The piranha problem: Large effects swimming in a small pond

Christopher Tosh[*], Philip Greengard[†], Ben Goodrich[‡], Andrew Gelman[§], Aki Vehtari[¶], and Daniel Hsu[‖]

8 Apr 2021

Abstract

In some scientific fields, it is common to have certain variables of interest that are of particular importance and for which there are many studies indicating a relationship with a different explanatory variable. In such cases, particularly those where no relationships are known among explanatory variables, it is worth asking under what conditions it is possible for all such claimed effects to exist simultaneously. This paper addresses this question by reviewing some theorems from multivariate analysis that show, unless the explanatory variables also have sizable effects on each other, it is impossible to have many such large effects. We also discuss implications for the replication crisis in social science.

1   Introduction

This paper reviews a few mathematical theorems with implications for quantitative social science. Identifying and measuring the effects of explanatory variables are central problems in statistics and drive much of the world's scientific research. Despite the substantial effort spent on these tasks, there has been comparatively little work on addressing a related question: how many explanatory variables can have large effects on an outcome? An answer to this question may be helpful in assessing new claims.

Consider, by way of example, the problem of explaining voters' behaviors and choices. A multitude of researchers have identified and tested the effects of internal factors such as fear, hope, pride, anger, anxiety, depression, and menstrual cycles (Parker and Isbell, 2010; Ladd and Lenz, 2011; Obschonka et al., 2018; Durante et al., 2013), as well external factors such as droughts, shark attacks, and the performance of local college football games (Achen and Bartels, 2002; Healy et al., 2010; Fowler and Hall, 2018; Fowler and Montagnes, 2015). Many of these particular findings have been questioned on methodological grounds (Fowler and Montagnes, 2015; Fowler and Hall, 2018; Clancy, 2012; Gelman, 2015), but beyond the details of these particular studies, it is natural to ask if all of these effects are actually real in the sense of representing patterns that will consistently appear in the future.

The implication of the claims regarding ovulation and voting, shark attacks and voting, college football and voting, etc., is not merely that some voters are superficial and fickle. No, these papers claim that seemingly trivial or irrelevant factors have *large and consistent* effects, and this runs into the problem of interactions. For example, the effect on your vote of the local college football team losing could depend crucially on whether there's been a shark attack lately, or on what's up with your hormones on election day. Or the effect could be positive in an election with a female candidate and negative in an election with a male candidate. Or the effect could interact with your parents' socioeconomic status, or whether your child

---

[*]Data Science Institute, Columbia University, New York.
[†]Department of Statistics, Columbia University, New York.
[‡]Department of Political Science, Columbia University, New York.
[§]Department of Statistics and Department of Political Science, Columbia University, New York.
[¶]Department of Computer Science, Aalto University, Espoo, Finland.
[‖]Department of Computer Science, Columbia University, New York.

is a boy or a girl, or the latest campaign ad, or any of the many other factors that have been studied in the evolutionary psychology and political psychology literatures. Again, we are not saying that psychological factors have no effect on social, political, or economic decision making; we are only arguing that such effects, if large, will necessarily interact in complex ways. Similar reasoning has been used to argue against naive assumptions of causal identification in economics, where there is a large literature considering rainfall as an instrumental variable, without accounting for the implication that these many hypothesized causal pathways would, if taken seriously, represent violations of the assumption of exclusion restriction (Mellon, 2020).

In this work, we demonstrate that there is an inevitable consequence of having many explanatory variables with large effects: the explanatory variables must have large effects on each other. We call this type of result a "piranha theorem" (Gelman, 2017), the analogy being the folk wisdom that if one has a large number of piranhas (representing large effects) in a single fish tank, then one will soon be left with far fewer piranhas (Anonymous, 2021). If there is some outcome on which a large number of studies demonstrate an effect of a novel explanatory variable, then we can conclude that either some of the claimed effects are smaller than claimed, or some of the explanatory variables are essentially measuring the same phenomenon.

There are a multitude of ways to capture the dependency of random variables, and thus we should expect there to be a correspondingly large collection of piranha theorems. We formalize and prove piranha theorems for correlation, regression, and mutual information in Sections 2 and 3. These theorems illustrate the general phenomena at work in any setting with multiple causal or explanatory variables. In Section 4, we examine typical correlations in a finite sample under a simple probabilistic model.

At first glance, it may be somewhat surprising that one can rule out the possibility of many independent variables having large effects on a single outcome. Indeed, there is nothing functionally, or causally, that excludes such dependencies. Statistically speaking, however, it is a different story. To see the difference, consider the following (somewhat silly) example of voting behavior. There are p independent and identically distributed Bernoulli(0.5) random variables, standing for truly independent causes such as whether or not there was a shark attack that day or whether or not the local college football team won that week. Now, say that a voter votes for the Democrat if exactly an even number of these causes occur, and they vote Republican otherwise. Functionally, there is a strong dependence of voting behavior on each of the causes: for any fixed assignment of causes, flipping the value of any one of these will completely flip the voter's choice. However, for statistical notions of effect such as correlation or regression coefficients, it is a completely different story. Indeed, the correlation between voting behavior and any one of the causes is zero. Moreover, it is also easily verified that the coefficient vector obtained by performing a linear regression of voting behavior on these causes is the all-zero vector. Thus, there are situations where functional and causal dependencies can coexist with a lack of statistical effects; and the upshot of our current paper is that when those causal influences are independent of each other, then they cannot all induce large statistical effects.

Our results are partly motivated by the replication crisis, which refers to the difficulties that many have had in trying to independently verify established findings in social and biological sciences (Ioannidis, 2005). Many of the explanations for the crisis have focused on various methodological issues, such as researcher degrees of freedom (Simmons et al., 2011), underpowered studies (Button et al., 2013), and data dredging (Head et al., 2015). In some cases, solutions to these issues have also been proposed, notably good practice guidelines for authors and reviewers (Simmons et al., 2011) and preregistration of studies (Miguel et al., 2014). Beyond the criticisms of practice and suggested fixes, these works have also provided much needed statistical intuition.

