Quasi-regression for heritability

Art B. Owen
Stanford University

March 2012

Abstract

We show in an idealized model that the narrow sense (linear) heritability from d autosomal SNPs can be estimated without bias from n independently sampled individuals, by a method of moments strategy based on quasi-regression. The variance of this estimator is below $C_1/n + C_2 d^2/n^3$ for constants $C_j < \infty$, uniformly in $n \geqslant 2$ and $d \geqslant 1$. In particular $d \gg n$ is allowed, and heritability can be consistently estimated in some limits where the effects of individual SNPs cannot be consistently estimated due to non-sparsity. Furthermore, when the $C_2 d^2/n^3$ term dominates, then doubling the sample size reduces the variance by almost eight–fold.

The idealization is that the model assumes complete linkage equilibrium, although it does allow for an arbitrary pattern of regression coefficients. In particular the coefficients need not be sparse or Gaussianly distributed, nor independent of the minor allele frequency.

The method and the rate of convergence extend to additive heritability. For a full quadratic model encompassing pairwise interactions with $d^* = 2d + d(d-1)/2$ coefficients the error variance is $O(d^4/n^3)$.

1 Introduction

A missing heritability problem has emerged in genome wide association studies (Manolio et al., 2009). A phenotype known to be largely heritable, for example height, may be found to have only a few significantly predictive SNPs and those few may explain in total only a small portion of the known heritability.

A typical dataset setup has measurements of d SNPs on n subjects where frequently $d \gg n$. Multiple linear regression of the phenotype on SNPs is a useful way to study their effects. Even when $d \gg n$, it is still possible to estimate the needed regression coefficients, if the true coefficient vector is sparse. See for example Zhao and Yu (2006). When that coefficient vector is not sparse, then it is not possible to estimate it accurately (Candès and Davenport, 2011).

2 Notation and models

In this paper, the meaning of $a(n,d) = O(b(n,d))$ is that there is some constant $C < \infty$ for which $|a(n,d)| \leqslant C \times b(n,d)$ holds for all $n \geqslant 2$ and all $d \geqslant 1$. That is, we use non-asymptotic order bounds.

Let y_i be a phenotype for subject i. We assume that the population average of this phenotype has been subtracted from y_i so that $\mathbb{E}(y_i) = 0$. We will suppose that the phenotype is bounded, $|y_i| \leqslant c < \infty$. Boundedness is not necessary, but we prefer to write c^3 instead of, for example, $\mathbb{E}(y^{12})^{1/4}$. Conversely, we find $\mathbb{E}(y^2)$ more informative than c^2, so we do not use the bound everywhere.

Let $\widetilde{x}_{ij}$ be a genetic measure for SNP j on subject i. For an autosomal SNP, $\widetilde{x}_{ij}$ might be 1 if subject i has one or more copies of the minor allele and zero otherwise. Alternatively, $\widetilde{x}_{ij} \in \{0,1,2\}$ might be the number of copies of minor allele j in subject i. Let $x_{ij} = (\widetilde{x}_{ij} - \mathbb{E}(\widetilde{x}_{ij}))/\sqrt{\mathrm{Var}(\widetilde{x}_{ij})}$ so that $\mathbb{E}(x_{ij}) = 0$ and $\mathrm{Var}(x_{ij}) = 1$ in the relevant population.

We assume additionally that the components $x_1, \ldots, x_d$ of $\boldsymbol{x}$ are independent of each other. This neglects linkage disequilibrium (LD). LD only affects a small fraction of the $O(d^2)$ SNP pairs, but we expect it could be a significant consideration in aggregate because there are so many such pairs.

The k'th moment of x_j is $\mu_{jk} = \mathbb{E}(x_j^k)$. We will need μ_{j4} for each SNP. If SNP j has minor allele frequency $\epsilon_j > 0$, then $\mu_{4j} = O(\epsilon_j^{-1})$. The variable x_j is bounded. In the motivating examples, $|x_j| \leqslant c_j$ where $c_j = O(\epsilon_j^{-1/2})$.

In most work, we assume that $\epsilon_j \geqslant \epsilon_0$ holds for some $\epsilon_0 > 0$. Then μ_{4j} are uniformly bounded. Some intermediate expressions use the average $\bar{\mu}_4 = (1/d)\sum_{j=1}^d \mu_{j4}$. These allow us to consider a limit in which alleles with smaller MAF become eligible for use as n increases. For example, if $\epsilon_j \geqslant n^{-1/2}$ then $\mu_{4j} \leqslant n^{1/2}$, and if we only assume $\epsilon_j \geqslant m/n$ (perhaps the smallest reasonable MAFs to use) for some $m > 0$, then $\mu_{4j} = O(n)$.

The Euclidean norm of $\boldsymbol{x}$ is denoted $\|\boldsymbol{x}\|$. We easily find that $\mathbb{E}(\|\boldsymbol{x}\|^2) = d$ and $\mathbb{E}(\|\boldsymbol{x}\|^4) = d^2 + d(\bar{\mu}_4 - 1)$, so $\mathrm{Var}(\|\boldsymbol{x}\|) = d(\bar{\mu}_4 - 1)$. As a result $\|\boldsymbol{x}\|/d = 1 + O_p(d^{-1/2}\bar{\mu}_4^{1/2})$. In particular

$$\mathbb{E}\biggl(\sum_{i=1}^n\Bigl(\frac{\|\boldsymbol{x}_i\|^2}{d} - 1\Bigr)^2\biggr) = \frac{n}{d}(\bar{\mu}_4 - 1),$$

and so if $d \gg n$, then $\|\boldsymbol{x}_i\|^2 \approx d$ is a good approximation and one that underlies some of our methods.

