Research Article

The paradox of big data

Gary Smith[1]

Received: 20 December 2019 / Accepted: 29 April 2020 / Published online: 8 May 2020

Abstract
Background The data deluge seemingly makes it more likely that data mining will discover new, heretofore unknown relationships. *Findings* Monte Carlo simulations demonstrate the paradox of big data: the data deluge makes it more likely that the patterns and relationships discovered by data mining are spurious. *Conclusion* Models are more likely to be reliable if expert opinion is used in their specification, instead of viewing human expertise as an unhelpful constraint on knowledge discovery.

Keywords Data mining · Big data · Holdout data

1 Introduction

Kitchin [1] provides an insightful comparison of inductive, deductive, abductive, and data-driven epistemologies in the era of big data. He concludes that "it seems likely that the data-driven approach will eventually win out." Monte Carlo Simulations are used here to demonstrate and gauge some of the weaknesses in both the purely inductive and data-driven approaches.

2 Theory before data

Classical statistical tests begin with a falsifiable theory, followed by the collection of data for a statistical test of the theory. For example, it is known that heart attacks and strokes can be triggered by blood clots, and that aspirin inhibits blood clotting. A research hypothesis is whether regular doses of aspirin will reduce the chances of heart attacks and strokes. This conjecture was tested in the 1980s by a five-year, double-blind randomized control trial involving 22,000 doctors, with half taking an aspirin tablet every other day, and half taking a placebo. The doctors taking placebos had more than three times as many fatal heart attacks and nearly twice as many nonfatal heart attacks as the treatment group [2].

3 Data mining

Data mining goes in the other direction, putting data before theory [3, 4], indeed viewing the use of a priori knowledge as an unwelcome constraint that limits the possibilities for knowledge discovery [5, 6].

In an article titled, "The End of Theory: The data deluge makes the scientific method obsolete." Anderson [7] argued that

> Petabytes allow us to say: "Correlation is enough." We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot.

In the heart attack example, a data miner might compile a database of 1000 personal characteristics of 10,000 people who suffered heart attacks and 10,000 people who did not, and then use a data mining algorithm to identify

✉ Gary Smith, gsmith@pomona.edu | [1]Department of Economics, Pomona College, 425 N. College Avenue, Claremont, CA 91711, USA.

distinguishing characteristics. It might turn out that the heart attack victims were more likely to have a fondness for apples, work for city government, live in a small town, have green eyes, and use the word *great!* on Facebook.

The data mining researcher might conclude that apples, government jobs, small towns, green eyes, and Facebook *great!*s are unhealthy, and concoct some fanciful stories to explain these correlations. Or the data miner might argue that the data speak for themselves and that is all that needs to be said. Correlation supersedes causation.

Once considered a sin akin to plagiarism [8], data mining has been advocated in a wide variety of disciplines [9–11]. For example Liu and Tsyvinski [12] estimated 810 regression coefficients relating daily bitcoin returns to a wide variety of variables, including the effect of bitcoin returns on stock returns in the beer, book, and automobile industries. They reported that 63 of the 810 estimated coefficients were statistically significant at the 10 percent level, including the effect of bitcoin returns on stock returns in the paperboard-containers-and-boxes industry. They were undaunted by the likelihood that such correlations were likely temporary coincidences or the fact that they had fewer significant coefficients than would be expected if they had just correlated bitcoin returns with random numbers: "We don't give explanations, we just document this behavior." Smith [13] gives several other examples of unanchored data-mined research.

4 The Feynman trap

Richard Feynman [14], a Nobel Laureate physicist, once told an audience,

> You know, the most amazing thing happened to me tonight. I was coming here, on the way to the lecture, and I came in through the parking lot. And you won't believe what happened. I saw a car with the license plate ARW 357. Can you imagine? Of all the millions of license plates in the state, what was the chance that I would see that particular one tonight? Amazing!

The Feynman trap is the Achilles heal of data mining. A specific license plate may be highly improbable, a priori, but the existence of some license plate, after the fact, is a certainty. In the same way, a specific pattern may be highly improbable, a priori, but the existence of some pattern, after the fact, is a certainty. Calude and Longo [15] prove that "Complete disorder is an impossibility. Every large set of numbers, points or objects necessarily contains a highly regular pattern."

