2020

In praise of partially interpretable predictors

Tri Le

Bertrand S. Clarke

Received: 16 August 2017 | Revised: 6 April 2019 | Accepted: 18 December 2019

DOI: 10.1002/sam.11450

WILEY

RESEARCH ARTICLE

In praise of partially interpretable predictors

Tri Le[1] | **Bertrand Clarke**[2]

[1]Department of Mathematics, Science, and Informatics, Mercer University, Atlanta, Georgia

[2]Department of Statistics, University of Nebraska-Lincoln, Lincoln, Nebraska

Correspondence
Tri Le, Department of Mathematics, Science, and Informatics, Mercer University, 3001 Mercer University Dr., Atlanta, GA 30341.
Email: le_tm@mercer.edu

Funding information
the NSF, DMS-1419754

Abstract

Often there is an uninterpretable model that is statistically as good as, if not better than, a successful interpretable model. Accordingly, if one restricts attention to interpretable models, then one may sacrifice predictive power or other desirable properties. A minimal condition for an interpretable, usually parametric, model to be better than another model is that the first should have smaller mean-squared error or integrated mean-squared error. We show through a series of examples that this is often not the case and give the asymptotic forms of a variety of interpretable, partially interpretable, and noninterpretable methods. We find techniques that combine aspects of both interpretability and noninterpretability in models seem to give the best results.

KEYWORDS

mean-squared error, model average, Nadaraya–Watson, Priestley–Chao, stacking

1 | INTRODUCTION

Extracting information from a data set requires an analyst to choose judiciously from among potentially applicable statistical procedures and balance competing factors. Often the procedures are forms of regression: fixed-effects linear regression and nonparametric regression are perhaps the most common but there are many others. Factors to be balanced include model bias, interpretability, variability, complexity, sample size, and the mean-squared error (MSE) of the chosen predictor. In such situations, many scientists engage in model building. By this, they mean they have a data generator (DG) that produces outcomes Y, and they think they have some relevant information about it. For instance, there may be explanatory variables $X = (X_1, \ldots, X_d)$ that are potentially involved in determining Y.

Scientists often turn to statisticians to help them use the X_j's to "model" Y, i.e., to find a mathematical expression that encapsulates Y in the sense of representing it in terms of the explanatory variables, apart from random error. In this context, we focus on the case where the information in the data is regarded as having been summarized by the predictor from a regression function that is assessed by its MSE. Of the numerous other settings that could have been chosen, we have chosen this because it is often used, is relatively simple, and is a convenient paradigm case.

One of the main factors often only tacitly examined in this approach is interpretability. This is so because scientists often simply impose extra constraints such as requiring the terms appearing in the final model be physically meaningful, i.e., have a real-world interpretation, and be consistent with accepted models for closely related phenomena. Scientists are generally not satisfied with just any expression that helps them predict Y; they want to "understand" the DG.

A key point of this paper is that, as appealing as conventional modeling is, it is also often suboptimal predictively,

specifically in terms of MSE. We argue that extending an interpretable model to include uninterpretable components often increases its flexibility and so often gives better MSE properties for realistic sample sizes. Recalling that MSE is a sum of squared bias and variability, we suggest that the key improvement from using more flexible models is mostly in reducing bias since more flexibility often increases variability. The net result of lower MSE hopefully also provides improved prediction.

Indeed, the bias of a model has been identified as a key source of lack of reproducibility of inferences. For instance, Milkowski et al. [21] focus on reproducing an entire experiment and implicitly acknowledge that bias is a central problem in neuroscience. More explicitly, Ransohoff [26], amongst others, focuses on this point as he offers a detailed critique on the role of modeling in cancer experimentation and biomarkers, concluding "all models are guilty of bias until proved innocent".

Thus our basic question is: If one insists that the final announced model be interpretable physically, as opposed to merely providing good predictions, what is the apparent cost in MSE? In answering this question, one is led to conclude that the physical justifications for models may seem valid but often not to the exactitude or precision required especially when good prediction is a goal. This is seen in Weng et al. [33] and more generally is consistent with Milkowski et al. [21] (and the references therein). In addition, analysts and experimentalists alike want to avoid being misled by insisting on a level of interpretability that cannot be justified. Formally, we focus on the difference between seeking interpretable models to generate predictors and merely seeking good predictors.

One way to answer this question is to identify an interpretable model and then compare its MSE to an alternative model from a larger class that may not be fully interpretable. If sample size considerations are not a concern, then the flexible model should never perform worse than the interpretable model – assuming that the interpretable model is simply an interpretable member of the larger class. The bigger the model class searched, the better the model found should be, provided sample sizes are adequate.

Ideally, the interpretable model should reflect the best physical modeling that can be done, and the more flexible model class should extend the best physical model class substantially in some relevant sense. Thus, here we compare the performance of an interpretable model with either an uninterpretable or partially interpretable model based on it to assess the cost of interpretability. Built into this view is that interpretable models are rarely, if ever, true to arbitrary precision.

The sense in which we use the term interpretable is given precisely in (1). We start with a model that is interpretable by construction in that sense, i.e., its components and operations have physical correlates. Then, rather than using model selection techniques to choose extra (interpretable) terms to improve prediction, we use statistical operations, such as model averaging and nonparametrics, to improve the interpretable model-based predictor. The extra components and operations do not admit interpretations, i.e. have no necessary physical correlates; however, they often improve predictions. For convenience, we use MSE as a proxy for predictive error. This is not the same as assessing variance-bias tradeoffs for models because the goal is not modeling; it is improving prediction and assessing the degree of improvement over a model that is taken as interpretable and appropriate – perhaps not the "best possible" but certainly not discredited on physical grounds. We leave the concept of a best possible model unexamined apart from referencing discussions of $\mathcal{M}$-closed, -complete, and -open problems; see Clyde and Iversen [7] Clarke et al. [6], Clarke et al. [5], Le and Clarke [14], and Clarke and Clarke [4].

