Corrected Abstract for: Dissecting high-dimensional traits with Bayesian sparse factor analysis of genetic covariance matrices

Daniel E. Runcie, Sayan Mukherjee

March 19, 2017

1 Appendix

1.1 Posterior sampling:

We estimate the posterior distribution of the Bayesian genetic sparse factor model with an adaptive partially collapsed Gibbs sampler (van Dyk and Park 2011) based on the procedure proposed by Bhattacharya and Dunson (2011). The value k^* at which columns in $\mathbf{\Lambda}$ are truncated is set using an adaptive procedure (Bhattacharya and Dunson 2011). Given a truncation point, the following conditionally posterior distributions are sampled from in order:

1. The full conditional posterior distribution of the truncated factor loading matrix $\mathbf{\Lambda}_{k^*}$ is dependent on the parameters $\mathbf{B},\ \mathbf{E}_a,\ \mathbf{F} = \mathbf{Z}\mathbf{F}_a + \mathbf{F}_r$, and $\mathbf{\Psi}_r = \text{Diag}(\psi_{r_j})$. The full density factors into independent multivariate normal densities (MVNs) for each row of $\mathbf{\Lambda}_{k^*}$:

   $$\pi(\boldsymbol{\lambda}_j \mid \mathbf{y}_j,\ \mathbf{b}_j,\ \mathbf{e}_{a_j},\ \ \mathbf{F},\ \psi_{r_j}) \sim \text{N}(\psi_{r_j}^{-1}\mathbf{C}^{-1}\mathbf{F}^T(\mathbf{y}_j - \mathbf{X}\mathbf{b}_j - \mathbf{Z}\mathbf{e}_{a_j}), \mathbf{C}^{-1}),$$

   where: $\mathbf{C} = \psi_{r_j}^{-1}\mathbf{F}^T\mathbf{F} + \text{Diag}(\psi_{jl}^{-1}\tau_j^{-1})$.

   To speed up the MCMC mixing, we partially collapse this Gibbs update step by marginalizing over $\mathbf{E}_a \sim \text{N}(\mathbf{0}, \mathbf{A}, \mathbf{\Psi}_a)$. Let $\mathbf{\Psi}_a = \text{Diag}(\psi_{a_j})$:

   $$\pi_{/\mathbf{e}_{a_j}}(\boldsymbol{\lambda}_j \mid \mathbf{y}_j,\ \mathbf{b}_j,\ \mathbf{F},\ \psi_{a_j},\ \psi_{r_j}) \sim \text{N}(\mathbf{C}^{*-1}\mathbf{F}^T(\psi_{r_j}\mathbf{I}_n + \psi_{a_j}\mathbf{Z}\mathbf{A}\mathbf{Z}^T)^{-1}(\mathbf{y}_j - \mathbf{X}\mathbf{b}_j), \mathbf{C}^{*-1}),$$

   where: $\mathbf{C}^* = \mathbf{F}^T(\psi_{r_j}\mathbf{I}_n + \psi_{a_j}\mathbf{Z}\mathbf{A}\mathbf{Z}^T)^{-1}\mathbf{F} + \text{Diag}(\psi_{jl}^{-1}\tau_j^{-1})$.

   The matrix sum $\psi_{r_j}\mathbf{I}_n + \psi_{a_j}\mathbf{Z}\mathbf{A}\mathbf{Z}^T$ can be efficiently inverted each MCMC iteration by pre-calculating a unitary matrix $\mathbf{U}$ and a diagonal matrix $\mathbf{S}$ such that $\mathbf{Z}\mathbf{A}\mathbf{Z}^T = \mathbf{U}\mathbf{S}\mathbf{U}^T$. Thus, $(\psi_{r_j}\mathbf{I}_n + \psi_{a_j}\mathbf{Z}\mathbf{A}\mathbf{Z}^T)^{-1} = \mathbf{U}\ \text{Diag}(1/(\psi_{r_j} + \psi_{a_j}s_{ii}))\mathbf{U}^T$ which does not require a full matrix inversion.

