SAGE Reference

The SAGE Encyclopedia of Research Design

Overfitting

By: Jay Hegdé
Edited by: Bruce B. Frey
Book Title: The SAGE Encyclopedia of Research Design
Chapter Title: "Overfitting"
Pub. Date: 2022
Access Date: July 3, 2022
Publishing Company: SAGE Publications, Inc.
City: Thousand Oaks,
Print ISBN: 9781071812129
Online ISBN: 9781071812082
DOI: https://dx.doi.org/10.4135/9781071812082.n423
Print pages: 1140-1142

Overfitting is a problem encountered in statistical modeling of data, where a model fits the data well because it has too many explanatory variables. Overfitting is undesirable because it produces arbitrary and spurious fits, and, even more importantly, because overfitted models do not generalize well to new data.

Overfitting is also commonly encountered in the field of machine learning, including learning by neural networks. In this context, a learner, such as a classifier or an estimator, is trained on an initial set of samples and later tested on a set of new samples. The learner is said to be overfitted if it is overly customized to the training samples and its performance varies substantially from one testing sample to the next.

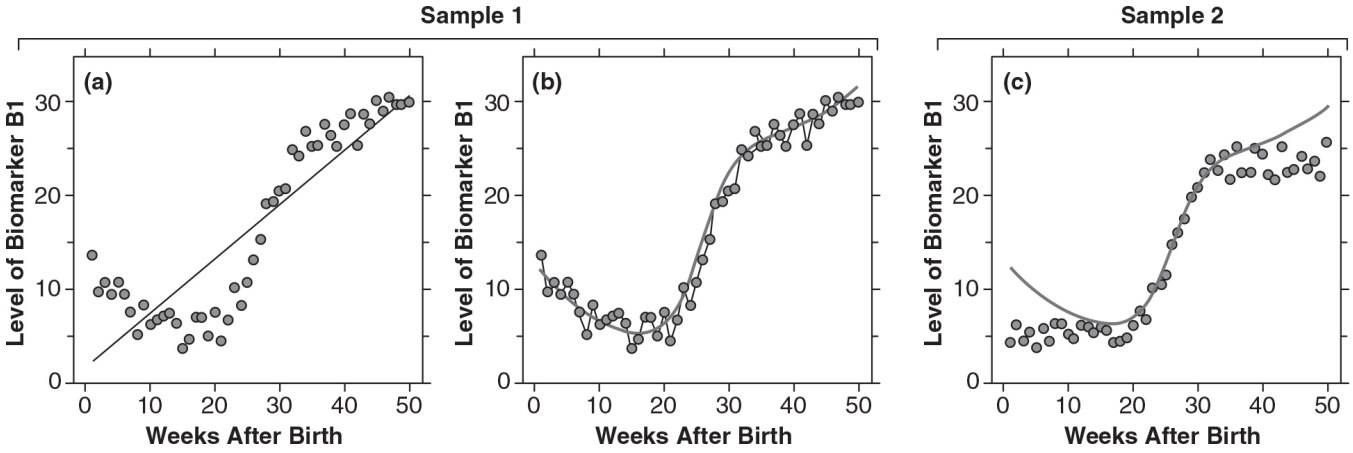
Because the nature of the overfitting problem and the methodologies of addressing it are fundamentally similar in the two fields, this entry examines overfitting mostly from the viewpoint of statistical modeling.

Relationship Between Model Fit and the Number of Explanatory Variables

In a statistical model, the fit of the model to the data refers to the extent to which the observed values of the response variable approximate the corresponding values estimated by the model. The fit is often measured using the coefficient of determination R2, the value of which ranges between 0 (no fit) and 1 (perfect fit).

As one adds new explanatory variables (or parameters) to a given model, the fit of the model typically increases (or occasionally stays the same, but will never decrease). That is, the increase in fit does not depend on whether a given explanatory variable contributes significantly to the overall fit of the model or adds to its predictive power. Therefore, all other things being equal, a larger number of explanatory variables typically means better fit, but it does not necessarily mean a better model.