This leads to the somewhat surprising operational point of this paper: Usually the best models one can find using data include some components that are interpretable (and not too far wrong) and some components that are not interpretable, i.e., the purely interpretable and the purely uninterpretable extremes are often suboptimal. One can argue that this point is implicitly accepted by the frequent use of, for instance, model averaging techniques. Our point, however, is more than this: Purely model-based techniques are not merely an ideal that is sometimes not achieved. Modeling by itself is, all too often, just suboptimal. Indeed, the common paradigm of proposing a model, falsifying it, and proposing a better model that can in turn be falsified eventually leading to truth is no longer a viable paradigm for much of contemporary research. In our view, despite being apparent, the severe limitations of this "model falsification" paradigm have not been studied sufficiently. So, let us turn the model falsification paradigm on itself: Given that the model falsification paradigm has itself largely been falsified, what do we propose instead? One answer is explicitly blending physical modeling and purely statistical techniques, as this paper proposes.

A feature of interpretability that is not formally discussed in the statistics literature as much as it could be is how detailed the modeling is. At one point does one draw a line and say: We will include these features in a model and ignore the others? For instance, it is common to write networks of reactions among hydrocarbons, DNA, RNA, and proteins ignoring the mechanism of protein synthesis. Likewise, we often speak of transcription of a gene in terms of nucleotides ignoring the regulatory role of chromosomal proteins. At root, a judgment call is being made about what is most important to include in a model so it will be useful.

We refer to this choice as the level of detail of modeling since the usual assumption is that the omitted features are less important, and presumed to be a finer scale, i.e., more detailed, than the features included. It is reasonable to suggest that the more detailed the interpretability required of a model is, the harder it will be to find good models. Indeed, the improvement from including noninterpretable components may compensate for omitted levels of modeling detail perhaps because more detailed modeling is infeasible.

One extreme case concerns models that have interpretable axiomatic derivations. For example, there are well-known axiomatic derivations for some stochastic models such as the binomial and for some deterministic models such as the Navier–Stokes equations. These are not counterexamples to the main point of this paper, since verifying that the axioms are satisfied to arbitrary precision in a given setting is exceedingly difficult and may even be wrong given the ubiquity of molecular effects, e.g., in liquids like protoplasm. Specifically, in practice it is often difficult to verify that Bernoulli outcomes are perfectly independent and identical or that a real fluid is perfectly isotropic. In either of these cases, approximations to the model would likely have to be used and their success assessed by comparison with other predictors. So, even in these cases, a more flexible predictor can improve on the predictor given by the axiomatically derived model by being more responsive to the data.

At the other extreme are purely nonparametric models: These models should be regarded as uninterpretable and usually have slower rates of convergence than fully interpretable models (unless the interpretable models are highly complex). The slower rate of convergence can make them inferior to good interpretable models, meaning that the interpretable model gives such a good approximation to the true model that it cannot be improved given the sample size. In these cases, constraining the model space by including interpretable components in an uninterpretable model may improve it if the interpretable components are accurate. That is, again, combining interpretable and uninterpretable aspects in a single model can give better results than purely uninterpretable models.

Ensemble methods are a class of predictive strategies that typically combine interpretable and uninterpretable aspects. So, given the discussion above, it is not a surprise that they often do well predictively, at least asymptotically. Specifically, since they combine models, they are usually only partially interpretable. Hence, they fall in the mid-range between the two extremes where we argue the best strategies often lie. Perhaps the earliest results on the optimality of ensemble methods are for Bayes model averages and are due to Skouras and Dawid [28] and Raftery and Zheng [25].

Model averaging methods are predicated on the idea that a (usually convex) mixture of models will outperform any of its component models. Otherwise put, enlarging the model space to include more models – which can lead to noninterpretability – improves the prediction provided by any component of the ensemble. Usually a "sanity" criterion is desired: If one of the component models in the average is "true" or at least closest to the true model, the ensemble defaults to it asymptotically. This does not contradict the fact that a simple model that is a good approximation may be better for small sample sizes than a correct but complex model.

These heuristics are consistent with the model average studied in Mays [17] and developed in successive works such as Mays et al. [19], Mays et al. [20], and Mays and Birch [18]. Their main point is to treat a response Y as the convex combination of a linear model (LM) and a Nadaraya–Watson (NW) estimator (see Nadaraya [22] and Watson [32]). The parameter controlling the tradeoff between the two function estimators, say the weight on the LM estimator, $\widehat{\gamma} \in [0, 1]$, is optimal under a squared error criterion. If the linear model is true to infinite precision, then we expect that $\widehat{\gamma} \to 1$ in limit of large n and the model average will reduce to the LM term. In this case, the coefficient on the term with the NW estimator is $1 - \widehat{\gamma} \to 0$, meaning the term vanishes. On the other hand, if the LM is not correct, we expect that scenarios in which $\widehat{\gamma} \to 0$ can be constructed. That is, $\widehat{\gamma}$ may converge to a constant in (0, 1) representing the most useful tradeoff between LM and a NW estimator even taking into account their different rates of convergence, see Clauses 1 and 3 of Theorem 3.1 below. (The limitation is that, if there are too many parameters in the linear model, its error term $\mathcal{O}(1/n)$ may be larger than the error term $\mathcal{O}(1/n^{4/5})$.) An optimal limiting value for $\widehat{\gamma}$ in (0, 1) would indicate that the best tradeoff between an interpretable LM and a noninterpretable NW estimator identifies a partially interpretable predictor better than either, consistent with the intuition in the paper.

Much more remains to be said. First, instead of choosing $\widehat{\gamma}$ as before, we use a more general and successful model average, namely, stacking. Stacking (STK) was invented in Wolpert [34] and obtains model coefficients (for any number of models) using a cross-validation optimization. Importantly, stacking coefficients have a formal consistency property. Thus we can show in Theorem 3.2 (Clause 1) that (*STK*, *PC*) (see Section 2.2.4 for a definition) is never worse than PC alone in terms of asymptotic convergence rates. Here, PC stands for the Priestley–Chow estimator (the fixed design analog of the NW estimator). Second, we take a composition of functions so that, rather than expressing Y as a sum of two terms, we express it as a sequence of function compositions in which the first stages preprocess the data. That is, even though we

include a model averaging procedure, our overall predictor is an ensemble method but not a model average. Thus, our predictors are different from those in earlier work.

