Note

1. Stata code for this and other simulations discussed in this comment is available at http://stephenvaisey.com/documents/qcasimulations.do.

References

Eliason, Scott R. and Robin Stryker. 2009. "Goodness-of-fit Tests and Descriptive Measures in Fuzzy-set Analysis." *Sociological Methods and Research* 38(1):102–46.

Raftery, Adrian E. 1995. "Bayesian Model Selection in Social Research." Pp. 111–63 in *Sociological Methodology*, vol. 25, edited by Peter V. Marsden. Cambridge, MA: Blackwell.

Ragin, Charles C. 2008. *Redesigning Social Inquiry: Fuzzy Sets and Beyond*. Chicago: University of Chicago Press.

Vaisey, Stephen. 2007. "Structure, Culture, and Community: The Search for Belonging in 50 Urban Communes." *American Sociological Review* 72(6):851–73.

Vaisey, Stephen. 2009. "QCA 3.0: The 'Ragin Revolution' Continues." *Contemporary Sociology* 38(4):308–12.

Author Biography

Stephen Vaisey is an associate professor of sociology and a senior fellow at the Kenan Institute for Ethics at Duke University. The main goal of his research is to understand where people get their ideas about what a "good life" looks like and how these (usually implicit) ideas shape their strategies of action over time. He has also been writing lately about the promise and pitfalls of panel data for causal inference.

COMMENT: METHOD GAMES—A PROPOSAL FOR ASSESSING AND LEARNING ABOUT METHODS

Jake Bowers*

*University of Illinois at Urbana-Champaign, USA
Corresponding Author: Jake Bowers, jwbowers@illinois.edu
DOI: 10.1177/0081175014542078

Imagine assessing a promising method for pattern discovery using a game. One scholar would invent a true pattern of features, generate an outcome, and perhaps hide this pattern amid irrelevant information. For example, the game designer might provide 15 binary features of 40 cases to the players. Players would compete to discover the hidden truth. One version of the method game would require that participants use a particular algorithm. A second version would allow participants to choose their own algorithms. For example, some might choose a variant of qualitative comparative analysis (QCA; Rihoux and Ragin 2008), others would implement an adaptive lasso (Zou 2006), and still others might prefer one of the many competitors to the lasso, such as the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li 2001), random forests (Breiman 2001), or kernel-regularized least squares (Hainmueller and Hazlett 2012).[1]

In the first version of the competition, we would learn about craft: In different hands, the same method may perform differently. The results of this competition would teach us about the many kinds of substantive and methodological judgments required to use the method successfully. In the version of the game in which players choose different approaches, we could learn how different methods compare in their ability to address a given problem.[2]

If, however, time were short or players difficult to recruit, we could approximate such a game using a computer rather than a group of scholars: Lucas and Szatrowski (this volume, pp. 1–79) provide an example of how to evaluate one method in this way. That is, a single scholar could generate a true relationship as if kicking off a real method game but then write a computer program to compare the effectiveness of different algorithms in a simple situation. Imagine that scholarly literature focusing on 40 cases suggests that a complex dependent pattern of binary features $X_{i1}, \ldots, X_{iP}$, of a given case i, drives outcomes (say, for $P = 5$, $Y_i = \{(X_{i1} \cdot X_{i2} \cdot X_{i3})$ OR $(X_{i4} \cdot X_{i5})\}$ all $X_{ip} \in \{0,1\}$ and thus $Y_i \in \{0,1\}$). Furthermore, imagine that three methods suggest themselves as useful a priori: (1) QCA, (2) the adaptive lasso, and (3) iterative sure independence screening with a SCAD plugin (ISIS/SCAD; Fan and Lv 2008). Fan and Li (2001) proved that the SCAD penalty would correctly set false parameters to zero as $n \to \infty$ given a reasonable choice of tuning parameters in contrast to the simple lasso proposed by Tibshirani (1996); that is, SCAD has an oracle property but the simple lasso does not. Later, Zou (2006) showed that a modification of the lasso penalty (the adaptive lasso) does have an oracle property given well-chosen tuning parameters and weights. And Fan and Lv (2008) demonstrated that when the number of irrelevant features is much larger than the number of cases (e.g., when each case has 4,000 measured features but we observe only 40 cases), a preliminary dimension reduction step (ISIS) improves the performance of the SCAD penalty. Although QCA does not promise to find the truth as information increases, it appears, prima facie, well suited to discovering complex comparisons, and it does not require tuning parameters. This essay presents the results of one implementation of a machine version of the method game to compare QCA, the adaptive lasso, and ISIS/SCAD.[3]