2. The full conditional posterior distribution of the joint matrix $[\mathbf{B}^T\ \mathbf{E}_a^T]^T$ is dependent on the parameters $\mathbf{F}, \mathbf{\Lambda}, \mathbf{\Psi}_a$, and $\mathbf{\Psi}_r$. The full density factors into independent MVNs for each column of the matrix:

   $$\pi\left(\left[\begin{array}{c}\mathbf{b}_j\\ \mathbf{e}_{a_j}\end{array}\right] \mid \mathbf{y}_j,\ \boldsymbol{\lambda}_j,\ \mathbf{F},\ \psi_{a_j},\ \psi_{r_j}\right) \sim \text{N}\left(\psi_{r_j}^{-1}\mathbf{C}^{-1}\mathbf{W}^T(\mathbf{y}_j - \mathbf{F}\boldsymbol{\lambda}_j^T), \mathbf{C}^{-1}\right),$$

where $\mathbf{W}$ and $\mathbf{C}$ are defined as:

$$\mathbf{W} = [\mathbf{X}\ \mathbf{Z}]$$
$$\mathbf{C} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \psi_{a_j}^{-1}\mathbf{A}^{-1} \end{bmatrix} + \psi_{r_j}^{-1}\mathbf{W}^T\mathbf{W}.$$

The precision matrix $\mathbf{C}$ can be efficiently inverted each MCMC iteration by pre-calculating the unitary matrix $\mathbf{U}$ and diagonal matrices $\mathbf{S}_1$ and $\mathbf{S}_2$ as the generalized singular value decomposition of the two components of $\mathbf{C}$ such that $\mathbf{C}^{-1} = \mathbf{U}\ \mathrm{Diag}(1/(\psi_{a_j} s_{1_{ii}} + \psi_{r_j} s_{2_{ii}}))\mathbf{U}^T$ which does not require a full matrix inversion.

3. The full conditional posterior distribution of the latent factor heritabilities, $\mathbf{\Sigma}_a = \mathrm{Diag}(h_j^2)$, is dependent on $\mathbf{F}$ and $\mathbf{F}_a$. The density factors into independent distributions for each h_j^2, each of which has the form of a multinomial distribution since the prior on this parameter is discrete. This update step can be partially collapsed by marginalizing over $\mathbf{F}_a \sim \mathrm{N}(\mathbf{0}, \mathbf{A}, \mathbf{\Sigma}_a)$. The partially collapsed density is normalized by summing over all possibilities of h_j^2:

$$\pi_{/\mathbf{f}_{a_j}}\left(h_j^2 = h^2 \mid \mathbf{f}_j\right) = \frac{\mathrm{N}\left(\mathbf{f}_j \mid \mathbf{0}, h^2\mathbf{Z}\mathbf{A}\mathbf{Z}^T + (1-h^2)\mathbf{I}_n\right)\pi_{h_j^2}(h^2)}{\sum_{l=1}^{n_h} \mathrm{N}\left(\mathbf{f}_j \mid \mathbf{0}, h_l^2\mathbf{Z}\mathbf{A}\mathbf{Z}^T + (1-h_l^2)\mathbf{I}_n\right)\pi_{h_j^2}(h_l^2)}$$

where $\mathrm{N}(\mathbf{x} \mid \mu, \Sigma)$ is the MVN with mean μ and variance Σ, evaluated at $\mathbf{x}$, $h_l^2 = l/n_h$, and $\pi_{h_j^2}(h^2)$ is the prior probability that $h_j^2 = h^2$. Given this conditional posterior, h_j^2 is sampled from a multinomial distribution. The MVN densities can be calculated efficiently with the diagonalization matrices given in step 1.

4. The full conditional posterior distribution of the genetic effects on the factors, $\mathbf{F}_a$ depends on $\mathbf{F}$ and $\mathbf{\Sigma}_a$. This distribution factors into independent MVNs for each column $\mathbf{f}_{a_j}, j = 1 \ldots k^*$ st $h_j^2 \neq 0$:

$$\pi\left(\mathbf{f}_{a_j} \mid \mathbf{f}_j, h_j^2\right) \sim \mathrm{N}\left(\mathbf{C}^{-1}(1-h_j^2)^{-1}\mathbf{Z}^T\mathbf{F}_j, \mathbf{C}^{-1}\right)$$

where: $\mathbf{C} = (1-h_j^2)^{-1}\mathbf{Z}^T\mathbf{Z} + (h_j^2)^{-1}\mathbf{A}^{-1}$.

The precision matrix $\mathbf{C}$ can be efficiently inverted each MCMC iteration in the same manner as in step 2.

