Supplementary Materials for "Bayesian Combinatorial Multi-Study Factor Analysis"

A Illustration of Prior Model

We illustrate the effect of using both the multiplicative gamma process shrinkage prior on Λ, which encourages increasing shrinkage with each factor, and the Indian Buffet Process (IBP) prior on $\mathcal{A}$, which encourages sparsity in the number of active factors, using a toy example. In particular, we simulate Λ and $\mathcal{A}$ under their priors in a setting with 60 features and 4 studies, similar to Scenarios 1 and 2 of the simulation studies. We consider two sets of hyperparameter values for each of $\mathcal{A}$ and Λ. For $\mathcal{A}$, we consider $(\alpha, \beta) = (3, 1)$, which encourages a smaller number of factors, and $(\alpha, \beta) = (6, 1)$, which encourages a larger number of factors. For Λ, we consider $a_2 = 3.1$, which encourages a slower rate of increasing shrinkage, and $a_2 = 4.1$, which encourages a faster rate of increasing shrinkage; in both cases, we consider $a_1 = 2.1, b_1 = 1, b_2 = 1$, and $\nu = 3$.

Figure 1 plots the Frobenius norms of the columns of the loadings matrix produced under each combination of hyperparameter values. The norm of column i is computed using the ith column of Λ if $\mathcal{A}$ indicates that this is an active factor, and otherwise is set to 0. These results show that increasing the degree of sparsity in the IBP prior on $\mathcal{A}$ (in this case, when $\alpha = 3$) yields a faster rate of increasing shrinkage. The properties of these two priors enhance one another: the IBP prior increases the effective rate of shrinkage, which in turn encourages greater amounts of signal in a smaller number of factors.

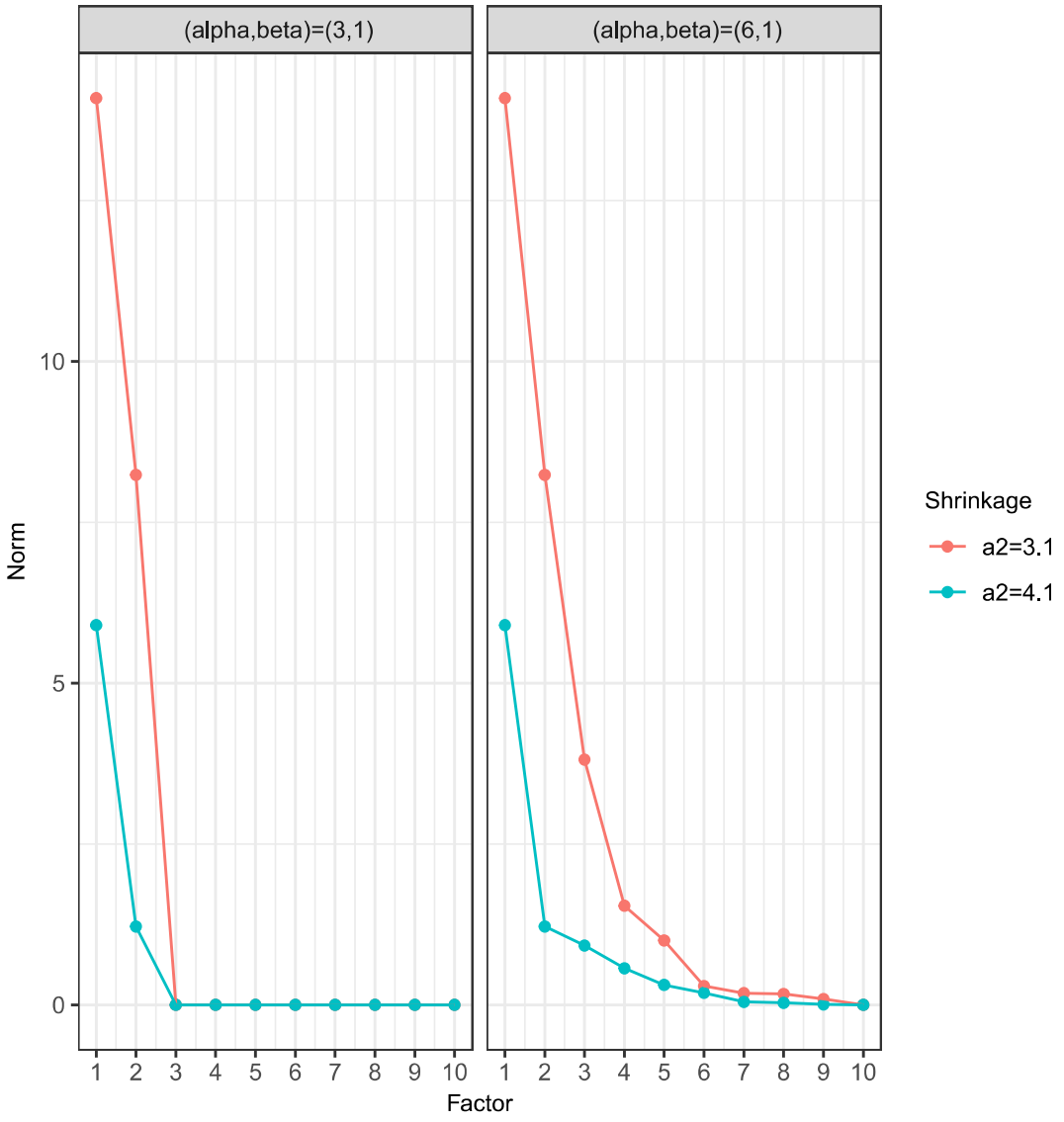

Figure 1: Frobenius norms of factor loadings when Λ and $\mathcal{A}$ are generated under their priors, using two sets of hyperparameter values each.

B Gibbs Sampling Algorithm for Posterior Inference

The Gibbs sampling algorithm for posterior inference with our model is as follows:

1. We follow Knowles and Ghahramani (2007) to update each entry of $\mathcal{A}$. For s ranging from 1 to S, for k ranging from $S+1$ to K, set $\mathcal{A}_{sk} = 1$ with probability

$r/(r+1)$, where

$$r = \tilde{m} \cdot \prod_{i=1}^{n_s} \exp\left(\frac{1}{2(\boldsymbol{\Lambda}_k^T \boldsymbol{\Psi}_s^{-1} \boldsymbol{\Lambda}_k + 1)^{-1}} \bar{l}^2\right) (\boldsymbol{\Lambda}_k^T \boldsymbol{\Psi}_s^{-1} \boldsymbol{\Lambda}_k + 1)^{-\frac{1}{2}},$$

for

$$\bar{l} = (\boldsymbol{\Lambda}_k^T \boldsymbol{\Psi}_s^{-1} \boldsymbol{\Lambda}_k + 1)^{-1} \boldsymbol{\Lambda}_k^T \boldsymbol{\Psi}_{\boldsymbol{s}}^{-1} (\boldsymbol{x}_{is} - \boldsymbol{\Lambda} \boldsymbol{A}_s^{k0} \boldsymbol{l}_{is}),$$

and $\tilde{m} = \Pr(\mathcal{A}_{sk} = 1 \mid \cdots)/\Pr(\mathcal{A}_{sk} = 0 \mid \cdots) = m_{k,-s}/(\beta + S - 1 - m_{k,-s})$. Here, $m_{k,-s}$ denotes the sum of column k of $\mathcal{A}$, subtracting the current value of $\mathcal{A}_{sk}$. Further, $\boldsymbol{A}_s^{k0}$ has the kth diagonal element of $\boldsymbol{A}_s$ equal to 0 (i.e. $\mathcal{A}_{sk} = 0$). New elements of $\boldsymbol{l}_{is}$ are sampled for each i, s as needed if $\mathcal{A}_{sk} = 1$ (see step 4 for more details).

2. Again as in Knowles and Ghahramani (2007) and as discussed by Doshi-Velez et al. (2009), we use a Metropolis-Hastings step to allow new factors to be added. For s ranging from 1 to S, to sample the number of new factors k_{new} for study s, sample $k_{\text{new}} \sim \text{Pois}\left(\frac{\alpha\beta}{\beta+S-1}\right)$, and the corresponding new elements ω_{new}, δ_{new}, and $\boldsymbol{l}_{s,\text{new}}$ from their respective priors outlined in the previous section. Then accept all these samples with probability $\min(1, r)$, for r equal to

$$\prod_{p=1}^{P} |2\pi \boldsymbol{D}_{p,\text{new}}|^{-\frac{1}{2}} |2\pi (\boldsymbol{D}_{p,\text{new}}^{-1} + \psi_{sp}^{-2} \boldsymbol{A}_{s,\text{new}} \boldsymbol{l}_{s,\text{new}}^T \boldsymbol{l}_{s,\text{new}} \boldsymbol{A}_{s,\text{new}})^{-1}|^{\frac{1}{2}}$$
$$\times \exp\left\{\frac{1}{2} \bar{\boldsymbol{\Lambda}}_{p,\text{new}} (\boldsymbol{D}_{p,\text{new}}^{-1} + \psi_{sp}^{-2} \boldsymbol{A}_{s,\text{new}} \boldsymbol{l}_{s,\text{new}}^T \boldsymbol{l}_{s,\text{new}} \boldsymbol{A}_{s,\text{new}}) \bar{\boldsymbol{\Lambda}}_{p,\text{new}}^T\right\}.$$

