Supplement for: Adjusting for principal components of molecular phenotypes induces replicating false positives

Andy Dahl[a,1], **Vincent Guillemot**[b], **Joel Mefford**[a], **Hugues Aschard**[b,c], **and Noah Zaitlen**[a,1]

[a]Department of Medicine, University of California San Francisco, San Francisco, 94158; [b]Centre de Bioinformatique, Biostatistique et Biologie Intégrative, Institut Pasteur, Paris, 15th Arrondissement; [c]Department of Epidemiology, Harvard TH Chan School of Public Health

This manuscript was compiled on September 3, 2018

1. Calculations used to derive PC conditioning bias

We use the same notation and data as in the main text:

- $Y \in \mathbb{R}^{N\times P}$ is the phenotype matrix ($N \leq P$). We use the model
$$Y = x\alpha^T + Y^0$$
- $Y^0 \in \mathbb{R}^{N\times P}$ is the noise matrix
- $x \in \mathbb{R}^{N\times 1}$ is the covariate of interest
- $\alpha \in \mathbb{R}^{P\times 1}$ is the vector of effect sizes for x
- $U \in \mathbb{R}^{N\times N}$, $\lambda \in \mathbb{R}^N$, and $V \in \mathbb{R}^{P\times N}$ are defined as the SVD
$$\frac{1}{\sqrt{p}}Y = U\text{diag}(\lambda)V^T$$
we define U^0, λ^0, and V^0 analogously for Y^0
- We also use the inverse eigengaps
$$c_j := \frac{1}{\lambda_1 - \lambda_j}$$
- Confounders are not discussed in this document but can be modelled without loss of generality as components of the general noise matrix Y^0.

A. Equation 7: two stage least squares. We first evaluate a few helpful terms using the first-order approximation to u_1 developed in equation 6 in the main text:

$$\begin{aligned}
x^T u_1 &= \tilde{x}_1 + a\sum_{j>1} c_j\tilde{x}_j(\tilde{y}_{1q}\tilde{x}_j + \tilde{y}_{jq}\tilde{x}_1) + O(a^2)\\
y_p^T u_1 &= \tilde{y}_{1p} + a\sum_{j>1} c_j\tilde{y}_{jp}(\tilde{y}_{1q}\tilde{x}_j + \tilde{y}_{jq}\tilde{x}_1) + O(a^2)\\
x^T u_1 y_p^T u_1 &= \tilde{x}_1\tilde{y}_{1p} + a\sum_{j>1} c_j(\tilde{x}_1^2\tilde{y}_{jp}\tilde{y}_{jq} + \tilde{x}_j^2\tilde{y}_{1p}\tilde{y}_{1q} + \tilde{x}_1\tilde{x}_j(\tilde{y}_{jp}\tilde{y}_{1q} + \tilde{y}_{1p}\tilde{y}_{jq})) + O(a^2)
\end{aligned}$$

We use these terms to expand the two-stage least squares expression for $\hat{\alpha}$:

$$\begin{aligned}
\hat{\alpha}_p &= \frac{y_p^T x - y_p^T u_1 x^T u_1}{N - (x^T u_1)^2}\\
&= \frac{(y_p^T x - y_p^T u_1^0 x^T u_1^0) + (y_p^T u_1^0 x^T u_1^0 - y_p^T u_1 x^T u_1)}{N - (x^T u_1)^2}\\
&= \frac{N - (x^T u_1^0)^2}{N - (x^T u_1)^2}\left(\hat{\alpha}_p^0 - \frac{a\sum_{j>1} c_j(\tilde{x}_1^2\tilde{y}_{jp}\tilde{y}_{jq} + \tilde{x}_j^2\tilde{y}_{1p}\tilde{y}_{1q} + \tilde{x}_1\tilde{x}_j(\tilde{y}_{jp}\tilde{y}_{1q} + \tilde{y}_{1p}\tilde{y}_{jq}))}{N - (x^T u_1^0)^2}\right) + O(a^2)\\
&= \hat{\alpha}_p^0(1+\gamma) - \frac{a(1+\gamma)}{N - \tilde{x}_1^2}\left(\tilde{x}_1^2\sum_{j>1} c_j\tilde{y}_{jq}\tilde{y}_{jp} + \tilde{x}_1\sum_{j>1} c_j\tilde{x}_j(\tilde{y}_{jq}\tilde{y}_{1p} + \tilde{y}_{1q}\tilde{y}_{jp}) + \tilde{y}_{1q}\tilde{y}_{1p}\sum_{j>1} c_j\tilde{x}_j^2\right) + O(a^2)
\end{aligned}$$

