Estimating Correlations in Low-Reliability Settings With Constrained Hierarchical Models

Mahbod Mehrvarz and Jeffrey N. Rouder

University of California, Irvine

Author Note

Version 1, February, 2024.

MM https://orcid.org/0009-0002-5426-5867

JNR https://orcid.org/0000-0003-2023-3891

Author Contributions: MM and JNR jointly conceived the project, developed the models, executed the simulations, and wrote the results.

Open Science Practices: The simulated data, analysis scripts, and code used to generate the figures are openly available at https://github.com/specl/ctx-pca-localize

This research was supported in part by NSF 2126976 and by ONR N00014-23-1-2792.

Correspondence concerning this article should be addressed to Mahbod Mehrvarz, Department of Cognitive Sciences, University of California, Irvine, CA, 92697. E-mail: mehrvarm@uci.edu

Abstract

It is popular to study individual differences in cognition with experimental tasks, and the main goal of such approaches is to analyze the pattern of correlations across a battery of tasks and measures. One difficulty is that experimental tasks are often low in reliability as effects are small relative to trial-by-trial variability. Consequently, it remains difficult to accurately estimate correlations. One approach that seems attractive is hierarchical modeling where trial-by-trial variability and variability across conditions, tasks, and individuals are modeled separately. Here we show that hierarchical models may reduce the error in estimating correlations up to 46%, but only if substantive constraint is imposed. The approach here is Bayesian, and we develop novel Bayesian hierarchical factor models for experiments where trials are nested in conditions, tasks, and individuals. The prior on covariances across tasks can either be unconstrained, in which there is little error reduction, or constrained, in which there is substantial error reduction. The constraints are: 1. There is a low-dimension factor structure underlying the covariation across tasks; and 2. All loadings are nonnegative leading to a positive manifold on correlations. We argue that both of these assumptions are reasonable in cognitive domains, and that with them, researchers may profitably use hierarchical models to estimate correlations across tasks in low-reliability settings.

Keywords: Individual Differences, Cognitive Control, Methodology, Factor models, Bayesian Hierarchical Models

Estimating Correlations in Low-Reliability Settings With Constrained Hierarchical Models

In the modern landscape, researchers study cognitive abilities such as perception, language, problem-solving, and memory through the lens of information processing. The main focus is on understanding how we represent, manage, combine, store, recall, and assess information from our surroundings. These abilities are often measured in experiments comprised of conditions that are contrasted to eliminate nuisance factors. For example, to measure **cognitive control**–the ability to inhibit proponent responses–tasks such as Stroop employ congruent and incongruent conditions. The difference in response times between the two conditions serves as a measure of cognitive control without the influence of factors that generically affect response times. In this manner, the experimental approach with contrasted conditions provides for theoretically valid and interpretable measures of cognitive abilities.

Over the last century, researchers have used individual differences across tasks to understand underlying mental processes and variations in them across populations. Continuing with the cognitive control example, we note that the structure and correlates of cognitive control have been studied repeatedly with individual differences using experimental tasks. However, finding consistent patterns of individual differences has proven difficult.

This difficulty perhaps reflects a statistical concern that goes by the moniker of the **reliability crisis**. Indeed, measures of cognitive abilities from experimental designs often lack sufficient reliability for latent variable modeling. Three characteristics observed in recent inquiries have exemplified this concern: First, several tasks that purportedly measure the same construct do not correlate well. For example, the correlation between flanker and Stroop effects in large studies is often near .1 and rarely greater than .25. (Enkavi et al., 2019; Rey-Mermet et al., 2019; Rouder et al., 2023). Second, factor-model analyses have not proven replicable or robust. The evidence comes from Karr et al. (2018), who showed that when bootstrapped simulated data were submitted to confirmatory-factor-model analysis, the best-fitting model from the original set was rarely recovered. Finally, it has been observed that individual differences in experiments have low reliability (Draheim et al., 2019; Hedge et al., 2018). How people differ in an experiment does not predict with high precision how they will differ again.

How bad is this reliability problem? The following brief example may serve as motivation. Consider an experiment with six tasks. Each task is comprised of two conditions—a congruent and an incongruent condition—with the latter requiring more cognitive control than the former. In this setup, trials are nested in conditions; for instance, an individual might perform 100 trials in each condition for each task. The focal point of interest here is the individual's score, calculated as the difference in average performance between the incongruent and congruent conditions, and the critical question is how these scores covary across the tasks. With six tasks at hand, we encounter 15 distinct inter-task correlations. For simplicity, let's posit that each task correlates with every other task at a true value of $\rho = .5$. This value is depicted in the first column of Figure 1A, labeled *Truth*. We simulated data from this setup for 200 hypothetical individuals using realistic values for true performance (details to be discussed later). The next column, labeled *Usual*, is the usual Pearson sample correlation estimate from individuals' scores. Note that these sample correlations are dramatically underestimated. Indeed, this attenuation of correlation is a well-known phenomenon that occurs when scores are particularly prone to measurement error (Spearman, 1904). A challenging aspect of this attenuation is that it is **asymptotic**. As the number of individuals in the experiment is increased, attenuation does not decrease. Rather, the analyst becomes increasingly confident in badly attenuated values.

One long-standing strategy to mitigate attenuation is to calculate how much attenuation may be present and adjust accordingly; this process is called **disattenuation**. The third column in Figure 1A shows Spearman's correction[1] for attenuation (Spearman,

[1] Spearman's correction for attenuation is $r^*_{xy} = \frac{r_{xy}}{\sqrt{r_{xx} \cdot r_{yy}}}$ where r_{xy} is the Pearson sample correlation

1904) for the 15 intertask correlations. As advertised, these corrected values are disattenuated and center around the true value. Yet, these values are surprisingly variable. The experiment simulated here is quite large comprising 240,000 observations from 200 individuals performing 100 trials across two conditions in six tasks. One might expect that estimates of 15 correlation coefficients from 240,000 observations would be highly accurate. However, the reality is strikingly different: the disattenuated correlations range widely from .2 to .8, deviating on average by .18. The extent of uncertainty in such a large data set is disconcerting. This inability to localize correlations in reasonable cognitive-control settings serves as the problem for this report.

Why is localizing these correlations important? There are two reasons. First, the correlation (or covariation) among tasks serves as a critical target for capturing the relations among tasks. Second, this covariation is often decomposed using latent variable models such as factor analysis and structural equation modeling. The effectiveness of these latent variable decompositions hinges critically on the precision with which these correlations may be localized. In cases where correlations are imprecisely localized or dramatically attenuated, the resulting latent structure may be only weakly identifiable, casting doubt on the validity of inferences drawn from it (Karr et al., 2018).

Several recent approaches have been proposed to address the issue of poor localization. Draheim et al. (2019) recommends measures that do not depend on contrasting conditions. Deveau et al. (2015), Kucina et al. (2023), and Wells et al. (2021) recommend gamification where experiments are integrated into an engaging computer video game format complete with stimulating music and a competitive points system. Haines et al. (2023), Matzke et al. (2017), and Rouder and Haaf (2019) recommend hierarchical model analysis where multiple sources are modeled separately. Hierarchical modeling approaches are explored here.

Yet, from our experience, the anticipated benefits of hierarchical models in

between two variables, and r_{xx} and r_{yy} is the reliability of each.

localizing correlations have not fully materialized. Rouder et al. (2023) used a simulation approach to assess whether hierarchical models could indeed localize correlations in reasonable designs with realistic effects. The fourth column of Figure 1A, labeled *Unconstrained* shows correlations from a hierarchical model, to be presented subsequently. Rouder et al. identified three key advantages of the hierarchical approach. First, hierarchical-model analysis typically leads to less attenuation compared to the usual approach. Second, it achieves somewhat more precise localization than Spearman's disattenuation method in repeated simulations (Rouder et al., 2023). Third, hierarchical models offer a realistic assessment of uncertainty in correlations. The last benefit is illustrated in Figure 1B. The histogram is the posterior distribution of one of the correlations. The dashed lines show the 95% credible interval. This interval is large, spanning from .27 to .76 in value. Its size shows the high degree of uncertainty, which reflects a high degree of variability in performance from trial to trial. In contrast, the usual estimate is shown as a point. The segments show the 95% confidence interval. The relatively narrow confidence interval reflects the large number of individuals without consideration of trial-to-trial variability. The usual correlation is greatly attenuated (true value is .5) indicating that this high level of confidence is misplaced.

