Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions

Sarah M. Urbut, Gao Wang, Peter Carbonetto Matthew Stephens

Nat Genet.* 51, 187–195 (2019)

目录

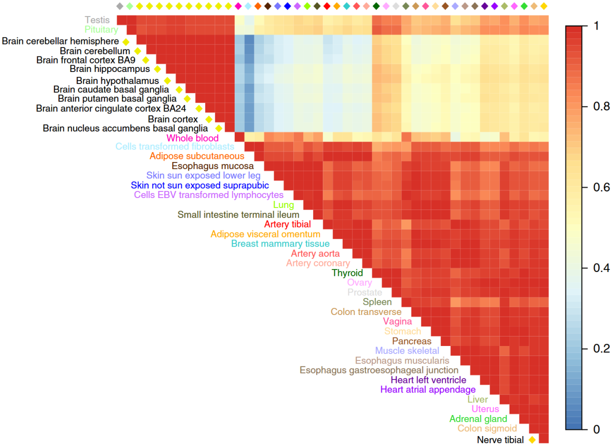
- ▶ Introduction
- ▶ Recap: Shrinkage Estimation
- ▶ Mash.method
- ▶ Mash.result

Motivation

- Genomic studies estimate effect (e.g., eQTL) of thousands of units (genes) across multiple conditions (tissues).
- In such multivariate settings, effects can be *sparse* (non-zero in few conditions), *shared* (common across conditions), or *correlated*.
- **mash** (multivariate adaptive shrinkage) learns a mixture of covariance structures to capture these patterns of heterogeneity and improve effect estimates.

Correlation Matrix of eQTL in 44 tissues

1 Introduction



目录

2 Recap: Shrinkage Estimation

- ▶ Introduction
- ▶ Recap: Shrinkage Estimation
- ▶ Mash.method
- ▶ Mash.result

spike and slab & ash

- (Spike-and-Slab): $b \sim \pi_0\delta_0 + (1-\pi_0)\mathcal{N}\left(0, \tau^2\right)$
- $w(\hat{b}) = P(b \neq 0 \mid \hat{b}) = \frac{(1-\pi_0)\varphi\left(\hat{b};0,\tau^2+\hat{s}^2\right)}{\pi_0\varphi\left(\hat{b};0,\hat{s}^2\right)+(1-\pi_0)\varphi\left(\hat{b};0,\tau^2+\hat{s}^2\right)}$
- $\mathrm{E}[b \mid \hat{b}] = w(\hat{b})\frac{\tau^2}{\tau^2+\hat{s}^2}\hat{b}, \quad \mathrm{Var}(b \mid \hat{b}, \text{ slab }) = \frac{\tau^2\hat{s}^2}{\tau^2+\hat{s}^2} < \hat{s}^2$
- ash(adaptive shrinkage): $b \sim \pi_0\delta_0 + \sum_{k=1}^{K} \pi_k\mathcal{N}\left(0, \sigma_k^2\right)$

Multivariate Gaussian

- likelihood: $\hat{\mathbf{b}} \mid \mathbf{b} \sim \mathcal{N}_R(\mathbf{b}, S), \quad S = \text{diag}\left(\hat{s}_1^2, \ldots, \hat{s}_R^2\right),$
- prior : $\mathbf{b} \sim \mathcal{N}_R(\mathbf{0}, U), \quad \text{rank}(U) = r < R$,
- posterior mean: $U\left(U + S\right)^{-1} \hat{\mathbf{b}}$

Subspace Shrinkage

Spectral Decomposition

$$U = V\Lambda V^\top, \quad Q = [V \;\; W],\; Q^\top Q = I_R$$

Orthogonalization

$$Q^\top (U+S) Q = \begin{pmatrix} \Lambda + \sigma^2 I_r & 0 \\ 0 & \sigma^2 I_{R-r} \end{pmatrix}$$

Inverse & Shrinkage

$$(U+S)^{-1} = Q \begin{pmatrix} (\Lambda + \sigma^2 I_r)^{-1} & 0 \\ 0 & \frac{1}{\sigma^2} I \end{pmatrix} Q^\top$$

$$U(U+S)^{-1} = V \operatorname{diag}\Big(\tfrac{\lambda_i}{\lambda_i + \sigma^2}\Big) V^\top$$

