratory analysis. We then discuss in detail how the proposed testing framework can be used

in an applied setting. We conclude with a discussion on the proposed methods including limitations and recommendations.

1.2. The Gaussian Graphical Model

The Gaussian graphical model (GGM) encapsulates conditional relations among multivariate normal data. These relations are customarily visualized to infer the underlying dependence structure (i.e., the partial correlation "network"; Højsgaard et al., 2012; Lauritzen, 1996). A GGM is an undirected graph that can be denoted by $G = (V, E)$, where $V = \{1, \ldots, p\}$ is a vertex set and $E \subseteq V \times V$ is an edge set. V refers to the p "nodes"' in the network, say, items on a depression scale, and E defines the estimated network structure. Let $\mathbf{y} = (y_1, ..., y_p)^\top$ be a vector of observed random variables that index the vertices in G, and assume it to be multivariate normal, $\mathbf{y} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Here $\boldsymbol{\mu}$ is a $p \times 1$ mean vector and $\boldsymbol{\Sigma}$ is a $p \times p$ positive definite covariance matrix.

Throughout the rest of this paper we will use $\mathbf{Y}$ to denote the $n \times p$ data matrix, where each row corresponds to observations from an individuals. Without loss of generality, we assume the data to be mean centered, that is, $\boldsymbol{\mu} = \mathbf{0}$. The undirected graph G is obtained by establishing the non-zero off-diagonal elements in the precision matrix, $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$. That is, $(i, j) \in E$ when nodes i and j are determined to be conditionally dependent and set to zero otherwise. The edges in a GGM correspond to partial correlations, ρ_{ij}, that is, the correlation between variables i and j after controlling for all other variables. These can be computed directly from the elements in $\boldsymbol{\Theta}$,

1.2.1. Formalizing Theoretical Models. Psychological theories can be expressed as hypotheses with constraints on the parameters of interest (Hoijtink, 2011). This translates into thinking of theories in terms of constraints among conditional (in)dependencies. In a GGM, for example, it may be expected that a set of partial correlations are approximately equal to each other, larger or smaller than another set of partial correlations, or larger or smaller than a constant (typically zero). These kinds of hypotheses can be derived from theory or an exploratory analysis. This work focuses on the latter. Here the goal is not to determine the graph (e.g., McNally et al., 2017), but rather the structure of interrelations among partial correlations.

A major hurdle to confirmatory testing in networks has been the lack of available methods. Recently, however, a Bayes factor approach was introduced specifically for this purpose (Williams

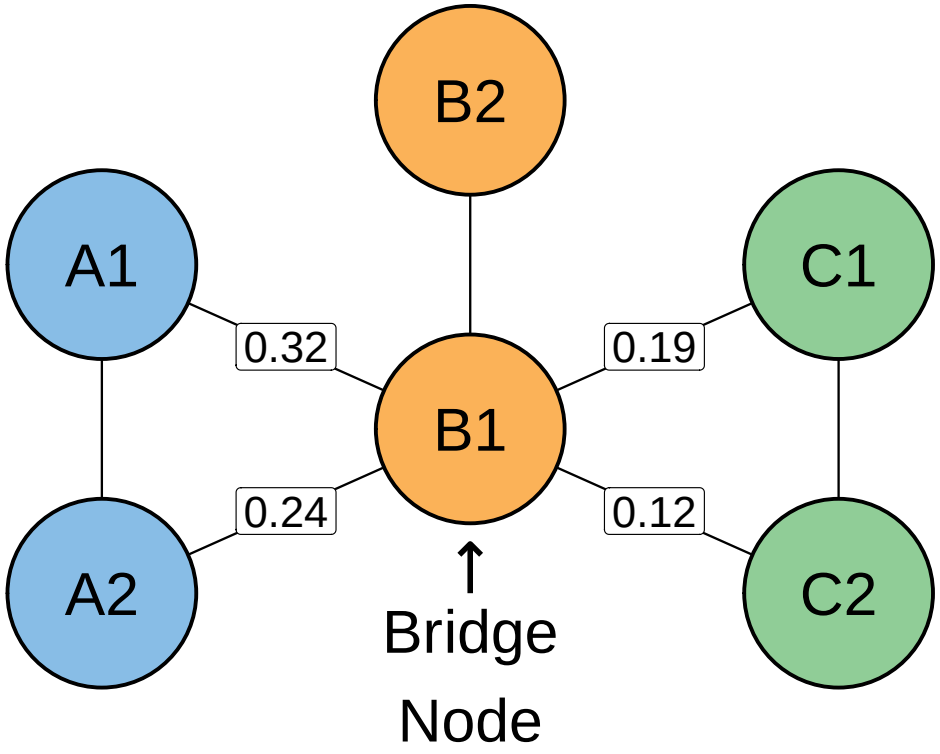

FIGURE 1.1. Example network with three communities: A, B, and C. Lines between two nodes indicates conditional dependence. It can be seen that B1 is a central node which bridges community B to communities A and C. Hypotheses can then be gleaned from the graph. For example, $\mathcal{H}_1 : \rho_{B1-A1} > \rho_{B1-A2} > \rho_{B1-C1} > \rho_{B1-C2} > 0$ or $\mathcal{H}_2 : (\rho_{B1-A1}, \rho_{B1-A2}, \rho_{B1-C1}, \rho_{B1-C2}) = 0$. The former tests the order of bridge edges and constrains them to be positive. The latter tests whether B1 is conditionally independent of nodes outside the B community. This captures how network structures encode information that can be used to formalize and test a model.

& Mulder, 2020; Williams, Rast, et al., 2020). This opens the door for testing hypotheses not currently possible with classical statistics (i.e., p-values). For instance, there is theoretical interest in characterizing central structures involving bridge nodes, or nodes that connect to multiple communities (i.e., clusters) within a network (c.f., Castro et al., 2019; Cramer et al., 2010a, 2010b; Jones et al., 2019). These nodes can be identified through visualizing a network (e.g., Beard et al.,

2016) or bridge centrality metrics (Jones et al., 2019). For example, by inspecting Figure 1.1, it is possible to formulate hypotheses relating to the order of edges or effect sizes within (or between) clusters, that is,

$$
\begin{aligned}
&\mathcal{H}_1 : \rho_{B1-A1} > \rho_{B1-A2} > \rho_{B1-C1} > \rho_{B1-C2} > 0 \\
&\mathcal{H}_2 : (\rho_{B1-A1}, \rho_{B1-A2}, \rho_{B1-C1}, \rho_{B1-C2}) = 0 \\
&\mathcal{H}_3 : \text{“not } \mathcal{H}_1 \text{ or } \mathcal{H}_2\text{.”}
\end{aligned} \tag{1.1}
$$

In (1.1), $\mathcal{H}_1$ captures the order of edges in Figure 1.1. Substantively, this hypothesis can be interpreted as capturing the order of importance (defined by effect size) for the bridging relations that connect clusters of nodes in a network. Furthermore, there is an additional constraint that all of the edges are positive. This reflects the expectation of a positive manifold that has a central role in network theory (Borsboom et al., 2011). $\mathcal{H}_2$ then tests whether all the nodes are actually conditionally independent, which also implies that there is no inherent ordering. Finally, $\mathcal{H}_3$ captures some yet to be hypothesized structure of relations. These hypotheses are formal models that can be evaluated. That is, one can directly quantify support for $\mathcal{H}_1$ versus $\mathcal{H}_2$ with a Bayes factor, a measure of relative support between competing hypotheses (Kass & Raftery, 1995). This demonstrative example captures the guiding idea of this work: network structures (e.g., Figure 1.1) encode information that can be used to formalize and test models.

1.2.2. Testing Strategy. We use the testing strategy in Williams and Mulder (2020) for confirmatory analyses. In this approach, hypotheses are expressed as (in)equality constraints on partial correlations. For example, $\mathcal{H}_1$ and $\mathcal{H}_2$ in (1.1) were expressed using inequality and equality constraints, respectively, on the edges in Figure 1.1. The evidence for such hypotheses can be quantified with the Bayes factor — a measure of relative support between competing hypotheses or models. In matrix notation, order constrained hypotheses can be written as

$$
\mathcal{H}_t : \mathbf{R}_t\boldsymbol{\rho} > \mathbf{r}_t, \tag{1.2}
$$

where $t = 1, \ldots, T$ denotes the competing hypotheses. In (1.2), $[\mathbf{R}_t\boldsymbol{\rho}|\mathbf{r}_t]$ is an augmented matrix that specifies the constraints under $\mathcal{H}_t$. In reference to (1.1), the system of inequalities under, say, $\mathcal{H}_1$, are formulated as

$$\mathbf{R}_{\mathcal{H}_1}\boldsymbol{\rho} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \rho_{B1-A1} \\ \rho_{B1-A2} \\ \rho_{B1-C1} \\ \rho_{B1-C2} \end{bmatrix} > \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \tag{1.3}$$

where $\mathbf{R}_{\mathcal{H}_1}$ denotes a matrix containing the coefficients for the contrasts of interest. Bayes factors can then be computed using the encompassing prior approach (Klugkist et al., 2005). The idea is to first specify an unconstrained (or encompassing) distribution for a hypothesis, $\mathcal{H}_u$, that does not place constraints on the partial correlations. This corresponds to an unconstrained network where theoretical expectations are not incorporated. The encompassing prior for GGMs is specified for the precision matrix, $\boldsymbol{\Theta}$, as a matrix-F distribution (Mulder & Pericchi, 2018). The implied marginal prior for the partial correlations is then

$$\rho_{ij} \sim beta(\frac{\delta}{2}, \frac{\delta}{2}) \text{ on } (-1, 1), \tag{1.4}$$

where δ is a prior hyperparameter that controls the standard deviation. The prior distribution for different values of δ can be seen in Figure 1.2. Note that it is not possible to place a beta prior on each ρ_{ij} directly because the resulting joint prior distribution for the partial correlation matrix would not be positive definite (for technical details see Mulder & Pericchi, 2018; Williams & Mulder, 2020).

The prior distributions under the constrained hypotheses are then obtained by truncating the encompassing prior according to the imposed constraints. Thus, instead of having to formulate T separate priors, only the unconstrained prior needs to be formulated. Furthermore, due to the encompassing prior approach, the Bayes factor of each constrained hypothesis against the unconstrained hypothesis $\mathcal{H}_u$ is straight forward to obtain, that is,

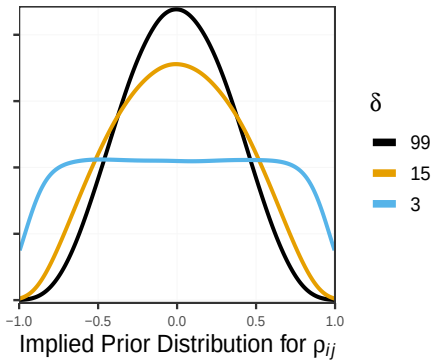

FIGURE 1.2. The implied marginal distribution for the encompassing prior on the partial correlations. The prior hyperparameter, δ, controls the standard deviation. Values of 99, 15, and 3 correspond to standard deviations of 0.10, 0.25, and 0.50, respectively.

$$\text{BF}_{tu} = \frac{\Pr(\boldsymbol{\rho} \in \Omega_t | \mathbf{Y}, \mathcal{H}_u)}{\Pr(\boldsymbol{\rho} \in \Omega_t | \mathcal{H}_u)}, \tag{1.5}$$

where Ω_t denotes the subspace under a constrained hypothesis $\mathcal{H}_t$ that satisfies the constraints on $\boldsymbol{\rho}$. In (1.5), the posterior probability in the numerator and the prior probability in the denominator can be understood as measures of 'relative fit' and 'relative complexity' of $\mathcal{H}_t$ relative to $\mathcal{H}_u$, respectively (Mulder, 2014). Once (1.5) has been obtained for all constrained hypotheses of interest, the Bayes factors between them can be computed using the transitivity property of the Bayes factor. For example, $\text{BF}_{12} = \frac{\text{BF}_{1u}}{\text{BF}_{2u}}$ provides the relative evidence in favor of $\mathcal{H}_1$. If we had, say, $\text{BF}_{12} = 5$, this would indicate the observed partial correlations are five times more likely under $\mathcal{H}_1$ than $\mathcal{H}_2$. Importantly, Bayes factors can also be viewed as measuring the relative success at predicting the observed data (Kass & Raftery, 1995). Once computed, Bayes factors can be used to obtain posterior model probabilities, that is, the probability that a hypothesis t is true given the data (Mulder, 2016). Assuming that all models have equal prior probabilities (i.e., $\frac{1}{T}$), the posterior probabilities for the $t = 1, \ldots, T$ hypotheses under consideration are given by

$$Pr(\mathcal{H}_t | \mathbf{Y}) = \frac{\text{BF}_{tu}}{\sum_{t'=1}^{T} \text{BF}_{t'u}}. \tag{1.6}$$

Whereas Bayes factors reflect the relative probability of the data under two hypotheses, posterior probabilities reflect relative support for a set of hypotheses given the data.

1.2.3. Summary. In this section we described a framework wherein Gaussian graphical models are used for both exploratory and confirmatory analyses. There are two aspects of this approach worth emphasizing. First, it allows for flexible testing of constraints in psychological networks. This readily allows for comparing theoretical models. For example, even for the relatively simple hypothesis in (1.1), testing whether the partial correlations are all greater than zero formally expressed the theoretical expectation of a positive manifold. Second, we demonstrated that the underlying network structure from an exploratory analysis encodes the necessary ingredients to generate hypotheses (e.g., Figure 1.1). This is a central idea of network analysis. The critical distinction is that we are presenting a comprehensive approach for formalizing and testing hypotheses generated from an exploratory analysis.

1.3. Empirical Applications

We now discuss in further detail how exploratory and confirmatory approaches can work in tandem to test hypotheses related to central structures. Recall that one motivation for network analysis was to generate hypotheses in an exploratory setting, and, in turn, a primary goal of this work is to bring this idea to fruition. To this end, we take on the perspective of a network researcher that formulates hypotheses based on an initial exploratory analysis and then tests them in a confirmatory setting. Note that Bayesian testing provides both the conditional dependence structure, $\mathbf{A}^{CD}$, and the conditional independence structure, $\mathbf{A}^{CI}$, which further opens the door for novel insights.

As mentioned above, there is theoretical and clinical interest in characterizing central structures involving bridge nodes (Castro et al., 2019; Cramer et al., 2010a, 2010b; Jones et al., 2019). This is because bridge symptoms are thought to drive the co-occurence of symptoms between communities and serve as targets for intervention (Beard et al., 2016; McNally et al., 2017). Thus, we focus on testing hypotheses related to bridge symptoms. To identify bridge symptoms, we rely on bridge strength rather than visual inspection. This is for two reasons. First, a node's bridge strength is defined as as the absolute sum of its inter-community edges, and therefore, highlights larger

effects (Jones et al., 2019). This brings in to focus central structures (i.e., a subsystem) on which hypotheses can be formulated. Second, the Bayesian framework we use places prior distributions on the partial correlations (Equation 1.4). With the exception of bridge expected influence, bridge strength is the only bridge statistic that accounts for these parameters[1].

1.3.1. Single Disorder. In the following, we estimate a GGM of PTSD symptoms and demonstrate how bridge strength can be used to identify central structures for which hypotheses can be formulated. In several examples, we discuss how to test these hypotheses in an independent dataset. We use data reported in Fried et al. (2018). Specifically, we use two samples of patients receiving treatment for PTSD ($n = 926$ and $n = 956$; Samples 3 and 4 in Table 1 in Fried et al. (2018)). The presence and severity of PTSD symptoms were assessed using the fourth edition of the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-IV*, American Psychiatric Association, 1994). Each of the 16 symptoms ($p = 16$) belonged to one of three communities (Re-experiencing, Avoidance, and Arousal).

1.3.1.1. *Exploratory Analysis.* We began by estimating the network structures $\mathbf{A}^{CD}$ and $\mathbf{A}^{CI}$ with the `R` package **BGGM** (Williams & Mulder, 2020) (panels A and B in Figure 1.3). Recall that there is strong theory in the network literature that expects all relations to be positive (i.e., a positive manifold, Borsboom et al., 2011; Horn & Cattell, 1966). This was formally incorporated into the analysis with a one-sided hypothesis test, $\mathcal{H}_0 : \rho_{ij} = 0$ versus $\mathcal{H}_1 : \rho_{ij} > 0$, for each partial correlation in the network (see The Gaussian Graphical Model). Hence, $\mathbf{A}^{CD}$ includes only positive relations. Because this analysis was used to formulate confirmatory hypotheses, we "erred on the side of discovery" (Bem, 2004) and used a Bayes factor threshold of three (this is considered "moderate" evidence, Lee & Wagenmakers, 2013) to determine the network structures.

With the network structures in hand, we proceeded to identify central nodes as indicated by bridge strength with the `R` package **networktools** (Jones et al., 2019). This is also the customary approach in network analysis, where, for example, the most central nodes are identified after estimating the structure (e.g., Beard et al., 2016; McNally et al., 2017). The results indicated that D1 ("sleep problems", bridge strength = 0.65) and B4 ("disinterest in activities", bridge strength =

[1]Bridge expected influence is identical to bridge strength but does not take the absolute value of edges before summing them (Jones et al., 2019)

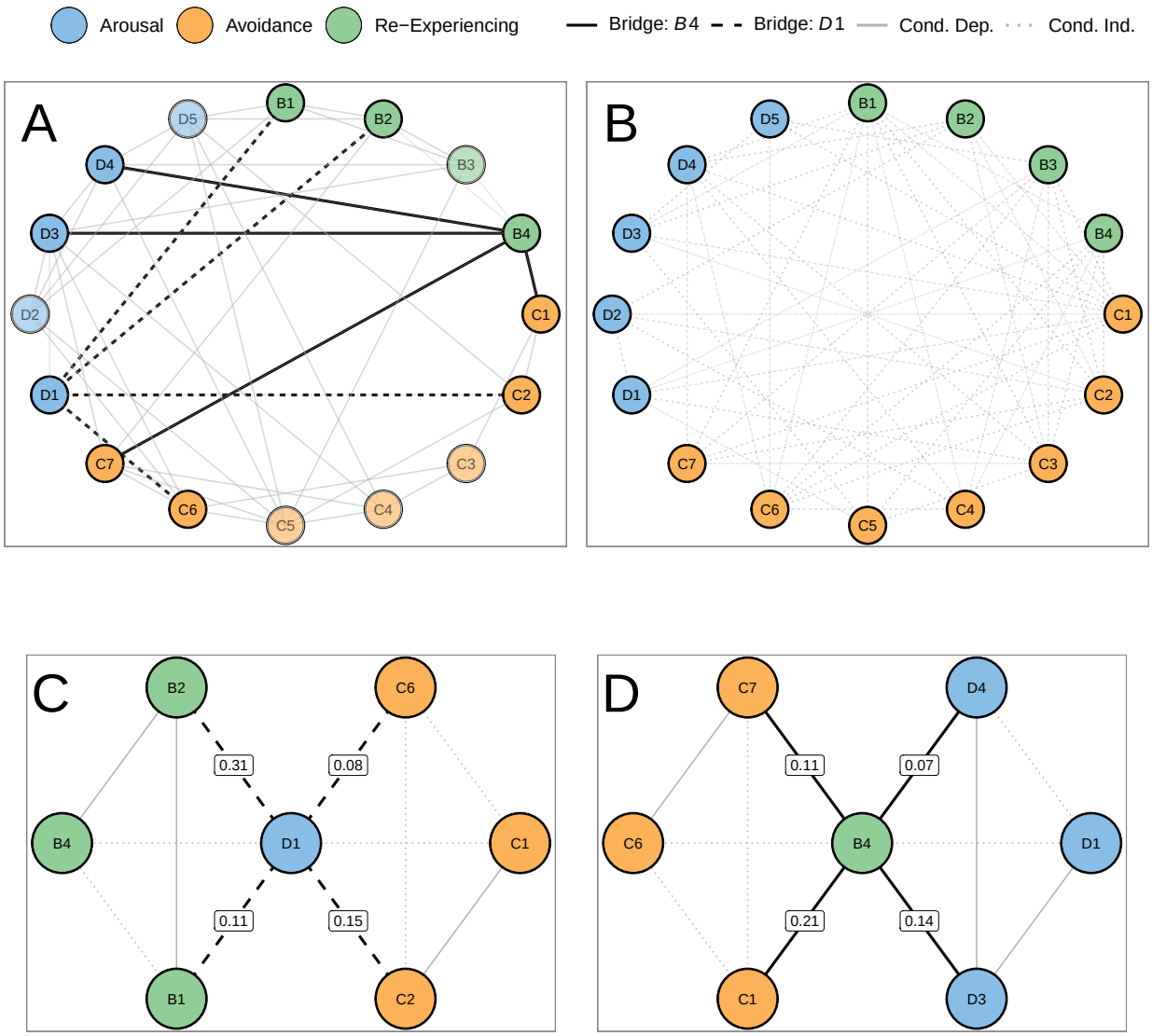

FIGURE 1.3. Exploratory network structure of PTSD symptoms. D1 ("sleep problems" and B4 ("disinterest in activities") emerged as bridge nodes. (**A**) The conditional dependence structure. Lines between two nodes indicate an association between them after controlling for all other nodes. This structure encodes information for testing associations such as $\mathcal{H}_1 : (\rho_{D1-C2}, \rho_{D1-C6}) > 0$. (**B**) The conditional independence structure. Dotted lines indicate that there is no association between two nodes after controlling for all other nodes. This structure encodes information for testing null associations such as $\mathcal{H}_1 : (\rho_{B4-C6}, \rho_{D1-B4}, \rho_{D1-C1}) = 0$. (**C** & **D**). A magnified look at the neighborhood of bridge relations for D1 and B4. We extracted the information encoded in these structures to formulate and test hypotheses (see Equations 1.7, 1.8, and 1.9).

0.53) were the most central nodes (see Table A.1 for full definitions). The neighborhood of bridge relations for both nodes can be seen in Figure 1.3 (panel A).

1.3.1.2. *Confirmatory Analysis.* We emphasize that centrality indices summarize an inherently exploratory analysis and only provide information from afar. For example, it is possible to have a top ranking bridge symptom emerge in two datasets with completely different bridging structures. Our approach extends the utility of bridge centrality metrics to confirmatory testing by using them to formulate hypotheses on the most central symptoms in the network. We thus focus on node D1 ("sleep problems") and node B4 ("disinterest in activities"). Figure 1.3 (panel C and D) zooms in on these top ranking bridge symptoms and their respective neighborhoods of bridge relations. The key idea is that honing into central symptoms allows researchers to easily formulate hypotheses of substantive or theoretical importance.

Varying degrees of replication. The topic of replicability has recently captivated the network literature (Forbes et al., 2017; Fried et al., 2018; Williams, 2020). To assess replicability, it is common to focus on individual edges with either classical (van Borkulo et al., 2016) or Bayesian testing (Williams, Rast, et al., 2020). Although the latter has the advantage of directly providing evidence for equality of partial correlations, it is possible to ask even more fine-grained questions about replication. For example, to what degree do central structures replicate? This can be expressed with formal models.

We first focused on node B4 (Figure 1.3, panel D) and tested the following hypotheses

$$
\begin{aligned}
&\mathcal{H}_1 : (\rho_{B4-C1}, \rho_{B4-C7}, \rho_{B4-D3}, \rho_{B4-D4}) > 0 \\
&\mathcal{H}_2 : \rho_{B4-C1} > (\rho_{B4-C7}, \rho_{B4-D3}, \rho_{B4-D4}) > 0 \\
&\mathcal{H}_3 : \text{“not } \mathcal{H}_1 \text{ or } \mathcal{H}_2\text{.”}
\end{aligned} \tag{1.7}
$$

In (1.7), $\mathcal{H}_1$ is testing for replication of all edges but is otherwise agnostic towards the interplay among bridge relations. $\mathcal{H}_2$ then provides a refined view into the bridge neighborhood by testing an additional constraint that the strongest edge replicated. That is, all of the bridge relations *and* the strongest edge re-emerged in an independent dataset. Furthermore, $\mathcal{H}_1$ and $\mathcal{H}_2$ both reflect a positive manifold. We also included $\mathcal{H}_3$ which accounts for structures that are not $\mathcal{H}_1$ or $\mathcal{H}_2$.

We compared the first two hypotheses against $\mathcal{H}_3$, where there was strong evidence for both $\mathcal{H}_1$ ($\text{BF}_{13} = 33.4$) and $\mathcal{H}_2$ ($\text{BF}_{23} = 120.4$). Hence, the data were more likely under the replication models than a model that did not include replication-based constraints. We then compared $\mathcal{H}_1$ to $\mathcal{H}_2$. Although the evidence was not strong, the data were more likely under $\mathcal{H}_2$ ($\text{BF}_{21} = 3.6$). This analysis indicates that (1) the bridge relations replicated in an independent dataset; and (2) the relation between "sleep problems" (node B4) and "avoidance of thoughts" (node C1) *could* be the strongest bridge between the Re-experiencing and Avoidance communities.

We then focused on node D1 (Figure 1.3, panel D) and tested the following hypotheses

$$\begin{aligned} \mathcal{H}_1 &: (\rho_{D1-C2}, \rho_{D1-C6}) > 0 \\ \mathcal{H}_2 &: (\rho_{D1-C2}, \rho_{D1-C6}) < 0 \\ \mathcal{H}_3 &: (\rho_{D1-C2}, \rho_{D1-C6}) = 0. \end{aligned} \tag{1.8}$$

In (1.8), the hypotheses are in relation to the Avoidance community and they again reflect network replication. For example, $\mathcal{H}_1$ expresses that both relations are positive, but does not impose an order restriction among bridge edges, whereas $\mathcal{H}_2$ expresses that both relations are negative and similarly does not impose an order restriction. Alternatively, $\mathcal{H}_3$ then captures the importance of ruling out conditional independence or that the effects are actually zero. Although the data were more likely under $\mathcal{H}_1$ than $\mathcal{H}_2$ ($\text{BF}_{12} = 12.3$), the evidence favored $\mathcal{H}_3$ over $\mathcal{H}_1$ ($\text{BF}_{31} = 11$). In other words, of the formal, replication models, there was evidence for null associations, or that these might *not* be bridge relations after all.

The same hypotheses in (1.8) were tested in relation to the Re-experiencing community. In this case, there was overwhelming evidence for $\mathcal{H}_1$. The posterior hypothesis probability was essentially 1, which translates into an infinite Bayes factor. This indicates that the bridge relations replicated for the Arousal and Re-experiencing communities, and demonstrates the utility of comparing formal models for clinical applications in particular. Although network analyses are often thought to identify target symptoms for interventions (e.g., Beard et al., 2016), we are not aware of any work

that has followed up an initial, exploratory analysis, with the goal of confirming hypotheses related to the potential targets. In this case, the results suggest that the symptom "sleep problems" may be useful in guiding interventions.

Ruling Out Bridges. It is important to rule out bridge relations in establishing the structure of inter-community relations. Here, the question of replication is concerned with null associations re-emerging in an independent dataset. To show this, we included two null associations in both panel C and D of Figure 1.3 (the dotted lines). We thus formulated the following hypotheses

$$\begin{aligned} \mathcal{H}_1 &: (\rho_{B4-C6}, \rho_{D1-B4}, \rho_{D1-C1}) = 0 \\ \mathcal{H}_2 &: (\rho_{B4-C6}, \rho_{D1-B4}, \rho_{D1-C1}) > 0 \\ \mathcal{H}_3 &: \text{"not } \mathcal{H}_1 \text{ or } \mathcal{H}_2\text{."} \end{aligned} \tag{1.9}$$

which represent a null model ($\mathcal{H}_1$), a positive manifold model ($\mathcal{H}_2$), and a model accounting for alternative structures ($\mathcal{H}_3$). The positive manifold model had a posterior hypothesis probability of essentially zero, indicating that positive associations can be ruled out. Further, the data were more likely under the conditional independence model, $\mathcal{H}_1$, than under $\mathcal{H}_3$ ($\text{BF}_{13} = 4.7$)[2]. This is striking because these very same relations have large bivariate correlations, yet, after controlling for the other symptoms in the network, there was evidence for conditional independence.

1.3.2. Multiple Disorders. Here we provide further examples our proposed framework using a comorbidity network. We estimate a GGM containing anxiety and depression symptoms and use bridge strength to identify central structures for which hypotheses can be formulated. In several examples, we discuss how to test these hypotheses in an independent dataset. We use data from Beard et al. (2016) that includes 16 symptoms gathered from 1029 patients receiving treatment for depression and anxiety. Symptoms were assessed using the Patient Health Questionnaire-9 (Kroenke et al., 2001) and the 7-item Generalized Anxiety Disorder Scale (Spitzer et al., 2006). Nine symptoms were in the "depression" community and seven symptoms were in the "anxiety" community. Because only one dataset was available, we performed exploratory analyses on one

[2]Changing the prior distribution resulted in *more* support for $\mathcal{H}_1$. This suggests $\text{BF}_{13} = 4.7$ is a lower bound for the evidence in favor of the null model.

half of the data, and used the other half for confirmatory testing (i.e., "data splitting"; Dahl et al., 2008; Faraway, 1995).

1.3.2.1. *Exploratory Analysis.* We followed the same procedure as above: (1) estimate the conditional dependence and independence structures; (2) identify the top scoring bridge symptoms; and (3) formulate hypotheses based on the results. The results indicated that node D8 ("motor", bridge strength = 0.40)[3] and node D6 ("guilt", bridge strength = 0.27) were the most central according to bridge strength (see Table A.2 for full definitions). Figure 1.4 displays the resulting (in)dependence structures (panels A and B) and the magnified neighborhood of bridge relations for nodes D8 and D6 (panel C).

1.3.2.2. *Confirmatory Analysis.* We reiterate that our confirmatory testing approach builds upon identifying bridge symptoms in an exploratory analysis to gain insights regarding central structures of a network. This can be done by developing hypotheses targeting the most central nodes as determined by centrality statistics (in this case bridge strength). Accordingly, we have magnified the neighborhood of bridge relations for nodes D8 and D6 (panel C Figure 1.4). This readily allows for devising hypotheses with substantive and theoretical relevance. Of course, an important first step is to investigate the extent to which the relations replicate. Note that we are not referring to simply detecting the effect, but testing the constrained models implied from the exploratory analysis.

Intra- and Inter-Bridge Sets. In addition to testing whether the edges for a bridge symptom simply re-emerge in a new dataset, it may be useful to test whether their exact order replicates. If the order of edges is known, and assuming a useful focal point is the strongest relation, this would imply an ordering among possible intervention targets. This notion is encoded in the exploratory analysis (panel C in Figure 1.4) which leads to the following hypotheses

$$\begin{aligned} &\mathcal{H}_1 : \rho_{D8-A5} = \rho_{D8-A7} > (\rho_{D6-A3}, \rho_{D6-A6}) > 0 \\ &\mathcal{H}_2 : \rho_{D8-A5} > \rho_{D8-A7} > \rho_{D6-A3} > \rho_{D6-A6} > 0 \\ &\mathcal{H}_3 : \text{"not } \mathcal{H}_1 \text{ or } \mathcal{H}_2\text{."} \end{aligned} \tag{1.10}$$

[3] "motor" refers to physical lethargy or restlessness (Kroenke et al., 2001)

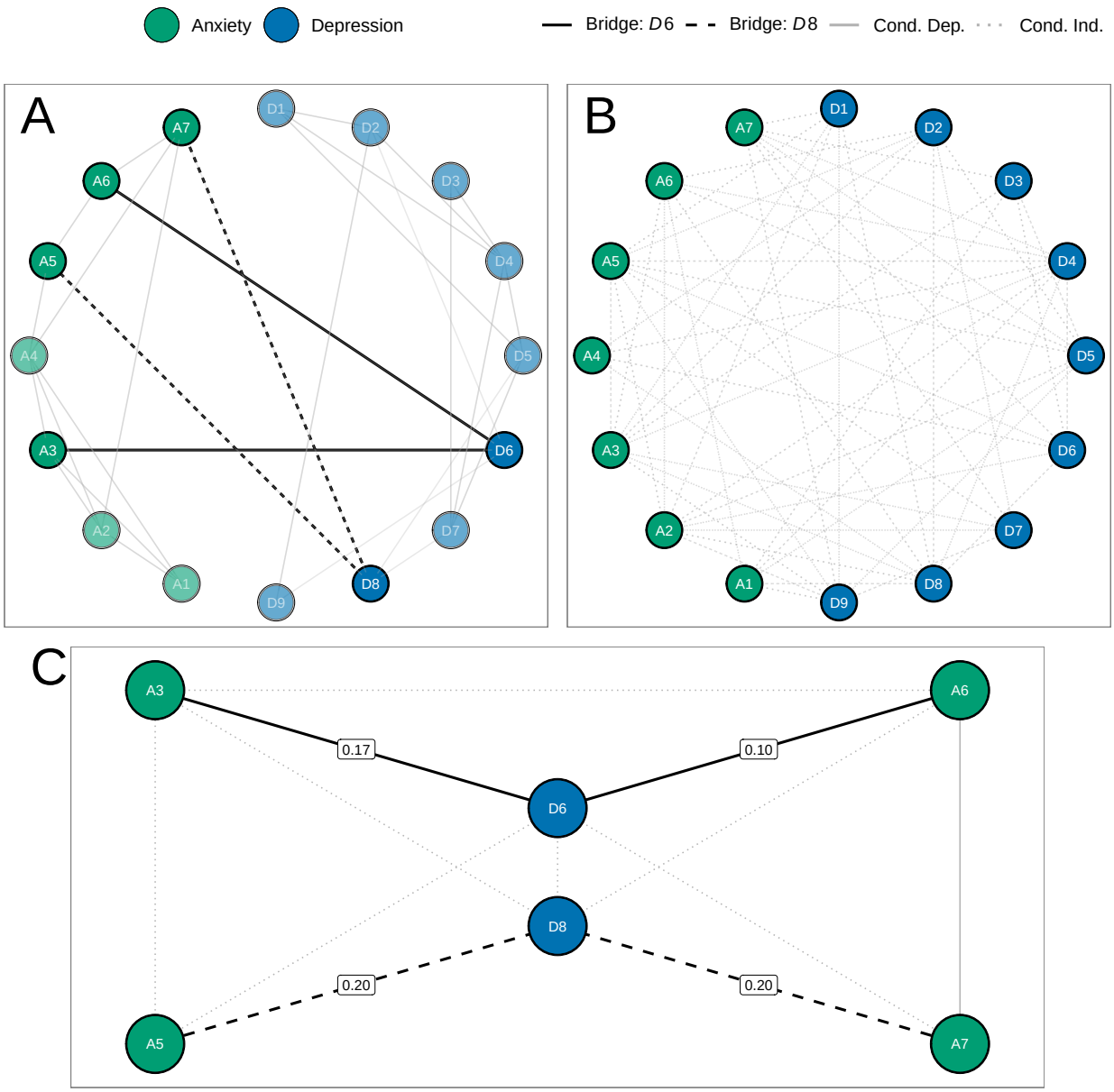

FIGURE 1.4. Exploratory network structure of depression and anxiety symptoms. D6 ("guilt") and D8 ("motor") emerged as bridge nodes. (**A**) The conditional dependence structure. Lines between two nodes indicate an association between them after controlling for all other nodes. (**B**) The conditional independence structure. Dotted lines indicate that there is no association between two nodes after controlling for all other nodes. (**C**). A magnified look at the neighborhood of bridge relations for D6 and D8. We extracted the information encoded in these structures to formulate and test hypotheses (see Equations 1.10, 1.11, and 1.12).

