Improved Estimation of Metabolite Effects Using Shared Feature Information

Harsh Vardhan Dubey, Gregory Farage, Śaunak Sen

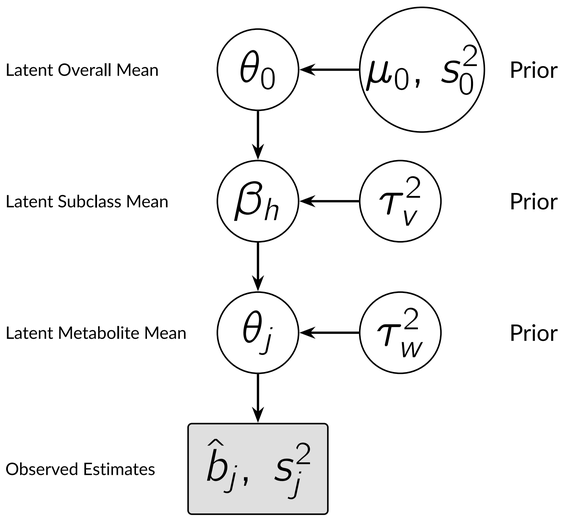
Division of Biostatistics, Department of Preventive Medicine,
The University of Tennessee Health Science Center

Introduction

- Metabolomics data are high-dimensional, noisy, and biologically structured; treating metabolites as independent often leads to unstable and high-variance effect estimates.
- Known biochemical organization such as metabolite subclasses and pathways are rarely exploited in existing modeling frameworks.
- We introduce a Bayesian hierarchical extension of MatrixLM that leverages this structure, enabling information sharing across related metabolites and yielding more stable, lower-MSE effect estimates while preserving interpretability.

Statistical framework

A hierarchical Bayesian model that pools information from individual metabolites up through subclasses, enabling stable and interpretable effect estimation across biological levels.



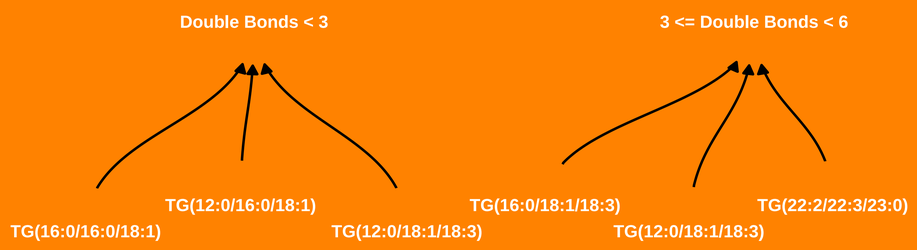
Bayesian Model: $\hat{b}_j$ are noisy observations of θ_j, partially pooled via subclass (β_h), and global mean (θ_0) with shrinkage through (τ_w, τ_v).

Take a picture for more details

THE UNIVERSITY OF TENNESSEE HEALTH SCIENCE CENTER

Borrowing strength across biologically related metabolites yields more stable and accurate effect estimates in metabolomics studies.

Individual metabolites are grouped into biologically meaningful subclasses, enabling related features to borrow strength during estimation.



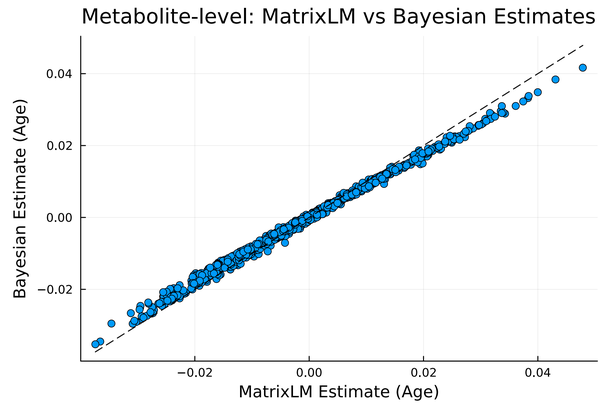
Applications

Across two independent metabolomics studies, we compare classical MatrixLM estimates with their Bayesian hierarchical counterparts for key biological covariates. We also run a diverse simulation study to probe when and why hierarchical shrinkage improves estimation, considering both *moderate* ($\tau_v = 0.12,\ \tau_w = 0.08$) and *large* ($\tau_v = 0.30,\ \tau_w = 0.12$) levels of hierarchical heterogeneity.

Simulation Study

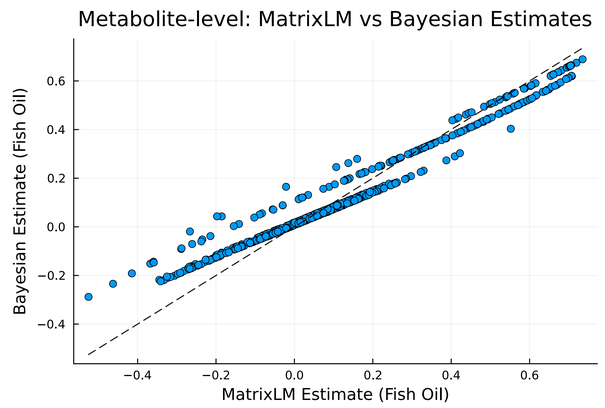
Heterogeneity	n	m	MSE Ratio	
			Estimate	Prediction
moderate	60	300	4.07	1.24
moderate	60	770	4.24	1.22
moderate	200	300	2.19	1.04
moderate	200	770	2.18	1.04
large	60	300	2.70	1.19
large	60	770	2.81	1.20
large	200	300	1.50	1.02
large	200	770	1.52	1.03

COPDGene Study Data (n = 784, m = 999)



COPDGene: Number of subclasses = 106, Number of covariates = 10.
Test MSE (MatrixLM)/Test MSE (Bayes) = 1.0004

SAMS Study Data (n = 98, m = 770)



SAMS: Number of subclasses = 4, Number of covariates = 4.
Test MSE (MatrixLM)/Test MSE (Bayes) = 1.037

Conclusion: By incorporating biochemical hierarchy, our Bayesian framework yields more stable and accurate metabolite effect estimates across a wide range of study settings. The magnitude of improvement depends on factors such as hierarchical heterogeneity and sample size n, and in some regimes existing methods may perform comparably or slightly better. Nevertheless, when meaningful biological structure is present, exploiting it is key to better inference in such studies.