We consider several models for heritability. The ***linear*** model is

$$y_i = \sum_{j=1}^d x_{ij}\beta_j + \varepsilon_i, \qquad i = 1, \ldots, n \tag{1}$$

where ε_i is a random error uncorrelated with $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{id})$ having mean 0 and variance σ^2. Effects of the environment, interactions among the SNPs

and interactions between SNPs and the environment are subsumed into ε_i. The pairs $(\boldsymbol{x}_i, y_i)$ are independent.

Interest naturally centers on the vector $\beta = (\beta_1, \ldots, \beta_d)$. Estimation of β is difficult because usually $n \ll d$. If β is sparse, then it can still be estimated using compressed sensing methods. But β might not be sparse and then there are lower bounds (Candès and Davenport, 2011) on its estimation error.

To study linear heritability we are interested in $\sigma^2_{\mathrm{L}} = \sum_{j=1}^d \beta_j^2 = \|\beta\|^2$. If β is not sparse enough to allow an accurate estimate $\hat{\beta}$ of β, then we cannot rely on $\|\hat{\beta}\|^2$ to be a good estimate of σ^2_{L}. Accordingly, we turn our attention to strategies for estimating $\sum_j \beta_j^2$ directly without requiring an accurate estimate of β.

Our starting point is the bias corrected quasi-regression estimator of σ^2_{L} from Owen (2000) given below. That algorithm is able to estimate σ^2_{L} in an asymptotic regime where $n/d^{2/3} \to \infty$, without assuming sparsity of β. That is, the sample size n can be sublinear in d. We choose an asymptotic regime with n and d growing together not because more SNPs will be found, but because the usual fixed d and $n \to \infty$ framework is not a good description for a finite data set with $n \ll d$.

In addition to the linear model, we may also be interested in an additive model which contains x_j^2 terms. If x_j only takes 2 levels, then the centered value $x_j^2 - 1$ is linearly dependent on x_j, and we gain nothing from incorporating it into the linear model. When SNPs are recorded at 3 levels, this additive model allows one to investigate dominance effects. Let

$$z_j = \frac{x_j^2 - \mu_{j3} x_j - 1}{\sqrt{\mu_{j4} - \mu_{j3}^2 - 1}}. \tag{2}$$

Then $\mathbb{E}(z_j) = 0$, $\mathbb{E}(z_j^2) = 1$ and $\mathbb{E}(x_j z_j) = 0$. The ***additive*** model for heritability is

$$y_i = \sum_{j=1}^d x_{ij}\beta_j + \sum_{j=1}^d z_{ij}\gamma_j + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{3}$$

where ε_i contains an error uncorrelated with the x_{ij} and the z_{ij}, and does not ordinarily take the same numerical value in models (1), (3) and others that we consider.

We are also interested in the ***squares*** model

$$y_i = \sum_{j=1}^d z_{ij}\gamma_j + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{4}$$

the pure ***interaction*** model

$$y_i = \sum_{j=1}^{d-1} \sum_{k=j+1}^{d} x_{ij}x_{ik}\beta_{jk} + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{5}$$

and the ***quadratic*** model

$$y_i = \sum_{j=1}^{d} x_{ij}\beta_{jk} + \sum_{j=1}^{d} z_{ij}\gamma_j + \sum_{j=1}^{d-1}\sum_{k=j+1}^{d} x_{ij}x_{ik}\beta_{jk} + \varepsilon_i, \qquad i = 1, \dots, n. \quad (6)$$

Some of these are of direct biological interest and others play a role in our proofs.

Corresponding to the model parameters above are the sums of squares

$$\sigma^2_{\mathrm{L}} = \sum_j \beta_j^2, \qquad \sigma^2_{\mathrm{S}} = \sum_j \gamma_j^2, \quad \text{and} \qquad \sigma^2_{\mathrm{I}} = \sum_{j<k} \beta_{jk}^2,$$

along with combinations $\sigma^2_{\mathrm{A}} = \sigma^2_{\mathrm{L}} + \sigma^2_{\mathrm{S}}$ and $\sigma^2_{\mathrm{Q}} = \sigma^2_{\mathrm{A}} + \sigma^2_{\mathrm{I}}$. These sums of squares take the same value in any model for which all their parts are defined. The following inequality

$$\max(\sigma^2_{\mathrm{L}}, \sigma^2_{\mathrm{S}}, \sigma^2_{\mathrm{I}}) \leqslant \sigma^2_{\mathrm{Q}} \leqslant \mathbb{E}(y^2) \quad (7)$$

will be useful in bounding some error terms. For instance, even though σ_I^2 is the sum of $d(d-1)/2$ nonnegative contributions, that sum is $O(1)$. Also, introducing artificial phenotypes like $y^2 - \mathbb{E}(y^2)$, and applying a version of (7) for them, yields further bounds that we use.

3 Quasi-regression for linear heritability

We begin with estimation of σ^2_{L}, for linear heritability. For $d \gg n$ it is infeasible to estimate β_j by least squares. The quasi-regression estimator of β_j is

$$\widetilde{\beta}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}y_i.$$

We easily find that $\mathbb{E}(\widetilde{\beta}_j) = \beta_j$. The naive estimator of σ^2_{L} is

$$\hat{\sigma}^2_{\mathrm{L}} = \sum_{j=1}^{d} \widetilde{\beta}_j^2.$$

Proposition 1 gives the bias of $\hat{\sigma}^2_{\mathrm{L}}$.