5. The residuals of the genetic effects on the factor scores, $\mathbf{F}_r$, can be calculated as $\mathbf{F} - \mathbf{F}_a$. The full conditional posterior distribution of $\mathbf{F}$ is a matrix variate normal distribution that depends on $\mathbf{\Lambda},\ \mathbf{B},\ \mathbf{E}_a, \mathbf{\Sigma}_{h^2}$ and $\mathbf{\Psi}_r$:

$$\begin{aligned}&\pi\left(\mathbf{F} \mid \mathbf{Y},\ \mathbf{\Lambda},\ \mathbf{B},\ \mathbf{E}_a, \mathbf{\Sigma}_{h^2},\ \mathbf{\Psi}_r\right)\\ &\sim \mathrm{MN}_{n,k^*}\left(\mathbf{C}^{-1}\left((\mathbf{Y} - \mathbf{XB} - \mathbf{ZE}_a)\mathbf{\Psi}_r^{-1}\mathbf{\Lambda}_{k^*} + \mathbf{ZF}_a(\mathbf{I}_{k^*} - \mathbf{\Sigma}_{h^2})^{-1}\right), \mathbf{C}^{-1}\right)\end{aligned}$$

where $\mathbf{C} = \mathbf{\Lambda}_{k^*}^T\mathbf{\Psi}_r^{-1}\mathbf{\Lambda}_{k^*} + (\mathbf{I}_{k^*} - \mathbf{\Sigma}_{h^2})^{-1}$.

6. The conditional posterior of the factor loading precision parameter ϕ_{ij} for trait i on factor j is:

$$\pi(\phi_{ij} \mid \tau_j, \lambda_{ij}) \sim \mathrm{Ga}\left(\frac{\nu+1}{2}, \frac{\nu + \tau_j\lambda_{ij}^2}{2}\right).$$

7. The conditional posterior of $\delta_m,\ m = 1 \ldots k^*$ is as follows. For δ_1:

$$\pi(\delta_1 \mid \phi, \tau_l^{(1)}, \mathbf{\Lambda}) \sim \mathrm{Ga}\left(a_1 + \frac{pk^*}{2}, b_1 + \frac{1}{2}\sum_{l=1}^{k^*}\tau_l^{(1)}\sum_{j=1}^{p}\phi_{jl}\lambda_{jl}^2\right)$$

and for $\delta_h, h \geq 2$:

$$\pi(\delta_h \mid \phi, \tau^{(h)_l}, \mathbf{\Lambda}) \sim \mathrm{Ga}\left(a_2 + \frac{p}{2}(k^* - h + 1), b_2 + \frac{1}{2}\sum_{l=h}^{k^*}\tau_l^{(h)}\sum_{j=1}^{p}\phi_{jl}\lambda_{jl}^2\right)$$

where $\tau_l^{(h)} = \prod_{t=1, t\neq h}^{l}\delta_t$.

8. The conditional posterior of the precision of the residual genetic effects of trait j is:

$$\pi(\psi_{a_j}^{-1} \mid \mathbf{e}_{a_j}) \sim \mathrm{Ga}\left(a_g + \frac{r}{2}, b_g + \frac{1}{2}\mathbf{e}_{a_j}^T\ \mathbf{A}^{-1}\mathbf{e}_{a_j}\right).$$

9. The conditional posterior of the model residual precision of trait j is:

$$\pi(\psi_{e_j}^{-1} \mid -) \sim \mathrm{Ga}\left(a_r + \frac{n}{2}, b_r + \frac{1}{2}\sum_{i=1}^{n}\left(y_{ij} - \mathbf{x}^{(i)}\mathbf{b}_j - \mathbf{f}^{(i)}\mathbf{\lambda}_j^T - \mathbf{z}^{(i)}\mathbf{e}_{a_j}\right)^2\right).$$

10. If missing observations are present, values are drawn independently from univariate normal distributions parameterized by the current values of all other parameters:

$$\pi(y_{ij} \mid -) \sim \mathrm{N}\left(\mathbf{x}^{(i)}\mathbf{b}_j + \mathbf{f}^{(i)}\boldsymbol{\lambda}_j^T + \mathbf{z}^{(i)}\mathbf{e}_{a_j}, \psi_j\right)$$

where y_{ij} is the imputed phenotype value for the j-th trait in individual i. The three components of the mean are: $\mathbf{x}^{(i)}$ the row vector of fixed effect covariates for individual i times $\mathbf{b}_j$, the jth column of the fixed effect coefficient matrix; $\mathbf{f}^{(i)}$, the row vector of factor scores on the k^* factors for individual i times $\boldsymbol{\lambda}_j^T$, the row of the factor loading matrix for trait j; and $\mathbf{z}^{(i)}$, the row vector of the random (genetic) effect incidence matrix for individual i times $\mathbf{e}_{a_j}$, the vector of residual genetic effects for trait j not accounted for by the k^* factors. Finally, ψ_j is the residual variance of trait j. All missing data is drawn in a single block update.