This is equivalent to the ratio of the likelihoods with and without the proposed factors, with the new elements of Λ marginalized out. Here, we define $\bar{\boldsymbol{\Lambda}}_{p,\text{new}}^T$ as

$$(\boldsymbol{D}_{p,\text{new}}^{-1} + \psi_{sp}^{-2} \boldsymbol{A}_{s,\text{new}} \boldsymbol{l}_{s,\text{new}}^T \boldsymbol{l}_{s,\text{new}} \boldsymbol{A}_{s,\text{new}})^{-1} \psi_{sp}^{-2} \boldsymbol{A}_{s,\text{new}} \boldsymbol{l}_{s,\text{new}}^T (\boldsymbol{x}_s^{(p)} - \boldsymbol{l}_s \boldsymbol{A}_s \boldsymbol{\Lambda}_p^T).$$

3. To update the loadings matrix elements and the associated parameters in the next 4 steps, we follow the main ideas of Bhattacharya and Dunson (2011) and the multi-study extension from De Vito et al. (2021). For p ranging from 1 to P, sample the transpose of the pth row of the factor loadings matrix

$$\boldsymbol{\Lambda}_p^T \sim \mathcal{N}\{\left(\boldsymbol{D}_p^{-1} + \sum_{s=1}^{S} \psi_{sp}^{-2} \boldsymbol{A}_s \boldsymbol{l}_s^T \boldsymbol{l}_s \boldsymbol{A}_s\right)^{-1} \left(\sum_{s=1}^{S} \psi_{sp}^{-2} \boldsymbol{A}_s \boldsymbol{l}_s^T \boldsymbol{x}_s^{(p)}\right),$$
$$\left(\boldsymbol{D}_p^{-1} + \sum_{s=1}^{S} \psi_{sp}^{-2} \boldsymbol{A}_s \boldsymbol{l}_s^T \boldsymbol{l}_s \boldsymbol{A}_s\right)^{-1}\},$$

where $\boldsymbol{l}_s$ is the $n_s \times K$ matrix of latent factors across samples for study s, $\boldsymbol{x}_s^{(p)}$ is the $n_s \times 1$ vector of values for feature p across samples for study s, and $\boldsymbol{D}_p^{-1} = \text{diag}(\omega_{p1}\tau_1, \ldots, \omega_{pK}\tau_K)$.

4. For s ranging from 1 to S, i ranging from 1 to n_s, sample

$$\boldsymbol{l}_{is} \sim \mathcal{N}\{(\boldsymbol{I}_K + \boldsymbol{A}_s\boldsymbol{\Lambda}^T\boldsymbol{\Psi}_s^{-1}\boldsymbol{\Lambda}\boldsymbol{A}_s)^{-1}(\boldsymbol{A}_s\boldsymbol{\Lambda}^T\Psi_s^{-1}\boldsymbol{x}_{is}), \\ (\boldsymbol{I}_K + \boldsymbol{A}_s\boldsymbol{\Lambda}^T\boldsymbol{\Psi}_s^{-1}\boldsymbol{\Lambda}\boldsymbol{A}_s)^{-1}\}.$$

   This samples a $K \times 1$ vector regardless of how many factors are shared by study s. If the kth factor is not shared by the study, then the kth diagonal element of $\boldsymbol{A}_s$ will be 0; as a result, this sampling step effectively draws the corresponding element of $\boldsymbol{l}_{is}$ from the prior, and that element of $\boldsymbol{l}_{is}$ will not affect the likelihood due to multiplication by $\boldsymbol{A}_s$.

5. For p ranging from 1 to P, for k ranging from 1 to K, sample

$$\omega_{pk} \sim \Gamma\left(\frac{\nu+1}{2}, \frac{\nu + \tau_k\Lambda_{pk}^2}{2}\right).$$

6. Sample $\delta_1 \sim \Gamma\left(a_1 + \frac{PK}{2}, 1 + \frac{1}{2}\sum_{k=1}^{K}\tau_k^{(1)}\sum_{p=1}^{P}\omega_{pk}\Lambda_{pk}^2\right)$, where $\tau_k^{(1)}$ represents $\frac{\tau_k}{\delta_1}$.

7. For l ranging from 2 to K, sample

$$\delta_l \sim \Gamma\left(a_2 + \frac{P}{2}(K-l+1), 1 + \frac{1}{2}\sum_{k=l}^{K}\tau_k^{(l)}\sum_{p=1}^{P}\omega_{pk}\Lambda_{pk}^2\right),$$

   where $\tau_k^{(l)}$ represents $\frac{\tau_k}{\delta_l}$.

8. For s ranging from 1 to S, for p ranging from 1 to P, sample

$$\psi_{sp}^{-2} \sim \Gamma\left(a_\psi + \frac{n_s}{2}, b_\psi + \frac{1}{2}\sum_{i=1}^{n_s}(\boldsymbol{x}_{is}^{(p)} - \boldsymbol{\Lambda}_p\boldsymbol{A}_s\boldsymbol{l}_{is})^2\right).$$

For the extension of Tetris where we assume the study labels z_i for each sample are not known, we add an additional step to the above algorithm. In particular, we define $\mathcal{L}_{is}$ equal to the density of $\boldsymbol{x}_i$ under $\mathcal{N}(0, \Lambda A_s\Lambda^T + \Psi_s)$. Then z_i is sampled from a categorical distribution with the probability of belonging to each study s proportional to $\mathcal{L}_{is}$.

C Computational Timing

In Table 1, we summarize the mean computational times for the main three steps of our approach using ten runs for at least one setting from each simulation scenario. All runs were carried out on a cluster computing node with one core. Specifically, these run times are carried out on a high-performance computing server managed and supported by the Dana-Farber Cancer Institute with 25 nodes, each with 385 Gb of memory and 40 cores per socket.

Scenario	S	P	n_s	K	**Step 1 (Mean Hours)**	**Step 2 (Mean Hours)**	**Step 3 (Mean Hours)**	**Total (Mean Hours)**
1	4	60	10	10	0.24	1.71	0.05	2.00
2	4	60	10	9	0.25	1.90	0.06	2.21
3	16	60	10	20	1.84	9.46	0.23	11.53
4	3	370	34-50	8	1.76	0.15	0.25	2.16
4	6	370	11-32	11	7.84	3.71	0.52	12.07

Table 1: Mean computational runtimes for running the sampler (Step 1), computing the point estimate of $\mathcal{A}$ (Step 2), and running the sampler conditional on this value (Step 3) based on ten runs for each listed setting. We describe the number of studies S, the number of features P, the number of samples per study n_s (the range is provided when n_s varies), and the ground-truth number of factors K.

D Additional Simulation Results

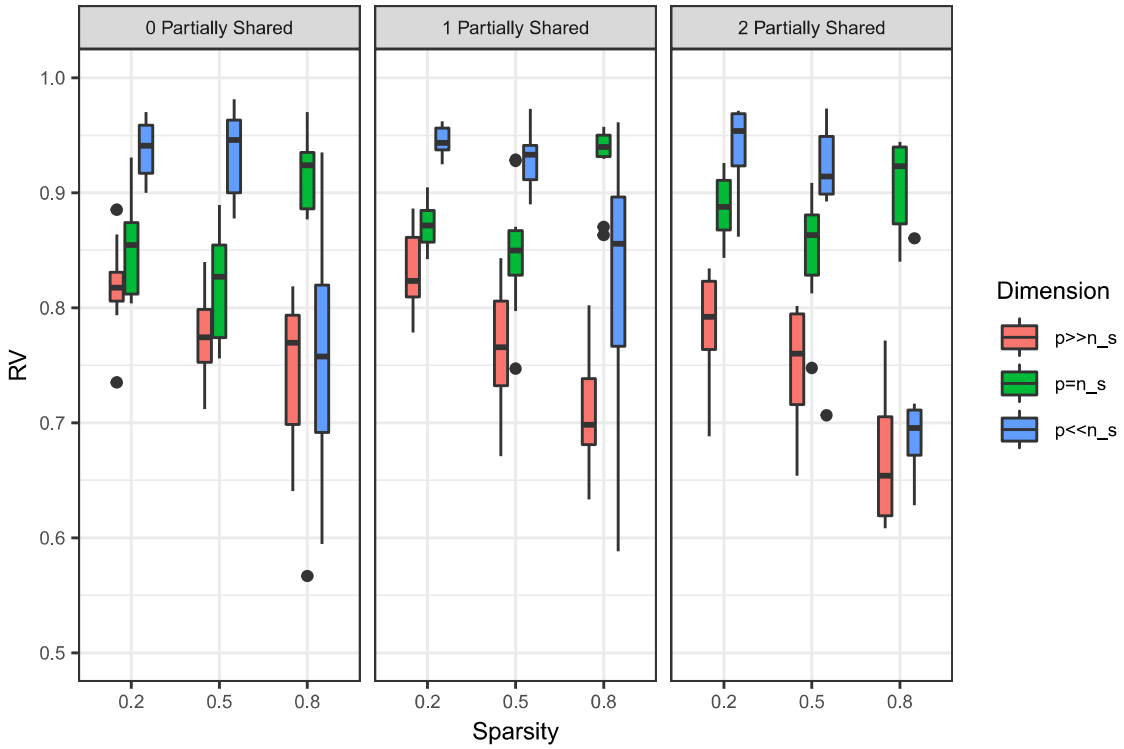

Figure 2: RV coefficients for the full loading matrix covariance across varying sparsity, data dimension, and number of partially shared factors.