Proposition 1.

$$\mathbb{E}(\hat{\sigma}^2_{\mathrm{L}}) = \frac{n-1}{n}\sigma^2_{\mathrm{L}} + \frac{1}{n}\mathbb{E}\big(\|\boldsymbol{x}\|^2 y^2\big).$$

Proof. First

$$\begin{aligned} n\mathbb{E}(\widetilde{\beta}_j^2) &= \frac{1}{n}\mathbb{E}\Big(\sum_i \sum_{i'} x_{ij}x_{i'j}y_i y_{i'}\Big) \\ &= \mathbb{E}\big(x_j^2 y^2\big) + (n-1)\mathbb{E}(x_j y)^2 \\ &= \mathbb{E}\big(x_j^2 y^2\big) + (n-1)\beta_j^2. \end{aligned}$$

Summing over j yields the result. □

Because $\mathbb{E}(\|\boldsymbol{x}\|^2) = d$, the bias in $\hat{\sigma}^2_{\mathrm{L}}$ is of order d/n, which is severe when $d \gg n$. But this bias is easily removed. Let

$$\widehat{B}_{\mathrm{L}} = \frac{1}{n^2}\sum_{i=1}^{n} \|\boldsymbol{x}_i\|^2 y_i^2. \tag{8}$$

We can obtain a bias corrected estimate

$$\hat{\sigma}^2_{\mathrm{L,BC}} = \frac{n}{n-1}\big(\hat{\sigma}^2_{\mathrm{L}} - \widehat{B}_{\mathrm{L}}\big), \tag{9}$$

because

$$\mathbb{E}\big(\hat{\sigma}^2_{\mathrm{L,BC}}\big) = \frac{n}{n-1}\Big(\frac{n-1}{n}\sigma^2_{\mathrm{L}} + \frac{1}{n}\mathbb{E}\big(\|\boldsymbol{x}\|^2 y^2\big)\Big) - \frac{1}{n-1}\mathbb{E}(\|\boldsymbol{x}\|^2 y^2) = \sigma^2_{\mathrm{L}}.$$

A simple approximation to the bias adjustment can be obtained via the concentration of measure approximation $\|\boldsymbol{x}_i\|^2 \approx d$. Letting,

$$\widetilde{B}_{\mathrm{L}} = \frac{d}{n^2}\sum_{i=1}^{n} y_i^2$$

the resulting estimate is

$$\hat{\sigma}^2_{\mathrm{L,BC,CM}} = \frac{n}{n-1}\big(\hat{\sigma}^2_{\mathrm{L}} - \widetilde{B}_{\mathrm{L}}\big). \tag{10}$$

Proposition 2. *Suppose that the phenotype y satisfies the bound $|y| \leqslant c$. Then*

$$\mathbb{E}\big((\widetilde{B}_{\mathrm{L}} - \widehat{B}_{\mathrm{L}})^2\big) \leqslant c^4\,\bar{\mu}_4\,\frac{d}{n^2}.$$

Proof. First $\widetilde{B}_{\mathrm{L}} - \widehat{B}_{\mathrm{L}} = n^{-2}\sum_{i=1}^{n}(d - \|\boldsymbol{x}_i\|)y_i^2$. Therefore

$$\begin{aligned}
\mathbb{E}\big((\widetilde{B}_{\mathrm{L}} - \widehat{B}_{\mathrm{L}})^2\big) &= \frac{1}{n^3}\mathbb{E}\big((d - \|\boldsymbol{x}\|^2)^2 y^4\big) + \frac{n-1}{n^3}\mathbb{E}\big((d - \|\boldsymbol{x}\|^2)y^2\big)^2 \\
&\leqslant c^4\bar{\mu}_4\frac{d}{n^3} + c^4\frac{1}{n^2}\mathbb{E}\big((d - \|\boldsymbol{x}\|^2)^2\big) \\
&\leqslant c^4\bar{\mu}_4\frac{d}{n^3} + c^4\frac{n-1}{n^3}\,d\bar{\mu}_4.
\end{aligned}$$

□

Proposition 2 shows that the concentration of measure approximation changes our estimate of σ^2_{L} by an amount with root mean square (RMS) $O(d^{1/2}n^{-1}\bar{\mu}_4^{1/2})$, where c is assumed to be constant in n and d for the phenotype of interest. In an asymptotic regime with $\bar{\mu}_4$ bounded we find that the concentration of measure approximation makes a vanishing RMS effect on our estimate if $d = o(n^2)$. Simulations in Owen (2000) showed very little difference between approximations with and without the concentration of measure approximation.

Theorem 1. *Let $\hat\sigma^2_{\mathrm{L}}$ be the naive quasi-regression variance estimate and assume that the phenotype satisfies the bound $|y| \leqslant c$. Then*

$$\mathbb{E}\big((\hat\sigma^2_{\mathrm{L,BC}} - \sigma^2_{\mathrm{L}})^2\big) = \mathrm{Var}(\hat\sigma^2_{\mathrm{L,BC}}) = O\Big(\frac{1}{n} + \frac{d^2}{n^3}\Big).$$

Proof. It suffices to show that $\mathrm{Var}(\hat\sigma^2_{\mathrm{L}} - \widehat{B}_{\mathrm{L}}) = O\big(n^{-1} + d^2n^{-3}\big)$. A lengthy derivation (see Lemma 3 of the appendix) yields

$$\begin{aligned}\mathrm{Var}(\hat\sigma^2_{\mathrm{L}}) &\leqslant \frac{4}{n}c^2\sigma^2_{\mathrm{L}} + \frac{1}{n^2}\Big(4\sigma_{\mathrm{L}}c^3\big(d^2 + d\bar\mu_4\big)^{1/2} + 2(d\bar\mu_4 + 2)\mathbb{E}(y^4)\Big) + \frac{1}{n^3}c^4(d^2 + d\bar\mu_4)\\ &= O(n^{-1} + dn^{-2} + d^2n^{-3}) = O(n^{-1} + d^2n^{-3}).\end{aligned}$$

