Philosophy of Science

Exploratory Research is More Reliable than Confirmatory Research

--Manuscript Draft--

Manuscript Number:	
Full Title:	Exploratory Research is More Reliable than Confirmatory Research
Article Type:	Article
Corresponding Author:	Clark Glymour, Ph.D. UNITED STATES
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	
Corresponding Author's Secondary Institution:	
First Author:	Clark Glymour, Ph.D.
First Author Secondary Information:	
Order of Authors:	Clark Glymour, Ph.D.
Order of Authors Secondary Information:	
Abstract:	Statistical methodology distinguishes between "confirmatory" research, in which individual hypotheses are proposed by a researcher and then tested with some data set, and "exploratory" research in which a very large number of hypotheses, generated from the data, are tested or otherwise assessed. The dominant opinion is that exploratory research is less reliable than confirmatory research, but several authors have argued that confirmatory research produces a high proportion of false, non-reproducible relationships because true hypotheses in a domain are rare, scientists are little better than chance at guessing true hypotheses to test among the many possibilities, and the conventional .05 level tests and low power result in many false positive relationships. It is argued and illustrated here that none of these objections bear against contemporary computerized search methods, and the rarity of true hypotheses among all of those possible in a domain is actually an advantage for these "exploratory" methods.

Exploratory Research is More Reliable than Confirmatory Research

Abstract

Statistical methodology distinguishes between "confirmatory" research, in which individual hypotheses are proposed by a researcher and then tested with some data set, and "exploratory" research in which a very large number of hypotheses, generated from the data, are tested or otherwise assessed. The dominant opinion is that exploratory research is less reliable than confirmatory research, but several authors have argued that confirmatory research produces a high proportion of false, non-reproducible relationships because true hypotheses in a domain are rare, scientists are little better than chance at guessing true hypotheses to test among the many possibilities, and the conventional .05 level tests and low power result in many false positive relationships. It is argued and illustrated here that none of these objections bear against contemporary computerized search methods, and the rarity of true hypotheses among all of those possible in a domain is actually an advantage for these "exploratory" methods.

1. Introduction.

For decades the statistical literature has contrasted "exploratory" research with "confirmatory" research, and claimed or implied that exploratory research is less reliable, more likely to lead to false claims of dependence or causality. Even in the face of the increasing necessity for computerized, data-driven inference methods to analyze 'big data," the drumbeat continues in a influential manifesto by Ioannidis (2005) claiming that "Most Published Research Is False," and in a recent statement of the American Statistical Association (2016) on the use of "p values" in statistical tests. In the first instance, these criticisms address conventional "confirmatory" hypothesis assessment, but they are commonly emphasized for "exploratory research" for causal relations. To the contrary, for causal discovery, exploratory research, done right, is more reliable than is confirmatory research; the objections made to confirmatory research are the virtues of exploratory research. "Confirmatory" research is the testing or (to include Bayesians) assessment on a body of data of one, or a handful, of hypotheses somehow obtained without regard for the data on which they will be tested. "Exploratory" research is any in which the hypotheses that are assessed on a body of data are functions of the data. The mathematical relation between a hypothesis and data used in assessments is often the same in exploratory and confirmatory research, but while confirmatory research examines one or a handful of hypotheses identified by an investigator, exploratory research sometimes assesses billions of hypotheses found by computer. The large data sets used in contemporary exploratory research in astronomy, climate science, genomics, neuroscience and, increasingly, social science, are typically not experimental, in the sense that the variables whose causal connections are investigated have not been directly controlled.

2. The Critical Arguments

John Ioannidis' essay "Most Published Research Findings are False" prompted a healthy focus on reliability and reinforced doubts about "confirmatory" and "exploratory" research. David Colquohoun (2014) argued by simulation that the false discovery rate of "significant" positive results using a .05 significance level in confirmatory research may be 30% or more. The reservations Ioannidis and others express about the reliability of exploratory research, however, rest on a misunderstanding of modern methods for estimating causal relations from very high dimensional observational and experimental data. Ioannidis and Colquohoun's arguments are one side of the argument given here that exploratory research from "big data" using appropriate methods is more reliable than confirmatory research when the frequency of true hypothesis among those that might be considered in a domain of inquiry is low. Ioannidis' and Colquhoun's concern was with the base rate of true causal relationships in a domain of inquiry. The methodology they criticize is: Select hypothesis, H, that variables A, B are causally connected; have data D; chose test statistic S and level alpha to test null hypothesis that A, B are independent; reject the null if S(D) < alpha; if the null is rejected, report that H is confirmed. They generalize an old argument that notes that reporting only "positive" results of hypothesis tests of multiple independent variables would result in reporting only false results. Assuming there are a great many unspecified claims that might be made in a subject domain, of which only a small fraction are true, and the selection of hypotheses to test is independent of their truth, Ioannidis argues that a positive finding of a relationship is more likely to be true than false just if $(1-\beta)R > \alpha$, where $(1-\beta)$ is the power of the test, α is the cut-off or alpha value for rejecting the null, and R is the proportion of true relationships to false (no relationship) in the domain. His claim is that in most studies the power may not be very high and R maybe very low and so the inequality will not hold. Colquhoun reinforces the argument with simulations, again assuming that the true hypotheses are rare and selection of hypotheses to test is independent of their truth. He estimates that 30% of positive results reported using a .05 significance level are false.