The larger the data set, the more likely we are to find patterns that are a priori highly improbable. A simple example is a streak of L consecutive heads (or tails) in a sequence of coin flips. Table 1 shows that as the number of flips increases, the probability of a streak of length 10 or more increases, and so does the length of the longest streak that has at least a 50 percent chance of occurring. If we consider all possible patterns (alternating heads and tails, two heads followed by one tail, three tails followed by two heads, and so on), some improbable pattern is inevitable in a large set of coin flips.

The same principle holds in models in which a set of explanatory variables is used to predict a response variable, in that the larger the number of potential explanatory variables, the higher is the probability that statistically significant coincidental relationships will be discovered. This is the paradox of big data: data mining is most seductive when there are a large number of variables, but the inevitable coincidental patterns in large data sets make it more likely that the statistical relationships discovered by data mining are spurious.

5 Prediction

Some data mining enthusiasts argue that the goal is prediction, not the estimation of casual effects [11, 16]. If there is a correlation between Facebook *great!*s and heart attacks, we do not need to know why these are correlated. It is enough to know that one predicts the other. The problem with this argument is that if there is no logical reason for a pattern, it is likely to be a temporary, spurious correlation—like heart attacks and Facebook *great!*s.

Consistently reliable predictions require an underlying causal structure. For example, a data mining program found a statistical relationship between the US stock prices and the daily high temperature in Curtin, Australia [17]. There was no underlying reason for this relationship, and it was useless for predicting stock prices with fresh data.

Models that reflect an underlying causal structure are useful precisely because they make trustworthy

Table 1 Longest streak, L

Number of flips	$P[L \geq 10]$	n such that $P[L \geq n] > 0.5$
10	0.002	3
100	0.087	7
250	0.212	8
500	0.385	9
1000	0.623	10
10,000	1.000	13
100,000	1.000	17
1,000,000	1.000	20

predictions. We do not need to know the exact reason why two things are related, but there needs to be a reason in order for the model to make reliable predictions.

Suppose that a thousand pebbles of various sizes, shapes, colors, and densities are created by a 3D printer that has been programmed so that the characteristics are independently determined. A data mining algorithm looking for patterns in these pebbles might discover that yellow pebbles happen to have bumps more often than do pebbles with other colors. When the algorithm is used with fresh data, it is likely to fail because the correlation between color and bumpiness was merely coincidental. Now suppose, instead, that pebbles are found at the bottom of a lake and there is a scientific reason why bumpy pebbles in this lake tend to be yellow. In this case, the correlation between bumpiness and yellowness may be a useful predictor even if we do not completely understand why bumpy pebbles are often yellow.

The crucial distinction between these two scenarios is whether there is a structural relationship—a reason why bumpy pebbles are often yellow. We may not know the reason, but it is the existence of a structural relationship that makes predictions useful. Correlation is not enough. Causality is crucial.

6 Holdout data

It has been argued that splitting the data into an in-sample training set and out-of-sample test set can be used to determine whether a discovered relationship is merely coincidental [11, 18]. More formally, a sequence of exploratory data analysis, hypothesis generation, and confirmation testing has been labeled a *data-intensive* or *data-driven* methodology [1, 19].

In the pebble example, training the algorithm on 500 in-sample pebbles and then testing the algorithm on 500 out-of-sample pebbles is likely to show the fragility of the yellow/bumpiness correlation if there is no underlying systematic relationship.

If that happens, however, the data mining algorithm can keep looking for other patterns until it finds one that makes successful predictions with both the in-sample and out-of-sample data—and it is certain to succeed if there are a large enough number of characteristics. Just as spurious correlations can be discovered with a subset of the data, so spurious correlations can be discovered with a full set of data.

Consider the calculation of the correlation between two variables with statistical significance considered to be a p value below 0.05. Even if every value of every variable is just an independently generated random number, we expect 1 out of every 20 correlations to be statistically significant in-sample and 1 out of every 400 to be statistically significant both in-sample and out-of-sample. Out-of-sample validation is not a guarantee of the validity of data-mined patterns that are discovered in-sample.

Out-of-sample tests are surely valuable; however, data mining with a holdout test is still data mining and still subject to the same pitfalls. Using data mining to identify a statistical relationship that holds for all the data is more difficult than identifying one for half the data, but it is nonetheless inevitable if a large enough data set is data-mined, and gives no assurance that the identified relationship will be useful for making predictions with fresh data.