In (1.10), the hypotheses focus on characterizing bridge sets, or the set of bridge edges belonging to a given symptom. For example, $\mathcal{H}_1$ posits that the bridge set for node D8 ("motor") is collectively greater than the set for node D6 ("guilt"), with constraint that the edges for node D8 are equal to each other. This effectively corresponds to testing whether node D8 has greater bridge strength than node D6. $\mathcal{H}_2$ then refines $\mathcal{H}_1$ by testing an exact order both between and within bridge sets. The data were more likely under both $\mathcal{H}_1$ ($\text{BF}_{13} = 4.4$) and $\mathcal{H}_2$ ($\text{BF}_{23} = 107$) than $\mathcal{H}_3$. Furthermore, there was more evidence supporting the hypothesis testing solely inequality constraints, $\mathcal{H}_2$, than the one including an equality constraint ($\text{BF}_{21} = 24.1$). This provides a clear characterization of the the bridge relations at hand — not only did the order of bridge strength replicate, but so did the order of the edges within the neighborhood of each bridge symptom. In this confirmatory test, the relationship between the depression symptom "motor" and anxiety symptom "restless" emerged as the top relation. This characterizes a central structure between anxiety and depression and can inform theory development with respect to these disorders.

Bridge Set Separation. It may further be of interest to identify whether bridge sets include common elements. That is, whether bridge symptoms connect to the same or different nodes. This may be useful in understanding whether bridge symptoms represent distinct central structures. As can be seen in panel C of Figure 1.4, the bridge sets for nodes D8 and D6 are mutually exclusive. This implies that there are two subsystems. Keeping this in mind, we formulated the following set of hypotheses

$$
\begin{aligned}
\mathcal{H}_1 &: (\rho_{D8-A3}, \rho_{D8-A6}) = 0 \\
\mathcal{H}_2 &: (\rho_{D8-A3}, \rho_{D8-A6}) > 0 \\
\mathcal{H}_3 &: \text{"not } \mathcal{H}_1 \text{ or } \mathcal{H}_2\text{."}
\end{aligned}
\tag{1.11}
$$

In (1.11), $\mathcal{H}_1$ tests conditionally independent associations between the symptom "motor" and the bridge set for "guilt" (i.e., nodes A3 and A6) versus $\mathcal{H}_2$, a positive manifold model, and $\mathcal{H}_3$, a model accounting for alternative structures. Although the data were more likely under $\mathcal{H}_1$ than $\mathcal{H}_3$ ($\text{BF}_{13} = 3.7$), there was support in favor of $\mathcal{H}_2$ compared to $\mathcal{H}_1$ ($\text{BF}_{21} = 2$). This analysis

suggests there is a small amount of evidence that "motor" has conditional dependent relations with the same nodes as "guilt".

We repeated the hypothesis tests in (1.11) for "guilt". Here, the findings differed slightly — the data were almost equally likely under $\mathcal{H}_1$ compared to $\mathcal{H}_3$ ($\text{BF}_{13} = 1.6$). Like above, however, the data were more likely under $\mathcal{H}_2$ than either $\mathcal{H}_1$ ($\text{BF}_{21} = 6.3$) or $\mathcal{H}_3$ ($\text{BF}_{23} = 10$). These analyses indicate that the null associations did not replicate, and instead support the idea that "motor" and "guilt" connect to the same symptoms. This information suggests, for example, that the nodes in panel C of Figure 1.4 make up a single central structure instead of two.

Bridge Node Separation. Thus far we have focused on testing multiple relationships simultaneously. While testing joint hypotheses is a key feature to our proposed testing strategy, it may be that a single parameter is of particular interest, say, due to theoretical importance. In this case, it is highly informative to test hypotheses focused on a single parameter. For example, panel C (Figure 1.4) indicates that nodes D8 and D6 are conditionally independent. However, this is in contrast to what might be expected from two symptoms in the same community. Accordingly, one can test a hypothesis focused solely on this relationship, for example

$$\begin{aligned} &\mathcal{H}_1 : \rho_{D8-D6} = 0 \\ &\mathcal{H}_2 : \rho_{D8-D6} > 0 \\ &\mathcal{H}_3 : \text{“not } \mathcal{H}_1 \text{ or } \mathcal{H}_2\text{.”} \end{aligned} \tag{1.12}$$

In (1.12), $\mathcal{H}_1$ tests a null association and $\mathcal{H}_2$ expresses a positive relationship. Though the data were more likely under $\mathcal{H}_3$ than $\mathcal{H}_2$ ($\text{BF}_{32} = 5.4$), there was more evidence in favor of $\mathcal{H}_1$ than $\mathcal{H}_3$ ($\text{BF}_{13} = 2.2$). Hence, there is some evidence in this confirmatory test that "motor" and "guilt" are conditionally independent symptoms. Importantly, focused hypotheses, such as in (1.12), can be used to draw powerful inferences with respect to relationships of particular interest.

1.4. Simulation Studies

Thus far we have provided a comprehensive framework for exploratory and confirmatory testing of central structures in partial correlation networks. Our hope is that researchers will integrate the

proposed methods into their own work. Therefore, it is important to understand how these methods behave under certain conditions. To this end, we emphasize a few important points:

(1) Bayes factors tend to infinity and posterior model probabilities tend to one in favor of the correct model with increasing sample size (O'Hagan, 1995). Although this property assumes the *true* model is being considered, recall that Bayes factors can also be interpreted as a measure for the relative success of predicting the observed data (Kass & Raftery, 1995). This perspective does not rely on the existence of a true model and is our preferred interpretation.

(2) More specific hypotheses result in higher degrees of evidence, given that they are supported by the data. This is due to being relatively less 'complex' and having relatively better 'fit' (e.g., Klugkist et al., 2005; Mulder, 2014). This was observed in (1.7), where the data were most likely under two replication-based hypotheses, but of these two, the more specific one yielded stronger evidence.

(3) The scale, or standard deviation, of the encompassing prior distribution influences the outcome when testing equality constrained hypotheses, but not when testing inequality constraints (i.e., the Jeffreys-Lindley paradox is not an issue, Mulder, 2014). This was seen in (1.9), where changing the prior distribution changed the evidence in support for $\mathcal{H}_1$, an equality constrained hypothesis. This is important to consider when testing confirmatory analyses because the resulting evidence can be considered objective for inequality constrained hypotheses[4]. That is, the Bayes factor is robust to the prior distribution.

We further examine these properties in two simulation studies. In each, data were generated with four variables and the true precision matrix

$$\boldsymbol{\Theta} = \begin{bmatrix} 1 & -0.10 & -0.14 & -0.18 \\ -0.10 & 1 & -0.22 & -0.26 \\ -0.14 & -0.22 & 1 & -0.3 \\ -0.18 & -0.26 & -0.30 & 1 \end{bmatrix}. \tag{1.13}$$

[4]In the case of equality constrained hypotheses, sensitivity analyses can be performed to determine the influence of the prior on the resulting Bayes factors (Hoijtink et al., 2019).

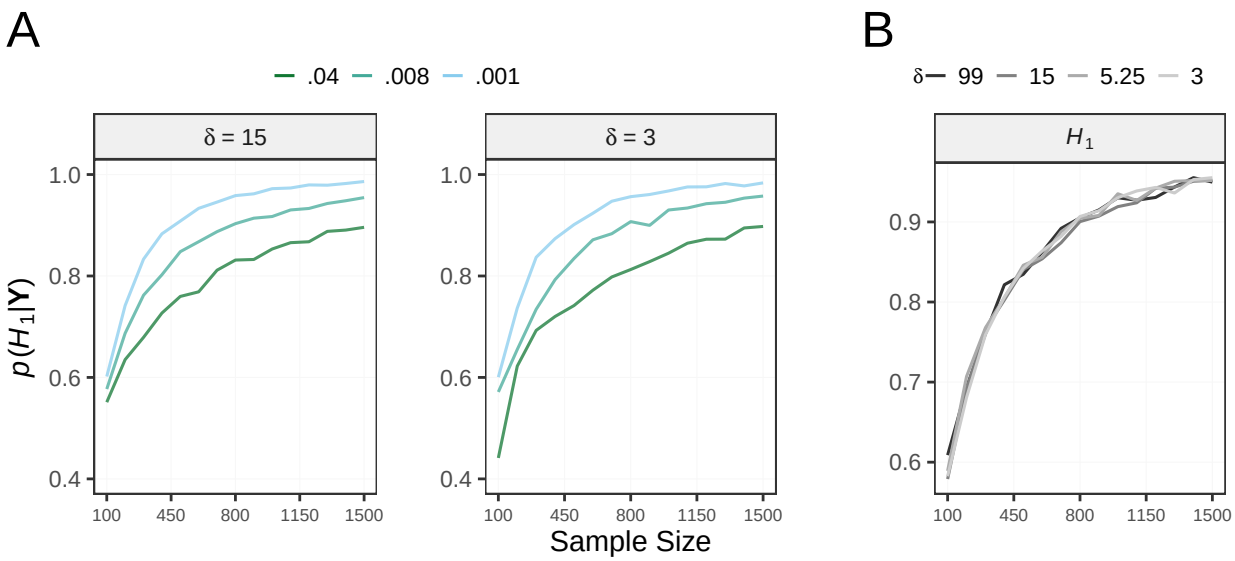

FIGURE 1.5. Results from the Simulation Studies The y-axes denote posterior model probabilities (PMPs). (**A**) In Study 1: Specific Hypotheses, three true and increasingly specific hypotheses were tested against competing hypotheses, across two values for the prior variance (δ). Each line captures the PMP for a true hypothesis and each color denotes its respective prior proportion in agreement with the unconstrained hypothesis. Though all PMPs tended to one with increasing sample size, more specific hypotheses needed less samples to receive strong evidence. The results support the notion that more specific hypotheses are preferred in the encompassing prior approach given that they are supported by the data. (**B**). In Study 2: Prior Specification, one true inequality constrained hypothesis was specified against a competing hypothesis across five values of δ. All PMPs tended to one with increasing sample size, and notably, overlap almost perfectly. This supports the notion that the evidence when testing inequality constrained hypotheses is robust to the prior specification.

The partial correlations then correspond to the off-diagonal elements of $\boldsymbol{\Theta}$ with the sign reversed. These values were based on the most common partial correlations we observed in our analyses. We then tested hypotheses based $\boldsymbol{\Theta}$ and computed posterior model probabilities. The first simulation study examines the advantage of testing precise hypotheses and the second evaluates the influence of the prior scale when testing inequality constrained hypotheses.

1.4.1. Study 1: Specific Hypotheses. In this study, we investigate posterior model probabilities for true hypotheses that have a different number of constraints. The idea is that more specific hypotheses (i.e, more constraints are placed) have smaller prior probabilities than less specific ones under the encompassing prior (see Equation 1.5). If a more specific model is supported

by the data, this will result in a larger Bayes factor, and accordingly, a larger posterior probability. Thus, there is a greater "reward" for formulating and testing specific hypotheses on central structures. We formulated $s = 1, 2, 3$ sets of hypotheses with $t = 1, 2, 3$ hypotheses in each. The *true* hypotheses corresponded to

$$\begin{aligned} &\mathcal{H}_1^1 : \rho_{23} > \rho_{14} > \rho_{13} > \rho_{12} \\ &\mathcal{H}_1^2 : \rho_{24} > \rho_{23} > \rho_{14} > \rho_{13} > \rho_{12} \\ &\mathcal{H}_1^3 : \rho_{34} > \rho_{24} > \rho_{23} > \rho_{14} > \rho_{13} > \rho_{12}. \end{aligned} \tag{1.14}$$

In (1.14), $\mathcal{H}_t^s$ denotes hypothesis t in set s. Each ρ_{ij} is the partial correlation corresponding to the element in the ith row and jth column of $\boldsymbol{\Theta}$ in (1.13). The number of constraints for the true hypotheses imply different prior proportions in agreement with the unconstrained parameter space, $\boldsymbol{\Omega}_t$, such that more constraints result in smaller prior proportions. For the hypotheses in (1.14), the proportions in agreement with $\boldsymbol{\Omega}_t$ are .04, .008, and .001, respectively. Each $\mathcal{H}_1^s$ was compared to a null and complement hypothesis. For example, $\mathcal{H}_1^1$ was compared to

$$\begin{aligned} &\mathcal{H}_2^1 : (\rho_{23}, \rho_{14}, \rho_{13}, \rho_{12}) = 0 \\ &\mathcal{H}_3^1 : \text{"Not } \mathcal{H}_1^1 \text{ or } \mathcal{H}_2^1\text{."}. \end{aligned} \tag{1.15}$$

Within each set, we assumed equal prior probabilities (i.e., $\frac{1}{3}$). Each hypothesis was first compared to the unconstrained model, $\mathcal{H}_u$, which resulted in BF_{tu}^s for all hypotheses. These Bayes factors are not of substantive interest, but they are needed for then computing BF_{12}^s, BF_{13}^s, and the posterior model probabilities. We considered sample sizes ranging from 100 to 1,500 (in increments of 100), and two values for the scale, $\delta \in \{3, 15\}$. The latter values correspond to prior standard deviations of 0.5 and 0.25, respectively, and were chosen because we view them as the most likely to be used in practice. The posterior hypothesis probabilities, $p(\mathcal{H}_1|\mathbf{Y})$, were averaged across 500 simulation trials.

1.4.1.1. *Results.* The results are shown in panel A Figure 1.5. The lines capture the posterior probability for the true hypotheses in (1.14) and each color denotes their respective prior complexity (i.e., the prior proportion in agreement with $\mathbf{\Omega}_t$). Across all conditions, the posterior probabilities tended towards one as sample size increased. Importantly, the most specific hypothesis (light blue line) received more support across all conditions. In fact, the posterior probability for $\mathcal{H}_1^3$ was over 0.9 with just 500 observations for both values of δ. However, note that the posterior probabilities differ according to the prior scale. This is due to the inclusion of each $\mathcal{H}_2^s$, an *equality* constrained hypothesis. This is a natural property of the methodology because the prior reflects the expected magnitude of the partial correlations when the equality does not hold. Together, these results show that when conducting confirmatory analyses, more specific hypotheses are preferred, and indeed result in greater posterior probabilities, given that they are supported by the data.

1.4.2. Study 2: Prior Specification. The use of Bayes factors has been critiqued for being overly sensitive to the choice of prior distribution (e.g, C. C. Liu and Aitkin, 2008, but see J. Rouder et al., 2016). However, a proposed advantage of the encompassing prior approach is that the resulting Bayes factors are robust to the prior variance when testing inequality constrained hypotheses (Hoijtink, 2011; Klugkist et al., 2005). This is an important consideration in confirmatory testing because the Bayes factor would be consistent regardless of how the scale for the encompassing prior is specified. Thus, we investigated the extent to which the prior hyperparameter δ influences posterior probabilities for inequality constrained hypotheses. We first formulated a single hypothesis set

$$\begin{aligned} &\mathcal{H}_1 : \rho_{34} > \rho_{24} > \rho_{23} > \rho_{13} > \rho_{12} \\ &\mathcal{H}_2 : \text{“not } \mathcal{H}_1\text{.”} \end{aligned} \tag{1.16}$$

with $\mathcal{H}_1$ as the true hypothesis. We then varied the prior hyperparameter $\delta \in \{99, 15, 5.25, 3\}$. These values correspond to prior standard deviations of 0.1, 0.25, 0.4, and 0.5, respectively. Prior probabilities for each hypothesis were assumed to be equal, and BF_{12} was computed. We considered

sample sizes ranging from 100 to 1,500 (in increments of 100), and posterior hypothesis probabilities, $p(\mathcal{H}_1|\mathbf{Y})$, were averaged across 500 simulation trials.

1.4.2.1. *Results.* The results are shown in Panel B Figure 1.5. Each line corresponds corresponds to the posterior probability of $\mathcal{H}_1$ versus $\mathcal{H}_2$ in (1.16) for a given value of δ. Importantly, the posterior hypothesis probabilities for each prior overlap almost perfectly across all sample sizes, and tended towards one as sample size increased. Together, the results indicate that the resulting Bayes factor is robust to different priors testing inequality constrained hypotheses.

1.5. Discussion

In this paper, we presented an innovative strategy for integrating exploratory analyses with confirmatory hypothesis testing in partial correlation networks. In doing so, one of the primary motivations for network analysis — hypothesis generation — has been fully realized. We began with an illustrative example based on a customary exploratory approach wherein, by simply plotting the network structure, we formulated several hypotheses regarding the most central node. This highlighted how information encoded by the partial correlations and the conditional (in)dependence structures can be employed to formalize clinical and theoretical expectations, that in turn, can be tested in a confirmatory setting.

The core contribution of this work demonstrated how centrality metrics can be used to guide hypothesis generation in exploratory network analyses. In extensive examples, bridge strength, a measure of inter-community connectivity, served to identify central bridge symptoms in two networks. Once identified, we formulated hypotheses with the goal of understanding various aspects of the bridging structures. For instance, whether the set of edges for one bridge symptom overlap with the set of another or whether two bridge symptoms are conditionally independent (see Empirical Applications). Together, these examples highlighted how inherently exploratory metrics can inform hypotheses aimed at replicating aspects of the network structure.

Testing specific structures can shed a new light upon the issue of replicability in networks. Indeed, formulating fine-grained hypotheses focused on characterizing central nodes is an active issue in the network literature (Epskamp, Borsboom, et al., 2018; Forbes et al., 2017, 2019). In this work, however, we successfully replicated multiple central structures across distinct datasets. Note

that we chose to focus on aspects we deemed most important. Namely, edge weights for the most central structures. Hence, it seems that network structures of interest can indeed be replicated.

1.5.1. Implications for Building Formal Models. The proposed testing strategy has several desirable qualities for building formal models. There is now a wealth of network analyses for several mental disorders (e.g., PTSD) and synthesizing this information to develop formal theories is a pressing challenge (for detailed discussions see Haslbeck et al. (2019) and Borsboom et al. (2020)). In order to move towards formal theories, researchers must move away from the traditional exploratory approach and begin testing confirmatory hypotheses. We argue that this is not only necessary for building formal models, but also thinking about clinical interventions. This important step is absent from the current literature, in that results from exploratory analyses are never confirmed. This is in contrast to other scientific disciplines that also use partial correlation networks. In biological fields, for instance, exploratory results are actually used to generate hypotheses that are then tested (Kelder et al., 2010; Krumsiek et al., 2012). These fields often conduct controlled experiments — perhaps a bridge too far for the most common applications in the social-behavioral sciences. However, as we demonstrated, it is certainly possible and quite useful to confirm findings that emerged in an exploratory context.

Furthermore, the Bayesian aspect of our approach is well-suited for constructing theories. Because we use the Bayes factor for confirmatory testing, we are quantifying the relative success of hypotheses at predicting the observed data (Kass & Raftery, 1995) — an important measure of explanatory power. Moreover, developing formal theory is an iterative process which requires updating as more data becomes available. Bayesian analyses naturally lend themselves to this because prior information can be incorporated. Specifically, the results of a Bayesian analysis (i.e, posterior odds) can be formally incorporated into subsequent analyses as prior odds. This allows for monitoring the evidence a given theory has amassed.

Finally, our approach facilitates testing "risky predictions" (Mayo, 1991; Meehl, 1967). That is, a prediction that extends beyond refuting a null hypothesis or simply testing a direction (e.g., the effect is positive). The idea behind our approach is that hypotheses can express precise expectations through (in)equality constraints. This was demonstrated in this work, for example, by testing whether an exact ordering of effects replicated in a new dataset. This is useful in developing

theories. Further, our approach can be used for testing theories by allowing researchers to be even more explicit about what should be observed. For instance, one could extend an exact ordering of edges by stipulating an additional constraint that they are all bounded between two values, say 0.10 and 0.20. This is a key aspect of theory building, that is, formulating and testing theoretical expectations.

1.5.2. Embracing the Gaussian Graphical Model. We urge researchers to embrace Gaussian graphical modeling. In our opinion, the focus on causality in psychological networks has led to an underappreciation of undirected networks as valuable tools for more than just exploratory data analysis. As we demonstrated, formalizing theoretical models can be accomplished by thinking in terms of constraints, on, say, the interactions between clinical symptoms. This allows researchers to establish, describe, and characterize important relations. This can be accomplished by adopting the powerful framework described in this work for exploratory and confirmatory testing. This is an important first step towards moving beyond the notion that GGMs are *merely* a stepping stone to directed networks.

1.5.3. Limitations. There are some notable limitations in this work. First, we only considered bridge strength as a metric to identify the most central structures in a network. We viewed this as a sensible choice because both the prior distributions for the confirmatory testing strategy and bridge strength focus on the partial correlations. Though this choice made it straight-forward to formulate hypotheses, it may not always be so clear what parameters to focus on when using alternative exploratory metrics (Bringmann et al., 2019; Jones et al., 2019). Second, because only covariance matrices were available we assumed multivariate normality when generating data. However, the data were collected as ordinal. In practice, ordinal data results in more sampling variability, and thus less statistical power to replicate effects[5] (Williams, 2020). Third, Bayes factor estimates may be unstable when testing overly specific hypotheses. This is is because the prior probability that the constraints are in agreement with the unconstrained parameter subspace becomes quite small, and thus, a prohibitively large number of samples are needed for accurate estimates (Mulder, 2016). In our experience, this issue typically arises when overly specific hypotheses are specified in conjunction with unordered groupings (e.g., (ρ_{12}, ρ_{13})). Fourth, when comparing nested hypotheses such as $\mathcal{H}_1$

[5]Methods for dealing with ordinal data in GGMs are implemented in the `R` package **BGGM**

and $\mathcal{H}_2$ in (1.7), the Bayes factor for the more specific hypothesis (e.g., BF_{21}) is bounded. As a result, the scale of the Bayes factor is difficult to interpret and the evidence for the true hypothesis does not tend to infinity with increasing sample size (if the more specific hypothesis is true, Mulder et al., 2010). Nested hypotheses can still be tested if there is a reason to do so, say, based on theoretical reasoning, but this caveat should be kept in mind when interpreting the evidence. Lastly, we did not conduct sensitivity analyses for any of our confirmatory hypothesis tests so it is uncertain to what extent the prior distribution influenced our results. While sensitivity analyses should be conducted in practice, we avoid doing so due to the demonstrative nature of this work.

1.5.4. Recommendations. We recommend that researchers make several considerations when using the exploratory and confirmatory strategies described in this paper. To start, researchers should carefully think about how exploratory metrics relate to the scientific question at hand when using them to guide the formulation of hypotheses. In particular it is not clear what centrality indices measure in psychological networks (Bringmann et al., 2019). For example, metrics using measures of "betweeness" and "closeness" assume the existence of a shortest path. Because shortest paths do not account for edge weight they may contradict how psychological variables are thought to interact. As such, it is important that researchers determine how exploratory metrics relate to their research prior to their use.

If independent samples are not available or if hypotheses cannot be derived from previous research, we recommend researchers take advantage of data splitting methods (Anderson & Magruder, 2017; Dahl et al., 2008; Faraway, 1995). We believe data splitting is underutilized in psychological research, and provides an accessible method for obtaining independent data on which to formulate and then test hypotheses. Indeed, data splitting could be called "one of the most seriously neglected ideas in statistics" (comment in Stone, 1974). Although this procedure results in a loss of statistical power, there are several ways to mitigate this. For example, power can be increased by focusing on inequality constrained hypotheses (e.g., Mulder & Raftery, 2019), or by focusing on the strongest edges. Additionally, a lower Bayes factor threshold can be used in determining the graph. Though lowering the threshold results in greater power, it also increases the rate of false positives. Thus, we recommend researchers focus on testing inequality constraints and large effects when splitting their data.

In fact, it may be useful to prioritize large effects in general. This is because there is an inherent limit to what can and cannot be confirmed in a data set. Characterizing large effects first then and smaller effects second can be thought of as a *top-down* approach that can guide exploratory and confirmatory analyses over time. This idea is not new. In the genetics literature, it has been suggested that focusing on large effects is a useful way to begin understanding a system. For example, Altay and Emmert-Streib (2010) state that

> "However, practically, no method can guarantee to [infer an entire network] for a given data set, not even for simulated data when a very large number of samples is available...For this reason, we lower the bar from the beginning by not aiming to infer the entire network, instead, [inferring] the strongest interactions among covariates only." (p. 2)

In fact, we attribute part of our success in replicating bridge relations to focusing on the strongest edges. Hence, we recommend that researchers focus on large effects when transitioning from exploratory to confirmatory analyses.

1.6. Conclusion

This work demonstrated that confirmatory testing can be woven into the very fabric of network analysis and theory. The ideas presented in this paper provide the foundation from which to begin comparing formalized expectations related to the (in)dependence structure of psychological constructs and mental disorders. We hope this bridges the gap between between hypothesis generation and testing in psychological networks. The testing strategy is implemented in the `R` package **BGGM**. A detailed tutorial is available on the Open Science Framework.

CHAPTER 2

Painless Posterior Sampling: Bayesian Bootstrapped Correlation Coefficients

2.1. Introduction

Correlation coefficients lie at the heart of research in the social–behavioral sciences (Chen et al., 2002; Cohen et al., 2013). They quantify the degree of association between variables, where hypotheses are often posited as correlational statements such as "there is a positive association between IQ and educational attainment." The most frequently used variant is the Pearson product–moment correlation, or Pearson correlation, that quantifies the strength of the linear association between two variables. Values of 1, -1, and 0, respectively, imply a perfectly positive, perfectly negative, and no relationship.

Although they play a leading role in psychological research, there is surprisingly little work done on estimating common correlation types in a Bayesian framework. To date, the Pearson correlation has received the bulk of attention (e.g., Mulder, 2016; Wagenmakers, Verhagen, et al., 2016; Wetzels & Wagenmakers, 2012), but research examining alternative types of correlations are scarce. This is unsurprising because the Pearson correlation is the most frequently used measure of association and it is also trivially estimated, say, by following the separation strategy of Barnard et al. (2000) or using the natural conjugate prior for the covariance matrix in a Gaussian model (i.e., the inverse-Wishart). Nevertheless, there are times when researchers would like to estimate a different type of correlation that may be better suited for their data. For example, Kendall's τ is a popular rank-based correlation method, but was not possible to estimate in a Bayesian framework until only recently (van Doorn et al., 2018; Yuan & Johnson, 2008). There are a variety of reasons for why this is the case, for instance, due to the lack of an explicit likelihood function and sensible choices for prior distributions (Yuan & Johnson, 2008). Furthermore, polychoric correlations, that are commonly used for ordinal data, can be challenging to implement and computationally expensive

to estimate (e.g., Lawrence et al., 2008). One such approach is the multivariate probit model (e.g., Albert, 1992; Chib & Greenberg, 1998), but this requires sampling latent (Gaussian) data and thresholds, both of which are not straightforward. These methodological challenges have resulted in a lack of software for estimating Bayesian correlations.

To overcome these hurdles, we propose the Bayesian bootstrap (BB, Rubin, 1981) as a simple and flexible approach to obtain a posterior distribution for a correlation matrix. This method is attractive in the sense that it avoids the direct specification of a prior and is straightforward to implement because it is operationally equivalent to the classical bootstrap (Efron, 1979). The key difference between them is that the BB attaches weights to the observed values from a uniform Dirichlet distribution, as opposed to the classical bootstrap that resamples the data. The main benefit of this weighting scheme is that the resulting samples can be used to approximate the posterior distribution of interest under a noninformative prior (Lo, 1987, 1988; Lyddon et al., 2019; Weng, 1989). The motivation behind the BB is nicely summarized in Kim and Lee (2003),

> "To circumvent such complications of the full Bayesian analysis, we propose Bayesian bootstrap (BB) procedures which, we believe, are easily accessible to practitioners and at the same time are reliable inference procedures...the BB procedures are conceptually parametric and conceptually simple but retain the flexibility of nonparametric models. Another advantage of the BB procedures is that it is unnecessary to elicit prior information..." (p. 1905)

Because the BB is flexible and does not require a prior to be explicitly specified by the analyst, it can be used to seamlessly estimate virtually any correlation matrix, including Kendall's τ and polychoric correlations. However, the BB remains relatively unknown in psychological contexts despite its simple form and utility with respect to simulating samples from the posterior distribution.

Naturally, a key attraction of the BB is that it shares important properties with traditional Bayesian inference.The benefits of adopting Bayesian approaches have been written about extensively in the psychological sciences (see e.g., Vandekerckhove et al., 2018, and other articles in that special issue). For instance, analysts commonly want to make statements about which parameter values are the most likely conditional on the observed data (Kruschke, 2018; Kruschke et al., 2012), but this privilege is reserved for Bayesian methods as opposed to classical inferential techniques.

Consequently, adopting a Bayesian approach necessarily results in a posterior distribution, and thus, statements can be made about the probability of specific parameter values, or a range of them (Wagenmakers et al., 2018; Wagenmakers, Morey, et al., 2016). Moreover, Bayesian inference allows for quantifying evidence in favor of a null hypothesis as opposed to more classical methods which typically only allow for (failing to) reject the null hypothesis.

Because the Bayesian bootstrap provides a valid posterior, it can be further employed to compare correlations. The problem of comparing correlations from the same sample has received ample attention in the literature (Dunn & Clark, 1969; X.-l. Meng et al., 1992; Mulder, 2016; Raghunathan et al., 1996; Steiger, 1980; Zou, 2007), and there are three main cases where comparing correlations is of interest (Krishnamoorthy & Xia, 2007): (1) overlapping dependent correlations, (2) non-overlapping dependent correlations, and (3) independent correlations from independent samples. Because the dependence structure is encoded in the posterior distribution, the BB can be employed in all of these situations.

2.1.1. Major Contributions. This work includes three major contributions. First, the Bayesian bootstrap is introduced as a method for approximating posterior distributions for several correlation coefficients. Namely, we describe the Bayesian bootstrap for the Pearson correlation, wherein the Spearman's and Gaussian rank correlations naturally arise as special cases. We further provide formulations to obtain Kendall's and polychoric correlation coefficients. We emphasize that these latter two coefficients, unlike the Spearman's and Gaussian rank correlations, cannot be trivially estimated in a Bayesian framework. Second, an approach is discussed for comparing two or more correlations, possibly with the region of practical equivalence (ROPE) of Kruschke (2018). This allows researchers to go beyond merely estimating correlations to making meaningful comparisons among them (e.g., establishing null associations). Third, to increase the availability of the proposed approach, Bayesian bootstrapped correlations have been implemented in the R package **BBcor**. For users who are unfamiliar with R, we have implemented a Shiny app[1] (Chang et al., 2021). The totality of these contributions places the Bayesian bootstrap into the toolbox of researcher psychologists.

[1]The Shiny app can be accessed at tinyurl.com/2nw33cu8

2.1.2. Overview. The outline of this article is as follows. We begin by delineating the Bayesian bootstrap procedure for different correlation types. Here it is shown how estimating correlations with the BB essentially amounts to calculating weighted correlations. Next, we demonstrate how two or more correlations can be compared with the resulting posterior distribution. We then move on to empirical illustrations of the method using two psychological datasets. These examples illustrate the utility of the proposed method in applied settings. We conclude with a brief discussion on the Bayesian bootstrap.

2.2. The Bayesian Bootstrap

There are at least three ways to view the Bayesian bootstrap (Kim & Lee, 2003): 1) as an extension of the classical bootstrap, 2) the limit of the full Bayesian posterior as the prior becomes completely uninformative (Gasparini, 1995, Theorem 2), and 3) a distribution that is proportional to the product of the empirical likelihood and an uninformative prior (Choudhuri, 1998; Lazar, 2003; Owen, 1990; Rubin, 1981). Because in psychology, most analysts are likely to have at least some familiarity with the classical bootstrap, we briefly describe this perspective here. Suppose $Y = (y_1, \ldots, y_n)$ is a random sample from an unknown distribution F and we are interested in estimating a functional of F, $T(F)$, say, the expected value of Y. The classical bootstrap entails resampling the data with replacement to obtain $Y_1^*, \ldots, Y_B^*$ where B is the number of bootstrap samples. Inferences are then drawn on the basis of $T(F_i^*)$, where F_i^* is the empirical distribution of the ith resampled dataset. Notice that the empirical distribution can be expressed as $F_i^* = \sum_j^n w_j \delta_{Y_j}$ where $n(w_1, \ldots, w_n) \sim \text{Multinomial}(n, 1/n, \ldots, 1/n)$. The weights w are discrete, considered to be known, and denote the proportion a distinct value of the original data, δ_{Y_j}, arises in the bootstrap sample. By instead considering the weights for each sample to be unknown, continuous, and distributed as $\text{Dirichlet}(1, \ldots, 1)$, the resulting empirical distribution F_i^* takes on a smoother shape (see Figure 1 in Rubin, 1981). Technical details of the connection between the Bayesian bootstrap and the usual posterior distribution are given in the appendix. For comprehensive mathematical treatments of the BB, we refer readers to Lo (1987, 1988), Newton and Raftery (1994), and the references therein.

