IES Institute of Education Sciences

Uncovering the Hidden Complexity in Statistical Models

Wes Bonifay, University of Missouri

Annual IES Principal Investigators Meeting
January 25-27, 2022
Advancing Equity and Inclusion in the Education Sciences

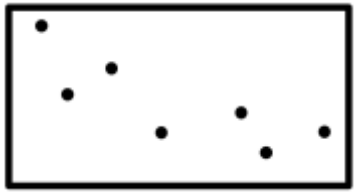
Complexity is “a model’s inherent flexibility that enables it to fit a wide range of data patterns”

Myung, Pitt, & Kim (2005), p. 12

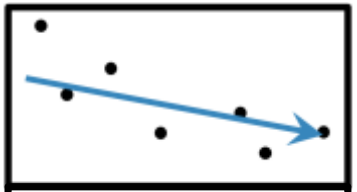
IES Institute of Education Sciences

Sources of Complexity

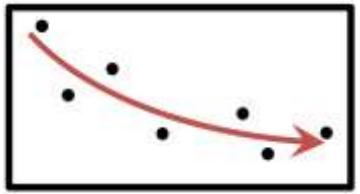
Observed Data

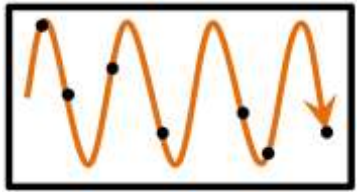


Model Complexity



$y = b_0 + b_1 x$





Complexity Evaluation:

More parameters:
(i.e., *parametric* complexity)

$y = b_0 + b_1 x + b_2 x^2$

$y = b_0 + b_1 x + \cdots + b_6 x^6$

e.g., AIC, BIC

Same # of parameters:
(i.e., *configural* complexity)

$y = \frac{b_0}{e^{b_1 x}}$

$y = b_0 \cdot \sin(b_1 x)$

e.g., Minimum Description Length

Model Evaluation Methods

Traditional

Goodness-of-fit testing, bootstrapping procedures, and other traditional evaluation methods do **not** consider the configural complexity of a model

Bayesian

Predictive model checking techniques and other Bayesian model evaluation methods offer additional insights, but they also fail to address configural complexity

Information-theoretic

The principle of minimum description length and the related notion of fitting propensity can be used to uncover and quantify the configural complexity

Minimum Description Length

Key insight:

Goodness-of-fit = fit to regularity + fit to noise

"Two-part" MDL (Rissanen, 1978) directly addresses the tradeoff between goodness-of-fit and complexity

- MDL focuses on encoding models as a sequence of *bits* (0s and 1s)
- The best model M to explain the data D is the one that minimizes the *total* description length: $\arg\min[L(\mathrm{M}) + L(\mathrm{D}|\mathrm{M})]$

  $L(\mathrm{M})$ = # of bits to describe the model, i.e., the regularity

  $L(\mathrm{D}|\mathrm{M})$ = # of extra bits to describe the data **after** it has been encoded by the model, i.e., the noise
- Vitányi & Li (2000) demonstrate that "compression of descriptions almost always gives optimal prediction" (p. 448), i.e., generalizability

Minimum Description Length (cont.)

MDL is a *principle* of data compression, and several formulations of this principle have been developed

e.g., Normalized maximum likelihood (Rissanen, 2001): $\mathrm{NML} = \frac{L(\mathrm{D}|\mathrm{M})}{\int_S L\left(\mathbb{D}\middle|\mathrm{M}(\mathbb{D})\right) d\mathbb{D}}$

$L(\mathrm{D}|\mathrm{M})$: the likelihood of observed data D given model M

$L(\mathbb{D}|\mathrm{M}(\mathbb{D}))$: the likelihood when the model is fit to all possible data $\mathbb{D}$ in the data space S

Thus, NML quantifies the information in a model, adjusted for the model's propensity to fit well to *all possible data*

- Unfortunately, the denominator involves integration across the complete data space, so it is usually intractable