In terms of our general thesis that predictors that combine some interpretable and noninterpretable aspects are better than either extreme, we find that for realistic sample sizes we get the best performance by enlarging LMs to $(STK, PC) - LM$; see Section 2.2.5 for a definition. We do not believe this will always be the case. However, $(STK, PC) - LM$ does provide an envelope around a LM that enlarges the search for good predictors. Our results also show that stacking function estimators tend to improve them while applying LM to a function estimator tends to worsen it, asymptotically. Even though this only holds for the specific LMs used in our examples, we suspect the principle holds more generally.

In Section 2 we begin with a theoretical example. Then, after defining our six function estimators that can be used as predictors, we give a real-data example to demonstrate our main points. In Section 3 we compare the asymptotic cost of our six predictors in terms of MSE and integrated MSE (IMSE) theoretically. In Section 4 we present two more examples with data, and in Section 5 we summarize our overall findings. Throughout, our examples with data are not applied in the sense that we are trying to solve a "real" problem. We are arguing that model-based prediction is improved by using statistical non-modeling-based techniques, at least in an MSE sense.

2 | MOTIVATION

Let us clarify our use of the term "interpretability." Suppose a model M consists of m components, say

$$M = \{c_1, \ldots, c_m\}. \tag{1}$$

The c_j's represent the ingredients that go into the formulation of a model such as variables, parameters, and operations indicating how the other elements of M are to be used. The model M is interpretable if and only if all its components, the c_j's, correspond to possible mechanisms within the DG that experts would not regard as pre-experimentally discredited on physical grounds. Conventionally, we say a model is valid if and only if it is interpretable and correct – at least to the degree that its predictions are sufficiently close to their corresponding future outcomes.

This definition of interpretability is very general so only a few cases of it can be examined in detail here. First, we look at one generic case – model averaging by stacking – and see theoretically that giving up some interpretability can be helpful in an MSE sense. Second, we use a dataset to see that under an MSE criterion the best function estimator (within a collection of six, each corresponding to a different model "M") uses both interpretable and noninterpretable stages in its composition.

2.1 | A theoretical result

First, we see a generic example in which stacking models outperforms any of the components in the average, asymptotically, in an MSE sense. That is, each model may be interpretable while the "stack" they give is not but still performs better predictively, or at least no worse, than any of its components. This result holds even if the individual models have physical interpretations that are mutually contradictory.

To form any model average, we must have a list of models and a way to combine them that will result in a predictor. So, let $\{f_1, \ldots, f_J\}$ be a uniformly bounded set of individually interpretable regression functions with $f_j : \Omega \to \mathbb{R}$ where $\Omega \subseteq \mathbb{R}^d$ is the closure of its interior. Often, the f_j's will have parameters in them, i.e., $f_j(x) = f_j(x|\,\theta_j)$, but for ease of exposition we will ignore this. Now, write $Y(x) = f_T(x) + \varepsilon$ for the true model of Y where ε is a mean-zero, finite-variance error term. This gives J candidate interpretable models $Y_j(x) = f_j(x) + \varepsilon$. With only minor loss of generality we assume that f_T and the f_j's are uniformly bounded elements of an L^2 space (denoted $\mathcal{L}^2$) that has a countable basis.

Denote a dataset of size n by $\mathcal{D} = \mathcal{D}_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, gathered sequentially, and assumed independent across (x, y) pairs. Now, we have J point predictors $Y_j(x_i)$ for the ith step, $i \geq k+1$, apart from a burn in period of length, say, k. To obtain a model average, write $\widehat{Y}_{n+1}(x_{n+1}) = \sum_{j=1}^{J} \alpha_j Y_j(x_{n+1})$ and suppose the α_j's are stacking weights, i.e., when estimated from $\mathcal{D}$, $\widehat{\alpha} = (\widehat{\alpha}_1, \cdots, \widehat{\alpha}_J)$ is

$$\widehat{\alpha} = \arg\min_{\alpha \in S} \sum_{i=1}^{n} \left(y_i(x_i) - \sum_{j=1}^{J} \alpha_j \widehat{f}_{j,-i}(x_i) \right)^2,$$

where $S \subset \mathbb{R}^J$. Then, for any x, the stacking model average point predictor is

$$\widehat{Y}_{stack}(x) = \sum_{j=1}^{J} \widehat{\alpha}_j f_j(x). \tag{2}$$

Using (1), given $\mathcal{D}$, let $M_j = \{f_j(\cdot), \varepsilon, +\}$ and $M_0 = \{f_1(\cdot), \ldots, f_J(\cdot), \varepsilon, STK\}$, i.e., the M_j's correspond to the J individual models and M_0 corresponds to the stacking model based on (2). The main result of this subsection is the following.

Theorem 2.1. *Suppose the true model can be written as* $f_T(x) = \sum_{j=1}^{J^*} \alpha_j^* f_j^*(x) + \varepsilon_{J^*}(x)$ *where the* f_j^**'s are bounded,*

orthonormal, continuous, and orthogonal to ε_{J^*} *and that* $\sup_x \varepsilon_{J^*}(x) \to 0$ *(in the* L^2*-norm) as* $J^* \to \infty$*. Write* $\widehat{Y}_{stack}(\cdot) = \sum_{j=1}^{J} \widehat{\alpha}_j f_j(\cdot)$ *in which the* f_j*'s are orthonormal and continuous and that as* $J, J^* \to \infty$*,* $\langle \{f_j\} \rangle$ *and* $\langle \{f_j^*\} \rangle \to \mathcal{L}^2$ *(in the* L^2 *norm). Then,*

$$\limsup_{J,J^*,n\to\infty} \left[\int E_Y(Y(x) - f_j(x))^2 dx - \int E_Y(Y(x) - \widehat{Y}_{stack}(x))^2 dx \right] \geq 0.$$

Remark 2.1. The intuitive content of this result is that M_0 is better than any M_j, in a limiting sense. That is, using stacking – which is uninterpretable – is better than using any of the f_j's that are assumed to be interpretable. We believe this result will extend to f_j's having finite-dimensional parameters θ_j.

