1 The global testing problem

Let $(X, Y) \in \mathbb{R}^{p+1}$, where X is p-dimensional. Is it of interest to study the testing problem:

$$H_0 : \max_{1 \le j \le p} |\mathrm{corr}(X_j, Y)| = 0, \quad \text{vs} \quad H_1 : \max_{1 \le j \le p} |\mathrm{corr}(X_j, Y)| > 0 \,?$$

Would GWAS and/or eQTL be possible applications?

2 Lasso tuning using only summary statistics

why summary stat available

Assume we have $Y \in \{0, 1\}^n$, $X \in \mathbb{R}^{n \times (p+1)}$, where the first column of X is $\mathbf{1}_n$, which corresponds to the intercept term. We want to learn a linear model that predicts Y from X. Consider the lasso estimator

$$\hat{\beta}_\lambda = \arg\min \frac{1}{n} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 \,,$$

which is equivalent to

$$\hat{\beta}_\lambda = \arg\min 2R^T\beta + \beta^T C \beta + \lambda \|\beta\|_1 \,,$$

where

$$R = \frac{1}{n} X^T Y, \quad C = \frac{1}{n} X^T X$$

are the empirical correlation between X and Y, and the empirical covariance of X. The following discussion will focus on these quantities, assuming that the original data (X, Y) is not accessible.

Remark: In the two-population problem, we may have $C = \gamma C_1 + (1 - \gamma) C_2$, $R = \gamma R_1 + (1 - \gamma R_2)$, where C_j, R_j are the sample correlations from population $j \in \{1, 2\}$. The following development is based on the two-population context. The one-population version is simpler, just dropping all the j indices.

Ideally the tuning of λ should aim at minimizing the predictive risk from one population:

$$\mathrm{Risk}_j(\hat{\beta}_\lambda) = \mathbb{E}\left[(Y_{\mathrm{new}}^{(j)} - (X_{\mathrm{new}}^{(j)})^T \hat{\beta}_\lambda)^2 \middle| \hat{\beta}_\lambda \right]$$

where $(X_{\mathrm{new}}^{(j)}, Y_{\mathrm{new}}^{(j)})$ is an independent draw from the jth population.

However this may be too much to ask. Even with fully accessible data, we need very strong assumptions to achieve this.

A more practical quantity is the "Same-X" predictive risk: **i should also condition on beta here?**

$$\text{Risk}_j(X) = \mathbb{E}\left[\|Y^{(j)}_{\text{new}}(X^{(j)}) - X^{(j)}\hat{\beta}_\lambda\|_2^2 \middle| X^{(1)}, X^{(2)}\right] \tag{1}$$

where $Y^{(j)}_{\text{new}}(X^{(j)})$ is a fresh draw of $Y^{(j)}$ from the conditional distribution of $(Y^{(j)}|X^{(j)})$ (treating the design matrix $X^{(j)}$ as fixed), and the expectation is over both $Y^{(j)}$ and $Y^{(j)}_{\text{new}}$. I.e., how accurate does $X^{(j)}\hat{\beta}_\lambda(Y, X)$ predict $Y^{(j)}_{\text{new}}(X^{(j)})$, the response in population j generated again from the same X?

Based on a result of Efron [2], the Same-X predictive risk can be related to the in sample predictive risk as follows:

$$\text{Risk}_j(X) = \|Y^{(j)} - X^{(j)}\hat{\beta}_\lambda\|_2^2 + 2\sum_{i=1}^{n_j} \text{Cov}(\hat{\mu}_i^{(j)}, Y_i^{(j)})\,, \tag{2}$$

where $\hat{\mu}_i^{(j)} = (X_i^{(j)})^T\hat{\beta}$ is the fitted mean value for $\mathbb{E}(Y_i^{(j)}|X_i^{(j)})$.

The term $\text{Cov}(\hat{\mu}_i, y_i)$ is called the "optimism" (also known as "covariance penalty"), which quantifies the difference between the actual predictive risk and the in-sample predictive risk. It is easy to check that the in-sample predictive risk $\|Y^{(j)} - X^{(j)}\hat{\beta}_\lambda\|_2^2$ can be computed using only the summary statistics:

$$\begin{aligned}\|Y^{(j)} - X^{(j)}\hat{\beta}_\lambda\|_2^2 =& (Y^{(j)})^TY^{(j)} - 2\hat{\beta}_\lambda^T(X^{(j)})^TY^{(j)} + \hat{\beta}_\lambda^T(X^{(j)})^TX^{(j)}\hat{\beta}_\lambda \\ =& (Y^{(j)})^TY^{(j)} - 2n_j\hat{\beta}_\lambda^TR_j + n_j\hat{\beta}_\lambda^TC_j\hat{\beta}_\lambda\,.\end{aligned}$$

Remark: The in-sample prediction error $\|Y^{(j)} - X^{(j)}\hat{\beta}_\lambda\|_2^2$ involves the term $(Y^{(j)})^TY^{(j)}$, which equals n_j times the average of $Y^{(j)}$ (because the entries are binary), which is also n_j times the first entry of R_j. Therefore, we only need to know n_j, the sample size of population j in the training sample, to compute this quantity.