Notice that this game is relatively easy. The only relevant case knowledge involves understanding that the causal features involve the *X*'s, and the outcomes are recorded in *Y*. Simplifying the case-knowledge requirements here allows an assessment of method rather than game player: All of the machine players have the same case knowledge and will use it in the same manner. Furthermore, machine players are naive. Assessing the performance of a machine will not tell us about the craft by which human scholars exploit a method. Furthermore, any single collection of case attributes can idiosyncratically advantage one method over another. For fairness, and to approximate the kind of natural variation we would see if different human players were involved in the game, the script generated a different set of features for each machine player, although the outcomes arose from the same deterministic true function as described above (i.e., the case knowledge is held constant across players and scenarios).

This competition involved 800 players, each using all three approaches to seek the truth. The script runs two contests. The easier of the two games presents players with a five-column data set: Each column represents a part of the truth, and players focus on finding the true combinations of the existing features. The hard game differs from the easy game only in that the data set contains 10 irrelevant case features in addition to the original 5. The script counts a player as successful if the player found the truth and only the truth. In the easy game, QCA, the adaptive lasso, and ISIS/SCAD found the truth and only the truth for 18%, 82%, and 96% of the players, respectively. In the hard game, QCA, the adaptive lasso, and ISIS/SCAD found the truth and only the truth for 0%, 33%, and 61% of the players, respectively.

One should not interpret these results as severe criticism of the adaptive lasso or QCA. Remember that in this essay, I propose a way to evaluate a method that is agnostic to the methods themselves. It proposes a machine-based approach only as a low-budget way to approximate the real competition among human players. However, in this essay I *do* claim that all methods must be able to be evaluated: As a minimal standard, given a true set of relationships, a good method should find the truth.

In the machine learning literature, such critical evaluation drives innovation. For example, a large and growing literature both criticizes and builds on the adaptive lasso: If the features are highly interdependent, we might prefer adaptive versions of the fused lasso (Rinaldo 2009), the grouped lasso (Wang and Leng 2008), or the elastic net (Ghosh 2011; Zou and Hastie 2004). Future methodology building on simple QCA might adapt insights from machine learning to overcome current shortcomings or advise against the use of QCA for particular designs and data. Most scholars would prefer a technique that recovers the truth more than 60% of the time, so the community of scholars might use these results to motivate work to improve the performance of ISIS/SCAD or to find a substitute. Obviously, a real competition with skilled human players with excellent judgment might have produced different results. After all, those who investigate and modify the script will notice little expertise and craft in the use of the techniques. For example, the tuning parameters for the adaptive lasso were chosen fairly naively, with only one open-source implementation of QCA being used and many other small but potentially consequential decisions appearing throughout. This essay also highlights the importance of the real, human-based, game: The simple machine game involved very simple case knowledge, and perhaps a scholar might have been able to identify the noise variables in some preliminary steps in the process if these data arose from real-world observations. A human game would also enable learning about methods for using case knowledge to guide and constrain pattern recognition and machine learning algorithms.

Fruitful communication about methods involves comparing the successes of different methods in the hands of different human scholars confronting specific research designs, theoretical goals, and existing observations. The method game proposed here would help us learn not only about a method in an abstract sense but about the craft of using said method in comparison with other methods. A machine version of the method game enables a fast, cheap, and controlled way to begin to build a comparative understanding of the methods and/or to motivate people to engage in a real method game.

Acknowledgments

David Collier, Stephanie Magean, Joanne Rodrigues, and Tim Liao helped improve this comment.