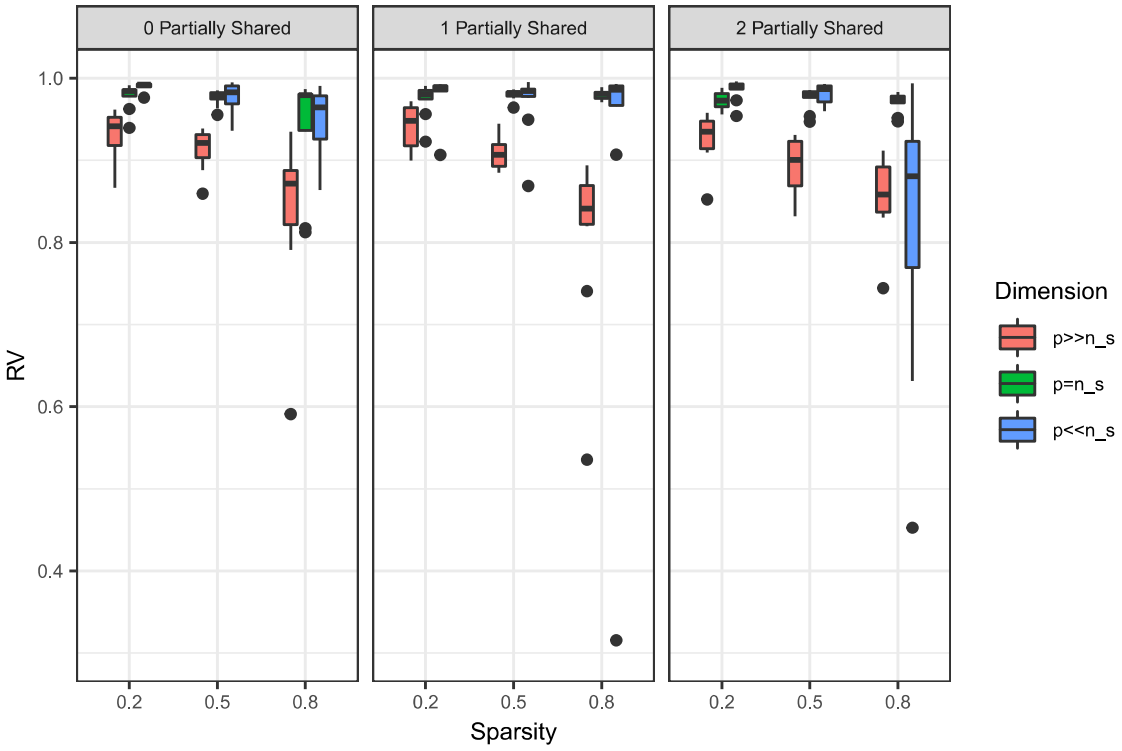

Figure 3: RV coefficients for the common loading matrix covariance across varying sparsity, data dimension, and number of partially shared factors.

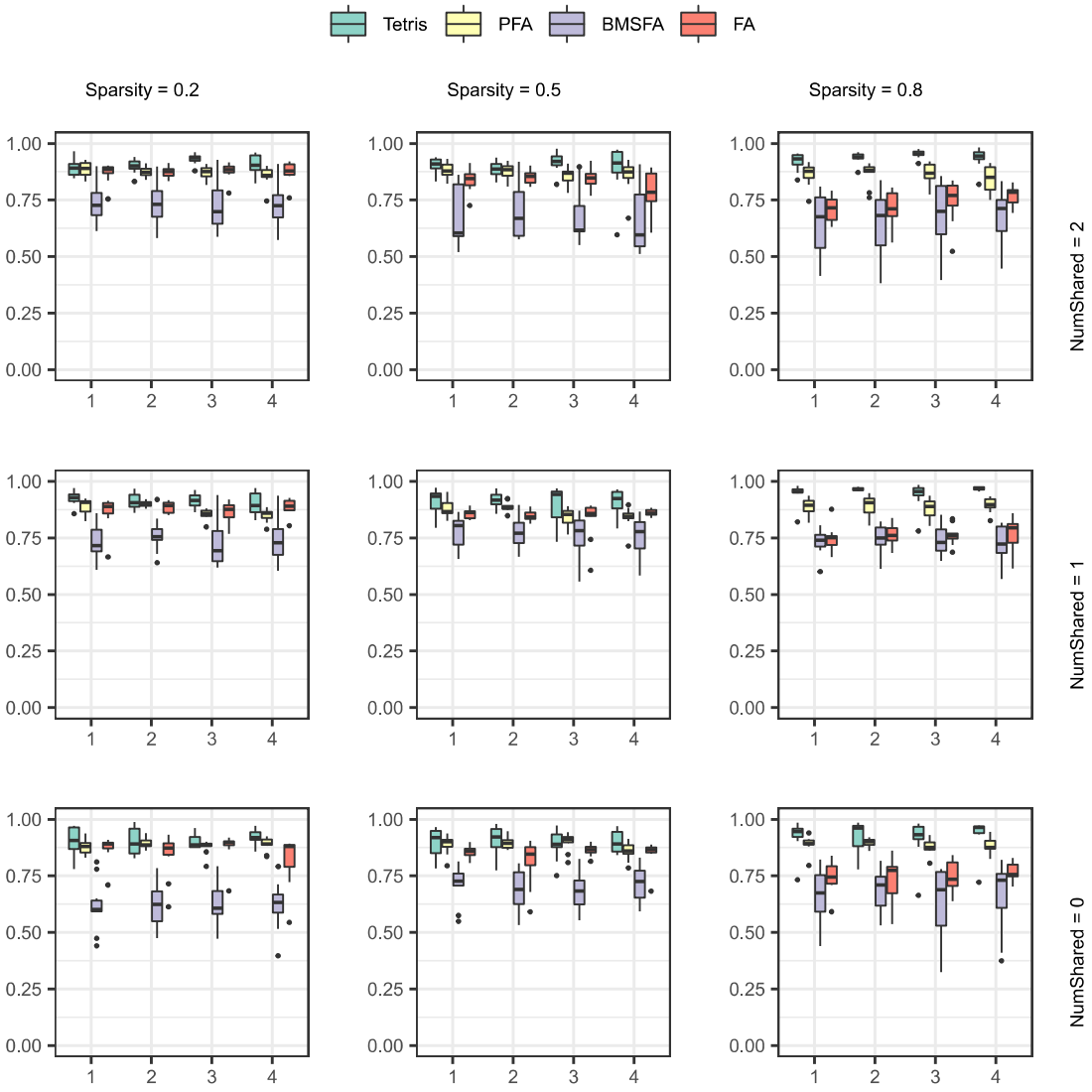

Figure 4: RV coefficients for study-specific covariances across varying sparsities and numbers of partially shared factors in the $p = n_s$ setting.

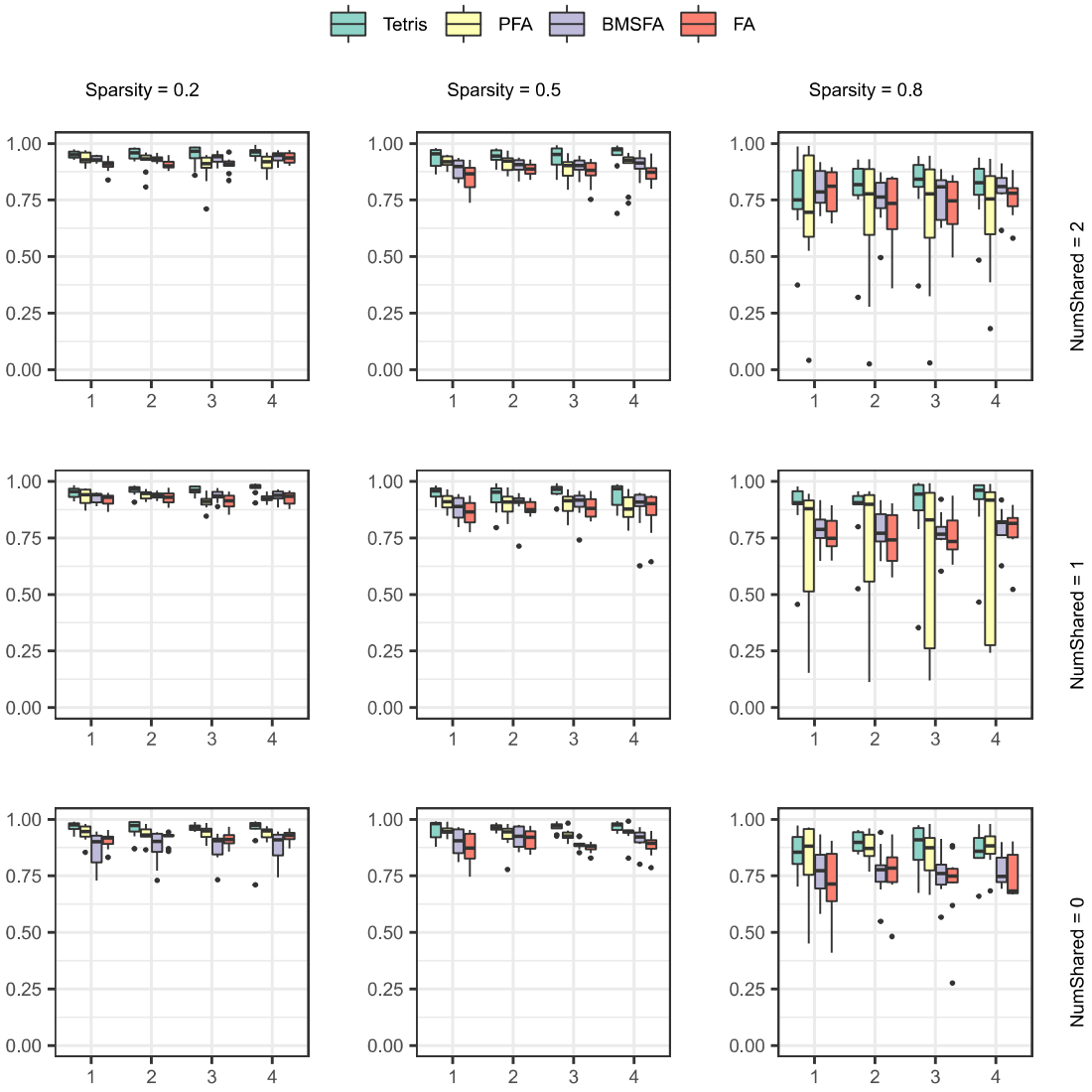

Figure 5: RV coefficients for study-specific covariances across varying sparsities and numbers of partially shared factors in the $p \ll n_s$ setting.