Next

$$\mathrm{Var}(\widehat{B}_{\mathrm{L}}) = \frac{1}{n^3}\Big(\mathbb{E}(\|\boldsymbol{x}\|^4y^4) - \mathbb{E}\big(\|\boldsymbol{x}\|^2y^2\big)^2\Big) \leqslant \frac{c^4}{n^3}(d^2 + d\bar\mu_4).$$

Thus $\mathrm{Var}(\widehat{B}_{\mathrm{L}}) = O(d^2n^{-3})$ and $\mathrm{Var}(\hat\sigma^2_{\mathrm{L}}) = O(n^{-1} + d^2n^{-3})$, so $\mathrm{Var}(\hat\sigma^2_{\mathrm{L}} - \widehat{B}_{\mathrm{L}}) \leqslant \mathrm{Var}(\hat\sigma^2_{\mathrm{L}}) + \mathrm{Var}(\widehat{B}_{\mathrm{L}}) + 2(\mathrm{Var}(\hat\sigma^2_{\mathrm{L}})\mathrm{Var}(\widehat{B}_{\mathrm{L}}))^{1/2} = O(n^{-1} + d^2n^{-3})$. □

When $d \gg n$, the dominant term in the bound for $\mathrm{Var}(\hat\sigma^2_{\mathrm{L,BC}})$ is $4c^4d^2n^{-3}$. In that case, doubling n reduces the variance by nearly a factor of eight.

The concentration of measure approximation is accurate to within $O(d^{1/2}n^{-1})$. The standard deviation of $\hat\sigma^2_{\mathrm{L,BC}}$ is $O(n^{-1/2} + dn^{-3/2})$. As a result, the bias introduced by concentration of measure is of order $\sqrt{n/d} + \sqrt{d/n^3}$ standard errors. It is thus negligible for the values of n and d used in many genome-wide association studies. Both $\hat\sigma^2_{\mathrm{L,BC}}$ and $\hat\sigma^2_{\mathrm{L,BC,CM}}$ require $O(nd)$ computation but the latter has a smaller implicit constant.

4 Additive and quadratic heritability

The additive model predicts y by a linear combination of x_j and z_j for $j = 1, \ldots, d$. Because $x_1, \ldots, x_d$ are independent, it follows that the entire list $x_1, \ldots, x_d, z_1, \ldots, z_d$ consists of $2d$ mutually uncorrelated predictors. The analysis of the additive model is similar to that of the linear model, except that there are twice as many predictors and the fourth moments of z_j are larger than those of x_j.

The quadratic model incorporates predictors x_jx_k for $j < k$. For distinct pairs $j < k$ and $j' < k'$ with $j \neq j'$ or $k \neq k'$ the predictors x_jx_k and $x_{j'}x_{k'}$ are uncorrelated. Similarly x_jx_k is uncorrelated with all of the $x_{j'}$ and $z_{j'}$ whether or not $j = j'$ or $k = k'$. The big difference in the quadratic model is that there are now more than $d^2/2$ coefficients to estimate instead of $O(d)$ coefficients.

For the additive model, we estimate

$$\widetilde{\beta}_j = \frac{1}{n}\sum_{i=1}^n x_{ij}y_i, \quad \text{and} \quad \widetilde{\gamma}_j = \frac{1}{n}\sum_{i=1}^n z_{ij}y_i,$$

and the naive estimate of σ^2_{A} is

$$\hat{\sigma}^2_{\mathrm{A}} = \sum_{j=1}^{d} (\widetilde{\beta}_j^2 + \widetilde{\gamma}_j^2).$$

Proposition 3.

$$\mathbb{E}(\hat{\sigma}^2_{\mathrm{A}}) = \frac{n-1}{n}\sigma^2_{\mathrm{A}} + \frac{1}{n}\mathbb{E}\big((\|\boldsymbol{x}\|^2 + \|\boldsymbol{z}\|^2)y^2\big).$$

Proof. The same argument used in Proposition 3 applies here. □

Letting

$$\widehat{B}_{\mathrm{A}} = \frac{1}{n^2}\sum_{i=1}^{n}(\|\boldsymbol{x}_i\|^2 + \|\boldsymbol{z}_i\|^2)y_i^2 \quad \text{and} \quad \widetilde{B}_{\mathrm{A}} = \frac{2d}{n^2}\sum_{i=1}^{n} y_i^2,$$

we obtain estimates

$$\hat{\sigma}^2_{\mathrm{A,BC}} = \frac{n}{n-1}\big(\hat{\sigma}^2_{\mathrm{A}} - \widehat{B}_{\mathrm{A}}\big) \quad \text{and}$$
$$\hat{\sigma}^2_{\mathrm{A,BC,CM}} = \frac{n}{n-1}\big(\hat{\sigma}^2_{\mathrm{A}} - \widetilde{B}_{\mathrm{A}}\big).$$

As in the linear case we find that $\mathbb{E}\big(\hat{\sigma}^2_{\mathrm{A,BC}}\big) = \sigma^2_{\mathrm{A}}$. Also from "A = S + L" we get

$$\mathbb{E}\big((\widetilde{B}_{\mathrm{A}} - \widehat{B}_{\mathrm{A}})^2\big) \leqslant c^4\big(\bar{\mu}_4(x) + \bar{\mu}_4(z) + 2\sqrt{\bar{\mu}_4(x)\bar{\mu}_4(z)}\big)\frac{d}{n^2}$$

where $\bar{\mu}_4(x)$ and $\bar{\mu}_4(z)$ are average fourth moments of x_j and z_j respectively.