Despite these three benefits, the problem of precise localization of correlations remains. In experiments of reasonable size, Root Mean Square Error (**RMSE**) remains high, often around .15, suggesting that even with a large data set of 240,000 observations, discerning whether a correlation was .3 or .7 remains elusive. Localizing correlations even with hierarchical models is challenging even in large data sets.

Constrained and Unconstrained Hierarchical Models

One strategy for increasing the precision of estimates is to use simpler and more constrained models. An example of this strategy comes from polynomial regression: the slope of a regression term will be estimated to higher precision when higher-order terms are excluded from the model. We propose using constrained hierarchical models that are designed to increase the precision of estimating covariation. The approach, of course, is not cost free—the constraints we impose have to be realistic and warranted. The constraints we explore are a positive manifold across covariation and a low-dimensional factor structure across tasks.

This pay-to-play strategy is best explained with the formal specification of models. Consider a battery of cognitive control tasks where I individuals each perform J tasks. Each task is comprised of a congruent and an incongruent condition, and each individual performs L_{ijk} trials in each condition. Let $i = 1, \ldots, I$ index individuals, $j = 1, \ldots, J$ index tasks, $k = 1, 2$ index conditions, and $\ell = 1, \ldots, L_{ijk}$ index replicate trials. Let $Y_{ijk\ell}$ denote an observation, for example, the response time on a given trial. We place the following linear model on Y:

$$Y_{ijk\ell} = \alpha_{ij} + x_k\theta_{ij} + \epsilon_{ijk\ell}. \tag{1}$$

In this model, x_k is the contrast code for the two conditions, set at -$\frac{1}{2}$ and $\frac{1}{2}$ for congruent and incongruent conditions, respectively. The parameter α_{ij} is the true overall speed of the ith individual on the jth task. The parameter θ_{ij} is the true difference between incongruent and congruent conditions for the jth task—it is the ith individual's true effect on the jth task and is the focal point of analysis. The error term is $\epsilon_{ijk\ell} \sim \text{Normal}(0, \tau_j^2)$, with τ_j^2 describing the variability of trial noise for the jth task. Errors are independent of one another.

At the next level, we place random effect models on individual true effects. Let $\boldsymbol{\theta} = (\theta_{i1}, \ldots, \theta_{iJ})'$ be a column vector true effects for the ith individual, which is modeled as a multivariate normal:

$$\theta_{\mathbf{i}} \sim \text{N}_{\mathbf{J}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{2}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_J)'$ is a vector of mean effects and $\boldsymbol{\Sigma}$ is a $J \times J$ covariance matrix.

The target for constraint in this context is the covariance matrix, $\boldsymbol{\Sigma}$ (Bollen, 1989). In the most general formulation, no constraint is placed on $\boldsymbol{\Sigma}$ other than it is valid covariance matrix. When no constraint is placed on $\mathbf{\Sigma}$, then any relations among the tasks may be accounted for. Constraint may be added by placing restrictions on the elements of $\mathbf{\Sigma}$. Consider, for example, the constraint from a one-factor model (see Figure 2A). Under this constraint, the covariance $\mathbf{\Sigma}$ is,

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 + \lambda_1^2 & \lambda_1\lambda_2 & \lambda_1\lambda_3 & \ldots & \lambda_1\lambda_J \\ \lambda_1\lambda_2 & \sigma_2^2 + \lambda_2^2 & \lambda_1\lambda_3 & \ldots & \lambda_1\lambda_J \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \lambda_1\lambda_J & \lambda_2\lambda_J & \lambda_3\lambda_J & \ldots & \sigma_J^2 + \lambda_J^2 \end{bmatrix}, \tag{3}$$

where λ_j and σ_j^2 are the factor loading and residual variance in true effects, respectively, for the jth task. This covariance matrix is highly constrained. For example, we can express $\text{Cov}(3,4) = \text{Cov}(1,3) \times \text{Cov}(2,4)/\text{Cov}(1,2)$ as $\lambda_3\lambda_4 = \lambda_1\lambda_3 \times \lambda_2\lambda_4/\lambda_1\lambda_2$. The one-factor constraint entails a reduction in the number of parameters. For example, if there are 6 tasks, then the unconstrained covariance matrix has 21 free parameters (6 diagonal elements and 15 off-diagonal elements) while the one-factor constrained version has 12 free parameters. The critical question then is whether adding constraint on $\mathbf{\Sigma}$ in hierarchical models can substantially improve the localization of correlations in realistic designs. To foreshadow, we show here that it can.

Hierarchical Models For Estimating Covariation

The hierarchical models in this report are the combination of Eq 1 and Eq 2 along with constraints on $\mathbf{\Sigma}$. These models are analyzed in the Bayesian framework because of computational tractability. In the Bayesian framework, priors are placed on all parameters. Our approach here is to use weakly informed subjective priors. For example, the prior on overall true speed, α_{ij} is a normal with a mean of .8 s and a standard deviation of 1 s. This specification is quite reasonable for cognitive control tasks like the Stroop or flanker where RTs are typically under a second. Likewise, priors of τ are broad with much mass between .01 and 1 s, which corresponds to broad prior on standard deviation of trial-to-trial noise with much mass between 100 ms and 1 s. Likewise, priors on μ in Eq 2 reflect the expected range of effects between 0 and 150 ms. The critical specifications are for $\mathbf{\Sigma}$, which follow:

The Inverse-Wishart, An Unconstrained Model

The unconstrained model we use comes from Haaf and Rouder (2017) and Rouder et al. (2023). Here, an inverse Wishart prior is placed on $\mathbf{\Sigma}$: $\mathbf{\Sigma}^{-1} \sim \text{Wishart}(m, \boldsymbol{S})$, where m and $\boldsymbol{S}$ serve as prior settings on the shape and scale of the inverse Wishart. The shape parameter is uninformative in this context when set to $m = J + 1$. The scale parameter is a bit trickier. A lack of *a prior* preference for negative or positive correlations comes when $\boldsymbol{S} = b^2 \times \boldsymbol{I}$ where $\boldsymbol{I}$ is the identity matrix (of size J) and $m \times b^2$ is the expected variance.[2] This expectation is made with reference to the analysis of existing data sets as discussed subsequently.

Orthogonal Factor Models

The orthogonal factor model serves as constraint on $\mathbf{\Sigma}$. A model of dimension D is given by

$$\mathbf{\Sigma} = \boldsymbol{\lambda}\boldsymbol{\lambda}' + \mathbf{\Delta}, \tag{4}$$

where $\boldsymbol{\lambda}$ is a $J \times D$ matrix of factor loadings and $\mathbf{\Delta}$ is a $J \times J$ diagonal matrix with a diagonal of $\sigma_1^2, \ldots, \sigma_J^2$. If $D = 1$, the model reduces to the one-factor model discussed above. The degree of constraint then is determined by D. When $D = J$, there is no constraint—any covariance may be represented with an equal number of factors to tasks. As D is reduced, more constraint is added, with the most constraint for $D = 1$. We develop one- and two-factor models here.

[2] There is an alternative choice of prior for covariance that we extensively explored, the LKJ prior (Lewandowski et al., 2009). This prior is less informative than the Wishart because, unlike the Wishart, the estimation of correlation is independent of the specification of scale. Consequently, this prior is recommended (McElreath, 2016), and implementation is convenient in the R-package rstan (Stan Development Team, 2020). Yet, we found better performance for the inverse Wishart in simulations in that the posterior credible intervals were smaller and better covered the true value. The increased performance of the Wishart reflects the fact that researchers have a rough idea about the scale of individual differences—it is on the order of tens of milliseconds—and this is enough information for the improved performance of the inverse Wishart.

The priors for Eq 4 are defined as follows. The factor loadings are distributed as $\lambda \sim \mathrm{N}(0, .03^2)$, representing a wide-ranging prior. Residual variances σ_j^2 are distributed as Inverse-Gamma $\left(\frac{1}{2}, \frac{1}{2}A^2\right)$, where A^2 is a prior setting that reflects the scale of variance. Inverse-gamma distributions with a shape of 1/2 are diffuse and only weakly informative. Additionally, the covariance element ($\mathbf{\Sigma}$) from the Wishart model was tuned in a similar manner.

Positive Manifold Constraints

The positive manifold refers to a well-known finding that cognitive-performance tasks tend to correlate positively. Carroll (1993), for example, showed that across a large survey of cognitive-performance tasks, people who performed well on one task performed well on the others. Ritchie (2015, p. 25) dramatically emphasizes the point: "Some researchers have even deliberately creative cognitive tests tapping different skills that they expected not to correlate together, but they always did."