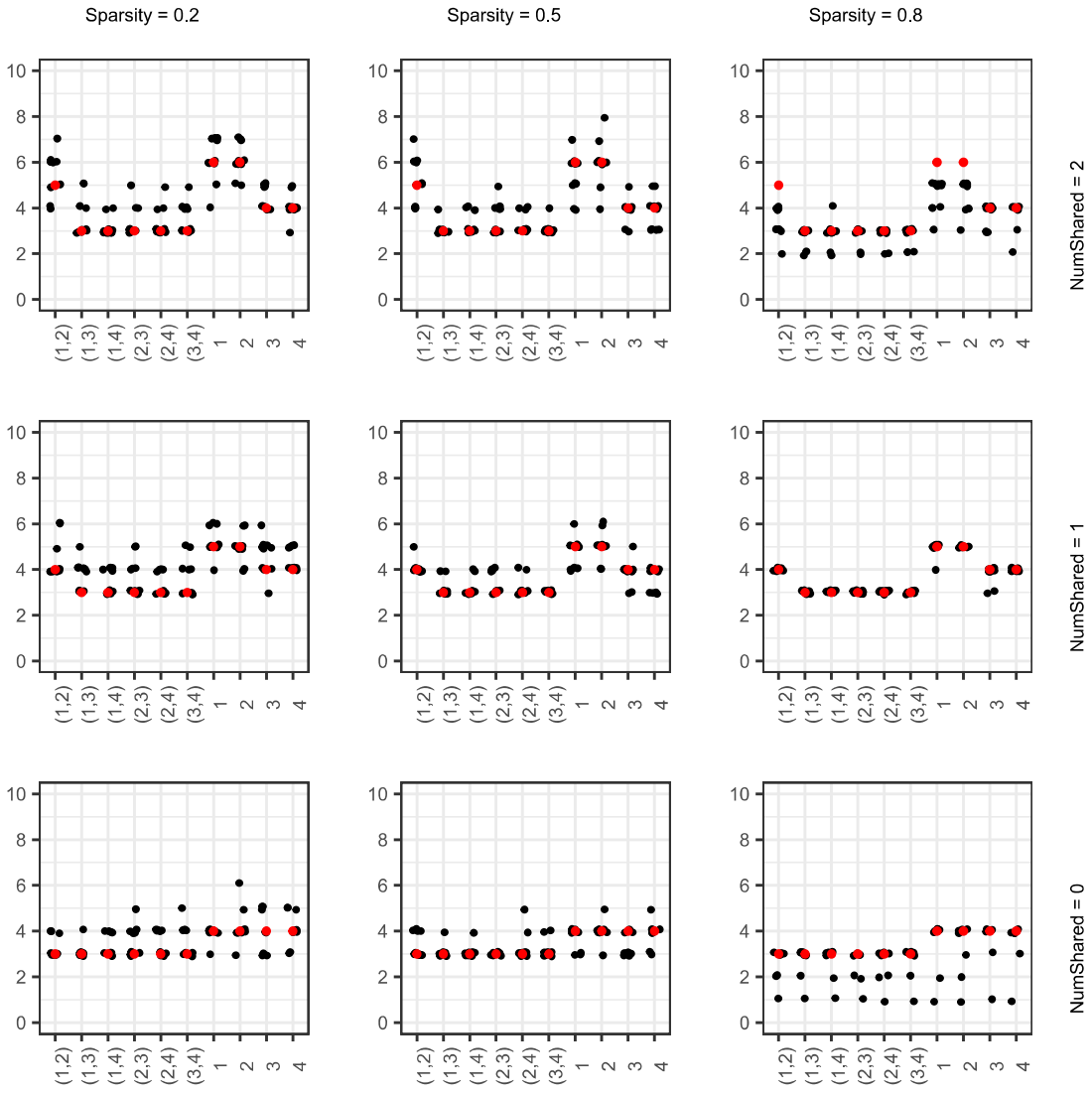

Figure 6: Number of factors shared by each pair of studies i and j, indicated by (i, j), and the number of total factors belonging to study i, indicated by i, for the $p = n_s$ simulations in Scenario 2. Estimated values are in black (with jitter, for visual clarity) and ground-truth values are in red.

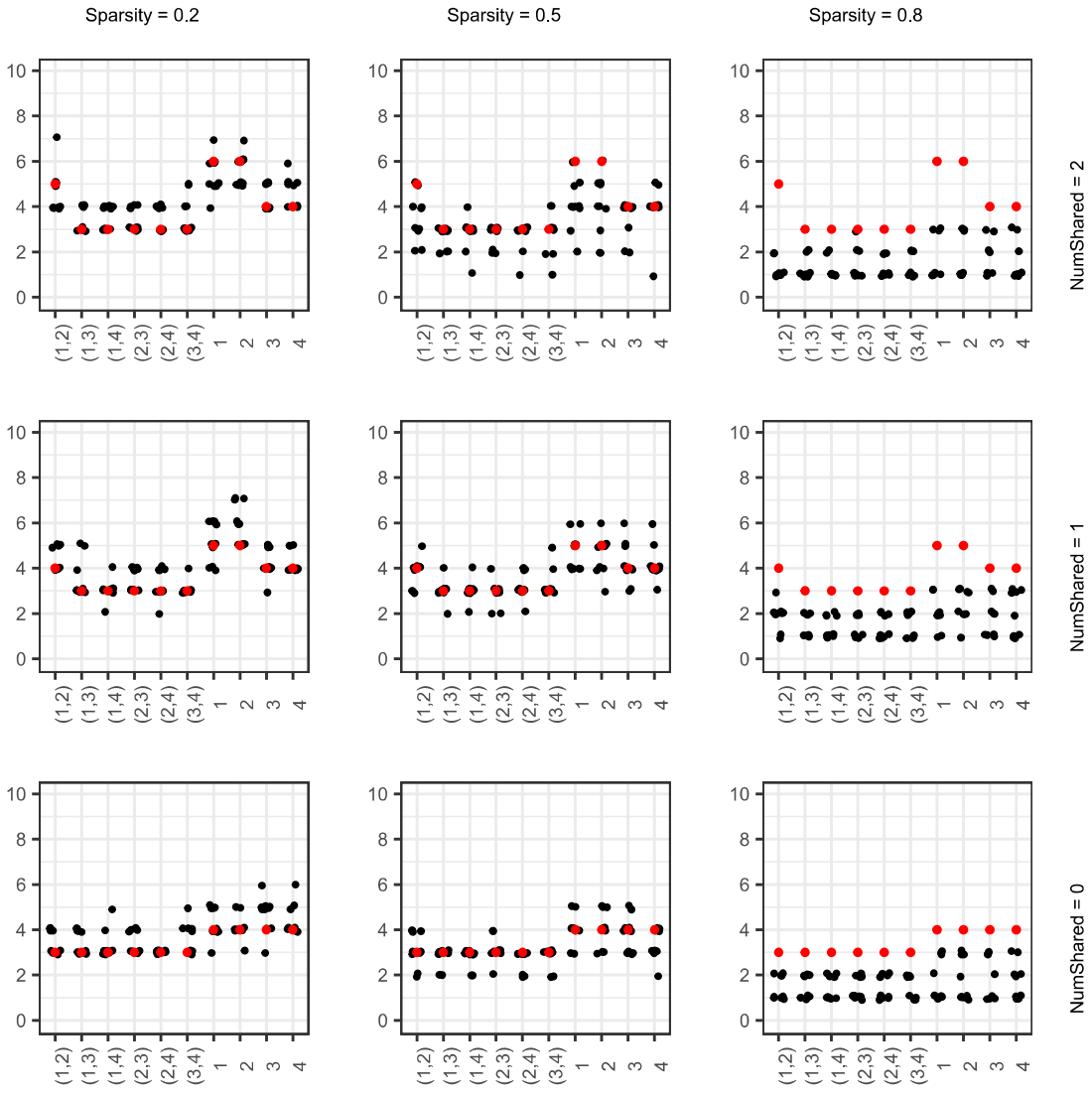

Figure 7: Number of factors shared by each pair of studies i and j, indicated by (i, j), and the number of total factors belonging to study i, indicated by i, for the $p \ll n_s$ simulations in Scenario 2. Estimated values are in black (with jitter, for visual clarity) and ground-truth values are in red.

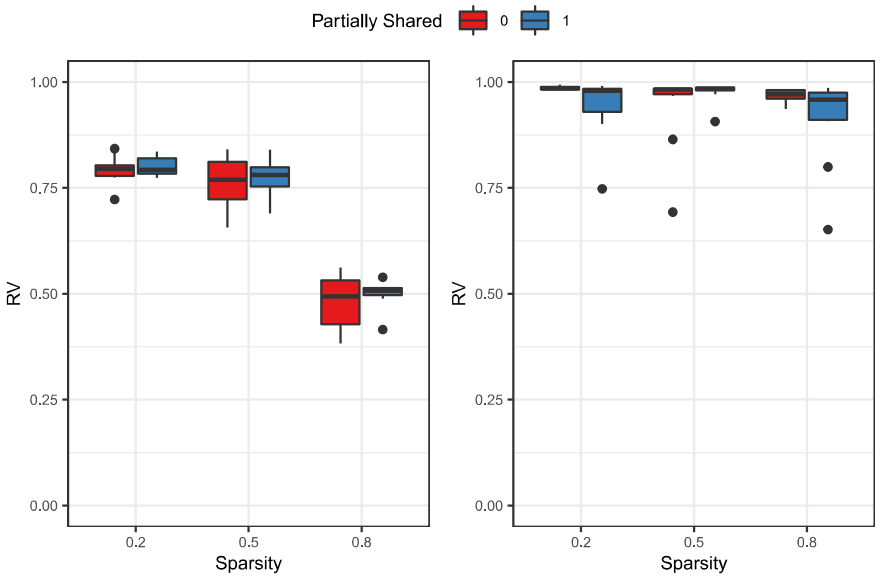

Figure 8: RV coefficients for the full (left) and common (right) loading covariances for the Scenario 3 simulation, which has 16 studies, across a range of sparsities and number of partially shared factors.