Theorem 2. *Let $\hat{\sigma}^2_{\mathrm{A}}$ be the naive quasi-regression variance estimate of σ^2_{A} and assume that the phenotype satisfies the bound $|y| \leqslant c$. Then*

$$\mathrm{Var}(\hat{\sigma}^2_{\mathrm{A,BC}}) = O\Big(\frac{1}{n} + \frac{d^2}{n^3}\Big).$$

Proof. The additive variance explained is $\sigma^2_{\mathrm{A}} = \sigma^2_{\mathrm{L}} + \sigma^2_{\mathrm{S}}$, and the estimate partitions accordingly, as $\hat{\sigma}^2_{\mathrm{A,BC}} = \hat{\sigma}^2_{\mathrm{L,BC}} + \hat{\sigma}^2_{\mathrm{S,BC}}$. We can apply Theorem 1 to both parts, letting z_j take the role of x_j in the second application. □

For the quadratic model, we incorporate estimates

$$\widetilde{\beta}_{jk} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}x_{ik}y_i$$

for $1 \leqslant j < k \leqslant d$. The naive estimate of σ^2_{Q} is

$$\hat{\sigma}^2_{\mathrm{Q}} = \sum_{j=1}^{d}(\widetilde{\beta}_j^2 + \widetilde{\gamma}_j^2) + \sum_{j<k}\widetilde{\beta}_{jk}^2.$$

Proposition 4.

$$\mathbb{E}(\hat{\sigma}^2_{\mathrm{Q}}) = \frac{n-1}{n}\sigma^2_{\mathrm{Q}} + \frac{1}{n}\mathbb{E}\big((\|\boldsymbol{x}\|^2 + \|\boldsymbol{z}\|^2 + \|\boldsymbol{x}\|^4)y^2\big).$$

Proof. Recall that $\sigma^2_{\mathrm{Q}} = \sigma^2_{\mathrm{L}} + \sigma^2_{\mathrm{S}} + \sigma^2_{\mathrm{I}}$. To take account of the σ^2_{I} component, write for $j < k$,

$$\begin{aligned} n\mathbb{E}\big(\widetilde{\beta}^2_{jk}\big) &= \frac{1}{n}\sum_i \sum_{i'} \mathbb{E}\big(x_{ij}x_{ik}x_{i'j}x_{i'k}y_i y_{i'}\big) \\ &= \mathbb{E}\big(x_j^2 x_k^2 y^2\big) + (n-1)\beta^2_{jk}. \end{aligned}$$

Summing over $1 \leqslant j < k \leqslant d$, yields

$$\sum_{j<k} \mathbb{E}\big(\widetilde{\beta}^2_{jk}\big) = \frac{1}{n}\mathbb{E}(\|\boldsymbol{x}\|^4 y^2) + \frac{n-1}{n}\sigma^2_{\mathrm{I}}. \quad \square$$

Once again, we get a bias correction and a concentration of measure approximation:

$$\begin{aligned} \widehat{B}_{\mathrm{Q}} &= \frac{1}{n^2}\sum_{i=1}^n (\|\boldsymbol{x}_i\|^4 + \|\boldsymbol{x}_i\|^2 + \|\boldsymbol{z}_i\|^2) y_i^2 \quad \text{and} \\ \widetilde{B}_{\mathrm{Q}} &= \frac{d^2 + (2+\bar{\mu}_4)d}{n^2}\sum_{i=1}^n y_i^2. \end{aligned}$$

The bias is now of much larger order d^2, consistent with there being more than $d^2/2$ parameters to estimate.

Using these expressions, we obtain estimates

$$\begin{aligned} \hat{\sigma}^2_{\mathrm{Q,BC}} &= \frac{n}{n-1}\big(\hat{\sigma}^2_{\mathrm{Q}} - \widehat{B}_{\mathrm{Q}}\big) \quad \text{and} \\ \hat{\sigma}^2_{\mathrm{Q,BC,CM}} &= \frac{n}{n-1}\big(\hat{\sigma}^2_{\mathrm{Q}} - \widetilde{B}_{\mathrm{Q}}\big). \end{aligned}$$

Here $\mathbb{E}\big(\hat{\sigma}^2_{\mathrm{Q,BC}}\big) = \sigma^2_{\mathrm{Q}}$.

In the model equation "$\mathrm{Q} = \mathrm{L} + \mathrm{S} + \mathrm{I}$" we have already seen how the linear and square parts have an accurate concentration of measure approximation. It remains to consider $\widetilde{B}_{\mathrm{I}} - \widehat{B}_{\mathrm{I}}$.

