Poor portability of polygenic scores even within an ancestry group

H. Mostafavi, A. Harpak, D. Conley, J.K. Pritchard, M. Przeworski

April 11, 2019

Contents

Contents

List of Figures

List of Tables

Key implications   ffdsafs

1 Matching sib-regression and standard GWAS sample sizes

We are looking for sample size n^* of a standard GWAS using unrelated individuals such that if we use the estimated effects in a polygenic score, the prediction accuracy matches the prediction accuracy of a polygenic score based on a sib regression with sample size n_{pairs}. Throughout, we assume that all causal sites i are known and unlinked, and that there is no population stratification or assortative mating. We begin by finding the error associated with estimating the effect for a single site in each type of GWAS. We will then examine (and match) the prediction accuracy of the polygenic scores composed from effects estimated in the unrelated/sib estimation sets, $\hat{\beta}_{ur}, \hat{\beta}_{sib}$—on a new, independent prediction set $\{(x', y')\}$.

1.1 Error of the estimated effect at a single site

In the unrelated sample GWAS, our model is

$$y = g + e$$

where e is the environmental effect and

$$g = \beta_0^{ur} + \sum_i \beta_i x_i$$

is the polygenic score. Our model for the effect of site i is

$$y = \beta_0 + \beta_i x_i + \epsilon_i, \tag{1}$$

where

$$\epsilon_i = g - \beta_i x_i + e,$$

with variance

$$Var[\epsilon_i] = Var[g - \beta_i x_i] + Var[e] = Var[y] - \beta_i^2 Var[x_i].$$

In OLS regression for the effect of site i, the standard error is

$$Var[\hat{\beta}_i^{ur}] = \frac{Var[\epsilon_i]}{(n-1)Var[x_i^{ur}]} = \frac{Var[y] - \beta_i^2 Var[x_i]}{(n-1)Var[x_i]},$$

where n is the sample size. In sib regression our model for site i is

$$\Delta y = \beta_0^s + \beta_i \Delta x_i + \Delta\epsilon_i,$$

with variance

$$Var[\Delta\epsilon_i] = Var[\Delta g - \beta_i \Delta x_i] + Var[\Delta e] =$$

$$Var[\Delta g] + \beta_i^2 Var[\Delta x_i] - 2\beta_i^2 var[\Delta x_i] + Var[\Delta e]$$

Using the A and B subscripts to denote sibs. Recall that for sibs we expect

$$Cov[x_{i,A}, x_{i,B}] = \frac{1}{2}Var[x_i],$$

$$Cov[g_A, g_B] = \frac{1}{2}Var[g].$$

Plugging this back in we get

$$Var[\Delta\epsilon_i] = Var[g] - \beta_i^2 var[x_i] + 2Var[e](1-\rho)$$

where $\rho = Cor[e_A, e_B]$ is the environmental correlation between sibs. The error of the estimated effect in sib regression is therefore

$$Var[\hat{\beta}_i^s] = \frac{Var[\Delta\epsilon_i]}{(n_{pairs}-1)Var[\Delta x_i]} = \frac{Var[y] - \beta_i^2 var[x_i] + Var[e](1-2\rho)}{(n_{pairs}-1)Var[x_i]}.$$

1.2 Sample size required for matched prediction accuracy

We wish to find a sample size n^* in a unrelated sample GWAS s.t. the prediction accuracy matches that of a polygenic score function $\hat{g}_{sib}$ derived from sib regression. We will measure prediction accuracy as the expected correlation between the polygenic score function $\hat{g}$ and the phenotype in an independent prediction set {(x',y')},

$$R = \frac{Cov[\hat{g}(x'), y']}{\sqrt{Var[y']Var[\hat{g}(x')]}},$$

by requiring

$$R(standard\ GWAS\ with\ sample\ size\ n^{*}) \overset{!}{=} R(sib\ regression\ with\ sample\ size\ n^{pairs})$$

or equivalently

$$\frac{Cov[\hat{g}_{sib}(x'), y']}{\sqrt{Var[\hat{g}_{sib}(x')]}} \overset{!}{=} \frac{Cov[\hat{g}_{ur}(x'), y']}{\sqrt{Var[\hat{g}_{ur}(x')]}}. \tag{2}$$

We will find a sample size n^{*} to use in the estimator $\hat{g}_{ur}$ that satisfies this condition. We first note that if the estimates $\hat{\beta}$ are given then

$$Cov[y', \hat{g}(x')|\hat{\beta}] = Cov[g(x'), g(x')] + \sum_{i}^{m} x'_i(\hat{\beta}_i - \beta_i)|\hat{\beta}] =$$

$$Var[g(x')|\hat{\beta}] + \sum_{i}^{m} Cov[\beta_i x'_i, (\hat{\beta}_i - \beta_i)x'_i|\hat{\beta}] = \sum_{i}^{m} Var[x'_i]\beta_i\hat{\beta}_i,$$