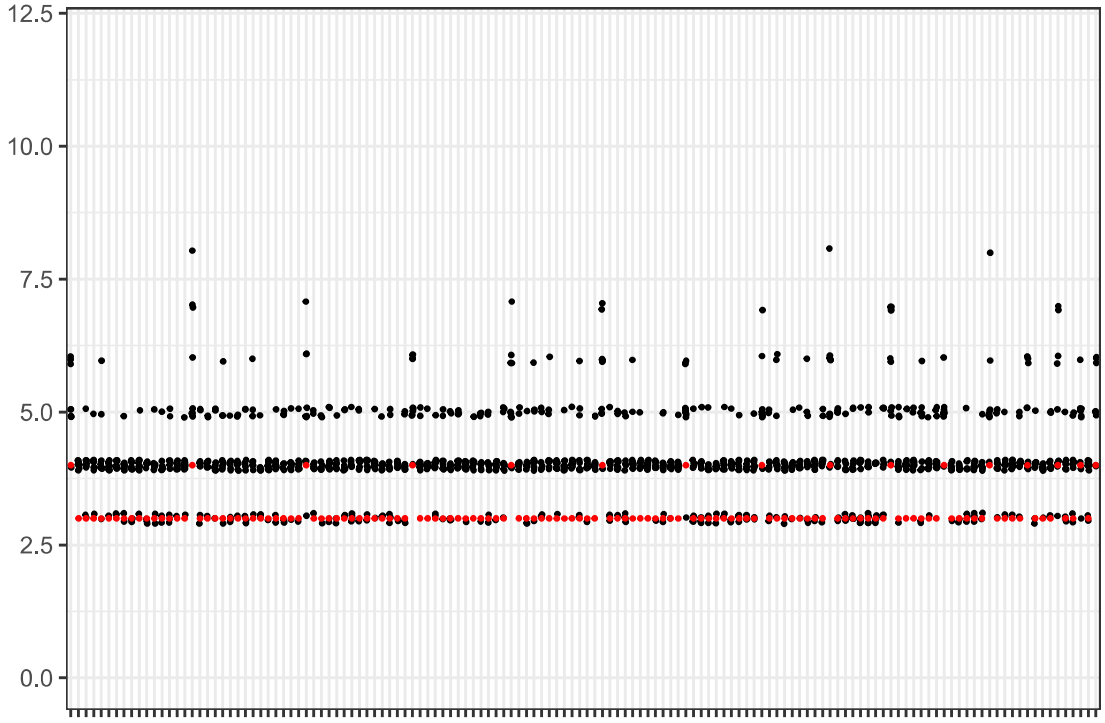

Figure 9: Number of factors shared by each pair of studies i and j, indicated by (i, j), and the number of total factors belonging to study i, indicated by i, for the Scenario 3 simulations with no shared factors and 80% sparsity. Estimated values are in black (with jitter, for visual clarity) and ground-truth values are in red.

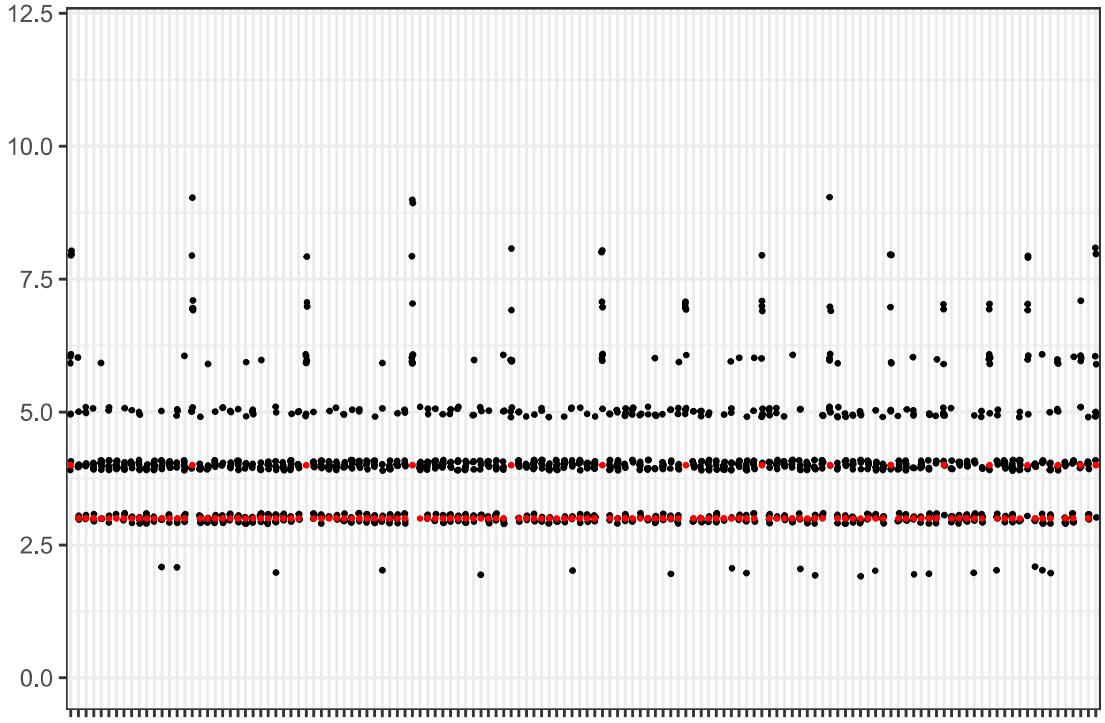

Figure 10: Number of factors shared by each pair of studies i and j, indicated by (i, j), and the number of total factors belonging to study i, indicated by i, for the Scenario 3 simulations with no shared factors and 50% sparsity. Estimated values are in black (with jitter, for visual clarity) and ground-truth values are in red.

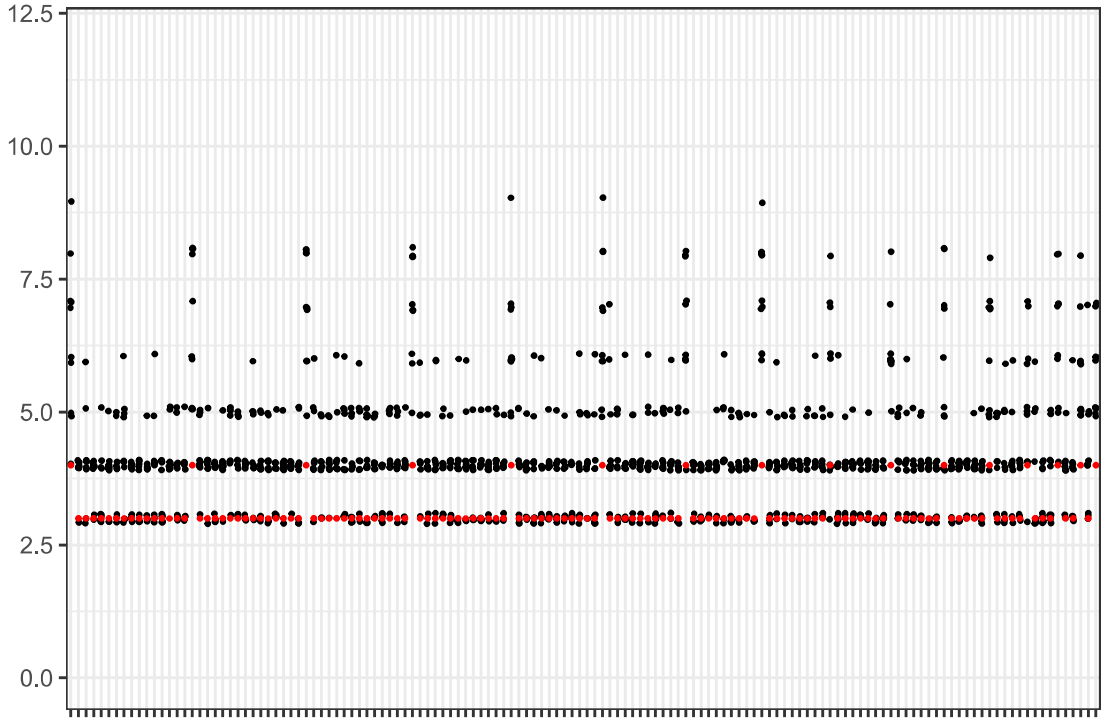

Figure 11: Number of factors shared by each pair of studies i and j, indicated by (i, j), and the number of total factors belonging to study i, indicated by i, for the Scenario 3 simulations with no shared factors and 20% sparsity. Estimated values are in black (with jitter, for visual clarity) and ground-truth values are in red.