To implement the positive-manifold constraint, we require that the factor loadings be nonnegative. This approach is analogous to nonnegative matrix factorization (Lee & Seung, 1999). In practice, we set a truncated prior $\lambda_{jd} \sim \mathrm{N}_+(0, .03^2)$

Analysis

The above hierarchical models were implemented in JAGS (Plummer, 2003). Many researchers who analyze factor models in the Bayesian framework use the conditional forms of factor models given by $\theta_{ij} \sim \mathrm{N}(\mu_j + \sum_d \lambda_{jd}\eta_{id}, \delta_j^2)$ (e.g. Ghosh & Dunson, 2009). The advantage of using this conditional form is that sampling in MCMC is done through repeated calls of independent normals. The disadvantage in this context is poor mixing; samples in chains are highly correlated from iteration to iteration. The marginal forms given in Eq. 4 are slower to sample in MCMC because samples from multivariate normals are needed. Yet, the mixing is markedly improved. In fact, we find that the effective sample size is often 90% or more of the number of iterations for all parameters. In the following simulations, we ran chains of 3000 iterations with the first 1000 iterations serving as burn-in.

Simulations

To evaluate the performance of constrained hierarchical models in localizing correlations, we conducted a series of 14 simulations encompassing a broad spectrum of ground truths. To ensure that our simulations closely mirrored real-world situations, we drew on empirical designs and observed effects from real conflict experiments (see Rouder et al., 2024). Each simulation was comprised of 200 hypothetical individuals performing six different tasks. For each task, individuals performed 100 trials in congruent and incongruent conditions.

Ground truth parameter values were chosen to reflect realistic data. The distribution of true individual overall speed was centered at 800 ms and had a standard deviation of 200 ms. Each task had a between-condition true effect of 60 ms and a between-individual true standard deviation of 25 ms. Trial-to-trial variability was 200 ms in standard deviation. The critical elements of the ground truth are the parameters that specify the relations among tasks. We discuss the choices which comprise the 14 simulations:

One-Factor, Equal-Loading Ground Truth

Previously, we explored the case where all true correlations were .5 in value (see Figure 1). The first set of simulations expands this case to other true correlation values. Panel A of Figure 2 depicts a factor model that generates correlation matrices with equal off-diagonal elements. Here, there is a single factor that loads equally on all tasks. Panels A_1 to A_3 in Figure 3 show three examples of correlation matrices from this one-factor, equal-load setup. The difference among the three variants is the degree of load, with the loadings increasing from A_1 to A_3.

From these true correlation matrices, we generated true individual effect values, θ_{ij} by sampling from a multivariate normal distribution. We next computed 15 correlation estimates among these true values and recorded the RMSE from the true population values. This RMSE is the expected error if true individual values are known, and it serves as an ideal lower bound on error when true individual effects must be estimated from the data. Panels A_1 to A_3 of Figure 4 display the RMSE across 108 runs in each of these variants, denoted as *True Ind.* The best-case RMSE varied from .04 to .07, depending on the factor loading.

For each run, we simulated observations for each individual, with each completing 100 trials in both conditions for each task, comprising a total of 240,000 observations. For every run, we computed the usual sample correlation and the Spearman-disattenuated correlation for observed effects. Figure 4 shows the distributions of RMSEs across the runs, denoted as *Usual* and *Spearman* respectively. Figure $4A_1$, for scenarios with low-value uniform true correlations, and the usual approach performed well whereas the Spearman-disattenuated method exhibited higher error. This result is expected as the downward attenuation in the usual approach works well when the true correlation is small in value. As true correlations are larger in value, as in simulations A_2 and A_3, the attenuation in the usual approach is in error, and the Spearman disattenuation correlation has a lower RMSE.

The simulated data were also submitted to a series of hierarchical models. This included the unconstrained Wishart model (referred to as *Wishart*) and one- and two-factor constrained models, with and without a positive constraint. These models are denoted as 1F, 2F, 1F+, and 2F+, respectively. Estimates from the Wishart model behave much like those from Spearman disattenuation, and this behavior is no coincidence as both are based on similar treatments of measurement noise. The constrained factor models exhibited consistently better performance for all three simulations. Here we see the gain from constraint—it will hold broadly.

One-Factor, Graded-Loading Ground Truth

Panel B of Figure 2 shows a one-factor setup where the loadings vary across tasks. Panels B_1 to B_3 of Figure 3 show the corresponding ground truth correlation matrices. We used three versions where the loadings increased in value. The data were simulated as before–there were 108 runs with each run being the simulation of 200 individuals who performed 100 trials in two conditions and across six tasks. The RMSEs for the true individual values (a lower bound), the usual approach, the Spearman disattenuation, and the five hierarchical models are shown in the right column of Figure 4. The same trends in the previous set of simulations are evident—(i) attenuation in the usual method is well suited for small true correlations; (ii) The Spearman-disattention and Wishart unconstrained model estimates are fairly error prone for small true correlation values and improve for larger true one; (iii) constrained hierarchical model estimates have the smallest error in all three variants.

Two-Factor, Perfect Cluster Ground Truth

We simulated a case where the tasks cluster on factors. The first factor loads on the first three tasks; the second factor loads on the second three tasks. We call this a two-factor perfect cluster configuration, and it is shown in Panel C of Figure 2. The corresponding true correlation matrices are shown in Panels C_1 to C_3 of Figure 3. Here the loadings from the factors are increased in value leading to within-cluster correlations of .1, .3, and .6, respectively. The left column of Figure 5 shows RMSEs for the true values, the usual approach, the Spearman disattenuation, and five hierarchical models in its left column.

The usual approach does better than the two previous configurations due to the prevalence of correlation of zero in value, which is accurately captured through attenuation. And as the correlation coefficients within clusters rise, so does the RMSE for the usual approach. In these simulations, the one-factor models are misspecified. They are too simple to capture the cluster configuration. This misspecification is especially acute when the within-cluster correlation is high in value. And, not too surprisingly, the one-factor models are particularly error prone in simulation C_3, where within-cluster correlation is high. When the correlation values are lower in value, the miss from the one-factor model is not as pronounced, and the performance is comparable to the two-factor models. It is not too surprising that the two-factor models do well in all the two-cluster simulations as there is a strong match between the generating model and the fitted model.

Two-Factor, Graded Loading Ground Truth

Figure 2D shows a two-factor model where factors load on tasks in a graded manner. The corresponding ground truth correlation matrices for this setup are shown in Panels D_1 to D_3 of Figure 3. The right column of Figure 5 shows RMSEs for the eight methods. Once again, the constrained hierarchical models outperform the other methods. Contrary to the perfect cluster configuration, the one-factor model does well as the ground truth can be approximated by a one-factor structure. The two-factor model consistently shows stable and robust performance across all variants.

Bi-Factor Ground Truth

In our final set of simulations, we examine two bi-factor configurations. The first, shown in Figure 2E, features a single general factor that equally loads on every task, resembling the one-factor equal-loading ground truth. Additionally, it includes two specific factors that uniquely impact three tasks each, without cross-loading, akin to the perfect-cluster ground truth. The second configuration, presented in Panel F of the same figure, also incorporates the same general factor loading on all tasks and includes an additional three specific factors, each uniquely loading on two tasks without cross-loading. Figure 6 displays the RMSEs for both configurations, as calculated using the ind-true, usual approach, the Spearman disattenuation, and five hierarchical models. Although the constrained models are misspecified, they maintain their competitive advantage in estimating correlation coefficients.

Trends Across All Simulations

Figure 7 shows a summary of the results. Plotted are RMSE values for the usual method, the Spearman disattenuation, the unconstranted (Wishart) hierarchical model, and a two-factor, positive manifold hierarchical model. The superiority of the constrained model is evident across all 14 simulation setups. The average RMSEs across these setups are .192, .172, .177, and .109 for the usual, Spearman, unconstrained, and constrained estimates, respectively. The gain for the constrained model, while varying across setup, is about 35-46 % on average compared to the other three. Hence, using constrained hierarchical models has the potential to markedly improve the localization of correlations in experimental psychology designs.

Assessing Uncertainty

One advantage of Bayesian modeling is that it provides an account of posterior uncertainty in parameters. Figure 1B provides an example—the wide range of the 95% credible interval of the select correlation coefficient shows a large degree of uncertainty. How accurate are these measures of uncertainty in the constrained Bayesian hierarchical models? We use simulation methods to assess the coverage of credible intervals here. Coverage in this context refers to how often the 95% credible interval covers the true value in simulation. Concerns over coverage have a long history in Bayesian analysis and serve as motivation for probability-matching priors (Datta & Mukerjee, 2004; Stein, 1985).