While the current work does not directly address the replication crisis, it gives reason to be suspicious of certain types of results. Groups of studies that claim to have found a variety of important explanatory variables for a single outcome should be scrutinized, particularly when the dependencies among the explanatory variables has not been investigated.

The research reviewed here is related to, but different from, the cluster of ideas corresponding to multiple comparisons, false discovery rates, and multilevel models. Those theories correspond to statistical inference in the presence of some specified distribution of effects, possibly very few nonzero effects (the needle-in-a-haystack problem) or possibly an entire continuous distribution, but without necessarily any concern about how these effects interact. Here we consider deterministic or probabilistic constraints on the distribution of

effects, with the common theme that only in very unusual settings is it possible to have many large effects coexisting.

2 Piranha theorems for correlation and linear regression

In this section, we present piranha theorems for two different ways of measuring linear effects. The first of these, correlation, is straightforward to interpret. We will show that it is impossible for a large number of explanatory variables to be correlated with some outcome variable unless they are highly correlated with each other. Our second piranha theorem examines linear regression coefficients. In particular, we will show that if a set of explanatory random variables is plugged into a regression equation, the ℓ_2-norm $\|\beta\|$ of the least squares coefficient vector β can be bounded above in terms of (the eigenvalues of) the second-moment matrix. Thus, there can only be so many individual coefficients with a large magnitude.

2.1 Correlation

The first type of pattern we consider is correlation,

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\,\text{var}(Y)}}.$$

In particular, we will show that if all the covariates are highly correlated with some outcome variable, then there must be a reasonable amount of correlation among the covariates themselves. This is formalized in the following theorem, which is known as Van der Corput's inequality (Tao, 2014). We offer a proof here for completeness.

Theorem 1 (Van der Corput's inequality)**.** *If $X_1, \ldots, X_p, Y$ are real-valued random variables with finite non-zero variance, then*

$$\sum_{i=1}^{p} |\text{corr}(X_i, Y)| \leq \sqrt{p + \sum_{i \neq j} |\text{corr}(X_i, X_j)|}.$$

In particular, if $|\text{corr}(X_i, Y)| \geq \tau$ for each $i = 1, \ldots, p$, then $\sum_{i \neq j} |\text{corr}(X_i, X_j)| \geq p(\tau^2 p - 1)$.

Proof. Without loss of generality, we may assume that $X_1, \ldots, X_p, Y$ have mean zero and unit variance. Define $Z_1, \ldots, Z_p$ by

$$Z_i = \begin{cases} X_i & \text{if } \mathbb{E}(YX_i) > 0, \\ -X_i & \text{else.} \end{cases}$$

Thus $\mathbb{E}(YZ_i) = |\mathbb{E}(YX_i)|$ and $\mathbb{E}(Z_i^2) = \mathbb{E}(X_i^2)$ for each $i = 1, \ldots, p$. By Cauchy-Schwarz,

$$\sum_{i=1}^{p} \mathbb{E}(YZ_i) = \mathbb{E}\left(Y \sum_{i=1}^{p} Z_i\right) \leq \sqrt{\mathbb{E}\left(\left(\sum_{i=1}^{p} Z_i\right)^2\right)}.$$

Therefore,

$$\sum_{i=1}^{p} |\mathbb{E}(YX_i)| = \sum_{i=1}^{p} \mathbb{E}(YZ_i) \leq \sqrt{\sum_{i=1}^{p} \mathbb{E}(Z_i^2) + \sum_{i \neq j} \mathbb{E}(Z_i Z_j)} \leq \sqrt{p + \sum_{i \neq j} |\mathbb{E}(X_i X_j)|}.$$

Rearranging gives us the theorem statement. $\qquad \square$

A direct consequence of Theorem 1 is that if $X_1, \ldots, X_p$ are independent random variables and have correlation at least τ with Y, then $p \leq 1/\tau^2$.

In some situations, the outcome variable may change from study to study. Some studies may look at the effect of a priming technique on mood, while others may look at at a different priming technique on life outlook. Although mood and life outlook are not exactly the same, we might reasonably expect them to be highly correlated. However, if we have mean-zero and unit-variance random variables X, Y, Z satisfying $\mathbb{E}(XY) \geq \tau$ and $\mathbb{E}(YZ) \geq 1 - \epsilon$, then

$$\mathbb{E}(XZ) = \mathbb{E}(X(Z - Y + Y)) \geq \tau + \mathbb{E}(X(Z - Y)),$$

and by Cauchy-Schwarz, we have

$$\mathbb{E}(X(Z - Y))^2 \leq \mathbb{E}(X^2)\mathbb{E}((Z - Y)^2) \leq 2 - 2(1 - \epsilon).$$

Thus, $\mathbb{E}(XZ) \geq \tau - \sqrt{2\epsilon}$. This gives the following corollary of Theorem 1.

Corollary 2. *Suppose $X_1, Y_1, \ldots, X_p, Y_p$ are real-valued random variables with finite non-zero variance. If $\mathrm{corr}(Y_i, Y_j) \geq 1 - \epsilon$ and $|\mathrm{corr}(X_i, Y_i)| \geq \tau$ for $i, j = 1, \ldots, p$, then $\sum_{i \neq j} |\mathrm{corr}(X_i, X_j)| \geq p((\tau - \sqrt{2\epsilon})^2 p - 1)$.*

The bound in Theorem 1 is essentially tight for large p. To see this, pick any $0 \leq \tau \leq 1$, and take $X_1, \ldots, X_p$ to be mean-zero random variables with covariance matrix Σ given by

$$\Sigma_{ij} = \begin{cases} 1 & \text{if } i = j, \\ \tau^2 & \text{if } i \neq j. \end{cases}$$

If $Y = \sum_{j=1}^{p} X_j$, then for each $i = 1, \ldots, p$,

$$\mathrm{corr}(X_i, Y) = \frac{\mathbb{E}\left(X_i \sum_{j=1}^{p} X_j\right)}{\sqrt{\mathbb{E}\left(\sum_{j,k} X_j X_k\right)}} = \frac{1 + (p-1)\tau^2}{\sqrt{p + p(p-1)\tau^2}} \xrightarrow{p \to \infty} \tau.$$

2.2 Linear regression

We next turn to showing that least squares linear regression solutions cannot have too many large coefficients. Specifically, letting $\beta = (\beta_1, \ldots, \beta_p)^\intercal \in \mathbb{R}^p$ denote the regression coefficients of least squared error,

$$\beta = \underset{\alpha = (\alpha_1, \ldots, \alpha_p)^\intercal \in \mathbb{R}^p}{\mathrm{argmin}} \mathbb{E}\left((\alpha_1 X_1 + \cdots + \alpha_p X_p - Y)^2\right), \tag{1}$$

we bound the number of β_i's that can have large magnitude. This is formalized in our next piranha theorem.

