The piranha problem: Large effects swimming in a small pond

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

4 Nov 2022

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 different explanatory variables. In such cases, particularly those where no relationships are known among the 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 showing that, unless the explanatory variables also have sizable dependencies with each other, it is impossible to have many such large effects. We discuss implications for the replication crisis in social science.

1   Background

In this work, we discuss an inevitable consequence of having a stable system in which many explanatory variables have large effects: these variables must have large interactions which will be unlikely to cancel either other out to the extent required for general stability or predictability. 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 for which a large number of studies demonstrate effects of novel explanatory variables, then we can conclude that either some of these effects are smaller than claimed or that multiple the explanatory variables are essentially measuring the same phenomenon.

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? The present work follows up on Cornfield et al. (1959) and Ding and Vanderweele (2014), considering quantitative constraints in the effects of additional variables.

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 teams (Achen and Bartels, 2002; Healy et al., 2010; Fowler and Hall, 2018; Fowler and Montagnes, 2015). Many of these findings have been questioned on methodological grounds (Fowler and Montagnes, 2015; Fowler and Hall, 2018; Clancy, 2012; Gelman, 2015a), but they remain in the public

---

[*]Memorial Sloan Kettering Cancer Center, 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.

discourse (e.g. Krugman, 2021). Beyond the details of these particular studies, it is natural to ask if all of these effects can be real in the sense of representing patterns that will consistently appear in the future.

The implication of the published and well-publicized 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. If such effects are large, it is necessary to consider their interactions, because the average effect of a factor in any particular study will depend on the levels of all the other factors in that environment. Similarly, Mellon (2020) has argued 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.

These concerns are particularly relevant in social science, where the search for potential causes is open-ended. Our work here is 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. 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.

This article collects several mathematical results regarding the distributions of correlations or coefficients, with the aim of fostering further work on statistical models for environments with a multiplicity of effects. What is novel in this paper is not the theorems themselves but rather viewing them in the context of trying to make sense of clusters of research studies that claim to have found large effects.

There are many ways to capture the dependence among 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 Section 4. These theorems illustrate the general phenomena at work in any setting with multiple causal or explanatory variables.

2   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 A, you can expect to see B.

In response to this attitude, we present 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, a "hall of mirrors" of interactions (Cronbach, 1975) that would make it hard to detect 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 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.

3   Example: hypothesized effect sizes in social priming

We demonstrate the possibility of quantitative analysis of the piranha problem using the example of 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 published in the social priming and related literatures as having large effects on behavior: male and female hormones (Petersen et al., 2013; Durante et al., 2013), subliminal images (Bartels, 2014; Gelman, 2015b), the outcomes of recent football games (Healy et al., 2010; Graham et al., 2022; Fowler and Montagnes, 2015, 2022), irrelevant news events such as shark attacks (Achen and Bartels, 2002; Fowler and Hall, 2018), a chance encounter with a stranger (Sands, 2017; Gelman, 2018b), parental socioeconomic status (Petersen et al., 2013), weather (Jessica L. Tracy, 2014; Gelman, 2018a), the last digit of one's age (Alter and Hershfield, 2014; Kühnea et al., 2015), the sex of a hurricane name (Jung et al., 2014; Freese, 2014), the sexes of siblings (Blanchard and Bogaert, 1996; Bogaert, 2006; Gelman and Stern, 2006), the position in which a person is sitting (Carney et al., 2010; Cesario and Johnson, 2018), and many others. A common feature of these examples is that the stimuli have no clear direct effect on the measured outcomes, and in many 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 principle. 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 outrageous to think that walking speed could be randomly doubled or halved based on a random collection of unnoticed and essentially irrelevant 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 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 five or ten 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 whose elements 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.