The significance level in psychological and social science research is commonly put at .05, although lower (.01) and higher (.1) values are sometimes used. By contrast, the "five sigma" cutoff used in experimental physics corresponds to 3×10^{-7}, a much more severe test. Why not use five sigma in social science and psychology? Reported p values of tests in psychology and social science are rarely so small. If Ioannidis and Colquohoun are correct that true hypotheses are rare in this subject, and researchers are no better than chance at selecting true hypotheses to test—and even if they are somewhat better than chance--then the answer seems plain: a psychologist or social scientist, testing one hypothesis at a time would rarely if ever in a career have a hypothesis to publish. The entire evaluation structure of academic social and behavioral science careers would collapse.

There is another respect in which confirmatory research is not severe, or not severe enough. A common complaint is that results found on one sample of subjects are not found with other samples from even slightly different populations. One obvious remedy is to have larger and more diverse samples, but that is not enough. A relationship might hold on average but fail to hold in relevant subpopulations. A relationship might seem to hold but not be causal because it is due to common causes that have not been measured. Not only are large, diverse samples of people needed, but also large, diverse measures of their diversity. But with more variables, there are more possible confoundings to test, exponentially more. The number of possible sets of common causes of any 2 of N variables is $2^{(N-2)}$. For big N, no one could do all of the tests one at a time, and even it such tests were done, Ioannidis and Colquhoun's arguments would bear on them. Severity in confirmatory research appears impossible.

3. Modern Search Methods

Ioannidis claims that false positives are especially likely to result from "exploratory" research where multiple hypotheses are examined: "The greater the number and the lesser [sic] the selection of tested relationships in a scientific field, the less likely the research findings are to be true. Thus research findings are more likely true in confirmatory designs…than in hypothesis generating designs" [because in exploratory studies a lot of false hypotheses are tested, and the more that are tested, the more errors will be made.] "Fields considered highly informative and creative given the wealth of the assembled and tested information, such as microarrays and other high-throughput discovery oriented research…should have extremely low PPV" [Positive Predictive Value, the probability that a reported result is true]. (p. 0698). The claim might be correct if modern search methods used the strategy Ioannidis and Colquohoun presuppose, but they do not. The first point to note is that search methods are in one respect familiar animals: they are statistical estimators. No one would seriously suggest that the outputs of statistical estimators should not depend on the data, or that the only appropriate procedure for parameter estimation is to first guess the value of a parameter in ignorance of the data and then test the estimate. But that is what critics wrongly assume the computerized search for causal relations must be. Exploratory research for causal relationships is statistical parameter estimation, sometimes with the aid of hypothesis tests, sometimes not. Statistical estimation concerns single parameters (e.g., a mean), vectors of parameters (e.g., mean and variance), and matrices of parameters (e.g., covariances). Causal inference in systems with a large number of variables ("high dimensional" systems) is matrix estimation. Suppose a large number N of variables are measured for each of n units or cases. Assume for the moment that unobserved common causes are not countenanced. Then a causal theory of the system is given by entries in an N X N matrix, where a 1 entry indicates that the row variable causes the column variable, and a 0 indicates otherwise. The inference problem is to estimate that matrix from the sample data. Where unmeasured common causes are entertained, third and fourth values for entries are allowed, indicating respectively a common cause but no cause from the row to the column variables, or both a common cause and a cause from the row to the column

variables. Variations on these forms of matrix estimation are common. Well-known procedures estimate the topological structure of a directed graph representing causal relations. For these procedures the matrix is different. An entry of 1 indicates that the row variable is a parent of the column variable in the graph. An entry of "-" indicates that *either* the row is a parent of the column *or* vice versa. Inclusion of unmeasured confounders is as above. Assuming transitivity, a matrix of causal connections can be generated from a matrix of direct causal connections. Still more elaborate estimation problems occur in exploratory research. Consider the problem of estimating from among a set of measured variables those subsets such that each subset has a single, unmeasured common cause, and estimating the direct causal connections among those unmeasured variables. This too, can be viewed as a parameter estimation problem for entries in a multidimensional matrix, but ugly enough that it need not be detailed. Methods for exploratory research are seldom presented as estimation problems, and while matrix representations of the parameters are sometimes referred to, they are never presented, but they lay behind the explorations.