Theorem 3. *Suppose $X_1, \ldots, X_p, Y$ are real-valued random variables with mean zero and unit variance. If $\beta \in \mathbb{R}^p$ satisfies equation (1), then the squared ℓ_2 norm of β satisfies*

$$\|\beta\|^2 \leq \frac{1}{\lambda_{\min}},$$

where $\lambda_{\min}$ is the minimum eigenvalue of the second-moment matrix $\mathbb{E}(XX^\intercal)$ of $X = (X_1, \ldots, X_p)^\intercal$.

Consider again the setting where $X_1, \ldots, X_p$ are independent. In this case, the second-moment matrix $\mathbb{E}(XX^\intercal)$ will be the identity matrix, and its minimum eigenvalue will be 1. Thus, Theorem 1 states for independent covariates, there may be at most $1/\tau^2$ regression coefficients β_i with magnitude larger than τ.

The proof of Theorem 3 relies on the following technical lemma, essentially a consequence of orthogonality.

4

Lemma 4. *If $U_1, \ldots, U_p, Y$ are real-valued random variables with mean zero and unit variance such that $\mathbb{E}(U_i U_j) = 0$ for all $i \neq j$, then*

$$\sum_{i=1}^{p} \left(\mathbb{E} U_i Y \right)^2 \; \leq \; 1.$$

Proof. Denote the covariance matrix of the random vector $(U_1, \ldots, U_p, Y)^\mathsf{T}$ as

$$\Sigma \; = \; \begin{pmatrix} I & a \\ a^\mathsf{T} & 1 \end{pmatrix},$$

where $a_i = \mathbb{E}\left(U_i Y \right)$ for $i = 1, \ldots, p$. Define the vector $v = (-a^\mathsf{T}, \|a\|)^\mathsf{T} \in \mathbb{R}^{p+1}$. Then

$$v^\mathsf{T} \Sigma v \; = \; 2(1 - \|a\|) \|a\|^2 \; \geq \; 0,$$

where the inequality follows from the fact that Σ is a covariance matrix and hence positive semi-defininte. We conclude that $\|a\| \leq 1$. $\qquad \square$

With the above in hand, we turn to the main proof of this section.

Proof of Theorem 3. The case where $\lambda_{\min} = 0$ is trivial. Thus, assume $\lambda_{\min} > 0$. In this case, the second-moment matrix $\mathbb{E}(XX^\mathsf{T})$ is invertible, its inverse has eigenvalues bounded above by $1/\lambda_{\min}$, and

$$\beta \; = \; (\mathbb{E}(XX^\mathsf{T}))^{-1} \mathbb{E}(YX).$$

Define $\tilde{X} = (\mathbb{E}(XX^\mathsf{T}))^{-1/2} X$, so $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_p)^\mathsf{T}$ is a vector of mean-zero and unit variance random variables with $\mathbb{E}(\tilde{X}_i \tilde{X}_j) = 0$ for all $i \neq j$. By Lemma 4,

$$\|\mathbb{E}(Y\tilde{X})\|^2 \; = \; \sum_{j=1}^{p} \mathbb{E}(Y\tilde{X}_j)^2 \; \leq \; 1.$$

Therefore

$$\|\beta\|^2 \; = \; \|(\mathbb{E}(XX^\mathsf{T}))^{-1/2} \mathbb{E}(Y\tilde{X})\|^2 \; = \; \mathbb{E}(Y\tilde{X})^\mathsf{T}(\mathbb{E}(XX^\mathsf{T}))^{-1} \mathbb{E}(Y\tilde{X}) \; \leq \; \frac{1}{\lambda_{\min}} \|\mathbb{E}(Y\tilde{X})\|^2 \; \leq \; \frac{1}{\lambda_{\min}},$$

where the first inequality uses the upper-bound of $1/\lambda_{\min}$ on the eigenvalues of $(\mathbb{E}(XX^\mathsf{T}))^{-1}$. $\qquad \square$

3 A piranha theorem for mutual information

Though many statistical analyses hinge on discovering linear relations among variables, not all do. Thus, we turn to a more general form of dependency for random variables: *mutual information*. Our mutual information piranha theorem will be of a similar form as the above results, namely that if many covariates share information with a common variable, then they must share information among themselves.

To simplify our analysis, we assume that all the random variables we consider in this section take values in discrete spaces. For two random variables X and Y, their mutual information is defined as

$$I(X; Y) \; = \; H(X) - H(X \mid Y) \; = \; H(Y) - H(Y \mid X),$$

where $H(\cdot)$ and $H(\cdot \mid \cdot)$ denote entropy and conditional entropy, respectively. These are defined as

$$H(X) \; = \; \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)},$$

$$H(Y \mid X) \; = \; \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x, y) \log \frac{p(x)}{p(x, y)},$$

where $\mathcal{X}$ (resp. $\mathcal{Y}$) is the range of X (resp. Y), $p(x, y)$ is the joint probability mass function of X and Y, and $p(x)$ is the marginal probability mass function of X.

We use the following facts about entropy and conditional entropy.

Fact (Chain rule of entropy). *For random variables $X_1, \ldots, X_p$,*

$$0 \leq H(X_1, \ldots, X_p) = \sum_{i=1}^{p} H(X_i \mid X_1, \ldots, X_{i-1}).$$

Moreover, we also have for any other random variable Y,

$$0 \leq H(X_1, \ldots, X_p \mid Y) = \sum_{i=1}^{p} H(X_i \mid Y, X_1, \ldots, X_{i-1}).$$

Fact (Conditioning reduces entropy). *For random variables X, Y, Z,*

$$H(X|Y, Z) \leq H(X \mid Y) \leq H(X).$$

Using these facts, we may prove the following mutual information piranha theorem.