The last line introduced γ, which is defined as

$$\gamma := \frac{(x^T u_1)^2 - (x^T u_1^0)^2}{N - (x^T u_1)}$$

Using the fact that $\gamma = O(a)$, derived in Section 1C, this gives the bias approximation

$$\text{Bias}_p = \hat{\alpha}_p^0 \gamma - \frac{a}{N - \tilde{x}_1^2} \left(\tilde{x}_1^2 \sum_{j>1} c_j \tilde{y}_{jq} \tilde{y}_{jp} + \tilde{x}_1 \sum_{j>1} c_j \tilde{x}_j (\tilde{y}_{jq} \tilde{y}_{1p} + \tilde{y}_{1q} \tilde{y}_{jp}) + \tilde{y}_{1q} \tilde{y}_{1p} \sum_{j>1} c_j \tilde{x}_j^2 \right) + O(a^2) \quad [1]$$

We now assume γ is approximately negligible, which follows from a central limit theorem approximation used below in Section 1C, to write

$$\text{Bias}_p \approx -\frac{a}{N - \tilde{x}_1^2} \left(\tilde{x}_1^2 \sum_{j>1} c_j \tilde{y}_{jq} \tilde{y}_{jp} + \tilde{x}_1 \sum_{j>1} c_j \tilde{x}_j (\tilde{y}_{jq} \tilde{y}_{1p} + \tilde{y}_{1q} \tilde{y}_{jp}) + \tilde{y}_{1q} \tilde{y}_{1p} \sum_{j>1} c_j \tilde{x}_j^2 \right)$$

We now drop middle term inside the parentheses that sums terms proportional to $\tilde{x}_1 \tilde{x}_j$. These summands are each products of normal random variables with mean zero and only have standard deviation $c_j(\tilde{y}_{jq}\tilde{y}_{1p} + \tilde{y}_{1q}\tilde{y}_{jp})$. In contrast, the summands in the third term in the parentheses have mean $c_j \tilde{y}_{1p}\tilde{y}_{1q}$ and variance $2c_j^2 \tilde{y}_{1p}^2 \tilde{y}_{q1}^2$. By the central limit theorem, for large N the comparison simplifies to comparing a $\mathcal{N}\left(0, \sum_{j>1} c_j^2 \frac{(\tilde{y}_{jq}\tilde{y}_{1p} + \tilde{y}_{1q}\tilde{y}_{jp})^2}{\tilde{y}_{1q}^2 \tilde{y}_{1p}^2}\right)$ to a $\mathcal{N}\left(\sum_{j>1} c_j, \sum_{j>1} c_j^2\right)$. We say the former is negligible because it has mean zero and its standard deviation is smaller than the mean of the latter:

$$\sqrt{\sum_{j>1} c_j^2 \frac{(\tilde{y}_{jq}\tilde{y}_{1p} + \tilde{y}_{1q}\tilde{y}_{jp})^2}{\tilde{y}_{1q}^2 \tilde{y}_{1p}^2}} \approx \|c_{-1}\|_2 \leq \|c_{-1}\|_1$$

The inequality is fully general and in our case holds loosely: in the GEUVADIS data (described below) and Marchenko-Pastur spectra with the same aspect ratio, $\frac{\|c_{-1}\|_1}{\|c_{-1}\|_2}$ is 19.3 and 9.1, respectively. Marchenko-Pastur is a useful reference distribution for the singular values from an unstructured matrix, as it is almost surely the limiting spectral measure for entry-wise i.i.d. random matrices as their dimensions increase to ∞ with a fixed aspect ratio.

