Quantifying terms from the bias-variance decomposition using metalog distributions

Symposium on Data Science and Statistics

Saint Louis, MO, June 2-5, 2021

Neil A. Hamlett, D.Sc., MBA

neil.hamlett@uncertainty-research.science https://www.linkedin.com/in/neil-hamlett-strategic-quant/

https://hamlett-neil-ur.github.io/ https://orcid.org/0000-0001-8278-0087

Abstract

The bias-variance decomposition (BVD) serves as a key guidepost in machine-learning practice. Comparing model scores for training, validation, and test data provide qualitative indications of possible model overfitting. Explicit calculation of BVD terms is generally problematic, absent assumptions about convenient probability distributions.

Metalog distributions, an emerging technique from the Decision-Analysis community, provide a mechanism for explicit calculation of BVD terms. Metalogs fit arbitrary empirical distributions to closed-form, multiply-differentiable, continuous functions represented as series.

The bias and variance terms result from metalog-distribution fits to prediction-model training- and test-data residual errors employing a method resembling Locally Weighted Scatterplot Smoothing (LOWESS). Instead of calculating means as is done with LOWESS, metalog-fit probability-density functions are instead obtained. This allows for some statistical inference using machine-learning results, including estimates of prediction and confidence intervals. Applying Kolmogorov-Smirnov or other tests to the resulting bias and variance distributions leads to a quantitative characterization of prediction-model overfitting.

Metalog distributions can similarly be employed to estimate Cramer-Row Lower Bounds (CRLB) on estimation-error variance. This occurs from fitting orthogonal-transform-space representations of the explanatory variables to metaled distributions. The CLRB offers a proxy for the irreducible-variance term in the BVD. Explicit calculation of BVD terms using metalog distributions affords the opportunity to extend Analysis of Variance (ANOVA) methods from classical statistics into the machine-learning realm.

Introduction.

The bias-variance decomposition is a key concept for machine learning and data science. Prominently described by [Hastie2009, p. 233], it is commonly used to evaluate whether a model is overfit. In standard practice, data sets are partitioned into training, validation, and test subsets. Statistical models are fit to the training set, often repeatedly through a cross-validation process. One or more statistical scores are calculated for each candidate model. The best model produces the most-favorable score.

The test data are held "in reserve". The model best fitting the training and validation data is finally applied to the test data and its score calculated. The score from the test data is compared to that from the training and validation subsets. If the latter score is significantly better than the former, the model is likely to be overfit. Under such circumstances, the model suboptimally explains data not used in its derivation.

This approach suffers from two limitations. First, it provides only a qualitative indication of possible overfitting, using model scores as a proxy for actual bias-variance analysis. In contrast, Analysis of Variance (ANOVA) in classical statistics produces extensive insight into the extent to which models' residuals contain unexplained variance (e.g., [Dielman2005, Olive2017,

Sahai2004]). Model-score comparisons alone yield substantially less insight.

Second, the practice described above only considers two of three terms in the bias-variance decomposition. It considers bias, the degree to which the model fails to the structure of the phenomena that produced the data. Also variance quantifies the degree to which the model fits the idiosyncratic randomness of the training data.

The bias-variance decomposition in [Hastie2009] however contains a third term, the "irreducible error". This is the extent to which inherent randomness in the data defies systematic explanation. Irreducible error resembles Information Theory's concept of entropy (e.g., [Cover2006, pp. 33-35]), the irreducible uncertainty in stochastic phenomena from which data result. Not considering irreducible error might lead to underappreciation of the variance not explained by a model.

Explicit calculation of terms of in the bias-variance decomposition is often inconvenient. The classical statistics community assumes convenient statistical distributions — often from a family based on exponential functions (e.g., [Agresti2013, chapter 4]). Machine learning relaxes such assumptions, which are often not valid in the real world anyway.