Exploratory causal search is just parameter estimation; a causal search algorithm is just an estimator. As with more familiar estimators, various questions about convergence to the truth as a function of sample size can be asked: Does the estimator give a point value or a set of values? Given a true joint distribution of the measured variables, under what conditions does the algorithm return true information? As the sample size from a population increases without bound, does the estimator converge to true information? If it converges, does it do so uniformly, that is provide a sample size for which the probability of error is within some bound? Such questions have been asked and answered for several causal search algorithms, and they are well-posed for any new ones that will surely be developed.

Methods for "exploratory" search for causal relations, available for 25 years but only recently reconfigured for "big data," do not follow the "one hypothesis—one test" paradigm, and with high dimensional data, the methods that use a sequence of hypothesis tests do not use a .05 significance level. Their estimates are made by strategies that, while some of them ultimately depend on a complex series of Bayesian estimates or hypothesis tests, impose severe conditions on any positive causal claim. In particular, for the two procedures illustrated in the subsequent examples:

- Each positive causal claim is tested or assessed multiple times, against multiple competing hypotheses.
- The procedures are biased *against* positive results.
- The procedures have an adjustable bias against weak effects and in favor of strong effects, and can be used in various ways to find the variables with the strongest total effect size for an outcome of interest.
- The reverse of the concern about rare positive relations holds: the procedures are most reliably accurate, most informative, and most feasible when the true positive causal relations are rare.

Search algorithms for causal relations can try to estimate a directed graph (DAG), or try to estimate a class of "indistinguishable" directed graphs. For acyclic directed graphs (DAGS) the Markov equivalence class of a DAG is the set of all graphs that imply the same conditional independence relations; two DAGs are in the same Markov equivalence class if they have the same adjacencies when directions are ignored, and if they share the same "unshielded colliders"—triples X -> Y <- Z where X and Z are not adjacent. A Markov equivalence class is essentially a set of alternative explanations of a data generating process. There are several automated strategies for identifying a Markov equivalence class from sample data. One procedure, PC-Max (Ramsey, 2016), uses a sequence of conditional independence tests. Another, the Fast Greedy Equivalence Search (FGES) algorithm (Ramsey, et al., 2016), uses a quasi-Bayesian score, the Bayes Information Criterion (BIC) score. Their search strategies are quite different. FGES starts with no connections between variables, builds up a Markov equivalence class edge by edge, and when the BIC score can no longer be improved, tries removing edges to improve the score. Given their assumptions, the PC-Max and FGES algorithms are asymptotically (in the sample size) "pointwise" correct ("consistent") in the sense that in the large sample limit they converge to returning the set of directed acyclic graphs data that imply the same conditional independence relations as the graph of the causal relations of the structure that generated the data. PC-Max, is uniformly consistent for sparse graphs and appropriate rates of growth of sample sizes. Aside from asymptotic theory, there are procedural advantages over the one hypothesis-one test methodology, advantages that defeat the objections to "exploratory methods. For example, FGES estimates a series of posterior probabilities as it dynamically changes the class of models it considers, starting with an empty graph and adding stepwise the edge that most improves the Bayes Information Criterion (BIC) score and then removing edges until that score cannot be further improved. The BIC score is-2ln L + c k ln(n) where ln is the natural logarithm, L is the maximum likelihood estimate, k is number of free parameters of the model, and n is the sample size. The factor c is usually set at 1, but can be any multiple of 1 without affecting asymptotic consistency. A lower BIC score is better. The c k ln(n) term constitutes a bias