Remark: If the goal is to select tuning parameters, then we do not even need to evaluate the term $(Y^{(j)})^TY^{(j)}$, because it does not involve λ, and hence is common for all λ's.

2.1 Approximating the covariance penalty

Let $(Y_1^{(1)*}, ..., Y_{n_1}^{(1)*})$, $(Y_1^{(2)*}, ..., Y_{n_2}^{(2)*})$ be a bootstrap sample, then we can approximate the optimsim term by

$$\widehat{\text{Cov}}(\hat{\mu}_i^{(j)}, Y_i^{(j)}) = \text{Cov}_*(\hat{\mu}_i^{(j)*}, y_i^{(j)*})$$

where $\mu_i^{(j)*} = (X_i^{(j)})^T\hat{\beta}^*$, and $\hat{\beta}^*$ is the bootstrap version of $\hat{\beta}$ using the bootstrap sample.

The bootstrap sample is generated by

$$y_i^{(j)*} = \hat{\mu}_{i,0}^{(j)} + \epsilon_i^{(j)*}$$

where $\epsilon_i^{(j)*}$ is a centered Bernoulli noise such that $\mathbb{E}_*\epsilon_i^{(j)*} = 0$ and $\mathbb{E}_*(\epsilon_i^{(j)*})^2 = \hat{\mu}_{i,0}^{(j)}(1 - \hat{\mu}_{i,0}^{(j)})$.

Pr(epsilon = 1-mu) = mu; Pr(epsilon = -mu) = 1-mu

Remark: Here $\hat{\mu}_0$ is a "preliminary estimate", which is expected to be fairly accurate although not optimal. Such an estimate can usually be optained by a small-ish λ.

Now

$$\begin{aligned}\sum_{i=1}^{n} \mathrm{Cov}_*(\hat{\mu}_i^{(j)*}, Y_i^{(j)*}) =& \sum_{i=1}^{n_j} \left[\mathbb{E}_*\hat{\mu}_i^{(j)*}(\hat{\mu}_{i,0}^{(j)} + \epsilon_i^{(j)*}) - \mathbb{E}_*\hat{\mu}_i^{(j)*}\mathbb{E}_*(\hat{\mu}_{i,0}^{(j)} + \epsilon_i^{(j)*})\right] \\ =& \sum_{i=1}^{n_j} \mathbb{E}_*\hat{\mu}_i^{(j)*}\epsilon_i^{(j)*} \\ =& \mathbb{E}_*(\hat{\beta}^*)^T(X^{(j)})^T\epsilon^{(j)*}\,.\end{aligned}$$

calculate

In order to obtain $\hat{\beta}^*$, we will need the bootstrap versions of R_1 and R_2, where

$$\begin{aligned}R_j^* = (X^{(j)})^T Y^{(j)*} =& (X^{(j)})^T(X^{(j)}\hat{\mu}_0 + \epsilon^{(j)*}) \\ =& n_j C_j \hat{\mu}_0 + (X^{(j)})^T\epsilon^{(j)*}\,.\end{aligned} \tag{3}$$

Therefore, in order to obtain the optimism estimate, we only need to generate

$$(X^{(j)})^T\epsilon^{(j)*}$$

for $j = 1, 2$.

Let $X^{(j)} = U_j D_j V_j^T$ be the singular value decomposition of $X^{(j)}$. If we only have the summary statistic C_j, then U_j is not accessible but (V_j, D_j) are. Here we keep in mind the setting that $X^{(j)} \in \mathbb{R}^{n_j \times (p+1)}$ where $p > n_j$. So U_j is full rank orthonormal.

Thus we only need to generate $\tilde{\epsilon}^*$:

$$\tilde{\epsilon}^* = U^T\epsilon^*\,.$$

The frist and second moments of $\tilde{\epsilon}^{(j)*}$ are

$$\mathbb{E}(\epsilon^{(j)*}) = 0$$

and

$$\mathbb{E}\epsilon^{(j)*}(\epsilon^{(j)*})^T = \mathrm{diag}(\hat{\mu}_{1,0}^{(j)} - (\hat{\mu}_{1,0}^{(j)})^2, ..., \hat{\mu}_{n,0}^{(j)} - (\hat{\mu}_{n,0}^{(j)})^2)$$

= E(U^t epsilon epsilon^t U) = U^t diag U

U = [[0,1], [1,0]] is a counter example?

take an average?