Coverage was assessed for the two-factor model for two ground truths. One ground truth was the two-factor graded (see Figure 2D), and for this truth, the two-factor model is well specified. The other ground truth was the bi-factor one (see Figure 1F). Here, the two-factor model is too simple, and the question is how this misspecification affects coverage. For each truth, we simulated data sets and estimated posteriors for 1000 runs. For each run, the 95% credible intervals were computed for all correlation coefficients, and the coverage proportion was the number of runs in which the true value was within the 95% credible interval. The top panel of Figure 8 shows these coverage proportions along with true values of the correlation coefficients in the simulation. Coverage is quite reasonable—nearly perfect for the well-specified truth and only somewhat biased downward for true correlations that are higher in value in the misspecified case.

The bottom panels of Figure 8 show 95% credible intervals across all 1000 runs for both simulations (well-specified on the left; misspecified on the right). These intervals, while wide, are well-behaved. The coverage misses in the misspecified case are due to some shrinkage toward moderate values rather than due to intervals that are too narrow. All in all, researchers using these models to assess uncertainty on correlations can have a high degree of confidence in the resulting intervals. The ability of Bayesian hierarchical models to provide accurate assessments of uncertainty stands as a substantial advantage of this approach over more popular competitor approaches.

Discussion

Although studying individual differences in psychological experiments offers many advantages, methodological challenges remain. The most important one is the attenuation of correlations reflecting unaccounted trial noise in task scores. These attenuations are difficult in that they are large, obscure the latent structure, and remain even in the limit of many individuals. The obvious way to deal with trial noise is to account for it within a hierarchical model, and several authors recommend just this treatment (Haines et al., 2023; Matzke et al., 2017). Previously, we assessed the ability of unconstrained hierarchical models to accurately localize correlations in designs with typical sample sizes and typical effect sizes (Rouder et al., 2023), and the results were that correlations could be localized with a RMSE of around .2. Such a large RMSE even in designs that had hundreds of thousands of observations spread across hundreds of people was discouraging.

In the current paper, we explore whether correlations may be better localized by introducing the substantive constraints of a low-dimensional structure and positive loadings. The results are quite encouraging. With this low-dimensional structure, RMSEs improved by 37%, and with the positive loadings, the gain was another 9%. There was even gain when the true structure was more complicated than the imposed factor structure. Importantly, hierarchical models provide for a sense of uncertainty, and we showed through simulation that resulting credible intervals have reasonable frequentist coverage properties. Overall, then, we think the cost of the low-dimensional and positive-loading assumptions is small compared to the gain in localization.

There are a few limitations to the approach as follows:

First, the approach is assumptive and works well when the substantive constraints hold. The obvious concern is when the constraints do not hold. The good news here is that the positive loading constraint may be jettisoned without much loss if it is suspected not to hold. The low-factor dimensionality seems more critical. If researchers suspect a high-dimensional structure, then assuming a low-dimensional structure may distort the results to some degree. Results were good with the two-factor structure even when a three-factor structure served as ground truth, though it may be that this result is more fortuitous than systematic. Of course, researchers who suspect high-dimension structure would need to collect data sets with larger sample sizes, trial sizes, and with larger effects and individual differences. Given the resolution of most data sets collected in cognitive abilities, a low-dimensional constraint strikes us as judicious.

Second, the main target of inquiry is the correlation matrix, and the goal is to provide the best estimates possible. The target is not the factor structure underlying these correlations. In this regard, the factor structure serves as a filter to gain regularization in estimation. Using factor models this way is common in many signal-processing, genetics, and machine-learning domains (Bhattacharya & Dunson, 2011; Du et al., 2023; Sanyal & Ferreira, 2012). The limitation, however, is that the current development is inappropriate for inference on the latent factor structure. There are two separate issues. One has to do with the quality of posterior estimates of factor loadings. The factor model has a nonidentifiability with regard to rotations—the covariance across task loadings is invariant to different rotations of the factor loadings. In conventional analysis, the rotation is fixed without any loss. In Bayesian analysis through MCMC, in contrast, each iteration potentially corresponds to a different rotation. Hence, the posteriors on factor loadings are too variable—they reflect some true variability as well as variability from different rotations on different iterations (Poworoznek et al., 2021). This issue does not affect estimates of the correlations as these are stable with any iteration-to-iteration rotation differences. In fact, though not discussed here, we find good mixing in chains of correlations even with this added variability in factor loadings. The second issue is that there is no development herein of model comparison. Model comparison on dimensionality in the Bayesian framework is topical (e.g., Oh & Kim, 2010) but the subject is outside the scope of this paper.

Third, the model of data developed here is a linear model. Linear models are convenient from a statistical perspective, but they are not plausible from a psychological-process perspective. Instead, one might wish to model response time and response choice jointly as arising from an evidence accumulation model. There are hierarchical evidence accumulation models (Vandekerckhove et al., 2011; Wiecki et al., 2013). In these models, individual parameters come from a parent distribution that is task specific. To our knowledge, evidence-accumulation data models have not yet been integrated with factor-analytic models of individual differences. Although this integration is outside the scope of the development herein, we suspect it may happen in the near future.

References

Bhattacharya, A., & Dunson, D. B. (2011). Sparse Bayesian infinite factor models. *Biometrika*, *98*(2), 291–306. https://doi.org/10.1093/biomet/asr013

Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley & Sons. https://doi.org/10.1002/9781118619179

Carroll, J. B. (1993, January 29). *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*. Cambridge University Press.

Datta, G. S., & Mukerjee, R. (2004). Probability Matching Priors: Higher Order Asymptotics. *178*. https://doi.org/10.1007/978-1-4612-2036-7

Deveau, J., Jaeggi, S. M., Zordan, V., Phung, C., & Seitz, A. R. (2015). How to build better memory training games. *Frontiers in Systems Neuroscience*, *8*. Retrieved December 10, 2023, from https://www.frontiersin.org/articles/10.3389/fnsys.2014.00243

Draheim, C., Mashburn, C. A., Martin, J. D., & Engle, R. W. (2019). Reaction time in differential and developmental research: A review and commentary on the problems and alternatives. *Psychological Bulletin*, *145*(5), 508–535. https://doi.org/10.1037/bul0000192

Du, K.-L., Swamy, M. N. S., Wang, Z.-Q., & Mow, W. H. (2023). Matrix Factorization Techniques in Machine Learning, Signal Processing, and Statistics. *Mathematics*, *11*(12), 2674. https://doi.org/10.3390/math11122674

Enkavi, A. Z., Eisenberg, I. W., Bissett, P. G., Mazza, G. L., MacKinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2019). Large-scale analysis of test–retest reliabilities of self-regulation measures. *Proceedings of the National Academy of Sciences*, *116*(12), 5472–5477. https://doi.org/10.1073/pnas.1818430116

Ghosh, J., & Dunson, D. B. (2009). Default Prior Distributions and Efficient Posterior Computation in Bayesian Factor Analysis. *Journal of Computational and Graphical Statistics: A Joint Publication of American Statistical Association, Institute of*

Mathematical Statistics, Interface Foundation of North America, *18*(2), 306–320. https://doi.org/10.1198/jcgs.2009.07145

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

Haines, N., Kvam, P. D., Irving, L. H., Smith, C., Beauchaine, T. P., Pitt, M. A., Ahn, W.-Y., & Turner, B. M. (2023). Theoretically Informed Generative Models Can Advance the Psychological and Brain Sciences: Lessons from the Reliability Paradox. https://doi.org/10.31234/osf.io/xr7y3

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

Karr, J. E., Areshenkoff, C. N., Rast, P., Hofer, S. M., Iverson, G. L., & Garcia-Barrera, M. A. (2018). The unity and diversity of executive functions: A systematic review and re-analysis of latent variable studies. *Psychological Bulletin*, *144*(11), 1147–1185. https://doi.org/10.1037/bul0000160

Kucina, T., Wells, L., Lewis, I., Salas, K. de, Kohl, A., Palmer, M., Sauer, J. D., Matzke, D., Aidman, E., & Heathcote, A. (2023). Calibration of Cognitive Tests to Address the Reliability Paradox for Decision-Conflict Tasks. https://doi.org/10.31234/osf.io/bc6nk