The above approximation assumes

$$\frac{\tilde{y}_{jq}\tilde{y}_{1p} + \tilde{y}_{1q}\tilde{y}_{jp}}{\tilde{y}_{1q}\tilde{y}_{1p}} = \frac{\sqrt{\lambda_j}}{\sqrt{\lambda_1}} \frac{V_{jq}V_{1p} + V_{1q}V_{jp}}{V_{1q}V_{1p}} \leq 1$$

which is reasonable so long as entries of V are bounded away from zero; further, we only require the inequality on average over j (after weighting by c_j^2).

Dropping the middle summand and the term in γ simplifies the bias expression considerably, giving

$$\begin{aligned} \text{Bias}_p &\approx -\frac{a}{N - \tilde{x}_1^2} \left(\tilde{x}_1^2 \sum_{j>1} c_j \tilde{y}_{jq} \tilde{y}_{jp} + \tilde{y}_{1q} \tilde{y}_{1p} \sum_{j>1} c_j \tilde{x}_j^2 \right) \\ &= -2a\bar{c} \frac{N-1}{N - \tilde{x}_1^2} \left(\sum_{j>1} w_j^{\tilde{x}} \tilde{y}_{jq} \tilde{y}_{jp} + w_1^{\tilde{x}} \tilde{y}_{1q} \tilde{y}_{1p} \right) \\ &\approx -2a\bar{c} \sum_{j=1}^{N} \tilde{y}_{jp} \tilde{y}_{jq} w_j^{\tilde{x}} \end{aligned} \quad [2]$$

As in the main text, the weights $w^{\tilde{x}}$ and $\bar{c}$ are defined by

$$w_1^{\tilde{x}} := \frac{\sum_{k>1} c_k \tilde{x}_k^2}{2(N-1)\bar{c}}; \quad w_j^{\tilde{x}} := \frac{c_j \tilde{x}_1^2}{2(N-1)\bar{c}} \quad \bar{c} := \frac{1}{N-1} \sum_{j>1} c_j$$

The approximation in Eq. (2) uses

$$\frac{1}{N - \tilde{x}_1^2} \approx \frac{1}{N-1} \left(1 + \frac{\tilde{x}_1^2 - 1}{N-1} \right) \approx \frac{1}{N-1}$$

which is correct to first order in the random variable $\frac{\tilde{x}^2 - 1}{N-1}$, which has mean zero and variance $\frac{2}{N-1}$.

[1] E-mail: andywdahl@gmail.com, noah.zaitlen@ucsf.edu

B. $\bar{c}$ approximation. We drop the superscript from λ^0 for this subsection.

We require a strong assumption that $\lambda_1 \gg \lambda_2$ for our first approximation below. Unfortunately, K eigenvalues will be large compared to all other if there are K strong confounders, so this approximation will be worse for more realistic data. In Supplementary Figure 11, for example, the top singular value is roughly 1.5 times larger than the second. Regardless, a regression conditioning on K PCs would presumably instead require the last $N-K$ to be small, which is far mot attainable. In fact, it automatically holds if K is chosen, for example, as an elbow of the $\{\lambda\}$ scree plot, which is very common in practice. Finally, these approximation are the last step in our overall bias calculation, and serve primarily to reduce our expression into interpretable forms–greater approximation accuracy can be achieved leaving the approximation in terms of the (inscrutable) parameter $\bar{c}$.

So, assuming $\lambda_1 \gg \lambda_2$,

$$\begin{aligned}
\bar{c} &:= \frac{1}{N-1}\sum_{j>1}\frac{1}{\lambda_1-\lambda_j} \approx \frac{1}{(N-1)\,\lambda_1}\sum_{j>1}\left(1+\frac{\lambda_j}{\lambda_1}\right) \\
&= \frac{1}{(N-1)\,\lambda_1}\left(N-1+\frac{\left(\sum_{j=1}^N \lambda_j\right)-\lambda_1}{\lambda_1}\right) \\
&\approx \frac{1}{\lambda_1}\left(1+\frac{N-\lambda_1}{(N-1)\,\lambda_1}\right) \\
&\approx \frac{1}{\lambda_1} \qquad [3]
\end{aligned}$$

This used the approximation that $\sum_{j=1}^N \lambda_1 = N$. This approximation is not very good–writing λ as a length-N vector, $N = \|\lambda\|_1^2 \le \|\lambda\|_1^2$, with approximate equality only when $\lambda_1 \gg \sum_{j>1}\lambda_j$, a much stronger condition than $\lambda_1 \gg \lambda_2$. However, unlike other approximations, this one is guaranteed to be conservative (in the sense that the true $\bar{c}$, and thus the resulting bias approximation, is even larger than our provided approximation), and so we do not overly worry about accuracy.

