Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

Subject Section

A convex optimization framework for gene-level tissue network estimation with missing data and its application in disease architecture

Kushal K. Dey [1,*], Rahul Mazumder [2, *]

[1]Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, MA and
[2]Sloan School of Management, Operations Research Center and Center for Statistics, MIT, Cambridge, MA.

* denotes authors to whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Genes with correlated expression across individuals in multiple tissues are potentially informative for systemic genetic activity spanning these tissues. In this context, the tissue-level gene expression data across multiple subjects from the Genotype Tissue Expression (GTEx) Project is a valuable analytical resource. Unfortunately, the GTEx data is fraught with missing entries owing to subjects often contributing only a subset of tissues. In such a scenario, standard techniques of correlation matrix estimation with or without data imputation do not perform well. To solve this problem, we propose `Robocov`, a novel convex optimization-based framework for robustly learning sparse covariance or inverse covariance matrices for missing data problems.

Results: `Robocov` produces more interpretable visual representation of correlation and causal structure in simulation settings and GTEx data analysis. We also show that `Robocov` estimators have a lower false positive rate than competing approaches for missing data problems. Genes prioritized by the average value of `Robocov` correlations or partial correlations across tissues are enriched for pathways related to systemic activities such as signaling pathways, circadian clock and immune function. SNPs linked to these prioritized genes showed high enrichment and unique information for blood-related traits; in comparison, no disease signal is observed for SNPs characterized analogously using standard correlation estimator.

Availability: `Robocov` is available as an R package `https://github.com/kkdey/Robocov`.

Contact: kdey@hsph.harvard.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The gene expression data from nearly 50 tissues across more than 500 post-mortem donor individuals from Genotype Tissue Expression (GTEx) project has proved to be a valuable resource for understanding tissue-specific and tissue-shared genetic architecture [1, 2, 3, 4]. Here we are interested in one specific aspect of tissue-shared gene regulation: the correlation and partial correlation in gene expression for different tissue pairs based on individual donor level data. A major challenge in this context is the extensive amount of missing entries in gene expression data—each donor contributes only a subset of tissues for sequencing. Common imputation based methods do not work well here as reported in [5], owing to stringent assumptions about missing entries being close to some central tendency (median) or adhering to some low-dimensional representation of the observed entries [6, 7]. Popular shrinkage and/or sparse correlation or partial correlation estimators such as *corpcor* [8, 9], GLASSO [10] or CLIME [11] are not designed for data with missing values.

A recently proposed approach, `CorShrink` [5], co-authored by one of the authors of this paper (Dey), accounts for missing data through adaptive shrinkage [12] of correlations. `CorShrink` does not guarantee a positive semidefinite (PSD) matrix as part of its EM-based framework, and necessitates a post-hoc modification to ensure a PSD correlation matrix. Furthermore, `CorShrink` does not extend to conditional graph or partial correlation estimation. Here, we propose a new approach based on convex optimization: `Robocov` that applies to both covariance and

inverse covariance matrix estimation in the presence of missing data under the following regularization principles: (a) the covariance matrix is sparse (i.e., has a few nonzero entries) or (b) the inverse covariance matrix is sparse.

`Robocov` does not *impute* missing values per-se[1]—it directly estimates the covariance or inverse covariance matrices in the presence of missing values. To handle missing values, we consider a loss function that depends upon the pairwise covariance terms (computed based on the observed samples) but incorporates an adjustment to guard against our lack of knowledge regarding the missing observations. For inverse covariance estimation, `Robocov` uses a robust optimization based approach [14, 15] that accounts for the uncertainty in estimating the pairwise sample covariance terms (due to the presence of missing values). Interestingly, both lead to convex optimization formulations that are amenable to modern optimization techniques [16]—they are scalable to moderate-large scale instances; and unlike conventional EM methods (that lead to nonconvex optimization tasks), our estimators attain the global solution of the optimization formulations defining the `Robocov` estimators.

Our experiments suggest that `Robocov` estimators for correlation and partial correlation matrices have lower false positive rate compared to competing approaches for missing data problems. When applied to the GTEx gene expression data with $\sim 70\%$ missing data, `Robocov` produced less cluttered and highly interpretable visualization of correlation and conditional graph architecture. From a biological perspective, a gene with high correlation in expression across many tissue pairs is potentially reflective of more systemic biological processes spanning multiple tissues. To this end, we prioritized genes based on the average `Robocov` estimated correlation (partial correlation) across all tissue-pairs; we call them `Robospan` (`pRobospan`) genes. A pathway enrichment analysis of `Robospan` (`pRobospan`) genes showed enrichment in systemic functional pathways and the immune system. SNPs linked to `Robospan` (`pRobospan`) genes were tested for autoimmune disease informativeness by applying Stratified LD-score regression (S-LDSC) to 11 common blood-related traits (5 autoimmune diseases and 6 blood cell traits; average N=306K), conditional on a broad set of annotations. `Robospan` and `pRobospan` genes showed high disease informativeness for blood-related traits. In comparison, `Corspan` genes defined similarly using the standard correlation estimator were non-informative. This highlights the biological and disease-level significance of our work.