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788–791. https://doi.org/10.1038/44565

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

Matzke, D., Ly, A., Selker, R., Weeda, W. D., Scheibehenne, B., Lee, M. D., & Wagenmakers, E.-J. (2017). Bayesian Inference for Correlations in the Presence of

Measurement Error and Estimation Uncertainty (S. Vazire & S. Bouwmeester, Eds.). *Collabra: Psychology*, *3*(1), 25. https://doi.org/10.1525/collabra.78

McElreath, R. (2016, December 30). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Chapman and Hall/CRC. https://doi.org/10.1201/9781315372495

Oh, H. S., & Kim, D.-G. (2010). Bayesian principal component analysis with mixture priors. *Journal of the Korean Statistical Society*, *39*(3), 387–396. https://doi.org/10.1016/j.jkss.2010.04.001

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

Poworoznek, E., Ferrari, F., & Dunson, D. (2021, July 29). *Efficiently resolving rotational ambiguity in Bayesian matrix sampling with matching*. arXiv: 2107.13783 [stat]. https://doi.org/10.48550/arXiv.2107.13783

Rey-Mermet, A., Gade, M., Souza, A. S., Von Bastian, C. C., & Oberauer, K. (2019). Is executive control related to working memory capacity and fluid intelligence? *Journal of Experimental Psychology: General*, *148*(8), 1335–1372. https://doi.org/10.1037/xge0000593

Ritchie, S. (2015, June 18). *Intelligence: All That Matters*. John Murray Press.

Rouder, J. N., & Haaf, J. M. (2019). A psychometrics of individual differences in experimental tasks. *Psychonomic Bulletin & Review*, *26*(2), 452–467. https://doi.org/10.3758/s13423-018-1558-y

Rouder, J. N., Kumar, A., & Haaf, J. M. (2023). Why many studies of individual differences with inhibition tasks may not localize correlations. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-023-02293-3

Rouder, J. N., Peña, A. F. C. D. la, Mehrvarz, M., & Vandekerckhove, J. (2024). On Cronbach's merger: Why experiments may not be suitable for measuring individual differences. https://doi.org/10.31234/osf.io/8ktn6

Sanyal, N., & Ferreira, M. A. R. (2012). Bayesian hierarchical multi-subject multiscale analysis of functional MRI data. *NeuroImage*, *63*(3), 1519–1531. https://doi.org/10.1016/j.neuroimage.2012.08.041

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*(1), 72–101. https://doi.org/10.2307/1412159

Stan Development Team. (2020). RStan: The R interface to Stan. https://doi.org/http://mc-stan.org/

Stein, C. M. (1985). On the coverage probability of confidence sets based on a prior distribution. *Banach Center Publications*, *16*(1), 485–514. Retrieved January 22, 2024, from https://eudml.org/doc/267838

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2011). Hierarchical diffusion models for two-choice response times. *Psychological Methods*, *16*(1), 44–62. https://doi.org/10.1037/a0021765

Wells, L., Kucina, T., Kohl, A., Lewis, I., de Salas, K., Aidman, E., & Heathcote, A. (2021). A flexible gaming environment for reliably measuring cognitive control. Retrieved December 10, 2023, from https://figshare.utas.edu.au/articles/conference_contribution/A_flexible_gaming_environment_for_reliably_measuring_cognitive_control/23112845/1

Wiecki, T., Sofer, I., & Frank, M. (2013). HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. *Frontiers in Neuroinformatics*, *7*. Retrieved February 4, 2024, from https://www.frontiersin.org/articles/10.3389/fninf.2013.00014

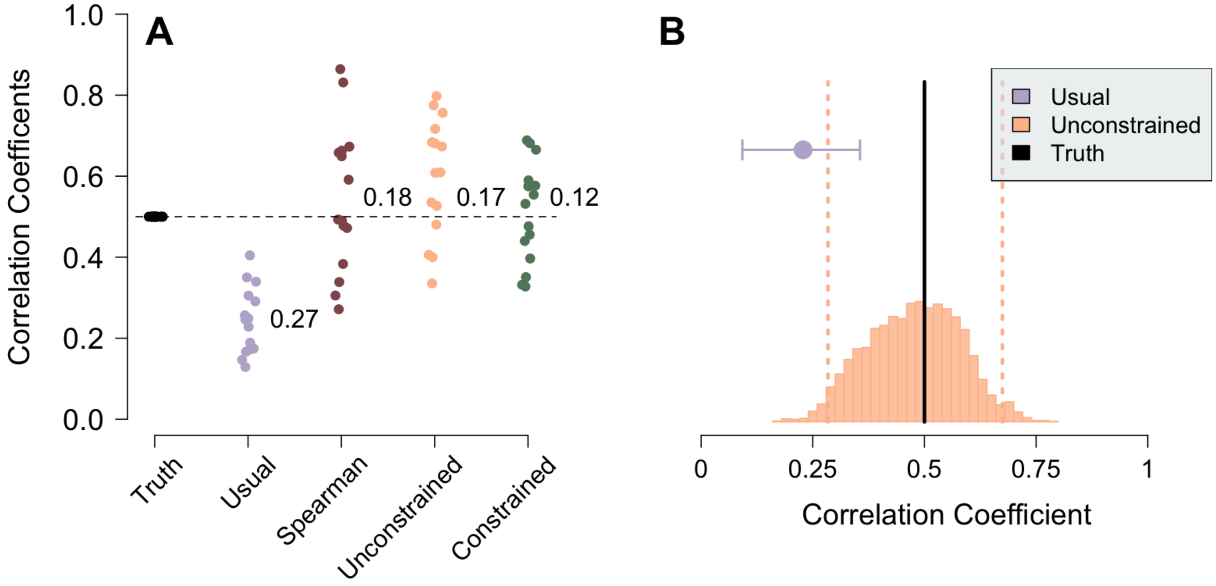


Figure 1

A. A single simulation run where all true correlations are .5. Shown are the usual Pearson sample correlations, the Spearman disattenuated correlations, and estimates from an unconstrained and constrained hierarchical model. The values next to the spread of points show the RMSEs in estimation. B. A comparison of uncertainty in correlation estimates. The point shows the usual estimate along with 95% confidence intervals. The distribution shows the posterior of the unconstrained model correlation estimate. The usual estimate not only is dramatically attenuated, but has narrow CIs that do not cover the true values. The posterior is centered on the truth, but is wide showing difficulty in localizing correlations without constraint.