A recent innovation originating in the decision-analysis community creates the opportunity to calculate decomposition terms explicitly. Applied decision practitioner Tom Keelin developed a method for making calculations based on empirical distributions. Metalog distributions provide an approach to fitting arbitrary empirical distributions to continuous, multiply-differentiable, closed-form functions [Keelin2016]. Closed-form representations for terms in the bias-variance decomposition can thereby be obtained.

Metalog distributions' formulations are based on the quantile function $M_n(y)$ defined as

$$x = M_n(y) \Leftrightarrow y = Pr\{x \leq \mathscr{X}\} = P_{\mathscr{X}}(x).$$

The quantile function is estimated using a series

$$\hat{M}_n(y) = \sum_{\nu=1}^{\frac{n}{2}} \left(a_{2\nu-1} \left(y - \frac{1}{2} \right)^{2\nu-1} + a_{2\nu} \left(y - \frac{1}{2} \right)^{2\nu-1} \ln\left(\frac{y}{1-y} \right) \right).$$

The coefficients result from either an ordinary-least-squares (OLS) solution or another linear solver. Probability-density and -distribution functions follow from applying elementary differential and integral calculus to $\hat{M}_n(y)$.

Target audience.

Two distinct constituencies benefit from explicit calculation of bias-variance decomposition terms. Machine-learning and data-science practitioners comprise the first. Explicit calculation of probability distributions associated with each decomposition term opens up the opportunity to more-thoroughly characterize models' degree of fit and residual uncertainty. Expressing bias as a probability distribution, for example, permits articulation of confidence intervals from machine-learning models. This allows for deeper insight and clearer communication about residual uncertainty than model scores alone.

Statistical inference — not generally attempted in machine-learning contexts — produces uncertainty insights such as confidence intervals. These allow explicit visualization of a model's residual uncertainty. Model-fit scores by themselves are more-abstract and consequently challenging to explain to model consumers who are less-quantitatively initiated.

Practitioners in adjacent disciplines of decision analysis and operations research represent a second constituency. These communities approach uncertainty from a perspective that is epistemically opposite to that commonly employed by the data-science community. Rooted deeply in frequentist paradigms, data scientists and machine-learning professionals are inclined to answer questions along the lines of, "Of what can I be certain?"

Decision and operations analysts tend to approach their practices from a Bayesian perspective. Epistemically, they focus on questions like, "To what extent am I uncertain?" Decision and operations analysts are as a result interested in probability distributions (e.g., [Spetzler1975]). Consequently, the ability to articulate prediction-model residual errors in terms of probability-density functions provides an essential bridge between machine learning and these other communities.

Summary of research.

This research demonstrates an approach to explicitly calculate bias-variance decomposition terms using Keelin's metalog-distributions framework. Closed-form, continuous-function representations of probability distributions result, in addition to descriptive statistics. It also produces a package using the python programming language.

Explicitly, [Hastie2009, eqn (7.10)] presents the decomposition as

$$Err\left(x_0\right) = \sigma_\varepsilon^2 + \left[E\hat{f}\left(x_0\right) - f(x_0)\right]^2 + E\left[\hat{f}\left(x_0\right) - E\hat{f}\left(x_0\right)\right]^2.$$

These terms are the unexplained error, the bias, and the variance terms, respectively. A metalog-based framework yields explicit probability distributions for each.

Bias and variance terms.

The bias and variance terms follow from model residual analysis. Writing in the context of linear regression, Olive encourages the consideration of response and residual plots [Olive2017]. Figure 1 illustrates. Although the figure presents results of a regression model, this analysis can be performed for any prediction model.

Similar techniques provide access to both the bias and variance terms. [Olive2017] also recommends applying Locally-Weighted Scatterplot Smoothing (LOWESS) to response and residual plots for both test and training data. LOWESS involves windowing the scatter under analysis and calculating the local mean [Cleveland1979]. Illustrated in Figure 1, this leads to a locus of points, which when compared to the zero-residual curve provides a visual indication of bias in predictions.