Remark 2.2. In the statement of this result, the f_j's are orthonormal. This can be readily generalized to allow any uniformly bounded set of functions with the property $\langle \{f_j\} \rangle \to \mathcal{H}$ where $\mathcal{H}$ is a sub-Hilbert space of $\mathcal{L}^2$ that has $f_T \in \mathcal{H}$.

Proof. An easy extension of Theorem 3.2 in Le and Clarke [14] gives that

$$\widehat{\alpha}_j \to \langle f_j, f_T \rangle = \alpha_j \quad \text{as} \quad n \to \infty,$$

in the probability associated with Y. Also, since the L^2 space has a countable basis, for any J^*, there is a J large enough that

$$\left| \sum_{j=1}^{J^*} \alpha_j^* f_j^*(x) - \sum_{j=1}^{J} \alpha_j f_j(x) \right|$$

can be made arbitrarily small in the L^2 norm. From the assumptions on the f_j's and ε_{J^*},

$$\langle f_j, e_{J^*} \rangle = \int f_j(x) e_{J^*}(x) dx \to 0 \quad \text{as } J^* \to \infty.$$

Taking E in f_T, i.e., $E_{Y,}$ for a new value of x

$$E_Y(Y(x) - \widehat{Y}_{stack}(x))^2 = E_Y(Y - E_Y Y)^2 + E_Y\left(E_Y Y - \sum_{j=1}^{J} \widehat{\alpha}_j f_j(x) \right)^2 + 2E_Y(Y - E_Y Y)\left(E_Y Y - \sum_{j=1}^{J} \widehat{\alpha}_j f_j(x) \right)$$

The last term is zero by the independence assumption. Adding and subtracting $\sum_{j=1}^{J} \alpha_j f_j(x)$ in the middle term on the right gives

$$E_Y(Y(x) - \widehat{Y}_{\text{stack}}(x))^2$$

$$= \sigma^2 + \left(\sum_{j=1}^{J^*} \alpha_j^* f_j^*(x) + e_{J^*} - \sum_{j=1}^{J} \alpha_j f_j(x) \right)^2 \tag{3}$$

$$+ 2\left(\sum_{j=1}^{J^*} \alpha_j^* f_j^*(x) + e_{J^*} - \sum_{j=1}^{J} \alpha_j f_j(x) \right)\left(\sum_{j=1}^{J} E(\widehat{\alpha}_j - \alpha_j) f_j(x) \right) \tag{4}$$

$$+ E\left(\sum_{j=1}^{J} \sum_{j'=1}^{J} (\widehat{\alpha}_j - \alpha_j) f_j(x) (\widehat{\alpha}_{j'} - \alpha_{j'}) f_{j'}(x) \right)^2 \tag{5}$$

Let $\mathbf{1}_{j,j'} = 1 \iff j = j'$ and zero otherwise. Now, doing the same for any f_j:

$$E_Y(Y - f_j(x))^2 = \sigma^2 + \left(\sum_{j=1}^{J^*} \alpha_j^* f_j^*(x) + e_{J^*} - \sum_{j=1}^{J} \alpha_j f_j(x) \right)^2 \tag{6}$$

$$+ 2\left(\sum_{j=1}^{J^*} \alpha_j^* f_j^*(x) + e_{J^*} - \sum_{j=1}^{J} \alpha_j f_j(x) \right)\left(\sum_{j'=1}^{J} (\mathbf{1}_{j,j'} - \alpha_{j'}) f_{j'}(x) \right) \tag{7}$$

$$+ \left(\sum_{j'=1}^{J} (\alpha_{j'} - \mathbf{1}_{j,j'}) f_{j'}(x) \right)^2 \tag{8}$$

It is seen that terms (3) and (6) are the same and that terms (4) and (7) go to zero. Since term (8) goes to a non-negative constant and term (5) goes to zero, the result follows. □

For completeness, we note two points: First, interpretability and complexity are different concepts. The complexity of the model M refers to the number of components that interact and the variety of ways in which they interact, regardless of any real-world correlates. Second, more complex DGs will often have higher bias than less complex DGs, and conversely, for fixed sample size. On the other hand, complex DGs may sometimes be approximated by less complex models, thereby giving low bias. Nevertheless, intuitively, it is easier to model less complex DGs so the bias can often be reduced by simply increasing sample size. To avoid excessive digression, we do not examine the interactions between complexity and bias here.

2.2 | Defining the function predictors

In this subsection we define six predictors. Two are familiar, namely the LM predictor and the PC predictor. The third is the PC linear model (PC-LM) predictor in which the fitted values of a PC predictor are fed into LMs as the "Y". The other three involve stacking but differ in the

details. The simplest is to stack PC predictors (STK, PC), i.e., use stacking on components found from applying PC to bootstrapped datasets. The fifth uses the fitted values from stacking PCs as the "Y" in LMs, (STK,PC)-LM. The sixth is to stack linearized PCs, STK,(PC-LM). The last three use bootstrapping to form components to stack; see Le and Clarke [14] for details.

Note that our formal definitions are for the case of a d-dimensional explanatory variable. Indeed, two explanatory variables will be used in the Tour de France data in Section 2.3, and the intuition in Section 1 is independent of dimension. The two examples in Section 4 also have $d = 2$.