Theorem 5. *Given random variables $X_1, \ldots, X_p$ and Y, we have*

$$\sum_{i=1}^{p} I(X_i; Y) \leq H(Y) + \sum_{i=1}^{p} I(X_i; X_{-i}),$$

where $X_{-i} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_p)$.

Proof. Using the definition of mutual information, we have

$$H(X_i \mid X_{-i}) \geq H(X_i) - I(X_i; X_{-i}).$$

Since conditioning reduces entropy, this implies

$$H(X_i \mid X_1, \ldots, X_{i-1}) \geq H(X_i \mid X_{-i}) = H(X_i) - I(X_i; X_{-i}).$$

Thus, we have by the chain rule of entropy

$$H(X_1, \ldots, X_p) = \sum_{i=1}^{p} H(X_i \mid X_1, \ldots, X_{i-1}) \geq \sum_{i=1}^{p} H(X_i) - I(X_i; X_{-i}). \tag{2}$$

The chain rule of entropy combined with the fact that conditioning reduces entropy implies

$$H(X_1, \ldots, X_p \mid Y) \leq \sum_{i=1}^{p} H(X_i \mid Y). \tag{3}$$

Plugging equations (2) and (3) into our formula for $I(X_1, \ldots, X_p; Y)$ gives us

$$
\begin{aligned}
I(X_1, \ldots, X_p; Y) &= H(X_1, \ldots, X_p) - H(X_1, \ldots, X_p \mid Y) \\
&\geq \sum_{i=1}^{p} H(X_i) - I(X_i; X_{-i}) - H(X_i \mid Y) \\
&= \sum_{i=1}^{p} I(X_i; Y) - I(X_i; X_{-i}).
\end{aligned}
$$

Now, we can also write

$$I(X_1, \ldots, X_p; Y) = H(Y) - H(Y \mid X_1, \ldots, X_p) \leq H(Y).$$

Rearranging gives us the theorem. $\square$

One corollary of Theorem 5 is that for any random variable Y, there can be at most $p \leq H(Y)/\alpha$ random variables $X_1, \ldots, X_p$ that (a) are mutually independent and (b) satisfy $I(X_i; Y) \geq \alpha$.

4   Correlations in a finite sample

Suppose we conduct a survey with data on p predictors X and one outcome of interest Y on a random sample of n people, and then we evaluate the correlations between the outcome and each of the predictors. In this section, we consider how collinearity in X relates to these correlations.

We first give bounds on the range of possible values for the sum of squared correlations between each predictor and the outcome. We then model the outcome as a random vector, uniform over the unit sphere, and determine the expected sum of squared correlations. The proofs are given in the appendix.

Throughout this section, we assume that the data collected are contained in an $n \times p$ matrix X with $n > p$, where each of the columns $X_1, \ldots, X_p \in \mathbb{R}^n$ of X has mean zero and unit ℓ_2 norm.

We use $\mathrm{corr}(x, y)$ for $x, y \in \mathbb{R}^n$ (neither in the span of the all-ones vector $\mathbb{1}$) to denote the sample correlation:

$$\mathrm{corr}(x, y) = \frac{\sum_{i=1}^n (x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^n (x_i - \mu_x)^2 \sum_{i=1}^n (y_i - \mu_y)^2}},$$

where $\mu_x = \frac{1}{n} \sum_{i=1}^n x_i$ and $\mu_y = \frac{1}{n} \sum_{i=1}^n y_i$. We let $\sigma_1 \geq \cdots \geq \sigma_p \geq 0$ denote the singular values of X.

4.1   Range of sum of squared correlations

We first show that the sum of squared correlations is bounded above by the square of the largest singular value of X. Heuristically, the bound tells us that if there are strong linear relationships between the components of X, then there exists an outcome, Y, that is highly correlated with many of these components.

Theorem 6. *For all $Y \in \mathbb{R}^n$ not in the span of the all-ones vector $\mathbb{1}$,*

$$0 \leq \sum_{i=1}^p \mathrm{corr}(X_i, Y)^2 \leq \sigma_1^2.$$

Furthermore,

$$\sum_{i=1}^p \mathrm{corr}(X_i, U_1)^2 = \sigma_1^2,$$

where U_1 is the left singular vector of X corresponding to the singular value σ_1. Finally, any $y \in \mathbb{R}^n$ not in the span of $\mathbb{1}$ for which $y - \mu_y \mathbb{1}$ is orthogonal to the range of X satisfies

$$\sum_{i=1}^p \mathrm{corr}(X_i, y)^2 = 0.$$

4.2   Expected value of sum of squared correlations

In the previous section we established bounds on the sum of squared correlations correlations between the columns of X and any vector Y. In this section, we model Y as a random vector and find the expected value of the sum of squared correlations between the columns of X and Y. We assume Y is uniformly distributed on the unit sphere in $\mathbb{R}^n$, but the following theorem holds for any choice of radially symmetric random vector Y. We choose a radially symmetric distribution because we have no reason to give preference to one direction over another.

The following theorem shows that the maximum sum of squared correlations is generally much larger than the expected sum of square correlations. The theorem also shows that, assuming each row of X (each person surveyed) is independent and identically distributed, that the expected sum of squared correlations decays like $1/n$ where n is the number of people surveyed.

Theorem 7. *Let Y be uniformly distributed on the unit sphere in $\mathbb{R}^n$. Then*

$$\mathbb{E}\left(\sum_{i=1}^p \mathrm{corr}(X_i, Y)^2\right) = \frac{1}{n-1} \sum_{i=1}^p \sigma_i^2.$$

If Y is uniformly distributed on the unit sphere in $\mathbb{R}^n$, then for large n, the distribution of Y is well approximated by the n-dimensional multivariate Gaussian with mean zero and covariance $\frac{1}{n}I$. As a consequence, for large n, the distribution of sum of squared correlations is well approximated by a linear combination of independent chi-squared random variables, each with one degree of freedom: $\frac{1}{n-1}(\sigma_1 \xi_1 + \cdots + \sigma_p \xi_p)$.