Sometimes, experiments can be run and rerun in order to generate essentially unlimited data that will eventually expose the transitory nature of patterns and relationships that are spurious. Repeated testing is more challenging for models that rely on observational data [20]. In addition, there is little incentive for researchers to spend time and resources attempting to replicate studies—either their own or others. When replication attempts are made, the results are often disappointing [21–26].

7 Crowding out

There is a more subtle problem with data mining tempered by holdout tests. Suppose that a data mining algorithm is used to select predictor variables from a data set that includes a relatively small number of "true" variables that are structurally related to the variable being predicted and a large number of "nuisance" variables that are independent of the variable being predicted. One problem is that some nuisance variables are likely to be coincidentally successful both in-sample and out-of-sample, but then flop when the model goes live with new data.

A second problem is that a data mining algorithm may select nuisance variables in place of true variables that could be used to make reliable predictions. The testing and retesting of a data-mined model may expose the nuisance variables as useless, but will not resuscitate the true variables that were crowded out by the nuisance variables. The more nuisance variables that are initially considered, the more likely it is that some true variables will disappear without a trace.

8 Methods

Monte Carlo simulations were used to explore the perils of data mining. A total of n observations for each of m candidate explanatory variable were determined by random draws from a normal distribution with mean 0 and standard deviation σ_x:

$$X_{i,j} = \varepsilon_{i,j}\,\varepsilon \sim N[0, \sigma_x] \quad (1)$$

The independence of the explanatory variables ensures that there are no structural relationships among the explanatory variables that might cause some variables to be proxies for others.

The central question is how effective the estimated model is at making reliable predictions with fresh data. So, in each simulation, half the observations were used to estimate the model's coefficients, and the remaining half were used to test the model's reliability.

All the data were centered by subtracting the sample means. The in-sample data were centered on the in-sample means and the out-of-sample data were centered on the out-of-sample means so that the out-of-sample predictions would not be inflated if the in-sample and out-of-sample means differed.

Five randomly selected explanatory variables (the *true* variables) were used to determine the value of a dependent variable *Y*,

$$Y_j = \sum_{i=1}^{5} \beta_i X_{i,j} + v_j \quad v \sim N[0, \sigma_y] \quad (2)$$

where the value of each β coefficient was randomly determined from a uniform distribution ranging from 2 to 4, and *v* is normally distributed with mean 0 and standard deviation σ_y. The range 0 to 2 was excluded so that the true variables would have substantial effects on the dependent variable. The other candidate variables are *nuisance* variables that have no effect on *Y*, but might be coincidentally correlated with *Y*.

The base case was $\sigma_x = 5$, $\sigma_y = 20$, $m = 100$ candidate variables; and $n = 200$ observations, but I also considered all combinations of $\sigma_x = 5$, 10, or 20; $\sigma_y = 10$, 20, or 30; $m = 5$, 10, 50, 100, 500, or 1000; and $n = 100$, 200, 500, or 1000. For the range of values considered here, the results were robust with respect to the values of σ_x and σ_y, so I only report results for the base case, $\sigma_x = 5$ and $\sigma_y = 20$.

9 Stepwise regression

One obstacle to data mining for the best-fitting model is the sheer number of models that can be considered. A researcher who wants to choose up to 10 out of 100 possible explanatory variables has 19.4 trillion possible combinations to choose from. With 1000 possible explanatory variables, there are 2.66×10^{23} combinations of up to 10 variables. So, in practice, various work-arounds are used.

Forward-selection stepwise regression chooses the explanatory variables from a group of candidate variables by starting with no explanatory variables and then adding variables, one by one, based on which variable is the most statistically significant, until there are no remaining statistically significant variables. The use of stepwise regression to identify the ten most statistically significant explanatory variables requires only 955 regressions if there are 100 candidate variables, and 9955 regressions if there are 1000 candidates. Thus, stepwise regression is often recommended as an efficient way of using data mining for knowledge discovery [27–30].

One hundred thousand simulations were done for each parameterization of the model. Every step of the stepwise regression procedure added the candidate explanatory variable with the lowest two-sided *p* value if this *p* value was less than 0.05. The results were not due to the use of stepwise regression to select the explanatory variables, but rather to data mining. Stepwise regression is simply a practical data mining tool for identifying explanatory variables that are statistically correlated with the variable being predicted.