Other random effects, such as the line $\times$ sex effects modeled in the gene expression example of this paper can be incorporated into this sampling scheme in much the same way as the residual genetic effects, $\mathbf{E}_a$, are included here.

2 Acknowledgments

We would like to thank Barbara Engelhardt, Iulian Pruteanu-Malinici, Jenny Tung, and two anonymous reviewers for comments and advice on this method. We would like to thank Mark Grotte and David Katz for pointing out errors in the original text.

References

Ayroles, J. F., M. A. Carbone, E. A. Stone, K. W. Jordan, R. F. Lyman, M. M. Magwire, S. M. Rollmann, L. H. Duncan, F. Lawrence, R. R. H. Anholt, and T. F. C. Mackay, 2009 Systems genetics of complex traits in Drosophila melanogaster. Nat Genet *41*(3)**:** 299–307.

Bhattacharya, A. and D. B. Dunson, 2011 Sparse Bayesian infinite factor models. Biometrika *98*(2)**:** 291–306.

Blows, M. W., S. F. Chenoweth, and E. Hine, 2004 Orientation of the genetic variance-covariance matrix and the fitness surface for multiple male sexually selected traits. The American Naturalist *163*(3): 329–340.

Krzanowski, W. J., 1979 Between-Groups Comparison of Principal Components. J Am Stat Assoc *74*(367): 703–707.

Stone, E. A. and J. F. Ayroles, 2009 Modulated modularity clustering as an exploratory tool for functional genomic inference. PLoS Genet *5*(5): e1000479.

van Dyk, D. A. and T. Park, 2011 Partially collapsed Gibbs sampling & path-adaptive Metropolis-Hastings in high-energy astrophysics. In S. Brooks, A. Gelman, G. Jones, and X.-L. Meng (Eds.), *Handbook of Markov Chain Monte Carlo*, pp. 383–397. New York, NY: Chapman & Hall/CRC Handbooks of Modern Statistical Methods.