2 Methods

Let $X_{N\times P}$ be a data matrix with N samples and P features, where some of the entries X_{np} may be missing, denoted here by NA. Let X^f denote the fully-observed version of the partially-observed data matrix[2] X. We assume that samples are independent and follow a Multivariate Normal distribution: i.e., $X^f_{n,*} \sim \text{MVN}(0,\Sigma)$ where $\Sigma_{P\times P}$ and $\Omega := \Sigma^{-1}$ (also of size $P\times P$) denote the model covariance and inverse covariance matrices respectively. Based on the observed entries, we obtain a matrix $\hat{\Sigma}$ of pairwise covariances such that for all $i,j \in \{1,\ldots,P\}$:

$$\hat{\Sigma}_{ij} := \frac{1}{n_{ij}-1} \sum_{n: X_{ni}\neq \text{NA}, X_{nj}\neq \text{NA}} (X_{ni}-\bar{X}_i)(X_{nj}-\bar{X}_j) \quad (1)$$

where, $\bar{X}_k$ denotes the sample mean of feature k based on the observed entries; and n_{ij} is the number of samples n with non-missing entries in both features i and j. Let n_i denote the number of observed samples (i.e., not missing) for feature i. For our analysis, we will assume[3] that $n_{ij} > 2$ for all i,j. We note that the matrix of all pairwise covariance terms: $\hat{\Sigma} = ((\hat{\Sigma}_{ij}))$, as defined in (1), need not be positive semidefinite due to the presence of missing values in the data matrix.

2.1 Robocov covariance estimator

We first present the `Robocov` covariance matrix estimator—this leads to an estimate of Σ via the following regularized criterion:

$$\min \sum_{i<j} |\Sigma_{ij}| \quad \text{s.t.} \quad \Sigma \succeq 0,\ |\hat{\Sigma}_{ij} - \Sigma_{ij}| \le C_{ij},\ \forall i,j \quad (2)$$

where Σ is the optimization variable and C_{ij}s are data-driven constants that control the amount by which Σ_{ij} can differ from the sample version $\hat{\Sigma}_{ij}$. Problem (2) minimizes a convex penalty function (this encourages sparsity [17] in Σ_{ij}s) subject to convex constraints (note that Σ is positive-semidefinite i.e., $\Sigma \succeq 0$). Problem (2) is a convex semidefinite optimization problem [16]; and can be solved efficiently by modern semidefinite optimization algorithms for moderately large instances (e.g, $P \sim 1000$) using (for example) the SCS solver in CVX software [16, 18, 19, 20].

We compute C_{ij} based on the Fisher's Z-Score [21, 22] (for a complete derivation see Supplementary Note):

$$C_{ij} = \hat{\sigma}_i\hat{\sigma}_j \min\left(2, \eta(n_{ij})\left\{3(1-\hat{R}^2_{ij}) + 2\sqrt{3}\eta(n_{ij})\right\}\right) \quad (3)$$

where $\eta(n_{ij}) = \sqrt{1/(n_{ij}-1) + 2/(n_{ij}-1)^2}$; and $\hat{R}$ is the pairwise sample correlation matrix derived from $\hat{\Sigma}$

In summary, we note that our proposed `Robocov` estimator does not impute missing values per-se — it directly leads to an estimate for the covariance matrix Σ while taking into account the presence of missing-values in the data matrix.

While (2) leads to a covariance matrix estimator, this can be modified to deliver a correlation matrix instead of a covariance matrix:

$$\begin{aligned} \min \quad & \sum_{i<j} |\mathcal{R}_{ij}| \\ \text{s.t.} \quad & \mathcal{R} \succeq 0, \mathcal{R}_{ii} = 1, \forall i, \quad |\hat{R}_{ij} - \mathcal{R}_{ij}| \le C^{(R)}_{ij},\ \forall i,j \end{aligned} \quad (4)$$

where $\mathcal{R}$ is the optimization variable and $C^{(R)}_{ij} = \frac{\hat{C}_{ij}}{\hat{\sigma}_i\hat{\sigma}_j}$. See the Supplementary Note for additional details.

In the Supplementary Note, we present a framework given by the minimization of a regularized loss function that generalizes the estimator presented in (2).

2.2 Robocov inverse covariance estimator

We present a regularized likelihood framework to estimate the inverse covariance matrix (Ω) under a sparsity constraint. Our optimization criterion is convex in Ω (and not Σ which was the case in Section 2.1).

Recall that GLASSO minimizes an ℓ_1-norm regularized negative log-likelihood criterion (fully observed case); and is given by:

$$\min_{\Omega\succ 0} \quad -\log\det(\Omega) + \langle\tilde{\Sigma},\Omega\rangle + \lambda\sum_{ij}|\Omega_{ij}| \quad (5)$$

where, $L(\Omega;\tilde{\Sigma}) := -\log\det(\Omega) + \langle\tilde{\Sigma},\Omega\rangle$ is the negative log-likelihood (ignoring irrelevant constants), $\tilde{\Sigma}$ is the fully observed sample covariance

[1] Expectation Maximization (EM) [13] methods typically used for estimation with missing values depend upon probabilistic modeling assumptions and lead to highly nonconvex problems posing computational challenges.

[2] Note that X is a restriction of X^f to the observed entries.

[3] If necessary, as a pre-processing step, we remove features so that the condition $n_{ij} > 2$ is satisfied for all i,j.

matrix and $\lambda \geq 0$ is the regularization parameter. Replacing $\tilde{\Sigma}$ by the observed matrix $\hat{\Sigma}$ in (5) is problematic due to the error in estimating the pairwise covariances arising from the missing values (different cell entries of the sample covariance matrix involve different effective sample sizes n_{ij}s leading to varying accuracies in estimating $\tilde{\Sigma}_{ij}$s). To account for this uncertainty, we use ideas from robust optimization [14, 15]—to the best of our knowledge, this approach has not been used earlier in the context of sparse inverse covariance estimation (in the presence of missing values). Our robust optimization approach minimizes the worst-case loss arising from the errors in estimating the cell entries $\tilde{\Sigma}_{ij}$s. This leads to a min-max optimization problem of the form:

$$\min_{\Omega \succeq 0} \max_{\substack{\Delta \\ |\Delta_{ij}| \leq D_{ij}, \forall i,j}} \left\{ -\log\det(\Omega) + \langle \Omega, \hat{\Sigma} + \Delta \rangle \right\} + \lambda \sum_{ij} |\Omega_{ij}|. \tag{6}$$

which is convex [16] in Ω (See Supplementary Note). Convexity ensures that a global minimum to the problem can be obtained reliably—making our approach different from traditional missing data techniques based on the EM algorithm [13] that often lead to complex nonconvex optimization tasks with multiple local solutions.