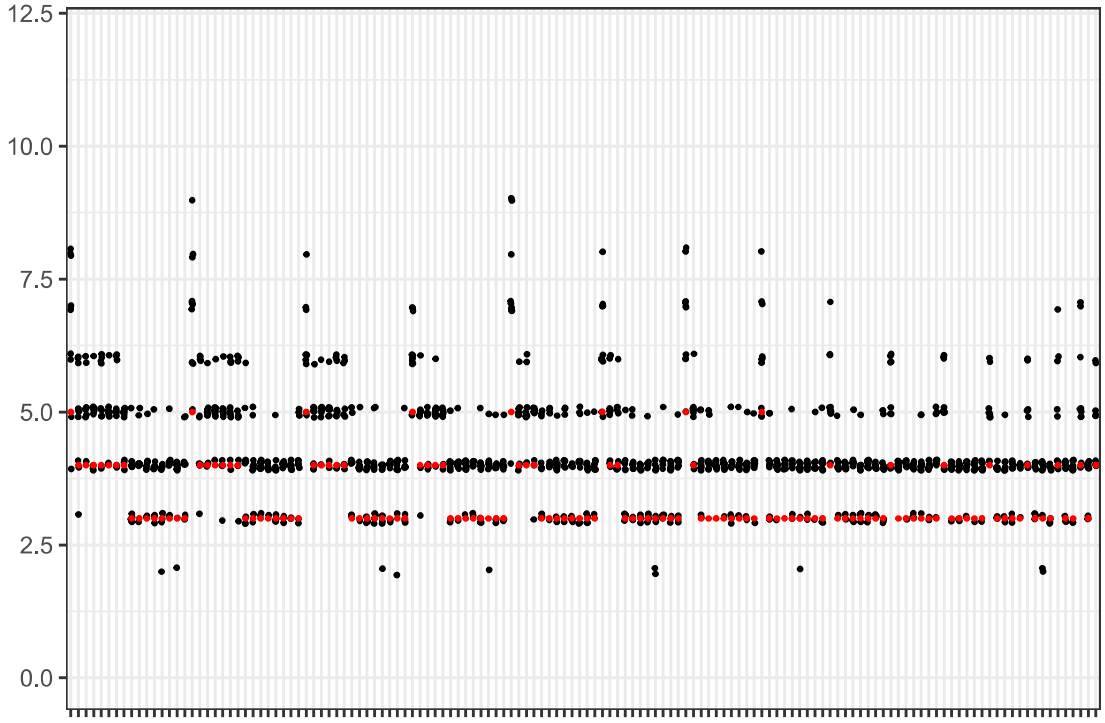

Figure 12: Number of factors shared by each pair of studies i and j, indicated by (i, j), and the number of total factors belonging to study i, indicated by i, for the Scenario 3 simulations with one shared factor and 80% sparsity. Estimated values are in black (with jitter, for visual clarity) and ground-truth values are in red.

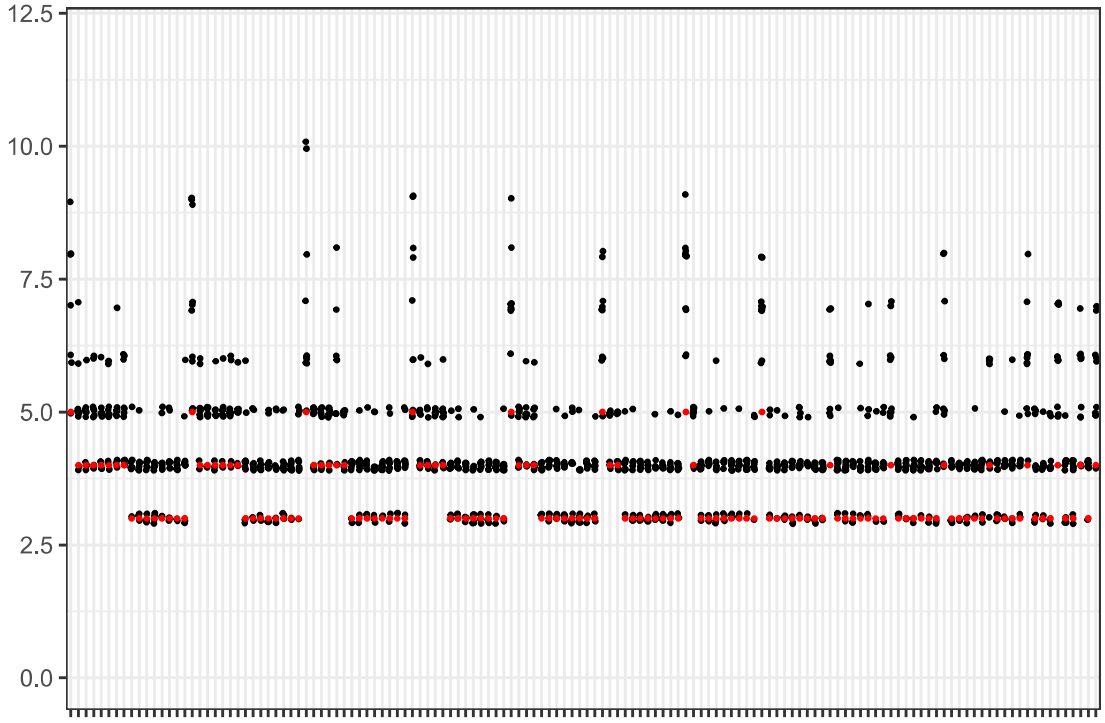

Figure 13: Number of factors shared by each pair of studies i and j, indicated by (i, j), and the number of total factors belonging to study i, indicated by i, for the Scenario 3 simulations with one shared factor and 50% sparsity. Estimated values are in black (with jitter, for visual clarity) and ground-truth values are in red.

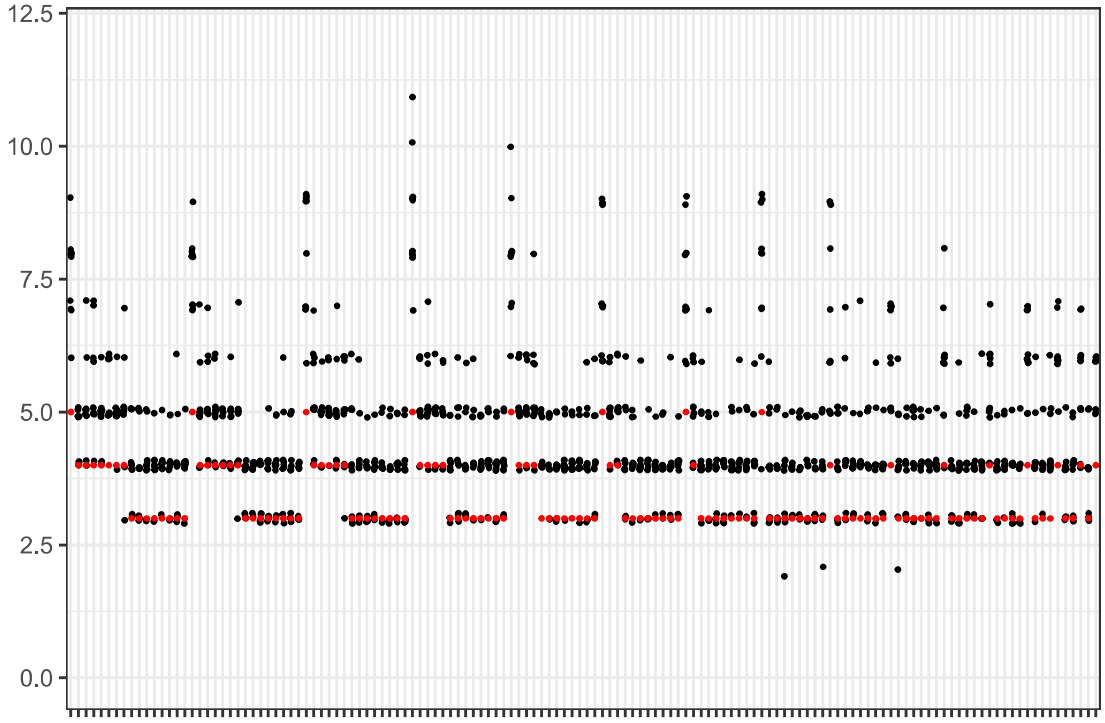

Figure 14: Number of factors shared by each pair of studies i and j, indicated by (i, j), and the number of total factors belonging to study i, indicated by i, for the Scenario 3 simulations with one shared factor and 20% sparsity. Estimated values are in black (with jitter, for visual clarity) and ground-truth values are in red.

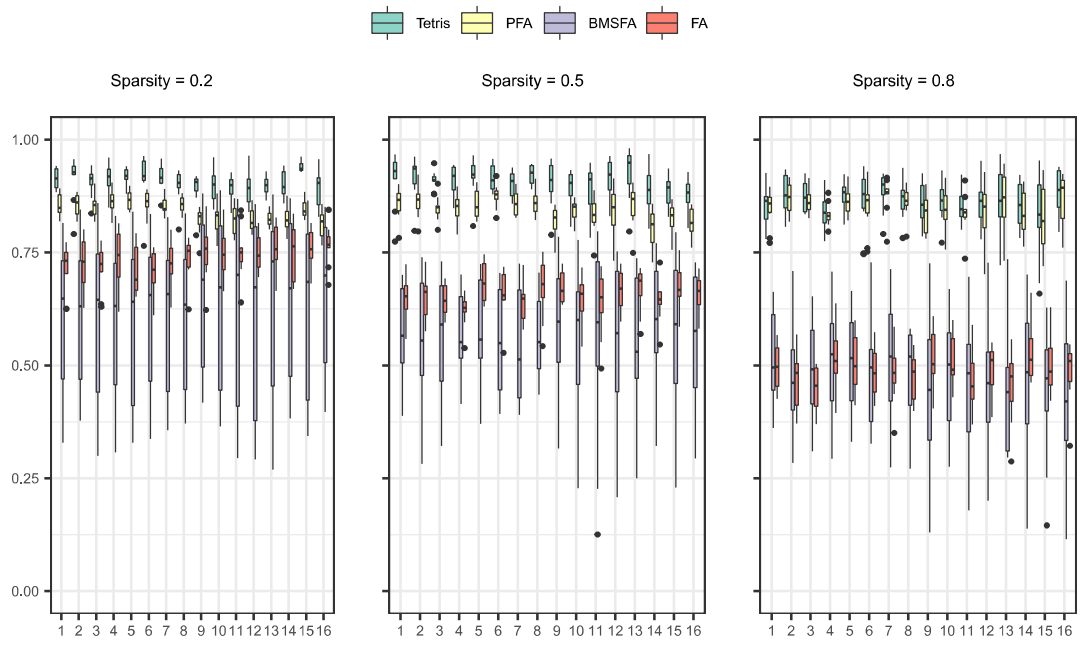

Figure 15: RV coefficients for study-specific covariances for the Scenario 3 simulation in the setting with one partially shared factor, across varying sparsities.