2.2.1. Illustration. To illustrate the process of obtaining a BB posterior, suppose that we have n observations of a random variable Y. The BB generates a posterior probability for each

observation $y_1, \ldots, y_n$, where unobserved values have zero posterior probability. Specifically, one BB sample is obtained by drawing n weights from a uniform Dirichlet distribution and attaching them to the data. The generated weights can be interpreted as the probabilities that $Y = y_i$ in each sample (Rubin, 1981). In practice, these weights are easily generated using draws from an exponential distribution (see e.g., Devroye, 1986, p. 594). If this process is repeated S times, then the distribution of all S samples is the BB distribution of Y. More often, however, we are interested in estimating the parameter of a distribution, say, the mean. For each s sample ($s = 1, \ldots, S$), the steps for estimating the mean of Y are as follows:

(1) Draw n exponential variates

$$z_i^{(s)} \sim \text{Exp}\,(1)\,, \quad i = 1, \ldots, n \tag{2.1}$$

(2) Generate the weights

$$w_i^{(s)} = \frac{z_i^{(s)}}{\sum_{i=1}^{n} z_i^{(s)}} \tag{2.2}$$

(3) Calculate the weighted sample mean

$$\bar{y}^{(s)} = \sum_{i=1}^{n} w_i^{(s)} y_i \tag{2.3}$$

The empirical distribution of $\{\bar{y}^{(1)}, \ldots, \bar{y}^{(S)}\}$ is the BB approximation to the posterior of the mean of Y. A visual comparison between a BB distribution and an analytical posterior for this scenario is shown in Figure 2.1. Note steps 1 and 2 can be merged if a uniform Dirichlet distribution random number generator is directly available, as is the case in many programming platforms. Further, a subscript can be added $\bar{y}_g^{(s)}$ ($g = 1, \ldots, G$) in each step to distinguish means between groups. This opens up the possibility to obtain a posterior distribution for mean differences (e.g., $\delta^{(s)} = \bar{y}_1^{(s)} - \bar{y}_2^{(s)}$). In what follows, we demonstrate how these ideas can be harnessed to estimate and compare a variety of correlation coefficients.

2.2.2. Pearson, Spearman's, and Gaussian Rank Correlation Coefficients.

2.2.2.1. *Background.* The most popular correlation is the Pearson product-moment correlation coefficient, or Pearson correlation, which captures the linear relationship between two variables.

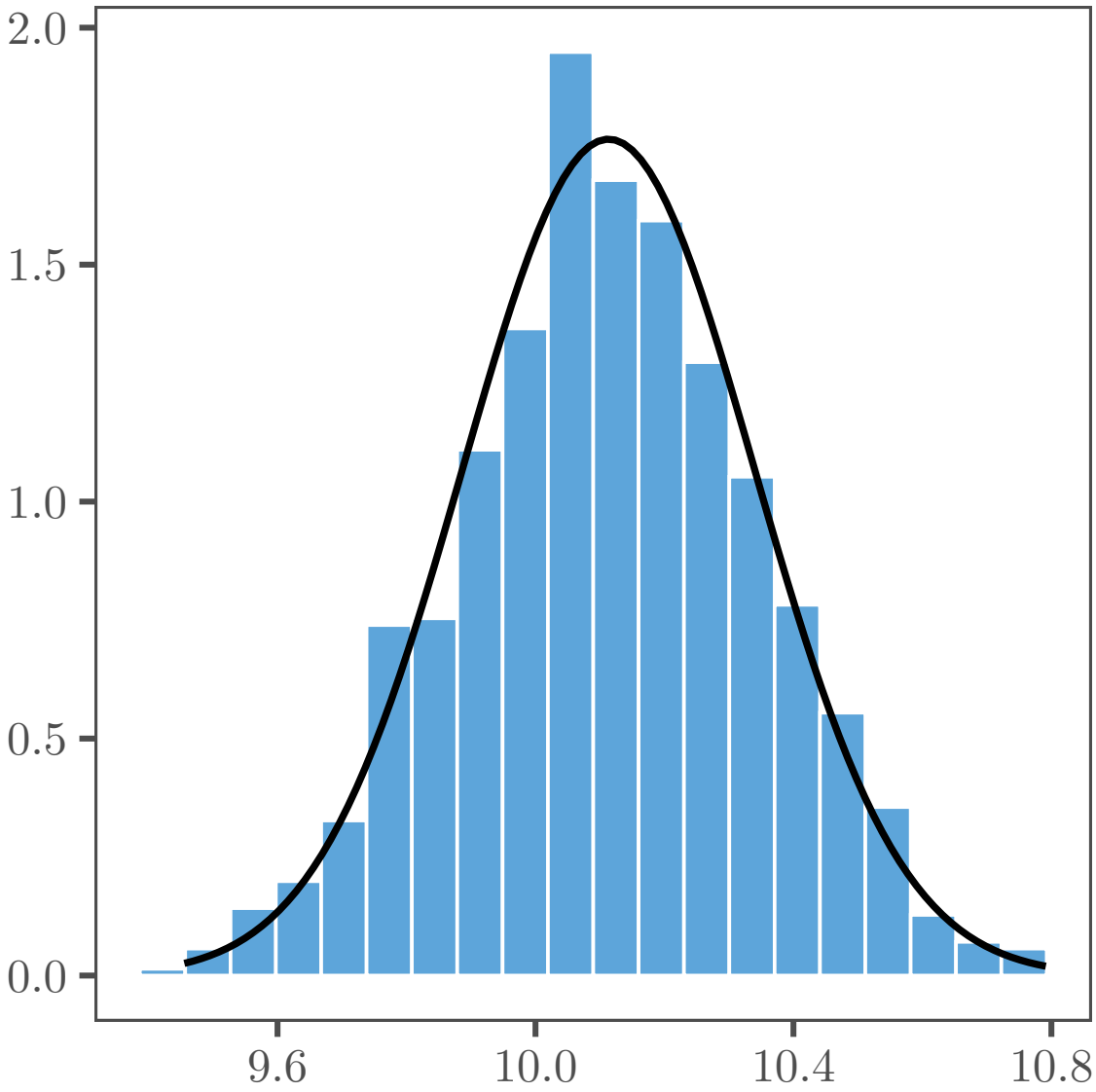

FIGURE 2.1. Density histogram of 1000 Bayesian bootstrapped means using the steps outlined in (2.1) – (2.3). The black line is the posterior density of the mean resulting from a normal prior with mean equal to zero and standard deviation equal to 10. The sample data ($n = 500$) was generated from a Normal distribution with mean equal to 10 and standard deviation equal to 5.

When the data are ordinal, it is common to use the nonparametic Spearman's correlation, which is defined as the Pearson correlation between the ranks of two variables and describes their monotonic

relationship. Although conceptually easy to understand and compute, using Spearman's correlation results in a small loss of statistical efficiency. A recently proposed alternative is the Gaussian rank correlation (Boudt et al., 2012). The Gaussian rank correlation is defined as the Pearson correlation of the normalized ranks (i.e., their Van der Waerden scores). The advantage of normalizing the variables prior to computing their correlation is that there is a small gain in statistical efficiency (for Gaussian data) when estimating the monotonic relationship between them. Interestingly, the main difference between the Pearson's, Spearman's, and Gaussian rank correlations is whether the raw, ranked, or normalized rank observations are being correlated. Hence, only a formulation for the Pearson correlation is needed to obtain any of the three correlation types. Note that Rubin (1981) described the BB for a single Pearson's correlation, but did not consider the full correlation matrix or other correlation types.

In a Bayesian framework, the Pearson correlation matrix is traditionally estimated by modeling the covariance matrix $\boldsymbol{\Sigma}$. To this end, the legacy inverse-Wishart prior has been the de facto standard. Due to its conjugacy, computation can be relatively efficient and thus it is widely implemented in Bayesian software (e.g., Plummer, 2003). However, the inverse-Wishart prior has been criticized for several reasons: the uncertainty for all variances is controlled by a single degrees of freedom parameter (Barnard et al., 2000), the marginal distribution for the variances have low density near zero (Gelman, 2006), and there is an *a priori* dependence between the resulting correlations and variances (Tokuda et al., 2011). Separation strategies exist to deal with the dependence between the variances and correlations (e.g., Barnard et al., 2000), but suffer from similar problems as the inverse-Wishart. Alternative distributions exist that circumvent these issues, such as the LKJ (Lewandowski et al., 2009) or matrix-F (Mulder & Pericchi, 2018) prior distributions. Although they are more flexible than the inverse-Wishart, the incurred expense is that they are more computationally complex and, additionally, are not yet widely available in Bayesian software. For instance, the LKJ prior is mostly restricted to programs that interface with Stan (Carpenter et al., 2017) and do not readily provide the full correlation matrix. The matrix-F prior has been implemented for a full correlation matrix, but first requires estimating the partial correlations and thus the prior cannot be placed directly over the correlation matrix (Williams et al., 2019). By

instead employing the Bayesian bootstrap, an approximate posterior for the full correlation matrix can be obtained painlessly.

2.2.2.2. *Bayesian Bootstrap Steps.* We now describe the necessary ingredients for obtaining Bayesian bootstrapped samples of Pearson, Spearman's, and Gaussian rank correlations. Without a loss of generality, assume $\mathbf{Y}$ to be a mean-centered $n \times p$ data matrix with sample covariance matrix $\mathbf{S}$. The Pearson correlation matrix for $\mathbf{Y}$ is given by

$$\mathbf{R} = \mathbf{D}^{-\frac{1}{2}}\mathbf{S}\mathbf{D}^{-\frac{1}{2}} \tag{2.4}$$
$$\mathbf{S} = (n-1)^{-1}\left(\mathbf{Y}'\mathbf{Y}\right)$$

where $\mathbf{D}^{-\frac{1}{2}}$ is a diagonal matrix containing the inverse square roots of the diagonal elements of $\mathbf{S}$ and each r_{ij} element of $\mathbf{R}$ indicates the correlation between the ith and jth column of $\mathbf{Y}$. The Spearman's correlation matrix is obtained when each (i, j)th element in $\mathbf{Y}$ is replaced with its rank, $R(Y_{ij})$. Similarly, if the elements are replaced with their Van der Waerden scores, $\Phi^{-1}\left(\frac{R(Y_{ij})}{n+1}\right)$, where Φ^{-1} denotes the quantile function for a standard normal curve, then $\mathbf{R}$ contains the Gaussian rank correlations.

A simple modification of (2.4) yields a posterior sample of $\mathbf{R}$. Mainly, for each s sample, the values drawn from the uniform Dirichlet distribution are used to center the columns of $\mathbf{Y}$ at their *weighted* mean, and then further to obtain a *weighted* covariance matrix $\mathbf{S}_{\text{w}}$. With this modification, a Bayesian bootstrapped sample for $\mathbf{R}$ is obtained by computing

$$\mathbf{R}_{\text{w}}^{(s)} = \left(\mathbf{D}_{\text{w}}^{(s)}\right)^{-\frac{1}{2}} \mathbf{S}_{\text{w}}^{(s)} \left(\mathbf{D}_{\text{w}}^{(s)}\right)^{-\frac{1}{2}} \tag{2.5}$$
$$\mathbf{S}_{\text{w}}^{(s)} = \left[1 - \sum_{i=1}^{n}\left(w_i^{(s)}\right)^2\right]^{-1} \left(\mathbf{Y}_{\text{w}}^{(s)\prime}\mathbf{Y}_{\text{w}}^{(s)}\right) \tag{2.6}$$
$$\mathbf{Y}_{\text{w}}^{(s)} = \mathbf{Y} \circ \mathbf{w}_*^{(s)}\mathbf{1}_p' \tag{2.7}$$

where $\mathbf{R}_{\text{w}}^{(s)}$ is a weighted correlation matrix, $\left(\mathbf{D}_{\text{w}}^{(s)}\right)^{-\frac{1}{2}}$ is a diagonal matrix containing the inverse square roots of the diagonal elements of $\mathbf{S}_{\text{w}}^{(s)}$, and $\mathbf{Y}_{\text{w}}^{(s)}$ is a weighted version of the data matrix.

The symbol "$\circ$" denotes the Hadamard product, $\mathbf{w}_*^{(s)}$ is an n-dimensional vector with elements $w_{*,i}^{(s)} = \sqrt{w_i^{(s)}}$, and $\mathbf{1}_p$ is a p-dimensional vector containing 1's. If $\mathbf{R}_{\text{w}}^{(s)}$ is computed S times, then the distribution of $\{\mathbf{R}_{\text{w}}^{(1)}, \ldots, \mathbf{R}_{\text{w}}^{(S)}\}$ is the BB distribution of $\mathbf{R}$. Similarly, the BB distribution of each r_{ij} is the empirical distribution of $\{r_{\text{w},ij}^{(1)}, \ldots, r_{\text{w},ij}^{(S)}\}$. Notice that computing a posterior sample with the BB requires only a few steps and does not involve explicitly invoking a prior distribution. In this way, the Bayesian bootstrap provides a seamless method for obtaining posterior distributions for the Pearson, Spearman's, and Gaussian rank correlation matrices.

2.2.3. Kendall's Rank Correlation Coefficient.

2.2.3.1. *Background.* A similar approach can be taken to obtain posterior samples for Kendall's rank correlation coefficient (Kendall, 1938), or Kendall's τ, a widely used measure of association in nonparametric statistics. Like Spearman's correlation, it is a robust measure that captures monotonic relationships between two variables, but has some advantages. It is asymptotically more efficient and has an appealing interpretation. Kendall's τ can be interpreted as follows. Suppose we have n observations for two random variables X and Y. A pair of differences $(x_i - x_j)$ and $(y_i - y_j)$ is said to be concordant if they share the same sign and discordant if they do not. Kendall's τ is obtained by taking the difference between concordant and discordant pairs and dividing this quantity by the number of all possible pairs. When $\tau = 1$ (-1) all pairs of observations are concordant (discordant).

Despite its popularity, there is a dearth of literature on Bayesian inference for Kendall's rank correlation. The main reason for this is that nonparametric tests in Bayesian settings have historically been limited by a lack of prior distributions and an explicit likelihood function (Yuan & Johnson, 2008) — without which a model cannot be formulated in a Bayesian framework. Recently, van Doorn et al. (2018) developed a method for deriving a posterior distribution for Kendall's τ based on its standardized test statistic T^*. However, this method only considers a single correlation at a time. That is, the full correlation matrix is not readily estimated, which, in turn, prevents easily comparing correlations. In contrast, a Bayesian bootstrap approach to estimating Kendall's τ circumvents this concern because it readily estimate the full correlation matrix.

2.2.3.2. *Bayesian Bootstrap Steps.* For the case of X and Y, Kendall's τ is defined as

$$\tau = \frac{\sum_{1 \leq i < j \leq n}^{n} \operatorname{sgn}(x_i - x_j) \operatorname{sgn}(y_i - y_j)}{k}, \tag{2.8}$$

where $k = \frac{n(n-1)}{2}$ is the number of distinct pairs. The above is commonly referred to as τ_A and does not account for ties. When ties are present, the denominator is adjusted to correct for this and is defined as $\sqrt{(k - t_x)(k - t_y)}$ where t_x and t_y denote the number of ties in X and Y, respectively. This version is commonly known as τ_B and because this is the version we consider here, we simply refer to it as τ.

A Bayesian bootstrapped sample for Kendall's rank correlation between X and Y can be computed by first drawing values from a uniform Dirichlet distribution and weighting the numerator to obtain

$$\tau_{\mathrm{w}}^{(s)} = \sum_{1 \leq i < j \leq n}^{n} w_i^{(s)} w_j^{(s)} \operatorname{sgn}(x_i - x_j) \operatorname{sgn}(y_i - y_j). \tag{2.9}$$

This expression is nice because the connection to the original formulation is clear, but it can be generalized to obtain the full correlation matrix (Pozzi et al., 2012, pp. 15 – 17). Let $\mathbf{Z}$ be a $k \times p$ matrix where each (l, v) element is associated to $\operatorname{sgn}(y_i^v - y_j^v)$ where $l = 1, \ldots, k$, $v = 1, \ldots, p$, $i = 2, \ldots, n$, and $j = 1, \ldots, n-1$, or in words, a matrix where each element indicates the sign for the difference of the observation pair (i, j) on variable v. With this definition of $\mathbf{Z}$, a Bayesian bootstrap sample for the Kendall's τ correlation matrix can be obtained as follows

$$\boldsymbol{\tau}_{\mathrm{w}}^{(s)} = \left(\mathbf{D}_{\mathrm{w}}^{(s)}\right)^{-1/2} \mathbf{S}_{\mathrm{w}}^{(s)} \left(\mathbf{D}_{\mathrm{w}}^{(s)}\right)^{-1/2} \tag{2.10}$$

$$\mathbf{S}_{\mathrm{w}}^{(s)} = \mathbf{Z}^{(s)\prime} \mathbf{Z}_{\mathrm{w}}^{(s)} \tag{2.11}$$

$$\mathbf{Z}_{\mathrm{w}}^{(s)} = \mathbf{Z} \circ \mathbf{w}_*^{(s)} \mathbf{1}_p' \tag{2.12}$$

where $\boldsymbol{\tau}_{\mathrm{w}}^{(s)}$ is the weighted correlation matrix, $\left(\mathbf{D}_{\mathrm{w}}^{(s)}\right)^{-1/2}$ is a diagonal matrix containing the inverse square roots of the diagonal elements of $\mathbf{S}_{\mathrm{w}}^{(s)}$, and $\mathbf{Z}_{\mathrm{w}}^{(s)}$ is a weighted version of $\mathbf{Z}$. The k-dimensional vector $\mathbf{w}_*^{(s)}$ contains the elements $w_{*,l}^{(s)} = \sqrt{w_i^{(s)} w_j^{(s)}}$, and $\mathbf{1}_p$ is p-dimensional vector containing 1's. When written this way, Kendall's rank correlation can be conceptualized as a Pearson correlation computed with $\mathbf{Z}$. If $\boldsymbol{\tau}_{\mathrm{w}}^{(s)}$ is computed S times, then the distribution of $\{\boldsymbol{\tau}_{\mathrm{w}}^{(1)}, \ldots, \boldsymbol{\tau}_{\mathrm{w}}^{(S)}\}$ is the BB

distribution of $\boldsymbol{\tau}$. Like each $r_{\mathrm{w},ij}$, the BB distribution of each $\tau_{\mathrm{w},ij}$ is their empirical distribution over all S samples.

2.2.4. Polychoric Correlation Coefficient.

2.2.4.1. *Background.* An important measure of association in the field of psychometrics is the polychoric correlation coefficient (Jöreskog, 1994; Olsson, 1979). Like correlations that describe monotonic relationships, the polychoric correlation is often used with ordinal data. The key difference here is that the ordinal data are considered to be the result of discretizing continuous variables. Accordingly, the polychoric correlation captures the linear association between two latent continuous variables underlying the observed ordinal data. Note that we assume that the joint distribution of the two latent variables is Gaussian, but other distributions can be used (e.g., bivariate t, Albert, 1992).

Getting Bayesian estimates of polychoric correlations can be difficult due to their computational expense. Computing the likelihood requires iteratively sampling from truncated Gaussian distributions and the covariance matrix is typically restricted to be a correlation matrix for identifiability reasons (Albert, 1992; Chib & Greenberg, 1998). Further, nuisance parameters, termed thresholds, must be estimated for each variable. More efficient MCMC algorithms have been developed, for example, by using parameter expansion for data augmentation (Lawrence et al., 2008; Talhouk et al., 2012) or parameterising the precision matrix of the latent variables in terms of the Cholesky decomposition (Webb & Forster, 2008), but these techniques introduce computational complexities of their own and remain unavailable in statistical software (to our knowledge). Thus, for polychoric correlations, the Bayesian bootstrap again provides a relatively simple solution.

2.2.4.2. *Bayesian Bootstrap Steps.* For ease of exposition, we focus on estimating the polychoric correlation between two variables, but the following can be applied for the entire correlation matrix. Suppose that two ordinal variables X and Y are expressed in a two-way contingency table with R rows and C columns. That is, there are R levels in X and C levels in Y. If the data is collected on n individuals and classified with respect to the rows and columns, then the cell counts, n_{rc} $(r = 1, \ldots, R,\ c = 1, \ldots, C)$ have respective probabilities π_{rc}. The typical estimation approach is then to assume that the ordinal variables correspond to continuous Gaussian variables ξ and η. The n pairs (ξ_i, η_i) can likewise be placed in an $R \times C$ contingency table using row thresholds $-\infty =$

$a_0 < a_1 < \cdots < a_{R-1} < a_R = \infty$ and column thresholds $-\infty = b_0 < b_1 < \cdots < b_{C-1} < b_C = \infty$. The relationship between X and ξ is

$$x_i = \begin{cases} 1 & \text{if } \xi_i < a_1 \\ 2 & \text{if } a_1 \leq \xi_i < a_2 \\ \vdots & \\ R & \text{if } a_{R-1} \leq \xi_i \end{cases}, \tag{2.13}$$

and similarly for Y and η.

The polychoric correlation can then be estimated in two steps (Olsson, 1979). The thresholds are first estimated as

$$a_r = \Phi^{-1}\left(\frac{\sum_{i=1}^{n} \mathbb{I}(x_i \leq r)}{n}\right), \quad r = 1, \ldots, R-1 \tag{2.14}$$

$$b_c = \Phi^{-1}\left(\frac{\sum_{i=1}^{n} \mathbb{I}(y_i \leq c)}{n}\right), \quad c = 1, \ldots, C-1, \tag{2.15}$$

where Φ denotes the bivariate standard normal cumulative density function with correlation ρ and the symbol $\mathbb{I}(\cdot)$ denotes the indicator function. Then, the likelihood of the sample

$$\sum_{r=1}^{R}\sum_{c=1}^{C} n_{rc} \ln \pi_{rc} \tag{2.16}$$

is maximized with respect to ρ. Above, n_{rc} is the number of observations in the (r, c)th cell of the contingency table and π_{rc} is the probability that (ξ_i, η_i) belongs to that cell

$$\begin{aligned} \pi_{rc} = & \\ & \Phi(a_r, b_c) - \Phi(a_{r-1}, b_c) - \\ & \Phi(a_r, b_{c-1}) + \Phi(a_{r+1}, b_{c+1}). \end{aligned} \tag{2.17}$$

The value of ρ that maximizes the log-likelihood is the estimate for the polychoric correlation between X and Y.

A Bayesian bootstrapped sample of the polychoric coefficient can be obtained through a reweighting scheme applied to the $R \times C$ contingency table. To obtain the weighted cell probabilities, the thresholds are first estimated based on the simulated Dirichlet weights (Bailey et al., 2018)

$$a_{\mathrm{w},r}^{(s)} = \Phi^{-1}\left(\sum_{i=1}^{n} w_i^{(s)} \mathbb{I}(x_i \leq r)\right) \tag{2.18}$$

$$b_{\mathrm{w},c}^{(s)} = \Phi^{-1}\left(\sum_{i=1}^{n} w_i^{(s)} \mathbb{I}(y_i \leq c)\right). \tag{2.19}$$

Similarly, the term n_{rc} in (2.16) is replaced with

$$n_{\mathrm{w},rc}^{(s)} = \sum_{i=1}^{n} w_i^{(s)} \mathbb{I}(x_i = r)\mathbb{I}(y_i = c). \tag{2.20}$$

The weighted probabilities for each sample $\pi_{\mathrm{w},rc}^{(s)}$ are computed using the expression in (2.17), but with the weighted thresholds so that the log-likelihood for each sample is given by

$$\sum_{i=1}^{s}\sum_{j=1}^{r} n_{\mathrm{w},rc}^{(s)} \ln \pi_{\mathrm{w},rc}^{(s)}. \tag{2.21}$$

Finally, the Bayesian bootstrapped sample for the polychoric correlation, $\rho^{(s)}$, is the one that maximizes (2.21). If this procedure is carried out S times, then $\{\rho^{(1)}, \ldots, \rho^{(S)}\}$ is the BB distribution of the polychoric correlation between X and Y.

2.2.5. Comparing Correlations. Once a set of correlations has been estimated, a common next step is to make comparisons among them, say, to determine which association is the largest. This can be done by computing the posterior distribution for comparisons of interest. The main advantage of doing so is that standard deviations (analogous to standard errors) are available in situations where they would otherwise be difficult to obtain (e.g., the difference between two

polychoric correlations with the same matrix). Fortunately, the Bayesian bootstrapped posterior distribution can be used to make such comparisons.

Using the Bayesian bootstrap, the posterior can be obtained for linear combinations of correlations by manipulating the posterior samples of the individual correlations. Say we have estimated a $p \times p$ correlation matrix and are interested in their pairwise differences. Let $\boldsymbol{\rho}^{(s)}$ be a vector containing the sth sample for the $G = p(p-1)/2$ distinct correlations and $\mathbf{C}$ be a matrix of coefficients capturing the pairwise differences. Each element of $\mathbf{C}$ is either a 1, -1, or 0. A posterior sample for these differences can be obtained by expressing them as a linear combination

$$\boldsymbol{\delta}^{(s)} = \mathbf{C}\boldsymbol{\rho}^{(s)} \tag{2.22}$$

$$\boldsymbol{\rho}^{(s)} = \begin{bmatrix} \rho_1^{(s)} \\ \rho_2^{(s)} \\ \vdots \\ \rho_G^{(s)} \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} 1 & -1 & 0 & \dots & 0 \\ 1 & 0 & -1 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & 1 & -1 \end{bmatrix}. \tag{2.23}$$

The distribution of all $\{\boldsymbol{\delta}^{(1)}, \dots, \boldsymbol{\delta}^{(S)}\}$ approximates the posterior distribution for the comparisons between the correlations in $\boldsymbol{\rho}$. Now, means, standard deviations, and credible intervals can be computed directly for the posterior of $\boldsymbol{\delta}$. The subscripts of the G correlations can denote distinct correlations within the same group, the same correlation for distinct groups, or distinct correlations from distinct groups. Although we focused on pairwise differences here, this idea can be extended to more general linear combinations.

An additional advantage of Bayesian analysis is the ability to "accept" parameter values that provide support either for or against a null hypothesis. For instance, if one wants to conclude that there is no difference between the magnitude of two correlations, then this can be done using a formal procedure such as the region of practical equivalence (ROPE) approach (Kruschke, 2018). The ROPE approach is similar in spirit to a frequentist approach wherein a prespecified parameter value is rejected if it is not covered by a $100(1-\alpha)\%$ confidence interval. The difference is that a range of parameter values (i.e., a ROPE) is stipulated where values in this range are treated as as equivalent to a null value (e.g., 0). Once this region is established and the posterior distribution of

$\boldsymbol{\delta}$ has been computed, a $100(1-\alpha)\%$ credible interval (CrI) can be constructed for each comparison. If the computed interval lies entirely inside the ROPE bounds, then the estimated parameter value is treated as equivalent to the null value and conversely, if the interval completely excludes the ROPE, then the null value is rejected. This is because a $100(1-\alpha)\%$ CrI contains the $100(1-\alpha)\%$ most probable values (assuming a symmetric distribution). Thus, if the CrI is entirely inside of the ROPE, then we can interpret the parameter value as being practically equivalent to the null value and vice versa. A decision is withheld if there is overlap between the interval and the ROPE. Adopting the framework described above permits researchers to utilize the BB to make meaningful comparisons between associations using a variety of different correlation coefficients with the goals of either parameter estimation, making decisions about a parameter value, or both.

2.2.6. Summary. In this section, we described how posterior distributions for several different correlations can be obtained in a straightforward manner via the Bayesian bootstrap. The central theme was that simulating posterior samples for correlations boils down to repeatedly calculating weighted correlations where the weights are uniform Dirichlet distributed. In each iteration of the bootstrap, the resulting weighted correlation constitutes a draw from the correlation's posterior distribution, and when done repeatedly, the distribution of the calculated statistics approximates the posterior of interest. The main advantage of this method is that posterior inference for correlations can be done "painlessly". That is, obtaining BB estimates for the correlations does not require specifying a prior distribution or complex sampling techniques. Altogether, the BB provides a powerful tool for approximate Bayesian inference of popular correlation types in social-behavior sciences.

2.3. Empirical Application

Below we discuss an empirical example where we illustrate how the **BBcor**[2] package can be applied to obtain and compare Bayesian bootstrapped correlations in practice. We utilize data that were first analyzed in Šrol et al., 2021 to compare dependent correlations from the same sample. The data were collected to study the negative social consequences of Covid-19 related conspiracy beliefs. Slovakian participants ($N = 501$) completed survey items measuring their prejudiced

[2]The **BBcor** package can be downloaded from CRAN or from https://github.com/donaldRwilliams/BBcor.

and discriminatory views against three social outgroups associated with the pandemic in Slovakia. Specifically, data were collected on negative feelings, social distance, and discriminatory views towards Chinese, Roma, and Italian people. Further, measurements were taken on the degree of belief in general Covid-19 conspiracies (e.g., "Covid-19 is a biological weapon intended to eliminate the overcrowded human population") and Chinese-specific Covid-19 conspiracies (e.g., "the Chinese created [SARS-CoV-2] as a biological weapon which then got out of hand"). As part of the analysis in this study, the three measures of prejudice and discrimination were each correlated with the measures of conspiracy belief, yielding six correlations per outgroup. The resulting correlations were then compared using Steiger's z-test (Steiger, 1980). For example, the correlation between negative feelings towards Italians and general Covid-19 conspiracy beliefs was compared to the correlation between negative feeling towards Italians and Chinese-specific Covid-19 conspiracies.

There are two details to note here. First, the z-test used to compare correlations makes the assumption that the underlying data are Gaussian. Second, failing to reject the null hypothesis does not provide support in favor of no difference (i.e., absence of evidence is not evidence of absence). Thus, it may be desirable to use a method of comparison that accommodates a measure of association more appropriate for Likert-type data such as the data collected (e.g., Kendall's τ), and that allows for statements in favor of the null hypothesis. This can easily be accomplished with the Bayesian bootstrap methodology outlined in this article.

2.3.1. Calculating the correlations. We assume the reader to have some familiarity with the R programming language (R Core Team, 2021). To begin, the **BBcor** package must be installed and loaded, and the data must be read into R.

```
# install and load BBcor
install.packages("BBcor")
library(BBcor)
# read in data set
data("srol2021")
str(srol2021)
> 'data.frame': 501 obs. of 11 variables:
```

```
> $ neg_feelings_china : int 100 96 75 50 42 68 50 80 ...
> $ social_distance_china : num 7 7 5.33 2.67 1 ...
> $ discrimination_china : int 7 7 5 3 1 6 1 2 3 3 ...
> $ neg_feelings_italy : int 67 50 55 50 68 38 50 20 ...
> . . .
> $ discrimination_roma : int 7 5 3 3 1 7 1 2 2 3 ...
> $ china_Covid_conspiracy : num 2.5 4.25 3.25 2.25...
> $ generic_Covid_conspiracy: num 2.62 3.25 2.75 2.38 ...
```

The Bayesian bootstrapped Kendall's τ correlation matrix for this data is trivially obtained via the `bbcor` function:

```
bb_tau <- bbcor(srol2021, method = "kendall", iter = 1000, cores = 1)
bb_summary <- summary(bb_tau, ci = 0.9, decimals = 2)
```

Here, the `bbcor` function samples the posterior for the correlation matrix, and takes as arguments the data, the desired correlation type, the number of samples to draw, and the number of cores to use when parallel computing is employed. Printing the returned object outputs the mean correlation matrix. Running `summary` on the returned object and specifying the desired credible interval returns a data frame summarising the posterior with means, standard deviations, and bounds for the credible intervals. For instance, previewing the summary object with `head(bb_summary)` prints

```
> Relation Post.mean Post.sd Cred.lb Cred.ub
> 1 neg_feelings_china--social_distance_china 0.16 0.03 0.10 0.21
> 2 neg_feelings_china--discrimination_china 0.19 0.03 0.13 0.24
> 3 social_distance_china--discrimination_china 0.15 0.04 0.09 0.20
> 4 neg_feelings_china--neg_feelings_italy 0.43 0.03 0.38 0.49
> 5 social_distance_china--neg_feelings_italy 0.06 0.04 0.00 0.11
> 6 discrimination_china--neg_feelings_italy 0.17 0.03 0.11 0.22
```

Depending on the precision of of measurements being considered, it can be desirable to obtain more than two decimal points (Cousineau, 2020). This can easily be done by adjusting the `decimals` setting in the `summary` method. The posterior means for the correlations and respective intervals

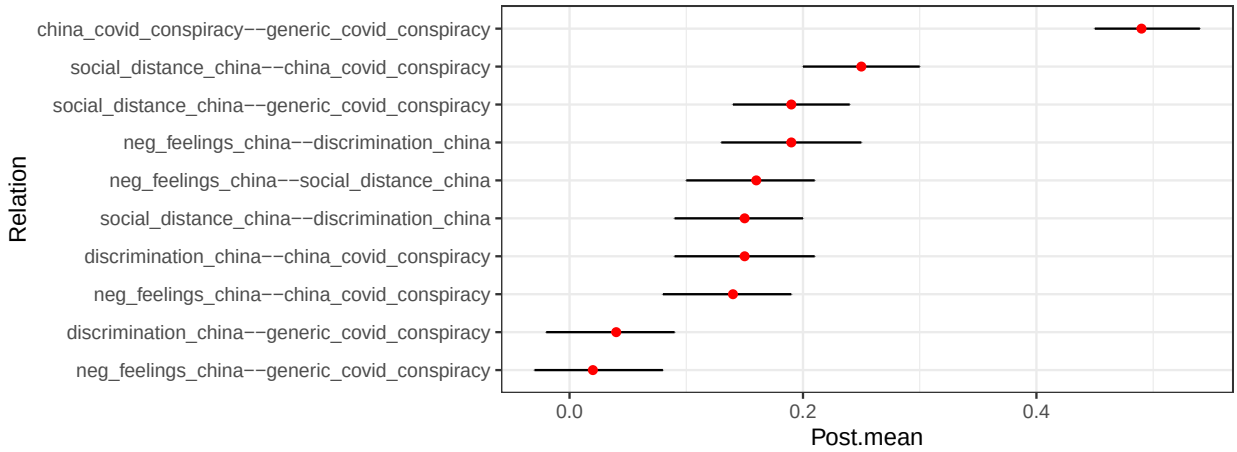

FIGURE 2.2. Output of the `plot` method for Bayesian bootstrapped (BB) correlations obtained with the `bbcor` function. The red dots indicate BB posterior means for the correlations and the bars denote their respective BB 90% credible intervals.

can easily be visualized using syntax from the **ggplot2** library (Wickham, 2016). For example, if we subset the data to only include the prejudice and discrimination measures for China and the two conspiracy theory variables, then the following code returns a plot for the ten resulting correlations which can be seen in Figure 2.2

```
library(ggplot2)
bb_tau_china <- bbcor(Covid_china_subset, method = "kendall")
plot(bb_tau_china) + theme_bw()
```

2.3.2. Analyzing comparisons. The Bayesian bootstrapped correlations can be compared using the `compare` function. The correlations to be compared can be specified either using a character string or by providing a contrast matrix as detailed in Comparing Correlations. For example, if the focus is on comparing the correlation between negative feelings towards China and belief in China-specific Covid-19 conspiracies to the correlation between negative feelings towards China and belief in generic Covid-19 conspiracies, then one can specify the following,

```
comparison <- "neg_feelings_china--china_Covid_conspiracy > neg_feelings_china--
    ↪ generic_Covid_conspiracy"
compare(comparison, obj = bb_tau_china)
```

which yields a summary of the comparison when printed.

```
> Call:
> lin_comb.bbcor(lin_comb = lin_comb, obj = obj, ci = ci, rope = rope,
> contrast = contrast)
> ------
> Combinations:
> C1: neg_feelings_china--china_Covid_conspiracy > neg_feelings_china--
    ↪ generic_Covid_conspiracy
> ------
> Posterior Summary:
>
> Post.mean Post.sd Cred.lb Cred.ub Pr.less Pr.greater
> C1 0.12 0.03 0.07 0.17 0 1
> ------
> Note:
> Pr.less: Posterior probability less than zero
> Pr.greater: Posterior probability greater than zero
```

Above, the `comparison` object is a string that states the comparison to be made is that `neg_feelings_china--china_Covid_conspiracy` is greater than `neg_feelings_china--generic_Covid_conspiracy`. This string is passed along to the `compare` function along with the name of the object containing the correlations. The output displays several summary statistics for the posterior of this comparison such as the mean difference, standard deviation, credible interval bounds, and the proportion of posterior mass that is greater or less than zero. In this case, the difference between the two correlations is 0.12, 90% CrI [0.07, 0.17], and the entirety of the posterior mass is above zero.