against positive results. The greater the value of c, the greater is that bias; with higher values of c, fewer connections are found, and those which give the data the greater likelihood are the more likely to be found. A sense of both the severity and computational demands of the FGES criteria can be gained by considering what it requires to add a first edge in a million variable problem. There are approximately 500,000,000,000 + 1 possible edges to add, including adding no edge (the + 1). Only the edge (or no edge) with the highest BIC score is added. The BIC score has an adjustable term, c above, that penalizes a model for its number of edges. By increasing the penalty, the search can (and for very high dimensional data generally must) be biased *against* positing causal connections. In contrast, PC-Max begins with a complete, undirected graph and removes edges sequentially. Undirected edges are then directed, where possible, to form a Markov equivalence class. In the first step, all edges are removed between any pair of uncorrelated (or more generally, independent) variables, again requiring 500,000,000,000 tests in a million variable problem. Of course, false connections may remain. The procedure next tests the independence of every pair of variables left adjacent from the first step, conditional on each single variable remaining adjacent to any one of them. For variables left adjacent after the second step, the procedure tests their independence conditional on every *pair* of variables adjacent to one or the other of them. And so on. The undirected connections are then directed, where possible, by attending to whether in any surviving triple X – Y – Z of undirected edges, with X, Z not adjacent, Y was or was not conditioned on in removing the X – Z adjacency. Each connection reported has survived a multitude of tests, effectively on different subsets of the data.

4. Examples

Precision is the frequency with which a reported result is true. Recall is the frequency with which true relations are found. In the examples that follow, all of the precisions and recalls are with respect to the Markov equivalence class.

For two random directed acyclic graphs of average degree 2 (the average number of edges attached to a variable) with 1000 Gaussian variables, parameterized as a linear model with unit variance and mean zero disturbance terms, coefficients drawn uniformly from [-1.5 - -0.5] ∪ [0.5 – 1.5], and sample size 1000, testing for independence of 100 randomly chosen pairs of variables using Fisher Z and significance cutoffs of .01 and .05, respectively, resulted in 57% and 59% false positives, in accord with the concerns of Ioannidis and Colquohoun. Lowering the alpha level further would of course reduce the proportion of false positives at the cost of decreasing recalls. With contemporary search methods on similar and even more difficult problems the results are quite different. On the same problem an improvement on the PC algorithm, PC-Max, returns 100 percent precision over 10 repetitions with novel graphs and parameter values in each case, and 96% recall. Alpha is chosen roughly by order of magnitude of the number of variables, in this case .001. When the density of true positives is doubled by randomly generating ten directed acyclic graphs of average degree 4, the precision is unchanged and recall falls to 85%. These numbers do not reflect the accuracies with which the directions of edges are found. Those are 98% and 97% average precision in the 1000 variable degree 2 and degree 4 cases respectively, and 99% in the 20,000 variable case. Recalls are lowest, 81%, for the degree 4, 1000 variable case.

Besides search with appropriate sequences of tests, there are quasi-Bayesian search algorithms suitable for high dimensional problems, notably the Fast Greedy Equivalence Search. Ramsey randomly generated sparse, linear causal models with 1,000 to 1,000,000 Gaussian variables, parameterized as above. Sample size was 1,000 in all runs. The condition warned against by Ioannidis holds: the sought after positive truths of causal connections are extremely rare. The FGES results are given in Table 1 as percentages of directed and directed edges in the Markov equivalence class of the true directed acyclic graph (Adj is adjacencies, meaning one or the other variables of a pair directly influences the other, Arr is direction of influence; Prec is precision, the frequency with which claimed relationships are true; Rec is recall, the frequency with which true relationships are claimed; Rep is number of repetitions with independently generated directed acyclic graphs and independently, randomly selected values of their linear coefficients and disturbance variances; Elapsed is wall time for a single run on the Pittsburgh Super Computer, or average of wall times for multiple runs with a quad core MacBook Pro. All numbers are averages over the runs.)

# Nodes	# Edges	# Rep	Adj Prec	Adj Rec	Arr Prec	Arr Rec	# Processors	Elapsed
1000	1000	100	98.92 %	94.77 %	98.92 %	90.05 %	2	1.2 s
1000	2000	100	98.43 %	88.04 %	96.27 %	85.74 %	4	8.5 s
30,000	30,000	10	99.77 %	94.60 %	99.04 %	89.97 %	120	53.5 s
30,000	60,000	10	99.81 %	86.72 %	99.23 %	84.47 %	120	3.4 m
1,000,000	1,000,000	1	93.90 %	94.83 %	83.11 %	90.57 %	120	11.0 h

Table 1: Average over random repetitions at sample size 1000 of accuracies and run times of the Fast Greedy Equivalence Search Algorithm (FGES) for Gaussian data using a BIC score with penalty 4.

According to Ioannidis' conclusion, Table 1, should be impossible. The source code and simulation facilities used are available at https://github.com/cmu-phil/tetrad.

Direct comparisons of FGES and PC-Max with samples of 1000 units and 20,000 variables and with graphs parameterized as above finds comparable precisions but better recall with FGES.