5 Scenarios of application

Although we cannot directly apply these piranha theorems to data, we see them as providing some relevance to social science reasoning. The motivation of this review article is to collect several interesting results regarding the distributions of correlations or coefficients, with the aim of fostering further work on mathematical and statistical models for environments with a multiplicity of effects.

For example, an influential experiment from 1996 reported that participants were given a scrambled-sentence task and then were surreptitiously timed when walking away from the lab (Bargh et al., 1996). Students whose sentences included elderly-related words such as "worried," "Florida," "old," and "lonely" walked an average of 13% more slowly than students in the control condition, and the difference was statistically significant.

This experimental claim is of historical interest in psychology in that, despite its implausibility, it was taken seriously for many years (for example, "You have no choice but to accept that the major conclusions of these studies are true" (Kahneman, 2011)), but it failed to replicate (Harris et al., 2013) and is no longer generally believed to represent a real effect; for background see Wagenmakers et al. (2015). Now we understand such apparently statistically significant findings as the result of selection with many researcher degrees of freedom (Simmons et al., 2011).

Here, though, we will take the published claim at face value and also work within its larger theoretical structure, under which weak indirect stimuli can produce large effects.

An effect of 13% on walking speed is not in itself huge; the difficulty comes when considering elderly-related words as just one of many potential stimuli. Here are just some of the factors that have been presented in the social priming literature as having large effects on behavior: hormones (male and female), subliminal images, the outcomes of recent football games, irrelevant news events such as shark attacks, a chance encounter with a stranger, parental socioeconomic status, weather, the last digit of one's age, the sex of a hurricane name, the sexes of siblings, the position in which a person is sitting, and many others. A common feature of these examples is that the stimuli have no clear direct effect on the measured outcomes, and in most cases the experimental subject is not even aware of the manipulation. Based on these examples, one can come up with dozens of other potential stimuli that fit the pattern. For example, in addition to elderly-related words, one could also consider word lengths (with longer words corresponding to slower movement), sounds of words (with smooth sibilance motivating faster walking), subject matter (sports-related words as compared to sedentary words), affect (happy words compared to sad words, or calm compared to angry), words related to travel (inducing faster walking) or invoking adhesives such as tape or glue (inducing slower walking), and so on. Similarly, one can consider many different sorts of incidental events, not just encounters with strangers but also a ringing phone or knocking at the door or the presence of a male or female lab assistant (which could have a main effect or interact with the participant's sex) or the presence or absence of a newspaper or magazine on a nearby table, *ad infinitum*.

Now we can invoke the piranha theorem. Suppose we can imagine 100 possible stimuli, each with an effect of 13% on walking speed, all of which could arise in a real-world setting where we encounter many sources of text, news, and internal and external stimuli. If the effects are independent, then at any given time we could expect, on the log scale, a total effect with standard deviation $0.5\sqrt{100}\log(1.13) = 0.61$, thus walking speed could easily be multiplied or divided by $e^{0.61} = 1.8$ based on a collection of arbitrary stimuli that are imperceptible to the person being affected. And this factor of 1.8 could be made arbitrarily large by simply increasing the number of potential primes.

It is ridiculous to think that walking speed could be randomly doubled or halved based on a random collection of unnoticed stimuli—but that is the implication of the embodied cognition literature. It is basically a Brownian motion model in which the individual inputs are too large to work out.

We can think of four ways to avoid the ridiculous conclusion. The first possibility is that the different factors could interact or interfere in some way so that the variance of the total effect is less than the sum of the variances of the individual components. Second, effects could be much smaller. Change those 13% effects to 1% effects and you can get to more plausible totals, in the same way that real-world Brownian oscillations are tolerable because the impact of each individual molecule in the liquid is so small. Third, one could reduce the total number of possible influences. If there were only 10 possible stimuli rather than 100 or 1000 or more, then the total effect could fall within the range of plausibility. Fourth, there could be a distribution of effects with a few large influences and a long tail of relatively unimportant factors, so that the infinite sum has a reasonable bound.

All four of these options, or combinations of them, have major implications for the study of social priming and, more generally, for causal inference in an open-ended setting with large numbers of potential influences. First, if large interactions are possible, this suggests that stable individual treatment effects might be impossible to find: a 13% effect of a particular intervention in one particular experiment might be -18% in another context or $+2\%$ in the presence of some other unnoticed factor, and this in turn raises questions about the relevance of any particular study. Second, if effects are much smaller than reported, this suggests that existing studies are extremely underpowered (Button et al., 2013), so that published estimates are drastically overestimated and often in the wrong direction (Gelman and Carlin, 2014), thus essentially noise. Third, a restriction of the universe of potential stimuli would require an overhaul of the underlying theoretical framework in which just about any stimulus can cause a noticeable change. For example, if we think there cannot be more than about 10 large effects on walking speed, it would seem a stretch that unnoticed words in a sentence scrambling test would be one of these special factors. Fourth, if the distribution of effects is represented by a long series, most of which are tiny, this implies a prior distribution with a spike near zero, which in turn would result in near-zero estimated effect sizes in most cases. Our point is not that all effects are zero but rather that in a world of essentially infinite possible causal factors, some external structure must be applied in order to be able to estimate stable effects from finite samples.

6 Discussion and directions for future work

6.1 Bridging between deterministic and probabilistic piranha theorems

Are there connections between the worst-case bounds in Sections 2 and 3, the probabilistic bounds in Section 4 of this paper, and priors for the effective number of nonzero coefficients (Piironen and Vehtari, 2017) and models such as R^2 parameterization of linear regression as proposed in (Zhang et al., 2020)? We can consider two directions. The first is to consider departures from the parametric models such as the multivariate normal and t distributions and work out their implications for correlations and regression coefficients. The second idea is to obtain limiting results in high dimensions (that is, large numbers of predictors), by analogy to central limit theorems of random matrices. The idea here would be to consider a $n \times (p+1)$ matrix and then pull out one of the columns at random and consider it as the outcome, Y, with the other p columns being the predictors, X.