2.2.1 | Linear model

The LM predictor of the response for a new value of the explanatory variable x_{new} is interpretable and, in the usual notation, is

$$\widehat{Y}_{\mathrm{LM}}(x_{\mathrm{new}}) = x'_{\mathrm{new}}\widehat{\beta}_{\mathrm{LM}} = x'_{\mathrm{new}}(X'X)^{-1}X'Y,$$

where $Y = (Y_1, \ldots, Y_n)'$, $x'_{\mathrm{new}} = (1, x_{\mathrm{new}})$,

$$X' = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ x_1 & x_2 & \ldots & x_n \end{pmatrix},$$

and $x_{\mathrm{new}}, x_1, \ldots, x_n \in \Omega \subseteq \mathbb{R}^d$.

2.2.2 | Priestley–Chao

First, we define the PC predictor when $d = 1$ and then generalize it when $d \geq 2$.

Assume $x_1, \ldots, x_n \in \mathbb{R}$ are univariate and $(x_1, y_1), \ldots, (x_n, y_n)$ are generated by the model $Y_i = f(x_i) + \varepsilon_i$, $i = 1, \ldots, n$ where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. $(0, \sigma^2)$, and the design points are equidistant in [0, 1], i.e.,

$$x_i = \frac{i-1}{n-1}, \quad i = 1, \ldots, n. \tag{9}$$

Let $f : [0, 1] \to \mathbb{R}$ be the underlying function to be estimated, and choose a fixed kernel K symmetric about zero. The PC predictor of the response for a new value $f(x_{\mathrm{new}})$, see Priestley and Chao [24], for the deterministic design (9) is

$$\begin{aligned}\widehat{Y}_{\mathrm{PC}}(x_{\mathrm{new}}) &= \frac{1}{n}\sum_{i=1}^{n} K_{h_n}(x_{\mathrm{new}} - x_i)Y_i \\ &= \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x_{\mathrm{new}} - x_i}{h_n}\right)Y_i,\end{aligned}$$

where $x_{\mathrm{new}} \in [0, 1]$, and h_n is a positive real number. Asymptotic expressions for h_n will be seen to follow from a variance–bias tradeoff argument. Sometimes, the bandwidth parameter h_n is regarded as a complexity because the curve tends to oscillate more as $h_n \to 0$ for fixed n. However, this interpretation is not necessary, and here we simply regard h_n as a parameter indexing a collection of curves.

A related function estimator is the NW one defined by

$$\widehat{Y}_{\mathrm{NW}}(x_{\mathrm{new}}) = \frac{\sum_{i=1}^{n} K_{h_n}(x_{\mathrm{new}} - x_i)Y_i}{\sum_{i=1}^{n} K_{h_n}(x_{\mathrm{new}} - x_i)}.$$

The difference between PC and NW is that NW is typically derived under the assumption that the x_i's are probabilistically generated whereas PC treats the x_i's as design points. For full comparability, we have only computed estimators that treat the x's as design points. However, mathematically, we could, for example, replace each occurrence of PC with NW. The asymptotics of PC and NW are nearly identical, and neither PC nor NW can be reasonably regarded as interpretable. Consequently, their performance should be very similar and this is noted in Theorems 3.1, 3.2, and 3.3 below.

Now, assume $x_1, \ldots, x_n \in \Omega \subseteq \mathbb{R}^d$ and the kernel K is also d-variate, $d \geq 2$. In this case, the bandwidth h_n becomes a $d \times d$ positive definite matrix H_n and the function

$$K_{H_n}(u) = |H_n|^{-1}K(H_n^{-1}u)$$

is now used to assign weights. (Here, $|\cdot|$ denotes determinant.) For a uniform random design, the PC predictor of $f(x_{\mathrm{new}})$ is

$$\widehat{Y}_{PC}(x_{\mathrm{new}}) = \frac{b}{n}\sum_{i=1}^{n} K_{H_n}(x_{\mathrm{new}} - x_i)Y_i,$$

where b is the volume of the design region $\Omega \subseteq \mathbb{R}^d$. Usually, K is bounded, has all odd moments zero (i.e., $\int u_1^{\ell_1} \cdots u_d^{\ell_d} K(u)\ du = 0$), and satisfies $\int uu^T K(u)\ du = \mu_2(K) I_d$ (i.e., the integral of the outer product is a constant depending on the second moment of K times the d-dimensional identity matrix). In the multivariate case, the NW predictor is defined by

$$\widehat{Y}_{\mathrm{NW}}(x_{\mathrm{new}}) = \frac{\sum_{i=1}^{n} K_{H_n}(x_{\mathrm{new}} - x_i)Y_i}{\sum_{i=1}^{n} K_{H_n}(x_{\mathrm{new}} - x_i)}.$$

2.2.3 | Priestley-Chao linear model

The PC-LM predictor of the response for a new value of f at the explanatory variable x_{new} is

$$\widehat{Y}_{\text{PC-LM}}(x_{\mathrm{new}}) = x'_{\mathrm{new}}\widehat{\beta}_{\text{PC-LM}} = x'_{\mathrm{new}}(X'X)^{-1}X'\widehat{Y}_{\mathrm{PC}},$$

where $\widehat{Y}_{\mathrm{PC}} = (\widehat{Y}_{\mathrm{PC}}(x_1), \ldots, \widehat{Y}_{\mathrm{PC}}(x_n))'$, $x'_{\mathrm{new}} = (1, x_{\mathrm{new}})$,

$$X' = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ x_1 & x_2 & \ldots & x_n \end{pmatrix}.$$

and $x_{\mathrm{new}}, x_1, \ldots, x_n \in \Omega \subseteq \mathbb{R}^d$. Since LM is interpretable, PC-LM is partially interpretable.

2.2.4 | Stacking Priestley-Chao

Stacking was first introduced by Wolpert [34] and studied primarily as a predictor in numerous contexts such as regression Breiman [1], Clarke [3], Sill et al. [27], density estimation Smyth and Wolpert [29], classification and distance learning Ozay and Vural [23].