However, to incorporate the uncertainty both in the estimation set (summarized by the Multivariate Normal distribution of $\hat{\beta}$) and the prediction set (the randomness of (x', y')) we will use the law of total covariance,

$$Cov[y', \hat{g}(x')] = E_{\hat{\beta}}[Cov_{(x',y')}[y', \hat{g}(x')|\{\hat{\beta}\}]] + Cov_{\hat{\beta}}[E_{(x',y')}[Cov[y'|\{\hat{\beta}\}], E_{(x',y')}[\hat{g}(x')|\{\hat{\beta}\}]] =$$

$$E_{\hat{\beta}} \sum_{i}^{m} Var[x'_i]\beta_i\hat{\beta}_i + Cov_{\hat{\beta}}[\sum_{i}^{m} E[x'_i]\beta_i, \sum_{i}^{m} E[x'_i]\hat{\beta}_i] = \sum_{i}^{m} Var[x_i]\beta_i^2.$$

Plugging this back into eq. 2, we are left with the requirement

$$Var[\hat{g}_{sib}(x')] \overset{!}{=} Var[\hat{g}_{ur}(x')]. \tag{3}$$

Applying the law of total variance for each estimated polygenic score $\hat{g}$,

$$Var[\hat{g}_{sib}(x')] = Var_{\hat{\beta}}[E_{x'}[\hat{g}_{sib}(x')|\hat{\beta}]] + E_{\hat{\beta}}[Var_{x'}[\hat{g}_{sib}(x')|\hat{\beta}]] =$$

$$\sum_{i}^{m} E[X'_i]Var[\hat{\beta}_i] + \sum_{i}^{m} Var[X'_i]Var[\hat{\beta}_i].$$

After plugging this back into 3 and reordering,

$$\frac{n^* - 1}{n_{sib} - 1} = \frac{1}{1 + \frac{Var[e](1-2\rho)}{Var[y] - \frac{(\sum_i^m E[x_i^2]\beta_i^2)}{(\sum_i^m \frac{E[x_i^2]}{Var[x_i]})}}},$$

or, assuming

$$\frac{(\sum_i^m E[x_i^2]\beta_i^2)}{(\sum_i^m \frac{E[x_i^2]}{Var[x_i]})} << Var[y],$$

there's probably a nicer way to present this assumption. but should be a valid assumption

we find

$$\frac{n^*}{n^{pairs}} \approx \frac{1}{1 + (1 - h_g^2)(1 - 2\rho)}. \tag{4}$$

1.3 Empirical matching of standard errors

Note that the result of eq. 4 the same as we would get if we required

$$\forall i \; Var[\hat{\beta}_i^{sib}(x_i)] \overset{!}{=} Var[\hat{\beta}_i^{ur}(x_i^{sib})] \tag{5}$$

without taking randomness in the prediction set into account. In practice (in the main text), we have no prior knowledge on rho and instead we find a sample size n^* for the standard GWAS s.t.

$$median_i(Var[\hat{\beta}_i^{sib}(x)]) \overset{!}{=} median_i(Var[\hat{\beta}_i^{ur}(x)]) \tag{6}$$

The reason that eq. 5 is approximately met is that if we assume that y is a highly polygenic trait where

$$\forall i \; \beta_i^2 Var[xi] << Var[y],$$

then

$$\forall i \; Var[\hat{\beta}_i^{sib}(x)] = Var[\hat{\beta}_i^{ur}(x)] = \frac{D}{Var[x_i]}$$

where D is the same for sib- and standard GWAS estimates (as n^* is used), and approximately independent of β_i. Now, eq. 6 can be thought of as a weighted-median of D. In conclusion, the requirement of eq. 6 leads to equal prediction accuracy under the model assumptions.

2 Indirect parental effects

2.1 Distribution of effect estimate at a single site

We start by considering the model

$$y = \beta_0 + g + n + e$$

where g is a direct-effects polygenic score of an individual with genotypes (effect-allele count) x_i at each site i,

$$g = \sum_i^m \beta_i x_i,$$

and

$$n = \sum_i^m \eta_i (x_i + \tilde{x}_i^m + \tilde{x}_i^p)$$

is the indirect-effects polygenic score of an individual with parental allele counts $x_i + \tilde{x}_i^p + \tilde{x}_i^m$ at each site where $\tilde{x}_i^m$ is the untransmitted maternal effect allele count, and $\tilde{x}_i^p$ is the untransmitted paternal effect allele count, with $\tilde{x}_i^m, \tilde{x}_i^p \in \{0, 1\}$. We will show that if we take the strategy of matching effect estimate errors between sib regressuib and standard GWAS then the prediction accuracy of the two polygenic score functions in an independent sample can differ. Specifically in the case of a large positive correlation between indirect and direct effects, the standard GWAS polygenic score is expected to outperform the sib-based polygenic score.