Proposition 5. *For the interaction model, the bias adjustments satisfy*

$$\mathbb{E}\big((\widetilde{B}_{\mathrm{I}} - \widehat{B}_{\mathrm{I}})^2\big) = O\Big(\frac{d^3}{n^2}\Big).$$

Proof.

$$\mathbb{E}\big((\widetilde{B}_{\mathrm{I}} - \widehat{B}_{\mathrm{I}})^2\big) = \frac{1}{n^4}\mathbb{E}\Big(\Big(\sum_{i=1}^n (d^2 + d\bar{\mu}_4 - \|\boldsymbol{x}_i\|^4) y_i^2\Big)^2\Big)$$

$$= \frac{1}{n^3}\mathbb{E}\Big((d^2 + d\bar{\mu}_4 - \|\boldsymbol{x}\|^4)^2 y^4\Big) + \frac{n-1}{n^3}\mathbb{E}\Big((d^2 + d\bar{\mu}_4 - \|\boldsymbol{x}\|^4) y^2\Big)^2$$
$$= \frac{c^4}{n^2}\mathrm{Var}(\|\boldsymbol{x}\|^4)$$
$$= O\Big(\frac{d^3}{n^2}\Big). \qquad \square$$

Theorem 3. *Let $\hat{\sigma}^2_{\mathrm{Q}}$ be the naive quasi-regression variance estimate of σ^2_{Q} and assume that the phenotype satisfies the bound $|y| \leqslant c$. Then*

$$\mathrm{Var}(\hat{\sigma}^2_{\mathrm{Q,BC}}) = O\Big(\frac{1}{n} + \frac{d^4}{n^3}\Big).$$

Proof. The error from the additive parts is already seen to be $O(1/n + d^2/n^3)$. Therefore it suffices to show that $\mathrm{Var}(\hat{\sigma}^2_{\mathrm{I,BC}}) = O(1/n + d^4/n^3)$. This is exactly what we would expect from a linear model with $d^* = d(d-1)/2$ independent predictors. But the predictors in the interaction model, while uncorrelated, are not independent.

We can follow the argument for the linear case using d^* predictors of the form $x_j x_k$ for $j < k$ in Lemma 3, replacing each x_k by $x_r x_s$ for $r < s$ and then each x_j by $x_j x_k$ for $j < k$ and each $\boldsymbol{x}$ by $\boldsymbol{x}^* = (x_1x_2,\, x_1x_3,\, \ldots,\, x_{d-1}x_d) \in \mathbb{R}^{d^*}$. We then use $\mathbb{E}(x_j x_k y) = \beta_{jk}$ and replace ε_{L} by $\varepsilon_{\mathrm{I}} = y - \sum_{j<k} x_j x_k \beta_{jk}$.

The analogy holds in a straightforward way except for the term which becomes

$$\frac{n-1}{n^3}\sum_{j<k}\sum_{r<s}\mathbb{E}\big(x_j x_k x_r x_s y^2\big)^2$$

which now involves four way sums over components of $\boldsymbol{x}$. We need this to be $O(d^*/n^2) = O(d^2/n^2)$. If we use the artificial phenotype $y^2 - \mathbb{E}(y^2)$ and consider a pure quartic model with $d(d-1)(d-2)(d-3)/24$ predictors $x_j x_k x_r x_s$ with j, k, r, and s all distinct, then the terms with no ties among j, k, r, s sum to at most $\mathbb{E}(y^2)^2$. What remains are the terms where $\{j,k\} \cap \{r,s\} \neq \varnothing$. There are $O(d^2)$ such terms and they are uniformly bounded. $\square$

Thus while it is possible to estimate the interaction inheritance and hence the quadratic inheritance with $n \ll d^2$, we still need n of larger order than d. Practically, this implies that we may be able to investigate interactions among strategically chosen subsets of SNPs but perhaps not the entire interaction at practical sample sizes.

The concentration of measure approximation for the interaction inheritance is $\|\boldsymbol{x}\|^4 \approx d^2 + d\bar{\mu}_4$. The root mean square change from this shortcut is $d^{3/2}/n$, while the standard deviation of the estimate is $O(n^{-1/2} + d^2 n^{-3/2})$. Suppose that n is just barely large enough to allow estimation of σ^2_{I}, perhaps $n \propto d^{4/3+\epsilon}$. Then the RMS difference between $\hat{\sigma}^2_{\mathrm{I,BC}}$ and $\hat{\sigma}^2_{\mathrm{I,BC,CM}}$ is of order $d^{3/2}n^{-1} = d^{1/6-\epsilon}$ which diverges. As a result, the concentration of measure shortcut is not recommended for the interaction or quadratic model.

Acknowledgments

I thank Or Zuk for helpful discussions. This work was supported by the NSF under grant DMS-0906056.

References

Candès, E. and Davenport, M. A. (2011). Howe well can we estimate a sparse vector? Technical report, Stanford University.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S. Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., Mackay, T. F. C., McCarroll, S. A., and Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, 461:747–753.

Owen, A. B. (2000). Assessing linearity in high dimensions. *The Annals of Statistics*, 28(1):1–19.

Zhao, P. and Yu, B. (2006). On model selection consistency of the lasso. *Journal of Machine Learning Research*, 7:2541–2563.

Appendix

Here we prove some basic Lemmas used in the main Theorem. The following basic properties of $\|\boldsymbol{x}\|^2$ are used below:

$$\mathbb{E}\big(\|\boldsymbol{x}\|^2\big) = d, \qquad \mathbb{E}\big(\|\boldsymbol{x}\|^4\big) = d^2 + d\bar{\mu}_4$$
$$\mathbb{E}\big((\|\boldsymbol{x}\|^2 - d)^2\big) = d\bar{\mu}_4, \quad \text{and} \quad \mathbb{E}\big((\|\boldsymbol{x}\|^2 - d)^r\big) = O(d^{r-1}), \quad \text{integer } r \geqslant 3.$$

Lemma 1 uses the artificial phenotype $y^2 - \mathbb{E}(y^2)$ to bound a quantity which appears in $\mathrm{Var}(\hat{\sigma}^2_{\mathrm{L}})$.