3 Figures

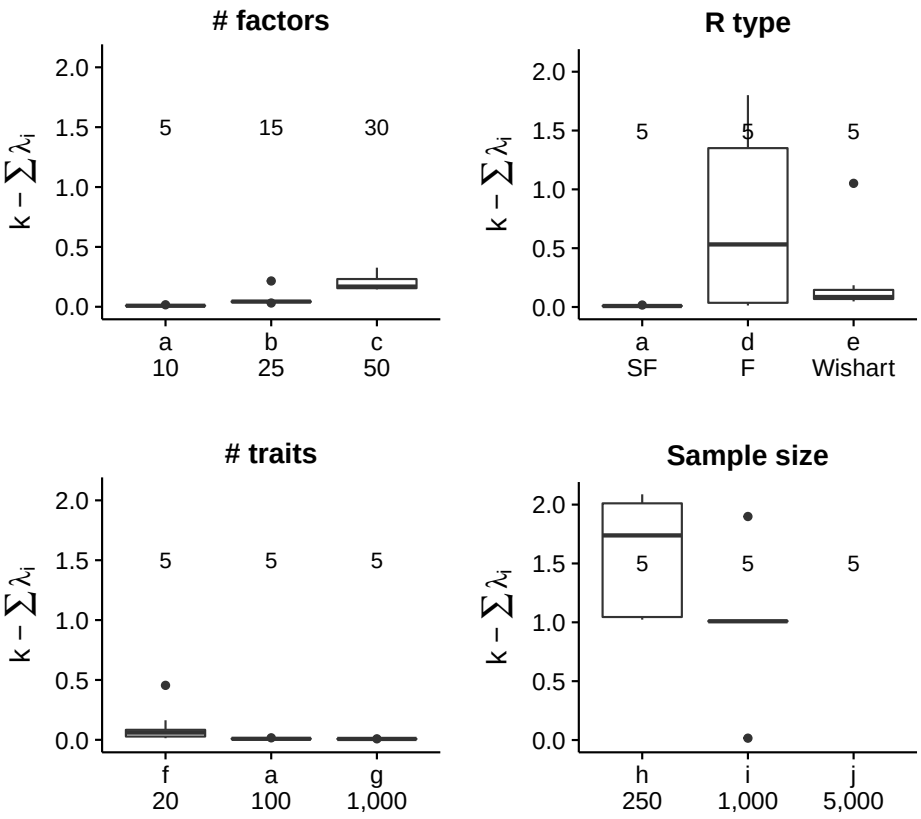

Figure 1: **The Bayesian genetic sparse factor model accurately estimates the dominant subspace of high-dimensional G matrices.** Each subplot shows the distribution of Krzanowski's statistics ($\sum \lambda_{s_i}$, Krzanowski 1979; Blows, Chenoweth, and Hine 2004) calculated for posterior mean estimates of **G** across a related set of scenarios. Plotted values are $k - \sum \lambda_{s_i}$ so that statistics are comparable across scenarios with different subspace dimensions. On this scale, identical subspaces have a value of zero and values increase as the subspaces diverge. The value of k used in each scenario is listed inside each boxplot. The difference from zero roughly corresponds to the number of eigenvectors of the true subspace missing from the estimated subspace. Different parameters were varied in each set of simulations as listed below each box. **A**. Increasing numbers of simulated factors. **B**. Different properties of the **R** matrix. "SF": a sparse-factor form for **R**, "F": a (non-sparse) factor form for **R**, "Wishart": **R** was sampled from a Wishart distribution. **C**. Different numbers of traits. **D**. Different numbers of sampled individuals. Note that in scenarios *h-j*, factor h^2s ranged from 0.0 to 0.9. Complete parameter sets describing each simulation are described in Table **??**.

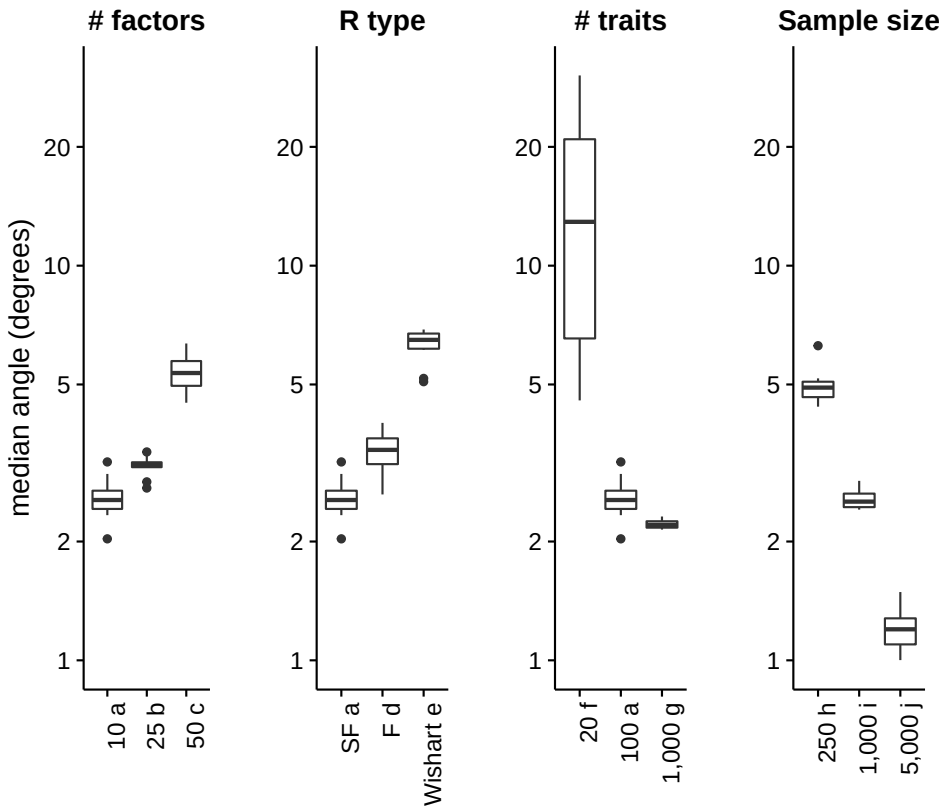

Figure 2: **Latent factors were accurately recovered in most simulations.** The true factors in each simulation were matched to the most similar estimated factor by calculating the minimum vector angle between each true factor and an estimated factor. The median error angle across factors in each simulation is plotted. Boxplots show the distribution of median error angles by scenario. Two identical vectors have an angle of zero. Completely orthogonal vectors have an angle of 90. **A**. Increasing numbers of simulated factors. **B**. Different properties of the **R** matrix. **C**. Different numbers of traits. **D**. Different numbers of sampled individuals.