6.2 Regularization, sparsity, and Bayesian prior distributions

There has been research from many directions on regularization methods that provide soft constraints on models with large numbers of parameters. By "soft constraints," we mean that none of the parameters is literally constrained to fall within any finite range, but the estimates are pulled toward zero and can only take on large values if the data provide strong evidence in that direction.

Examples of regularization in non-Bayesian statistics include wavelet shrinkage (Donoho and Johnstone, 1994), lasso regression (Tibshirani, 1996), estimates for overparameterized image analysis and deep learning networks (Bora et al., 2017), and models that grow in complexity with increasing sample size (Geman and Hwang, 1982; Li and Meng, 2020). In a Bayesian context, regularization can be implemented using weakly informative prior distributions (Greenland and Mansournia, 2015; van Zwet, 2019) or more informative

priors that can encode the assumed sparsity (Mitchell and Beauchamp, 1988; George and McCulloch, 1993; Carvalho et al., 2009, 2010; Polson and Scott, 2011; Bhattacharya et al., 2015; Ghosh et al., 2018; Zhang et al., 2020) or assumed correlation and sparsity (Liu et al., 2018). Classical regularization is motivated by the goal of optimizing long-run frequency performance, and Bayesian priors represent additional information about parameters, coded as probability distributions. The various piranha theorems correspond to different constraints on these priors and thus even weakly informative priors should start by encoding these constraints.

From a different direction is the "bet on sparsity principle" based on the idea that any given data might allow some only some small number of effects or, more generally, a low-dimensional structure, to be reliably learned (Hastie et al., 2001; Tibshirani, 2014).

6.3   Implications for social science research

As noted at the beginning of this article, there has been a crisis in psychology, economics, and other areas of social science, with prominent findings and apparently strong effects that do not appear in attempted replications by outside research groups (e.g., Open Sciecnce Collaboration, 2015; Altmejd et al., 2019; Gordon et al., 2020)). The discussion of the replication crisis has touched on many aspects of the problem, including estimating its scale and scope, identifying the statistical errors and questionable research practices that have led researchers to systematically overestimate effect sizes and be overconfident in their findings, and studying the incentives of the scientific publication process that can allow entire subfields to get lost in the interpretation of noise.

The present article goes in a different direction, asking the theoretical question: under what conditions is it possible to have many large effects in a multivariate system? In different ways, our results rule out the possibility of multiple large effects or "piranhas" among a set of random variables. These theoretical findings do not directly call into question any *particular* claimed effect, but they do raise suspicions about a model of social interactions in which many large effects are swimming around, just waiting to be captured in quantitative studies.

To more directly connect our theorems with social science would require some modeling of the set of candidate predictor and outcome variables in a subfield, similar to multiverse analysis (Steegen et al., 2016). Any general implications for social science would only become clear after consideration of particular research areas.

6.4   Piranhas and butterflies

A fundamental tenet of social psychology and behavioral economics, at least how it is presented in the news media, and taught and practiced in many business schools, is that small "nudges," often the sorts of things that we might not think would affect us at all, can have big effects on behavior.

The model of the world underlying these claims is not just the "butterfly effect" that small changes can have big effects; rather, it's that small changes can have big and predictable effects, a sort of "button-pushing" model of social science, the idea that if you do X, you can expect to see Y.

In response to this attitude, we have presented the piranha argument, which states that there can be some large and predictable effects on behavior, but not a lot, because, if there were, then these different effects would interfere with each other, and as a result it would be hard to see any consistent effects of anything in observational data.

In a similar vein, Cook (2018) writes, "The butterfly effect is the semi-serious claim that a butterfly flapping its wings can cause a tornado half way around the world. It's a poetic way of saying that some systems show sensitive dependence on initial conditions, that the slightest change now can make an enormous difference later ... Once you think about these things for a while, you start to see nonlinearity and potential butterfly effects everywhere. There are tipping points everywhere waiting to be tipped!" But, Cook continues, it's not so simple: "A butterfly flapping its wings usually has no effect, even in sensitive or chaotic systems. You might even say especially in sensitive or chaotic systems. Sensitive systems are not always and everywhere sensitive to everything. They are sensitive in particular ways under particular circumstances,

and can otherwise be quite resistant to influence. ...The lesson that many people draw from their first exposure to complex systems is that there are high leverage points, if only you can find them and manipulate them. They want to insert a butterfly to at just the right time and place to bring about a desired outcome. Instead, we should humbly evaluate to what extent it is possible to steer complex systems at all. We should evaluate what aspects can be steered and how well they can be steered. The most effective intervention may not come from tweaking the inputs but from changing the structure of the system." Whether thinking in terms of butterflies or piranhas, we can think of an infinite series of potential effects, which imply that only a few can be large and also create the possibility of interactions that, after some point, overwhelm any main effects.

Acknowledgements

References

Achen, C. H. and Bartels, L. M. (2002). Blind retrospection: Electoral responses to drought, flu, and shark attacks. *Presented at the Annual Meeting of the American Political Science Association.*

Altmejd, A., Dreber, A., Forsell, E., Huber, J., Imai, T., Johannesson, M., Kirchler, M., Nave, G., and Camerer, C. (2019). Predicting the replicability of social science lab experiments. *PLoS One*, 14:e0225826.

Anonymous (2021). Piranha problem. *TV Tropes*, https://tvtropes.org/pmwiki/pmwiki.php/Main/PiranhaProblem.

Bargh, J. A., Chen, M., and Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype-activation on action. *Journal of Personality and Social Psychology*, 74:230–244.

Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015). Dirichlet-Laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, 110(512):1479–1490.

Bora, A., Jalal, A., Price, E., and Dimakis, A. G. (2017). Compressed sensing using generative models. In *International Conference on Machine Learning*, pages 537–546.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., and Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5):365–376.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). Handling sparsity via the horseshoe. In van Dyk, D. and Welling, M., editors, *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 73–80. PMLR.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480.

Clancy, K. (2012). Hot for Obama, but only when this smug married is not ovulating. *Scientific American*, https://blogs.scientificamerican.com/context-and-variation/hot-for-obama-ovulation-politics-women/.

Cook, J. (2018). The other butterfly effect. *John D. Cook Consulting*, https://www.johndcook.com/blog/2018/08/07/the-other-butterfly-effect/.