This research calculates metalog distributions for locally-windowed scatters. This produces conditional distributions of the type $p_{\varepsilon_i|\hat{y}_i}\left(\varepsilon_i|\hat{y}_i\right)$ for the residual-plot case. A metalog can similarly be fit to the marginal to obtain $p_{\hat{y}_i}\left(\hat{y}_i\right)$. A joint distribution straightforwardly follows. One consequently obtains a residual-squared statistic

$$\xi_\varepsilon = \int\int \varepsilon^2\, p_{\varepsilon|\hat{y}}\left(\varepsilon|\hat{y}\right)\, p_{\hat{y}}\left(\hat{y}\right)\, d\varepsilon\, d\hat{y}.$$

Distinct probability-density functions $p_{\text{train}}\left(\varepsilon|\hat{y}\right)$ and $p_{\text{test}}\left(\varepsilon|\hat{y}\right)$ characterize residual errors from applying the model to the training and test data sets, respectively. Other descriptive statistics indicate the divergence between the two distributions. Adaptions of the Kolmogorov-Smirnov test [NIST2012] lead to quantitative indications of extent of conformity between the two. It becomes possible to quantitatively characterize errors attributable to overfitting. This research explores such adaptations.

Irreducible-error term.

Quantifying the irreducible-error team follows from using metalog distributions to calculate the Cramer-Rao Lower Bound (CRLB) on estimation-error variance [Kay1993, Hogg2013]. In its most-general form, the CRLB appears in [Kay1993, pp. 39-45]. Defining the *Fisher information matrix* as

$$\left[\mathbf{I}\left(\theta\right)\right]_{i,j} = -\mathcal{E}\left[\frac{\partial^2\, \ln\left(p\left(\mathrm{x};\, \theta\right)\right)}{\partial\theta_i\, \partial\theta_j}\right],$$

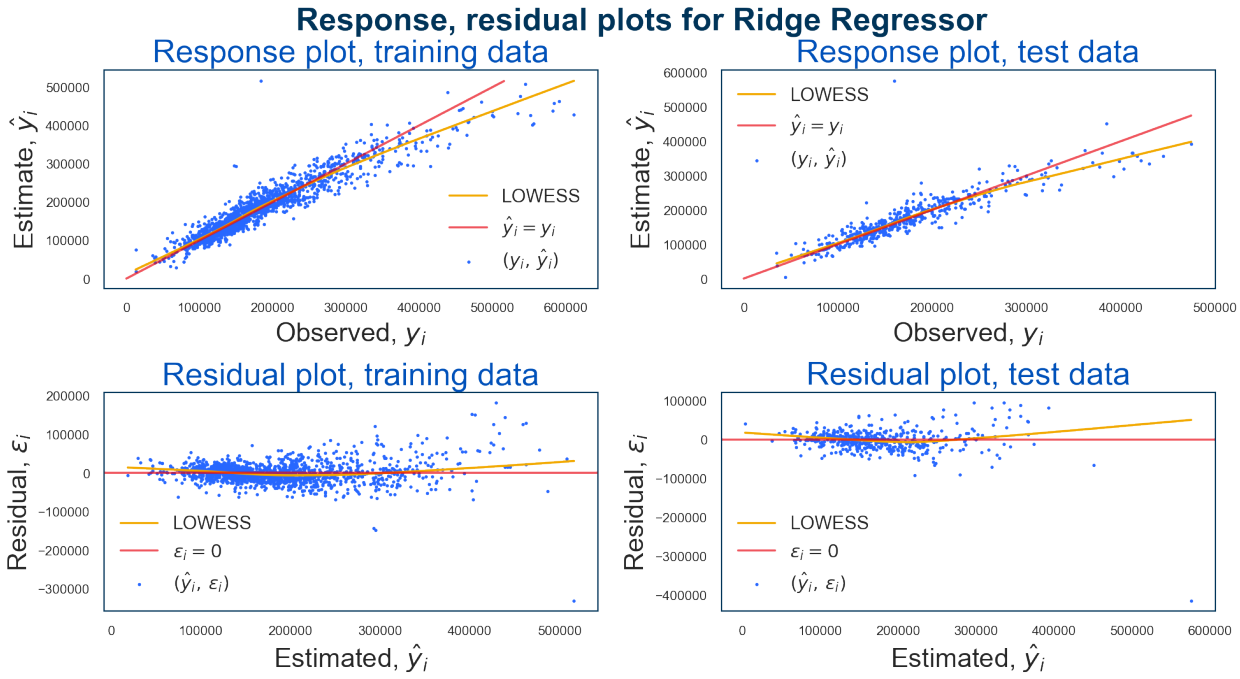
a lower bound on the estimation-error variance for θ_i is calculated using