We note that $\sum_{j=1}^N \lambda_j^2 = \|\lambda\|_2^2 = N$ because we assumed the matrix Y is normalized to have columns with mean zero and variance 1, giving

$$\sum_{j=1}^N \lambda_1 = \mathrm{tr}\left(\frac{1}{P}Y^TY\right) = \frac{1}{P}\sum_{i,p}Y_{ip}^2 = \frac{1}{P}\sum_p \|Y_{,p}\|_2^2 = N$$

C. Simplifying computations for γ. Defining and simplifying γ gives

$$\begin{aligned}
\gamma &:= \frac{N-\tilde{x}_1^2}{N-(x^Tu_1)^2} - 1 \\
&= \frac{1}{1-2a\tilde{x}_1\sum_{j>1}c_j\tilde{x}_j(\tilde{y}_{1q}\tilde{x}_j+\tilde{y}_{jq}\tilde{x}_1)(N-\tilde{x}_1^2)^{-1}+O(a^2)} - 1 \\
&= 2a\tilde{x}_1\sum_{j>1}c_j\tilde{x}_j(\tilde{y}_{1q}\tilde{x}_j+\tilde{y}_{jq}\tilde{x}_1)(N-\tilde{x}_1^2)^{-1}+O(a^2) \\
&= \frac{2a}{N-\tilde{x}_1^2}\left(\tilde{x}_1^2\sum_{j>1}c_j\tilde{y}_{jq}\tilde{x}_j+\tilde{x}_1z_{1q}\sum_{j>1}c_j\tilde{x}_j^2\right)+O(a^2) \\
&\approx 2a\tilde{x}_1\tilde{y}_{1q}\bar{c}
\end{aligned}$$

The last line uses the approximations that $\frac{N-1}{N-\tilde{x}_1^2} \approx 1$ and replaces the sums over j by their expectations, which is reasonable as

$$\sum_{j>1}c_j\tilde{x}_j^2 \sim \mathcal{N}\left((N-1)\bar{c}, 2\sum_{j>1}c_j^2\right) \qquad \text{[by CLT]}$$

$$\sum_{j>1}c_j\tilde{y}_{jq}\tilde{x}_j \sim \mathcal{N}\left(0, \sum_{j>1}c_j^2\tilde{y}_{jq}^2\right)$$

The CLT approximation should be good unless a very small number of eigenvalues (other than the first) are far larger than the rest, similar to our requirement for the $\bar{c}$ approximation that $\lambda_1 \gg \lambda_2$ (but weaker).

Using Eq. (3) to simplify $\bar{c}$, γ can then be approximated by

$$\gamma \approx 2a\tilde{x}_1V_{q1}\frac{\sqrt{P}}{\sqrt{\lambda_1}} \qquad [4]$$

In practice, this term is negligible: a is assumed small, V_{q1} is on the order of $P^{-1/2}$, and $\sqrt{\lambda_1}$ tends to be large in real data, e.g. 8 in GEUVADIS.

D. PCA on residuals. To derive the condition for the inequality $\hat{\sigma}^2_{cond} \leq \hat{\sigma}^2_{cond}$,

$$\begin{aligned}
\hat{\sigma}^2_{cond} &= \frac{1}{N-(K+1)} \| \left(I - xx^T - UU^T\right) y_p \|^2 \\
&= \frac{N-1}{N-(K+1)} \hat{\sigma}^2_{uncond} - \frac{1}{N-(K+1)} \|U^T y_p\|^2 \\
&= \hat{\sigma}^2_{uncond} + \frac{1}{N-(K+1)} \left(K\hat{\sigma}^2_{uncond} - P\sum_{k=1}^{K} \lambda_k V^2_{pk} \right)
\end{aligned}$$

