The piranha problem: Large effects swimming in a small pond

Christopher Tosh, Philip Greengard, Ben Goodrich, Andrew Gelman, and Daniel Hsu

Columbia University, New York

5 Nov 2020

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 with formal theorems that show, unless the explanatory variables also have sizable effects on each other, it is impossible to have many such large effects.

1 Introduction

This paper offers 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 [25, 19, 23, 7], as well external factors such as droughts, shark attacks, and the performance of local college football games [1, 17, 8, 9]. Many of these particular findings have been questioned on methodological grounds [9, 8, 5, 10], 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 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

E-mail: ct2915@columbia.edu, pg2118@columbia.edu, benjamin.goodrich@columbia.edu, gelman@stat.columbia.edu, aki.vehtari@aalto.fi, djhsu@cs.columbia.edu

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 [21].

In this work, we prove 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" [11], the analogy being that if one has a large number of piranhas (representing large effects) in a single fish tank, then one will soon be left with at most one piranha. 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.

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 [18]. Many of the explanations for the crisis have focused on various methodological issues, such as researcher degrees of freedom [26], underpowered studies [4], and data dredging [16]. In some cases, solutions to these issues have also been proposed, notably good practice guidelines for authors and reviewers [26] and preregistration of studies [22]. 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.

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,

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

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* [28]. We offer a proof here for completeness.

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

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

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

Proof. Without loss of generality, we may assume that $X_1, \dots, X_p, Y$ have mean zero and unit variance. Define $Z_1, \dots, Z_p$ 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 $i = 1, \dots, p$. By Cauchy-Schwarz,

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

Therefore,

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

Rearranging gives us the theorem statement. □

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

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 $\mathbb{E}(XY) \ge \tau$ and $\mathbb{E}(YZ) \ge 1 - \epsilon$, then

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

and by Cauchy-Schwarz, we have

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

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

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

The bound in Theorem 1 is essentially tight for large p. To see this, pick any $0 \le \tau \le 1$, and take $X_1, \dots, X_p$ 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 $i = 1, \dots, p$,

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

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,

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

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.

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*

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

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

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

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

$$v^\intercal \Sigma v \;=\; 2(1 - \|a\|)\|a\|^2 \;\geq\; 0,$$

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

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 $\mathbb{E}(XX^\intercal)$ is invertible, its inverse has eigenvalues bounded above by $1/\lambda_{\min}$, and

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

Define $\tilde{X} = (\mathbb{E}(XX^\intercal))^{-1/2}X$, so $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_p)^\intercal$ 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

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

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

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

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

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

$$\begin{aligned} H(X) &= \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)}, \\ H(Y \,|\, X) &= \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x,y) \log \frac{p(x)}{p(x,y)}, \end{aligned}$$

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 $X_1, \ldots, X_p$,*

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

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

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

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

$$H(X|Y,Z) \;\le\; H(X \,|\, Y) \;\le\; H(X).$$

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*

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

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

Proof. Using the definition of mutual information, we have

$$H(X_i \,|\, X_{-i}) \;\ge H(X_i) - I(X_i; X_{-i}).$$

Since conditioning reduces entropy, this implies

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

Thus, we have by the chain rule of entropy

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

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

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

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

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

Now, we can also write

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

Rearranging gives us the theorem. □

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 $\text{corr}(x, y)$ for $x, y \in \mathbb{R}^n$ (neither in the range of $\mathbb{1}$) to denote the sample correlation:

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

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 range of the all-ones vector $\mathbb{1}$,*

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

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 range 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 square 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 multvariate Gaussian with mean zero and covariance $(1/n)I$. As a consequence, for large n, the distribution of sum of square correlations is well-approximated by the following 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 Discussion and directions for future work

5.1 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 specific models such as the R^2 parameterization of linear regression as proposed in [32]? 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.

5.2 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 [6], lasso regression [29], estimates for overparameterized image analysis and deep learning networks [3], and models that grow in complexity with increasing sample size [12, 20]. In a Bayesian context, regularization can be implemented using weakly informative prior distributions; e.g., [14, 31].

Classical regularization is motivated by the goal of optimizing long-run frequency performance, and Bayesian priors represent additional information about parameters, coded as a prior distribution. The various piranha theorems correspond to different constraints on these priors.

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 [15, 30].

5.3 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., [24, 2, 13]). 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 [27]. Any general implications for social science would only become clear after consideration of particular research areas.

Acknowledgements

We thank Aki Vehtari and Lauren Kennedy for helpful discussions. CT and DH acknowledge support from NSF grant CCF-1740833, a JP Morgan Faculty Award, and a Sloan Research Fellowship. AG acknowledges support from ONR grant N00014-19-1-2204.

References

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

[2] Adam Altmejd, Anna Dreber, Eskil Forsell, Juergen Huber, Taisuke Imai, Magnus Johannesson, Michael Kirchler, Gideon Nave, and Colin Camerer. Predicting the replicability of social science lab experiments. *PLoS One*, 14:e0225826, 2019.

[3] Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G. Dimakis. Compressed sensing using generative models. In *International Conference on Machine Learning*, pages 537–546, 2017.

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

[5] Kate Clancy. 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/`, 2012.