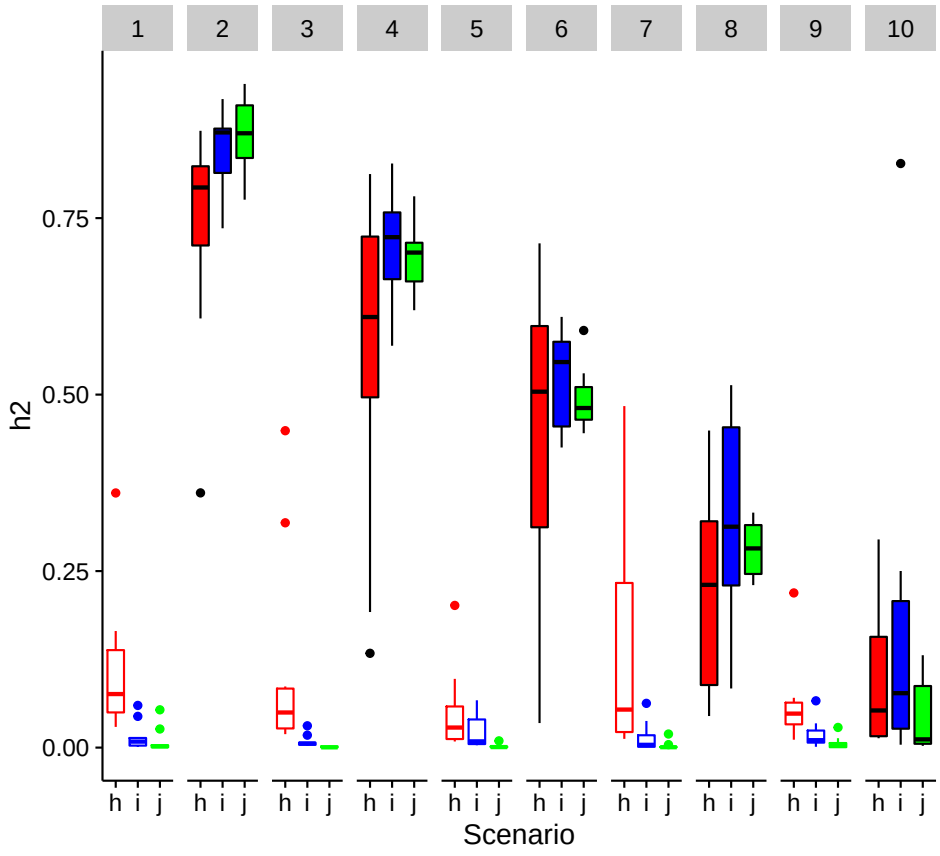

Figure 3: **Latent factor heritabilities were accurately recovered.** Distributions of factor h^2 estimates for scenarios *h*-*j*. These scenarios differed in the number of individuals sampled. 10 factors were generated in each simulation and assigned h^2s between 0.0 and 0.9. After fitting our factor model to each simulated dataset, the simulated factors were matched to estimated factors based on the trait-loading vector angles. Each boxplot shows the distribution of h^2 estimates for each simulated factor across 10 simulations. Note that the trait-loadings for each factor differed in each simulation; only the h^2 values remained the same. Thin horizontal lines in each column show the simulated h_j^2 values. Colors correspond to the scenario, and filled boxes/circles are used for factors with $h_j^2 > 0$.