We first examine the distribution of an estimate of the effect of x_i on the phenotype. The OLS regression for a single site in a standard GWAS follows eq. 1 and can be rewritten as

$$y = \beta_0 + (\beta_i + \eta_i)x_i + \eta_i(\tilde{x}_i^p + \tilde{x}_i^m) + \epsilon_i \tag{7}$$

with

$$\epsilon_i = g + n + e - (\beta_i + \eta_i)x_i - \eta_i(\tilde{x}_i^p + \tilde{x}_i^m).$$

From eq. 7 and the assumption of no assortative mating or other population structure, giving

$$Cov[\tilde{x}_i^p, \tilde{x}_i^m] = Cov[x_i, \tilde{x}_i^m] = Cov[x_i, \tilde{x}_i^p] = 0, \tag{8}$$

it directly follows that $\hat{\beta^{ur}}_i$ is Normally distributed and unbiased–with expectation

$$E[\hat{\beta}_i] = \beta_i + \eta_i.$$

We next calculate $\hat{\beta}_i^{ur}$'s variance. From assumption 8 and

$$Var[\tilde{x}_i^m + \tilde{x}_i^p] = Var[x_i],$$

we get

$$Var[\epsilon_i] = Var[y]+(\beta_i+\eta_i)^2Var[x_i]+\eta_i^2Var[x_i]-2Cov[g+n,(\beta_i+\gamma_i)x_i]-2Cov[n,\eta_i(\tilde{x}_i^m+\tilde{x}_i^p)] =$$

$$= Var[y] - (\beta_i + \eta_i)^2 Var[x_i].$$

Finally,

$$Var[\hat{\beta}_i^{ur}] = \frac{Var[\epsilon_i]}{(n-1)Var[x_i]} = \frac{Var[y] - (\beta_i + \eta_i)^2 Var[x_i]}{(n-1)Var[x_i]}.$$

In sib regression we have

$$\Delta y = \Delta g + \Delta e$$

as indirect parental effects completely cancel out in taking the difference between sibs, because the sibs have an equal paternal effect allele count. Thus, as in the case of direct effects alone,

$$\hat{\beta}_i^{sib} \sim N(\beta_i, \frac{Var[y] - \beta_i^2 var[x_i] + Var[e](1-2\rho)}{(n_{pairs}-1)Var[x_i]})$$

2.2 Polygenic score prediction accuracy

We now examine the difference in prediction accuracy of $\hat{g}^{ur}$ and $\hat{g}^{sib}$ after matching

$$Var[\hat{\beta}_i^{ur}] \overset{!}{=} Var[\hat{\beta}_i^{sib}] \tag{9}$$

by choosing a standard GWAS effect size n^* that empirically satisfies the condition–as we do in the main text (see also section 1.3); we get

$$Cov[y', \hat{g}(x')] = \sum_i^m (\beta_i + \eta_i) Cov[\hat{\beta}_i x'_i, x'_i] + \eta_i Cov[\hat{\beta}_i x'_i, (\tilde{x}_i'^m + \tilde{x}_i'^p)].$$

Taking eq. 8 into account, the second term is zero. By the law of total covariance,

$$Cov[y', \hat{g}(x')] = \sum_i^m (\beta_i + \eta_i) E_{\hat{\beta}}[Cov_{x'}[\hat{\beta}_i x'_i, x'_i] | \hat{\beta}] + \sum_i^m (\beta_i + \eta_i) Cov_{\hat{\beta}}[E_{x'}[\hat{\beta}_i x'_i | \hat{\beta}], E_{x'}[x'_i] | \hat{\beta}]] =$$

$$= \sum_i^m (\beta_i + \eta_i) E_{\hat{\beta}}[Var[x_i] \hat{\beta}_i] + \sum_i^m (\beta_i + \eta_i) Cov_{\hat{\beta}}[\hat{\beta}_i E[x_i], E[x'_i]],$$

$$= \sum_i^m Var[x_i] (\beta_i + \eta_i) E[\hat{\beta}_i]. \tag{10}$$

Similarly, by the law of total variance,

$$Var[\hat{g}(x')] = Var_{\hat{\beta}}[E_{x'}[\hat{g}(x') | \hat{\beta}]] + E_{\hat{\beta}}[Var_{x'}[\hat{g}(x') | \hat{\beta}]] = Var_{\hat{\beta}}[\sum_i^m \hat{\beta}_i E[x_i]] + E_{\hat{\beta}}[\sum_i^m Var[x_i] \hat{\beta}_i^2] =$$

$$\sum_i^m E[x_i]^2 Var[\hat{\beta}_i] + \sum_i^m Var[x_i] E[\hat{\beta}_i^2]. \tag{11}$$