- **投影**：用 $V^\top \hat{b}$ 把观测映到 Col(U)
- **收缩**：$\lambda_i/(\lambda_i + \sigma^2)$
- **映回原空间**：用 V 再做一次投影
- **后验均值**：补空间方向成分收缩为 0
- **后验协方差**：$V \operatorname{diag}\big(\tfrac{\lambda_i \sigma^2}{\lambda_i + \sigma^2}\big) V^\top$

目录

3 Mash.method

- Introduction
- Recap: Shrinkage Estimation
- Mash.method
- Mash.result

Notation and Setup

- $p(\boldsymbol{b}; \boldsymbol{\pi}, \boldsymbol{U}) = \sum_{k=1}^{K} \sum_{l=1}^{L} \pi_{k,l} N_R \left(\boldsymbol{b}; 0, \omega_l U_k\right)$
- $b_{jr}(j = 1, \ldots, J; \quad r = 1, \ldots, R)$ the true value of effect j in condition r.
- let $\hat{b}_{jr}$ denote the observed estimate of this effect, and let $\hat{s}_{jr}$ be the standard error of this estimate, so $z_{jr} = \widehat{b}_{jr}/\hat{s}_{jr}$ is the standard Z statistic used to test whether b_{jr} is zero. Let $B, \hat{B}, S$ and Z denote the corresponding $J \times R$ matrices, and let $\boldsymbol{b}_j$ (respectively, $\hat{\boldsymbol{b}}_j, \boldsymbol{z}_j$) denote the j th row of B (respectively, $\hat{B}, Z$).

cont.

what is exactly b_{jr}

在 GTEx 数据里，不是对每个基因在每个组织都各选一个 topSNP

- 1. 对每个组织选择一个 topSNP
- 2. 对这些 snp 选择一个有最大的 Z-score 的 snp
- 3，以这个 snp 代表这个基因在所有组织上的基因表达量

也就是对于一个基因，所有组织，选择一个 snp

cont.

- Step 1 learn patterns of sparsity, sharing and correlations by estimating covariance matrices $\boldsymbol{U}$ and mixture proportions $\boldsymbol{\pi}$ in two substeps:
- **Step 1a**: Generate candidate covariance matrices $\boldsymbol{U} = (U_1, \ldots, U_K)$. This list includes both data-driven matrices that are estimated from the strongest signals in the condition-by-condition results and canonical matrices that have simple interpretations
- Step 1b: Given $\boldsymbol{U}$, estimate $\boldsymbol{\pi}$ by maximum likelihood
- step 2: Compute the posterior distribution for each effect given the condition-by-condition results and the fitted prior. These posterior distributions yield improved effect estimates—posterior means and standard deviations—that account for sparsity and correlations among effects.

Matrix Factorization of Z

- principal components analysis: 捕捉主要的模式
- sparse factor analysis (SFA)： 允许几个特殊的模式：（脑组织）

Singular value decomposition yields a set of singular values and singular vectors of $\tilde{Z}$. Let λ_p, v_p denote the p th singular value and corresponding right singular vector. SFA yields matrix factorization

$$\widetilde{Z} = LF + E$$

where L is a sparse $\tilde{J} \times Q$ matrix of loadings and F is a $Q \times R$ matrix of factors. We use $Q = 5$.

Data Driven Matrix U

3 Mash.method

- 但是我们的数据是 $\hat{b}_j$ 和 S_j 关心的是 b_j
- 对 $\hat{b}_j$ 做 PCA SFA 也不恰当
- 这里还存在一个 gap
- 将 Z 矩阵 PCA SFA 的结果作为 EM 算法的初始值 fit：
$p\left(\boldsymbol{b}_j \mid \boldsymbol{\pi}, U\right)=\sum_{k=1}^K \pi_k N_R\left(\boldsymbol{b}_j ; 0, U_k\right)$
- 初始值的作用：限制了 U_K 的秩
- Extreme Deconvolution: inferring complete distribution functions from noisy, heterogeneous and incomplete observations. Ann. Appl. Stat. 5, 1657–1677 (2011).

cont.