Often, analysts are interested in making more than one comparison. For example, Šrol et al. (2021) repeated the same comparison as above for each country (China, Roma, and Italy) and for each measure of prejudice and discrimination. Thus, there were three comparisons made per

country. To avoid tediously typing long character strings, it can be useful to specify a contrast matrix to encode the comparisons of interest. For the subset of variables for China, we must specify a 3×10 matrix corresponding to the three comparisons and ten unique correlations. Additionally, a region of practical equivalence (ROPE) may be stipulated as above, say [-0.10, 0.10]. In R, the analogous code is written as follows

```
contrast_vec <- c(0, 0, 0, 1, 0, 0, -1, 0, 0, 0,
                  0, 0, 0, 0, 1, 0, 0, -1, 0, 0,
                  0, 0, 0, 0, 0, 1, 0, 0, -1, 0 )

contrast_mat <- matrix(constrast_vec, nrow = 3, ncol = 10, byrow = TRUE)

compare(obj = bb_tau_china, contrast = contrast_mat, ci = 0.9, rope = c(-0.10,
    ↪ 0.10))
> ------
> Call:
> lin_comb.bbcor(lin_comb = lin_comb, obj = obj, ci = ci, rope = rope,
> contrast = contrast)
> ------
> Combinations:
> C1: C1
> C2: C2
> C3: C3
> ------
> Posterior Summary:
>
> ROPE: [ -0.1 , 0.1 ]
>
> Post.mean Post.sd Cred.lb Cred.ub Pr.in
> C1 0.12 0.03 0.07 0.17 0.2762
```

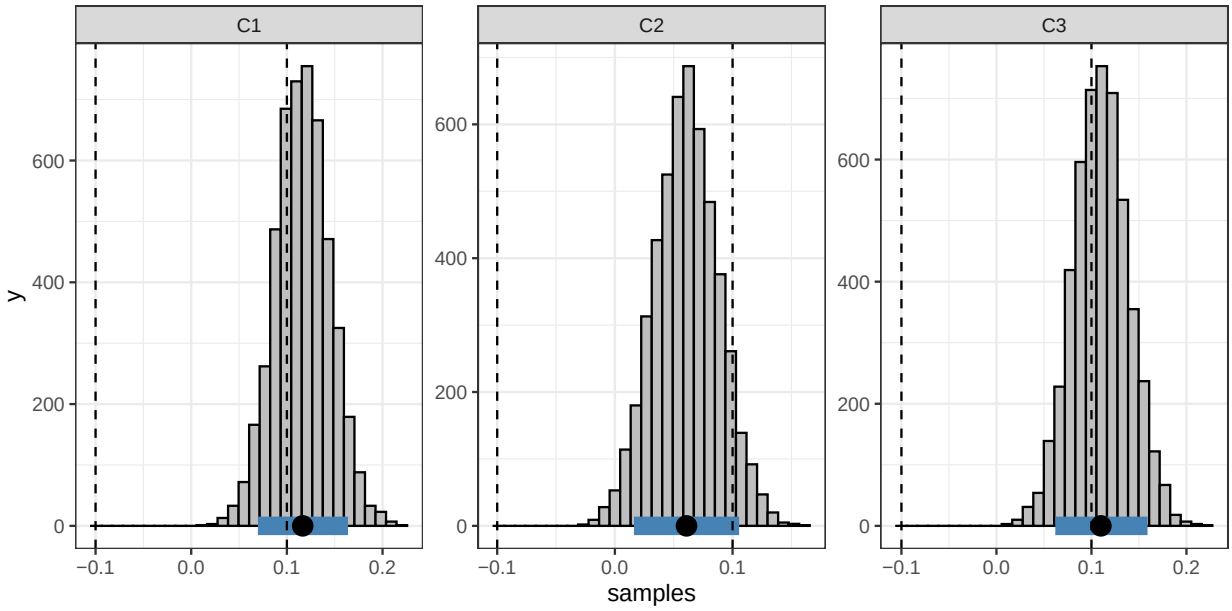

FIGURE 2.3. Output from plotting comparisons with the `compare` function. The histograms represent BB posterior samples for the comparisons, black dots indicate the BB mean, and blue bars denote BB 90% CrIs. The dotted black lines capture the bounds for the ROPE of [-0.1, 0.1].

```
> C2 0.06 0.03 0.02 0.11 0.9162
> C3 0.11 0.03 0.06 0.16 0.3544
> ------
> Note:
> Pr.in: Posterior probability in ROPE
```

In this output, there are three rows of summary statistics, corresponding to the three comparisons specified in `contrast_mat`. The column `Pr.in` contains the proportion of the posterior mass contained in the ROPE. These combinations can also be visualized via a plotting method. If the `compare` object is saved into an object named `china_comparison`, then calling `plot(china_comparison)` produces the plot in Figure 2.3. As can be seen, the intervals for all combinations overlap with the ROPE and thus the evidence is ambiguous as to whether these correlations differ. This is a slightly different conclusion than the original analysis where the null

hypothesis of no difference was rejected for all three combinations at an $\alpha = 0.05$ level. The results for these comparisons, along with the ones for Italy and Roma can be seen in Table 2.1.

	China-specific covid-19 CTs	Generic covid-19 CTs	Mean Difference
Negative Feelings (China)	0.14 [0.09, 0.19]	0.02 [-0.03, 0.07]	0.12 [0.7, 0.17]
Social Distance (China)	0.25 [0.19, 0.30]	0.19 [0.14, 0.24]	0.06 [0.02, 0.11]
Discrimination (China)	0.15 [0.19, 0.20]	0.04 [-0.02, 0.09]	0.11 [0.06, 0.16]
Negative Feelings (Italy)	0.07 [0.02, 0.12]	0.06 [0.00, 0.11]	0.01 [-0.03, 0.06]
Social Distance (Italy)	0.22 [0.17, 0.28]	0.19 [0.14, 0.24]	0.04 [-0.01, 0.08]
Discrimination (Italy)	0.15 [0.10, 0.21]	0.13 [0.07, 0.18]	0.02 [-0.03, 0.07]
Negative Feelings (Roma)	0.06 [0.01, 0.12]	0.14 [0.08, 0.19]	-0.07 [-0.12, -0.03]
Social Distance (Roma)	0.14 [0.08, 0.19]	0.14 [0.08, 0.19]	0.00 [-0.05, 0.04]
Discrimination (Roma)	0.16 [0.10, 0.22]	0.22 [0.16, 0.27]	-0.06 [-0.11, -0.01]

TABLE 2.1. Bayesian bootstrapped posterior mean estimates and 90% credible intervals (brackets) for Kendall's τ correlations between measures of discrimination and prejudice and belief in conspiracy theories.

Recall that the original analysis computed Pearson's correlations, for which there are many tests to probe the difference between two correlations (Diedenhofen & Musch, 2015). The data, however, were measured using ordinal scales. As such, a measure of association like Kendall's τ may be more useful than a linear association, but this introduces a separate problem due to the lack of a standard error for the difference between two Kendall's τ's. In our example above, the BB methodology was used overcome this issue as we trivially estimated and compared the Kendall's τ estimates.

Numerically, the computed correlations and their comparisons were similar to the original, but the resulting interpretations differed. With respect to the magnitude of the associations, some correlations were roughly the same as their Pearson's counterparts, but the majority were slightly weaker, with the differences between the Pearson's and Kendall's τ values ranging from 0.01 to 0.08. Despite these discrepancies, the estimates for the magnitude of the differences were approximately equal between the Pearson and Kendall's correlations. On the other hand, the interpretation of results between the z-test and the ROPE approach diverged. For example, the original analysis failed to reject the null hypothesis for all three comparisons involving Italy. Again, this does not allow statements to be made in support of equality between both correlations. In contrast, the 90% intervals for all three comparisons were trapped completely between [-0.1, 0.1] and under this decision rule, we can conclude that there is evidence to support the respective correlations as

practically equivalent. The credible intervals for the remaining comparisons all overlapped with the ROPE and thus there is no decisive evidence for or against equality of the correlations. This also differed from the original analysis in that the majority of these tests were rejected.

2.4. Discussion

In this article we aimed to show how the Bayesian bootstrap can be applied to obtain Bayesian posteriors for correlation coefficients. We began with a concise introduction to the Bayesian bootstrap and provided formulations to obtain Bayesian bootstrapped versions of the Pearson, Spearman's, Gaussian rank, Kendall's, and polychoric correlation coefficients. The main advantage of the BB is that it considerably simplifies obtaining the posterior for the full correlation matrix. A method for comparing correlations was then introduced based on the region of practical equivalence (ROPE) approach (Kruschke & Liddell, 2018). In an empirical application, we demonstrate how a typical analysis of correlations may be carried out using the **BBcor** package. We supplied R code to 1) estimate and visualize posterior estimates for the correlations discussed in the paper and 2) compare correlations using the ROPE approach and visualize the posterior for their difference. Consequently, this example also serves as a tutorial for readers who wish to implement the methodology outlined in this article.

The methods we proposed in this paper contribute to two bodies of literature. The majority of work in psychology examining correlations within a Bayesian framework has focused on hypothesis testing with the Bayes factor and thus attention is typically restricted to estimating one or two correlations at a time instead of the full correlation matrix. A complementary view in psychology has called for an increased focus on parameter estimation (Kruschke & Liddell, 2018; J. N. Rouder et al., 2018). Introducing the BB for correlations adds to the literature on Bayesian inference of correlations with a focus on parameter estimation because it is a flexible method capable of estimating the full correlation matrix for a variety of correlation types, and can easily be extended beyond those examined in this paper. Second, a considerable amount of work has been done examining methods for comparing correlations, but this work is focused almost exclusively on the Pearson correlation. By providing a framework wherein a variety of correlations may be compared,

the present article also adds to this literature. This is especially valuable for correlations involving ordinal data because of their ubiquity in the social-behavioral sciences.

In our view, the BB is a valuable tool that is best applied when the goal at hand is explicitly parameter estimation as opposed to Bayesian hypothesis testing. Much work at the intersection of psychology and Bayesian statistics has held an eye towards the Bayes factor (BF) for the latter purpose. Although BFs can be derived using the BB posterior (Newton & Raftery, 1994), it is suboptimal because it depends upon the harmonic mean; a method long known to be problematic (Diciccio et al., 1997; Lenk, 2009). If one wants to use the BB to make a decision with respect to a null parameter value, then we view the ROPE approach as a reasonable way of doing so. Moreover, one may want to use an alternative, informed prior when testing a hypothesis. This is challenging with the the Bayesian bootstrap because many hyperparameters must be introduced in order to accomplish this (e.g., Poirier, 2011). Thus, the BB shines in the exploratory stages of research because it employs an uninformative prior and can be used to quickly estimate the posterior for the full correlation matrix.

It is important to keep in mind certain aspects of the BB in practice. Practitioners should be wary of applying the BB to small samples (e.g, $n = 10$), as the resulting credible intervals may be more narrow than those obtained, say, with an MCMC Bayesian analysis. The width of the intervals are crucial when using the ROPE approach as overly narrow intervals may result in overconfident inferences. This may be due to the questionable assumption pointed out by Rubin (1981) that values for unobserved data receive no prior, and hence, no posterior, support (but see Hjort, 1991). Thus, if a sample does not include observations from the tails of the population under study (as might often be the case in small samples), then the variance of the BB posterior may be underestimated. However, as n increases, credible intervals based on the BB posterior will converge on those obtained using traditional Bayesian techniques (assuming a uniform prior).

Further the BB diverges from traditional Bayesian methods in some important ways. Most notably, no prior is explicitly elicited by the analyst. Although the subjective choice of specifying a prior can be seen as a core component of Bayesian inference (Savage, 1954), it is often desirable to eliminate this subjectivity in prior specification (Berger, 2006; Ghosh, 2011). It is interesting to note that Bayesian methods are often favored because they are more consistent with the likelihood

principle (Berger & Wolpert, 1988): all the evidence in a sample that is relevant to model parameters is contained in the likelihood function. But the BB violates this principle because the estimation of parameters relies on aggregating datasets which were not observed. An advantageous difference of the BB lies in the computational speed. Many common methods for Bayesian inference are based on MCMC sampling. Because these draws are serially dependent, many samples are typically required for a consistent estimate of the posterior. On the other hand, samples drawn using the BB are independent and thus fewer of them are required. Despite these differences, the BB is a reliable procedure for obtaining a valid posterior distribution.

2.4.1. Conclusion. We discussed a generic and simple approach to obtaining posterior distributions via the Bayesian bootstrap (BB) for a variety of correlation coefficients. It is generic because it can be applied broadly to different measures of associations and simple because it amounts to calculating weighted correlations. We further discussed a flexible approach to comparing correlations, or linear combinations thereof. Altogether, the BB provides a powerful tool for approximate Bayesian inference of popular correlation types in social-behavior sciences.

CHAPTER 3

Who Is and Is Not "Average"? Random Effects Selection with Spike-and-Slab Priors

Mixed-effects models are being increasingly used in the social-behavior sciences. Their use spans many areas in psychology from observational inquiries that track individuals over an extended period of time, to controlled settings that can include hundreds of experimental trials for each person. Their rise in popularity is mainly due to their ability to partition and account for different sources of variation, for instance, in the experimental effect (Aarts et al., 2014), stimulus type (Wolsiefer et al., 2017), or group membership (Raudenbush & Bryk, 2001). Adequately accounting for these sources of variability leads to the desired inference by ensuring that nominal error rates are maintained (Aarts et al., 2014; Barr et al., 2013; Judd et al., 2012; Williams et al., 2017; Wolsiefer et al., 2017). The idea is that variance components are often considered nuisance parameters that must be controlled or corrected for in order to draw valid inferences. Consequently, the primary inferential targets from mixed-effects models tend to be concerned with population averages, or fixed effects, while variance components play a secondary role. For instance, in a review of papers employing linear mixed-effects models, it was found that less than 10% reported the random effect variances (Meteyard & Davies, 2020), and similarly, only 32% of papers using generalized mixed models reported these variance components (Bono et al., 2021). On the other hand, however, these same sources of variations can provide valuable insights into individual differences in psychological processes (e.g., Haaf & Rouder, 2017; S. Liu et al., 2012; Williams & Mulder, 2019)

When individual differences are of central interest, it is customary to test the variance of the random effects. For example, in determining whether there is variation between individuals in a random intercepts model, one would fit two (nested) models — one with and one without the random intercepts — and perform a likelihood ratio test. If the test is not rejected, one would settle with the simpler model without the random intercept term (i.e., no individual differences).

Conversely, if the test is rejected then the random effect term is retained in the model. In order to explain the individual differences, the latter scenario may be followed up with the inclusion of covariates. In this work, we find a common ground between these two options. Because some individuals are best described by the fixed effect while others may differ drastically from it, we propose a method wherein some individual effects are allowed to deviate from the average and others are not. For example, it may be useful to describe which, if any, individuals depart from a typical learning trajectory (Estrada et al., 2018). As such, we propose a method that offers a more nuanced view of individual differences compared to the classical mixed effect vs fixed effect duality.

The need for more refined views of individual differences is reflected in recent efforts to extend methodological approaches for understanding individual differences. For example, Grice et al. (2020) point out that even though study results, when taken in aggregate, reflect theoretical expectations, it may be that only a few individuals actually behaved in the expected manner. One could imagine that an intervention is shown to alleviate depression on average, but this does not necessarily imply that the intervention is effective for a given individual. As a step towards understanding whether individuals behaved in a hypothesized manner, they propose adopting person-centered effect sizes, wherein effects are computed for each individual. These effects can in turn be used to quantify the proportion of observed effects that were in line with the hypothesized outcome.

In a similar spirit, J. N. Rouder and Haaf (2020) advocate for a Bayesian model comparison approach to distinguish situations where: all individuals have true effects in the same direction, individuals have true effects in differing directions, or all individual effects are equal to an average effect (also see Haaf & Rouder, 2017). This method involves fitting mixed-effects models that reflect each of these settings and comparing them. The underlying aim is to determine if there is support for individual differences in the data, and if so, which model best describes them

To date, however, no general approach has been provided to formally address the individual in individual differences. For instance, the person-centered effect sizes are general in that they can be applied across a wide variety of settings, but are computed in a somewhat ad-hoc manner with a focus on description. The approach in J. N. Rouder and Haaf (2020) allows analysts to quantify evidence for whether individual differences align with a particular pattern, but ultimately relies on

global descriptions of individual differences in linear models. Thus, it is desirable to have a framework that fulfills the desiderata of being applicable across the multitude of settings encountered in psychological science while simultaneously allowing researchers to rigorously evaluate individual effects.

3.0.1. Main Contribution. The main contribution of this work is the introduction of a Bayesian mixed-effects framework that may allow novel inferences in individual differences research. In mixed-effects models, there are fixed effects (averages across individuals), and there are random effects (deviations away from those averages). The main advantage of our proposed methodology is that it allows a more nuanced view of individual differences by quantifying evidence for or against *individual* random effects. In addition, because it can be fit using standard statistical software, it is flexible enough to be applied to a broad class of models (i.e., generalized linear mixed models).

With this framework we explicitly address the individual by providing a tool that is capable of answering which individuals are "average" and which ones are not. Intuitively, if β is a fixed effect, θ_j is the corresponding random effect for the jth individual, and $\beta_j = \beta + \theta_j$ is the total effect for the jth individual, then the problem we are interested in can be thought of as evaluating whether $\beta_j = \beta$ or $\beta_j \neq \beta$. As for implementation, the models we describe in this paper can easily be fit in the common programming languages R (R Core Team, 2021) and Python (Van Rossum & Drake, 2009), or by using the R package **SSranef**[1].

To answer the question of who is "average", we build upon spike-and-slab priors for Bayesian variable selection (George & McCulloch, 1993; Kuo & Mallick, 1998; Mitchell & Beauchamp, 1988). Traditionally used in the canonical regression setting to select predictors that are likely to have a non-zero effect, our innovation is to apply the spike-and-slab to select which *random effects* are likely non-zero in a mixed-effects model. A similar approach has been applied in psychological settings (e.g., Williams et al., 2021), but was restricted to random intercepts in linear mixed models whereas, in practice, the primary interest is often the random slopes. Further, it is common to estimate models with non-Gaussian likelihoods (e.g., mixed-effects logistic regression). Thus, a

[1]The **SSranef** R package can be downloaded from GitHub at https://github.com/josue-rodriguez/SSranef. An example illustrating how to use **SSranef** can be found in the Appendix.

novel aspect of this work is the extension of the spike-and-slab to random effects on slopes and generalized linear mixed models.

3.0.2. Overview. In what follows, we first present a motivating example where we introduce the central ideas underlying the spike-and-slab prior in the context of a generalized mixed-effects model. We show the value in using the spike-and-slab on random intercepts and that it is trivial for this approach to be applied to a variety of model types. We then demonstrate how the idea of random effect selection can be extended to random slopes. This allows researchers to, for example, answer how many individuals differed from a common experimental effect. This approach is illustrated in two empirical psychological data sets where we show how individual differences in the random slopes can be comprehensively disentangled. In two simulation studies, we assess the ability of our the proposed method to correctly identify (non)-average individuals without compromising the mixed-effects estimates. We conclude with a discussion on the implications of the current work and future directions.

3.1. Background

We employ the spike-and-slab approach for variable selection. In this approach, the selection problem is formulated in terms of a two-component mixture: 1) a 'spike' that is either a distribution centered narrowly around zero (George & McCulloch, 1993, 1997) or a Dirac measure at zero (Kuo & Mallick, 1998; Mitchell & Beauchamp, 1988) and 2) a diffuse 'slab' component surrounding zero. The former allows the shrinkage of small effects to zero and the latter prevents heavy shrinkage of larger effects. A central aspect of this approach is the addition of an indicator variable (Kuo & Mallick, 1998), which allows for switching between the spike and the slab throughout the MCMC sampling process (i.e., transdimensional sampling; Heck et al., 2019). The proportion of MCMC samples spent in each component can then be used to approximate the respective posterior model probabilities or the marginal Bayes factor for whether an effect should be included. In the context of random effects selection, this Bayes factor expresses the evidence for whether the random effect for a given individual should be included in the model. Interested readers can find an excellent introduction to the spike-and-slab prior for psychology in J. N. Rouder et al. (2018) and in-depth overview of its various specifications in O'Hara and Sillanpää (2009).

Importantly, much of the literature on spike-and-slab priors has been concerned with model selection and comparison (George & McCulloch, 1997; O'Hara & Sillanpää, 2009). This is distinct from our application in this paper as we do not focus on model selection in a traditional sense. Our goal is not to make judgments with respect to quality of fit among models with different variables, prior distributions, or functional forms, but rather we seek to use spike-and-slab priors as a means of understanding which individuals' effects deviate from a population-average estimate.

3.1.1. Illustrative Example. We begin our exposition by considering the work of Frühwirth-Schnatter and Wagner (2011), who used spike-and-slab priors with the overarching goal of

> [making] unit-specific selection of random effects in order to identify units which are 'average' in the sense that they do not deviate from the overall mean. (p. 2)

Specifically, they provided examples of random effect selection with a focus on logistic models. However, their approach relied on custom MCMC sampling schemes, rendering the techniques inaccessible to all but those who are comfortable implementing the algorithms on their own. Williams et al. (2021) introduced the idea of selecting unit-specific random effects to psychology with the goal of determining individual reliability, but they did not consider models outside of a classical random intercepts model. Because, to our knowledge, these are the only works to consider random effect selection, we view this as a good place to begin our exposition of the spike-and-slab. Using a random intercepts logistic regression model, we highlight key ideas relevant to our approach for random effect selection.

3.1.1.1. *Model Formulation.* For our illustrative example we use data from a linguistics experiment that were first reported in Caplan et al. (2021, Experiment 1). The participants ($N = 128$) in this study were presented with acoustically ambiguous audio involving minimal pairs of words (e.g., time/dime) along with disambiguating information that biased the audio to be interpreted as /t/ or /d/. The outcome for the ith trial and jth person is coded as a 1 or a 0 and represents whether participants heard a /t/ (1) or a /d/ (0) for a given word during the test phase (see original text for full details). For illustrative purposes, we adopt a simpler version of the full analysis in that we only consider a random intercept without covariates or additional random effects. To facilitate spike-and-slab selection, we employ the non-centered parameterization (Papaspiliopoulos et al., 2007), that is,

$$
\begin{aligned}
y_{ij} &\sim \text{Bernoulli}\left(\pi_j\right) \\
\text{logit}\left(\pi_j\right) &= \alpha + \tau z_j \\
\alpha &\sim \text{Normal}(0, 1) \\
\tau &\sim \text{St}^{+}(\nu = 3, 0, 1) \\
z_j &\sim \text{Normal}(0, 1),
\end{aligned} \tag{3.1}
$$

where y_{ij} is the outcome, α is the overall intercept, τ is the standard deviation of the random effects, z_j is a standardized effect size, and the product τz_j constitutes the random effect. Here, we are not modeling the random effects directly, but rather inferring them from a latent variable z_j. There our two main reasons for this: 1) it may lead to more efficient sampling of the posterior and 2) it allows us to think about the random effects in terms of standardized effect sizes. Further, we set standard normal priors for α and z_j, and a half Student-t distribution with three degrees of freedom for τ. Our choice for the half Student-t prior distribution stems from it having better properties than common alternatives for variance parameters in hierarchical models (e.g., inverse-gamma Gelman, 2006). The model in (3.1) estimates the baseline log-odds of hearing a /t/ (intercept), but allows for each individual to deviate away from it (random effect).

In such an analysis, it might be natural to ask whether each individual does indeed deviate from the overall log-odds, α, in hearing a /t/. This question can be addressed by adding an indicator variable $\gamma_j \in \{0, 1\}$ to the above model that governs, for each individual, whether the random effect is in the spike ($\gamma_j = 0$) or the slab ($\gamma_j = 1$) portion of the model in each MCMC iteration. Introducing this variable only requires the following modifications to (3.1)

$$
\begin{aligned}
\text{logit}\left(\pi_j\right) &= \alpha + \tau\left(z_j \gamma_j\right) \\
\gamma_j &\sim \text{Bernoulli}\left(\delta\right)
\end{aligned} \tag{3.2}
$$

while everything else remains the same. In (3.2), δ represents the *prior* inclusion probability, or the a priori probability that the jth random effect is non-zero. Choosing $\delta = 0.5$ expresses a lack of a priori preference for whether a random effect should be included or excluded, and it is the choice

we make throughout this article. Notice that when $\gamma_j = 0$, the random effect for the jth individual random effect drops out of the model, and when $\gamma_j = 1$, it is retained. If there is prior information that indicates whether individuals are more or less likely to deviate away the average, then this information can be included in the analysis by modifying δ to be greater than or less than 0.5.

The proportion of MCMC samples in which γ_j is equal to one is referred to as the *posterior* inclusion probability (PIP) of the jth random effect,

$$\Pr(\gamma_j = 1|\mathbf{Y}) = \frac{1}{S}\sum_{s=1}^{S}\gamma_j^{(s)}, \tag{3.3}$$

where $s = 1, \ldots, S$ indexes the MCMC samples and $\mathbf{Y}$ denotes the data. When there is strong support for including the jth random effect, its PIP will be large, and when there is little support for inclusion, its PIP will be small. PIPs of 0 and 1 indicate complete posterior support for excluding and including the jth random effect, respectively. Additionally, Bayes factors (Kass & Raftery, 1995) can be computed based on the PIPs. Assuming equal prior odds, the Bayes factor in favor of the random effect being non-zero rather than zero can be calculated as

$$BF_{10} = \frac{\Pr(\gamma_j = 1|\mathbf{Y})}{1 - \Pr(\gamma_j = 1|\mathbf{Y})}. \tag{3.4}$$

The ability to compute posterior inclusion probabilities and Bayes factors allows for the direct quantification of evidence for whether an individual's baseline log-odds are different than the "average" baseline log-odds of hearing a /t/.

Although it is not the only way to formulate a spike-and-slab prior in a Bayesian model (O'Hara & Sillanpää, 2009), our approach carries some distinct advantages. First, by using a point-mass at zero for the spike instead of a continuous distribution with small variance, we explicitly consider whether a given random effect is equal to zero instead of just nearly zero. Further, the prior probability of drawing a one for γ_j (i.e., the prior inclusion probability) is fixed at 0.5. This is equivalent to setting equal prior odds for whether a random effect is non-zero or zero, and simplifies the expression for the Bayes factor. Note that allowing the prior probability δ to be a random variable by endowing it with a prior (e.g., Beta) may result in superior selection for point-mass spikes (Ley & Steel, 2009). For these reasons, the above formulation of the spike-and-slab is the one we use throughout the paper.

3.1.1.2. *Software and Estimation.* We fit the model using the JAGS language in R (Plummer, 2003) because of its ability to easily fit spike-and-slab models (Ntzoufras, 2002; O'Hara & Sillanpää, 2009)[2]. The fitted model used four chains of 25,000 iterations after a burn-in period of 5,000 iterations which resulted in a total of 100,000 samples from the posterior distribution. This number of samples provided a good quality of the parameter estimates (all $\hat{R}$s $= 1$; Brooks & Gelman, 1998).

3.1.1.3. *Results.* The results are displayed in Figure 3.1. Panel A shows the prior distribution for the random effects and Panel B shows the posterior for the random effect of the 56th and 78th participants, respectively. Note that the spike (black arrow) and slab (blue bars) both constitute roughly half of the prior density. Panel C displays the point estimates of the random effects for all 128 participants and their respective 90% credible intervals (CrIs). The individuals from panel B are represented by the green (participant 78) and orange (participant 56) dots.

Recall that the goal of fitting this model was to determine the evidence for whether a given individual deviates from the overall log-odds, or intercept. If an individual does not differ from the intercept, then most of the of posterior mass should be in the spike for the random effect. If an individual does differ from the intercept, then there should be a lot of posterior mass in the slab. This can be clearly seen in Figure 3.1 where most of the posterior mass is in the spike for participant 56 and, conversely, none at all for participant 78. For the former, there was a 0.23 posterior inclusion probability, or a Bayes factor of roughly 3 in favor of the spike. This can be considered moderate evidence in favor of the participant being "average" (Lee & Wagenmakers, 2013). For the latter, the posterior inclusion probability was 1 and is equivalent to a Bayes factor of infinity that this individual differs from the "average".

The shapes of these posteriors have a straightforward relationship with the size of the random effect. This correspondence is shown in Figure 3.1 (panel C) where the orange dot (participant 56) is near zero and the green dot (participant 78) is far away from zero. This makes sense intuitively; if a random effect is near zero, then there will be little to no evidence that a participant differs from the intercept, and conversely, there will be stronger evidence that a participant differs from the intercept with larger random effects.

[2]All code to reproduce the analyses and figures in this paper are available on the Open Science Framework at https://osf.io/n2z49/.