$$\text{var}\left\{\hat{\theta}_i\right\} \geq \left[\mathbf{I}^{-1}\left(\theta\right)\right]_{i,i}.$$

Now, metalog distributions are presently only available for univariate distributions. The Fisher information matrix is generally defined in terms of multivariate distributions $p\left(\mathrm{x};\, \theta\right)$. Metalog distributions can obtained for distinct, individual, univariate components of a multivariate distribution represented in an orthogonal-transform space. QR and singular-value decomposition are examples (e.g., [Horn2013, Jolliffe2002]).

Obviously $Pr\left\{\mathrm{x} \leq \mathscr{X}\right\}$ is associated with an infinite quantity of $Pr\left\{u_i \leq \mathscr{U}_i\right\}$ values for $\mathscr{X} \subset \mathscr{U}_1 \times \ldots \times \mathscr{U}_N$, where the $\left\{\mathscr{U}_n\right\}$ are the components of the orthogonal-transform-space components of $\mathscr{X}$. This problem occurs when seeking interval probabilities in $\mathscr{X}$. This complication can be handled using optimization techniques, such as Data-Envelope Analysis (e.g., [Ozcan2008]).

Figure 1: Quantitative characterization of bias-variance decomposition terms is accessible from predictive-model residual analysis.



References

[Agresti2013] A. Agresti, *Categorical Data Analysis*, third edition, Hoboken, NJ: Wiley, 2013.

[Cleveland1979] W. S. Cleveland, "Robust locally weighted regression and smoothing scatterplots", *Journal of the American Statistical Association*, 74(368)829-836, December 1979.

[Cover2006] T. M. Cover, J. A. Thomas, *Elements of Information Theory*, second edition, Hoboken, NJ: Wiley, 2006.

[Dielman2005] T. E. Dielman, *Applied Regression Analysis*, Mason, OH: South-Western Cengage Learning, 2005.

[Hastie2009] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning*, second edition, New York: Springer Science, 2009.

[Hogg2013] R. V. Hogg, J. W. McKean, A. T. Craig, *Mathematical Statistics*, seventh edition, Boston, MA: Pearson, 2013.

[Horn2013] R. A. Horn, C. R. Johnson, *Matrix Analysis*, second edition, Cambridge, UK: Cambridge University Press, 2013.

[Jolliffe2002] I. T. Jolliffe, *Principal Component Analysis*, second edition, New York: Springer-Vaerlag, 2002.

[Kay1993] S. M. Kay, *Signal Processing: Estimation Theory*, Englewood Cliffs, NJ: Prentice Hall, 1993.

[Keelin2016] T. W. Keelin, "The metalog distributions", *Decision Analysis*, INFORMS, 13(4):243-277.

[NIST2012] *Engineering Statistics Handbook*, National Institute for Standards and Technology, 2012, https://bityl.co/4YRI.

[Olive2017] D. J. Olive, *Linear Regression Analysis*, Cham, CH: Springer, 2017.

[Ozcan2008] Y. A. Ozcan, *Health Care Benchmarks and Performance Evaluation*, New York: Springer Science + Business, 2004.

[Sahai2004] H. Sahai, M. M. Oheda, *Analysis of Variance for Random Models: Volume I*, New York: Springer Science + Business, 2004.

[Spetzler1975] Carl S. Spetzler, Carl-Axel S. Staël Von Holstein, (1975) "Exceptional Paper—Probability Encoding in Decision Analysis", *Management Science*, INFORMS, 22(3):340-358.