In one set of simulations, all of the candidate dependent variables were nuisance variables. In the second set of simulations, five of the candidate variables were the five true variables that were used to generate the values of the dependent variable. The first set of simulations, with entirely spurious variables, explores the extent to which coincidental correlations with the dependent variable can create an illusion of a successful prediction model. The second set of simulations, in which the true explanatory variables are among the candidate variables, explores how well a data mining procedure is able to distinguish between meaningful and meaningless variables.

The predictive success of the model was gauged by the correlation between the actual values of the dependent variable and the model's predicted values. The square of the in-sample correlation is the coefficient of multiple determination, R^2, for the estimated model. The out-of-sample correlation is the corresponding statistic using the out-of-sample data with the in-sample estimated coefficients.

10 Results

10.1 Including no true variables

Tables 2 and 3 report the results of the Monte Carlo simulations with no true variables included among the candidate variables. Table 2 shows that even though all of the potential explanatory variables are nuisance variables, a data mining procedure that selects variables based on statistical significance can include a substantial number of nuisance variables. With a small number of observations, the number of selected variables is constrained by

Table 2 Average number of variables per equation, no true variables

Observations to used estimate model	Number of candidate variables					
	5	10	50	100	500	1000
50	1.12	1.29	3.40	9.32	47.58	47.99
100	1.11	1.27	3.05	6.63	97.79	96.79
500	1.11	1.25	2.78	5.30	36.38	98.70
1000	1.11	1.25	2.74	5.18	29.76	73.92

Table 3 In-sample correlation, no true variables

Observations to used estimate model	Number of candidate variables					
	5	10	50	100	500	1000
50	0.344	0.365	0.551	0.784	1.000	1.000
100	0.244	0.258	0.385	0.549	1.000	1.000
500	0.109	0.115	0.169	0.234	0.582	0.852
1000	0.078	0.081	0.119	0.164	0.392	0.586

Table 4 Average number of variables per equation, five true variables

Observations to used estimate model	Number of candidate variables					
	5	10	50	100	500	1000
50	3.15	3.39	5.90	12.06	47.66	47.99
100	4.50	4.74	6.99	10.71	97.84	97.88
500	5.00	5.25	7.30	9.96	40.89	97.68
1000	5.00	5.25	7.28	9.83	34.66	80.54

Table 5 Chances of being a nuisance variable, five true variables

Observations to used estimate model	Number of candidate variables					
	5	10	50	100	500	1000
50	0.000	0.084	0.520	0.783	0.968	0.979
100	0.000	0.055	0.371	0.599	0.962	0.969
500	0.000	0.047	0.315	0.498	0.878	0.949
1000	0.000	0.047	0.313	0.491	0.856	0.938

the number of observations but, otherwise, the number of included variables increases as the number of candidate variables increases and declines as the number of observations increases—presumably because the more accurate estimates are more likely to be close to zero. Yet, even with 1000 observations on each variable, many nuisance variables are included in the model when a large number of candidate variables are considered, confirming the argument of Calude and Longo.

Table 3 shows how the data mining algorithm discovered coincidental patterns that create a false impression of success. For example, with 100 nuisance variables, the average in-sample correlation between the predicted and actual value of the response variable was 0.549 with 100 observations and 1.000 with 1000 observations. No matter what the in-sample correlation, it is misleading large because these are, after all, nuisance variables that are independent of the dependent variable. The out-of-sample correlations between the dependent variable and the selected explanatory variables averaged zero, yet there were many individual models in which the out-of-sample correlations were as high or higher than the in-sample correlations, suggesting that the model is useful for predictions when it is, in fact, useless.

10.2 Including five true variables

Tables 4, 5, 6, 7 report the results of the Monte Carlo simulations with the five true variables included among the candidate variables. Table 4 shows that the inclusion of five true variables did not eliminate the selection of nuisance variables. More often, the total number of variables that

Table 6 In-sample correlation, five true variables

Observations to used estimate model	Number of candidate variables					
	5	10	50	100	500	1000
50	0.639	0.652	0.758	0.884	1.000	1.000
100	0.657	0.663	0.714	0.780	1.000	1.000
500	0.650	0.651	0.661	0.674	0.788	0.911
1000	0.648	0.649	0.654	0.660	0.715	0.791