4   Piranha theorems

In this section, we present piranha theorems for linear and nonlinear effects. We consider two different ways of measuring linear effects. The first of these, correlation, is straightforward to interpret. We 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. The second examines linear regression coefficients. We 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 of the predictors. Thus, there can only be so many individual coefficients with a large magnitude. Finally, we consider a general (nonlinear) form of dependency, mutual information, and present a corresponding piranha theorem for that measure.

4.1   Correlation

The first type of pattern we consider is correlation. 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 nonzero variance, then*

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

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

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,

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

This is also easily seen from applying Cauchy-Schwarz to the $p + 1$-dimensional correlation matrix via $e_1' R a$ with $e_1' = (1, 0, \ldots, 0)$ and $a' = (0, b')$ with b_j set to ± 1 depending on the sign of the y, X_j correlation. 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. □

A direct consequence of Theorem 1 is that if $X_1, \ldots, X_p$ are independent (or uncorrelated) random variables and each has correlation at least τ with y, then $\tau \leq 1/\sqrt{p}$.

In some situations, the outcome variable may change from study to study, for example a program evaluation in economics might look at employment, income, or savings; a political intervention might target turnout or vote choice; or an education experiment might look at outcomes on several tests. Although the different outcomes in a study 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 nonzero variance. If $\operatorname{corr}(Y_i, Y_j) \geq 1 - \epsilon$ and $|\operatorname{corr}(X_i, Y_i)| \geq \tau$ for $i, j = 1, \ldots, p$, then $\sum_{i \neq j} |\operatorname{corr}(X_i, X_j)| \geq p((\tau - \sqrt{2\epsilon})^2 p - 1)$.*

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 $i = 1, \ldots, p$,

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

One drawback of Theorem 1 is that the upper bound depends on a coarse measure of interdependence of the covariates, namely the sum of all pairwise correlations $\sum_{i,j} |\operatorname{corr}(X_i, X_j)|$. One might hope that if we have a finer-grained control on the correlation matrix, we should be able to get a stronger result. This is accomplished by the following piranha theorem, which shows that we can instead get an upper bound that depends on the largest eigenvalue of the correlation matrix. However, this comes at the expense of bounding the sum of squared correlations $|\operatorname{corr}(X_i, Y)|^2$, rather than the sum of their absolute values.

Theorem 3. *If $X_1, \ldots, X_p, y$ are real-valued random variables with finite nonzero variance, then*

$$
\sum_{i=1}^{p} |\operatorname{corr}(X_i, y)|^2 \;\leq\; \lambda_{\max},
$$

where $\lambda_{\max}$ is the maximum eigenvalue of the correlation matrix $\Sigma_{i,j} = \operatorname{corr}(X_i, X_j)$.

Observe that Theorems 1 and 3 are generally incomparable since $\sum_{i=1}^{p} |\operatorname{corr}(X_i, y)|^2 \leq \sum_{i=1}^{p} |\operatorname{corr}(X_i, y)|$ but

$$
\lambda_{\max} \leq \sqrt{\sum_{i,j} |\operatorname{corr}(X_i, X_j)|^2} \leq \sqrt{\sum_{i,j} |\operatorname{corr}(X_i, X_j)|}.
$$

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

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-definite. We conclude that $\|a\| \leq 1$. $\qquad \square$

With the above in hand, we turn to the proof of Theorem 3.

Proof of Theorem 3. Assume without loss of generality that $X_1, \ldots, X_p, y$ have mean zero and unit variance. Denote the eigendecomposition of Σ as $Q \operatorname{diag}(\lambda_1, \ldots, \lambda_p)Q^T$, where $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ and Q is orthogonal.

