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
[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.