Table 7 Out-of-sample correlation, five true variables

Observations to used estimate model	Number of candidate variables					
	5	10	50	100	500	1000
50	0.509	0.491	0.380	0.280	0.131	0.097
100	0.606	0.600	0.543	0.478	0.266	0.245
500	0.642	0.641	0.631	0.618	0.505	0.400
1000	0.645	0.644	0.639	0.632	0.577	0.503

were included increased. Table 5 confirms that an expansion in the number of candidate variables increased the chances that an included variable is a nuisance variable. For example, with 100 observations for 100 candidate variables (5 true and 95 nuisance), the probability that an included variable is a nuisance variable was 0.599. This probability approached 1 as the number of candidate variables increased. With 100 observations on 500 variables, the average number of included variables was 96.82 and, on average, 0.963(96.82) = 93.23 of the included variables were nuisance variables and 3.59 were true variables.

These simulations also document how a plethora of nuisance variables can crowd out true variables. With 100 observations and 100 candidate variables, for example, one or more true variables were crowded out 50.2 percent of the time, and two or more variables were crowded out 16.0 percent of the time. There were even occasions when all five true variables were crowded out.

Tables 6 and 7 show the in-sample and out-of-sample correlations. Comparing Tables 3 and 6, the in-sample correlations were substantially higher with the five true variables among the candidate variables if there were few candidate explanatory variables or a large number of observations. The out-of-sample correlations (Table 7) with five true variables were substantially lower than the in-sample correlations (Table 6) when the number of candidate variables was large relative to the number of observations. For example, with 50 observations for 1000 candidate variables, the average correlation was 1.000 in-sample and 0.097 out-of-sample.

11 Discussion

The simulations reported here involve up to 1000 observations on up to 1000 explanatory variables, but the conclusions clearly apply to even larger data sets.

Data-mined models can make statistically persuasive in-sample predictions even when they only consider irrelevant explanatory variables. Out-of-sample validation is unreliable because some data-mined models using nothing but nuisance variables will, by luck alone, be statistically impressive both in-sample and out-of-sample.

The inclusion of true explanatory variables among the candidate explanatory variables makes it more likely that a model will be useful for out-of-sample predictions, but this usefulness is undermined by the inclusion of nuisance variables that crowd out true variables.

Data mining algorithms cannot effectively distinguish between true variables and nuisance variables because computer algorithms do not know what variables are in any meaningful sense and cannot assess whether there is an underlying causal relationship [17]. Currently, only human expertise can do that.

The increasing reliance on data mining algorithms to build models unguided by human expertise may be partly responsible for the reproducibility crisis, in which attempts to replicate published research findings often fail. Results reported with data-mined models are

inherently not reproducible, since they will almost certainly include nuisance variables that improve the model's in-sample fit, while undermining the out-of-sample success.

12 Conclusion

The Monte Carlo simulations reported here illustrate the validity of the theoretical argument of Calude and Longo [9] that spurious correlations are endemic in large data sets and, consequently, "Too much information tends to behave like very little information," in that the performance of data-mined models is more likely to be misleading when a large number of variables are analyzed.

Models are more likely to be useful and the results are more likely to be reproducible if expert opinion is used to select a plausible list of explanatory variables, instead of viewing human expertise as an unhelpful constraint on knowledge discovery. This is a corollary of the paradox of big data: the larger the number of possible explanatory variables, the more important is human expertise.

Bayesian methods provide a flexible alternative to using domain knowledge to select explanatory variables. Instead of forcing variable coefficients to either be zero or unconstrained, prior distributions can be used to convey expert knowledge and uncertainty. Ridge regression and other penalized regression procedures that shrink the estimated coefficients toward zero implicitly use spherical prior distributions for the coefficients, though other distributions may be preferred [31, 32].

The reliance on computer algorithms to select useful explanatory variables will continue to be problematic until computers acquire the common sense, wisdom, and expertise needed to distinguish between meaningful and meaningless patterns.

Availability of data and materials The data came from millions of Monte Carlo simulations. The computer code is available on request.

Compliance with ethical standards

Conflict of interest There are no financial or non-financial competing interests.