In words, the inner maximization over Δ in Problem (6) gives the largest (or worst-case) value of the negative log-likelihood—$\max_{\Delta} L(\Omega; \hat{\Sigma} + \Delta)$ where, Δ captures the uncertainty involved in estimating the entries of the sample covariance matrix $\tilde{\Sigma}$ due to the presence of missing values. The outer minimization problem (wrt Ω) considers the minimum of the *adjusted* loss function in addition to an ℓ_1-penalization on Ω that encourages a sparse estimate of Ω.

The so-called uncertainty set [15] in Δ is given by: $|\Delta_{ij}| \leq D_{ij}$ (for all i, j) where, the upper bound D_{ij} arises from a probability computation using the Fisher's Z-score criterion (see Supplementary Note):

$$\begin{aligned} D_{ij} &= C_{ij} + \tilde{C}_{ij} \\ \tilde{C}_{ij} &= \hat{\sigma}_i \hat{\sigma}_j \min\left\{2, \eta(N)\left\{3(1-\hat{R}_{ij}^2) + 2\sqrt{3}\eta(N)\right\}\right\}. \end{aligned} \tag{7}$$

Above, the value of the error D_{ij} will be large if n_{ij} is small, and equal to zero when $n_{ij} = n$ (with no missing entries).

The seemingly complicated min-max optimization problem in (6) reduces to a cousin of the GLASSO criterion (See Supplementary Note for details) — we use a weighted version of the ℓ_1-norm penalty on Ω:

$$\min_{\Omega \succeq 0} \left\{ -\log\det(\Omega) + \langle \Omega, \hat{\Sigma} \rangle + \sum_{ij} (\lambda + D_{ij}) |\Omega_{ij}| \right\}. \tag{8}$$

Problem (8) is a nonlinear semidefinite optimization problem in Ω—and the constraint $\Omega \succeq 0$ leads to a positive semidefinite inverse covariance matrix[4]. Problem (8) uses a weighted ℓ_1-norm on Ω where the penalty weights are adjusted to account for the uncertainty due to the presence of missing values. Note that the penalty parameter λ accounts for the sparsity in Ω arising from our prior sparsity assumption on Ω—the overall penalty weight for the (i, j)-th entry: $(\lambda + D_{ij})$ adds further regularization due to the presence of missing values.

Note that, as in Section 2.1, the `Robocov` inverse covariance estimator, bypasses the task of imputing the missing values. Our main goal is to directly estimate Ω from a partially observed data-matrix X. In this way, we can potentially mitigate the limitations of a sub-optimal imputation procedure. See Section 3 for an empirical validation.

The inverse covariance estimate Ω from (8), can be used to obtain the partial correlation estimator $\mathcal{W}$ as follows

$$\mathcal{W}_{ij} := -\frac{\Omega_{ij}}{\sqrt{\Omega_{ii}\Omega_{jj}}}. \tag{9}$$

Problem (8) was solved using R implementation of the CVX software [19, 20]. This was sufficient for the problem-scales we are dealing with.

In all our subsequent analysis and numerical results, we use the `Robocov` correlation estimator $\mathcal{R}$ (see Problem (4)) and partial correlation estimator $\mathcal{W}$ (9).

3 Results

Simulation Experiments: Synthetic and Real Data

We applied `Robocov` on simulated multivariate normal data from three population correlation structure models (hub, Toeplitz and 1-band precision matrix) with N samples, P features and π proportion of missing entries randomly distributed throughout the data matrix (Supplementary Note). Figure 1 shows results for all three model-settings with $N = 500, P = 50, \pi = 0.5$. In all cases, `Robocov` generated a sparse estimate of the population correlation $\mathcal{R}$ (Section 2.1) or partial correlation $\mathcal{W}$ (Section 2.2). The `Robocov` correlation estimator captured population structure more effectively for all three models compared to the standard pairwise sample correlation estimator (Figure S1). The `Robocov` partial correlation estimator also accurately captured the causal structure in the hub and 1-band precision matrix models; for the Toeplitz matrix, it recovered the high partial correlation band immediately flanking the diagonal but not the other alternating positive and negative low correlation bands (Figure 1).

Recent work [5] has shown hub-like patterns in expression correlation across tissue pairs for most genes. To this end, we applied `Robocov` on simulated data for hub population correlation matrix structure for different settings of N, P and π (Supplementary Note). Two metrics of particular interest were the false positive rate (FPR) and the false negative rate (FNR) (Supplementary Note). We used these metrics to compare `Robocov` correlation estimator with both the pairwise sample correlation estimator and `CorShrink`[5]. Across different (N, P, π)-settings, the `Robocov` correlation estimator had lower FPR than `CorShrink`. In comparison, for data with a large number of missing entries (i.e., high π), FNR for `Robocov` was worse compared to `CorShrink` (Table 1). We did not compare against other shrinkage-based correlation estimators such as *PDSCE*[23] and *corpcor*[24, 9] as (i) they do not account for missing entries in the data and have been shown to be sub-optimal to `CorShrink` for fully observed data (see Figure 4 from ref.[5]).