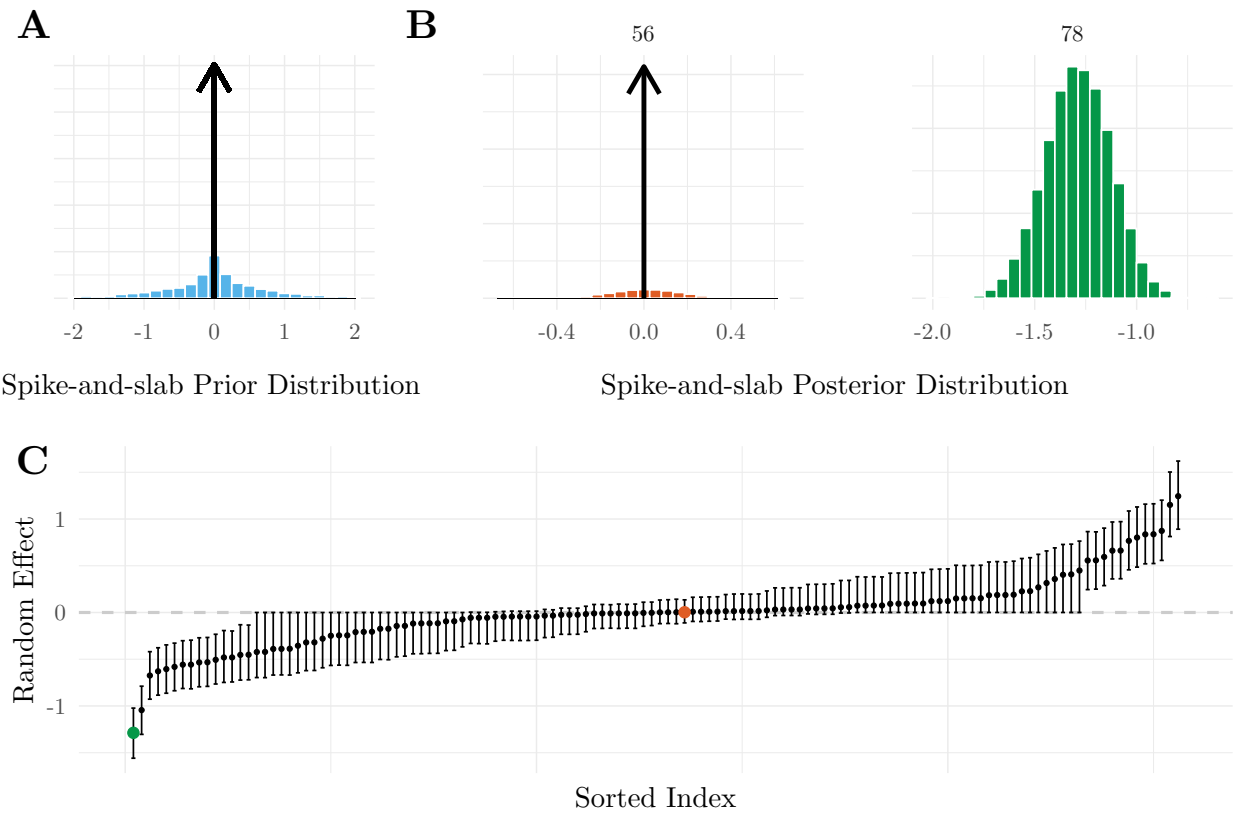

FIGURE 3.1. A) The point-mass spike-and-slab prior distribution. The spike (arrow) and the slab (blue bars) each take up half the prior density. When a random effect is sampled from the spike, it is zero and the effect for that individual is equal to the fixed effect. When it is sampled from the slab, it will take a non-zero value and the individual effect will deviate from the fixed effect. The proportion of MCMC iterations that a random effect is sampled from the slab is called its posterior inclusion probability (PIP). B) The posterior distribution for the random effects of the 56th and 78th participants. For the former, the majority of the posterior mass is in the spike (PIP = 0.23) where there is little mass in the slab (orange bars). For, the latter the entire posterior is in the slab (green bars, PIP = 1). Thus, participant 56 can be considered "average", and participant 78 can be considered not "average". C) Posterior means and respective 90% credible intervals for the random effects. The orange point (participant 56) is centered near zero and the green point (participant 87) is far from zero. This matches the corresponding posterior mass in the spike for each of these random effects.

3.1.1.4. *Summary.* The purpose of this illustrative example was to build the foundation for the following methodology. We highlighted the central idea behind the spike-and-slab prior, and in particular, how it can be leveraged to select individual random effects. The results indicated that this methodology can be profitably applied to determine which individuals differ from the overall

intercept in a logistic regression setting. The remainder of this paper will extend this idea to include random slopes to determine whether individuals differ from the average experimental effect.

3.2. Extension to Random Slopes

In psychology, it is more common for the random slopes to be of focal interest, not the random intercepts. The reason for this is that the slopes often corresponds to the effect of condition or manipulation in experimental settings. Accordingly, random effects in the slope encode individual differences in experimental effects. Thus, we seek to extend the application of the spike-and-slab prior to the random effects in slopes. Placing the spike-and-slab on the slopes allows evidence to be obtained for which individuals differ from the average experimental effect and which do not. As above, our exposition of this extension will be through applied examples.

It has recently been argued that there is low reliability in popular cognitive tasks for studying individual differences (Hedge et al., 2018; J. N. Rouder et al., 2019). The main explanation for low reliability among such tasks is that there exists little individual differences. In this context, individual differences are defined in reference to the ratio of between-subject variance to total variance. In what follows, we are not interested in individual differences in this sense, but whether there are individual differences in these tasks with respect to who deviates from the overall experimental effect, and then determining the kind of insights that may follow.

3.2.1. Empirical Application. We apply the proposed methodology to data from two classical inhibition tasks. These data were first analyzed in Hedge et al. (2018) and again in J. N. Rouder et al. (2019).

3.2.1.1. *Dataset 1: Stroop task.* Participants ($N = 47$) responded to the color of a centrally presented word which was red, blue, green or yellow. The word could be the same as the font color (congruent condition), different from the font color (incongruent condition), or one of four non-color words (neutral condition). Each participant completed 240 trials for each condition with the primary outcome being reaction time. For illustrative purposes, we focus simply on the congruent and incongruent conditions.

3.2.1.2. *Dataset 2: Flanker task.* The same 47 participants responded to the direction of a centrally presented arrow (left or right). On each trial, the central arrow was flanked above and

below by two other symbols. Flanking stimuli were arrows pointing in the same direction as the central arrow (congruent condition), arrows in the opposite direction as the central arrow (incongruent condition), or straight lines (neutral condition). Again, each participant completed 240 trials for each condition and the primary outcome was reaction time. As above, we only give consideration to the congruent and incongruent conditions.

3.2.1.3. *Model Formulation.* Because these tasks are similar both in what they are thought to measure and in their design, each dataset contains the same variables on which we focus: outcome (reaction time) and condition (in/congruent). Accordingly, we define a single model formulation that can be seamlessly applied to each dataset without modifying anything except for the data. For the ith trial and the jth person, we can define the likelihood for the reaction time as

$$
\begin{aligned}
y_{ij} &\sim \text{Normal}\left(\alpha_j + x_{ij}\beta_j, \sigma^2\right) \\
\alpha_j &= \alpha + \theta_{1j} \\
\beta_j &= \beta + \theta_{2j},
\end{aligned}
\tag{3.5}
$$

where for the jth person, α_j is the random intercept and encodes the average response time for the congruent condition, and β_j is the random slope which captures the difference in response time in the incongruent condition, relative to the congruent condition. The term x_{ij} encodes the condition (0 = congruent; 1 = congruent), and σ^2 is the residual variance. The terms θ_{1j} and θ_{2j} indicate the random effects for the intercept and slope, respectively.

For the model parameters defined in (3.5), we set the priors as follows:

$$
\begin{aligned}
\alpha, \beta &\sim \text{ Normal}\,(0, 1) \\
\boldsymbol{\theta}_j &\sim \text{Normal}(\mathbf{0}, \boldsymbol{\Sigma}) \\
\boldsymbol{\Sigma} &= \boldsymbol{\tau}\boldsymbol{\Omega}\boldsymbol{\tau} \\
\boldsymbol{\Omega} &\sim \text{LKJ}\,(\eta = 1) \\
\sigma, \tau_{11}, \tau_{22} &\sim \text{St}^{+}(\nu = 3, 0, 1).
\end{aligned}
\tag{3.6}
$$

Here, we place uninformative normal priors over the fixed effects and a multivariate normal prior with covariance matrix $\mathbf{\Sigma}$ for the random effects. We model the covariance matrix using the separation strategy discussed in Barnard et al. (2000) where $\mathbf{\Omega}$ is a 2×2 correlation matrix of the random effects and $\boldsymbol{\tau}$ is a 2×2 diagonal matrix whose elements are the standard deviations of the random effects. The prior for the correlation matrix is the LKJ distribution (Lewandowski et al., 2009) and is governed by a single parameter η. Setting $\eta = 1$ places a uniform prior over all correlation matrices. We set Half Student-t priors for all variance parameters for the reasons discussed in the Illustrative Example.

Of central interest is the parameter β_j, which corresponds to the experimental effect for the jth person. Recall that we want to know whether each individual differs from the common effect, β, and that we can use a spike-and-slab prior to answer this question. Thus, we can modify the above to include a spike-and-slab prior on the random slopes

$$\begin{aligned} \beta_j &= \beta + \theta_{2j}\gamma_j \\ \gamma_j &\sim \text{Bernoulli}\left(\delta\right), \end{aligned} \tag{3.7}$$

where $\delta = 0.5$ and everything else remains the same.

3.2.1.4. *Model Selection.* Up to this point, we have not discussed a decision rule for actually determining which individuals differ from the average effect. This is because Bayesian inference is not focused on making discrete choices, but rather considering the weight of evidence (Morey et al., 2016). In any case, there are times when it is desirable to do so. For instance, in addition to reporting random effect variances, one can report for example that 30% of the random effects differed from the average effect. Reporting such a number is in the same spirit as the metrics described in Grice et al. (2020), but supported by formal evidence (i.e., posterior inclusion probability). This might be especially insightful in situations with low between-person variance, a scenario that typically implies a lack of individual differences. This type of information can also be useful in other fields such as clinical or educational psychology, where one can identify a subset of individuals who respond differently to an intervention compared to the average response. Identifying individuals who display unusual behavior via random effects can be extended to models of variability as well

(e.g., Rast & Ferrer, 2018). For example, in cognitive aging research, random effects in the residual variance can be used to capture differences in behavioral "consistency" of cognitive ability (Rast & Zimprich, 2011; Watts et al., 2016). Here, identifying individuals with above or below average residual variance could serve as an early warning sign to the onset of Alzheimer's Disease (Lövdén et al., 2013; MacDonald et al., 2008).

Because in our above example we place the spike-and-slab prior on $N = 47$ random effects, there are 2^{47} distinct combinations of random effects that can be considered for inclusion in the final model. That is, there are 2^{47} possible models from which to choose. Thus, the issue that presents itself is how to choose which model should be used to determine who is "average". An intuitive choice would be to select the highest probability model (HPM), or the model containing the combination of random effects selected most frequently throughout the MCMC sampling process. In fact, it is the *median probability model* (MPM, Barbieri & Berger, 2004; Barbieri et al., 2021) that is more often considered. The MPM, which is used in the present paper, is defined to be the one including only those random effects with posterior inclusion probabilities (Equation 3.3) of at least 0.5. Several motivations underlie the MPM, including that it is the best single-model approximation to Bayesian model averaging and it is optimally predictive for linear models with respect to squared error loss under orthogonal designs. This does not mean the HPM should never be used, however. Indeed, the HPM can be used when the goal is explicitly to compute a Bayes factor of interest for hypothesis testing. That is, if one has a priori predictions about which individuals differ from the fixed effect. Further, once individuals have been classified as "average" or not, then it is straightforward to compute the proportion of the sample that differed from the common effect.

3.2.2. Software and Estimation. We fit the model above to both the Stroop and Flanker data using the **pymc3** (Salvatier et al., 2016) package in the Python programming language (Van Rossum & Drake, 2009). This was primarily because it allows the use of more efficient MCMC sampling schemes (e.g., Hoffman & Gelman, 2011) while retaining the ability to accommodate the point-mass spike-and-slab prior[3]. The fitted models used four chains of 10,000 iterations after a

[3]This model converged in JAGS without issues, but we fit it in pymc3 to demonstrate how to employ these models when more efficient samplers are desired.

tuning period of 2,000 iterations which resulted in a total of 40,000 samples from the posterior distribution. This number of samples provided a good quality of the parameter estimates (all $\hat{R}$s = 1).

3.2.2.1. *Results.* The main results are displayed in Figure 3.2. Panel A shows the point estimates of the slope random effects for all 47 participants and their respective 90% CrIs. Throughout the rest of this section, we will simply use "random effects" as shorthand for the slope random effect. Panel B displays the posterior inclusion probabilities (PIPs) as a function of the magnitude of the random effect. Upon visual inspection, it is easy to see which individuals have more evidence supporting that they differ from the average experimental effect. The PIPs make a V-shape in that they decrease as the magnitude of the effect approaches zero and increase again as they move away from zero. This is again unsurprising. Individuals with larger random effects should have more evidence to support that they differ from the average effect.

For the Stroop task, the mean posterior estimate for the overall experimental effect, β, was 0.07 and had a corresponding 90% CrI of [0.062, 0.076]. That is, on average, participants' reaction time was slower by 0.07 seconds in the presence of incongruent stimuli. Notably, the mean posterior estimates for the random effects ranged from -0.02 to 0.07, and their corresponding PIPs ranged from 0.22 to 0.99 (Bayes factors factors of 0.28 to over 13,000), in support of including the random effect. This spread of PIPs indicates considerable fluctuations in the level of support for whether individuals differ from the average experimental effect. They span from "moderate" evidence in favor of belonging to the average experimental effect on one end to "extreme" evidence in favor of different from it on the other (Lee & Wagenmakers, 2013). This spread was even wider in the Flanker task, where the PIPs for covered values from 0.19 to 1.

As previously mentioned, it may sometimes be desirable to categorize individuals as being "average" or not. When using the median probability model, individuals with PIPs over 0.5 can be thought of as being different from the average effect. In Figure 3.2 (Panel B), these two groups are separated by the dark dotted gray line. It is intriguing that for both tasks, quite a few points lie above this line. Specifically, 12 and 13 participants are above this line for the Stroop and Flanker tasks, respectively[4]. In other words, there is evidence that despite the belief that few individual

[4]We also examined PIP cut-offs of 0.75 and 0.9 (light gray dotted lines). For the former, this corresponds to 11% of the sample differing from the average experimental affect and roughly 7% for the latter.

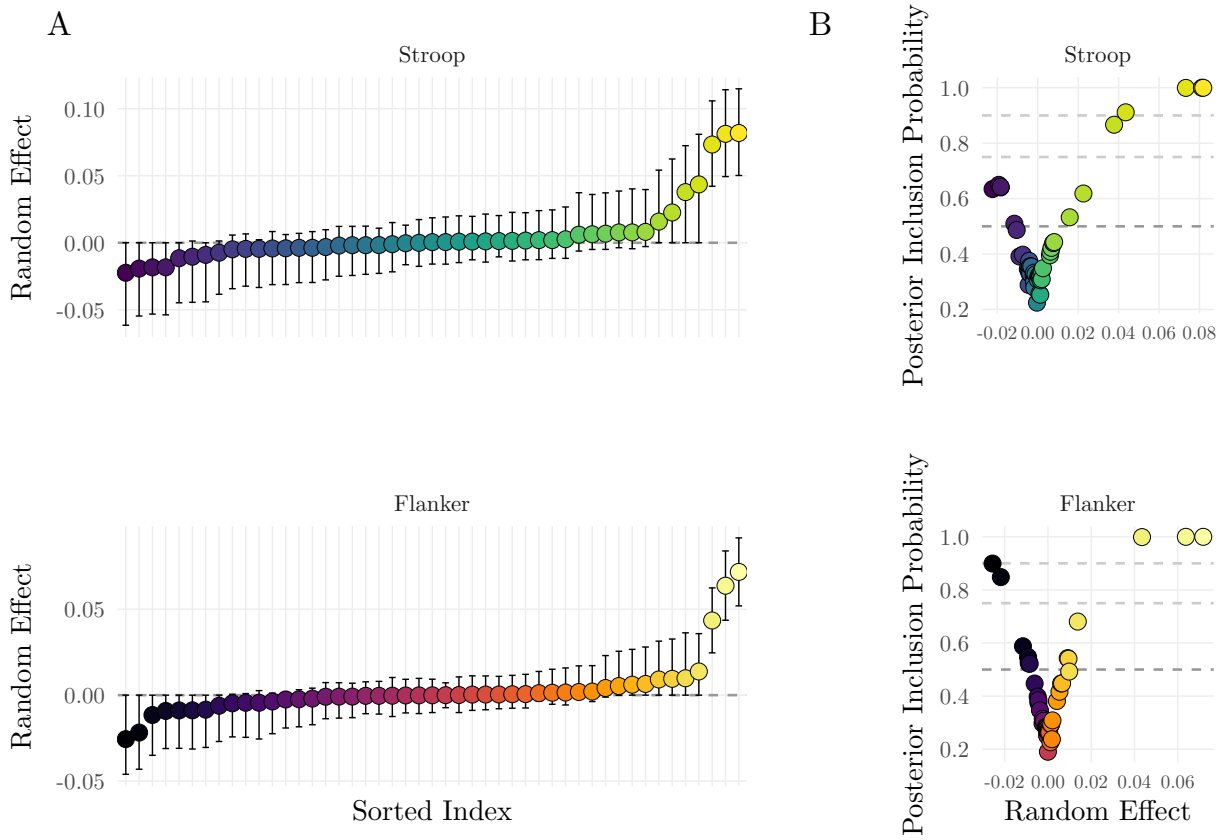

FIGURE 3.2. A) Posterior means and corresponding 90% CrIs for the random effects for the slopes (or experimental effects) in the Stroop and Flanker data, sorted in ascending order. B) The corresponding posterior inclusion probabilities or each random effect. The dark gray dotted line indicates a PIP of 0.5. The two light gray lines denoted PIPs of 0.75 and 0.90. Random effects that are closer to zero have whose posterior estimates have lower PIPs. If one were to use the median probability model as a decision then everyone above the dotted gray line would be considered as different than the "average". For both the Stroop and Flanker tasks, over 25% of the points lie above the dotted line. This clearly demonstrates individual differences in these tasks. Across both panels, distinct colored points refer to the same random effect.

differences exist in these kind of data, over a quarter of the sample diverged from the average experimental effect in each task.

Taken together, these results not only attest to the existence of individual differences in these two experiments, but speak to which individuals (and how many) differed from the average effect.

3.2.2.2. *Individual Performance Across Tasks.* The Stroop and Flanker tasks have long been considered to be measures of inhibition (Friedman et al., 2004). It is consequently natural to

think that individuals who differ from the average experimental effect in one task should also differ from the average effect in the other. In contrast, recent work has suggested that the correlations among inhibition tasks are low (Hedge et al., 2018; J. N. Rouder et al., 2019). That is to say, that performance on a given task is not necessarily predictive of performance on another. Because we examine individual differences in the sense of differing from a fixed effect and not in terms of the amount of variance, we look at whether the PIPs were comparable for individuals across task. Note that it would be possible to fit a multivariate model with the reaction times for both tasks as the outcome, and directly apply the spike-and-slab formulation to the random slopes for each task. In order to keep the exposition manageable, we opt for simple description.

Figure 3.3 displays a funnel plot containing the PIPs of the random slope effects for individuals on both tasks, sorted in descending order of PIPs for the Stroop model. The idea here is that if performance on these tasks are related, then we should see a funnel shape that starts wide at the top (i.e., individuals who had large PIPs in both tasks) and becomes narrow at the bottom (i.e., individuals who had small PIPs for both tasks). However, upon visual inspection, there is no apparent relation between the PIPs. For instance, participant 24 had a PIP of 0.99 for their random effect in the Stroop model, but a PIP of 0.31 in the Flanker model. On the other hand, Participant 35 had PIPs of near 1 on both tasks. Hence, whether an individual differs from the average experimental effect in one task may not be predictive of whether they differ from the average experimental effect in another.

3.2.2.3. *Posterior Predictive Check.* Lastly, an important aspect of Bayesian inference is model checking. This is typically done with the posterior predictive distribution (Gelman et al., 1996; X.-L. Meng, 1994). The main idea behind a "posterior predictive check" is that data generated from the model should resemble the observed data. The posterior predictive check thus entails generating datasets from the predictive distribution of the fitted model and comparing them to the observed dataset in order to evaluate the model's goodness-of-fit. Importantly, posterior predictive checks should capture aspects of the model which are of particular interest (Gelman & Hill, 2006, p. 514).

A principal quantity here is the Bayesian p-value, which can be defined as the proportion of times a quantity of interest calculated from the posterior predictive distribution exceeds the

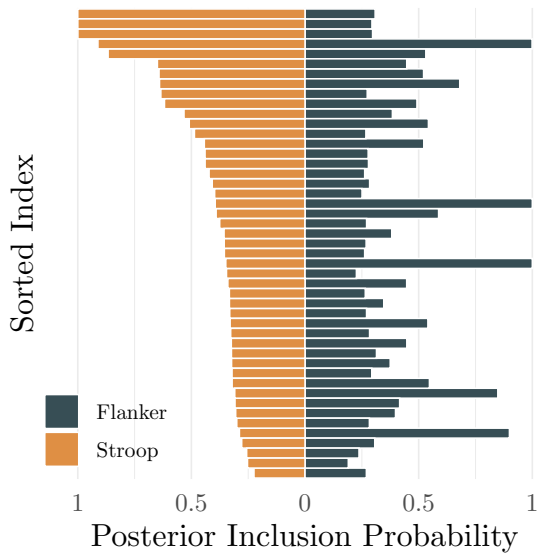

FIGURE 3.3. Funnel plot of the PIPs in the Stroop and Flanker tasks. For each individual, the orange bar indicates the PIP for their random effect on slope in the Stroop task. The opposite-side blue bar indicates that individual's PIP for their random effect on slope in the Flanker task. Because the plot does not produce a funnel shape, this suggests that whether an individual deviated from the average experimental effect in one task may not be predictive of whether they deviated in the other. Hence, although there were individual differences insofar as who was "average" in each task, it seems that reliability was low.

observed quantity. If the model is adequately capturing the data, then the p-values should be relatively close to 0.5 (Gelman, 2013). Values near 0 or 1 would indicate systematic misfits of the model to the data. Because the models we fit are focused on the mean differences in reaction time between two experimental conditions, as opposed to, say, the shape of the reaction time distributions, we perform a posterior predictive check on the subject-specific mean differences. If the model adequately captures these mean differences, the p-values should be dispersed around 0.5.

For each of 2,000 draws from the predictive distribution, we calculated the mean difference in reaction time between conditions for each of the 47 subjects. The resulting values were then compared to the empirical mean differences. The results of the posterior predictive checks are shown in Figure 3.4. The empirical mean differences are represented by red points and posterior predictive mean differences are indicated by the black points. The numbers on the right-hand side are the corresponding Bayesian p-values. Across both tasks, the p-values span from 0.16 to 0.84, with most of them between 0.25 and 0.75. These results can be viewed as evidence that the

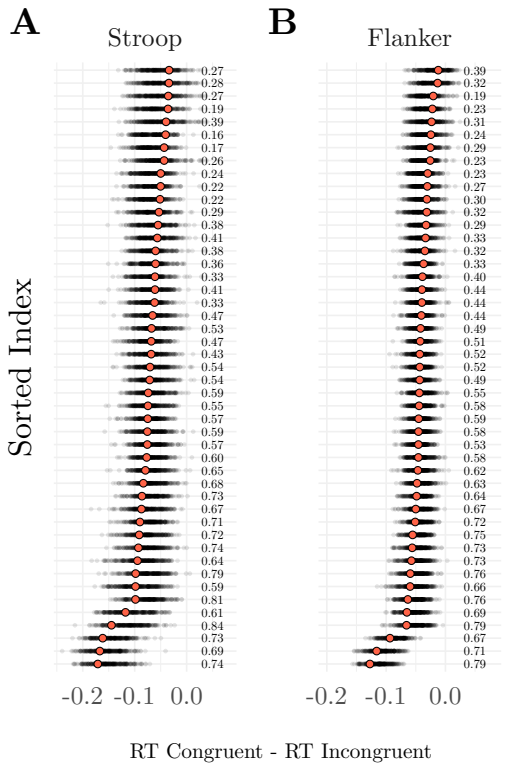

FIGURE 3.4. A posterior predictive check on the mean difference in reaction time between the congruent and incongruent conditions for each individual in the Stroop and Flanker tasks. The red points indicate the observed mean difference in reaction time and the black dots are draws from the posterior predictive distribution. The numbers on the right-hand side of each panel correspond to the Bayesian p-value for these predictive checks. Bayesian p-values that are closer to 0.5 than 0 or 1 suggest the model is successfully capturing the mean differences. As can be seen, the spike-and-slab formulation for these models adequately captures the mean differences.

fitted model adequately captures mean differences between conditions in the data and "passes" this posterior predictive check.

3.3. Simulation Studies

Up to this point, we have demonstrated how the spike-and-slab prior can be applied to gain new insights into individual differences in psychology. We now focus on better understanding the properties of the spike-and-slab prior when placed on random effects by way of two simulation studies. The first aims to support our claim that the spike-and-slab prior on the random effects is indeed capable of correctly identifying those who differ from the average and those who do not.

In the second simulation study, we address a potential issue noted by reviewers. As shrinkage is already an inherent part of mixed-effects models (Gelman & Hill, 2006; Raudenbush & Bryk, 2001), the inclusion of a spike-and-slab prior could incur "double shrinkage". That is, the random effects may be biased due to shrinkage in both the slab (as in a typical mixed-effects model) and spike components of the prior. The second simulation study investigates this possibility. To situate the findings within a familiar context, we include a standard mixed-effects model (i.e., a normal prior on the random effects) for comparison in both simulation studies.

3.3.1. Study 1. The goal of this study was to assess the classification performance of the spike-and-slab prior with respect to average and non-average random effects. Accordingly, we simulated data for a random intercepts model with $n = 100$ units of interest (e.g., people) and varied the number of observations per unit $n_j = 5, 10, 25$. For each $j = 1, \ldots, n$ unit, each $i = 1, \ldots, n_j$ observation, y_{ij}, was generated as

$$
\begin{aligned}
y_{ij} &= \alpha_j + \epsilon_{ij} \\
\alpha_j &= \alpha + \theta_j \\
\epsilon_{ij} &\sim \text{Normal}(0, 1),
\end{aligned}
\tag{3.8}
$$

where $\alpha = 1$ and θ_j captures the random effect for the jth person. The θ_j were systematically varied to be either 0, +1, or −1. The proportion of random effects that were exactly zero was set to be either 0.94, 0.74, or 0.5. The remaining random effects were set to either +1 or −1 in equal proportions.[5] These proportions translate to between-unit variances τ^2 of approximately 0.05, 0.2, and 0.25. Further, by setting $\sigma^2 = 1$, the resulting intraclass correlation coefficients (ICCs) are approximately 0.05, 0.15, and 0.2, respectively, where the ICC is defined as (Raudenbush & Bryk, 2001)

$$
\text{ICC} = \frac{\tau^2}{\tau^2 + \sigma^2}. \tag{3.9}
$$

[5]These values were chosen so that approximately, 50%, 75%, and 95% of the random effects were exactly zero, but an even number of non-zero random effects remained.

The ICC plays a key role in mixed-effects models because it captures test-retest and inter-rater reliability (Shrout & Fleiss, 1979; Weir, 2005), and is the proportion of total variance accounted for by the between-group variance, τ^2. As we will discuss below, the ICC is also of particular interest because it determines, in part, the amount of shrinkage that occurs.

For each of 200 iterations, data were generated as previously described and two mixed-effects model were fit: one employing the spike-and-slab prior on the random effects and another using the customary normal prior on the random effects. For the spike, a point-mass at zero was used whereas a diffuse normal distribution was used for the slab (as in Equations 3.1 and 3.2). The latter distribution was also used as the prior for the random effects in the standard mixed-effects model. For both models, the likelihood and remaining priors were specified as

$$
\begin{aligned}
y_{ij} &\sim \text{Normal}\left(\alpha + \theta_j, \sigma^2\right) \\
\alpha &\sim \text{Normal}\left(0, 1\right) \\
\sigma, \tau &\sim \text{St}^{+}\left(\nu = 3, 0, 1\right).
\end{aligned} \tag{3.10}
$$

All models were fit in R using the JAGS language. The fitted models used four chains of 5,000 iterations after a burn-in period of the same length.

Once the models were fit, each random effect θ_j was classified as average or differing from the average. A correct classification occurred when a non-zero random effect was included in the final model or when a zero random effect was excluded. For the model with the spike-and-slab prior, we considered two thresholds for inclusion: 1) a posterior inclusion probability (PIP) of 0.5 (i.e., the median probability model) and 2) a PIP of 0.75 (i.e., a Bayes factor of 3). For the model with the normal prior on the random effects, a 90% credible interval was used to classify the random effects. If the interval for the jth random effect included 0, then it was excluded from the final model, and included otherwise. Model performance was considered in terms of specificity[6], the proportion of truly zero random effects that were correctly classified, and sensitivity, the proportion of truly non-zero random effects that were correctly classified.

[6]Note that (1− specificity) is the false positive rate.

3.3.1.1. *Results.* The results are displayed in Figure 3.5. Panel A displays the average sensitivity for the random effects across ICCs, observations per unit, and priors. Across ICCs, all priors tended towards a sensitivity rate of 1, however, there were some discrepancies in conditions with few observations per unit. When n_j was either five or ten, the spike-and-slab prior using a PIP of 0.5 as the inclusion threshold ($SS_{0.5}$) was superior to both the spike-and-slab model using a PIP of 0.75 ($SS_{0.75}$) and the normal model using the 90% CrI. Interestingly, with relatively little between-unit variance (ICC = 0.05) and few units per observation ($n_j = 5$), the $SS_{0.5}$ model was 3.5 and 11 times more accurate in detecting non-average units than the $SS_{0.75}$ and normal models, respectively. This suggests that the spike-and-slab may be fruitfully applied to detect non-average individuals even when between-person variance is low. In sum, with sufficient observations, all models performed comparably well in detecting non-average units, but the $SS_{0.5}$ model (i.e., median probability model) was superior when either the ICC or number of observations was small. [7]

The average specificity is similarly displayed in panel B. Across all conditions, the worst specificity was observed for the $SS_{0.50}$ model with the ICC set to 0.05 and n_j set to 5. Here, the specificity for the $SS_{0.50}$ model was 0.75, while it was 0.99 for both the $SS_{0.75}$ and the normal model. As n_j increased, specificity for the normal model decreased and stabilized near a specificity of 0.9, or a false positive rate of 0.1. This is unsurprising as the specificity for credible interval approaches should be roughly equal to the width of the credible interval (Rubin, 1984). In contrast, the specificity for both spike-and-slab models were stable near one or tended to one. This finding hints at the model selection consistency property the spike-and-slab prior. Recall that, assuming prior equal odds, the PIP for each random effect corresponds to the Bayes factor (see Equation 3.4). Bayes factors tend to infinity and posterior model probabilities tend to one in favor of the "true" model as the sample size increases (O'Hagan, 1995). Therefore, with a sufficiently large sample size, the spike-and-slab approach will completely avoid false positives and false negatives, whereas the same cannot be said for random effect selection under the credible interval strategy.

Further, the classification results help clarify the trade-off in choosing different values for the PIP. Using a lower threshold, such as PIP ≥ 0.5, results in better sensitivity (i.e., detecting who *is not* average) at the cost of lower specificity (detecting who *is* average). As the PIP threshold

[7] Because these results may have been due to the discrete nature of the random effects, an alternative simulation study was conducted using continuous random effects, and its results can be found in the Appendix.

increases (e.g., PIP ≥ 0.75), this relationship reverses. Although not included in our results above, a similar relationship would be observed for the credible interval approach. Using a more narrow credible interval would result in higher sensitivity, at the cost of lower specificity, and vice versa for a wider interval. In studying variable selection, Li and Lin (2010) found that for a credible interval approach, a 50% CrI provided the best balance between sensitivity and specificity. Though such narrow intervals are not commonly used in psychological science, the Appendix contains the results from Study 1 using 50% CrIs instead of a 90% CrIs, but they do not shift the main conclusions from our results here. Taken together, our results here suggest that a strategy utilizing a spike-and-slab prior on the random effects is preferable to one using a customary normal prior on the random effects for detecting who is and is not "average".

3.3.2. Study 2. We now tackle the issue of double shrinkage in the random effects. Recall that the potential issue here is that the random effects may be biased towards zero due to shrinkage occurring both within the slab, as is typical in an ordinary mixed-effects model, and in the spike. In a customary random intercepts model with a normal prior on the random effects, the amount of shrinkage that occurs can be precisely determined through the so-called shrinkage factor, ω_j, which is given by

$$\lambda_j = \frac{\tau^2}{\tau^2 + \sigma^2/n_j} \tag{3.11}$$

$$\omega_j = 1 - \lambda_j. \tag{3.12}$$

Notice here that λ_j is calculated just as the ICC with the exception that the within-unit variance σ^2 is divided by n_j. Thus, holding n_j constant, larger ICCs imply smaller shrinkage factors and vice versa. Further, units with more observations will have smaller shrinkage factors. When all j units have equal observations ($n_1 = \cdots = n_j$), then there is a constant amount of shrinkage applied to all random effects ($\omega_1 = \cdots = \omega_j$).