References

1. Kitchin R (2014) Big data, new epistemologies and paradigm shifts. Big Data Soc. https://doi.org/10.1177/2053951714528481
2. Steering Committee of the Physicians' Health Study Research Group (1988) Preliminary report: findings from the aspirin component of the ongoing physicians' health study. N Engl J Med 28:262–264
3. Fayyad U, Piatetsky-Shapiro G, Smyth P (1996) From data mining to knowledge discovery in databases. AI Mag 17(3):37–54
4. Begoli E, Horsey J (2012) Design principles for effective knowledge discovery from big data .In: Software architecture (WICSA) and European conference on software architecture (ECSA), 2012 Joint working IEEE/IFIP conference
5. Piatetsky-Shapiro G (1991) Knowledge discovery in real databases: a report on the IJCAI-89 workshop. AI Mag 11(5):68–70
6. Sagiroglu S, Sinanc D (2013) Big data: a review, collaboration technologies and systems (CTS). In: 2013 International conference
7. Anderson C (2008) The end of theory, will the data deluge make the scientific method obsolete? Wired Mag 16(7):16
8. Smith G, Cordes J (2019) The 9 pitfalls of data science. Oxford University Press, Oxford
9. Kohler TA (2018) Our unfinished agenda (what i have learned). SAA Archaeol Rec 18(5):37–42
10. Grover V, Lyytinen K (2015) New state of play in information systems research: the push to the edges. MIS Q 39(2):271–296
11. Athey S (2018) The impact of machine learning on economics. In: Agrawal A, Gans J, Goldfarb A (eds) The economics of artificial intelligence: an agenda. University of Chicago Press, Chicago
12. Liu Y, Tsyvinski A. 2018. Risks and returns of cryptocurrency, working paper, August 13, https://ssrn.com/abstract=3226952, NBER Working Paper No. 24877, August 2018
13. Smith G (2020) Data mining fool's gold. J Inf Technol, forthcoming
14. Goodstein DL (1989) Richard P. Feynman, teacher. Phys Today 42(2):70–75
15. Calude CS, Longo G (2017) The deluge of spurious correlations in big data. Found Sci 22(3):595–612. https://doi.org/10.1007/s10699-016-9489-4
16. Mullainathan S, Spiess J (2017) Machine learning: an applied econometric approach. J Econ Perspect 31(2):87–106
17. Smith G (2018) The AI delusion. Oxford University Press, Oxford
18. Egami N, Fong CJ, Grimmers J, Roberts, ME, Stewart BM (2018) How to make causal inferences using text. arXiv:1802.02163 v12016
19. Kelling S, Hochachka W, Fink D, Riedewald M, Caruana R, Ballard G, Hooker G (2009) Data-intensive science: a new paradigm for biodiversity studies. Bioscience 59(7):613–620. https://doi.org/10.1525/bio.2009.59.7.12
20. Arnott RD, Harvey CR, Markowitz H (2018) A backtesting protocol in the era of machine learning, November 21. SSRN: https://ssrn.com/abstract=3275654
21. Open Science Collaboration (2015) Estimating the reproducibility of psychological science. Science 349:aac4716
22. Camerer CF et al (2016) Evaluating replicability of laboratory experiments in economics. Science 351:1433–1436
23. Camerer CF, Dreber A, Holzmeister F et al (2018) Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. Nat Hum Behav 2(9):637–644
24. Ioannidis JA (2005) Contradicted and initially stronger effects in highly cited clinical research. J Am Med Assoc 294(2):218–228
25. Pashler H, Wagenmakers E (2012) Editors' introduction to the special section on replicability in psychological science: a crisis of confidence? Perspect Psychol Sci 7(6):528–530
26. Baker M (2017) 1,500 scientists lift the lid on reproducibility. Nature 533(7604):452–454
27. Cios KJ, Pedrycz W, Swiniarski RW, Kurgan LA (2007) Data mining: a knowledge discovery approach. Springer, New York
28. Varian HR (2014) Big data: new tricks for econometrics. J Econ Perspect 28(2):3–27
29. Hastie T, Tibshirani R, Friedman J (2016) The elements of statistical learning, 2nd edn. Springer, New York

30. Bruce P, Bruce A (2017) *Practical statistics for data scientists: 50 essential concepts*. O'Reilly Media, Sebastopol
31. Smith G, Campbell F (1980) A critique of some ridge regression methods. J Am Stat Assoc Discuss Rejoinder 75(369):74–81
32. Leamer EE (1981) Coordinate-free ridge regression bounds. J Am Stat Assoc 76(376):842–849