Donoho, D. L. and Johnstone, J. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455.

Durante, K. M., Rae, A., and Griskevicius, V. (2013). The fluctuating female vote: Politics, religion, and the ovulatory cycle. *Psychological Science*, 24:1007–1016.

Fowler, A. and Hall, A. B. (2018). Do shark attacks influence presidential elections? Reassessing a prominent finding on voter competence. *Journal of Politics*, 80:1423–1437.

Fowler, A. and Montagnes, B. P. (2015). College football, elections, and false-positive results in observational research. *Proceedings of the National Academy of Sciences*, 112:13800–13804.

Gelman, A. (2015). The connection between varying treatment effects and the crisis of unreplicable research: A Bayesian perspective. *Journal of Management*, 41:632–643.

Gelman, A. (2017). The piranha problem in social psychology / behavioral economics: The "take a pill" model of science eats itself. *Statistical Modeling, Causal Inference, and Social Science*, https://statmodeling.stat.columbia.edu/2017/12/15/piranha-problem-social-psychology-behavioral-economics-button-pushing-model-science-eats/.

Gelman, A. and Carlin, J. B. (2014). Beyond power calculations: Assessing type s (sign) and type m (magnitude) errors. *Perspectives on Psychological Science*, 9:641–651.

Geman, S. and Hwang, C.-R. (1982). Nonparametric maximum likelihood estimation by the method of sieves. *Annals of Statistics*, 10:481–484.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

Ghosh, J., Li, Y., Mitra, R., et al. (2018). On the use of Cauchy prior distributions for Bayesian logistic regression. *Bayesian Analysis*, 13(2):359–383.

Gordon, M., Viganola, D., Bishop, M., Chen, Y., Dreber, A., Goldfedder, B., Holzmeister, F., Johannesson, M., Liu, Y., Twardy, C., Wang, J., and Pfeiffer, T. (2020). Are replication rates the same across academic fields? Community forecasts from the DARPA SCORE programme. *Royal Society Open Science*, 7:200566.

Greenland, S. and Mansournia, M. A. (2015). Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions. *Statistics in Medicine*, 34:3133–3143.

Harris, C. R., Coburn, N., Rohrer, D., and Pashler, H. (2013). Two failures to replicate high-performance-goal priming effects. *PLOS One*, 8:e72467.

Hastie, T. J., Tibshirani, R. J., and Friedman, J. H. (2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., and Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology*, 13(3).

Healy, A. J., Malhotra, N., and Mo, C. H. (2010). Irrelevant events affect voters' evaluations of government performance. *Proceedings of the National Academy of Sciences*, 107(29):12804–12809.

Ioannidis, J. P. A. (2005). Contradicted and initially stronger effects in highly cited clinical research. *Journal of the American Medical Association*, 294:218–228.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Allen Lane.

Ladd, J. M. and Lenz, G. S. (2011). Does anxiety improve voters' decision making? *Political Psychology*, 32(2):347–361.

Li, X. and Meng, X.-L. (2020). A multi-resolution theory for approximating infinite-p-zero-n: Transitional inference, individualized predictions, and a world without bias-variance trade-off. *Journal of the American Statistical Association*.

Liu, C., Yang, Y., Bondell, H., and Martin, R. (2018). Bayesian inference in high-dimensional linear models using an empirical correlation-adaptive prior. *arXiv:1810.00739*.

Mellon, J. (2020). Rain, rain, go away: 137 potential exclusion-restriction violations for studies using weather as an instrumental variable. `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3715610`.

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., Glennerster, R., Green, D. P., Humphreys, M., Imbens, G., et al. (2014). Promoting transparency in social science research. *Science*, 343(6166):30–31.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1036.

Obschonka, M., Stuetzer, M., Rentfrow, P. J., Lee, N., Potter, J., and Gosling, S. D. (2018). Fear, populism, and the geopolitical landscape: The "sleeper effect" of neurotic personality traits on regional voting behavior in the 2016 Brexit and Trump elections. *Social Psychological and Personality Science*, 9(3):285–298.

Open Sciecnce Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349:aac4716.

Parker, M. T. and Isbell, L. M. (2010). How I vote depends on how I feel: The differential impact of anger and fear on political information processing. *Psychological Science*, 21(4):548–550.

Piironen, J. and Vehtari, A. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, 11(2):5018–5051.

Polson, N. G. and Scott, J. G. (2011). Shrink globally, act locally: sparse Bayesian regularization and prediction. In Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West, M., editors, *Bayesian Statistics 9*, pages 501–538. Oxford University Press, Oxford.

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11):1359–1366.

Steegen, S., Tuerlinckx, F., Gelman, A., and Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11:702–712.

Tao, T. (2014). When is correlation transitive? *What's New*, `https://terrytao.wordpress.com/2014/06/05/when-is-correlation-transitive`.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tibshirani, R. J. (2014). In praise of sparsity and convexity. In *Committee of Presidents of Statistical Societies (COPSS) 50th anniversary volume*. Wiley.

van Zwet, E. (2019). A default prior for regression coefficients. *Statistical Methods in Medical Research*, 28:3799–3807.

Wagenmakers, E., Wetzels, R., Borsboom, D., Kievit, R., and van der Maas, H. L. J. (2015). A skeptical eye on psi. In *Extrasensory Perception: Support, Skepticism, and Science*, pages 153–176. Praeger.

Zhang, Y., Naughton, B. P., Bondell, H. D., and Reich, B. J. (2020). Bayesian regression using a prior on the model fit: The R2-D2 shrinkage prior. *Journal of the American Statistical Association*.

A Proofs of theorems from Section 4

In this section, we give the proofs of Theorems 6 and 7.