When a spike-and-slab prior is instead placed on the random effects, determining the shrinkage involves an additional consideration. For every MCMC iteration, each random effect is either included (slab) or excluded (spike). All else being equal, the slab portion of the prior has the

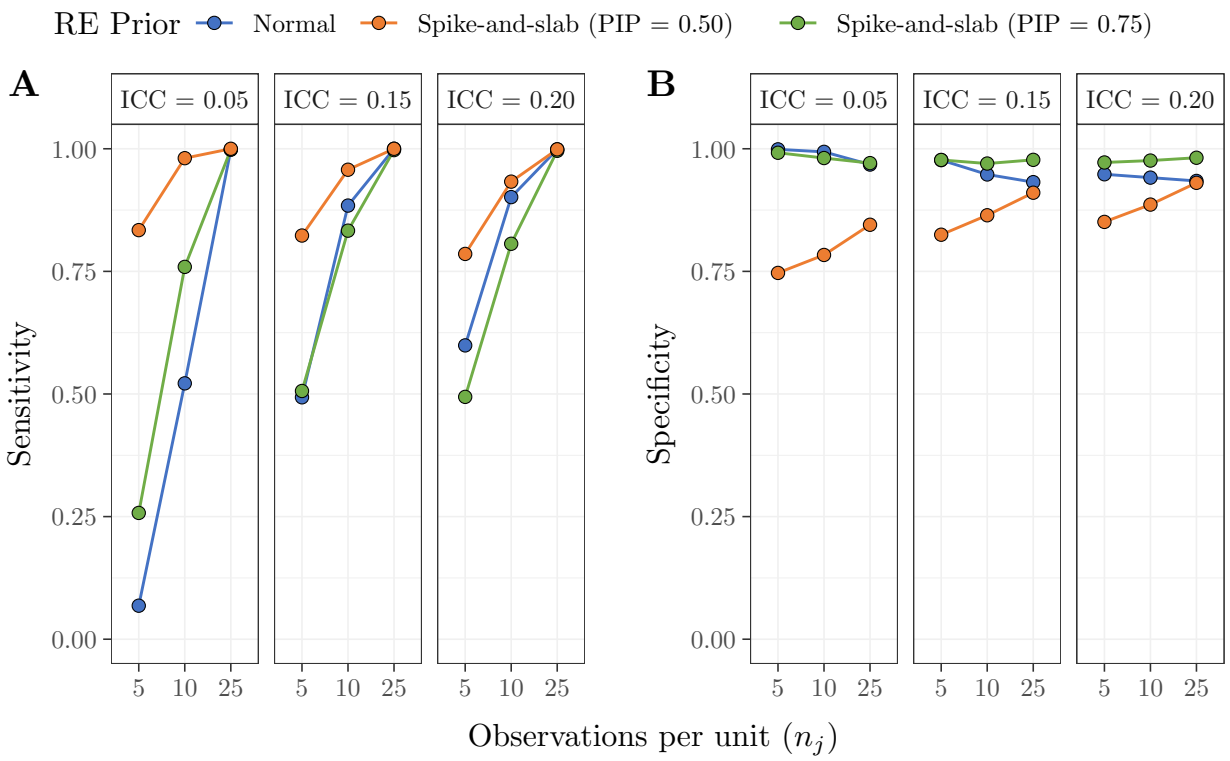

FIGURE 3.5. Classification rates of random effects under normal and spike-and-slab priors. For the normal prior, a 90% CrI was used to determine whether a random effect was "average" or not. For the spike-and-slab, two thresholds were used: a PIP of 0.50 and PIP of 0.75. A) Sensitivity between the three methods. Sensitivity tended to one as n_j increased for all methods, but the spike-and-slab combined with a PIP of 0.5 generally had the best sensitivity. B) Specificity between the three methods. Under the normal prior and 90% CrI, specificity was high with fewer n_j, but decreased to 0.9 as n_j increased. The spike-and-slab with a PIP of 0.75 had specificity of 1 or near 1 across all conditions, whereas using a PIP of 0.50 resulted in worse specificity. However, specificity still tended to 1 for the latter. This is a benefit of using the spike-and-slab prior — it will converge on the "true" model as the sample size grows.

effect of applying stronger shrinkage to larger random effects relative to smaller random effects. Conversely, the spike has the effect of subjecting small random effects to more extreme shrinkage, relative to larger random effects. Dropping the notational dependence on the iteration index s, λ_j is calculated in each MCMC iteration as a piecewise function of the form

$$\lambda_j = \begin{cases} 0 & \text{if } \gamma_j = 0 \\ \frac{\tau^2}{\tau^2 + \sigma^2/n_j} & \text{if } \gamma_j = 1 \end{cases}, \tag{3.13}$$

where γ_j denotes whether the jth random effect is included in the model. The final estimate for each λ_j can be calculated as the average of (3.13) across all MCMC iterations. Finally, the shrinkage factor can then be computed as $\omega_j = 1 - \lambda_j$. Because the posterior inclusion probability for the jth random effect is defined as the proportion of MCMC iterations where $\gamma_j = 1$, then keeping all else constant, using a spike-and-slab prior results in stronger shrinkage for estimates that have lower posterior inclusion probabilities.

With the shrinkage factors in hand, the estimate of each unit-specific intercept, α_j, can be computed by

$$\hat{\alpha}_j = \omega_j \cdot \bar{y}_{\cdot} + (1 - \omega_j) \cdot \bar{y}_j, \tag{3.14}$$

where $\bar{y}_{\cdot}$ indicates the grand mean of the outcome and $\bar{y}_j$ denotes the unit-specific mean of y. A shrinkage factor ω_j of 1 indicates total shrinkage towards the grand mean ($\hat{\alpha}_j = \bar{y}_{\cdot}$), and conversely, a shrinkage factor of zero indicates no shrinkage towards the grand mean ($\hat{\alpha}_j = \bar{y}_j$). By comparing the estimated α_j between mixed-effects models with normal and spike-and-slab priors, in addition to the shrinkage factors they produce, we can thoroughly investigate the impact of double shrinkage on the resulting random effects. To accomplish this, we followed the same set up as in Study 1. However, rather than focusing on the classification rates, we recorded the posterior estimates for the random intercept α_j and the shrinkage factors ω_j.

3.3.2.1. *Results.* The average estimates for the α_j are displayed in Figure 3.6. Columns differentiate between ICCs, rows differentiate between n_j, color differentiates between prior, and shape differentiates between (non-)zero random effects. The dashed line denotes $\alpha = 1$. As expected, the estimated α_j are subject to less shrinkage towards α as the ICC increases, and similarly, as n_j increases, regardless of the prior. Further, for units where $\theta_j = 0$, the estimated α_j were estimated

to be near the fixed effect α regardless of ICC, n_j, or prior. On the other hand, there were discrepancies in shrinkage between the spike-and-slab and normal priors when considering random effects that were set to either -1 or $+1$. For these random effects, the spike-and-slab prior often resulted in *less* shrinkage for the $\hat{\alpha}_j$ than the standard normal prior. For example, when the ICC was set to 0.05 and $n_j = 25$, the estimates $\hat{\alpha}_j$ for non-zero random effects were approximately 0.75 and 1.75 under the spike-and-slab prior. Meanwhile, the same estimates were roughly 0.5 and 1.5 under the normal prior. That is, the spike-and-slab prior allowed non-zero estimates to be closer to their actual values (0 and 2, respectively) than the normal prior. This result displays a nice property of the spike-and-slab in that the shrinkage is adaptive; larger effects receive little shrinkage whereas there is strong shrinkage for small effects (J. N. Rouder et al., 2018).

In order to better understand the differences in shrinkage between the priors, the average shrinkage factors ω_j for each condition are displayed in Table 3.1. As implied by the $\hat{\alpha}_j$ in Figure 3.6, the shrinkage factors decreased with increasing ICCs and n_j regardless of the prior that was used. Note, however, that the shrinkage factor under the normal prior is constant in each condition regardless of whether the random effect was actually equal to zero or not. Because shrinkage under the spike-and-slab prior is adaptive, the shrinkage factors were larger when $\theta_j = 0$ than when $\theta_j \neq 0$. Relatedly, under the spike-and-slab prior, there was relatively strong shrinkage for the random effects equal to zero, regardless of ICC or n_j, but for non-zero random effects, the shrinkage dissipated with increasing ICC and n_j. Generally speaking, the spike-and-slab prior applied more shrinkage to random effects that were truly zero and less shrinkage for non-zero random effects, relative to the normal prior.

Part of our results here are due to setting the prior inclusion probability for each random effect to 0.5 (see Equation 3.2). In practice, this is the most common choice because it expresses equal prior odds for whether a given random effect should be included or excluded from the model. Choosing alternative values would alter the amount of shrinkage observed in Figure 3.6 and Table 3.1. In practice, researchers applying the spike-and-slab prior to random effects should bear this in mind when setting the prior inclusion probabilities. To provide some intuition on the impact of choosing alternative values for the prior inclusion probabilities, we conducted additional simulation studies. The results are included in the Appendix.

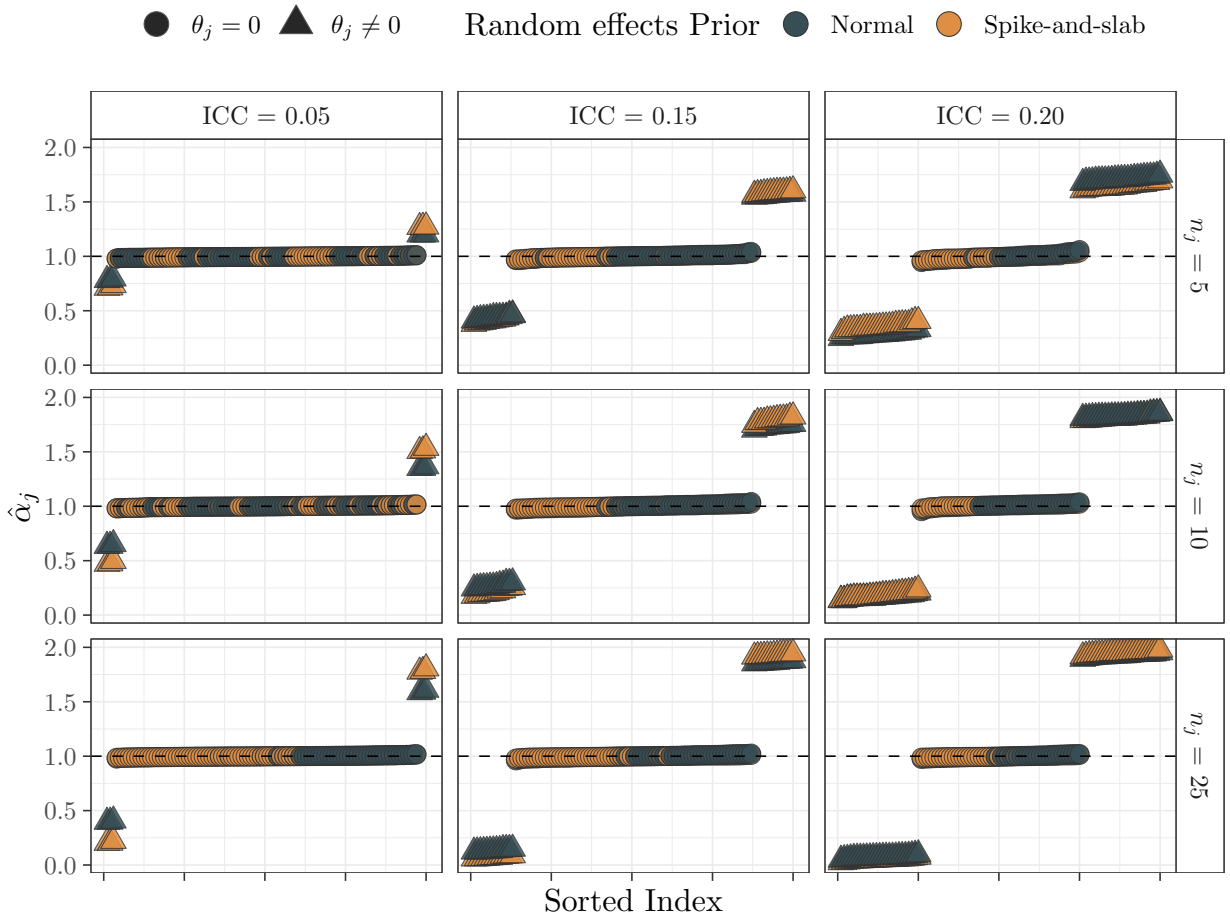

FIGURE 3.6. Estimates of the random intercepts α_j for mixed-effects models under normal and spike-and-slab priors, sorted in ascending order. The dashed lined denotes $\alpha = 1$. As expected, less shrinkage occurred as ICC and n_j increased regardless of prior, but there were differences in the amount of shrinkage. When the random effects were zero, the $\hat{\alpha}_j$ were highly similar between the priors across all conditions, but there were pronounced differences in the estimates for non-zero random effects. When $\theta_j \neq 0$, the spike-and-slab prior typically applied less shrinkage than the normal prior, such that estimates were closer to their true values. This is especially noticeable with smaller ICCs.

In summary, we observed that the double shrinkage induced from the spike-and-slab did not bias the random effects relative to a standard mixed-effects model by applying too much shrinkage. Rather, in many cases, the shrinkage applied by the spike-and-slab prior was preferable in that it applied weak shrinkage to non-zero effects and stronger shrinkage to truly zero random effects.

3.4. Discussion

In this work, we provided a general spike-and-slab formulation for random effect selection in mixed-effects models. The empirical application evidenced the utility of the proposed methodology for addressing individual differences in psychological science. Two simulation studies were conducted that illustrated key properties of the approach. Although spike-and-slab priors are not new in psychology research, their advantages were thought to be limited to exploratory variable selection and big-data contexts, such as fMRI analyses (J. N. Rouder et al., 2018). As we illustrated in this article, however, the spike-and-slab is also valuable in the context of "small-data" which is common in the social-behavioral sciences.

In the empirical application, we performed posterior predictive checks on the models for the cognitive tasks in order to inspect their adequacy in capturing important patterns in the data. While model checking is indeed an important part of statistical modeling, an additional motivation for performing the posterior predictive check was to address the concern of whether the spike-and-slab, "taken globally, [can] provide a good description of the structure in the data" (Haaf & Rouder, 2017, p. 794). As was shown in Figure 3.4, our formulation did a good job of describing the experimental effects, or mean differences between conditions. This ability of the spike-and-slab prior to provide trustworthy estimates was also observed in Study 2. Placing a spike-and-slab prior on the random effects does not necessarily compromise the model estimates.

The data we used in this paper came from experiments in psycholinguistics and cognitive psychology. We chose these data because: 1) they are typical representations of research that is done in the realm of individual differences with an emphasis on mixed-effects methodology and 2) data from cognitive tasks have been recently used in the context of reliability research. Mixed-effects models are routinely employed to analyze individual differences in this context. Given the history of individual differences in cognitive research, finding little individual differences in these tasks is somewhat unexpected. This perhaps points to the rather restrictive nature of the standard approach for probing individual differences in mixed-effects models. That is, if there is little between-subject variability, then a researcher might conclude that there are no individual differences. The spike-and-slab approach, in turn, offers a more nuanced view as it allows the differentiation between those who are and are not "average", even in low ICC settings. This was

TABLE 3.1. Average Shrinkage Factors (ω_j)

		Normal		Spike-and-Slab	
ICC	n_j	$\theta_j = 0$	$\theta_j \neq 0$	$\theta_j = 0$	$\theta_j \neq 0$
0.05	5	0.80	0.80	0.84	0.76
	10	0.65	0.65	0.74	0.51
	25	0.40	0.40	0.68	0.21
0.15	5	0.45	0.45	0.70	0.48
	10	0.28	0.28	0.68	0.25
	25	0.13	0.13	0.72	0.07
0.20	5	0.29	0.29	0.68	0.43
	10	0.17	0.17	0.69	0.23
	25	0.07	0.07	0.74	0.05

Note. Larger values indicate more shrinkage of the random effects towards zero. The shrinkage applied by the normal prior is constant regardless of whether $\theta_j = 0$ or $\theta_j \neq 0$, but the shrinkage applied by the spike-and-slab prior is adaptive.

clearly seen in Study 1, where the spike-and-slab prior had good performance in detecting non-average units even when the ICC was as low as 0.05 and in the empirical application, where over a quarter of the experimental effects for individuals did not conform with the average experimental effect.

3.4.1. Future Directions. An oft-overlooked aspect of mixed-effects models is that the residual variance (σ^2) and between-subject variance (τ^2) are considered to be constant across subjects. This can result in an improper amount of shrinkage (Hoff, 2009, Ch. 8), in essence, distorting the model estimates and their variability. This assumption can be relaxed so that the within- and between-subject variances can be allowed to vary as a function of predictors. Such models have been introduced to psychology under the name of mixed-effects location-scale models (Hedeker et al., 2012; Rast & Ferrer, 2018; Williams & Mulder, 2019). By allowing non-constant variances, individual differences may be more pronounced (Williams, Mulder, et al., 2020). Applying the spike-and-slab prior to the random effects in these models remains an interesting direction for future work because of the potential to tease apart individual differences in even finer detail.

The methodology we discuss in this paper also has promising potential in clinical fields. In this domain, there has been increasing interest in idiographic methods, or methods focused on

individuals (see for example, the models described in Piccirillo & Rodebaugh, 2019). The motivation for their use is often to identify individuals for whom a treatment may have different levels of efficacy. The use of mixed-effects models (and also mixed-effects location-scale models) in combination with spike-and-slab prior may provide an interesting avenue of research in idiographic studies because information is not lost by fitting separate models, but individuals who deviate from an average treatment effect may still be identified.

3.5. Summary

In this work, we discussed a general strategy to apply the spike-and-slab prior to the random effects in mixed-effects models for individual differences research. Importantly, this method allows researchers to gain a more nuanced view of individual differences than traditional approaches. By going beyond the testing of variance components to using the spike-and-slab for random effect selection, researchers can determine which individuals differ from an average effect. The methods discussed in article have been implemented in the R package **SSranef**.

CHAPTER 4

Discussion

4.1. Overview

The preceding chapters discussed distinct applications of statistical methodology that can be applied to elucidate rich and multifaceted inferences in psychological science. Importantly, the methodologies discussed are rooted in the rich soil of the Bayesian framework. In a Bayesian framework, a prior distribution, when combined with a likelihood function, yields a posterior distribution [1]. It is the flexibility afforded by these features of Bayesian analysis that allows each method discussed in this dissertation to enable statistical inferences that would otherwise be impractical, if not impossible, under traditional statistical methods. Below, each chapter is recapitulated and then discussed.

4.1.1. Chapter 1. Chapter 1 presented a unique framework that marries exploratory analyses with confirmatory hypothesis testing in partial correlation networks. The goal was to provide a means for researchers to formalize and test the intricate hypotheses that can arise from studying phenomena under a network approach. The first step in the proposed framework is an exploratory stage, where GGMs are estimated on an initial dataset with the sole purpose of generating hypotheses. We argue that, although such models have been traditionally proposed to be used in exploratory settings, they can and should also be considered in confirmatory settings. Thus, we delineate the second step in the framework, wherein the set of hypotheses developed from step 1 are simultaneously expressed via a mix of equality and inequality constraints on the edges in a GGM. Although hypotheses are often thought of in this way, there is a dearth of statistical methods that capture such constraints. Statistically, we capture ordered hypotheses via the so-called encompassing prior approach, where the encompassing prior is specified as a matrix-F distribution over the precision matrix of a multivariate normal distribution. By using the encompassing prior

[1] A normalizing constant is also needed in order for the posterior distribution to also be a proper probability distribution. However, this constant is not usually needed to arrive at a valid inference.

approach, it is possible to obtain Bayes Factors as measures of evidence for each hypothesis under investigation.

To demonstrate the utility of this framework, we presented several illustrative examples using psychopathology symptom networks. In each example, we begin with estimating a GGM that may be conceptualized as representing symptoms that collectively represent one or more disorders. The initial GGM's may be inspected and summarized using indices that describe its central structures, such as bridge nodes. We formulate various hypotheses, ranging from whether a given network structure can replicate (and to what degree) to which symptoms in a network may serve as the best targets in an intervention program. These examples provide support that the proposed approach can result in generating comprehensive hypotheses that can then be rigorously tested in a way that is infeasible outside of a Bayesian framework.

To better understand the behavior of the encompassing prior distribution, we conducted two simulation studies that together examined: (a) the behavior of the Bayes factor in relation to sample size, (b) the behavior of the Bayes factor in relation to the complexity of sample size, and (c) the sensitivity of the Bayes factor to the encompassing prior's variance. In the first simulation study, we formulated different *sets* of hypotheses, where hypotheses within a set varied in their complexity. The results showed that the Bayes factors for the "true" hypotheses tended to infinity with increasing sample size, and accordingly the encompassing prior will select the correct hypothesis in the limit. Further, the results revealed that fewer samples are needed to obtain large Bayes factors for a hypothesis when it is more complex, or specific, and it is supported by the data. In the second simulation study, we formulated a single hypothesis set, but introduced different conditions for the variance of the encompassing prior distribution used to test the hypotheses. The results indicated that the Bayes factor in favor of the true hypothesis was not affected by the choice of prior variance. This is a desirable quality for the prior because a main critique of using Bayes factors is that they can be overly sensitive to the choice of prior distributions.

4.1.2. Chapter 2. Chapter 2 introduces the concept of the Bayesian bootstrap as non-parametric method of sampling from the posterior distribution for various correlation coefficients commonly used in the social-behavioral sciences. Here, we aimed to provide the benefits of Bayesian inference — such as the probability of a parameter value, conditional on the data, or quantifying evidence in

favor of the null hypotheses — while overcoming the challenges inherent to estimating certain kinds of correlation coefficients. For example, there does not exist a suitable prior distribution for some correlation coefficients, or computational expense for obtaining posterior distributions can become prohibitively burdensome.

Due to the diversity of data types inherent to psychology, Pearson's, Spearman's, Gaussian rank, Kendall's τ, and polychoric correlations, and correlation matrices are often measures of associations that researchers want to estimate. Of these correlation types, only Pearson's correlation is easily estimated with a "normal" Bayesian approach. However, the Bayesian bootstrap provides a simple alternative to sampling the posterior of the remaining correlation types. Moreover, if the Bayesian bootstrap is applied to obtain a posterior distribution for the entire correlation matrix, then this facilitates making inferences on the differences between correlations. That is, not only can a point estimate and credible interval for, say, a Spearman's correlation be obtained, but a point estimate and credible interval for the difference between two Spearman's correlations can be obtained as well. Historically, the comparison of correlations in non-Bayesian frameworks has been challenging due to a lack of known sampling distributions for correlation differences. Further, it is possible to gain evidence for the *absence* of the correlation (i.e., evidence in favor of the null) by utilizing the Region of Practical Equivalence (ROPE) technique to the posterior of the difference of two correlations — another possible inference under a Bayesian approach that is difficult under a frequentist one.

We illustrate the method by revisting a series of correlation comparisons conducted in a previous study, but we instead use the Bayesian bootstrap approach to perform the comparisons. Here we found that the Bayesian bootstrap provided several advantages over the original analyses. First, we were able to use a more appropriate measure of association than the original study. Although the data represented ordinal variables, the original analysis estimated Pearson's correlations and consequently tested the difference between various Pearson's correlations. To our knowledge, Pearson's correlations are the only type of correlation for which tests of differences exist. With the Bayesian bootstrap, however, we were able to use a more appropriate measure of association (Kendall's τ) and still test the difference between these correlations. Using a ROPE approach, we found evidence for the equality of correlations — an inference that was not possible under the original analysis.

4.1.3. Chapter 3. In Chapter 3, we advocate for the spike-and-slab prior distribution as an effective prior for studying individual differences with mixed-effects models. By adopting this prior, researchers can ask — and subsequently address — the question of "who is average?" in the sense that they are represented by a population-average estimate. We consider that the typical approach to studying individual differences can be rather coarse in that it can be broadly concluded that individual differences either exist or they do not (with respect to a population-average estimate). We argue that there are more nuanced approaches to be considered. Specifically, it may be that *some* individuals are "average" while others are not.

Using data from two cognitive tasks, we study how the spike-and-slab prior can be used to formally capture this notion. These data were chosen because it has been argued that few individual differences exist in such data. We found that contrary to conventional opinion, there were many individual differences in task performance. Remarkably, we found that 25% of the sample in each dataset could be considered a "non-average" individual. In practice, this could allow researcher to identify which, and how many, individuals may be considered different from the average effect in a particular study. We concluded the empirical application of the method with a posterior predictive check to ensure that the spike-and-slab model fit the data well, therefore validating the plausibility of the individual differences we described.

We conducted two simulation studies to understand: (1) whether the spike-and-slab prior accurately identifies average and non-average individuals and (2) whether using the spike-and-slab prior could result in poor model parameter estimates due to "double shrinkage". A Gaussian prior distribution (as is used in a standard mixed-effects model) was used for comparison in both studies. The results of study 1 indicated that with many observations per individual, the spike-and-slab prior and a Gaussian prior distribution could be used to achieve similar levels of accuracy in identifying non-average individuals. However, in situations with few observations per individual or with low-between person variance, using the spike-and-slab prior resulted in superior classification accuracy. This simulation also showed that the spike-and-slab prior had the desirable property of model selection consistency. The results of study two indicated that the "double shrinkage" did not negatively impact the model estimates, and in some cases, was actually beneficial. That is, the spike-and-slab prior resulted in more accurate parameter estimates for "non-average" individuals

than the standard Gaussian prior. This result illustrated that the shrinkage under a spike-and-slab prior is adaptive; larger effects receive little shrinkage whereas there is strong shrinkage for small effects, which is desirable quality.

4.2. The Flexibility of a Bayesian Analysis

The methods described above were all predicated on a Bayesian approach to statistics. Although the wide-ranging benefits of a Bayesian approach, especially as applied in psychology, have been extensively discussed (e.g., Etz et al., 2018; Kruschke & Liddell, 2018; J. N. Rouder et al., 2018; Wagenmakers et al., 2018), we did not develop the methods in this dissertation merely because they are Bayesian and carry Bayesian advantages. Rather, we were driven by a belief that the Bayesian machinery allows a level of flexibility that cannot be found using traditional methods, and that this flexibility could be leveraged to craft methods with the explicit intent of helping psychology researchers obtain inferences that transcend conventional boundaries. In my view, there are two aspects of a Bayesian analysis that facilitate this flexibility in these methods. First and foremost is the ability to specify a prior distribution that reflects a researcher's knowledge, and the second is the ability to perform inferences on arbitrary quantities of interest based on posterior distributions.

4.2.1. The Prior Distribution. The prior distribution is perhaps the greatest advantage of a Bayesian analysis. The prior distribution reflects uncertainty, or a researcher's beliefs, about a model's parameters, prior to observing data that will be analyzed. In practice, this often takes the form of specifying a conventional parametric distribution (i.e., a Normal distribution), and perhaps constraining what values would be deemed "plausible" under that distribution. For example, suppose a researcher posits a Normal $(0, 1)$ prior in an effort to estimate a population mean. This prior places 95% of the prior density over $[-1.96, 1.96]$, and reflects the researcher's conviction that the most likely value for the population mean is 0, but values up to ± 1.96 are also reasonable, and a population mean of 5 is virtually impossible. But a prior distribution offers more utility than this.

In Chapter 1, the scale of encompassing prior can be set to reflect a researcher's beliefs about plausible values for the edge weights in a Gaussian Graphical Model, but the encompassing prior (in conjunction with a set of prior model probabilities) can also be used to reflect beliefs about the *ordering* of edges in a GGM. This augments the conventional use of the prior in two ways. First, it

adds a layer of knowledge that can be incorporated into an analysis. Instead of only reflecting beliefs about the values of edge weights, beliefs about the structure between edges can be incorporated, and thus the prior reflects a more holistic view of the network to be estimated. Second, it may simply be easier to reason at the level of order than at the level of specific parameter value. Even if a researcher has difficulty incorporating prior knowledge about plausible edge weights, they may still incorporate prior knowledge through the ordering of edge weights.

A similar rationale applies to our proposal of the spike-and-slab prior discussed in Chapter 3. Recall that the spike-and-slab prior was a two-component mixture: a point-mass spike centered at zero and a diffuse slab that captures non-zero values. This prior too reflects a belief beyond those concerning parameter values. In this case, this prior explicitly instantiates the belief that some individuals are average (the spike) and that some individuals are not average (the slab). The flexibility to integrate such beliefs or problem-specific knowledge through the prior distinguishes Bayesian inference from simply an analytical framework, but as a tool to foster richer and more comprehensive inferences than those obtained via traditional statistical methods.

4.2.2. Inference on Arbitrary Quantities. A noteworthy feature of having a joint posterior distribution for multiple variables is the capacity to perform inferences on quantities derived from the posterior. To illustrate, suppose a researcher runs a multiple regression with two predictor variables, and thus obtains two estimates of the (standardized) coefficients β_1 and β_2. It is common to follow up by comparing the absolute values of the coefficients to determine which of the predictors is more important. A more rigorous approach would be to subject the coefficients to a statistical test to establish whether they significantly differ. In a Bayesian framework, testing this difference can be accomplished by taking samples from the posterior distribution of β_1 and subtracting from them samples from the posterior distribution of β_2 to obtain a posterior for $\delta = \beta_1 - \beta_2$. Information can then be extracted from this new posterior as usual, including point estimates or credible intervals.

In Chapter 2, the posterior distributions obtained via the Bayesian bootstrap for different correlation coefficients are used to make comparisons in precisely this manner. However, unlike the example above for regression coefficients, sampling distributions for the difference of two correlation

coefficients are severely limited [2]. Thus, inferences can be achieved for quantities that cannot otherwise be easily evaluated, but may anyhow be of interest to a researcher. This idea is not limited to differences between parameter estimates. For instance, study 2 in Chapter 3 examined the degree of shrinkage applied by both the spike-and-slab prior and a Normal prior. To do so, we evaluated the posterior distribution for the shrinkage factor for a set of j random effects, ω_j, which required taking the posterior for different variance parameters and then calculating a posterior for their ratio. Once again, obtaining inferences on such quantities that do not have readily available sampling distributions is made easy by the flexibility of a Bayesian approach.

4.3. Conclusion

In conclusion, this dissertation has explored a range of statistical methodologies within a Bayesian framework, each offering unique insights and advantages for advancing research in psychological science. These methodologies have been designed to provide flexible tools for researchers to address complex questions that may be impractical or impossible to tackle using traditional statistical methods.

In Chapter 1, we introduced a novel framework for integrating exploratory and confirmatory analyses in partial correlation networks. This approach allows researchers to formalize and test intricate hypotheses arising from network-based phenomena. By utilizing an encompassing prior distribution, we demonstrated the capability of generating comprehensive hypotheses and rigorously testing them with Bayes factors. This methodology offers a valuable tool for researchers studying complex interrelationships in various domains of psychology. Chapter 2 introduced the Bayesian bootstrap as a powerful non-parametric method for estimating correlation coefficients commonly used in psychological research. This approach allows for Bayesian estimation for correlations that are difficult to estimate in a Bayesian setting, and further equips researchers with the tooling to compare those correlations. In Chapter 3, we explored the spike-and-slab prior in the context of

[2] One can reasonably object that this is not necessarily a Bayesian feature and that a standard bootstrap can accomplish much of the same. However, the standard bootstrap does not bring with it the interpretational advantages of a posterior distribution.

mixed-effects models, providing a strategy to identify and analyze individual differences in psychological research. This method offers a nuanced approach to understanding who does and does not deviate from the average as a way of adding nuance to the study of individual differences.

The methodologies presented in this dissertation illustrate the advantages of a Bayesian framework in addressing complex challenges within psychological science. By harnessing the flexibility of Bayesian analysis, researchers can advance their understanding of intricate phenomena with richer inferences that may better suit their research needs. These tools empower researchers to extract more out of their data, and ultimately enhance the quality and depth of knowledge psychological science.

Appendix A

Node	Symptom	Community
B1	Intrusive thoughts	Re-Experiencing
B2	Nightmares	Re-Experiencing
B3	Flashbacks	Re-Experiencing
B4	Physiological/psychological reactivity	Re-experiencing
C1	Avoidance of thoughts	Avoidance
C2	Avoidance of situations	Avoidance
C3	Amnesia	Avoidance
C4	Disinterest in activities	Avoidance
C5	Feeling detached	Avoidance
C6	Emotional numbing	Avoidance
C7	Foreshortened future	Avoidance
D1	Sleep problems	Arousal
D2	Irritability	Arousal
D3	Concentration problems	Arousal
D4	Hypervigilance	Arousal
D5	Startle response	Arousal

TABLE A.1. Definitions for nodes in Figure 1.3.

Node	Symptom	Community
D1	Lower interest or pleasure	Depression
D2	Feeling down, hopeless	Depression
D3	Trouble sleeping	Depression
D4	Tired or little energy	Depression
D5	Poor appetite/overeating	Depression
D6	Guilt	Depression
D7	Trouble concentrating	Depression
D8	Moving slowly/restless	Depression
D9	Suicidal thoughts	Depression
A1	Nervous, anxious, on edge	Anxiety
A2	Uncontrollable worry	Anxiety
A3	Worry about different things	Anxiety
A4	Trouble relaxing	Anxiety
A5	Restless	Anxiety
A6	Irritable	Anxiety
A7	Afraid something awful might happen	Anxiety

TABLE A.2. Node definitions for Figure 1.4.

Appendix B

Following Rubin (1981), let $\boldsymbol{d} = (d_1, \ldots, d_K)'$ be the vector of all K possible distinct values in $\mathbf{x} = (x_1, \ldots, x_n)'$ and let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)'$ be a vector of probabilities associated with $\boldsymbol{d}$ such that

$$p(x_i = d_k|\boldsymbol{\theta}) = \theta_k, \quad i = 1 \ldots, n;\ k = 1, \ldots, K, \tag{B.1}$$

and the sum of all probabilities equal one. If $\mathbf{x}$ is an i.i.d. sample from (B.1) and n_k is the number of values in $\mathbf{x}$ equal to d_k, then the prior for $\boldsymbol{\theta}$ under the Bayesian bootstrap is the so called Haldane prior (Haldane, 1932)

$$p\left(\boldsymbol{\theta}\right) \propto \prod_{k=1}^{K} \theta_k^{-1}, \tag{B.2}$$

and corresponds to the improper prior Dirichlet distribution $\text{Dir}(\boldsymbol{\alpha})$ with $\boldsymbol{\alpha} = (0, \ldots, 0)$. When this prior is combined with a multinomial likelihood, it yields a posterior for $\boldsymbol{\theta}$ which follows the Dirichlet distribution with $\boldsymbol{\alpha} = (1, \ldots, 1)$, that is,

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{x}) &\propto p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \\ &\propto \prod_{k=1}^{K} \theta_k^{n_k} \prod_{k=1}^{K} \theta_k^{-1} \\ &\propto \prod_{k=1}^{K} \theta_k^{n_k - 1}. \end{aligned} \tag{B.3}$$

A BB prior distribution (using $\alpha_i = 0.1$) and a corresponding posterior distribution are plotted in Figure B.1. As can be seen, the prior mass is mostly placed over probabilities near zero and one.