The basic idea is that J signal plus noise models of the form $Y = f_j(x) + \varepsilon, f_j : \Omega \subseteq \mathbb{R}^d \to \mathbb{R}$, for $j = 1, \ldots, J$ can be usefully combined to give the predictor (2). Here, we generalize (2) to allow f_j to be of the form $f_j(x_{\mathrm{new}}) = f_j(x_{\mathrm{new}}, \beta)$ where β is a parameter. Thus, $\widehat{f}_j(x_{\mathrm{new}}) = f_j(x_{\mathrm{new}}, \widehat{\beta})$ where $\widehat{\beta}$ is an estimator of β. Expression (2) corresponds to leave-one-out CV but can be readily modified to correspond to leave-K-out CV. Here, $S = \mathbb{R}^J$, but other choices are possible.

Depending on the $\widehat{f}_j$'s, STK is partially interpretable or not. Here we stack PC predictors $\widehat{Y}_{\mathrm{PC},j}$, $j = 1, \ldots, J$, obtained by drawing J bootstrap samples from the original data, so the STK-PC predictor of the response for a new value of the explanatory variable x_{new} is

$$\widehat{Y}_{\mathrm{STK,PC}}(x_{\mathrm{new}}) = \sum_{j=1}^{J} \widehat{\alpha}_j \widehat{Y}_{\mathrm{PC},j}(x_{\mathrm{new}})$$

and is not interpretable.

2.2.5 | (Stacking Priestley–Chao) linear model

The (STK,PC)-LM predictor of the response for a new value of the explanatory variable at x_{new} is

$$\begin{aligned} \widehat{Y}_{(STK,PC)-LM}(x_{new}) &= x'_{new}\widehat{\beta}_{(STK,PC)-LM} \\ &= x'_{new}(X'X)^{-1}X'\widehat{Y}_{STK,PC}, \end{aligned}$$

where $\widehat{Y}_{\mathrm{STK,PC}} = (\widehat{Y}_{\mathrm{STK,PC}}(x_1), \ldots, \widehat{Y}_{\mathrm{STK,PC}}(x_n))'$, $x'_{\mathrm{new}} = (1, x_{\mathrm{new}})$,

$$X' = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ x_1 & x_2 & \ldots & x_n \end{pmatrix}.$$

and $x_{\mathrm{new}}, x_1, \ldots, x_n \in \Omega \subseteq \mathbb{R}^d$. As in Section 2.2.4, $\widehat{Y}_{\mathrm{STK,PC}}$ is formed by using bootstrap samples. This is only slightly interpretable. Note that it is difficult to quantify such assessments.

2.2.6 | Stacking (Priestley-Chao linear model)

If we stack PC-LM predictors $\widehat{Y}_{\mathrm{PC\text{-}LM},j}$, $j = 1, \ldots, J$, then the STK-(PC, LM) predictor of the response for a new value of the explanatory variable $x_{\mathrm{new}} \in \Omega \subseteq \mathbb{R}^d$ is

$$\widehat{Y}_{\mathrm{STK,(PC\text{-}LM)}}(x_{\mathrm{new}}) = \sum_{j=1}^{J} \widehat{\alpha}_j \widehat{Y}_{\mathrm{PC\text{-}LM},j}(x_{\mathrm{new}}),$$

where $\widehat{Y}_{\mathrm{PC\text{-}LM},j}$ is the PC-LM predictor for model j defined in Section 2.2.3 formed by bootstrapping. As in Section 2.2.5, STK-(PC,LM) is only slightly interpretable.

2.2.7 | MSE and IMSE

Recall that, for good prediction, the MSE of a predictor must be small. For function estimation, the MSE of the predictor $\widehat{Y}$ at x is given by

$$MSE(\widehat{Y}(x)) = E[(\widehat{Y}(x) - f(x))^2].$$

This breaks down into two parts. The bias of $\widehat{Y}$ at x is

$$Bias(\widehat{Y}(x)) = E(\widehat{Y}(x)) - f(x);$$

the variance of $\widehat{Y}$ at x is

$$Var(\widehat{Y}(x)) = E[(\widehat{Y}(x) - E(\widehat{Y}(x)))^2];$$

and the MSE can be decomposed as

$$MSE(\widehat{Y}(x)) = Var(\widehat{Y}(x)) + [Bias(\widehat{Y}(x))]^2.$$

Now, the IMSE of the predictor $\widehat{Y}$ is defined by

$$IMSE(\widehat{Y}) = \int MSE(\widehat{Y}(x))dx.$$

In the next section we compare asymptotic expansions for various MSEs and then give inequalities they or the IMSEs they generate satisfy.

2.3 | A real-data example

As an example of the class of phenomena to which our ideas apply, consider the Tour de France data. We will see

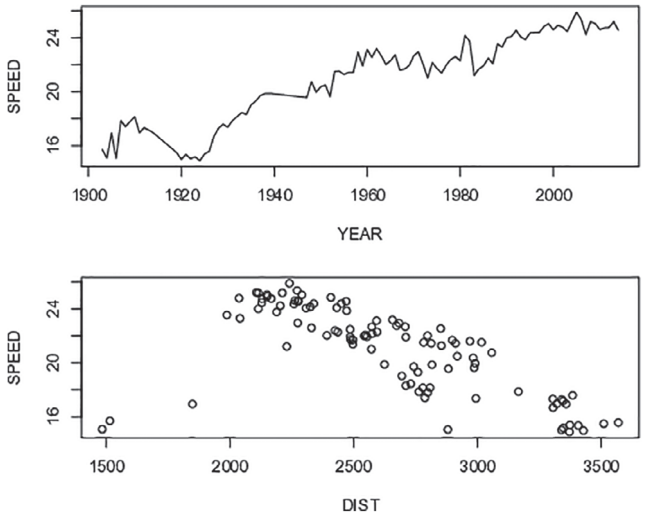

FIGURE 1 Scatter plots of SPEED vs. YEAR and SPEED vs. DIST for the Tour de France data. The dip in the top graph from 1915 to 1925 is due to the death toll of the First World War. The three data points at the lower left of the bottom graph are from the first 3 years of the Tour de France.