Notes

1. Hastie, Tibshirani, and Friedman (2009) and James et al. (2013) provide an excellent overview of many of the techniques known as "machine learning" or "statistical learning." The proposal for method assessment in this essay uses two simple techniques (adaptive lasso and SCAD) for illustration that are similar to common tools used by social scientists, but the idea of the method game is not limited to supervised linear model–based algorithms. The particular procedures evaluated here work by fitting penalized linear models to outcomes aiming to return coefficients of zero for irrelevant features, thereby revealing relevant features. These algorithms choose coefficients $\beta_1, \ldots, \beta_P$ for P features (and arbitrary combinations thereof), $X_{i1}, \ldots, X_{iP}$, related to some outcome, y_i, to minimize a function of the sum of squared prediction error (i.e., least squares) plus a penalty function that rewards solutions with some collection of β_p set to zero. The objective function tends to look like $\sum_{i=1}^{N} [y_i - (\beta_0 + \beta_1 X_{i1} + \cdots + \beta_P X_{iP})]^2 + \sum_{p=1}^{P} p(\lambda, \beta_p)$, where the tuning parameter, λ, determines the relative importance of the penalty function in the objective function compared with the least squares function. The adaptive lasso penalty is $p(\lambda, \beta_p, w_p) = \lambda w_p |\beta_p|$, where $w_p = 1/\hat{\beta}_p$ and $\hat{\beta}_p$ arises from a previous linear model (here a ridge regression but often an ordinary least squares regression). The SCAD replaces the lasso penalty with a function designed to have no penalty when $\beta_p = 0$ (like the adaptive lasso) but then to rise smoothly to penalize large β_p at a decreasing rate such that
$$p(\lambda, \beta_p, a) = \begin{cases} \lambda|\beta_p|, & \text{if} |\beta_p| \leq \lambda \\ -\left(\frac{|\beta_p|^2 - 2a\lambda|\beta_p| + \lambda^2}{2(a-1)}\right), & \text{if} \lambda < |\beta_p| \leq a\lambda \\ \frac{(a+1)\lambda^2}{2}, & \text{if} |\beta_p| > a\lambda \end{cases},$$
where a $>$ 2 and $\lambda > 0$. Some see adaptive lasso as an approximation of or competitor to the SCAD penalty (Hastie et al. 2009:92).
2. Although we might also confuse learning about method with a discovery that some researchers have excellent methodological judgment and luck.
3. Interested readers can download the open-source code from https://github.com/jwbowers/MethodGames.

References

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45(1):5–32.

Fan, Jianqing and Runze Li. 2001. "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties." *Journal of the American Statistical Association* 96(456):1348–60.

Fan, Jianqing and Jinchi Lv. 2008. “Sure Independence Screening for Ultrahigh Dimensional Feature Space.” *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 70(5):849–911.

Ghosh, Samiran. 2011. “On the Grouped Selection and Model Complexity of the Adaptive Elastic Net.” *Statistics and Computing* 21(3):451–62.

Hainmueller, Jens and Chad Hazlett. 2012. “Kernel Regularized Least Squares: Moving beyond Linearity and Additivity without Sacrificing Interpretability.” Retrieved June 20, 2014 (http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2046206).

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. 2nd ed. New York: Springer.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning*. New York: Springer.

Rihoux, Benoit and Charles C. Ragin. 2008. *Configurational Comparative Methods: Qualitative Comparative Analysis (QCA) and Related Techniques*. Thousand Oaks, CA: Sage.

Rinaldo, Alessandro. 2009. “Properties and Refinements of the Fused Lasso.” *Annals of Statistics* 37(5B):2922–52.

Tibshirani, Robert. 1996. “Regression Shrinkage and Selection via the Lasso: A Retrospective.” *Journal of the Royal Statistical Society, Series B (Methodological)* 73(3):267-88.

Wang, Hansheng and Chenlei Leng. 2008. “A Note on Adaptive Group Lasso.” *Computational Statistics and Data Analysis* 52(12):5277–86.

Zou, Hui. 2006. “The Adaptive Lasso and Its Oracle Properties.” *Journal of the American Statistical Association* 101(476):1418–29.

Zou, Hui and Trevor Hastie. 2003. “Regression Shrinkage and Selection via the Elastic Net, with Applications to Microarrays.” *Journal of the Royal Statistical Society, Series B* 67:301–20.

Author Biography

Jake Bowers (http://jakebowers.org) is an associate professor of political science and statistics and a research scientist at the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign. He has published work on the design and analysis of multilevel and cluster randomized studies in the *Journal of the American Statistical Association, Political Analysis, Statistical Science*, and *Electoral Studies*.