We can use the following a "partially-second-order" Gaussian approximation

$$\tilde{\epsilon}^{(j)*} \sim N(0, \tau_j^2 I_{n_j}) \tag{4}$$

where

$$\begin{aligned}\tau_j^2 &= \left((\hat{\mu}_0^{(j)})^T \mathbf{1}_{n_j} - (\hat{\mu}_0^{(j)})^T \hat{\mu}_0^{(j)}\right)/n_j \\ &= \left(\hat{\beta}_0^T (X^{(j)})^T \mathbf{1}_{n_j} + \hat{\beta}_0^T (X^{(j)})^T X^{(j)} \hat{\beta}_0\right)/n_j \\ &= \hat{\beta}_{1,0} + \hat{\beta}_0^T C_j \hat{\beta}_0 \,. \end{aligned} \tag{5}$$

Here $\hat{\beta}_{1,0}$ is the first coordinate (i.e., intercept) of the preliminary estimate $\hat{\beta}_0$, and we used the fact that the other columns in $X^{(j)}$ are centered and hence sum to 0.

is this always true in practice??

Procedure: Input: Summary statistics R_j, C_j for $j = 1, 2$ and corresponding sample sizes n_j; Preliminary estimate $\hat{\beta}_0$; bootstrap sample size B

For $b = 1, ..., B$, $j = 1, 2$

1. Generate $\tilde{\epsilon}^{j(*)}$ according to (4) with τ_j^2 given in the formula (5).
2. Generate (emulate) $(X^{(j)})^T \epsilon^{(j)*}$ by $V_j D_j \tilde{\epsilon}^{(j)*}$ where $V_j D_j^2 V_j^T$ is the SVD of nC_j.
3. Compute $\hat{\beta}^*$ using (R_1^*, C_1, R_2^*, C_2), with R_j^* determined by (3).

The covariance panelty in (2) is approximated by the average of $(\hat{\beta}^*)^T (X^{(j)})^T \epsilon^{(j)*}$ over the bootstrap repetitions.

3 Variable selection and fine mapping

Consider regression model

$$Y = X\beta + \epsilon$$

where $\beta \in \mathbb{R}^p$ is sparse, with $\|\beta\|_0$ likely being only 1, 2, or 3. The problem of testing whether $\beta = 0$ is the global effect test (which is related to Section 1), and can also be solved using the method presented here.

A main challenge in the fine mapping problem is that the columns of X can be highly correlated. Making it impossible to consistently recover the support of β: $S = \{1 \le j \le p : \beta_j \neq 0\}$. Instead, we aim at finding promising candidates (e.g., confidence sets) of S.

A Bayesian method SuSiE considers a similar problem, but outputs a posterior credible set that covers at least one element of S.

where the minimization is taken over all joint distributions c on $\{1, ..., p\}^2$ such that its marginal distributions are π and κ respectively. Each $q \geq 1$ defines a Wasserstein distance. For simplicity we can start from the case $q = 1$.

The Wasserstein-1 distance has a clear geometric interpretation (the shortest distance that one needs to move the probability mass to change π to κ), and has a natural range: $W_1(\pi, \kappa) = 0$ if and only if $\pi = \kappa$ and $W_1(\pi, \kappa) = 1$ if and only if π and κ are most different.

Challenges: (1) How does this work if there are multiple effects? The SuSiE PIP will no longer be a probability distribution over the SNPs. (2) How does this work if we use CVC-MPS? In the single effect case, the CVC outputs a collection of confidence sets indexed by the confidence level, but these are not probability distributions. One possibiliy is to aggregate the differences between the confidence set at different levels.

Using CVC for co-localization. The output of CVC is a collection of confidence sets at different confidence levels. We can convert such a collection of confidence sets to a probability distribution and then use the Wasserstein distance to compare the proximity of the two distributions for GWAS and eQTL, using the same idea described above for the PIP distributions. Here we describe how to convert the confidence sets to a probability distribution.

Let A be the set of SNPs. Since there are finitely many SNPs, conceptually we can find a sequence of confidence levels $0 = a_0 < a_1 < a_2 ... < a_K = 1$ such that for each $1 \leq k \leq K-1$ the corresponding level a_k confidence set A_k records the sequence of distinct confidence sets when we increase the confidence level from 0 to 1. The probability distribution $\pi = [\pi(j) : j \in A]$ corresponding to this collection of confidence sets is

$$\pi(j) = \sum_{k=1}^{K} (a_k - a_{k-1}) \frac{\mathbb{1}(j \in A_k \backslash A_{k-1})}{|A_k \backslash A_{k-1}|}$$

with the default $A_0 = \emptyset$, $A_K = A$.

References

[1] Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

[2] Bradley Efron. The estimation of prediction error: covariance penalties and cross-validation. *Journal of the American Statistical Association*, 99(467):619–632, 2004.

[3] Nicholas Kissel and Lucas Mentch. Forward stability and model path selection. *arXiv preprint arXiv:2103.03462*, 2021.

[4] Jing Lei. Cross-validation with confidence. *Journal of the American Statistical Association*, 115(532):1978–1997, 2020.

[5] Victor M Panaretos and Yoav Zemel. *An invitation to statistics in Wasserstein space*. Springer Nature, 2020.

[6] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.