Next, we assess the performance of the `Robocov` partial correlation estimator for the same simulation settings (Table 1). We are not aware of a sparse conditional graph or partial correlation estimation method that directly takes into account missing entries. Nevertheless, we compare the `Robocov` partial correlation estimator with (i) GLASSO on the pairwise sample correlation estimator $\hat{\Sigma}$ and (ii) CLIME on an imputed data matrix where, the imputation is performed using SoftImpute [7]. In the presence of missing data, `Robocov` partial correlation estimator showed better FPR and FNR compared to both GLASSO and CLIME-based estimators (Table 1). The underperformance of CLIME may be attributed to the error arising from the imputation step (Table 1).

Next, we evaluate the predictive performance of `Robocov` correlation estimator with pairwise sample correlation estimator and `CorShrink`. We considered the GTEx gene expression data for an example gene (ARHGAP30) across 544 donors and 53 tissues with close to 70% missing data owing to subjects contributing only a small fraction of tissues. We split the individual by tissue data for the gene into two equal groups and compared the estimated correlation matrix (we used different estimators: `Robocov`, `CorShrink` and pairwise sample correlation matrix) computed on one half of the individuals with the pairwise sample

[4] We get a positive semidefinite (PSD) estimate for Ω even if $\hat{\Sigma}$ is not PSD. The $\log\det$-term in the objective encourages an optimal solution to (8) to be positive definite (i.e, of full rank).

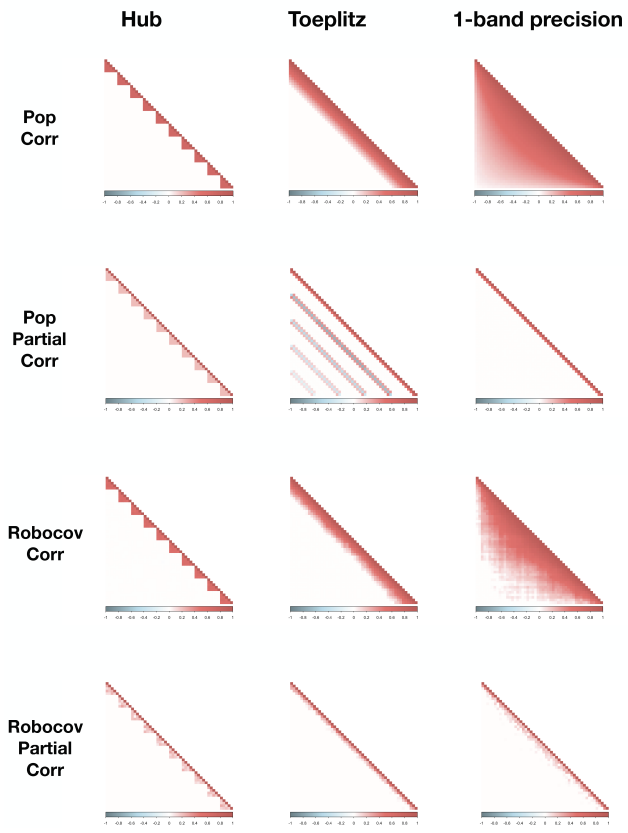

Fig. 1. We applied `Robocov` correlation and partial correlation estimators on data generated from Hub, Toeplitz or 1-band precision matrix based population models (Supplementary Note) with $N = 500$ samples, $P = 50$ features and $\pi = 0.5$ proportion of missing data. We present the population correlation matrix, population partial correlation matrix, `Robocov` correlation matrix and `Robocov` partial correlation matrix sequentially from first to last row.

Table 1. We compare three metrics: FP2 (False Positive 2-norm), FPR (False Positive Rate) and FNR (False Negative Rate) (Supplementary Note) to compare (i) the `Robocov` correlation estimator (Cor) against `CorShrink` and the standard pairwise sample correlation estimator; and (ii) the `Robocov` partial correlation estimator (PCor) against estimators available from GLASSO and CLIME. Data was generated for different (N, P, π) settings and results were averaged over 50 replications from same model. Optimal λ was chosen by cross-validation.

Hub: N = 50, P=50										
		π=0			π=0.25			π=0.5		
Type	Method	FP2	FPR	FNR	FP2	FPR	FNR	FP2	FPR	FNR
Cor	Robocov	0.05	0	0	0.14	0	0.14	0.26	0	0.19
	CorShrink	1.4	0.01	0	2.2	0.04	0.03	4	0.07	0.09
	Standard	6.7	0.24	0	8.8	0.30	0	15	0.28	0
PCor	Robocov	0.08	0	0.07	0.27	0.01	0.13	0.47	0	0.09
	GLASSO	0.12	0	0.15	0.29	0.01	0.15	0.59	0.02	0.12
	CLIME	1.5	0.09	0.07	1.4	0.07	0.08	1.3	0.08	0.07
Hub: N = 100, P=50										
		π=0			π=0.25			π=0.5		
Type	Method	FP2	FPR	FNR	FP2	FPR	FNR	FP2	FPR	FNR
Cor	Robocov	0.05	0	0	0.06	0	0	0.18	0	0.15
	CorShrink	0.9	0	0	1.3	0.02	0	2.9	0.03	0.01
	Standard	4.8	0.17	0	6.2	0.20	0	10	0.31	0
PCor	Robocov	0.23	0	0.06	0.21	0	0.09	0.18	0.03	0.11
	GLASSO	0.11	0	0.16	0.23	0	0.22	0.29	0.01	0.24
	CLIME	1.8	0.12	0.08	1.8	0.14	0.09	1.8	0.16	0.11
Hub: N = 500, P=50										
		π=0			π=0.25			π=0.5		
Type	Method	FP2	FPR	FNR	FP2	FPR	FNR	FP2	FPR	FNR
Cor	Robocov	0.03	0	0	0.01	0	0	0.08	0	0
	CorShrink	0.21	0	0	0.32	0	0	0.83	0	0
	Standard	2.1	0.01	0	2.8	0.05	0	4.4	0.14	0
PCor	Robocov	0.12	0	0.11	0.16	0	0.12	0.11	0	0.14
	GLASSO	0.16	0	0.19	0.29	0	0.20	0.19	0.02	0.20
	CLIME	2.1	0.11	0.16	2.0	0.14	0.18	2.0	0.15	0.17

correlation matrix computed from the other half. Both `Robocov` and `CorShrink` estimators considerably outperformed the pairwise sample correlation estimator, with `CorShrink` having slightly better predictive accuracy (Figure S2 and Table S1). As `Robocov` and `CorShrink` predictive performances are similar, the former may be preferable as it results in sparse estimates, leading to better interpretability.