Let $\tilde{X} = Q^T X$, where $X = (X_1, \ldots, X_p)$. Then $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_p)$ is a mean-zero random vector whose covariance matrix is $\operatorname{diag}(\lambda_1, \ldots, \lambda_p)$. For $j \in \{1, \ldots, p\}$ with $\lambda_j = \operatorname{var}(\tilde{X}_j) = 0$, we have $\tilde{X}_j = 0$ almost surely. Thus, we may apply Lemma 4 to get

$$\|\mathbb{E}(y\tilde{X})\|^2 \;=\; \sum_{j=1}^{p} \mathbb{E}(y\tilde{X}_j)^2 \;=\; \sum_{j:\lambda_j>0} \lambda_j \mathbb{E}(y\tilde{X}_j/\sqrt{\lambda_j})^2 \;\leq\; \lambda_1 \sum_{j:\lambda_j>0} \mathbb{E}(y\tilde{X}_j/\sqrt{\lambda_j})^2 \;\leq\; \lambda_1.$$

Thus, we have

$$\sum_{i=1}^{p} |\operatorname{corr}(X_i, y)|^2 \;=\; \|\mathbb{E}(yX)\|^2 \;=\; \|QQ^T\mathbb{E}(yX)\|^2 \;=\; \|Q\mathbb{E}(y\tilde{X})\|^2 \;=\; \|\mathbb{E}(y\tilde{X})\|^2 \;\leq\; \lambda_1,$$

where we have used the fact that Q is orthogonal. □

4.2 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,

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

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

Theorem 5. *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 5 states for independent covariates, there may be at most $1/\tau^2$ regression coefficients β_i with magnitude larger than τ.

Proof of Theorem 5. 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 $(\mathbb{E}(XX^\intercal))^{-1}$. □

Theorem 5 implies that we cannot have a regression coefficient bigger than $1/\sqrt{\lambda_{\min}}$. If the predictors are standardized and uncorrelated, $\lambda_{\min} = 1$. In general $\lambda_{\min}$ cannot get small without the explanatory variables having sizable correlations with each other.

4.3   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 previous 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

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

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

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 $X_1, \ldots, X_p$,*

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

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

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

Fact (Conditioning reduces entropy). *For random variables x, y, z,*

$$H(x|y, z) \;\leq\; H(x \,|\, y) \;\leq\; H(x).$$

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

Theorem 6. *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

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

Since conditioning reduces entropy, this implies

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

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

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

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. $\qquad\square$

One corollary of Theorem 6 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 $I(X_i; y) \geq \alpha$.

5   Correlations in a finite sample

We now turn our focus back to correlations, this time 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.

We collect the data 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, and we will 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 $\mu_y = \frac{1}{n} \sum_{i=1}^{n} y_i$.

An application of Theorem 3 tells us that any non-constant vector $y \in \mathbb{R}^n$ satisfies

$$0 \;\leq\; \sum_{j=1}^{p} |\mathrm{corr}(X_i, y)|^2 \;\leq\; \sigma_1^2,$$

where $\sigma_1 \geq \cdots \geq \sigma_p \geq 0$ denote the singular values of X. Moreover, it is not hard to see there exists a vector that achieves the upper bound, namely the top singular vector of X.

This analysis shows a *worst-case* piranha theorem: a bound on the number of large correlations with all possible response vectors. Stronger results can be obtained if we consider average behavior. Here, we consider a *stochastic* piranha theorem in which we assume that y is uniformly distributed on the unit sphere in $\mathbb{R}^n$. Our result will hold for any choice of radially symmetric random vector y (that is independent of

X), but we state it for the uniform distribution over the unit sphere for concreteness. We choose a radially symmetric distribution because we have no reason to give preference to one direction over another. Recall the value of studying average as well as worst-case behavior in areas of numerical analysis such as random matrix theory (Edelman and Rao, 2005).

The following theorem demonstrates this principle, showing that the maximum sum of squared correlations, an $O(1)$ quantity in n, is generally much larger than the expected sum of square correlations. Specifically, the following theorem shows that the expected sum of squared correlations decays like $1/n$.