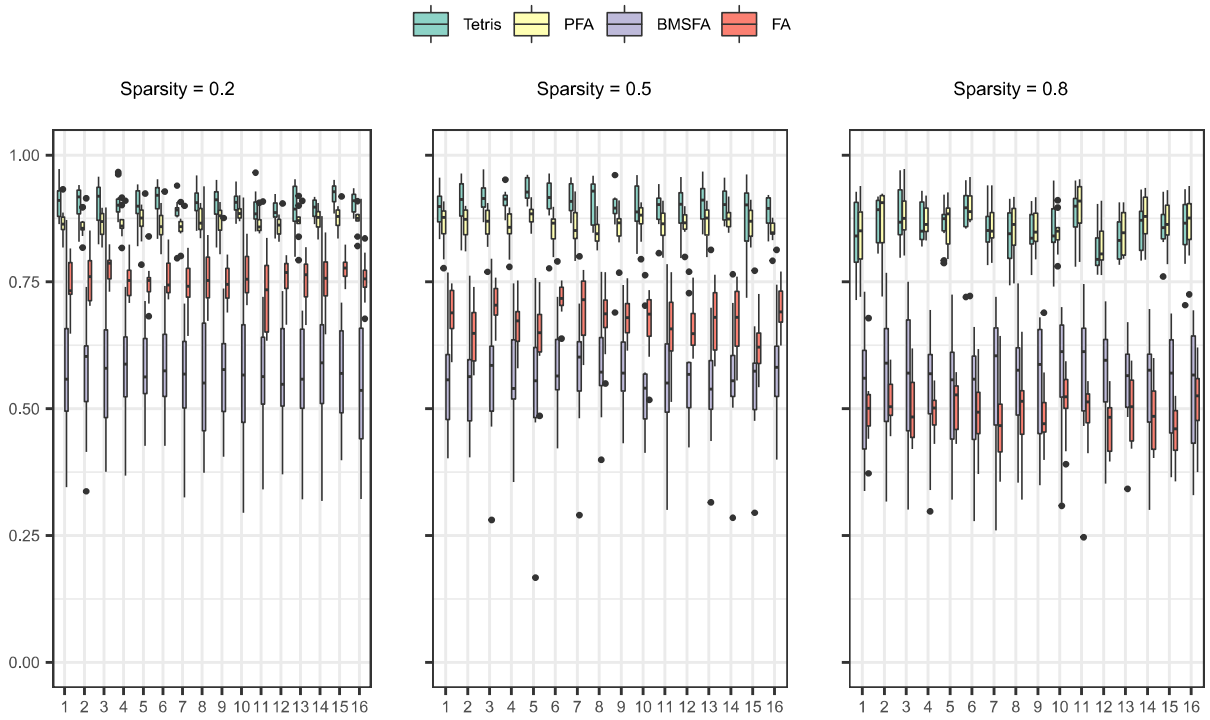

Figure 16: RV coefficients for study-specific covariances for the Scenario 3 simulation in the setting with no partially shared factors, across varying sparsities.

E Zero-Factor Cases

As described in the main text, there were five simulation runs resulting in Tetris identifying a solution with zero factors in most or all of the studies. We examine these runs here and analyze each with a likelihood ratio test to assess the extent to which these realizations of the simulated datasets could be plausibly explained with a noise-only model instead of the full factor model. Here, we consider the full factor model to be the actual data-generating distribution

$$\boldsymbol{X}_s \sim \mathcal{MVN}(\mathbf{0}, \boldsymbol{\Lambda}\boldsymbol{A}_s\boldsymbol{\Lambda} + \boldsymbol{\Psi}_s)$$

for each s, and the noise-only model is described by

$$\boldsymbol{X}_s \sim \mathcal{MVN}(\mathbf{0}, \boldsymbol{\Psi}_s^*),$$

where $\boldsymbol{\Psi}_s^*$ is the matrix consisting only of the diagonal elements of $\boldsymbol{\Lambda}\boldsymbol{A}_s\boldsymbol{\Lambda} + \boldsymbol{\Psi}_s$. The latter corresponds to the model implied by assuming that no columns of $\boldsymbol{\Lambda}$ contribute to study s, and all variance arises from a diagonal noise matrix. We perform the likelihood ratio test by computing the log-likelihood difference between the two models over all studies estimated to have zero factors, multiplying this difference by -2 to obtain the likelihood ratio test statistic, and computing a p-value by comparing this test statistic to a chi-squared distribution with $(\tilde{S} + P)K$ degrees of freedom where $\tilde{S}$ is the number of studies estimated to have zero factors.

It should be noted that we use the term "likelihood ratio test" loosely here. To truly be a likelihood ratio test, we should estimate the maximum likelihood parameters of both models and then compute the likelihood ratio when conditioning on these parameters. Instead, we are simply using the ground-truth parameters directly (or in the case of the noise-only model, a function of the ground-truth parameters). Hence, the chi-squared distribution is not technically valid as the asymptotic distribution of this test statistic. Nevertheless, we use this only as a simple measure to illustrate our point of how plausible it is for the data to be explained with a noise-only model compared to the data-generating, full-factor model.

Details of the five runs resulting in zero factors for most or all of the studies and their results in the above test are reported in Table 2. We find that the noise-only model may explain the realizations of the simulated data reasonably as well as the full factor model in the majority (three out of five) of the runs.

Number of Partially Shared Factors	Dimensionality	Sparsity	Log-Likelihood Ratio	p-value
0	$p = n_s$	0.80	-755	0.00
2	$p << n_s$	0.80	-26	1.00
2	$p << n_s$	0.80	-53	0.89
0	$p = n_s$	0.80	-804	0.00
1	$p << n_s$	0.80	-45	0.49

Table 2: Log-likelihood ratio between a noise-only model and the full factor model, and the corresponding p-value, for each simulation run that resulted in a zero-factor solution for most or all of the studies.

F Pathway Abbreviations

Reactome Pathway Name	**Abbreviated Name**
Nonsense Mediated Decay Enhanced by the Exon Junction Complex	Nonsense Mediated Decay
Influenza Viral RNA Transcription and Replication	Influenza Transcription and Replication
Influenza Life Cycle	Influenza Life Cycle
Metabolism of RNA	RNA Metabolism
Metabolism of mRNA	mRNA Metabolism
3 UTR Mediated Translational Regulation	3 UTR Mediated
Peptide Chain Elongation	Peptide Chain Elongation
SRP Dependent Cotranslational Protein Targeting to Membrane	SRP Dependent Protein Targeting
Translation	Translation
Cytokine Signaling in Immune System	Cytokine Signaling
Adaptive Immune System	Adaptive Immune System
Immune System	Immune System
Interferon Signaling	Interferon Signaling
Interferon Gamma Signaling	Interferon Gamma Signaling
Costimulation by the CD28 Family	CD28 Family Costimulation
PD1 Signaling	PD1 Signaling
Generation of Second Messenger Molecules	Second Messenger Molecules
Translocation of ZAP 70 to Immunological Synapse	ZAP 70 Translocation
Phosphorylation of CD3 and TCR Zeta Chains	CD3 and TCR Zeta Chains
Downstream TCR Signaling	Downstream TCR Signaling
TCR Signaling	TCR Signaling
MHC Class II Antigen Presentation	MHC Class II Antigen

Table 3: Pathway abbreviations used in main text.

G Additional Data Analysis Results

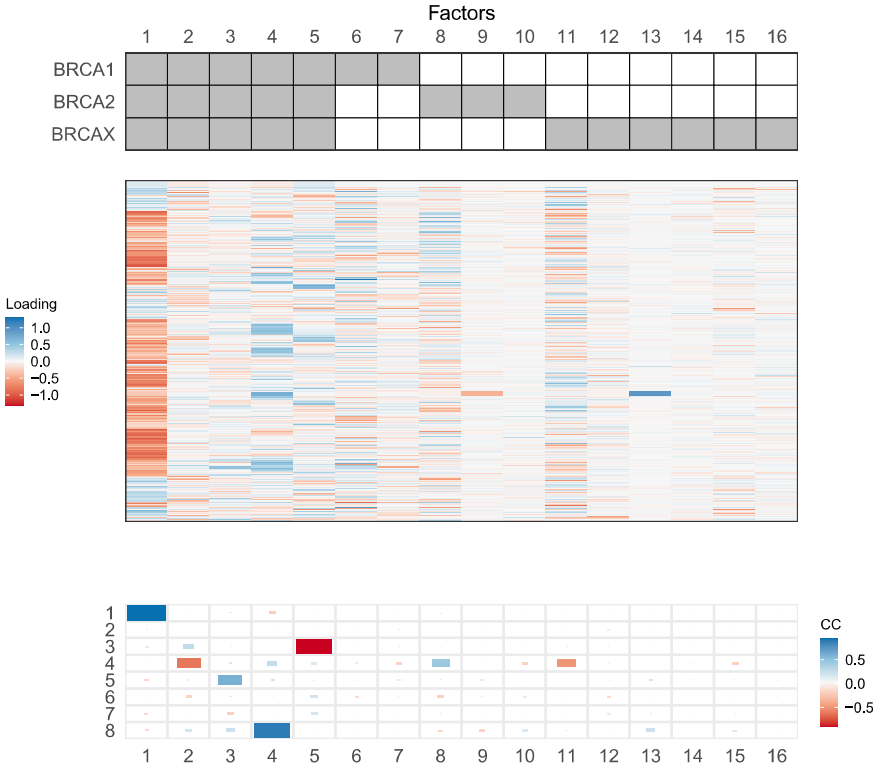

Figure 17: Visual summary of sharing pattern (top) and factor loadings (middle) for the analysis by genotype with BMSFA, and the congruence coefficients between each of BMSFA's factors and each of Tetris's factors (bottom).