An an alternative to `Robocov`, we may consider an estimator obtained by first imputing the missing entries in the data matrix and then estimating the correlation or partial correlation matrix for the complete data. For the same ARHGAP30 gene, we performed imputation by either a low rank factorization (SoftImpute[7], with or without scaling) or a median based approach (replacing the missing entries of a feature by the median value of the observed entries). The correlation matrix obtained by SoftImpute (both with and without scaling) showed artificial high negative and positive correlation sweeps between brain and non-brain tissues that were not observed in the pairwise correlation matrix (Figure S3). One possible explanation of this is that the data matrices in our case do not seem to have a low rank representation based on eigenvalue analysis (Figure S4). The median based imputation method on the other hand, is prone to showing false positives—for example, we see a high correlation between Fallopian tube and Cervix-Ectocervix, which is a consequence of only 3 individuals contributing both the tissues (Figure S3). `Robocov` can effectively get rid of these edge cases and generate sparser and more robust results compared to these imputation based approaches.

Gene Expression correlation analysis across tissue pairs

We applied `Robocov` to each of 16,069 cis-genes (genes with at least one significant cis-eQTL) from the GTEx v6 project [3] (see URLs). For each gene, the data matrix had 544 rows (post-mortem donors), 53 columns (tissues) and comprised of $\sim$ 70% missing entries. Figure 2 presents a visual comparison of `Robocov` correlation and partial correlation estimators with standard pairwise sample correlation matrix for two example genes (ARHGAP30 and GSTM1)—the `Robocov` estimators are sparse and visually less cluttered than the standard approach. The `Robocov` correlation structure across tissue pairs varied from one gene to another: some genes showed high correlation across all tissues (e.g. HBB, RPL9), some showed little to no correlation across tissues (e.g. NCCRP1), some showed high intra-Brain correlation but relatively low inter-Brain correlation (e.g. ARHGAP30) (Figures 3, S2 and S5). Additionally, two genes with similar correlation profiles may have very distinct expression profiles. For example, HBB and RPL9 both showed high correlation across all tissue pairs, but they had very distinct tissue-specific expression profiles. HBB showed high expression in Whole Blood relative to other tissues, while RPL9 had a more uniform expression profile across tissues (Figure 3). A similar pattern was observed also for two genes with negligible correlation across tissues, NCCRP1 and RPL21P11 (Figure S5).

Next, we assign to each gene, a prioritizing score defined by the average value of `Robocov` correlation (*Robospan-score*) or partial correlation (*pRobospan-score*) across all tissue pairs. Similarly, we also computed the