Alg	Ave Degree	AP	AR	AHP	AHR	E
PC-Max	2	1.00	0.94	0.99	0.87	225.91
FGS	2	1.00	0.98	1.00	0.98	175.24
PC-Max	4	1.00	0.79	0.99	0.74	335.25
FGS	4	1.00	0.93	1.00	0.93	264.50

Table 2: Comparison of constraint testing search (PC-Max) with scoring search (FGS) for time and recovery of 20,000 variable Gaussian Markov Equivalence Classes with alpha .00001 and BIC penalty 4 and sample sizes 1,000. Runtimes are for a 4 core MacBook Pro.

Kalisch and Buhlmann (2007) used improved tests in PC, the original asymptotically correct automated search algorithm for Markov equivalence classes, with sample sizes increasing with the number of variables, and graphs of degree 2 and degree 5; alpha for tests was 0.01 in all cases (they do not specify the range of linear coefficients or disturbance variances). With up to 20,000 variables their true positive rates approach 100% and their false positive rates are beneath 15% when the sample sizes approximate the number of variables.

Ioannidis and Colquhoun rightly note that effect size matters, but effect size and exploratory strategy can interact. Maatuis and colleagues (2010) have used a modification of PC, PC-Stable, together with the false discovery rate to compute a lower bound on the expected total effect of genes on phenotypes, and resampled to estimate the genes most likely to affect the phenotype. With more than 20,000 variables, in *S. cerevisiae* and *Arabidopsis thaliana* they find effects of genes that are known regulators and correctly predict new regulators, confirmed by knockout experiments. All of these examples assume Gaussian distributions and linear systems, but those assumptions are inessential. Both PC and FGS run with categorical variables, and conditional independence tests for non-Gaussian distributions have been developed (Ramsey, 2014l Gretton, et al. 2009; Zhang, et al. 2012). In fact, determination of directions of influence is more accurate and more informative when the variables have non-Gaussian distributions and non-Gaussian methods are used.

Any number of concerns arise: what about feedback relations, time series, unmeasured common causes? Building on work that has long been available for small problems, there are already methods in development for high dimensional stationary time series, feedback cycles and unmeasured confounding.

5. Conclusion

Automated search for causal relations has a long history of methods—factor analysis for example--that have no theoretical guarantees of accuracy and little confirmation of accuracy by simulation. That is no longer true. Contemporary misnamed "exploratory" methods for causal relations are better understood as consistent but complex statistical estimators. Ioannidis' argument against "exploratory" methods does not hold for them, and in fact the main assumption in his and Colquhoun's arguments, rarity of true relations among possible relations, is an advantage for the methods. With appropriate checks for distribution assumptions, use of background knowledge, and testing on simulated data from systems thought to approximate the distribution from target population, these methods have proved accurate and informative for high dimensional problems. Improvements on these and other search methods continue to appear regularly and rapidly.

If the social and behavioral sciences aim to begin to approximate the accuracies we expect from other sciences, it may well be that the fundamental designs of studies must change to include diverse populations, with measurements of every feasible feature, and analysis with accurate automated algorithms for estimating causal relationships.

References

ASA News (2016). http://www.amstat.org/asa/files/pdfs/P-ValueStatement.pdf

Colquhoun D. (2014) An investigation of the false discovery rate and the misinterpretation of p-values. Open Science. November 1;1(3):140216. DOI: 10.1098/rsos.140216.

Gretton A, (2009). Fukumizu K, Teo CH, Song L, B. Schölkopf B, Smola AJ. A kernel statistical test of independence. Adv Neural Inf Process Syst.;20: 585-592.

Ioannidis JP. (2005) Why most published research findings are false. PLoS Med. August 20;2(8): e124.

Kalisch M, Bühlmann P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. J Mach Learn Res. March; 8: 613-636.

Maathuis MH, Colombo D, Kalisch M, Bühlmann P. (2010). Predicting causal effects in large-scale systems from observational data. Nat Methods. ;7(4): 247-248.

Ramsey, J. (2016) Improving accuracy and scalability of the PC algorithm by maximizing p-value;. Preprint. Available: arXiv:1610.00378.

Ramsey J. (2014). A scalable conditional independence test for nonlinear, non-gaussian data;. Preprint. Available: arXiv:1401.5031. Accessed 08 September 2016.

Ramsey, J., M. Glymour, R. Sanchez-Romero, C. Glymour, (in press, 2016). A million variables and more: the Fast Greedy Equivalence Search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. International Journal of Data Science and Analytics.

Zhang K, Peters J, Janzing D, Schölkopf B. 2012. Kernel-based conditional independence test and application in causal discovery; 2012. Preprint. Available: arXiv:1202.3775.