A.1 Notation

For any $x = (x_1, \ldots, x_n)^\top \in \mathbb{R}^n$ such that $x \neq \lambda \mathbb{1}$ for all $\lambda \in \mathbb{R}$ (i.e., x is not in the span of $\mathbb{1}$), we write $x^* \in \mathbb{R}^n$ to denote the "standardized" vector given by the formula

$$x^* = \frac{x - \frac{1}{n}(x^\top \mathbb{1})\mathbb{1}}{\|x - \frac{1}{n}(x^\top \mathbb{1})\mathbb{1}\|} = \frac{x - (\frac{1}{n}\sum_{j=1}^n x_j)\mathbb{1}}{\sqrt{\sum_{i=1}^n (x_i - \frac{1}{n}\sum_{j=1}^n x_j)^2}}.$$

Note that x^* is a unit vector in $\mathbb{R}^n$ and is orthogonal to $\mathbb{1}$. Using this notation, we have

$$\mathrm{corr}(x, y) = (x^*)^\top (y^*) \tag{4}$$

for any $x, y \in \mathbb{R}^n$ not in the span of $\mathbb{1}$.

Write the singular value decomposition of X as

$$X = \sum_{k=1}^p \sigma_k U_k V_k^\top, \tag{5}$$

where $U_1, \ldots, U_p \in \mathbb{R}^n$ are orthonormal left singular vectors of X, $V_1, \ldots, V_p \in \mathbb{R}^p$ are orthonormal right singular vectors of X, and $\sigma_1 \geq \cdots \geq \sigma_p \geq 0$ are the singular values of X.

Recall that we assume $X_1, \ldots, X_p$ satisfy $\mathbb{1}^\top X_i = 0$ and $\|X_i\| = 1$ for all $i = 1, \ldots, p$. This implies the following lemma.

Lemma 8. $X_i = X_i^*$ for all $i = 1, \ldots, p$, and $U_k = U_k^*$ for all $k = 1, \ldots, p$.

Proof. The assumption on X_i implies that $X_i^* = X_i$ for each i. Moreover, the assumptions imply that the all-ones vector $\mathbb{1}$ is orthogonal to the range of X, which is spanned by $U_1, \ldots, U_p$. Hence $U_k = U_k^*$ for each k as well. $\qquad\square$

A.2 Proofs of Theorem 6 and Theorem 7

The proofs rely on the following lemma for expressing the sum of squared correlations.

Lemma 9. *For any vector $y \in \mathbb{R}^n$ such that $y \neq \lambda \mathbb{1}$ for all $\lambda \in \mathbb{R}$,*

$$\sum_{i=1}^p \mathrm{corr}(X_i, y)^2 = \sum_{k=1}^p \sigma_k^2 (U_k^\top y^*)^2.$$

Proof. By direct computation:

$$\sum_{i=1}^p \mathrm{corr}(X_i, y)^2 = \sum_{i=1}^p ((X_i^*)^\top (y^*))^2 \qquad \text{(by equation 4)}$$

$$= \sum_{i=1}^p (X_i^\top y^*)^2 \qquad \text{(by Lemma 8)}$$

$$= \|X^\top y^*\|^2$$

$$= \left\| \sum_{k=1}^p \sigma_k V_k U_k^\top y^* \right\|^2 \qquad \text{(by equation 5)}$$

$$= \sum_{k=1}^p \sigma_k^2 (U_k^\top y^*)^2 \qquad \text{(by Pythagorean theorem).} \qquad\square$$

Proof of Theorem 6. Since Y^* is a unit vector in $\mathbb{R}^n$, we have

$$\sum_{k=1}^{p}(U_k^\mathsf{T}Y^*)^2 \leq 1.$$

Hence, by Lemma 9,

$$0 \leq \sum_{i=1}^{p}\mathrm{corr}(X_i,Y)^2 = \sum_{k=1}^{p}\sigma_k^2(U_k^\mathsf{T}Y^*)^2 \leq \max_{k\in\{1,\dots,p\}}\sigma_k^2 = \sigma_1^2.$$

If $Y = U_1$ (which is U_1^* by Lemma 8), then

$$\sum_{i=1}^{p}\mathrm{corr}(X_i,U_1)^2 = \sum_{k=1}^{p}\sigma_k^2(U_k^\mathsf{T}U_1^*)^2 = \sum_{k=1}^{p}\sigma_k^2(U_k^\mathsf{T}U_1)^2 = \sigma_1^2.$$

If Y^* is orthogonal to the range of X, then

$$\sum_{i=1}^{p}\mathrm{corr}(X_i,Y)^2 = \sum_{k=1}^{p}\sigma_k^2(U_k^\mathsf{T}Y^*)^2 = 0. \qquad \square$$

Proof of Theorem 7. By Lemma 8, the vectors $U_1,\dots,U_p$ are orthogonal to the unit vector $\frac{1}{\sqrt{n}}\mathbb{1}$. We extend the collection of orthonormal vectors $U_1,\dots,U_p,\frac{1}{\sqrt{n}}\mathbb{1}$ with orthonormal unit vectors $U_{p+1},\dots,U_{n-1}$ to obtain an orthonormal basis for $\mathbb{R}^n$. With probability 1, the random vector Y is not in the span of $\mathbb{1}$. Hence, Y^* is well-defined and can be written uniquely as a linear combination of the aforementioned basis vectors:

$$Y^* = a_1 U_1 + \cdots + a_{n-1}U_{n-1} + a_n\frac{1}{\sqrt{n}}\mathbb{1},$$

where

$$a_k = \begin{cases} U_k^\mathsf{T}Y^* & \text{if } 1 \leq k \leq n-1, \\ 0 & \text{if } k = n \text{ (since } \mathbb{1}^\mathsf{T}Y^* = 0), \end{cases}$$

and

$$1 = a_1^2 + \cdots + a_{n-1}^2$$

(since Y^* is a unit vector). In particular,

$$1 = \mathbb{E}(a_1^2) + \cdots + \mathbb{E}(a_{n-1}^2),$$

which implies

$$\mathbb{E}(a_k^2) = \frac{1}{n-1}$$

for each $k = 1,\dots,n-1$, by symmetry. By Lemma 9,

$$\mathbb{E}\left(\sum_{i=1}^{p}\mathrm{corr}(X_i,Y)^2\right) = \mathbb{E}\left(\sum_{k=1}^{p}\sigma_k^2(U_k^\mathsf{T}Y^*)^2\right) = \sum_{k=1}^{p}\sigma_k^2\mathbb{E}(a_k^2) = \frac{1}{n-1}\sum_{k=1}^{p}\sigma_k^2. \qquad \square$$