Lemma 1.
$$\sum_j \sum_k \mathbb{E}\big(x_j x_k y^2\big)^2 \leqslant (d\bar{\mu}_4 + 2)\mathbb{E}(y^4).$$

Proof. We write
$$\sum_j \sum_k \mathbb{E}\big(x_j x_k y^2\big)^2 = \sum_j \sum_{k \neq j} \mathbb{E}\big(x_j x_k y^2\big)^2 + \sum_j \mathbb{E}\big(x_j^2 y^2\big)^2.$$
For the first term, consider an artificial phenotype $y^2 - \mathbb{E}(y^2)$. Then
$$\sum_j \sum_{k \neq j} \mathbb{E}\big(x_j x_k y^2\big)^2 = 2 \sum_{j<k} \mathbb{E}\big(x_j x_k (y^2 - \mathbb{E}(y^2))\big)^2$$

$$\begin{aligned}
&\leqslant 2\mathbb{E}\big((y^2 - \mathbb{E}(y^2))^2\big) \\
&\leqslant 2\mathbb{E}(y^4),
\end{aligned}$$

because of inequality (7) applied to the interaction model for this artificial phenotype. Next

$$\sum_j \mathbb{E}(x_j^2 y^2)^2 \leqslant \sum_j \mathbb{E}(x_j^4)\mathbb{E}(y^4) \leqslant \mathbb{E}(y^4) d\bar{\mu}_4. \qquad \square$$

Lemma 2.

$$\mathbb{E}\big((y-\varepsilon)\|\boldsymbol{x}\|^2 y^3\big) \leqslant \sigma_{\mathrm{L}} c^3 \big(d^2 + d\bar{\mu}_4\big)^{1/2}.$$

Proof. Using Cauchy-Schwarz, boundedness of y, and the above results on $\|\boldsymbol{x}\|$:

$$\begin{aligned}
\mathbb{E}\big((y-\varepsilon)\|\boldsymbol{x}\|^2 y^3\big) &\leqslant \mathbb{E}\big((y-\varepsilon)^2\big)^{1/2} \mathbb{E}\big(\|\boldsymbol{x}\|^4 y^6\big)^{1/2} \\
&\leqslant \sigma_{\mathrm{L}} c^3 \big(d^2 + d\bar{\mu}_4\big)^{1/2}. \qquad \square
\end{aligned}$$

Lemma 3 is the main lemma.

Lemma 3. *Under the conditions of Theorem 1*

$$\mathrm{Var}(\hat{\sigma}_{\mathrm{L}}^2) \leqslant \frac{4}{n} c^2 \sigma_{\mathrm{L}}^2 + \frac{1}{n^2}\Big(4\sigma_{\mathrm{L}} c^3 \big(d^2 + d\bar{\mu}_4\big)^{1/2} + 2(d\bar{\mu}_4 + 2)\mathbb{E}(y^4)\Big) + \frac{1}{n^3} c^4 (d^2 + d\bar{\mu}_4).$$

Proof. First,

$$\mathrm{Var}\big(\hat{\sigma}_{\mathrm{L}}^2\big) = \sum_j \sum_k \mathbb{E}\big(\widetilde{\beta}_j^2 \widetilde{\beta}_k^2\big) - \Big(\mathbb{E}\Big(\sum_j \widetilde{\beta}_j^2\Big)\Big)^2$$

We simplify some expressions below using $\mathbb{E}\big(x_j y\big) = \beta_j$ and $\varepsilon_{\mathrm{L}} \equiv y - \sum_j x_j \beta_j$.

For each pair $j, k \in \{1, 2, \ldots, d\}$,

$$\begin{aligned}
\mathbb{E}(\widetilde{\beta}_j^2 \widetilde{\beta}_k^2) &= \mathbb{E}\Big(\Big(\frac{1}{n}\sum_i x_{ij} y_i\Big)^2 \Big(\frac{1}{n}\sum_i x_{ik} y_i\Big)^2\Big) \\
&= \frac{1}{n^4} \sum_i \sum_{i'} \sum_{i''} \sum_{i'''} \mathbb{E}\big(x_{ij} x_{i'j} x_{i''k} x_{i'''k} y_i y_{i'} y_{i''} y_{i'''}\big) \\
&= \frac{(n-1)(n-2)(n-3)}{n^3} \mathbb{E}(x_j y)^2 \mathbb{E}(x_k y)^2 \\
&\quad + \frac{(n-1)(n-2)}{n^3}\Big(\mathbb{E}(x_j y)^2 \mathbb{E}(x_k^2 y^2) + \mathbb{E}(x_j^2 y^2)\mathbb{E}(x_k y)^2 + 4\mathbb{E}(x_j x_k y^2)\mathbb{E}(x_j y)\mathbb{E}(x_k y)\Big) \\
&\quad + \frac{n-1}{n^3}\Big(2\mathbb{E}\big(x_j y\big)\mathbb{E}\big(x_j x_k^2 y^3\big) + 2\mathbb{E}\big(x_k y\big)\mathbb{E}\big(x_k x_j^2 y^3\big) \\
&\qquad\quad + \mathbb{E}\big(x_j^2 y^2\big)\mathbb{E}\big(x_k^2 y^2\big) + 2\mathbb{E}\big(x_j x_k y^2\big)^2\Big) \\
&\quad + \frac{1}{n^3}\mathbb{E}\big(x_j^2 x_k^2 y^4\big).
\end{aligned}$$