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

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

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 $(Z_1, \ldots, Z_n)$ the n-dimensional multivariate Gaussian with mean zero and covariance $\frac{1}{n}I$. In particular, $(Z_1, \ldots, Z_n)$ is spherically symmetric and

$$\mathbb{E}(Z_1^2 + \cdots + Z_n^2) = 1 \qquad \text{and} \qquad \operatorname{var}(Z_1^2 + \cdots + Z_n^2) = O(1/n^2).$$

As a consequence, for large n, the distribution of sum of squared correlations is well approximated by a linear combination of independent χ^2 random variables, each with one degree of freedom: $\frac{1}{n-1}(\lambda_1^2 \xi_1 + \cdots + \lambda_p^2 \xi_p)$.

Combining this observation with Theorems 3 and 7, for any $n \times p$ matrix (or sample of data) X, if a vector y is distributed according to a spherically symmetric distribution, then

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

is supported on $[0, \sigma_1^2]$, has expectation $p/(n-1)$, and for large n has $O(1/n^2)$ variance.

6 Discussion and directions for future work

The piranha problem is a practical issue: as discussed in the references in Sections 1 and 3, it has interfered with research in fields including social priming, evolutionary psychology, economics, and voting behavior. An understanding of the piranha problem can be a helpful step in recognizing fundamental limitations of research in these fields along with related areas of application such as marketing and policy nudges (Carroll, 2017; Szászi et al., 2022). We suspect that a naive interpretation of the butterfly effect has led many researchers and policymakers to believe that there can be many large and persistent effects; thus, there is value in exploring the statistical reasons why this is not likely. In this way, the piranha problem resembles certain other statistical phenomena such as regression to the mean and the birthday coincidence problem (Mosteller, 1965), in that there is a regularity in the world that surprises people, and this regularity can be understood as a mathematical result. This motivates us to seek theorems that capture some of this regularity in a rigorous way. We are not all the way there, but this seems to us to be a valuable research direction.

6.1 Bridging between deterministic and probabilistic piranha theorems

Are there connections between the worst-case bounds in Section 4, the probabilistic bounds in Section 5, priors for the effective number of nonzero coefficients (Piironen and Vehtari, 2017), and models such as the R^2 parameterization of linear regression as proposed by 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 the parameters are not 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, 2021). 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; 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). More generally, models such as the horseshoe (Carvalho et al., 2010) assume a distribution of effect sizes with a sharp peak near zero and a long tail, which represent a solution to the piranha problem by allowing a large number of predictors without overflowing variance.

6.3 Nonlinear models

So far we have discussed linear regression, with theorems capturing different aspects of the constraint that the total R^2 cannot exceed 1. We can make similar arguments for nonlinear regression.

For example, consider a model of binary data with 20 causal inputs, each of which is supposed to have an independent effect of 0.5 on the logistic scale. Aligning these factors in the same direction would give an effect of 10, enough to change the probability from 0.01 to 0.99, which would be unrealistic in applied fields ranging from marketing to voting where no individual behavior can be predicted to that level of accuracy. One way to avoid these sorts of extreme probabilities would be to suppose the predictors are highly negatively correlated with each other, but in practice, input variables in social science tend to be positively, not negatively correlated (consider, for example, conservative political ideology, Republican party identification, and various issue attitudes that predict Republican vote choice and have positive correlations among the population of voters). The only other alternative that allows one to keep the large number of large effects is for the model to include strong negative interactions, but then the effects of the individual inputs would no longer be stable, and any effect would depend very strongly on the conditions of the experiment in which it is studied. It should be possible to express this reasoning more formally.

6.4 Implications for social science research

Although we cannot directly apply these piranha theorems to data, we see them as providing some relevance to social science reasoning.

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; see, for example, Open Science Collaboration (2015), Altmejd et al. (2019), and Gordon et al. (2020). Discussions of the replication crisis have 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 research reviewed in the present article 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.

The present article goes in a different direction, asking the theoretical question: under what conditions is it possible for many large effects to coexist in a multivariate system? In different ways, our results rule out or make extremely unlikely 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 raise suspicions about a model of social interactions in which many large effects are swimming around, just waiting to be captured by researchers who cast out the net of a quantitative study.