When Y has i.i.d. entries and N and P are large, we use the approximation $PV_{p,1:K} \overset{\text{iid}}{\sim} \chi^2_1$. Then

$$\begin{aligned}
\hat{\sigma}^2_{cond} &= \frac{1}{N-(K+1)} \| \left(I - xx^T - UU^T\right) y_p \|^2 \\
&= \frac{N-1}{N-(K+1)} \hat{\sigma}^2_{uncond} - \frac{1}{N-(K+1)} \|U^T y_p\|^2 \\
&= \hat{\sigma}^2_{uncond} + \frac{1}{N-(K+1)} \left(K\hat{\sigma}^2_{uncond} - P\sum_{k=1}^{K} \lambda_k V^2_{pk} \right)
\end{aligned}$$

This holds on average over x and choice of gene p,

$$\mathbb{E}\left(\hat{\sigma}^2_{uncond} - \frac{P}{K} \sum_{k=1}^{K} \lambda_k V^2_{pk} \right) = 1 - \frac{1}{K} \sum_{p=1}^{P} \sum_{k=1}^{K} \lambda_k V^2_{pk} = 1 - \frac{1}{K} \sum_{k=1}^{K} \lambda_k \leq 0$$

using the fact that Y is column-standardized, so $\mathbb{E}\left(\hat{\sigma}^2_{uncond}\right) = 1$.

We also empirically evaluate the probability the t-statistic is inflated by simulating $\hat{\sigma}^2_{uncond} \sim \chi^2_1$ and, independently, $PV^2_{pk} \overset{\text{iid}}{\sim} \chi^2_1$. We repeat the simulation 10,000 times for different choices of V and K and plotted the fraction of simulations where $t_{\text{uncond}} < t_{\text{cond}}$. The probability of inflation is always above 50%; peaks in the range of K used in practice (roughly, 5 to 20); and increases for more highly dispersed (i.e. realistic) eigenvalue spectra.

2. Accuracy of the approximations

We used the GEUVADIS data, described in the main text, as our Y^0 matrix to empirically assess the quality of the above approximations. We simulated 1,000 independent datasets from the model

$$\begin{aligned}
Y &\sim x\alpha^T + Y_{\text{GEUVADIS}} \\
x &\overset{\text{iid}}{\sim} \mathcal{N}(0,1) \\
\alpha &= ae_q = (0, \ldots, a, \ldots, 0) \\
q &\sim \text{Unif}(\{1, \ldots, P\}) \\
a &\in \{.0001, .001, .01, .1, .5\}
\end{aligned}$$

For each dataset, we compute the bias $\hat{\alpha}_p - \hat{\alpha}^0_p$ for each gene $p \neq q$ and our theoretical approximations to this bias. We then regress the observed biases on the theoretical biases, storing the regression coefficients and R^2.

The median R^2 is greater than .9999 for all a for both the "full" approximation given in Eq. (1) and the simplest approximation given in Equation 9 of the main text, i.e.

$$\text{Bias}_{-p} \approx aV_{p1}V_{-p1}$$

The empirical distribution of these regression coefficients is shown in Figure 2. The median coefficients are always negligibly far from 1, though the empirical 95% confidence intervals are nontrivial. These observations suggest our bias approximation is off by a scale factor near one, possibly because higher-order terms also predominantly scale in $V_{p1}V_{q1}$. We have not pursued more accurate approximations as the goal is only to demonstrate the existence and qualitative behavior of the bias; moreover, where the perturbation is truly small, we show in simulations below that the resulting bias is essentially negligible.

Overall, the final approximation given in Equation 9 of the main text appears to be a very good estimator, despite depending only on a, q and $V_{,1}$.