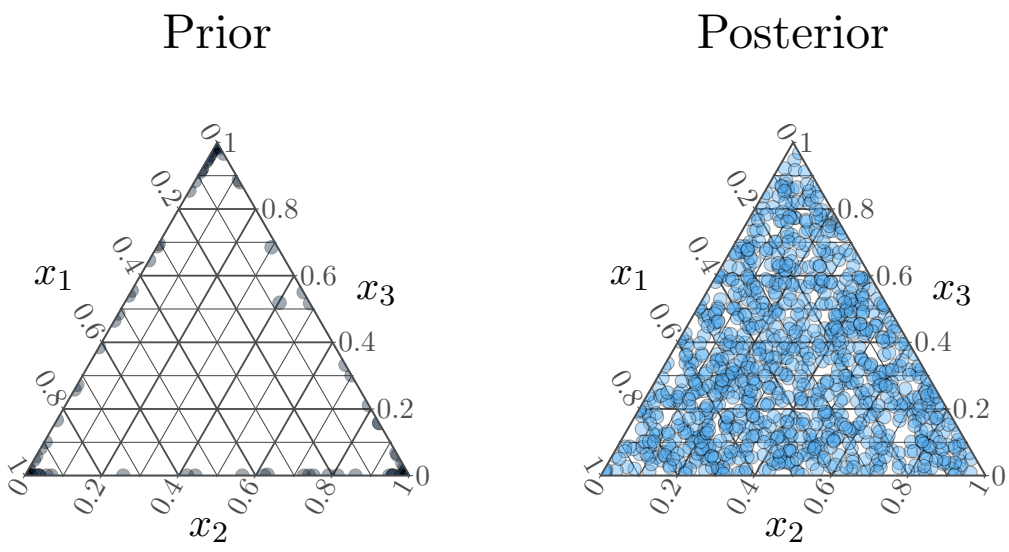

FIGURE B.1. Ternary plots of the prior (left) and posterior (right) distributions for the parameter $\boldsymbol{\theta}$ under the Bayesian bootstrap for three observations.

In the limit, as all $\alpha_i \to 0$, there is zero mass placed over θ's for unobserved data. The posterior distribution places mass uniformly on [0, 1] which indicates that any combination of θ's for the observed values is equally likely.

Appendix C

C.1. Normally Distributed Random Effects

In Study 1, the random effects were generated in a discrete manner to better assess their classification under a spike-and-slab prior. We repeated this simulation study such that all procedures remained the same, but instead of assigning the non-zero random effects values of ± 1, they were drawn from a standard normal distribution so that very small non-zero random effects would be introduced. Generating the random effects in this way reflects the commonly made assumption about the normality of random effects in mixed-effects models. Since random effects near zero are are likely to be absorbed by the spike component of the spike-and-slab prior, its classification performance depends on how well the θ_j's can be distinguished from zero (George & McCulloch, 1997). Therefore, a drop in sensitivity for the spike-and-slab may be expected with normally distributed random effects because small non-zero values may be considered to be zero.

The results are shown in Figure C.1. We once again found the $SS_{0.5}$ model to have the best sensitivity with few n_j and a small ICC, but it was the credible interval strategy under the normal prior that had superior sensitivity in all other conditions. As n_j increased, though, so did the the sensitivity for all three strategies. Although the normal prior generally had the best sensitivity with normally distributed random effects, it still fared poorly with respect to specificity. It was the $SS_{0.75}$ strategy that was the best in this regard. Of particular importance is that the model selection consistency property of the spike-and-slab prior was retained whereas it was still not applicable random effect selection with the credible interval approach. That is, both the sensitivity and the specificity tended to 1 under the spike-and-slab prior as the observations per unit increased, but specificity decreased with increasing n_j under the credible interval approach. The results from this simulation study suggest that if it the random effects are truly normally distributed and the goal is explicitly to maximize sensitivity, then using a credible interval to select non-zero random

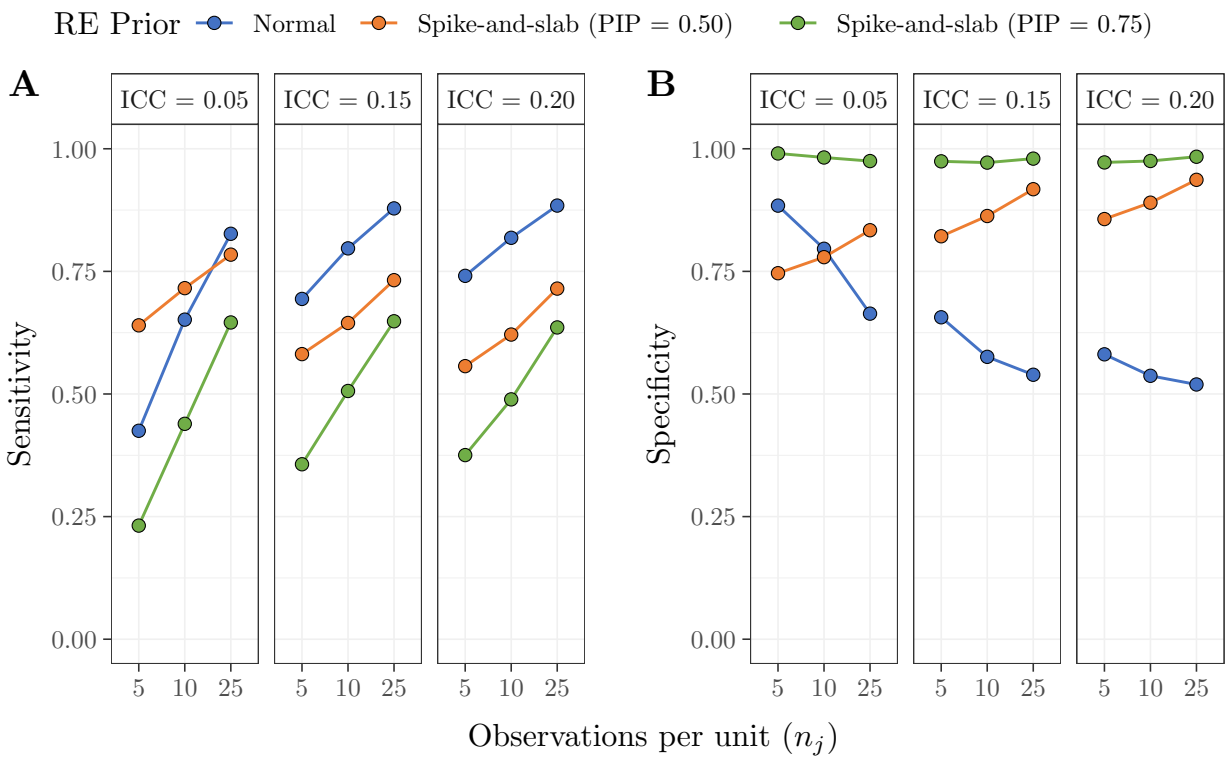

FIGURE C.1. Results from repeating Study 1, but with the random effects generated from a normal distribution.

effects may be used (at the cost of an increased false positive rate). If a balance of good sensitivity and specificity is instead desired, then random effect selection with the spike-and-slab prior is preferable.

C.2. Credible Interval Width

Figure C.2 displays the results from Study 1, but using 50% CrIs instead of 90% CrIs for the standard mixed-effects model. As described in the main text, random effect selection with narrower intervals leads to higher sensitivity in an exchange for lower specificity, and vice versa for wider intervals. Because it has previously been argued that a 50% CrI provides the best balance between sensitivity and specificity (Li & Lin, 2010), we compared the performance of a 50% CrI strategy for random effects selection to the spike-and-slab prior with PIP cut-offs of 0.5 ($SS_{0.5}$) and 0.75 ($SS_{0.75}$). However, the core conclusions from Study 1 did not change. In terms of sensitivity, the $SS_{0.5}$ model had the best sensitivity for the lowest ICC condition (ICC = 0.05), but now the CrI

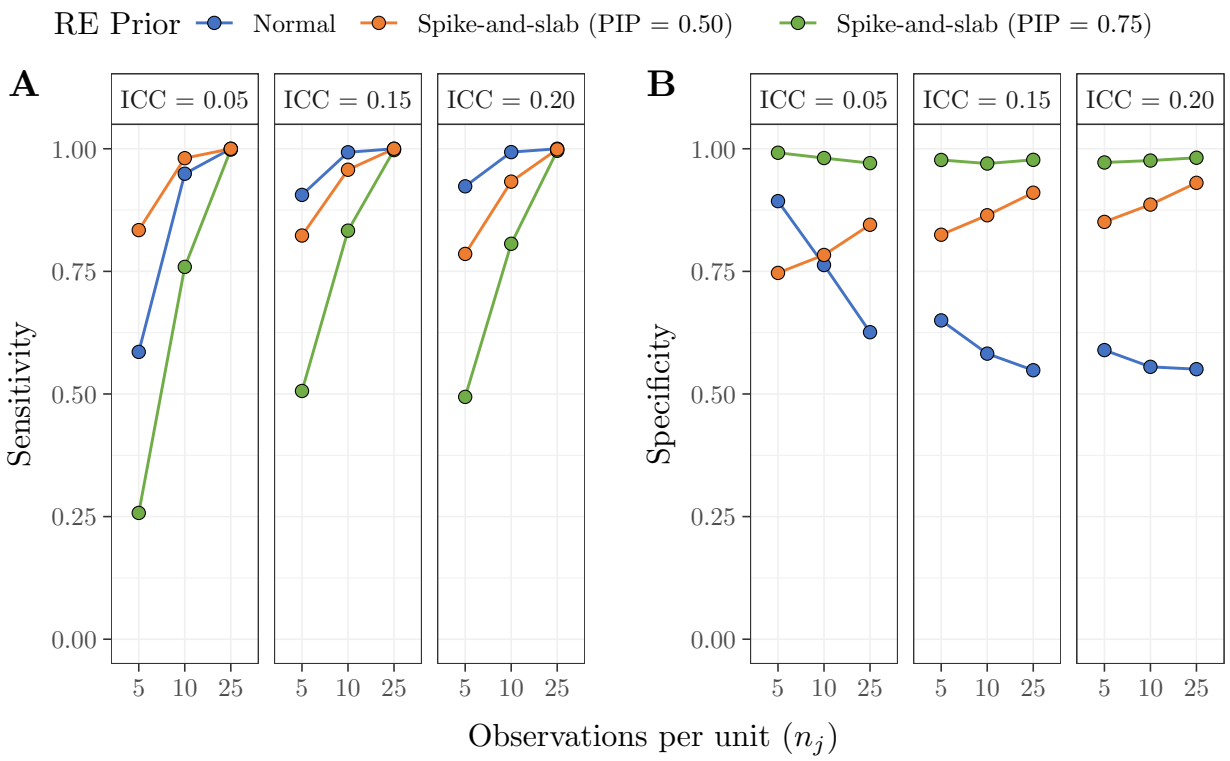

FIGURE C.2. Results from Study 1 with 50% CrIs instead of 90% CrIs for the model with the normal prior.

strategy had superior sensitivity as the ICC increased. In terms of specificity, though, the 50% CrI strategy performed worse than both spike-and-slab models in all but one condition (when the ICC was set to 0.05 and the observations per unit was set to 5). Importantly, the key difference remains that the spike-and-slab models are model selection consistent and will converge on the "true" model with increasing sample size while this does not hold for the credible interval strategy, regardless of the width that is chosen.

C.3. Varying Prior Inclusion Probabilities

To provide further intuition on the role of the prior inclusion probability of the random effects in classification and shrinkage of the random effects, we repeated Study 1 and Study 2 twice each. Once with a prior inclusion probability of 0.2 for all random effects and once with a prior inclusion probability of 0.8 for all random effects. The results for classification performance can be viewed in Figures C.3 and C.4, and the shrinkage results can be viewed in in Figures C.5 and C.6.

As shown in Figure C.3, reducing the prior inclusion probability had the effect of reducing the sensitivity for both spike-and-slab strategies. This resulted in the credible interval selection strategy having the highest sensitivity across all conditions. However, both models under the spike-and-slab prior ($SS_{0.5}$ and $SS_{0.75}$) had a specificity near one. The reverse pattern was true when the prior inclusion probability was 0.8 (Figure C.4). Here, the $SS_{0.5}$ model had a sensitivity of near 1 in all conditions, but a specificity of 0 in nearly all conditions. Again, the model selection consistency property of the spike-and-slab prior was observed. Together, these results show that the prior inclusion probability may be used to adjust the trade-off between sensitivity and specificity in classifying the average and non-average random effects.

With respect to the amount of shrinkage incurred by spike-and-slab and normal priors, many of the same general trends emerged occurred as in Study 2. Specifically, regardless of the prior, shrinkage decreased as n_j increased and as the ICC increased. Although there were again differences in shrinkage between the normal and spike-and-slab priors, they showed a different pattern than in Study 2. For instance, the top right panel in Figure C.5 shows that when the prior inclusion probability was 0.2, there was much stronger shrinkage induced under the spike-and-slab prior than the normal prior whereas in Study 2, there was not much difference in shrinkage between the priors. The reason for this additional shrinkage concerns the incongruence between the prior inclusion probability and the proportion of truly non-zeros. Setting the prior inclusion probability to 0.2 reflects that the expected a priori proportion of non-zero random effects is 0.2, but the data were generated such that the actual proportion of non-zero random effects was 0.5. The result of this mismatch is that more zeros were drawn during MCMC sampling for the γ_j's — reflecting the low inclusion probability — and in turn, this induced more shrinkage in the random effects (see Equation 3.13). This can be contrasted with the top left panel of Figure C.5, where the prior inclusion probability (0.2) was higher than the proportion of non-zero random effects (0.06). Here, the amount shrinkage is *less* than normal prior (mirroring the result in Study 2).

Figure C.6 shows the case for prior inclusion probabilities of 0.8 and so the prior expected proportion of non-zero random effects was higher than the actual proportion of non-zeros in all conditions. Now, the top right panel shows that the shrinkage is almost identical between the normal and spike-and-slab priors. Because the prior probability was relatively high, values of one

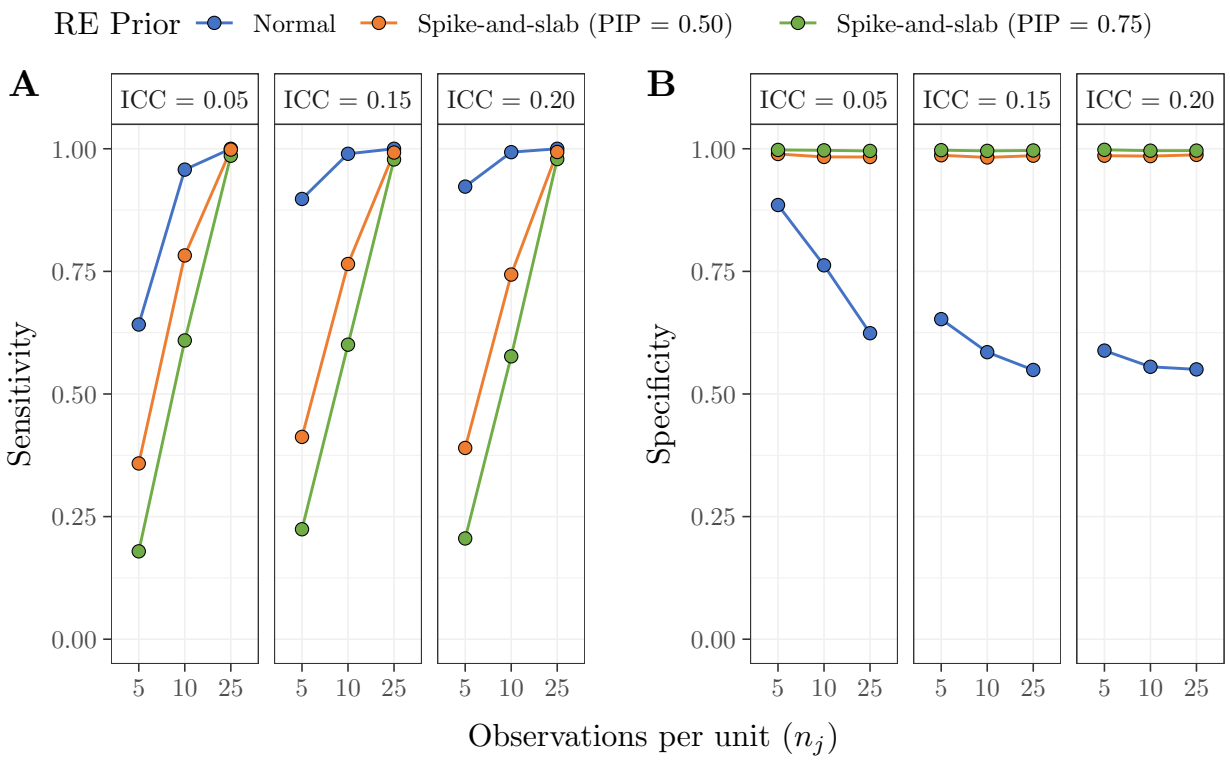

FIGURE C.3. Sensitivity and specificity for prior inclusion probabilities of 0.2.

were sampled more often for the γ_j's, and thus the amount of shrinkage induced under the spike-and-slab prior was nearly identical to that of the normal prior. These findings suggest that the spike-and-slab produces more shrinkage on the random effects than the normal prior when the prior inclusion probability is smaller than the true proportion of non-zero random effects, and less shrinkage when the prior inclusion probability is greater than the true proportion of non-zero random effects. Crucially, though, the influence of the prior inclusion probability vanished with increasing sample size n_j. This was especially pronounced in the lower right panels of Figures C.5 and C.6, where there was no shrinkage applied to the non-zero random effects, regardless of the prior inclusion probability.

C.4. Example Code

LISTING C.1. Example R code

```
# install package
```

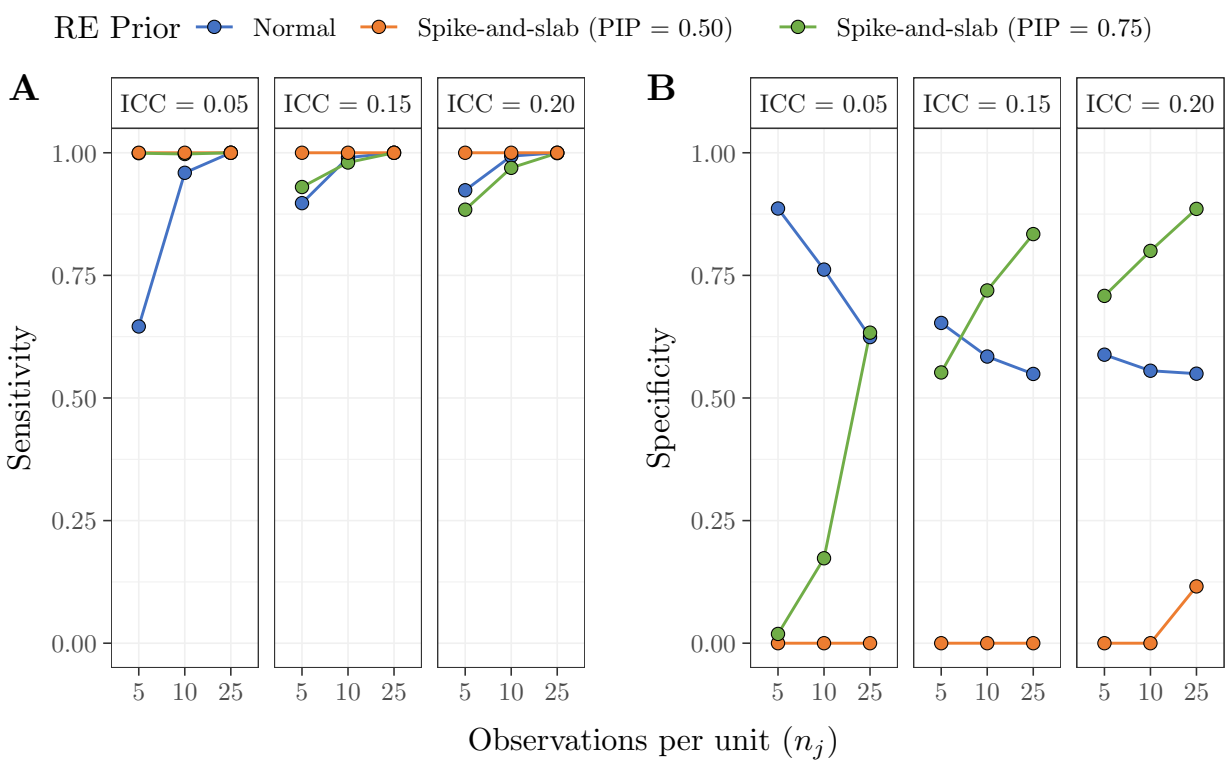

FIGURE C.4. Sensitivity and specificity for prior inclusion probabilities of 0.8.

```
remotes::install_github("josue-rodriguez/SSranef")


# fit mixed-effects model with spike-and-slab prior
# on random slopes
fit1 <- ss_ranef_beta(y = stroop$rt,
                      X = stroop$congruent,
                      unit = stroop$id)



# extract PIPs and calculate proportion
# of sample that differed from average
pips <- ranef_summary(fit1)$PIP
n_non_avg <- sum(pips > 0.5, na.rm = T)
```

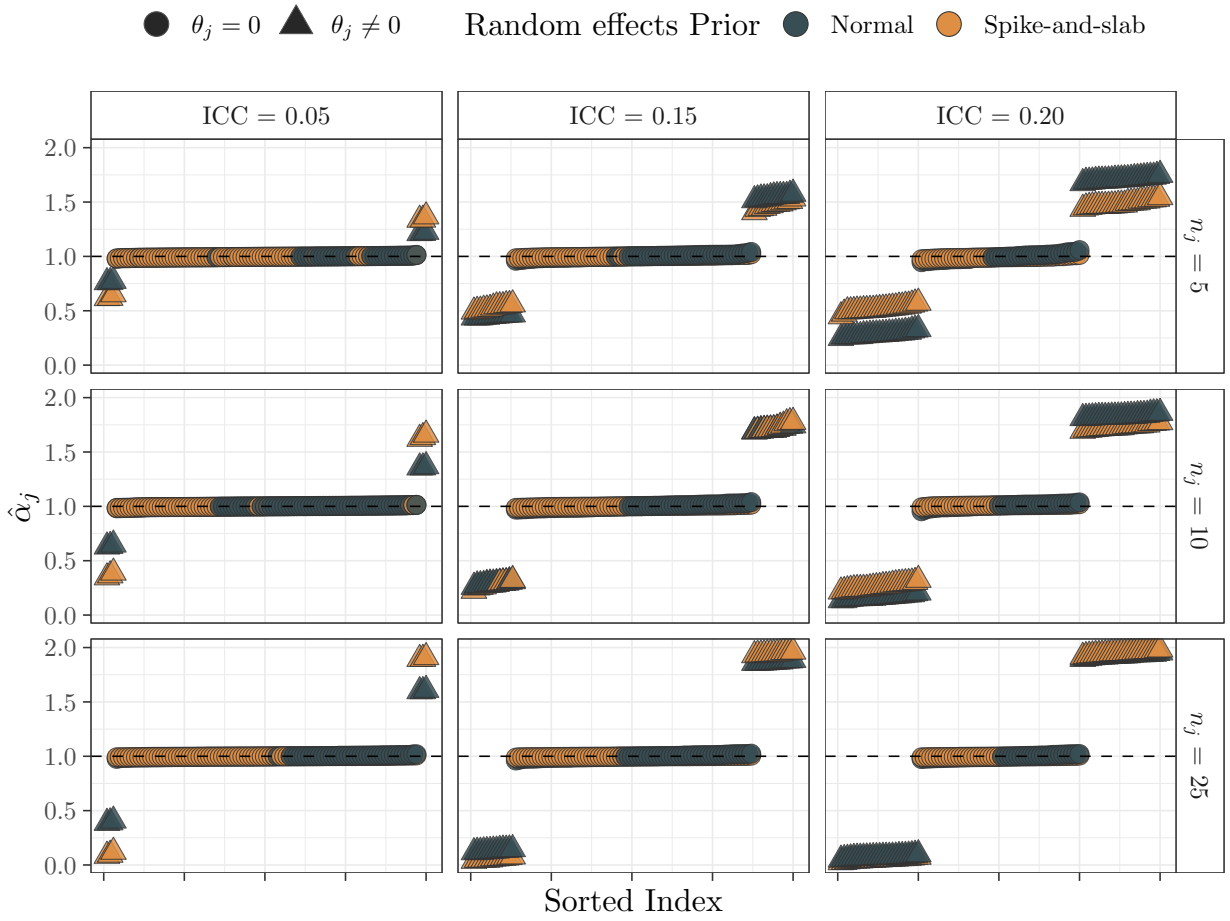

FIGURE C.5. Estimates of the random intercepts α_j for mixed-effects models under normal and spike-and-slab priors. The dashed lined denotes $\alpha = 1$. The prior inclusion probability for each random effect was set to 0.2.

```
n_total <- length(unique(stroop$id))

n_non_avg / n_total



 # re-fit model with different prior inclusion

 # probability for random effects

 priors <- list(gamma = "gamma[j] ~ dbern(0.8)")

 fit2 <- ss_ranef_beta(y = stroop$rt,
```

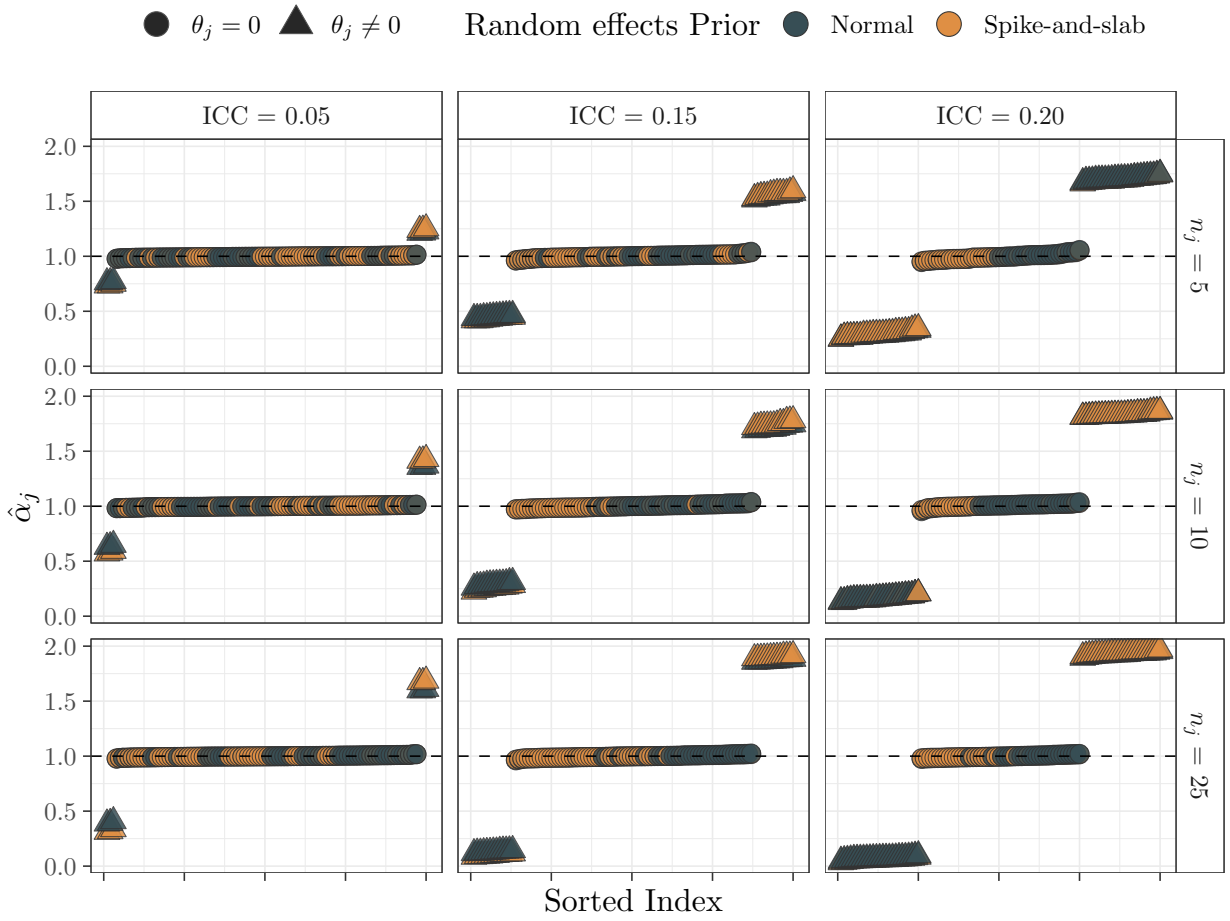

FIGURE C.6. Estimates of the random intercepts α_j for mixed-effects models under normal and spike-and-slab priors. The dashed lined denotes $\alpha = 1$. The prior inclusion probability for each random effect was set to 0.8.

```
                X = stroop$congruent,
                unit = stroop$id,
                prior = priors)
```

Bibliography

Aarts, E., Verhage, M., Veenvliet, J. V., Dolan, C. V., & van der Sluis, S. (2014). A solution to dependency: Using multilevel analysis to accommodate nested data. *Nature Neuroscience*, *17*(4), 491–496. https://doi.org/10.1038/nn.3648

Afzali, M. H., Sunderland, M., Teesson, M., Carragher, N., Mills, K., & Slade, T. (2017). A network approach to the comorbidity between posttraumatic stress disorder and major depressive disorder: The role of overlapping symptoms. *Journal of Affective Disorders*, *208*, 490–496. https://doi.org/10.1016/j.jad.2016.10.037

Albert, J. H. (1992). Bayesian estimation of the polychoric correlation coefficient. *Journal of Statistical Computation and Simulation*, *44*(1-2), 47–61. https://doi.org/10.1080/00949659208811448

Altay, G., & Emmert-Streib, F. (2010). Inferring the conservative causal core of gene regulatory networks. *BMC Systems Biology*, *4*(1), 132. https://doi.org/10.1186/1752-0509-4-132

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington DC.

Anderson, M., & Magruder, J. (2017, June). *Split-Sample Strategies for Avoiding False Discoveries* (w23544). National Bureau of Economic Research. Cambridge, MA. https://doi.org/10.3386/w23544

Armour, C., Fried, E. I., Deserno, M. K., Tsai, J., & Pietrzak, R. H. (2017). A network analysis of DSM-5 posttraumatic stress disorder symptoms and correlates in U.S. military veterans. *Journal of Anxiety Disorders*, *45*, 49–59. https://doi.org/10.1016/j.janxdis.2016.11.008

Bailey, P., Emad, A., Zhang, T., Xie, Q., & Sikali, E. (2018). Weighted and unweighted correlation methods for large-scale educational assessment: wCorr formulas. AIR–NAEP working paper no. 2018-01. NCES data R project series #02. *American Institutes for Research*.

Barbieri, M. M., & Berger, J. O. (2004). Optimal predictive model selection. *The Annals of Statistics*, *32*(3), 870–897. https://doi.org/10.1214/009053604000000238

Barbieri, M. M., Berger, J. O., George, E. I., & Ročková, V. (2021). The Median Probability Model and Correlated Variables. *Bayesian Analysis*, *16*(4), 1085–1112. https://doi.org/10.1214/20-BA1249

Barnard, J., McCulloch, R., & Meng, X.-L. (2000). Modeling Covariance Matrices in Terms of Standard Deviations and Correlations, With Application to Shrinkage. *Statistica Sinica*, *10*(4), 1281–1311. Retrieved November 3, 2019, from https://www.jstor.org/stable/24306780

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, *68*(3). https://doi.org/10.1016/j.jml.2012.11.001

Beard, C., Millner, A. J., Forgeard, M. J. C., Fried, E. I., Hsu, K. J., Treadway, M. T., Leonard, C. V., Kertz, S. J., & Björgvinsson, T. (2016). Network analysis of depression and anxiety symptom relationships in a psychiatric sample. *Psychological Medicine*, *46*(16), 3359–3369. https://doi.org/10.1017/S0033291716002300

Bem, D. J. (2004). Writing the empirical journal article. In *The compleat academic: A career guide, 2nd ed* (pp. 185–219). American Psychological Association.

Bentler, P. M. (1980). Multivariate Analysis with Latent Variables: Causal Modeling. *Annual Review of Psychology*, *31*(1), 419–456. https://doi.org/10.1146/annurev.ps.31.020180.002223

Berger, J. O. (2006). The case for objective Bayesian analysis. *Bayesian Analysis*, *1*(3), 385–402. https://doi.org/10.1214/06-BA115

Berger, J. O., & Wolpert, R. L. (1988). *The Likelihood Principle*. IMS.

Bono, R., Alarcón, R., & Blanca, M. J. (2021). Report Quality of Generalized Linear Mixed Models in Psychology: A Systematic Review. *Frontiers in Psychology*, *12*, 1345. https://doi.org/10.3389/fpsyg.2021.666182

Borsboom, D. (2017). A network theory of mental disorders. *World Psychiatry*, *16*(1), 5–13. https://doi.org/10.1002/wps.20375

Borsboom, D., Cramer, A. O. J., Schmittmann, V. D., Epskamp, S., & Waldorp, L. J. (2011). The Small World of Psychopathology. *PLoS ONE*, *6*(11). https://doi.org/10.1371/journal.pone.0027407

Borsboom, D., & Cramer, A. O. (2013). Network Analysis: An Integrative Approach to the Structure of Psychopathology. *Annual Review of Clinical Psychology*, *9*(1), 91–121. https://doi.org/10.1146/annurev-clinpsy-050212-185608

Borsboom, D., van der Maas, H., Dalege, J., Kievit, R., & Haig, B. (2020, February 29). *Theory Construction Methodology: A practical framework for theory formation in psychology* (preprint). PsyArXiv. https://doi.org/10.31234/osf.io/w5tp8

Boschloo, L., van Borkulo, C. D., Borsboom, D., & Schoevers, R. A. (2016). A Prospective Study on How Symptoms in a Network Predict the Onset of Depression. *Psychotherapy and Psychosomatics*, *85*(3), 183–184. https://doi.org/10.1159/000442001

Boudt, K., Cornelissen, J., & Croux, C. (2012). The Gaussian rank correlation estimator: Robustness properties. *Statistics and Computing*, *22*(2), 471–483. https://doi.org/10.1007/s11222-011-9237-0

Bringmann, L. F., Elmer, T., Epskamp, S., Krause, R. W., Schoch, D., Wichers, M., Wigman, J. T. W., & Snippe, E. (2019). What do centrality measures measure in psychological networks? *Journal of Abnormal Psychology*, *128*(8), 892–903. https://doi.org/10.1037/abn0000446

Bringmann, L. F., & Eronen, M. I. (2018). Don't blame the model: Reconsidering the network approach to psychopathology. *Psychological Review*, *125*(4), 606–615. https://doi.org/10.1037/rev0000108

Brooks, S., & Gelman, A. (1998). Some issues for monitoring convergence of iterative simulations. *Computing Science and Statistics*, 30–36. Retrieved January 17, 2020, from http://pages.cs.aueb.gr/~yiannisk/CVMCMC/CntrlVrtsPapers/brooks-gelman.pdf

Caplan, S., Hafri, A., & Trueswell, J. C. (2021). Now You Hear Me, Later You Don't: The Immediacy of Linguistic Computation and the Representation of Speech. *Psychological Science*, *32*(3), 410–423. https://doi.org/10.1177/0956797620968787

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software*, *76*(1), 1–32. https://doi.org/10.18637/jss.v076.i01

Castro, D., Ferreira, F., de Castro, I., Rodrigues, A. R., Correia, M., Ribeiro, J., & Ferreira, T. B. (2019). The Differential Role of Central and Bridge Symptoms in Deactivating Psychopathological Networks. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.02448

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2021). *Shiny: Web application framework for R*. manual. https://CRAN.R-project.org/package=shiny

Chen, P. Y., Smithson, M., & Popovich, P. M. (2002). *Correlation: Parametric and nonparametric measures*. Sage.