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.

Acknowledgements

References

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.

A. L. Alter and H. E. Hershfield. People search for meaning when they approach a new decade in chronological age. *Proceedings of the National Academy of Sciences*, 111:17066–17070, 2014.

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

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

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

L. Bartels. Here's how a cartoon smiley face punched a big hole in democratic theory. *Washington Post*, https://www.washingtonpost.com/news/monkey-cage/wp/2014/09/04/heres-how-a-cartoon-smiley-face-punched 2014.

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

R. Blanchard and A. F. Bogaert. Homosexuality in men and number of older brothers. *American Journal of Psychiatry*, 153:27–31, 1996.

A. F. Bogaert. Biological versus nonbiological older brothers and men's sexual orientation. *Proceedings of the National Academy of Sciences*, 103:10771–10774, 2006.

A. Bora, A. Jalal, E. Price, and A. G. Dimakis. Compressed sensing using generative models. *Proceedings of Machine Learning Research*, 70:537–546, 2017.

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*, 14:365–376, 2013.

D. R. Carney, A. J. C. Cuddy, and A. J. Yap. Power posing: Brief nonverbal displays affect neuroendocrine levels and risk tolerance. *Psychological Science*, 21:1363–1368, 2010.

A. E. Carroll. Don't nudge me: The limits of behavioral economics in medicine. *New York Times*, https://www.nytimes.com/2017/11/06/upshot/dont-nudge-me-the-limits-of-behavioral-economics-in-medicine, 2017.

C. M. Carvalho, N. G. Polson, and J. G. Scott. Handling sparsity via the horseshoe. *Proceedings of Machine Learning Research*, 5:73–80, 2009.

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

J. Cesario and D. J. Johnson. Power poseur: Bodily expansiveness does not matter in dyadic interactions. *Social Psychological and Personality Science*, 9:781–789, 2018.

K. 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.

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

J. Cornfield, W. Haenszel, E. C. Hammond, A. M. Lilienfeld, M. B. Shimkin, and E. L. Wynder. Smoking and lung cancer: Recent evidence and a discussion of some questions. *Journal of the National Cancer Institute*, 22:173–203, 1959.

L. J. Cronbach. Beyond the two disciplines of scientific psychology. *American Psychologist*, 30:116–127, 1975.

P. Ding and T. J. Vanderweele. Generalized Cornfield conditions for the risk difference. *Biometrika*, 101: 971–977, 2014.

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

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

A. Edelman and N. R. Rao. Random matrix theory. *Acta Numerica*, 14, 2005.

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

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

A. Fowler and B. P. Montagnes. Distinguishing between false positives and genuine results: The case of irrelevant events and elections. *Journal of Politics*, 2022.

J. Freese. The hurricane name people strike back! *Scatterplot*, https://scatter.wordpress.com/2014/06/16/the-hurricane-name-people-strike-back/, 2014.

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

A. Gelman. Disagreements about the strength of evidence. *Chance*, 28:55–59, 2015b.

A. 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-econom eats/, 2017.

A. Gelman. When you believe in things that you don't understand. *Statistical Modeling, Causal Inference, and Social Science*, https://statmodeling.stat.columbia.edu/2014/04/15/believe-things-dont-understand/, 2018a.

A. Gelman. Some experiments are just too noisy to tell us much of anything at all: Political science edition. *Statistical Modeling, Causal Inference, and Social Science*, https://statmodeling.stat.columbia.edu/2018/05/29/exposure-forking-paths-affects-support-publication/ 2018b.

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

A. Gelman and H. Stern. The difference between "significant" and "not significant" is not itself statistically significant. *American Statistician*, 60:328–331, 2006.

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

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

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

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

M. H. Graham, G. A. Huber, N. Malhotra, and C. H. Mo. Irrelevant events and voting behavior: Replications using principles from open science. *Journal of Politics*, 2022.