Figure 1 Overfitting and Underfitting



Notes: These hypothetical clinical data show the blood levels of a biomarker B1 in a group of newborn infants measured during the first 50 weeks of their lives. Panels A and B are from the first group of babies (Sample 1), and the data in Panel C are from a different, independent sample. The dots represent the observed values of the biomarker (in arbitrary units). The thin black lines represent the underfitted and overfitted models (Panels A and B), respectively. The thick gray line in Panel B denotes the parsimonious model that best fits Sample 1. The same model is also shown in Panel C for comparison. Note that even though the model fits the data in Sample 1 rather well, it fails to predict the new data in Sample 2.

How many variables are too many depends on the model and the data. There is no preset number of variables above which the model is considered overfitted. Rather, overfitting is a graded property of a model: A given model is more or less overfitted depending on the number of parameters it has relative to the number of parameters actually needed.

Why Overfitting Is Undesirable

The function of a statistical model is twofold: First, it should help account for, or explain, the observed values in the original sample(s) on which it is based. Second, and more importantly, it should reliably predict the observed values in new testing samples. Overfitting compromises both the explanatory and predictive abilities of a model.

To understand how, suppose one is interested in modeling the levels of a hypothetical biomarker in healthy babies as a function of the baby's age (Figure 1). Age-appropriate levels of the biomarker signify normal development. The goal of the modeling endeavor is to determine which factors affect the levels of this biomarker.

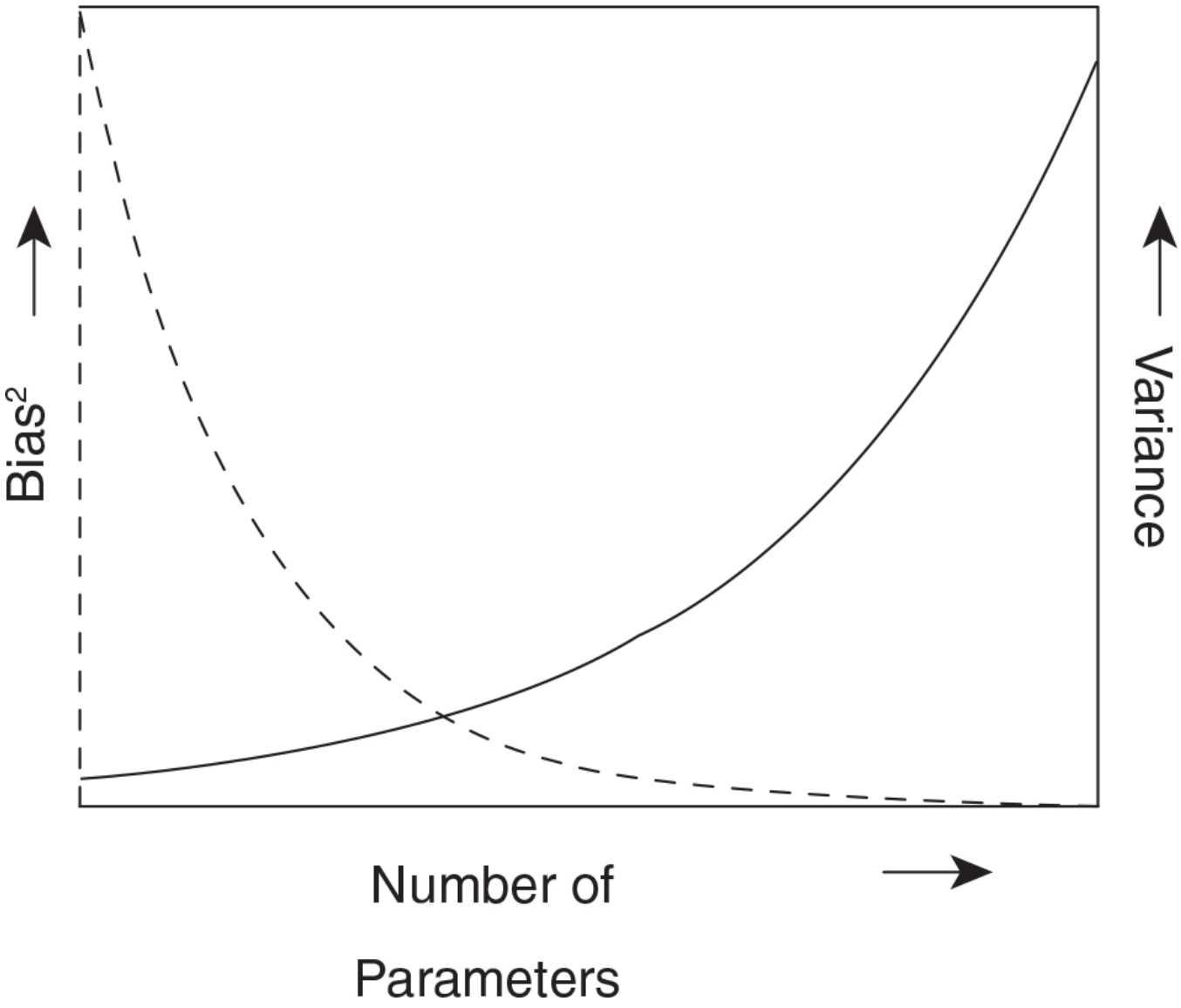
Figures 1a and 1b show the biomarker levels measured in one group of babies during the first 50 weeks of their lives. When baby's age is used as the sole explanatory variable—that is, when the model underfits the data—the model is consistently "off the mark"; it generally misestimates the response variable (Figure 1a). In this case, the model is said to have high bias.