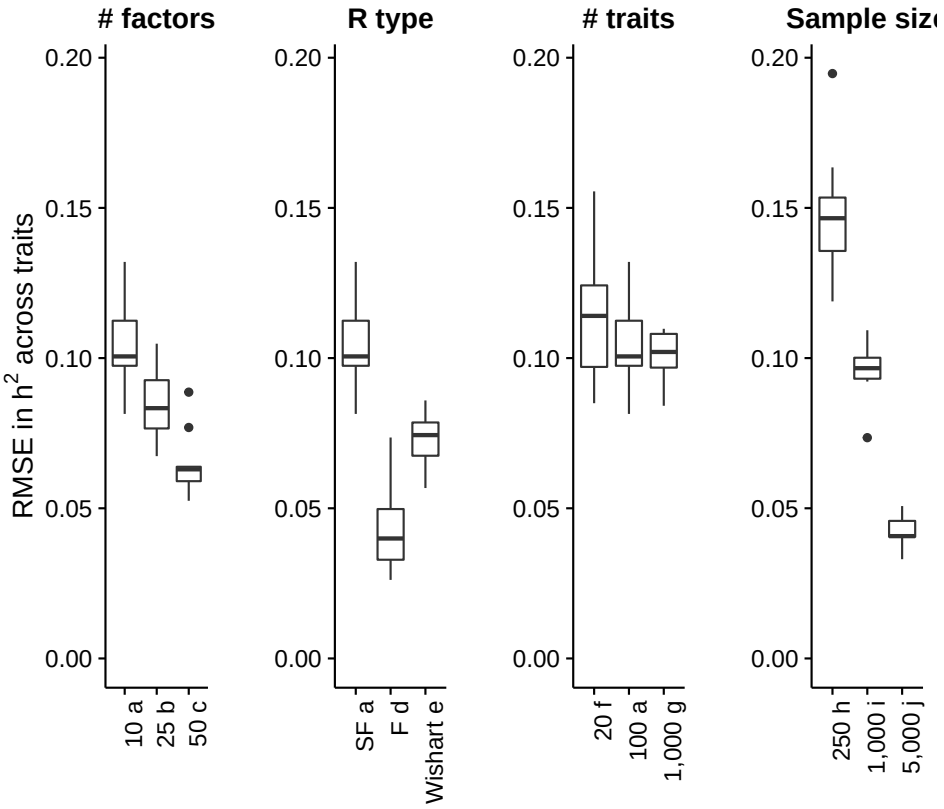

Figure 4: **Heritability estimates for each individual trait were accurate.** The heritability of each individual trait was estimated as $h_i^2 = \mathbf{G}_{ii}/\mathbf{P}_{ii}$. RMSE $= \sqrt{\frac{1}{p}\sum_{i=1}^{p}(\hat{h}_i^2 - h_i^2)^2}$ was calculated for each simulation. Boxplots show the distribution of RMSE values for each scenario. **A**. Increasing numbers of simulated factors. **B**. Different properties of the **R** matrix. **C**. Different numbers of traits. **D**. Different numbers of sampled individuals.