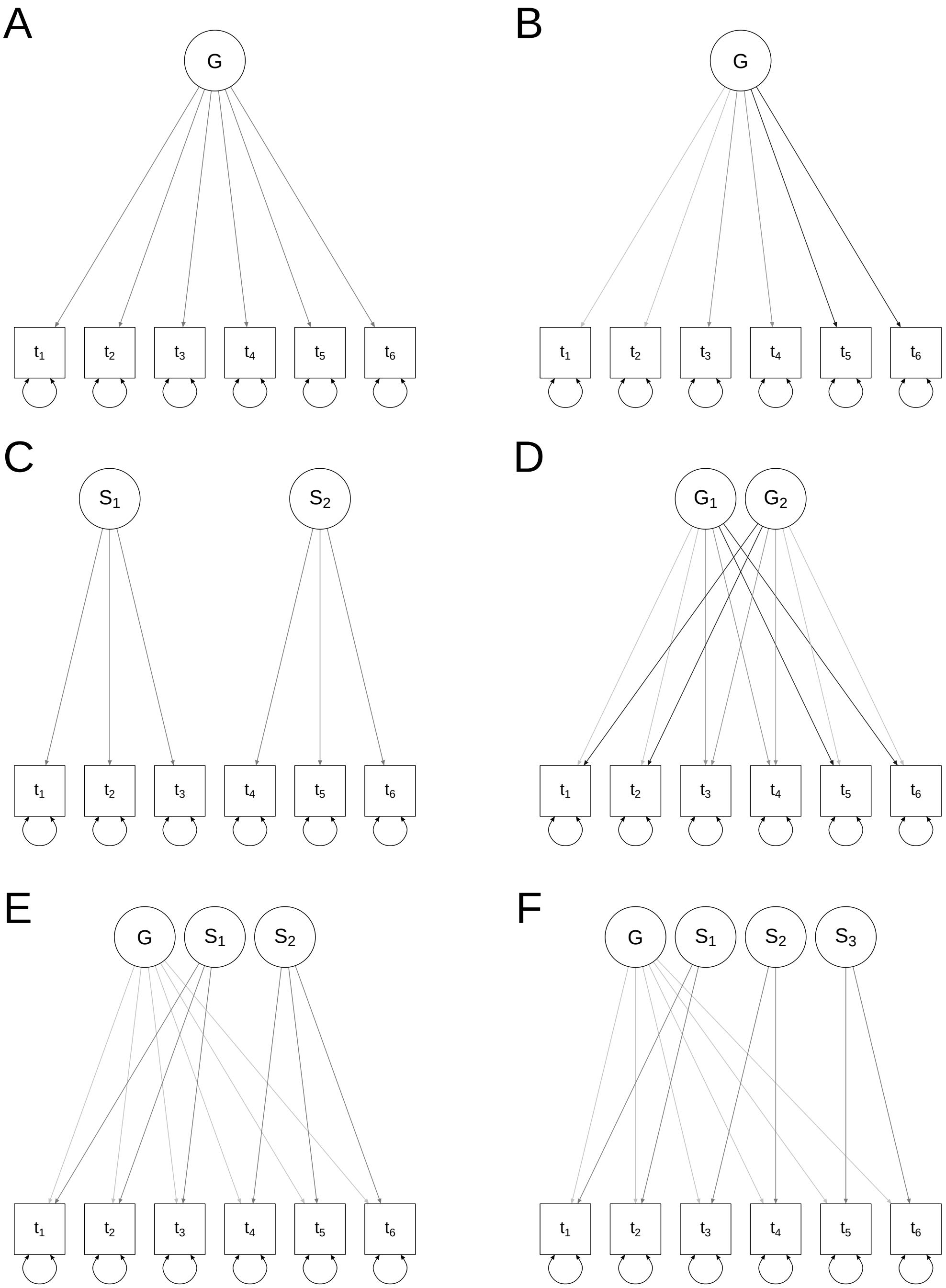
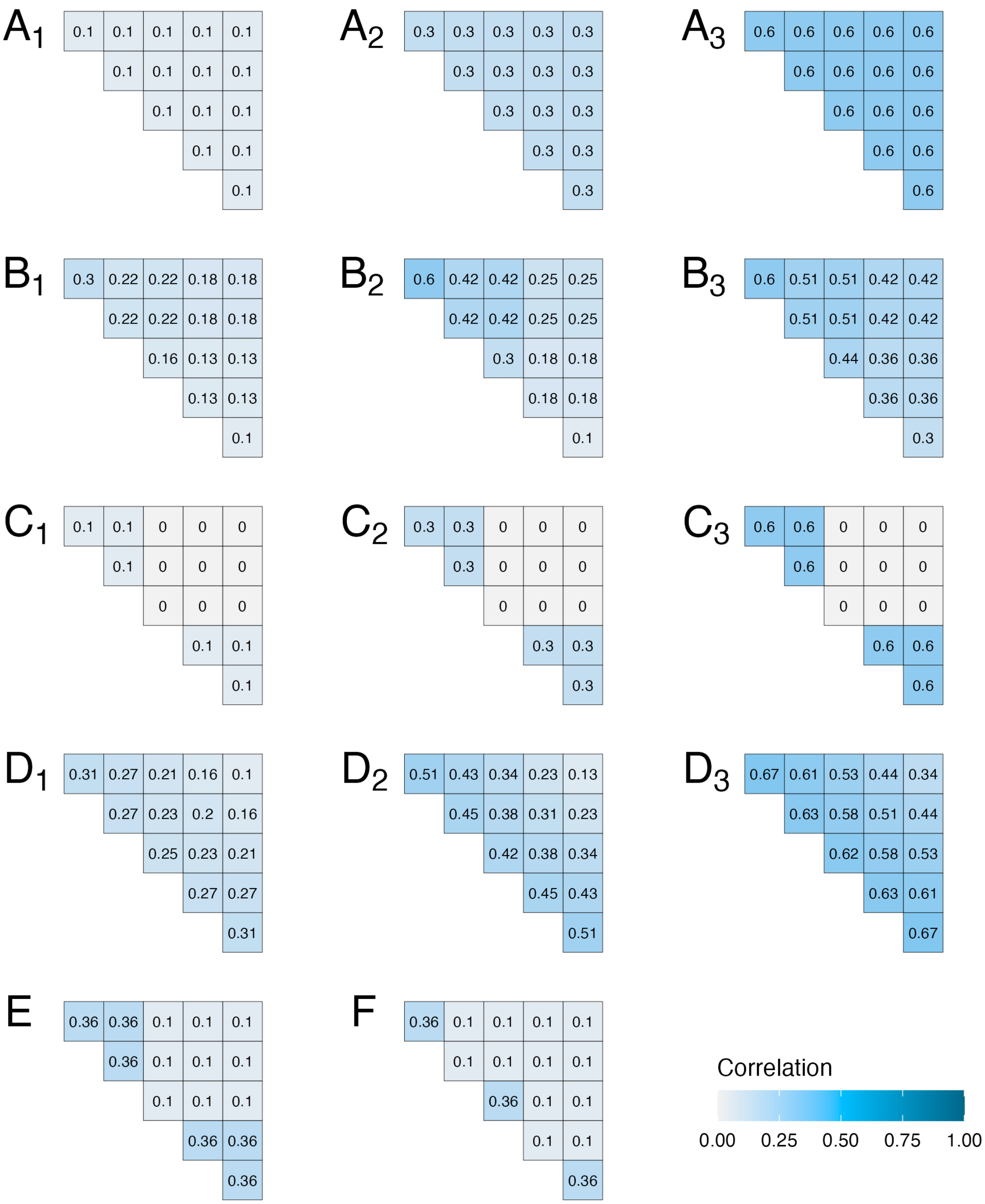


Figure 2

Latent variable representations of the ground truths used in simulations. See Figure 3 for the corresponding true correlation matrices.



A_1

0.1	0.1	0.1	0.1	0.1
	0.1	0.1	0.1	0.1
		0.1	0.1	0.1
			0.1	0.1
				0.1

A_2

0.3	0.3	0.3	0.3	0.3
	0.3	0.3	0.3	0.3
		0.3	0.3	0.3
			0.3	0.3
				0.3

A_3

0.6	0.6	0.6	0.6	0.6
	0.6	0.6	0.6	0.6
		0.6	0.6	0.6
			0.6	0.6
				0.6

B_1

0.3	0.22	0.22	0.18	0.18
	0.22	0.22	0.18	0.18
		0.16	0.13	0.13
			0.13	0.13
				0.1

B_2

0.6	0.42	0.42	0.25	0.25
	0.42	0.42	0.25	0.25
		0.3	0.18	0.18
			0.18	0.18
				0.1

B_3

0.6	0.51	0.51	0.42	0.42
	0.51	0.51	0.42	0.42
		0.44	0.36	0.36
			0.36	0.36
				0.3

C_1

0.1	0.1	0	0	0
	0.1	0	0	0
		0	0	0
			0.1	0.1
				0.1

C_2

0.3	0.3	0	0	0
	0.3	0	0	0
		0	0	0
			0.3	0.3
				0.3

C_3

0.6	0.6	0	0	0
	0.6	0	0	0
		0	0	0
			0.6	0.6
				0.6

D_1

0.31	0.27	0.21	0.16	0.1
	0.27	0.23	0.2	0.16
		0.25	0.23	0.21
			0.27	0.27
				0.31

D_2

0.51	0.43	0.34	0.23	0.13
	0.45	0.38	0.31	0.23
		0.42	0.38	0.34
			0.45	0.43
				0.51

D_3

0.67	0.61	0.53	0.44	0.34
	0.63	0.58	0.51	0.44
		0.62	0.58	0.53
			0.63	0.61
				0.67

E

0.36	0.36	0.1	0.1	0.1
	0.36	0.1	0.1	0.1
		0.1	0.1	0.1
			0.36	0.36
				0.36

F

0.36	0.1	0.1	0.1	0.1
	0.1	0.1	0.1	0.1
		0.36	0.1	0.1
			0.1	0.1
				0.36

Figure 3

True correlation matrices across the 14 simulations. See Figure 2 for the corresponding latent variable representation.