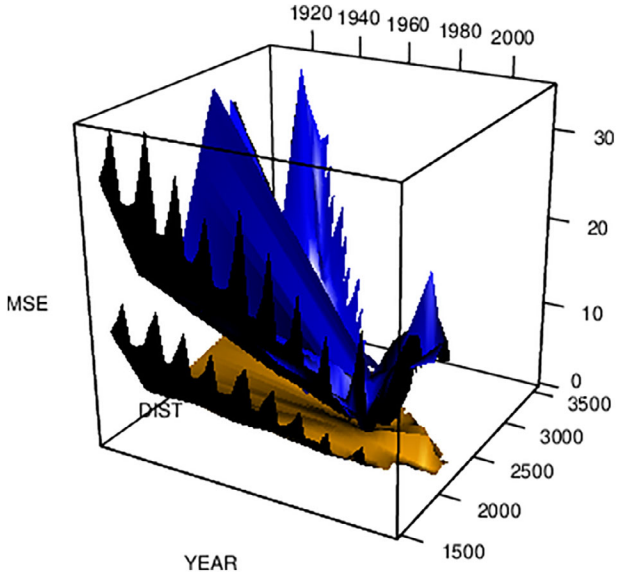

FIGURE 2 Graph of the MSEs of the two predictors in three dimensions using the `persp3d()` function in R. Away from (1960, 2500), the surface for LMs is much lower than the surface for (STK,NW)-LM.

that a fully interpretable LM has a very limited range of MSE optimality, while (STK-NW)-LM is a major improvement. The Tour de France data consists of triples (SPEED, YEAR, DISTANCE) in which SPEED is the winner's average speed over the tour for the given YEAR of the tour and the DISTANCE of the tour. This dataset has $n = 101$ and can be compiled simply by searching Wikipedia pages for the race results from 1903 to 2014; it is available from the authors on demand. Intuitively, one expects that over time the winning speed should increase as a result of improvements in athlete training, bicycle technology, and so forth. Likewise, one expects that winning speed will be a decreasing function of the distance simply because cyclists tire more as the race gets longer. These are seen in Figure 1.

We can also plot the MSEs of predictions from two function estimators using DIST and YEAR. The predictors for SPEED are LM and (STK,NW)-LM where we use the first two Legendre polynomials for DIST and YEAR as variables in the LM and NW. The predictor (STK,NW)-LM is partially interpretable since it stacks uninterpretable NWs but uses an interpretable LM. For the present, the MSEs of the two predictors are estimates of $MSE(x_i) = E(\widehat{f}_{-i}(x_i) - y_i)^2$ where $\widehat{f}_{-i}$ is the estimate of the function f from one of the two estimators using $101 - 1 = 100$ data points and dropping the ith one. We drew 100 bootstrap samples from the 100 data points and approximated $MSE(x_i)$ by the average $1/100 \sum_{j=1}^{100} (\widehat{f}_{-i,j}(x_i) - y_i)^2$ where j in an index for the bootstrap iteration.

In the notation of (1), LM would correspond to writing $M_{LM} = \{L_{11}, L_{12}, L_{21}, L_{22}, LM, \varepsilon\}$ where the explanatory variables $L_{jj'}$'s are the first two Legendre polynomials and "LM" means combining the explanatory variables in the usual LM formulation. The predictor (STK,NW)-LM would correspond to the model $M_{(\text{STK,NW})\text{-LM}}$ is $\{L_{11}, L_{12}, L_{21}, L_{22}, LM, B, NW, STK, \varepsilon\}$ where B denotes the bootstrapping used to form the NW estimators that are then stacked. Other predictors used later correspond to other models of the form (1).

Figure 2 is a perspective plot of two MSE surfaces. Each surface is a function of (YEAR, DIST). These functions are evaluated at the data points and the values smoothed. The overall lower surface shows the MSE for (STK,NW)-LM and the overall higher surface shows the MSE for LM by itself. We also generated plots like that in Figure 2 for the other four methods we define in Section 2.2. Their MSE plots were between the MSE plots of (STK,NW)-LM and LMs. So, Figure 2 merely shows the best and worst methods among these six. Note that in a region around (1960, 2500), the MSE for LM is a bit smaller than the MSE for (STK,NW)-LM. This small region where the interpretable function estimator LM is better than the partially interpretable one, may simply be where the LM is a good approximation to whatever the true model is.

In this case, we have for all practical purposes discredited the LM. That is, asking for too much interpretability of a predictor in a polynomial LM sense for the Tour de France data is unrealistic. This does not mean that every other interpretable model will fail – for instance, LMs based on other explanatory variables. However, enlarging the space of models by using a construction such as (STK, NW)-IM where IM indicates an interpretable model, remains likely

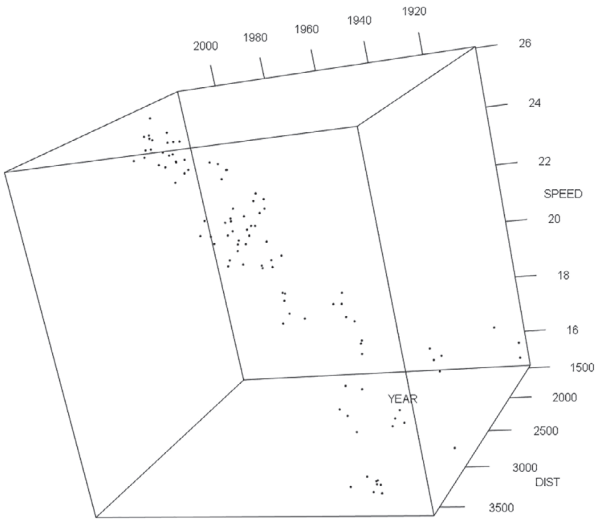

FIGURE 3 Scatter plot of the Tour de France data representing SPEED on the (most) vertical axis, i.e., as a variable dependent on YEAR and DIST.

to give an improvement because the use of (STK, NW) greatly increases the range of models available thereby permitting a reduced MSE.