- $U_1 = \widetilde{Z}^T\widetilde{Z}/\widetilde{J}$, the empirical covariance matrix of $\widetilde{Z}$.
- $U_2 = \sum_{p=1}^{P} \lambda_p v_p v_p^T/\widetilde{J}$, which is a rank- P approximation of the covariance matrix of $\widetilde{Z}$, with $P < Q$. We use $P = 3$.
- $U_3 = F^T L^T L F/\widetilde{J}$, which is a rank-Q approximation of the covariance matrix of $\widetilde{Z}$. The output of the EM algorithm defines U_1, U_2 and U_3 in the mash model
- Covariance matrices from the SFA results; specifically, the $Q = 5$ rank-1 matrices $F_q^T L_q^T L_q F_q$, with $q = 1, \ldots, Q$.

Canonical Covariance Matrices

3 Mash.method

名称	形式	含义
单位矩阵	I_R	各条件完全独立
单条件矩阵	$e_r e_r^\top$ （共 R 个）	只在一个组织里有 effect
全共享矩阵	$\mathbf{1}\mathbf{1}^\top$	所有组织有相同的效应
其他可选	（当 R 不大时可枚举 2^{R-1} 个）	各个 configuration

mash： I_R , $e_r e_r^\top$, $\mathbf{1}\mathbf{1}^\top$

Estimating π.

Assuming independence of the rows of $\hat{B}$, the likelihood for $\boldsymbol{\pi}$ is($\Sigma_{k,l} = \omega_l U_k, p = kl$)

$$L(\boldsymbol{\pi}) = p(\widehat{B} \mid \boldsymbol{\pi}, \boldsymbol{U}, \boldsymbol{V}) = \prod_{j=1}^{J} p\left(\widehat{\boldsymbol{b}}_j \mid \boldsymbol{\pi}, \boldsymbol{U}, V_j\right) = \prod_{j=1}^{J}\sum_{p=1}^{P} \pi_p N_R\left(\widehat{\boldsymbol{b}}_j; 0, \Sigma_p + V_j\right)$$

If the rows of $\hat{B}$ are not independent, this may be interpreted as a composite likelihood , which generally yields consistent point estimates. Maximizing $\mathrm{L}(\boldsymbol{\pi})$ is a convex optimization problem, which we solve using EM, accelerated using SQUAREM. If $\hat{B}$ has a large number of rows, we can reduce computational effort by taking a random subset of rows. In the GTEx analysis, we use a random subset of 20,000 rows. (It is important that this is a random subset, and not just the $\tilde{J}$ rows of strong effects.)

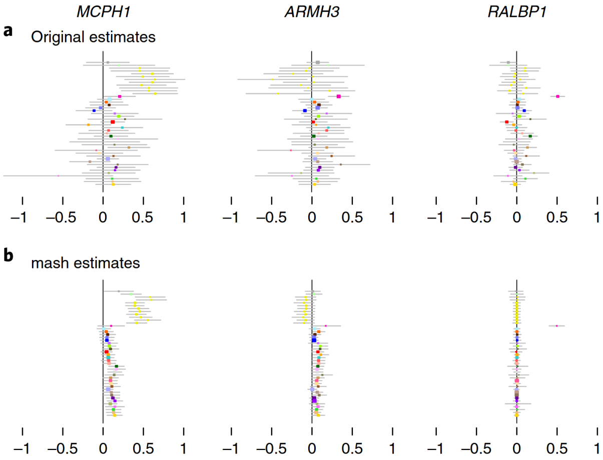
目录

4 Mash.result

- ▶ Introduction
- ▶ Recap: Shrinkage Estimation
- ▶ Mash.method
- ▶ Mash.result

Shrinkage

4 Mash.result



Overview

4 Mash.result

- mash placed 79% of the weight on the data-driven covariance matrices.
- mash framework essentially includes many methods as special cases (as well as simpler methods such as fixed-effects and random-effects meta-analyses)
- 非常 **genric,adaptive** 的方法，但是可解释性就比较差，不是很能从那么多矩阵里面看出组织之间的共享模式是什么，而是人为定义了两个量：share by magnitude, share by sign 从后验均值里来统计，总结组织之间的关系
- 注重 estimation 而不是 inference