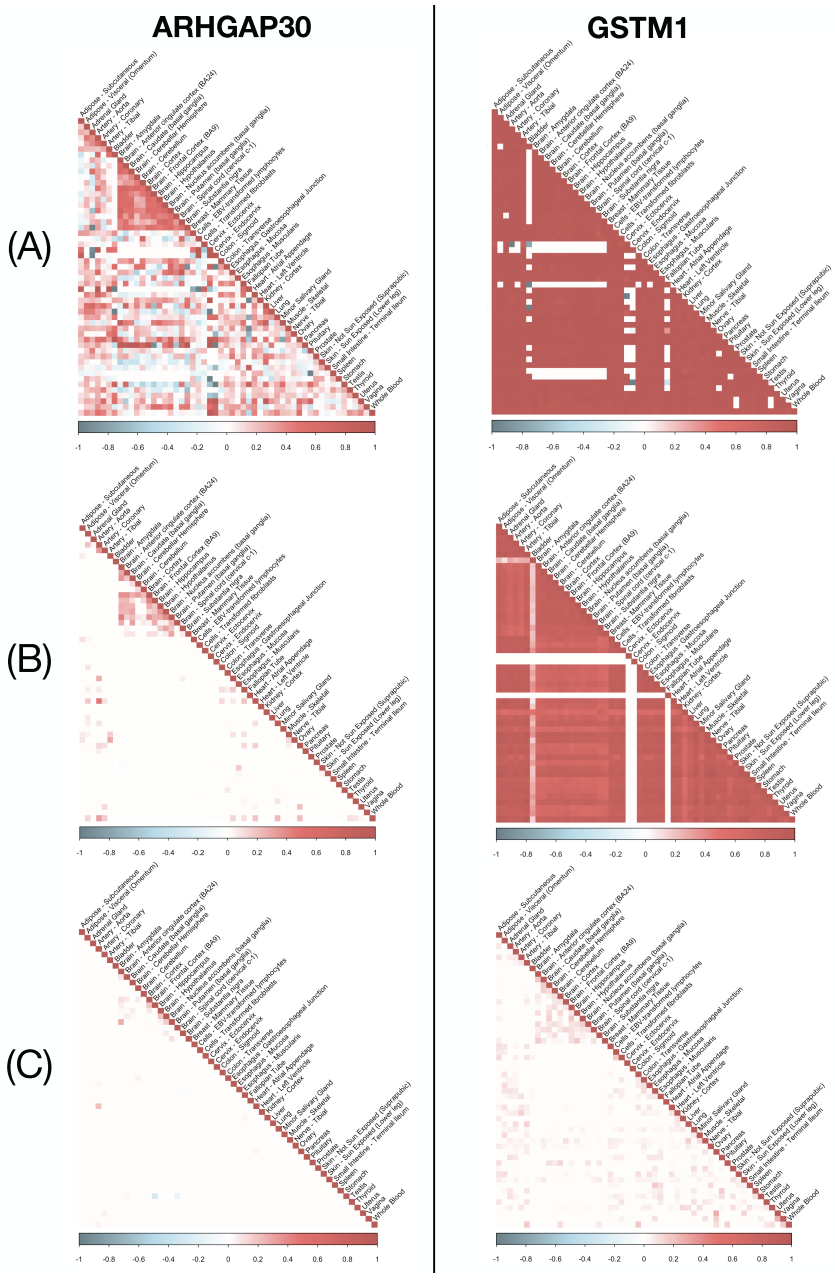

Fig. 2. Illustrative examples of pairwise sample correlation estimator, `Robocov` correlation and partial correlation estimators for 2 genes: (Left column) ARHGAP30 gene and (Right column) GSTM1 gene. Each column shows the (A) pairwise sample correlation estimator, (B) `Robocov` correlation estimator and (C) partial correlation estimator stacked from top to bottom.

average value of the pairwise sample correlation (*Corspan-score*) across tissues. Then we tested these gene scores for functional relevance. Contrary to expectation, none of the three scores showed significant enrichment in 3,804 housekeeping genes[25] (0.84x, 0.48x and 0.72x for Robospan-score, pRobospan-score and Corspan-score respectively). We compared these 3 gene scores with constraint-based metric of gene essentiality such as the absence of loss-of-function(LoF) variants (pLI[26] and s_het[27]). For each of the 50 quantile bins of pLI and s_het, we computed the median of each of these scores; and compared with the mid-value of the quantile bin. We observed a slight negative trend in all 3 scores with increasing quantile bins of both pLI and s_het (Figure S6). One possible explanation may be that genes with highly correlated expression across all tissues may be driven by tissue-shared regulation machinery which imposes lower selective constraints on these genes. The top 10% genes from each of the three gene prioritizing scores were used to define gene sets; we call them `Robospan`, `pRobospan` and `Corspan` genes. In a pathway enrichment analysis[28] of these gene sets, the top enriched pathways comprised of immune system, interferon signaling, heat stress factor (Table S2). Though not among the top 5 pathways, other interesting significant pathways included different signaling pathways (interleukin mediated signaling, NFkB signaling) and circadian clock related pathways (see URLs). The signifcance of pathway enrichment was stronger for `Robospan` and `pRobospan` genes compared to `Corspan`(Table S2). The enrichment of immune related pathways was further backed by high enrichment of

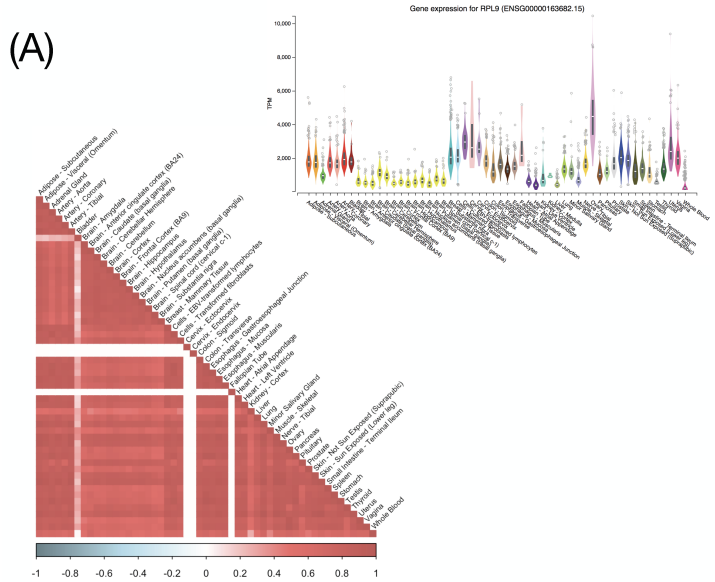

Fig. 3. Examples of genes with high average `Robocov` correlation across all tissue pairs but with distinct expression profiles. (A)RPL9 gene has uniformly high TPM (transcripts per million) values across most tissues (inset picture). (B) HBB shows high expression specifically in Whole Blood (inset picture). The expression profile plots for the genes have been fetched from the GTEx Portal (`https://gtexportal.org/home/`).

these genes in top 10% specifically expressed genes in Whole Blood (SEG-Blood[29]) (`Robospan`: 1.48x, `pRobospan`: 2.50x, `Corspan`: 1.45x). One may conjecture that this enrichment is an artifact caused by contamination of blood with GTEx tissue samples. This, however, is countered by examples of genes that have high correlation across all tissues but expression-wise, are specific to tissues that are not Whole Blood (Figure S7). We also see examples of specifically expressed genes in Whole Blood that have low Robospan-score (Figure S8).

Heritability analysis of blood-related traits

The enrichment of `Robospan`, `pRobospan` and `Corspan` genes with SEG-Blood genes and immune related pathways prompted us to test whether these genes are uniquely informative for blood-related complex diseases and traits.

For each gene set, we define SNP-level annotations to test for disease heritability. We define an *annotation* as an assignment of a numeric value to each SNP with minor allele count $\geq$5 in a 1000 Genomes Project European reference panel[30, 31]. For each gene set X, we generate two binary SNP-level annotations – we assign a value of 1 to a SNP if it lies within 5kb or 100kb window upstream and downstream of a gene in the gene set and 0 otherwise; this strategy has been used in several previous works[29, 32, 33].