MDL in Latent Variable Modeling

Although we cannot work with the NML directly, we can use the same *reasoning* to evaluate models that are commonly applied in quantitative psychology

- By generating the complete data space, we can investigate the inherent complexity of a model—its inbuilt tendency to fit well to any possible data

  This is not a new idea: Cutting et al. (1992) demonstrated that a baseline for model fit can be established by fitting models to random data

- Preacher (2006) used the term ***fitting propensity (FP)*** to denote a model's ability to fit diverse patterns of data, all else being equal

  He then examined the FPs of competing structural equation models that had the same number of parameters, but different structural configurations (i.e., functional forms)

Preacher (2006)

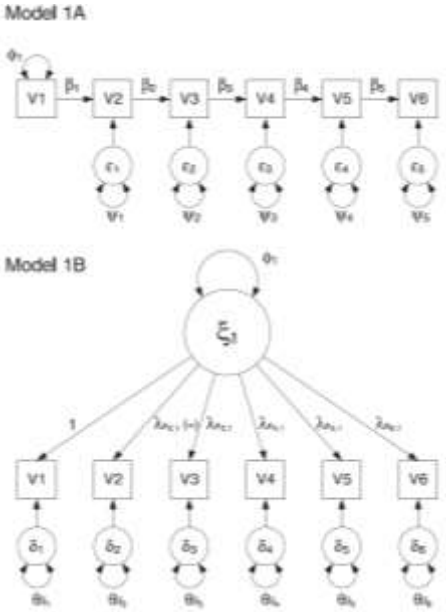


Figure 4. A simplex model (1A) and a factor model (1B), each with 11 freely estimated parameters