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

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

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

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.

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

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

A. T. B. Jessica L. Tracy. The impact of weather on women's tendency to wear red or pink when at high risk for conception. *PLoS One*, 9:e88852, 2014.

K. Jung, S. Shavitt, M. Viswanathan, and J. M. Hilbe. Female hurricanes are deadlier than male hurricanes. *Proceedings of the National Academy of Sciences*, 111:8782–8787, 2014.

D. Kahneman. *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux, 2011.

P. Krugman. Think of Mitch McConnell as a New Jersey shark. *New York Times*, 2021. URL https://www.nytimes.com/2021/09/28/opinion/republicans-government-shutdown-debt-limit.html.

S. Kühnea, T. Schneiderb, and D. Richter. Big changes before big birthdays? Panel data provide no evidence of end-of-decade crises. *Proceedings of the National Academy of Sciences*, 112:E1170, 2015.

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

X. Li and X.-L. 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*, 116:353–367, 2021.

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

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

E. Miguel, C. Camerer, K. Casey, J. Cohen, K. M. Esterling, A. Gerber, R. Glennerster, D. P. Green, M. Humphreys, G. Imbens, D. Laitin, T. Madon, L. Nelson, B. A. Nosek, M. P. andR. Sedlmayr, J. P. Simmons, U. Simonsohn, and M. van der Laan. Promoting transparency in social science research. *Science*, 343:30–31, 2014.

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

F. Mosteller. *Fifty Challenging Problems in Probability with Solutions*. New York: Dover, 1965.

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*, 9:285–298, 2018.

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

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*, 21:548–550, 2010.

M. B. Petersen, D. Sznycer, A. Sell, L. Cosmides, and J. Tooby. The ancestral logic of politics: Upper-body strength regulates men's assertion of self-interest over economic redistribution. *Psychological Science*, 24:1098–1103, 2013.

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

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

M. L. Sands. Exposure to inequality affects support for redistribution. *Proceedings of the National Academy of Sciences*, 114:663–668, 2017.

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.

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

B. Szászi, A. B. C. Anthony C. Higney, A. Gelman, I. Ziano, B. Aczel, D. G. Goldstein, D. S. Yeager, and E. Tipton. No reason to expect large and consistent effects of nudge interventions. *Proceedings of the National Academy of Sciences*, 119:e2200732119, 2022.

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

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

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

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

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

Y. D. Zhang, B. P. Naughton, H. 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*, 117:862–874, 2020.

A Proofs from Section 5

In this section, we give the proof of Theorem 7.

A.1 Notation

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

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

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

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

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

Write the singular value decomposition of X as

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

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 $\mathbb{1}^\mathsf{T}X_i = 0$ 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. $\square$

A.2 Proof of Theorem 7

We will take advantage of 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 $\lambda \in \mathbb{R}$,*

$$\sum_{i=1}^p \operatorname{corr}(X_i, y)^2 = \sum_{k=1}^p \lambda_k^2 (U_k^\mathsf{T}y^*)^2.$$

Proof. By direct computation:

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

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:

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

where

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

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,

$$\mathbb{E}\left(\sum_{i=1}^{p} \mathrm{corr}(X_i, y)^2 \right) \;=\; \mathbb{E}\left(\sum_{k=1}^{p} \lambda_k^2 (U_k^\mathsf{T} y^*)^2 \right) \;=\; \sum_{k=1}^{p} \lambda_k^2 \mathbb{E}(a_k^2) \;=\; \frac{1}{n-1} \sum_{k=1}^{p} \lambda_k^2$$

Since λ_i^2 are the eigenvalues of $X^t X$ and the columns of X have unit ℓ_2 norm,

$$\frac{1}{n-1} \sum_{k=1}^{p} \lambda_k^2 = \frac{p}{n-1}$$

because the trace of $X^t X$ is equal to the sum of its eigenvalues. $\qquad\square$