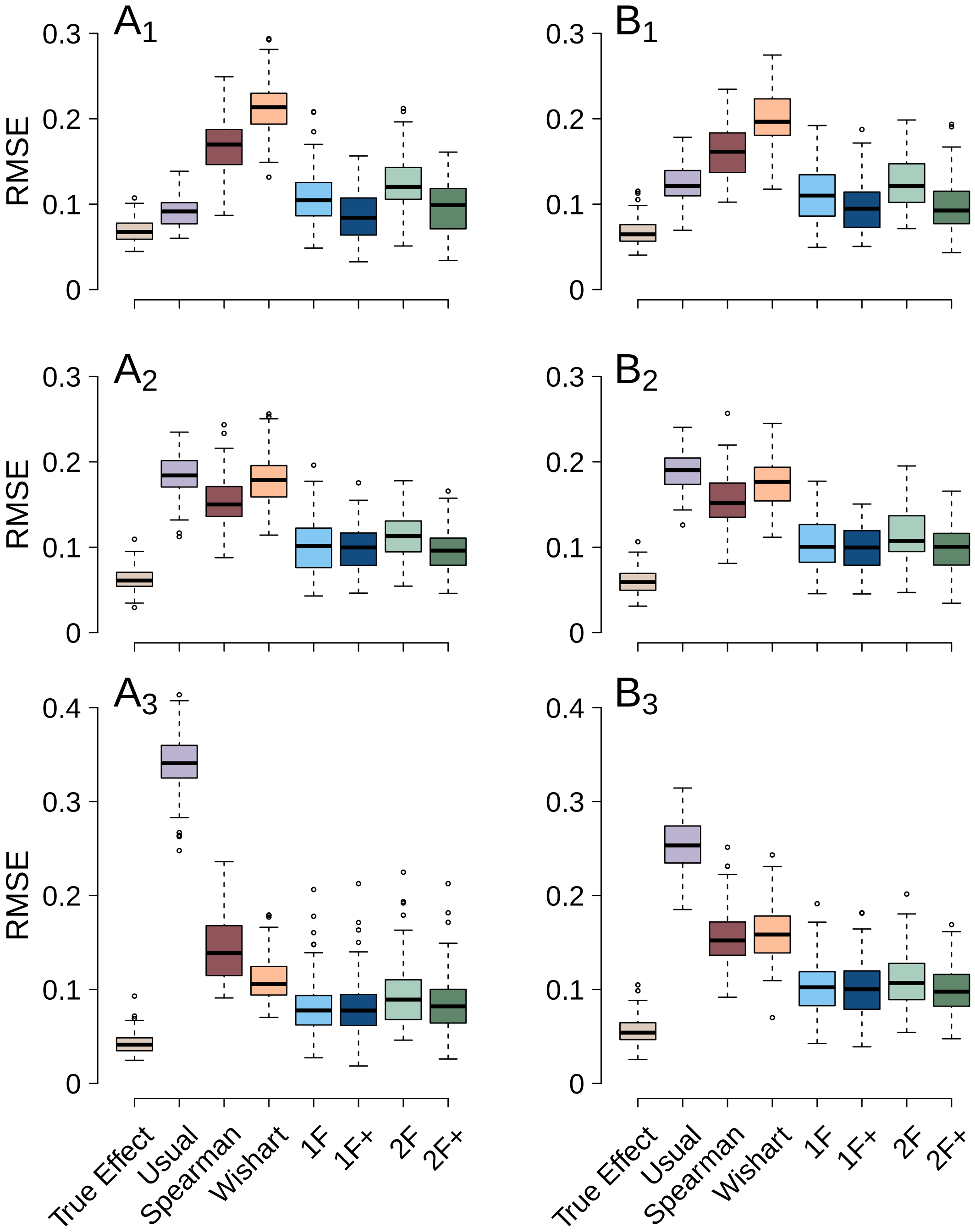


Figure 4

Distributions of RMSE across the 108 simulation runs for the six ground truths from the one-factor setup. Panel labels, such as A_1, correspond to true correlation matrices in Figure 3.

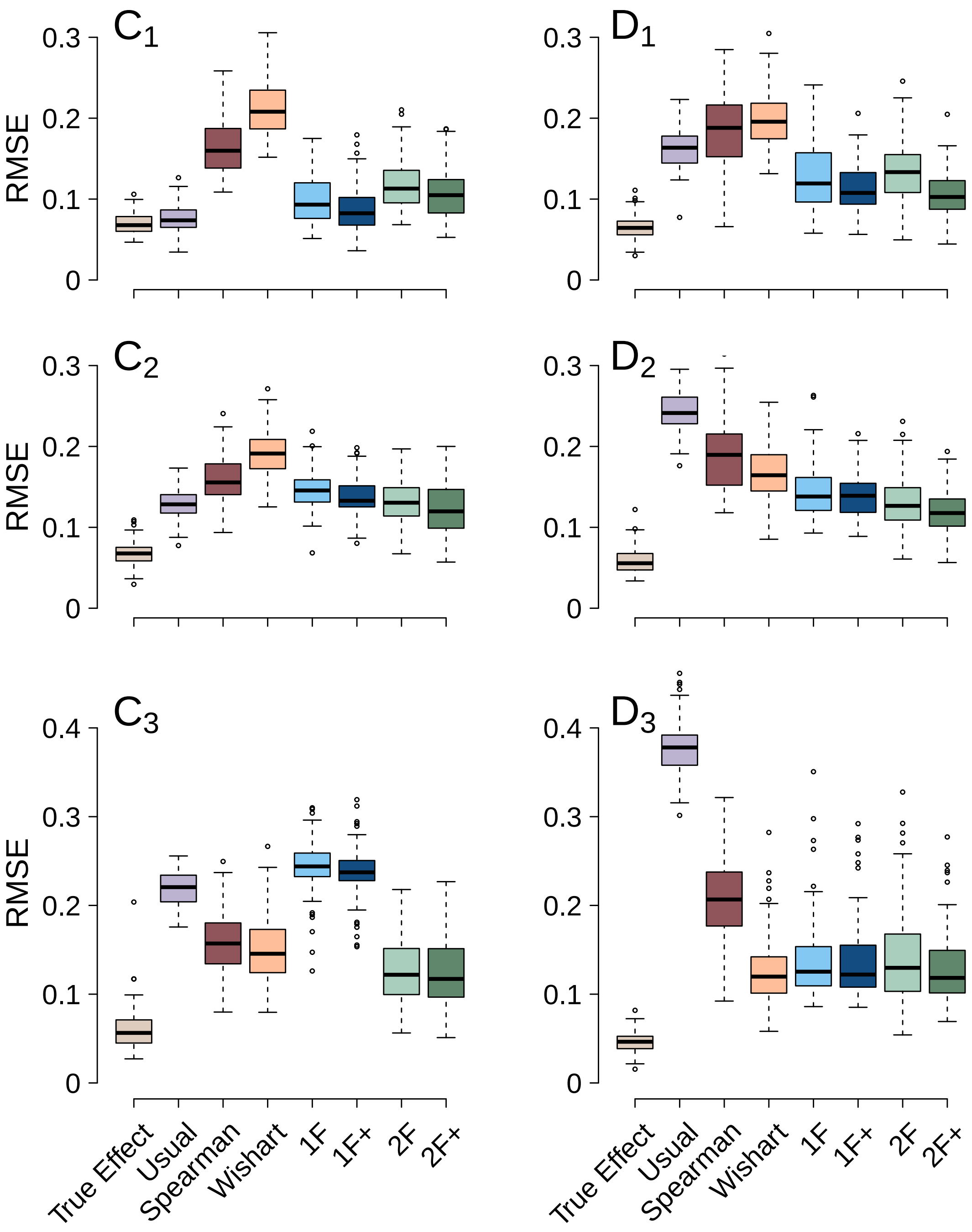


Figure 5

Distributions of RMSE across the 108 simulation runs for the six ground truths from the two-factor setup. Panel labels, such as C_1, correspond to true correlation matrices in Figure 3