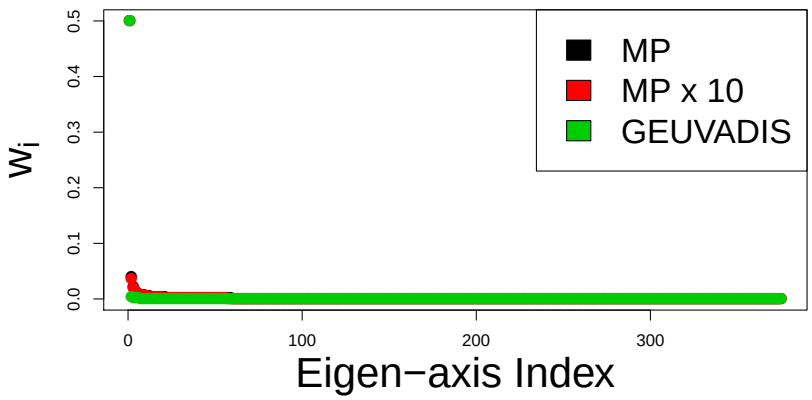

Fig. 1. Weights for the pseudo-correlation in three types of data. The GEUVADIS data is described in the main text; MP refers to Marchenko-Pastur with asymptotic aspect ratio equal to that in GEUVADIS; MPx10 is the same with aspect ratio ten times larger. In all cases, $w_1 = \frac{1}{2} \gg w_2$, especially in the real GEUVADIS data.

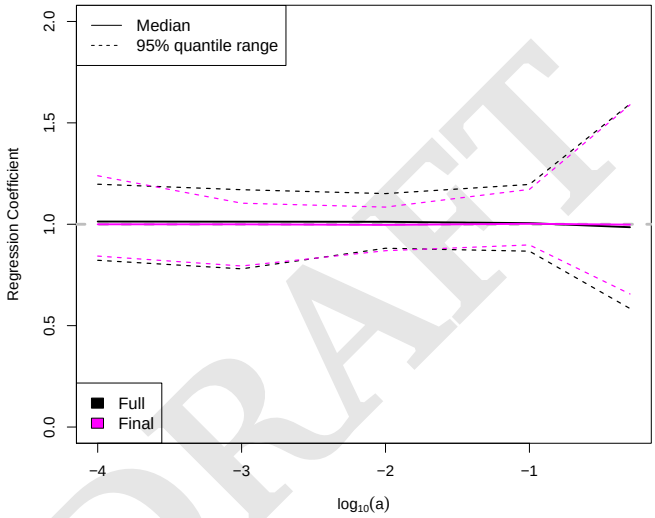

Fig. 2. Results from regressing the observed bias in $\hat{\alpha}$ on two novel approximations to the bias. The "final" approximation simplifies several negligible terms in the "full" approximation. The distribution of the regression coefficient is shown, which is identically 1 for perfect approximations. The x-axis varies a, the true signal strength, on a log scale.

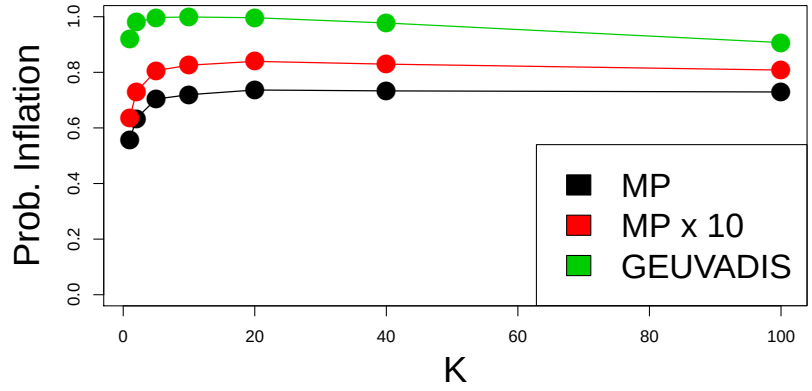

Fig. 3. Approximate probability that conditioning on supervised PCs inflates the nominally null t-statistic when regressing x on a column of Y. GEUVADIS, MP, and MPx10 are as in Supplementary Figure 1.

SI Figures.