Figure 1b shows the same data overfitted by blindly using every last bit of information about the babies as explanatory variables—including age, gender, ethnicity, diet, and hours of sleep per day—up to and including clearly absurd variables such as the color of the diaper, number of traffic deaths during the given week, and the phase of the moon. This model, although manifestly absurd, fits the observed data from this sample nearly perfectly (Figure 1b). Yet the model explains little of the observed biomarker levels. That is, a good fit does not necessarily make a good model.

The model is whittled down to retain only those factors that can account for a statistically significant portion of the observed data. This more parsimonious model provides a reasonable approximation to the data (Figure 1b, thick gray line). Therefore, this model fulfils the first of its aforementioned functions: It helps explain the data at hand.

However, this model still does not provide a good fit to a new sample (Figure 1c). This is the second, larger problem with the overfitted models: Because they are overly customized to the original samples, they are not sensitive to data patterns that occur only in some of the samples if those patterns happen not to be present in the original samples. Therefore, they fail to generalize well across all samples, so that their degree of fit tends to vary from one sample to the next. Such models are said to have high variance. That is, the second model, although more parsimonious, is still overfitted.

Figure 2 Model Selection as a Trade-Off Between Bias and Variability



Note: As the number of model parameters increases, the bias of the model decreases (dashed line), and the variance of the model increases (solid line). The point at which the two lines intersect represents the most optimal theoretical balance between underfitting and overfitting (i.e., optimal bias-variance trade-off) when all other things are equal.

Preventing Overfitting: The Principle of Parsimony

Clearly, a good model avoids both of these extremes of underfitting and overfitting. The general approach to building models that avoid these extremes is enunciated by the principle of parsimony, which states that a model should have the fewest possible number of parameters for adequate representation of the data. Achieving parsimony by "shaving away all that is unnecessary" is often referred to as Occam's razor, after the 14th-century English logician William of Occam.

Bias-Variance Trade-Off

A principled, and widely used, approach to determining the optimal number of parameters in a model is to

find the optimal trade-off point between bias and variance (Figure 2). As noted above, if the model has too few parameters (i.e., if it is underfitted), it tends to have a large bias. If it has too many parameters (i.e., if it is overfitted), it tends to have a large variance. Thus, the number of parameters that best balances bias with variance achieves the best balance between underfitting and overfitting.