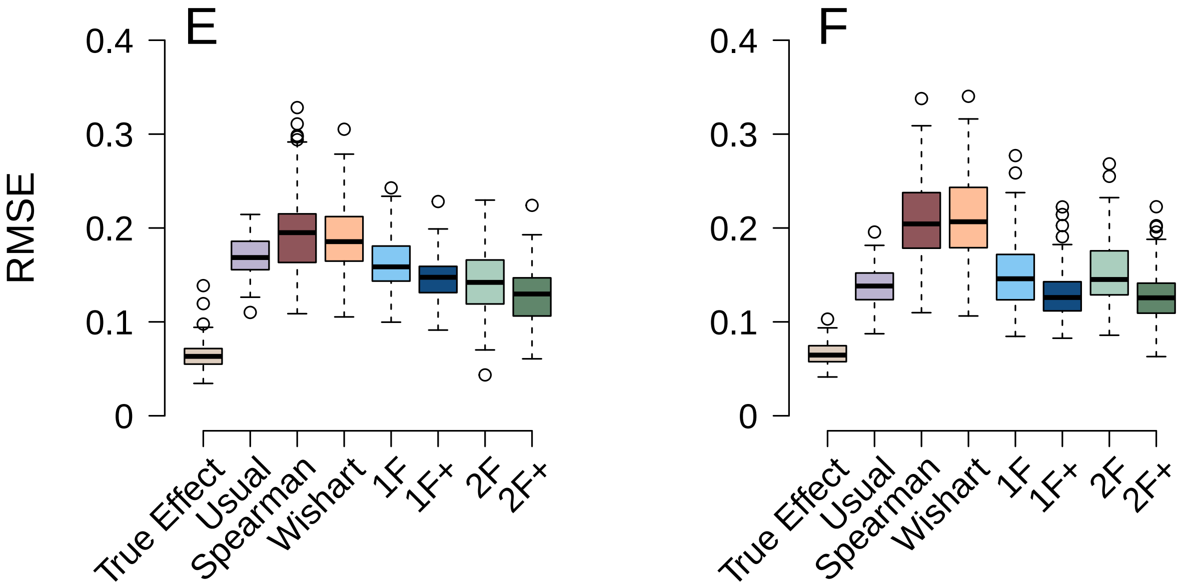


Figure 6

Distributions of RMSE across the 108 simulation runs for the two ground truths from the bi-factor setups. Panel labels, such as E, correspond to true correlation matrices in Figure 3

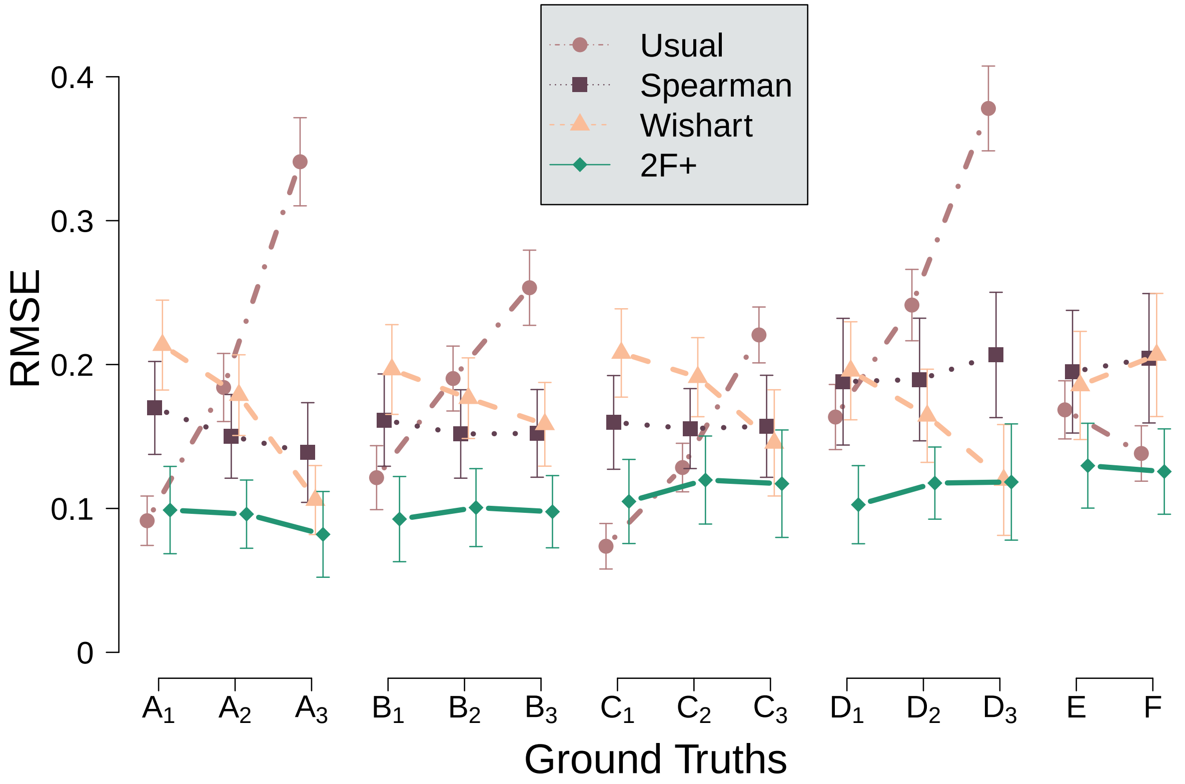


Figure 7

Summary of RMSE gain across all ground truths for select estimates. The two-factor positive constraint model does the best, overall, for localizing correlations.

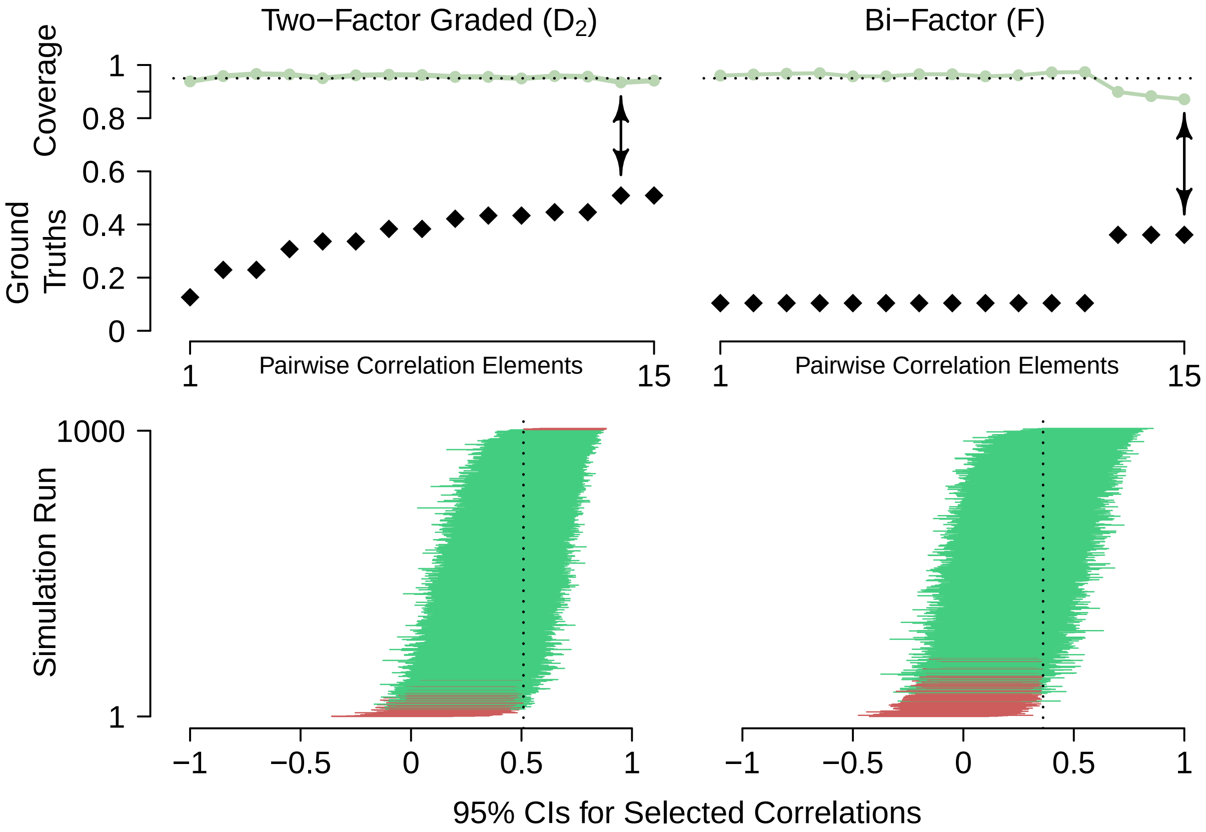


Figure 8

Coverage proportions and credible intervals for the two-factor model. Top: Coverage proportions and true correlation values for the two-factor graded ground truth (left) and bi-factor ground truth (right). The arrow indicates true values with the worst coverage. Bottom: Worst-case distribution of 95% credible intervals. There is one CI for each run. Green and red intervals respectively cover and fail-to-cover true values (black dotted line). These are worst-case distributions because they correspond to the true correlation value with the worst coverage.