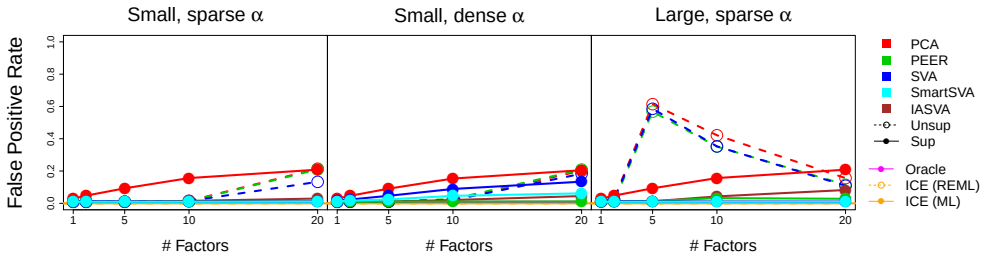

Fig. 4. False positive rates for a simulated, strong, *trans*-effect added to the GEUVADIS expression using a p-value threshold of 0.01. Three parameterizations for the causal effect α are shown. FPR is shown on the non-logged scale so that ICE is visible.

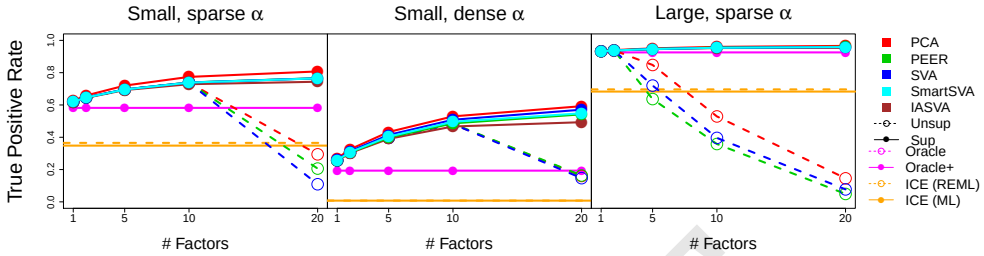

Fig. 5. True positive rates for a simulated, strong, *trans*-effect added to the GEUVADIS expression using a q-value threshold of 0.01. TPR is not the same as power for approaches that fail to control the FPR.

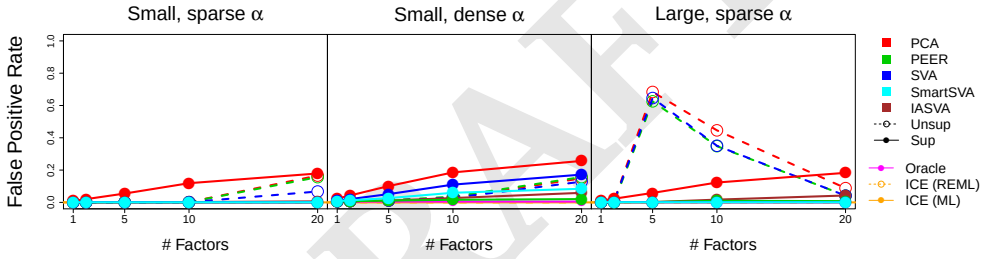

Fig. 6. False positive rates for a simulated, strong, *trans*-effect added to the GEUVADIS expression using a q-value threshold of 0.01. Three parameterizations for the causal effect α are shown. Using q-values rather than p-values seems to help PEER more than SVA.

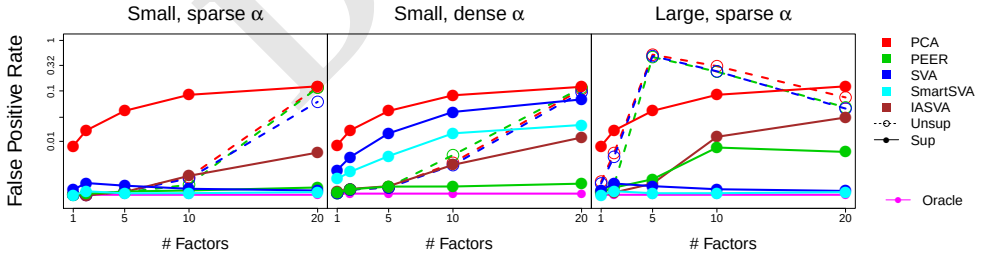

Fig. 7. False positive rates for a simulated, strong, *trans*-effect added to the GEUVADIS expression using a p-value threshold of 0.001 (on the log scale). Three parameterizations for the causal effect α are shown.

