Increasing generalizability via the principle of minimum description length

Wes Bonifay

University of Missouri

Missouri Prevention Science Institute

Note: This commentary has been accepted for publication in *Behavioral and Brain Sciences* in response to Tal Yarkoni's (2021) target article on "The Generalizability Crisis" (https://psyarxiv.com/jqw35). This version has not yet been copy-edited by the journal.

Abstract

Traditional statistical model evaluation typically relies on goodness-of-fit testing and quantifying model complexity by counting parameters. Both of these practices may result in overfitting and have thereby contributed to the generalizability crisis. The information-theoretic principle of minimum description length addresses both of these concerns by filtering noise from the observed data and consequently increasing generalizability to unseen data.

Increasing generalizability via the principle of minimum description length

As a remedy to the generalizability crisis, Yarkoni urges researchers to consider "cross-validation techniques that can minimize overfitting and provide alternative ways of assessing generalizability outside of the traditional inferential statistical framework" (Sec. 3.6.7). I believe this advice is valuable and worthy of elaboration.

Traditional model evaluation techniques are beset by (at least) two inconvenient truths. First, goodness-of-fit (GOF) and generalizability are inextricably tied to model complexity (defined by Myung, Pitt, and Kim (2005) as "a model's inherent flexibility that enables it to fit a wide range of data patterns" [p. 12]). As models become more complex, GOF to the observed data increases, but generalizability to unseen data decreases. Additionally, GOF indices conflate fit to the useful signal in the data with fit to the useless noise, and so must be adjusted to account for complexity. The widely used Akaike Information Criterion (Akaike, 1973), for example, mitigates the effects of complexity by penalizing for the number of parameters.

However, this leads to the second issue: Complexity cannot be fully assessed by simply counting parameters (and in fact, overfitting can occur with just one parameter; Piantadosi, 2018). Complexity is also affected by the configuration of variables in the model (Cutting et al., 1992): Models that organize the same number of parameters in different configurations may differ in terms of GOF. It follows from these two issues that researchers who rely exclusively on GOF and quantify complexity only by counting parameters are exacerbating the generalizability crisis.

A solution to these problems can be found by bypassing probability theory altogether and adopting a technique from information theory. The principle of *minimum description length* (MDL; Rissanen, 1978, 1989) aims to separate regularity (i.e., meaningful information) from

noise in the observed data and "squeeze out as much regularity as possible" (Grünwald, 2004, p. 15) via data compression. Suppose we have a sequence of 9 binary digits that contains a regularity: twice as many 1s as 0s. The complete data space *S* includes $2^9 = 512$ patterns, but the regularity only applies to 84 (or 16.4%) of those patterns. Thus, our sequence belongs to a relatively small subset of *S*. A description (e.g., programming code) that compresses the complete data in this manner would be quite useful: We would know, for example, that future use of that code would return only those sequences that contain the same regularity.

According to the MDL principle, the best description (or model) is that which maximizes compression of *S*. Our 9-digit sequence could be further compressed: The regularity of "twice as many 1s as 0s + the first three digits are 1s" describes just 20 patterns, compressing the data to less than 4% of *S*. That is, over 96% of sequences would not follow this more precise regularity, so we should be "impressed" (in the sense of Meehl's [1990] rainfall analogy or Lakatos's [1978] example of Halley's comet) when we find a sequence that does.

What does this have to do with the generalizability crisis? In his introduction to MDL, Grünwald (2004) described two relevant features. First, "MDL procedures automatically and inherently protect against overfitting" (p. 5). GOF statistics may overfit the data by capturing both signal and noise, whereas MDL methods filter out that noise through data compression, allowing researchers to focus only on the signal. Second, "MDL methods can be interpreted as searching for a model with good predictive performance on unseen data." (p. 6). Mathematical proof of this statement can be found in Vitányi and Li (2000), who concluded that "compression of descriptions almost always gives optimal prediction" (p. 448).

Although MDL may seem obscure, consider it in light of this statement from Roberts and Pashler (2000) in *Psychological Review*, following their declaration that good fit cannot clarify

what a theory predicts: “Without knowing how much a theory constrains possible outcomes, you cannot know how impressed to be when observation and theory are consistent” (p. 359). The phrase “a theory [that] constrains possible outcomes” can be rewritten in MDL terms as “a description that compresses the complete data space.” Through that translation, it becomes clear that the MDL principle encapsulates Meehl’s (1997) argument that “the narrower the tolerated range of observable values, the riskier the test, and if the test is passed, the stronger the corroboration of the substantive theory” (p. 407).