We assessed the informativeness of SNP annotations for disease heritability by applying stratified LD score regression (S-LDSC)[31] conditional on 86 baseline annotations comprising of coding, conserved,

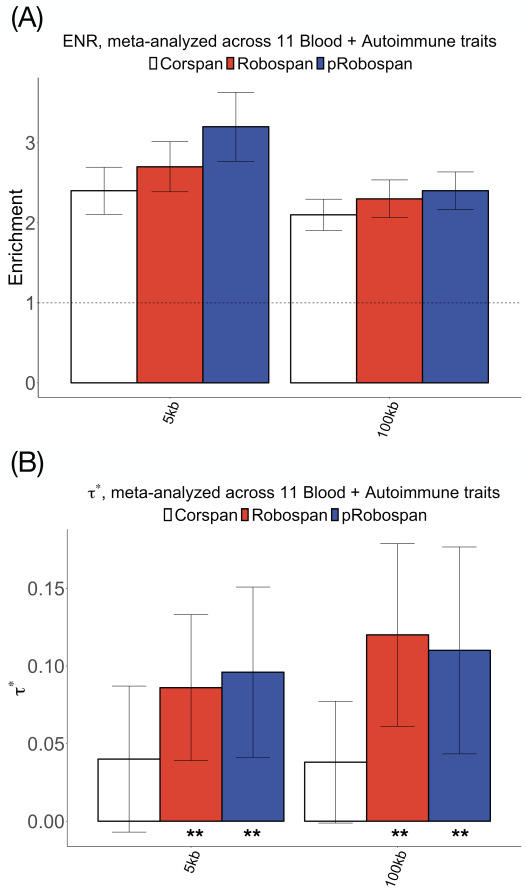

Fig. 4. Disease informativeness of 5kb and 100kb SNP annotations for **`Corspan`**, **`Robospan`** and **`pRobospan`** gene sets: (A) Heritability enrichment, conditional on baseline-LD model (v2.1). The base enrichment level is 1. (B) Standardized effect size ($\tau^\star$) conditional on baseline-LD model for `Corspan` (left column, white), `Robospan` (middle column, red) and `pRobospan` (right column, blue) gene sets. Results are meta-analyzed across 11 blood and autoimmune traits. ** denotes annotations that are significant after Bonferonni correction ($P < 0.05/8$) where 8 is the total number of SNP annotations tested. Error bars denote 95% confidence intervals. Numerical results are reported in Table S4.

epigenomic and LD related annotations (this is called the baseline-LD model; here we use version 2.1[34]). S-LDSC results were meta-analyzed across 11 relatively independent blood-related traits (5 autoimmune diseases and 6 blood traits (Table S3). We considered two S-LDSC metrics for comparison: enrichment and standardized effect size ($\tau^\star$) (Supplementary Note). Enrichment is defined as the proportion of heritability explained by SNPs in an annotation divided by the proportion of SNPs in the annotation[31]. Standardized effect size ($\tau^\star$) is defined as the proportionate change in per-SNP heritability associated with a 1 standard deviation increase in the value of the annotation, conditional on other annotations included in the model[34, 35]; unlike enrichment, $\tau^\star$ quantifies effects that are unique to the focal annotation and is a better metric for disease informativeness[5, 32, 29, 35].

All 6 annotations (5kb and 100kb for the 3 gene scores) were significantly enriched when meta-analyzed across 11 blood and autoimmune traits. However, SNP annotations corresponding to `Robospan` and `pRobospan` gene sets showed higher enrichment than `Corspan` genes (Figure 4 and Table S4). More importantly, 2 `Robospan`, 2 `pRobospan` and 0 `Corspan` annotations showed significant $\tau^\star$ conditional on the baseline-LD annotations after Bonferonni correction (Figure 4 and Table S4). When restricted to the 5 autoimmune traits, 2 `Robospan`, 0 `pRobospan` and 0 `Corspan` SNP annotations showed unique signal (Table S5). Even when these annotations were modeled jointly with SEG-Blood[29] genes and subjected to forward stepwise elimination similar to ref.[32, 5], 1 `Robospan` annotation (100kb) still remains significantly informative, suggesting unique disease information over SEG-Blood genes (Table S6).

4 Discussion

Here we present `Robocov`—a novel convex optimization-based framework for sparse estimation of covariance (correlation) and inverse covariance (partial correlation) matrix, given a data matrix with missing entries. Our approach does not rely on missing data imputation and hence mitigates the possible shortcomings of a sub-optimal imputation procedure (e.g., based on a low-rank assumption). Instead, `Robocov` directly estimates the correlation or partial correlation matrix of interest via a regularized loss minimization framework. Although here we focus our analysis on gene expression analysis, `Robocov` is a stand-alone generic tool that can be applied to any data with missing entries.

We have assessed the significance of our proposed `Robocov` framework over standard methods from a methodological, biological and disease analysis perspective. `Robocov` leads to sparse estimates and has a lower false positive rate compared to other competing methods. `Robocov` estimator is visually less cluttered and captures more robust biological signal. In terms of disease informativeness, `Robospan` and `pRobospan` gene sets, generated from the `Robocov` estimated correlation and partial correlation matrices, perform considerably better than the analogous `Corspan` gene set defined from standard correlation estimator.

There are several directions for future research. One such direction would be to incorporate covariate information underlying structured missing-ness to inform `Robocov` estimators. For GTEx data, donor metadata such as cause of death, age, gender etc can serve as important covariates. Second, we are interested in modifying `Robocov` to learn shared correlation structure between gene expression and other genetic and epigenomic data such as transcript level expression, ATAc-seq data etc. Third, from application standpoint, `Robocov` can also be used as an ingredient in item response models for large scale participant data that may contain extensive amount of missing entries, as in UK Biobank [36, 37].