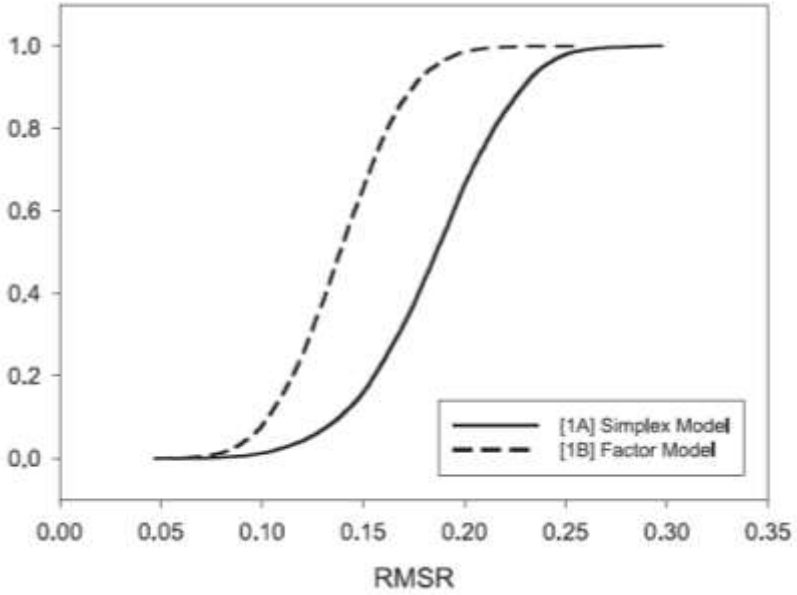


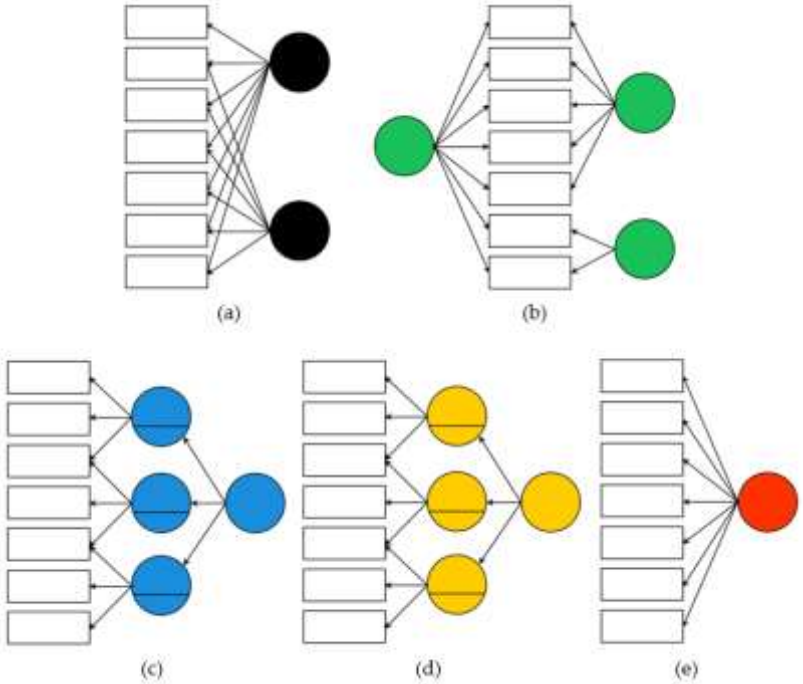
Figure 5. Cumulative frequency distributions of unadjusted goodness-of-fit (RMSR) for each model fit to the same 10,000 random correlation matrices

"The good fit of a hypothesized model to observed data, although desirable, can result from the model's inherent ability to predict data patterns and may have little to do with its value as a scientific tool. **Cherished models may have to be abandoned or replaced** if their past successes can be ascribed more to FP than to any insight they lend into the process that actually generated the data."

Preacher (2006), p. 254

MDL in item response theory modeling

Bonifay & Cai (2017) extended Preacher (2006) to the categorical case (item factor analysis)



5 measurement models:

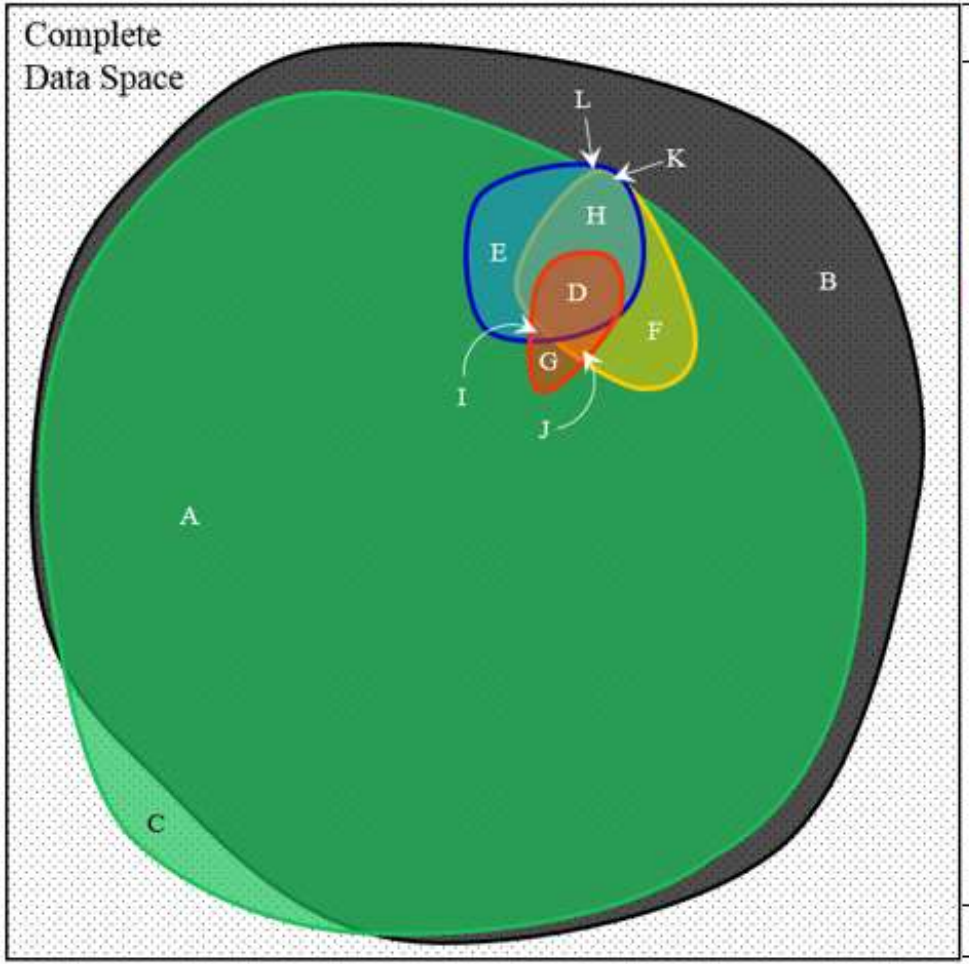
(a) EIFA
(b) Bifactor
(c) DINA
(d) DINO

20 free parameters

(e) Unidimensional (3PL)

21 free parameters

All models were fit to 1,000 data sets & evaluated via the Y2/N unadjusted fit index



Region	Y2/N ≤ .05
EIFA	79.2
Bifactor	63.5
DINA	5.1
DINO	5.2
Uni	2.3
A	51.8
B	19.2
C	3.8
D	1.1
E	1.9
F	2.0
G	0.7
H	1.7
I	0.2
J	0.3
K	0.1
L	0.1
Unoccupied	17.1

Bonifay & Cai (2017)

Results reiterate the importance of configural complexity:

- The configuration of the variables in a model affects its propensity to fit well
- The EIFA & bifactor models were the most configurally complex, as evidenced by their high propensity to fit *any possible* data
- DINA & DINO diagnostic classification models had low FP

  There is a distinct theoretical difference between these models, so they occupied different regions of the data space
- The unidimensional model had ***an additional free parameter***, but much lower FP!
- Strong implications re: model evaluation via goodness-of-fit

  For models with high FP, overfitting is a statistical inevitability

  For models with low FP, good fit is unlikely, so far more meaningful when obtained

Institute of Education Sciences

IES Award R305D210032

Innovative, Translational, and User-Friendly Tools for Comprehensive Statistical Model Evaluation

- Co-PI: Li Cai (UCLA)

Goal: To develop a statistical model evaluation framework that integrates traditional, Bayesian, and information-theoretic perspectives on model evaluation

- With regard to configural complexity, a major contribution of this project will be the development of a new statistical method for generating the synthetic datasets that are needed for thoroughly investigating model complexity

  The data generation method of Bonifay and Cai (2017) severely restricted the number and type of items that could be considered; the proposed work will introduce a statistical innovation for evaluations of complexity in models of more items and/or polytomous items

- In addition, this novel data generation method will entail special consideration of estimation via composite likelihood methods

IES Award R305D210032 (cont.)

In addition, this project will culminate in the ***CoSME*** (Comprehensive Statistical Model Evaluation) R package and Shiny app that will allow users to:

1. Specify a model and import data
2. Select their preferred model evaluation method(s):
    - Traditional (e.g., goodness-of-fit, parametric bootstrapping)
    - Bayesian (e.g., prior and/or posterior predictive model checking)
    - Information-theoretic (e.g., fitting propensity analysis)
3. Export a dynamic report

Further Resources

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. (in press). Parsimony in model selection: Tools for assessing fit propensity. *Psychological Methods.*

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

Myung, I. J., Pitt, M. A., & Kim, W. (2005). Model evaluation, testing and selection. In Lamberts, K. & Goldstone, R., (Eds.), *Handbook of Cognition*. London, UK: Sage Publications Ltd.

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

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

Preacher, K. J., Cai, L., & MacCallum, R. C. (2007). Alternatives to traditional model comparison strategies for covariance structure models. In T. D. Little, J. A. Bovaird, & N. A. Card (Eds.), *Modeling contextual effects in longitudinal studies* (pp. 33–62). Mahwah, NJ: Erlbaum.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica, 14*(5), 465-471.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psych 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.

THANKS!

Any questions?

Email: bonifayw@missouri.edu

Twitter: @wesbonifay