[6] David L. Donoho and Jain M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

[7] Kristina M. Durante, Ashley Rae, and Vladas Griskevicius. The fluctuating female vote: Politics, religion, and the ovulatory cycle. *Psychological Science*, 24:1007–1016, 2013.

[8] Anthony Fowler and Andrew B. Hall. Do shark attacks influence presidential elections? Reassessing a prominent finding on voter competence. *Journal of Politics*, 80:1423–1437, 2018.

[9] Anthony Fowler and B. Pablo Montagnes. College football, elections, and false-positive results in observational research. *Proceedings of the National Academy of Sciences*, 112:13800–13804, 2015.

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

[11] Andrew Gelman. 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/`, 2017.

[12] Stuart Geman and Chii-Ruey Hwang. Nonparametric maximum likelihood estimation by the method of sieves. *Annals of Statistics*, 10:481–484, 1982.

[13] Michael Gordon, Domenico Viganola, Michael Bishop, Yiling Chen, Anna Dreber, Brandon Goldfedder, Felix Holzmeister, Magnus Johannesson, Yang Liu, Charles Twardy, Juntao Wang, and Thomas Pfeiffer. Are replication rates the same across academic fields? Community forecasts from the DARPA SCORE programme. *Royal Society Open Science*, 7:200566, 2020.

[14] Sander Greenland and Mohammad Ali Mansournia. Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions. *Statistics in Medicine*, 34:3133–3143, 2015.

[15] Trevor J. Hastie, Robert J. Tibshirani, and Jerome H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, 2001.

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

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

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

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

[20] Xinran Li and Xiao-Li Meng. 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*, 2020.

[21] Jonathan Mellon. 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`, 2020.

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

[23] M. Obschonka, M. Stuetzer, P. J. Rentfrow, N. Lee, J. Potter, and S. D. Gosling. 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, 2018.

[24] Open Sciecnce Collaboration. Estimating the reproducibility of psychological science. *Science*, 349:aac4716, 2015.

[25] M. T. Parker and L. M. Isbell. 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, 2010.

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

[27] Sara Steegen, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11:702–712, 2016.

[28] Terence Tao. When is correlation transitive? *What's New*, `https://terrytao.wordpress.com/2014/06/05/when-is-correlation-transitive`, 2014.

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

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

[31] Erik van Zwet. A default prior for regression coefficients. *Statistical Methods in Medical Research*, 28:3799–3807, 2019.

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

A Proofs of theorems from Section 4

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

A.1 Notation

For any $x = (x_1, \ldots, x_n)^\intercal \in \mathbb{R}^n$ 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

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

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

$$\operatorname{corr}(x, y) \;=\; (x^*)^\intercal (y^*) \tag{4}$$

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

Write the singular value decomposition of X as

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

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 $\mathbb{1}^\intercal X_i = 0$ 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. □

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 $\lambda \in \mathbb{R}$,*

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

Proof. By direct computation:

$$\begin{aligned}
\sum_{i=1}^p \operatorname{corr}(X_i, y)^2 \;&=\; \sum_{i=1}^p ((X_i^*)^\intercal (y^*))^2 && \text{(by equation (4))} \\
&=\; \sum_{i=1}^p (X_i^\intercal y^*)^2 && \text{(by Lemma 8)} \\
&=\; \|X^\intercal y^*\|^2 \\
&=\; \left\| \sum_{k=1}^p \sigma_k V_k U_k^\intercal y^* \right\|^2 && \text{(by equation (5))} \\
&=\; \sum_{k=1}^p \sigma_k^2 (U_k^\intercal y^*)^2 && \text{(by Pythagorean theorem).}
\end{aligned}$$

□

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

$$\sum_{k=1}^{p}(U_k^\intercal Y^*)^2 \;\le\; 1.$$

Hence, by Lemma 9,

$$0 \;\le\; \sum_{i=1}^{p}\operatorname{corr}(X_i,Y)^2 \;=\; \sum_{k=1}^{p}\sigma_k^2(U_k^\intercal Y^*)^2 \;\le\; \max_{k\in\{1,\ldots,p\}}\sigma_k^2 \;=\; \sigma_1^2.$$

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

$$\sum_{i=1}^{p}\operatorname{corr}(X_i,U_1)^2 \;=\; \sum_{k=1}^{p}\sigma_k^2(U_k^\intercal U_1^*)^2 \;=\; \sum_{k=1}^{p}\sigma_k^2(U_k^\intercal U_1)^2 \;=\; \sigma_1^2.$$

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

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

Proof of Theorem 7. By Lemma 8, the vectors $U_1,\ldots,U_p$ are orthogonal to the unit vector $\frac{1}{\sqrt{n}}\mathbb{1}$. We extend the collection of orthonormal vectors $U_1,\ldots,U_p,\frac{1}{\sqrt{n}}\mathbb{1}$ with orthonormal unit vectors $U_{p+1},\ldots,U_{n-1}$ 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:

$$Y^* \;=\; a_1U_1 + \cdots + a_{n-1}U_{n-1} + a_n\tfrac{1}{\sqrt{n}}\mathbb{1},$$

where

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

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 $k=1,\ldots,n-1$, by symmetry. By Lemma 9,

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