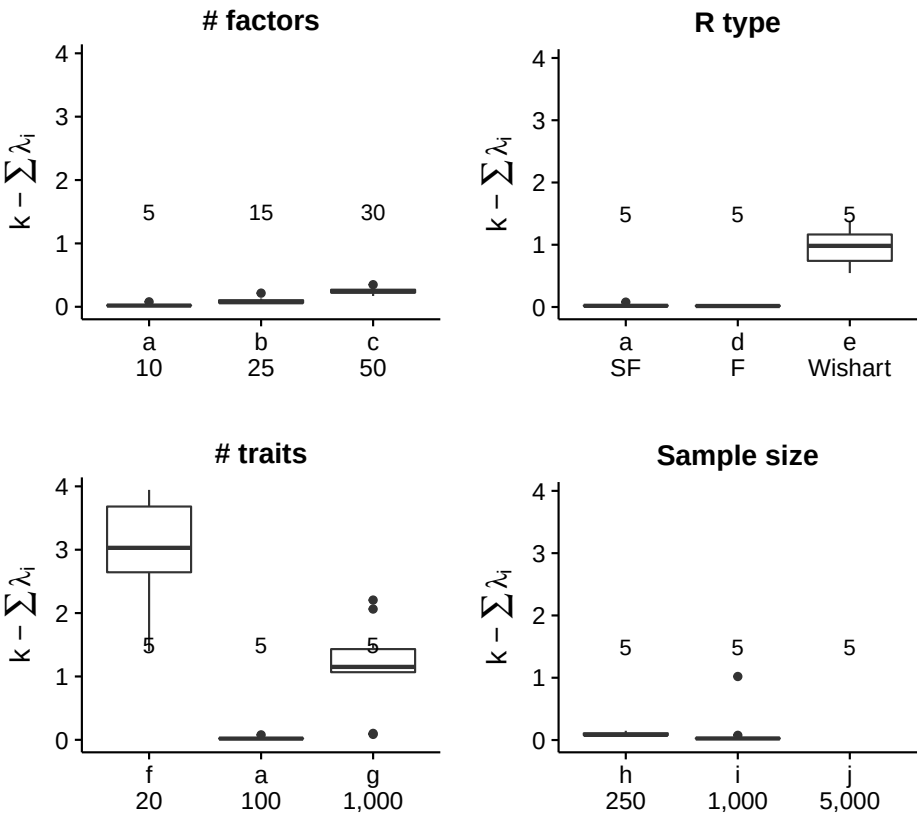

Figure S1: **P-matrix subspaces were accurately recovered.** This figure is identical to Figure 1 but for **P**. Each subplot shows the distribution of Krzanowski's statistics ($\sum \lambda_{s_i}$) calculated for posterior mean estimates of **P** across a related set of scenarios. The value of k used in each scenario is listed inside each boxplot. The parameter varied in each set of simulations is described at the bottom. (A) Increasing numbers of simulated factors. (B) Different properties of the **R** matrix. "SF": a sparse-factor form for **R**, "F": a (non-sparse) factor form for **R**, "Wishart": **R** was sampled from a Wishart distribution. In scenario *e*, the residual matrix did not have a factor form. Therefore, we chose $k = 19$ for the phenotypic covariance matrix because the corresponding eigenvectors each explained $> 1\%$ of total phenotypic variation. (C) Different numbers of traits. (D) Different numbers of sampled individuals. Complete parameter sets describing each simulation are described in Table 1.

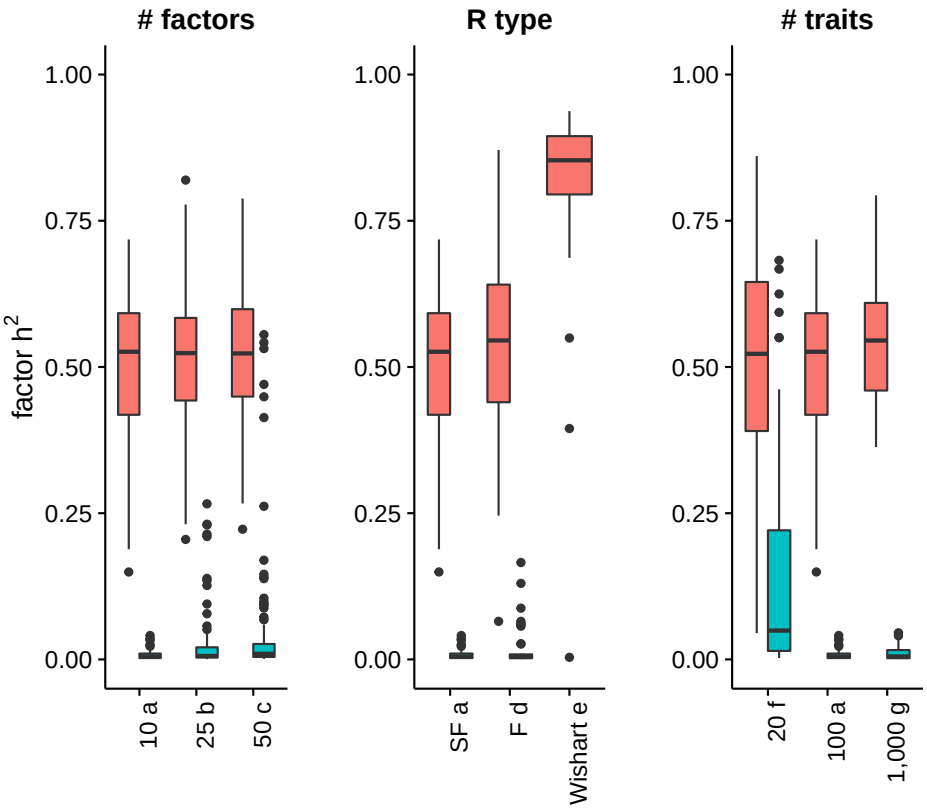

Figure S2: **Latent factor heritabilities were accurately recovered.** Distributions of factor h^2 estimates by simulation scenario. Each simulated factor was matched to the estimated factor with most similar trait-loadings as in Figure 3. Thin horizontal lines in each column show the simulated h^2_j values. Each simulated factor was assigned $h^2 = 0.0$ (black) or 0.5 (red), except in scenario *e* where all five factors were assigned $h^2 = 1$ (red). h^2 estimates are grouped across all 10 simulations of each scenario. (A) Increasing numbers of simulated factors. (B) Different properties of the **R** matrix. (C) Different numbers of traits.