Taken together, eq. 10 and eq. 11 give

$$R = \frac{Cov[\hat{g}(x'), y']}{\sqrt{Var[y'] Var[\hat{g}(x')]}} = \frac{\sum_i^m Var[x_i] (\beta_i + \eta_i) E[\hat{\beta}_i]}{\sqrt{Var[y]} \sqrt{\sum_i^m E[x_i]^2 Var[\hat{\beta}_i] + \sum_i^m Var[x_i] E[\hat{\beta}_i^2]}} =$$

$$= \frac{\sum_i^m Var[x_i] (\beta_i + \eta_i) E[\hat{\beta}_i]}{\sqrt{Var[y] (\sum_i^m Var[\hat{\beta}_i] E[x_i^2] + \sum_i^m Var[x_i] E[\hat{\beta}_i]^2)}} \tag{12}$$

We next note that

$$C := Var[y] \sum_i^m Var[\hat{\beta}_i] E[x_i^2]$$

is the same for sib regression and standard GWAS under the requirement of eq. 9. We therefore have

$$R^{ur} = \frac{\sum_i^m Var[x_i] (\beta_i + \eta_i)^2}{\sqrt{C + Var[y] \sum_i^m Var[x_i] (\beta_i + \eta_i)^2}},$$

and

$$R^{sib} = \frac{\sum_i^m Var[x_i](\beta_i + \eta_i)\beta_i}{\sqrt{C + Var[y]\sum_i^m Var[x_i]\beta_i^2}}.$$

We examined the fit of this prediction to a simulations of our setup: estimation in a sib-GWAS, estimation in a unrelated sample GWAS after choosing n^* to match the variance of effect estimates, and finally prediction in a sample of unrelated individuals. We used the following parameters:

- ratio of direct effects heritability to indirect effects heritability: 5

- narrow sense heritability from direct effects: 0.5

- number of loci each contributing equally: 1000

- number of sibling pairs for sib GWAS: Three different values: 15,000; 150,000 and 1,500,000.

- number of sibling pairs for sib validation: 5000

- number of unrelated individuals for validation: 20,000

- number of iterations (for estimating n^* and R under a given ρ (correlation between direct and indirect effect sizes) and heritabilities ratio: 1.

Fig. S1 shows the fit of the analytic derivation to simulated data with different correlation coefficients between indirect and direct effects. We note that even in the absence of correlation between indirect and direct effects, the polygenic score based on standard GWAS outperforms the sib-based polygenic score. This is a consequence of a nonzero C term in eq. 12. C is proportional to $Var[\hat{\beta}_i]$, and so vanishes when estimation sample sizes are very large. For realistic sample sizes, however, C is nonzero. As Fig. S1 demonstrates, unless the indirect and direct effects are strongly negatively correlated (< 0.5), prediction accuracy will be higher using the standard GWAS with unrelated individuals.

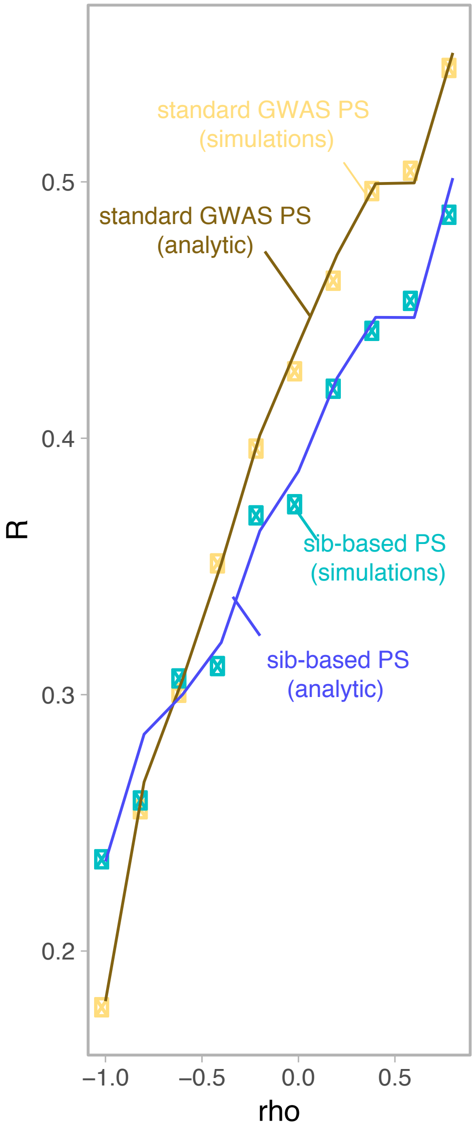

Figure S1: [XXX details on simulation params]