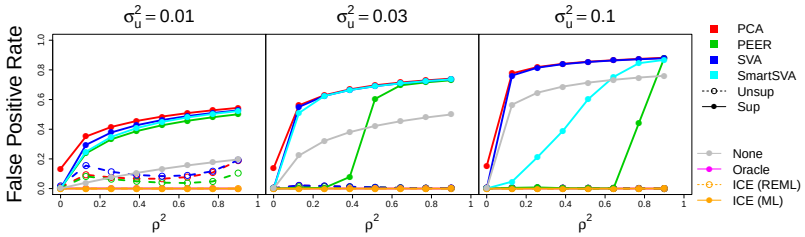

Fig. 8. False positive rates at a q-value threshold of .01 for testing a strong covariate x that is correlated with an unobserved confounder u. The squared correlation between x and u, ρ^2, is on the x-axis, and their respective transcriptome-wide variances explained are $\sigma_x^2 = 1\%$ and $\sigma_u^2 = 10\%$ (left) or $\sigma_u^2 = 40\%$ (right).

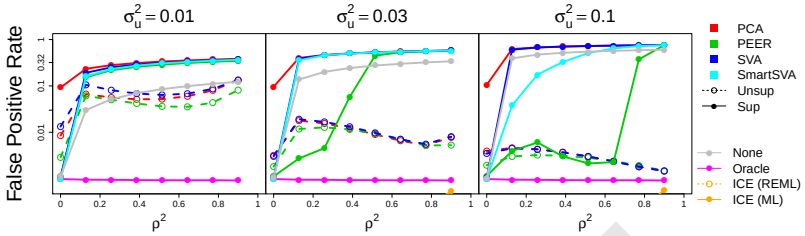

Fig. 9. False positive rates (on log scale) at a nominal significance threshold of .001 for testing a strong covariate x that is correlated with an unobserved confounder u. The squared correlation between x and u, ρ^2, is on the x-axis, and their respective transcriptome-wide variances explained are $\sigma_x^2 = 1\%$ and $\sigma_u^2 = 10\%$ (left) or $\sigma_u^2 = 40\%$ (right).

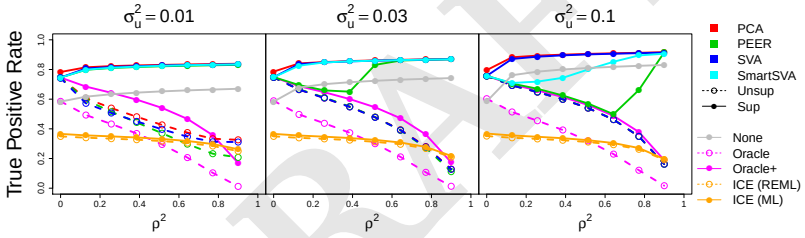

Fig. 10. True positive rates at a q-value threshold of .01 for testing a strong covariate x that is correlated with an unobserved confounder u. The squared correlation between x and u, ρ^2, is on the x-axis, and their respective transcriptome-wide variances explained are $\sigma_x^2 = 1\%$ and $\sigma_u^2 = 10\%$ (left) or $\sigma_u^2 = 40\%$ (right).

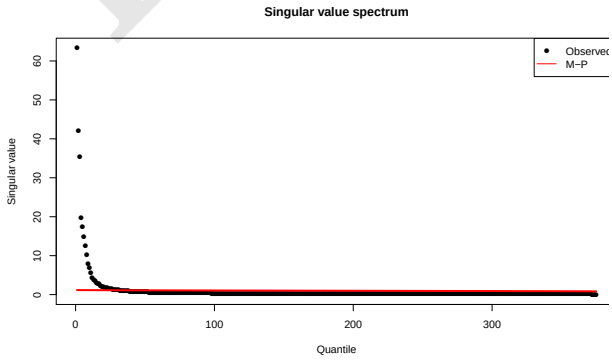

Fig. 11. Singular-value spectrum of our process gene expression matrix from GEUVADIS. M-P refers to the Marchenko-Pastur distribution with asymptotic aspect ratio matching the GEUVADIS data.