Summing over j and k,

$$\begin{aligned}\sum_j \sum_k \mathbb{E}(\widetilde{\beta}_j^2 \widetilde{\beta}_k^2) = \frac{(n-1)(n-2)(n-3)}{n^3}\sigma_{\mathrm{L}}^4 \\ + \frac{(n-1)(n-2)}{n^3}\Big(2\sigma_{\mathrm{L}}^2 \mathbb{E}\big(\|\boldsymbol{x}\|^2 y^2\big) + 4\mathbb{E}\big((y-\varepsilon_{\mathrm{L}})^2 y^2\big)\Big) \\ + \frac{n-1}{n^3}\Big(4\mathbb{E}\big((y-\varepsilon_{\mathrm{L}})\|\boldsymbol{x}\|^2 y^3\big) + \mathbb{E}(\|\boldsymbol{x}\|^2 y^2)^2 + 2\sum_j \sum_k \mathbb{E}(x_j x_k y^2)^2\Big) \\ + \frac{1}{n^3}\mathbb{E}\big(\|\boldsymbol{x}\|^4 y^4\big).\end{aligned}$$

Next $\mathbb{E}(\widetilde{\beta}_j^2) = \mathbb{E}(x_j^2 y^2)/n + (n-1)\beta_j^2/n$. Therefore

$$\begin{aligned}\Big(\mathbb{E}\Big(\sum_j \widetilde{\beta}_j{}^2\Big)\Big)^2 &= \frac{1}{n^2}\sum_j \sum_k \mathbb{E}\big(x_j^2 y^2\big)\mathbb{E}\big(x_k^2 y^2\big) \\ &\quad + \frac{(n-1)^2}{n^2}\sum_j \sum_k \beta_j^2 \beta_k^2 + 2\frac{n-1}{n^2}\sum_j \sum_k \mathbb{E}\big(x_j^2 y^2\big)\beta_k^2 \\ &= \frac{1}{n^2}\mathbb{E}\big(\|\boldsymbol{x}\|^2 y^2\big)^2 + \frac{(n-1)^2}{n^2}\sigma_{\mathrm{L}}^4 + 2\frac{n-1}{n^2}\mathbb{E}\big(\|\boldsymbol{x}\|^2 y^2\big)\sigma_{\mathrm{L}}^2.\end{aligned}$$

Subtracting,

$$\begin{aligned}\mathrm{Var}(\hat{\sigma}_{\mathrm{L}}^2) = \Big(\frac{(n-1)(n-2)(n-3)}{n^3} - \frac{(n-1)^2}{n^2}\Big)\sigma_{\mathrm{L}}^4 \\ + 2\Big(\frac{(n-1)(n-2)}{n^3} - \frac{n-1}{n^2}\Big)\mathbb{E}\big(\|\boldsymbol{x}\|^2 y^2\big)\sigma_{\mathrm{L}}^2 \\ + \Big(\frac{n-1}{n^3} - \frac{1}{n^2}\Big)\mathbb{E}\big(\|\boldsymbol{x}\|^2 y^2\big)^2 \\ + 4\frac{(n-1)(n-2)}{n^3}\mathbb{E}\big((y-\varepsilon_{\mathrm{L}})^2 y^2\big) \\ + \frac{n-1}{n^3}\Big(4\mathbb{E}\big((y-\varepsilon_{\mathrm{L}})\|\boldsymbol{x}\|^2 y^3\big) + 2\sum_j \sum_k \mathbb{E}(x_j x_k y^2)^2\Big) \\ + \frac{1}{n^3}\mathbb{E}\big(\|\boldsymbol{x}\|^4 y^4\big).\end{aligned}$$

The first three terms above are less than or equal to zero. Therefore

$$\begin{aligned}\mathrm{Var}(\hat{\sigma}_{\mathrm{L}}^2) \leqslant 4\frac{(n-1)(n-2)}{n^3}\mathbb{E}\big((y-\varepsilon_{\mathrm{L}})^2 y^2\big) \\ + \frac{n-1}{n^3}\Big(4\mathbb{E}\big((y-\varepsilon_{\mathrm{L}})\|\boldsymbol{x}\|^2 y^3\big) + 2\sum_j \sum_k \mathbb{E}(x_j x_k y^2)^2\Big) \\ + \frac{1}{n^3}\mathbb{E}\big(\|\boldsymbol{x}\|^4 y^4\big).\end{aligned}$$

We can simplify this bound using $\mathbb{E}\big((y-\varepsilon_{\mathrm{L}})^2 y^2\big) \leqslant c^2\sigma_{\mathrm{L}}^2$, Lemmas 1 and 2, and $\mathbb{E}\big(\|\boldsymbol{x}\|^4 y^4\big) \leqslant c^4(d^2 + d\bar{\mu}_4)$, yielding

$$\begin{aligned}
\mathrm{Var}(\hat{\sigma}_{\mathrm{L}}^2) &\leqslant 4\frac{(n-1)(n-2)}{n^3}c^2\sigma_{\mathrm{L}}^2 \\
&\quad + \frac{n-1}{n^3}\Big(4\sigma_{\mathrm{L}}c^3\big(d^2 + d\bar{\mu}_4\big)^{1/2} + 2(d\bar{\mu}_4 + 2)\mathbb{E}(y^4)\Big) + \frac{1}{n^3}c^4(d^2 + d\bar{\mu}_4) \\
&\leqslant \frac{4}{n}c^2\sigma_{\mathrm{L}}^2 + \frac{1}{n^2}\Big(4\sigma_{\mathrm{L}}c^3\big(d^2 + d\bar{\mu}_4\big)^{1/2} + 2(d\bar{\mu}_4 + 2)\mathbb{E}(y^4)\Big) + \frac{1}{n^3}c^4(d^2 + d\bar{\mu}_4).
\end{aligned}$$

□