It is important to note that the bias-variance trade-off represents only a conceptual guideline for preventing overfitting and not an actual methodology for doing it. This is because, for one thing, it is impossible to precisely determine the point of optimal trade-off with a finite number of samples and without knowing the "ground truth," that is, what the "correct model" actually is. For another, the optimal balance between bias and variance often depends on many factors, including the goal of modeling. For instance, when simplicity of the model is paramount, a researcher may select the smallest number of parameters that still provide a significant fit to the data. On the other hand, when bias is more "costly" than variance, one may retain more parameters in the model than strictly necessary. Thus, a model that is not the most parsimonious is not necessarily a bad model.

Model Selection Methods

Just as there is no single, universally applicable definition of what a good model is, there is no single, universally acceptable method of selecting model parameters. Rather, there is a variety of available methods. The methods vary widely in how they operate and how effective they are in preventing overfitting. It is beyond the purview of this entry to survey them all. Instead, some main approaches are briefly outlined.

As noted above, one can start with an overfitted model and test each explanatory variable for whether it contributes significantly to the model fit at a given alpha level (e.g., 0.05) and retain it in the model only if it does. Alternatively, one can start with an underfitted model and incorporate only those variables that significantly improve the overall fit of the model. These step-up (or forward) versus step-down (or backward) parameter selection methods generally (although not always) produce largely similar models. Such significance-testing (or hypothesis-testing) methods, although widely used in the biological and social sciences, are considered inadequate by themselves for model selection. This is ultimately because these methods only address the ability of a given variable to explain the original data and do not test the ability of the model to predict new ones. Cross-validation addresses some of these shortcomings, where the selected model is tested against new samples.

Akaike's Information Criterion (AIC) is an information theory–based index that measures the goodness of fit of a given model. It, along with its many variations, is a measure of the trade-off between bias and variance. For this and many other reasons, this measure has been a highly regarded method of preventing overfitting. Mallows' Cp statistic is another goodness-of-fit measure that generally, but not always, produces results similar to AIC.

Using Bayesian principles of inference is one of the most promising new approaches to the prevention of overfitting. In machine learning, Bayesian principles have been used in two main ways to prevent overfitting. First, prior marginal probability distribution of a given explanatory variable has been used to help infer the extent to which the variable "belongs" in the model. Second, model selection itself can be treated as a Bayesian decision problem, so as to select a model that is most likely to be the "correct" model. These techniques, however, are yet to be widely used for preventing overfitting in the biological and social sciences.

It should be clear from the foregoing discussion that preventing overfitting is as much an art as it is a science because it involves exercising subjective judgment based on objective statistical principles. The fact that the optimal model is often in the eye of the beholder means that model selection methods can be abused to select one's favorite model. But such a fishing expedition of methods to support one's favorite model is inappropriate. The researcher's question ought to be which model best represents the data at hand and not which method supports the model of choice.

See also Inference: Deductive and Inductive; Loglinear Models; Mixed Model Design; Models; Polynomials; SYSTAT

Jay Hegdé
http://dx.doi.org/10.4135/9781071812082.n423
10.4135/9781071812082.n423

Akaike, H. (1983). Information measures and model selection. Bulletin of the International Statistical Institute, 44, 277–291.
Bishop, C. M. (2007). Pattern recognition and machine learning. New York: Springer.
Burnham, K. P., & Anderson, D. R. (2003). Model selection and multi-model inference: A practical information-theoretic approach. New York, NY: Springer.
Draper, N. R., & Smith, H. (1998). Applied regression analysis. New York, NY: Wiley.
Duda, R. O., Hart, P. E., & Stork, D. G. (2000). Pattern classification. New York, NY: Wiley Interscience.
Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. Neural Computation, 4, 1–58.
Harrell, F. E., Jr. (2001). Regression modeling strategies. New York, NY: Springer.
Hawkins, D. M. (2004). The problem of overfitting. Journal of Chemical Information and Computer Sciences, 44, 1–12.