Various methods have been developed to quantify the MDL principle (see Myung et al., 2006; Navarro, 2004; Pitt et al., 2002), but their formulations involve statistical obstacles such as integration across the complete data space. To sidestep this intractability, quantitative psychologists have relied on simulation methods to gain MDL-type insights regarding latent variable models. Preacher (2006), generated 10,000 random correlation matrices to simulate the complete continuous data space and fit competing structural equation models with the same number of parameters but different configurations to each matrix (interested readers can conduct similar MDL-type studies using the ockhamSEM package in R; Falk & Muthukrishna, 2020). Despite the fact that the number of parameters was held constant, certain models had an inherent tendency to fit better than others (termed “fitting propensity”).

Bonifay and Cai (2017) expanded upon this work by considering the fitting propensity of several categorical data models. Among other findings, their analysis revealed that the confirmatory bifactor model achieved good fit to an excessively wide range of random datasets. The model was so deficient at compressing the data space (i.e., filtering out noise) that it accommodated an extremely wide range of data patterns, including many that were nonsensical. This MDL-inspired work demonstrated that good fit is essentially built into the bifactor model,

so if the goal is to ensure generalizability, GOF testing should not be considered risky or severe (Watts et al., 2019).

In summary, the information-theoretic principle of MDL offers insights into overfitting and generalizability that are not possible using traditional methods. Although this principle may not address many of the generalizability issues described in the target article, it should be considered by researchers who wish to avoid overfitting and thereby enhance predictive accuracy.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov, & F. Csaki (Eds.), *Second International Symposium on Information Theory* (pp. 267–281). Budapest: Akademiai Kiado.

Bonifay, W., & Cai, L. (2017). On the complexity of item response theory models. *Multivariate Behavioral Research*, *52*(4), 465-484.

Cutting, J. E., Bruno, N., Brady, N. P., & Moore, C. (1992). Selectivity, scope, and simplicity of models: A lesson from fitting judgments of perceived depth. *Journal of Experimental Psychology: General, 121*(3), 364-381.

Falk, C., & Muthukrishna, M. (2020). Parsimony in Model Selection: Tools for Assessing Fit Propensity. *arXiv preprint arXiv:2007.03699*.

Grunwald, P. (2004). A tutorial introduction to the minimum description length principle. *arXiv preprint math/0406077*.

Lakatos, I. (1978). Introduction: science and pseudoscience. In J. Worrall & G. Currie (Eds.), *The Methodology of Scientific Research Programs* (pp. 1-8). Cambridge University Press, Cambridge, UK.

Meehl, P. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if There Were No Significance Tests?* (pp. 393–425). Erlbaum.

Meehl, P. E. (1990a, April). Appraising and amending theories: The strategy of lakatosian defense and two principles that warrant it. *Psychological Inquiry*, *1*(2), 108–141.

Myung, J. I., Navarro, D. J., & Pitt, M. A. (2006). Model selection by normalized maximum likelihood. *Journal of Mathematical Psychology*, *50*(2), 167-179.

Myung, I. J., Pitt, M. A., & Kim, W. 2004. Model evaluation, testing and selection. In K. Lambert & R. Goldstone (Eds.), *The Handbook of Cognition* (pp. 422-436). Thousand Oaks, CA: Sage.

Navarro, D. J. (2004). A note on the applied use of MDL approximations. *Neural Computation*, *16*(9), 1763-1768.

Navarro, D. J. (2019). Between the devil and the deep blue sea: Tensions between scientific judgement and statistical model selection. *Computational Brain & Behavior*, *2*(1), 28-34.

Piantadosi, S. T. (2018). One parameter is always enough. *AIP Advances*, *8*(9), 095118.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, *109*(3), 472-491.

Preacher, K. J. (2006). Quantifying parsimony in structural equation modeling. *Multivariate Behavioral Research*, *41*(3), 227-259.

Rissanen, J. (1978). Modeling by the shortest data description. *Automatica, 14,* 465–471.

Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry.* World Scientific Publishing: Singapore.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*(2), 358-367.

Vitányi, P. M., & Li, M. (2000). Minimum description length induction, Bayesianism, and Kolmogorov complexity. *IEEE Transactions on Information Theory*, *46*(2), 446-464.

Watts, A. L., Poore, H. E., & Waldman, I. D. (2019). Riskier tests of the validity of the bifactor model of psychopathology. *Clinical Psychological Science*, *7*(6), 1285-1303.