URLs

- `Robocov` software
  `https://github.com/kkdey/Robocov`
- GTEx v6 data analysis, gene list, pathway enrichment results, gene sets, annotations
  `https://github.com/kkdey/Robocov-pages`
- Baseline-LD annotations:
  `https://data.broadinstitute.org/alkesgroup/LDSCORE/`
- Summary statistics:
  `https://data.broadinstitute.org/alkesgroup/sumstats_formatted/`

Acknowledgements

We thank Alkes L. Price, Bryce van de Geijn and Rajarshi Mukherjee for helpful comments. Rahul Mazumder was partially supported by the Office of Naval Research ONR-N000141512342, ONR-N000141812298 (Young Investigator Award), the National Science Foundation (NSF-IIS-1718258) and IBM.

References

[1]GTEx Consortium. The genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans. *Science*, 348(6235):648–660, 2015.

[2]GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204, 2017.

[3]K.K. Dey, C.J. Hsiao, and M. Stephens. Visualizing the structure of rna-seq expression data using grade of membership models. *PLoS genetics*, 13 (3):p.e1006599, 2017.

[4]F. Aguet et al. The gtex consortium atlas of genetic regulatory effects across human tissues. *BioRxiv*, page 787903, 2019.

[5]K.K. Dey and M Stephens. Empirical bayes shrinkage estimation of correlations, with applications. *bioRxiv*, 2018.

[6]R. Mazumder, T. Hastie, and R. Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of machine learning research*, 11(Aug):2287–2322, 2010.

[7]T. Hastie and R. Mazumder. softimpute: Matrix completion via iterative soft-thresholded svd. *R package version, 1.*, 2015.

[8]O. Ledoit and M. Wolf. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of empirical finance*, 10(5):603–621, 2003.

[9]J. Schäfer and K. Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 4(1), 2005.

[10]J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[11]T. Cai, W. Liu, and X. Luo. A constrained l1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011.

[12]M. Stephens. False discovery rates: a new deal. *Biostatistics*, 18(2):275–294, 2016.

[13]A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):p.1–22, 1977.

[14]A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust optimization*, volume 28. Princeton University Press, 2009.

[15]D. Bertsimas, D.B. Brown, and C. Caramanis. Theory and applications of robust optimization. *SIAM review*, 53(3):464–501, 2011.

[16]Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004.

[17]Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC, 2015.

[18]Brendan O'donoghue, Eric Chu, Neal Parikh, and Stephen Boyd. Conic optimization via operator splitting and homogeneous self-dual embedding. *Journal of Optimization Theory and Applications*, 169(3):1042–1068, 2016.

[19]S. Boyd, S.P. Boyd, and L Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[20]A. Fu, B. Narasimhan, and S. Boyd. Cvxr: An r package for disciplined convex optimization. *arXiv preprint arXiv:1711.07582*, 2017.

[21]R.A. Fisher. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10(4):507–521, 1915.

[22]R.A. Fisher. On the probable error of a coefficient of correlation deduced from a small sample. *Metron*, 1:3–32, 1921.

[23]A J Rothman. Positive definite estimators of large covariance matrices. *Biometrika*, 99(3):733–740, 2012.

[24]J. Schäfer and K. Strimmer. An empirical bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21(6):754–764, 2004.

[25]E. Eisenberg and E.Y. Levanon. Human housekeeping genes, revisited. *TRENDS in Genetics*, 29(10):569–574, 2013.

[26]M. Lek et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285, 2016.

[27]C.A. Cassa et al. Estimating the selective effects of heterozygous protein-truncating variants from human exome data. *Nature genetics*, 49(5):806, 2017.

[28]A. Kamburov et al. The consensuspathdb interaction database: 2013 update. *Nucleic acids research*, 41(D1):D793–D800, 2012.

[29]H.K. Finucane, Y.A. Reshef, V. Anttila, K. Slowikowski, A. Gusev, A. Byrnes, et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nature genetics*, 50:621, 2018.

[30]1000 Genomes Project Consortium. A global reference for human genetic variation. *Molecular cell*, 526(7571):p.68, 2015.

[31]H.K. Finucane, B. Bulik-Sullivan, A. Gusev, G. Trynka, Y. Reshef, P.R. Loh, V. Anttila, H. Xu, C. Zang, K. Farh, and S. Ripke. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature genetics*, 47:1228, 2015.

[32]S.S. Kim et al. Genes with high network connectivity are enriched for disease heritability. *The American Journal of Human Genetics*, 104:pp.896–913, 2019.

[33]C.A. de Leeuw et al. Magma: generalized gene-set analysis of gwas data. *PLoS computational biology*, 11(4), 2015.

[34]S. Gazal et al. Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. *Nat. Genet*, 49 (10):1421, 2017.

[35]F. Hormozdiari et al. Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits. *Nature genetics*, 50(7):1041, 2018.

[36]I. Sulis and M. Porcu. Handling missing data in item response theory. assessing the accuracy of a multiple imputation procedure based on latent class analysis. *Journal of Classification*, 34(2):p.327–359, 2017.

[37]S. Bauermeister and J. Gallacher. A psychometric evaluation of the 12-item epq-r neuroticism scale in 384,183 uk biobank participants using item response theory (irt). *BioRxiv*, page p.741249, 2019.

[38]C. Bycroft et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):p.203, 2018.

[39]L. Jostins et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*, 491:119–124, 2012.

[40]Y. Okada et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*, 506:376–381, 2014.

[41]P.C. Dubois et al. Multiple common variants for celiac disease influencing immune gene expression. *Nature genetics*, 42(4):p.295, 2010.

[42]J. Bentham et al. Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nature genetics*, 47(12):p.1457, 2015.