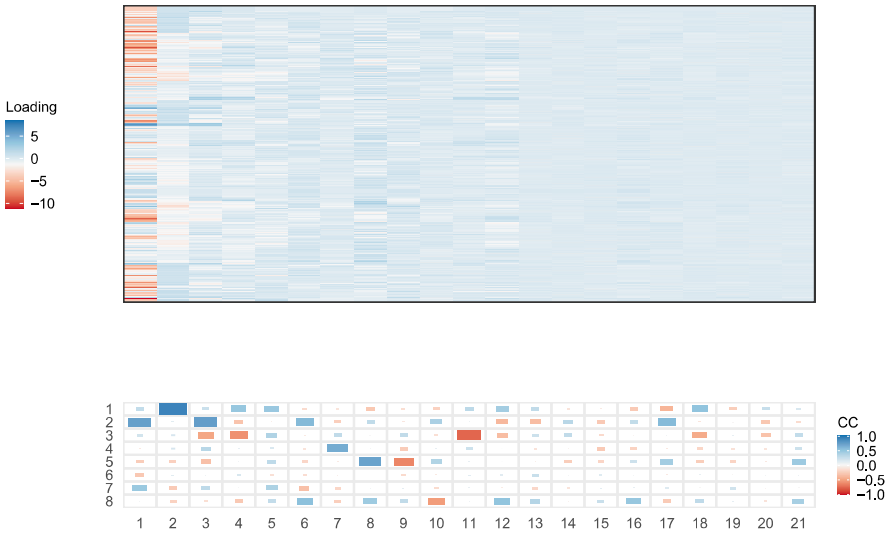

Figure 18: Factor loadings (top) for the analysis by genotype with PFA, and the congruence coefficients between each of PFA's factors and each of Tetris's factors (bottom).

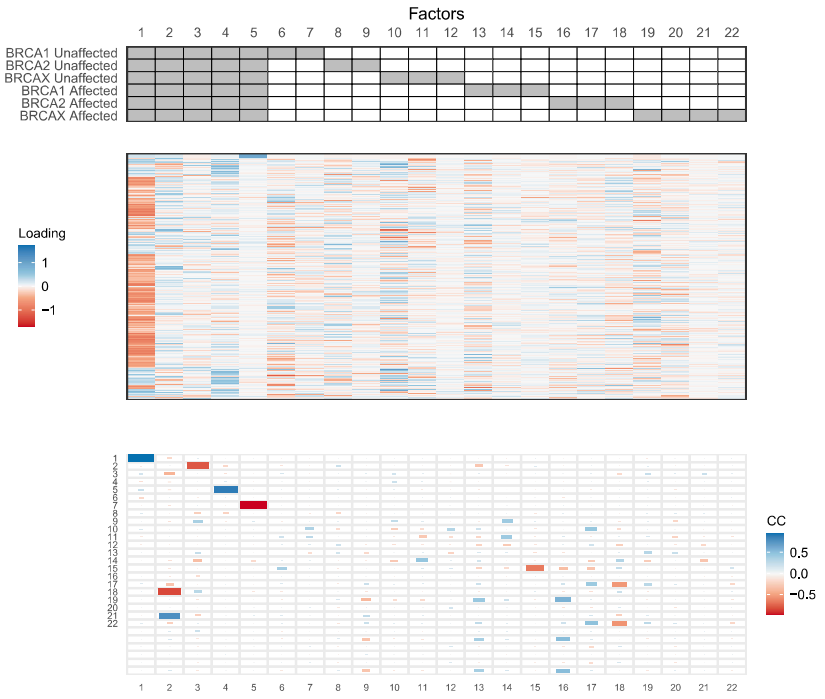

Figure 19: Visual summary of sharing pattern (top) and factor loadings (middle) for the analysis by genotype and affected status with BMSFA, and the congruence coefficients between each of BMSFA's factors and each of Tetris's factors (bottom).

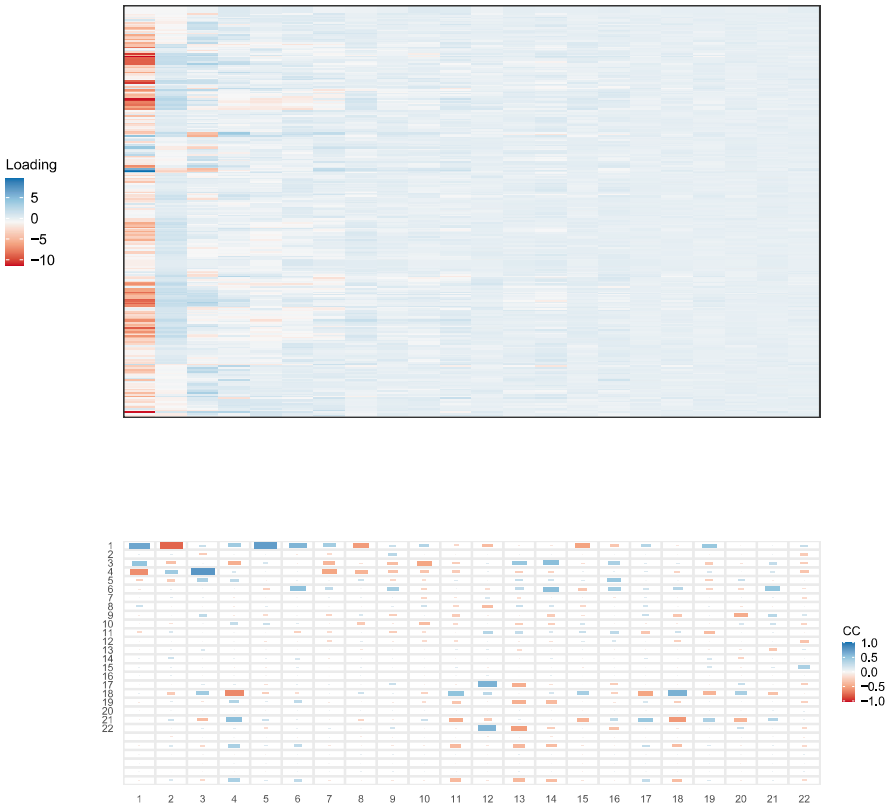

Figure 20: Visual summary of factor loadings (top) for the analysis by genotype and affected status with PFA, and the congruence coefficients between each of PFA's factors and each of Tetris's factors (bottom).

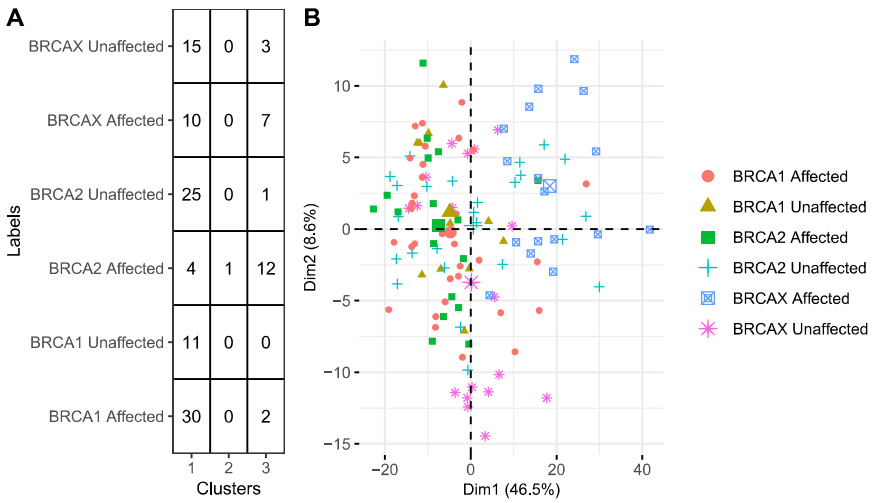

Figure 21: (A) Correspondence between the six-group labels and the modal clustering estimated by the extension of Tetris. (B) The first two PCs, colored by the six-group labels.

References

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

De Vito, R., Bellio, R., Trippa, L., and Parmigiani, G. (2021). Bayesian multistudy factor analysis for high-throughput biological data. *The Annals of Applied Statistics*, 15(4):1723–1741.

Doshi-Velez, F. et al. (2009). The Indian buffet process: Scalable inference and extensions. *Master's thesis, The University of Cambridge*.

Knowles, D. and Ghahramani, Z. (2007). Infinite sparse factor analysis and infinite independent components analysis. In *International Conference on Independent Component Analysis and Signal Separation*, pages 381–388. Springer.