Chib, S., & Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika*, *85*(2), 347–361. https://doi.org/10.1093/biomet/85.2.347

Choudhuri, N. (1998). Bayesian bootstrap credible sets for multidimensional mean functional. *Annals of statistics*, 2104–2127. https://doi.org/10.1214/aos/1024691463

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013, June 17). *Applied Multiple Regression/-Correlation Analysis for the Behavioral Sciences*. Routledge.

Cousineau, D. (2020). How many decimals? Rounding descriptive and inferential statistics based on measurement precision. *Journal of Mathematical Psychology*, *97*, 102362. https://doi.org/10.1016/j.jmp.2020.102362

Cramer, A. O. J., van der Sluis, S., Noordhof, A., Wichers, M., Geschwind, N., Aggen, S. H., Kendler, K. S., & Borsboom, D. (2012). Dimensions of Normal Personality as Networks in Search of Equilibrium: You Can't Like Parties if You Don't Like People. *European Journal of Personality*, *26*(4), 414–431. https://doi.org/10.1002/per.1866

Cramer, A. O. J., Waldorp, L. J., van der Maas, H. L. J., & Borsboom, D. (2010a). Comorbidity: A network perspective. *Behavioral and Brain Sciences*, *33*(2-3), 137–150. https://doi.org/10.1017/S0140525X09991567

Cramer, A. O. J., Waldorp, L. J., van der Maas, H. L. J., & Borsboom, D. (2010b). Complex realities require complex theories: Refining and extending the network approach to mental disorders. *Behavioral and Brain Sciences*, *33*(2-3), 178–193. https://doi.org/10.1017/S0140525X10000920

Dablander, F., & Hinne, M. (2019). Node centrality measures are a poor substitute for causal inference. *Scientific Reports*, *9*(1), 1–13. https://doi.org/10.1038/s41598-019-43033-9

Dahl, F. A., Grotle, M., Šaltytė Benth, J., & Natvig, B. (2008). Data splitting as a countermeasure against hypothesis fishing: With a case study of predictors for low back pain. *European Journal of Epidemiology*, *23*(4), 237–242. https://doi.org/10.1007/s10654-008-9230-x

Dalege, J., Borsboom, D., van Harreveld, F., & van der Maas, H. L. J. (2019). A Network Perspective on Attitude Strength: Testing the Connectivity Hypothesis. *Social Psychological and Personality Science*, *10*(6), 746–756. https://doi.org/10.1177/1948550618781062

Devroye, L. (1986). *Non-Uniform Random Variate Generation*. Springer Science & Business Media.

Diciccio, T. J., Kass, R. E., Raftery, A., & Wasserman, L. (1997). Computing Bayes Factors by Combining Simulation and Asymptotic Approximations. *Journal of the American Statistical Association*, *92*(439), 903–915. https://doi.org/10.1080/01621459.1997.10474045

Diedenhofen, B., & Musch, J. (2015). Cocor: A Comprehensive Solution for the Statistical Comparison of Correlations. *PLoS ONE*, *10*(4). https://doi.org/10.1371/journal.pone.0121945

Dunn, O. J., & Clark, V. (1969). Correlation Coefficients Measured on the Same Individuals. *Journal of the American Statistical Association*, *64*(325), 366–377. https://doi.org/10.1080/01621459.1969.10500981

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *Annals of Statistics*, *7*(1), 1–26. https://doi.org/10.1214/aos/1176344552

Epskamp, S., Borsboom, D., & Fried, E. I. (2018). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods*, *50*(1), 195–212. https://doi.org/10.3758/s13428-017-0862-1

Epskamp, S., & Fried, E. I. (2018). A tutorial on regularized partial correlation networks. *Psychological Methods*, *23*(4), 617–634. https://doi.org/10.1037/met0000167

Epskamp, S., Maris, G. K. J., Waldorp, L. J., & Borsboom, D. (2018, June 7). *Network Psychometrics*. Retrieved February 6, 2020, from http://arxiv.org/abs/1609.02818

Epskamp, S., Rhemtulla, M., & Borsboom, D. (2017). Generalized Network Psychometrics: Combining Network and Latent Variable Models. *Psychometrika*, *82*(4), 904–927. https://doi.org/10.1007/s11336-017-9557-x

Epskamp, S., van Borkulo, C. D., van der Veen, D. C., Servaas, M. N., Isvoranu, A.-M., Riese, H., & Cramer, A. O. J. (2018). Personalized Network Modeling in Psychopathology: The Importance of Contemporaneous and Temporal Connections. *Clinical Psychological Science*, *6*(3), 416–427. https://doi.org/10.1177/2167702617744325

Epskamp, S., Waldorp, L. J., Mõttus, R., & Borsboom, D. (2018). The Gaussian Graphical Model in Cross-Sectional and Time-Series Data. *Multivariate Behavioral Research*, *53*(4), 453–480. https://doi.org/10.1080/00273171.2018.1454823

Estrada, E., Ferrer, E., Shaywitz, B. A., Holahan, J. M., & Shaywitz, S. E. (2018). Identifying atypical change at the individual level from childhood to adolescence. *Developmental Psychology*, *54*(11), 2193–2206. https://doi.org/10.1037/dev0000583

Etz, A., Haaf, J. M., Rouder, J. N., & Vandekerckhove, J. (2018). Bayesian Inference and Testing Any Hypothesis You Can Specify. *Advances in Methods and Practices in Psychological Science*, *1*(2), 281–295. https://doi.org/10.1177/2515245918773087

Faraway, J. J. (1995). *Data Splitting Strategies for Reducing the Effect of Model Selection on Inference*.

Fisher, R. A. (1924). The Distribution of the Partial Correlation Coefficient. *Metron*, *3*. Retrieved April 23, 2020, from https://digital.library.adelaide.edu.au/dspace/handle/2440/15182

Forbes, M. K., Wright, A. G. C., Markon, K. E., & Krueger, R. F. (2017). Evidence that psychopathology symptom networks have limited replicability. *Journal of Abnormal Psychology*, *126*(7), 969–988. https://doi.org/10.1037/abn0000276

Forbes, M. K., Wright, A. G. C., Markon, K. E., & Krueger, R. F. (2019). Quantifying the Reliability and Replicability of Psychopathology Network Characteristics. *Multivariate Behavioral Research*, *0*(0), 1–19. https://doi.org/10.1080/00273171.2019.1616526

Fried, E. I., Eidhof, M. B., Palic, S., Costantini, G., Huisman-van Dijk, H. M., Bockting, C. L. H., Engelhard, I., Armour, C., Nielsen, A. B. S., & Karstoft, K.-I. (2018). Replicability and Generalizability of Posttraumatic Stress Disorder (PTSD) Networks: A Cross-Cultural Multisite Study of PTSD Symptoms in Four Trauma Patient Samples. *Clinical Psychological Science*, *6*(3), 335–351. https://doi.org/10.1177/2167702617745092

Fried, E. I., Epskamp, S., Nesse, R. M., Tuerlinckx, F., & Borsboom, D. (2016). What are 'good' depression symptoms? Comparing the centrality of DSM and non-DSM symptoms of depression in a network analysis. *Journal of Affective Disorders*, *189*, 314–320. https://doi.org/10.1016/j.jad.2015.09.005

Friedman, N. P., Profile, S., Friedman, N. P., & Miyake, A. (2004). The relations among inhibition and interference control functions: A latent-variable analysis. *Journal of Experimental Psychology; General*, 101–135.

Frühwirth-Schnatter, S., & Wagner, H. (2011, October 6). Bayesian Variable Selection for Random Intercept Modeling of Gaussian and Non-Gaussian Data. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, & M. West (Eds.), *Bayesian Statistics 9* (pp. 165–200). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199694587.003.0006

Gasparini, M. (1995). Exact multivariate Bayesian bootstrap distributions of moments. *The Annals of Statistics*, *23*(3), 762–768. https://doi.org/10.1214/aos/1176324620

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, *1*(3), 515–534. https://doi.org/10.1214/06-BA117A

Gelman, A. (2013). Two simple examples for understanding posterior p-values whose distributions are far from uniform. *Electronic Journal of Statistics*, *7*, 2595–2602. https://doi.org/10.1214/13-EJS854

Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models* (1st ed.).

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior Predictive Assessment of Model Fitness Via Realized Discrepancies. *Statistica Sinica*, *6*(4), 733–760. Retrieved May 14, 2020, from https://www.jstor.org/stable/24306036

George, E. I., & McCulloch, R. E. (1993). Variable Selection via Gibbs Sampling. *Journal of the American Statistical Association*, *88*(423), 881–889. https://doi.org/10.1080/01621459.1993.10476353

George, E. I., & McCulloch, R. E. (1997). Approaches For Bayesian Variable Selection. *Statistica Sinica*, *7*(2), 339–373.

Ghosh, M. (2011). Objective Priors: An Introduction for Frequentists. *Statistical Science*, *26*(2), 187–202. https://doi.org/10.1214/10-STS338

Grice, J. W., Medellin, E., Jones, I., Horvath, S., McDaniel, H., O'lansen, C., & Baker, M. (2020). Persons as Effect Sizes. *Advances in Methods and Practices in Psychological Science*, *3*(4), 443–455. https://doi.org/10.1177/2515245920922982

Haaf, J. M., & Rouder, J. N. (2017). Developing Constraint in Bayesian Mixed Models. *Psychological Methods*, *22*(4), 779–798. https://doi.org/10.1037/met0000156

Haldane, J. B. S. (1932). A note on inverse probability. *Mathematical Proceedings of the Cambridge Philosophical Society*, *28*(1), 55–61. https://doi.org/10.1017/S0305004100010495

Haslbeck, J. M. B., Ryan, O., Robinaugh, D., Waldorp, L., & Borsboom, D. (2019, December 10). *Modeling Psychopathology: From Data Models to Formal Theories* (preprint). PsyArXiv. https://doi.org/10.31234/osf.io/jgm7f

Heck, D. W., Overstall, A. M., Gronau, Q. F., & Wagenmakers, E.-J. (2019). Quantifying uncertainty in transdimensional Markov chain Monte Carlo using discrete Markov models. *Statistics and Computing*, *29*(4), 631–643. https://doi.org/10.1007/s11222-018-9828-0

Hedeker, D., Mermelstein, R. J., & Demirtas, H. (2012). Modeling between-subject and within-subject variances in ecological momentary assessment data using mixed-effects location scale models. *Statistics in Medicine*, *31*(27), 3328–3336. https://doi.org/10.1002/sim.5338

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Hjort, N. L. (1991). Bayesian and empirical Bayesian bootstrapping. *Preprint series. Statistical Research Report http://urn. nb. no/URN: NBN: no-23420.*

Hoff, P. D. (2009). *A first course in Bayesian statistical methods*. Springer
OCLC: ocn310401109.

Hoffman, M. D., & Gelman, A. (2011, November 17). *The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo*. Retrieved September 19, 2019, from http://arxiv.org/abs/1111.4246

Hoijtink, H. (2001). Confirmatory Latent Class Analysis: Model Selection Using Bayes Factors and (Pseudo) Likelihood Ratio Statistics. *Multivariate Behavioral Research*, *36*(4), 563–588. https://doi.org/10.1207/S15327906MBR3604_04

Hoijtink, H. (2011). *Informative hypotheses: Theory and practice for behavioral and social scientists*. CRC.

Hoijtink, H., Mulder, J., van Lissa, C., & Gu, X. (2019). A tutorial on testing hypotheses using the Bayes factor. *Psychological Methods*, *24*(5), 539–556. https://doi.org/http://dx.doi.org/10.1037/met0000201

Højsgaard, S., Edwards, D., & Lauritzen, S. (2012, February 18). *Graphical Models with R*. Springer Science & Business Media.

Hoorelbeke, K., Marchetti, I., De Schryver, M., & Koster, E. H. W. (2016). The interplay between cognitive risk and resilience factors in remitted depression: A network analysis. *Journal of Affective Disorders*, *195*, 96–104. https://doi.org/10.1016/j.jad.2016.02.001

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology*, *57*(5), 253–270. https://doi.org/10.1037/h0023816

Jones, P. J., Ma, R., & McNally, R. J. (2019). Bridge Centrality: A Network Approach to Understanding Comorbidity. *Multivariate Behavioral Research*, *0*(0), 1–15. https://doi.org/10.1080/00273171.2019.1614898

Jöreskog, K. G. (1994). On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika*, *59*(3), 381–389. https://doi.org/10.1007/BF02296131

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*(1), 54–69. https://doi.org/10.1037/a0028347

Kalisch, M., & Bühlmann, P. (2007). Estimating High-Dimensional Directed Acyclic Graphs with the PC-Algorithm. *Journal of Machine Learning Research*, *8*, 613–636.

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Kelder, T., Conklin, B. R., Evelo, C. T., & Pico, A. R. (2010). Finding the Right Questions: Exploratory Pathway Analysis to Enhance Biological Discovery in Large Datasets. *PLoS Biology*, *8*(8). https://doi.org/10.1371/journal.pbio.1000472

Kendall, M. G. (1938). A New Measure of Rank Correlation. *Biometrika*, *30*(1/2), 81–93. https://doi.org/10.2307/2332226

Kim, Y., & Lee, J. (2003). Bayesian bootstrap for proportional hazards models. *Annals of Statistics*, *31*(6), 1905–1922. https://doi.org/10.1214/aos/1074290331

Klugkist, I., Kato, B., & Hoijtink, H. (2005). Bayesian model selection using encompassing priors. *Statistica Neerlandica*, *59*(1), 57–69. https://doi.org/10.1111/j.1467-9574.2005.00279.x

Krishnamoorthy, K., & Xia, Y. (2007). Inferences on correlation coefficients: One-sample, independent and correlated cases. *Journal of Statistical Planning and Inference*, *137*(7), 2362–2379. https://doi.org/10.1016/j.jspi.2006.08.002

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9. *Journal of General Internal Medicine*, *16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Krumsiek, J., Suhre, K., Evans, A. M., Mitchell, M. W., Mohney, R. P., Milburn, M. V., Wägele, B., Römisch-Margl, W., Illig, T., Adamski, J., Gieger, C., Theis, F. J., & Kastenmüller, G. (2012). Mining the Unknown: A Systems Approach to Metabolite Identification Combining Genetic and Metabolic Information. *PLoS Genetics*, *8*(10). https://doi.org/10.1371/journal.pgen.1003005

Kruschke, J. K. (2018). Rejecting or Accepting Parameter Values in Bayesian Estimation. *Advances in Methods and Practices in Psychological Science*, *1*(2), 270–280. https://doi.org/10.1177/2515245918771304

Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The Time Has Come: Bayesian Methods for Data Analysis in the Organizational Sciences. *Organizational Research Methods*, *15*(4), 722–752. https://doi.org/10.1177/1094428112457829

Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, *25*(1), 178–206. https://doi.org/10.3758/s13423-016-1221-4

Kuo, L., & Mallick, B. (1998). Variable Selection for Regression Models. *Sankhyā: The Indian Journal of Statistics, Series B (1960-2002)*, *60*(1), 65–81. Retrieved March 17, 2021, from https://www.jstor.org/stable/25053023

Lauritzen, S. L. (1996, May 2). *Graphical Models*. Clarendon Press.

Lawrence, E., Bingham, D., Liu, C., & Nair, V. N. (2008). Bayesian Inference for Multivariate Ordinal Data Using Parameter Expansion. *Technometrics*, *50*(2), 182–191. https://doi.org/10.1198/004017008000000064

Lazar, N. A. (2003). Bayesian empirical likelihood. *Biometrika*, *90*(2), 319–326. https://doi.org/10.1093/biomet/90.2.319

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press. https://doi.org/10.1017/CBO9781139087759

Lenk, P. (2009). Simulation Pseudo-Bias Correction to the Harmonic Mean Estimator of Integrated Likelihoods. *Journal of Computational and Graphical Statistics*, *18*(4), 941–960. https://doi.org/10.1198/jcgs.2009.08022

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*(9), 1989–2001. https://doi.org/10.1016/j.jmva.2009.04.008

Ley, E., & Steel, M. F. J. (2009). On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. *Journal of Applied Econometrics*, *24*(4), 651–674. https://doi.org/10.1002/jae.1057

Li, Q., & Lin, N. (2010). The Bayesian elastic net. *Bayesian Analysis*, *5*(1), 151–170. https://doi.org/10.1214/10-BA506

Liu, C. C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, *52*(6), 362–375. https://doi.org/10.1016/j.jmp.2008.03.002

Liu, S., Rovine, M. J., & Molenaar, P. C. M. (2012). Selecting a linear mixed model for longitudinal data: Repeated measures analysis of variance, covariance pattern model, and growth curve approaches. *Psychological Methods*, *17*(1), 15–30. https://doi.org/10.1037/a0026971

Lo, A. Y. (1987). A Large Sample Study of the Bayesian Bootstrap. *Annals of Statistics*, *15*(1), 360–375. https://doi.org/10.1214/aos/1176350271

Lo, A. Y. (1988). A Bayesian Bootstrap for a Finite Population. *The Annals of Statistics*, *16*(4), 1684–1695. https://doi.org/10.1214/aos/1176351061

Lövdén, M., Schmiedek, F., Kennedy, K. M., Rodrigue, K. M., Lindenberger, U., & Raz, N. (2013). Does variability in cognitive performance correlate with frontal brain volume? *NeuroImage*, *64*, 209–215. https://doi.org/10.1016/j.neuroimage.2012.09.039

Lyddon, S. P., Holmes, C. C., & Walker, S. G. (2019). General Bayesian updating and the loss-likelihood bootstrap. *Biometrika*, *106*(2), 465–478. https://doi.org/10.1093/biomet/asz006

MacDonald, S. W., Hultsch, D. F., & Dixon, R. A. (2008). Predicting Impending Death: Inconsistency in Speed is a Selective and Early Marker. *Psychology and aging*, *23*(3), 595–607. https://doi.org/10.1037/0882-7974.23.3.595

Mayo, D. G. (1991). Novel Evidence and Severe Tests. *Philosophy of Science*, *58*(4), 523–552. Retrieved April 10, 2020, from https://www.jstor.org/stable/188479

McNally, R. J., Mair, P., Mugno, B. L., & Riemann, B. C. (2017). Co-morbid obsessive–compulsive disorder and depression: A Bayesian network approach. *Psychological Medicine*, *47*(7), 1204–1214. https://doi.org/10.1017/S0033291716003287

McNally, R. J. (2016). Can network analysis transform psychopathology? *Behaviour Research and Therapy*, *86*, 95–104. https://doi.org/10.1016/j.brat.2016.06.006

McNally, R. J., Robinaugh, D. J., Wu, G. W. Y., Wang, L., Deserno, M. K., & Borsboom, D. (2015). Mental Disorders as Causal Systems: A Network Approach to Posttraumatic Stress Disorder. *Clinical Psychological Science*, *3*(6), 836–849. https://doi.org/10.1177/2167702614553230

Meehl, P. E. (1967). Theory-Testing in Psychology and Physics: A Methodological Paradox. *Philosophy of Science*, *34*(2), 103–115. Retrieved April 10, 2020, from https://www.jstor.org/stable/186099

Meng, X.-L. (1994). Posterior Predictive p-Values. *Annals of Statistics*, *22*(3), 1142–1160. https://doi.org/10.1214/aos/1176325622

Meng, X.-l., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin*, *111*(1), 172–175. https://doi.org/http://dx.doi.org/10.1037/0033-2909.111.1.172

Meteyard, L., & Davies, R. A. I. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language*, *112*, 104092. https://doi.org/10.1016/j.jml.2020.104092

Mitchell, T. J., & Beauchamp, J. J. (1988). Bayesian Variable Selection in Linear Regression. *Journal of the American Statistical Association*, *83*(404), 1023–1032. https://doi.org/10.1080/01621459.1988.10478694

Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, *72*, 6–18. https://doi.org/10.1016/j.jmp.2015.11.001

Mulder, J., & Raftery, A. E. (2019). BIC Extensions for Order-constrained Model Selection. *Sociological Methods & Research*, 0049124119882459. https://doi.org/10.1177/0049124119882459

Mulder, J. (2014). Bayes factors for testing inequality constrained hypotheses: Issues with prior specification. *British Journal of Mathematical and Statistical Psychology*, *67*(1), 153–171. https://doi.org/10.1111/bmsp.12013

Mulder, J. (2016). Bayes factors for testing order-constrained hypotheses on correlations. *Journal of Mathematical Psychology*, *72*, 104–115. https://doi.org/10.1016/j.jmp.2014.09.004

Mulder, J., Hoijtink, H., & Klugkist, I. (2010). Equality and inequality constrained multivariate linear models: Objective model selection using constrained posterior priors. *Journal of Statistical Planning and Inference*, *140*(4), 887–906. https://doi.org/10.1016/j.jspi.2009.09.022

Mulder, J., & Pericchi, L. R. (2018). The Matrix-F Prior for Estimating and Testing Covariance Matrices. *Bayesian Analysis*, *13*(4), 1193–1214. https://doi.org/10.1214/17-BA1092

Newman, M. E. J. (2010). *Networks: An introduction*. Oxford University Press
OCLC: ocn456837194.

Newton, M. A., & Raftery, A. E. (1994). Approximate Bayesian Inference with the Weighted Likelihood Bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological)*, *56*(1), 3–48. https://doi.org/10.1111/j.2517-6161.1994.tb01956.x

Ntzoufras, I. (2002). Gibbs Variable Selection using BUGS. *Journal of Statistical Software*, *7*(1), 1–19. https://doi.org/10.18637/jss.v007.i07

O'Hagan, A. (1995). Fractional Bayes Factors for Model Comparison. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 99–118. https://doi.org/10.1111/j.2517-6161.1995.tb02017.x

O'Hara, R. B., & Sillanpää, M. J. (2009). A review of Bayesian variable selection methods: What, how and which. *Bayesian Analysis*, *4*(1), 85–117. https://doi.org/10.1214/09-BA403

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, *44*(4), 443–460. https://doi.org/10.1007/BF02296207

Owen, A. (1990). Empirical likelihood ratio confidence regions. *The Annals of Statistics*, *18*(1), 90–120. https://doi.org/10.1214/aos/1176347494

Papaspiliopoulos, O., Roberts, G. O., & Sköld, M. (2007). A General Framework for the Parametrization of Hierarchical Models. *Statistical Science*, *22*(1), 59–73. Retrieved April 12, 2021, from https://www.jstor.org/stable/27645805

Pearl, J. (2009, September 14). *Causality*. Cambridge University Press.

Piccirillo, M. L., & Rodebaugh, T. L. (2019). Foundations of idiographic methods in psychology and applications for psychotherapy. *Clinical Psychology Review*, *71*, 90–100. https://doi.org/10.1016/j.cpr.2019.01.002

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Working Papers*, 8.

Poirier, D. J. (2011). Bayesian Interpretations of Heteroskedastic Consistent Covariance Estimators Using the Informed Bayesian Bootstrap. *Econometric Reviews*, *30*(4), 457–468. https://doi.org/10.1080/07474938.2011.553542

Pozzi, F., Di Matteo, T., & Aste, T. (2012). Exponential smoothing weighted correlations. *The European Physical Journal B*, *85*(6), 175. https://doi.org/10.1140/epjb/e2012-20697-x

R Core Team. (2021). *R: A language and environment for statistical computing*. manual. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Raghunathan, T. E., Rosenthal, R., & Rubin, D. B. (1996). Comparing correlated but nonoverlapping correlations. *Psychological Methods*, *1*(2), 178–183. https://doi.org/http://dx.doi.org/10.1037/1082-989X.1.2.178

Rast, P., & Ferrer, E. (2018). A Mixed-Effects Location Scale Model for Dyadic Interactions. *Multivariate Behavioral Research*, *53*(5), 756–775. https://doi.org/10.1080/00273171.2018.1477577

Rast, P., & Zimprich, D. (2011). Modeling within-person variance in reaction time data of older adults. *GeroPsych*, *24*(11), 169–176. https://doi.org/10.1024/1662-9647/a000045

Raudenbush, S. W., & Bryk, A. S. (2001, December 19). *Hierarchical Linear Models: Applications and Data Analysis Methods* (2nd edition). SAGE Publications, Inc.

Robinaugh, D. J., Millner, A. J., & McNally, R. J. (2016). Identifying Highly Influential Nodes in the Complicated Grief Network. *Journal of abnormal psychology*, *125*(6), 747–757. https://doi.org/10.1037/abn0000181

Rouder, J., Morey, R., & Wagenmakers, E.-J. (2016). The Interplay between Subjectivity, Statistical Practice, and Psychological Science. *Collabra: Psychology*, *2*(1), 6. https://doi.org/10.1525/collabra.28

Rouder, J. N., & Haaf, J. M. (2020). Are There Reliable Qualitative Individual Difference in Cognition? https://doi.org/10.31234/osf.io/3ezmw

Rouder, J. N., Haaf, J. M., & Vandekerckhove, J. (2018). Bayesian inference for psychology, part IV: Parameter estimation and Bayes factors. *Psychonomic Bulletin & Review*, *25*(1), 102–113. https://doi.org/10.3758/s13423-017-1420-7

Rouder, J. N., Kumar, A., & Haaf, J. M. (2019). Why Most Studies of Individual Differences With Inhibition Tasks Are Bound To Fail. https://doi.org/10.31234/osf.io/3cjr5

Rubin, D. B. (1981). The Bayesian Bootstrap. *Annals of Statistics*, *9*(1), 130–134. https://doi.org/10.1214/aos/1176345338

Rubin, D. B. (1984). Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician. *The Annals of Statistics*, *12*(4), 1151–1172. Retrieved May 14, 2020, from https://www.jstor.org/stable/2240995

Ryan, O., Bringmann, L. F., & Schuurman, N. K. (2019, October 1). *The Challenge of Generating Causal Hypotheses Using Network Models* (preprint). PsyArXiv. https://doi.org/10.31234/osf.io/ryg69

Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic programming in python using PyMC3. *PeerJ Computer Science*, *2*, e55. https://doi.org/10.7717/peerj-cs.55

Savage, L. J. (1954). *The foundations of statistics*. John Wiley & Sons (Oxford, England). Retrieved February 11, 2021, from https://search.proquest.com/psycinfo/docview/615273673/64A78B70F9C54005PQ/14

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological bulletin*, *86*(2), 420.

Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A Brief Measure for Assessing Generalized Anxiety Disorder: The GAD-7. *Archives of Internal Medicine*, *166*(10), 1092–1097. https://doi.org/10.1001/archinte.166.10.1092

Šrol, J., Cavojova, V., & Mikušková, E. B. (2021). Social consequences of COVID-19 conspiracy beliefs: Evidence from two studies in Slovakia. https://doi.org/10.31234/osf.io/y4svc

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*(2), 245–251. https://doi.org/10.1037/0033-2909.87.2.245

Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, *36*(2), 111–147. Retrieved April 10, 2020, from https://www.jstor.org/stable/2984809

Talhouk, A., Doucet, A., & Murphy, K. (2012). Efficient Bayesian Inference for Multivariate Probit Models With Sparse Inverse Correlation Matrices. *Journal of Computational and Graphical Statistics*, *21*(3), 739–757. https://doi.org/10.1080/10618600.2012.679239

Tokuda, T., Goodrich, B., Van Mechelen, I., Gelman, A., & Tuerlinckx, F. (2011). Visualizing distributions of covariance matrices. *Columbia Univ., New York, USA, Tech. Rep*, 18–18.

Van Rossum, G., & Drake, F. L. (2009). *Python 3 reference manual*. CreateSpace.

van Borkulo, C. D., Boschloo, L., Kossakowski, J. J., Tio, P., Schoevers, R. A., Borsboom, D., & Waldorp, L. J. (2016). *Comparing network structures on three aspects: A permutation test* [Submitted for publication]. https://doi.org/10.13140/RG.2.2.29455.38569

Vandekerckhove, J., Rouder, J. N., & Kruschke, J. K. (2018). Editorial: Bayesian methods for advancing psychological science. *Psychonomic Bulletin & Review*, *25*(1), 1–4. https://doi.org/10.3758/s13423-018-1443-8

van Doorn, J., Ly, A., Marsman, M., & Wagenmakers, E.-J. (2018). Bayesian Inference for Kendall's Rank Correlation Coefficient. *The American Statistician*, *72*(4), 303–308. https://doi.org/10.1080/00031305.2016.1264998

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J. N., & Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*(1), 35–57. https://doi.org/10.3758/s13423-017-1343-3

Wagenmakers, E.-J., Morey, R. D., & Lee, M. D. (2016). Bayesian Benefits for the Pragmatic Researcher. *Current Directions in Psychological Science*, *25*(3), 169–176. https://doi.org/10.1177/0963721416643289

Wagenmakers, E.-J., Verhagen, J., & Ly, A. (2016). How to quantify the evidence for the absence of a correlation. *Behavior Research Methods*, *48*(2), 413–426. https://doi.org/10.3758/s13428-015-0593-0

Watts, A., Walters, R. W., Hoffman, L., & Templin, J. (2016). Intra-Individual Variability of Physical Activity in Older Adults With and Without Mild Alzheimer's Disease. *PLOS ONE*, *11*(4), e0153898. https://doi.org/10.1371/journal.pone.0153898

Webb, E. L., & Forster, J. J. (2008). Bayesian model determination for multivariate ordinal and binary data. *Computational Statistics & Data Analysis*, *52*(5), 2632–2649. https://doi.org/10.1016/j.csda.2007.09.008

Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *The Journal of Strength & Conditioning Research*, *19*(1), 231–240.

Weng, C.-S. (1989). On a Second-Order Asymptotic Property of the Bayesian Bootstrap Mean. *The Annals of Statistics*, *17*(2), 705–710. https://doi.org/10.1214/aos/1176347136

Wetzels, R., & Wagenmakers, E.-J. (2012). A default Bayesian hypothesis test for correlations and partial correlations. *Psychonomic Bulletin & Review*, *19*(6), 1057–1064. https://doi.org/10.3758/s13423-012-0295-x

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Williams, D. R., Carlsson, R., & Bürkner, P.-C. (2017). Between-litter variation in developmental studies of hormones and behavior: Inflated false positives and diminished power. *Frontiers in Neuroendocrinology*, *47*, 154–166. https://doi.org/10.1016/j.yfrne.2017.08.003

Williams, D. R., Martin, S. R., & Rast, P. (2021). Putting the individual into reliability: Bayesian testing of homogeneous within-person variance in hierarchical models. *Behavior Research Methods*. https://doi.org/10.3758/s13428-021-01646-x

Williams, D. R., & Mulder, J. (2020). Bayesian hypothesis testing for Gaussian graphical models: Conditional independence and order constraints. *Journal of Mathematical Psychology*, *99*, 102441. https://doi.org/10.1016/j.jmp.2020.102441

Williams, D. R., Mulder, J., Rouder, J. N., & Rast, P. (2020). Beneath the surface: Unearthing within-person variability and mean relations with Bayesian mixed models. *Psychological Methods*. https://doi.org/10.1037/met0000270

Williams, D. R., Rast, P., Pericchi, L. R., & Mulder, J. (2020). Comparing Gaussian graphical models with the posterior predictive distribution and Bayesian model selection. *Psychological Methods*. https://doi.org/10.1037/met0000254

Williams, D. R., Zimprich, D. R., & Rast, P. (2019). A Bayesian nonlinear mixed-effects location scale model for learning. *Behavior Research Methods*. https://doi.org/10.3758/s13428-019-01255-9

Williams, D. R. (2020, January 27). *Learning to Live with Sampling Variability: Expected Replicability in Partial Correlation Networks* (preprint). PsyArXiv. https://doi.org/10.31234/osf.io/fb4sa

Williams, D. R., & Mulder, J. (2019, January 14). *Bayesian Hypothesis Testing for Gaussian Graphical Models: Conditional Independence and Order Constraints* (preprint). PsyArXiv. https://doi.org/10.31234/osf.io/ypxd8

Wolsiefer, K., Westfall, J., & Judd, C. M. (2017). Modeling stimulus variation in three common implicit attitude tasks. *Behavior Research Methods*, *49*(4), 1193–1209. https://doi.org/10.3758/s13428-016-0779-0

Yuan, Y., & Johnson, V. E. (2008). Bayesian Hypothesis Tests Using Nonparametric Statistics. *Statistica Sinica*, *18*(3), 1185–1200. Retrieved December 1, 2020, from https://www.jstor.org/stable/24308537

Yule, G. U. (1897). On the Theory of Correlation. *Journal of the Royal Statistical Society*, *60*(4), 812–854. https://doi.org/10.2307/2979746

Zou, G. Y. (2007). Toward using confidence intervals to compare correlations. *Psychological Methods*, *12*(4), 399–413. https://doi.org/10.1037/1082-989X.12.4.399