Here we have not addressed the question of what components should be used to construct a model average or what form of model averaging should be used. Obviously, there are numerous choices for each. Determining which choices are best for a specific problem is problem-specific and may involve subject matter knowledge (physical modeling) as well as purely statistical knowledge; see Clarke et al [5], Clyde and Iversen [7], and Le and Clarke [14], among others.

To reinforce this point, Figure 3 that shows a scatter plot of SPEED vs (YEAR, DIST). It looks like it could be approximated by a one-dimensional curve in three dimensions. Mentally removing the outliers, one can loosely perceive a "curve" that snakes irregularly back and forth with DIST and rises with YEAR. However, the high variability and/or substantial degree of approximation in this sort of lower dimensional structure suggests that this approach will do no better than the two-dimensional approach we have taken. In fact, the waviness of the "curve" suggests a surface.

Finally, simply looking at the scatter plot does not suggest an obvious interpretable model class to use. For instance, using different functions of the explanatory variables could be more successful. Although possibly good over a larger region, non-integer powers, for instance, would have to be chosen and these could suffer the same problem as Legendre polynomials: the plot for SPEED as a function of YEAR increases to $+\infty$ for fixed DIST, and SPEED would still decrease to $-\infty$ as DIST increases for fixed YEAR. Using a function that has an asymptote in YEAR would require estimating the asymptotic value, the rate of increase, and this would still neglect the interaction between YEAR and DIST. Analogous problems would occur if more explanatory variables were used. So, a valid interpretable model seems very hard to find, and even improving the LM on which we based our estimates may require more estimates and other assumptions. That is, the domain of validity (here, given by minimal MSE) of the more interpretable model may increase as more terms are included, but that does not necessarily mean that the interpretation is actually correct; it may only be providing a better approximation to the true model, assuming one exists. (For instance, if x is an explanatory variable and we are fitting LMs, it can be unclear on small domains whether the correct term is x or $\sqrt{x}$ because the coefficients on x and $\sqrt{x}$ may make them hard to distinguish.)

3 | COMPARING MSEs AND IMSEs

In this section we present asymptotic expansions and comparisons for the MSEs and IMSEs of the predictors defined in Section 2.2. The first subsection treats the case of a univariate ($d = 1$) explanatory variable. The second subsection treats the case $d \geq 2$, a much harder setting in which many results are unavailable. The third subsection tries to provide some intuition as to what the formal results mean.

3.1 | Univariate Case, $d = 1$

In this subsection we provide the asymptotic expansions for MSE of the four predictors LM, PC, PC-LM, and (STK,PC)-LM.

Theorem 3.1. *(asymptotic expansions for MSEs). Suppose f is twice continuously differentiable, $f \in C^2[0, 1]$. Then*

1. *The asymptotic mean squared error of the LM predictor at x_{new} is*

$$MSE(\widehat{Y}_{LM}(x_{\text{new}})) = [f(x_{\text{new}}) - c_1(x_{\text{new}})]^2 + \frac{4(1 - 3x_{\text{new}} + 3x_{\text{new}}^2)\sigma^2}{n} + O\left(\frac{1}{n^2}\right),$$

where $c_1(x_{\text{new}}) \in [a,\ b]$ is defined in the proof and $a \leq f(x) \leq b$.

2. *(Gasser and Müller [9]) Suppose K has compact support and is Lipschitz-continuous on supp(K). The asymptotic mean squared error of the PC predictor at $x_{\text{new}} \in [0, 1]$ is*

$$MSE(\widehat{Y}_{\text{PC}}(x_{\text{new}})) = \frac{(\mu_2(K)f''(x_{new}))^2}{4}h_n^4 + \frac{1}{h_n}\sigma^2 S(K),$$

where $S(K) = \int K(t)^2 dt$ *and* $\mu_2(K) = \int t^2 K(t) dt$. *The optimal bandwidth* h_{opt} *decreases at the rate* $n^{-1/5}$.

3. *For the sake of completeness we note that the asymptotics for NW are almost the same as for PC. This is seen in the following, modified from Tsybakov* [31]. *The asymptotic mean squared error of the NW predictor at* x_{new} *is*

$$MSE(\widehat{Y}_{\text{NW}}(x_{\text{new}})) = \frac{h_n^4}{4}(\mu_2(K))^2 \times\left(f''(x_{\text{new}}) + 2f'(x_{\text{new}})\frac{p'(x_{\text{new}})}{p(x_{\text{new}})}\right)^2 + \frac{\sigma^2 S(K)}{nh_n}\frac{1}{p(x_{\text{new}})},$$

where $x_1, \ldots, x_n$ *have common density p(x). The optimal bandwidth* h_{opt} *decreases at rate* $n^{-1/5}$.

4. *The asymptotic mean squared error of the PC-LM predictor at* x_{new} *is*

$$MSE(\widehat{Y}_{\text{PC-LM}}(x_{\text{new}})) = [f(x_{\text{new}}) - c_1(x_{\text{new}})]^2 - \frac{2(f(x_{\text{new}}) - c_1(x_{\text{new}}))c_2(x_{\text{new}}) - c_3}{n^{1/2}} + O\left(\frac{1}{n^{3/4}}\right).$$

where $c_2(x_{new}) \in [1/2c \int t^2 K(t)dt, 1/2d \int t^2 K(t)dt]$, $c \le f''(x) \le d$, *and* c_3 *is some constant. Since the statement for PC and NW are very similar, we omit NW*.

5. *The asymptotic mean squared error of the (STK,PC)-LM predictor at* x_{new} *is*

$$MSE(\widehat{Y}_{\text{(STK,PC)-LM}}(x_{\text{new}})) = [f(x_{\text{new}}) - mc_1(x_{\text{new}})]^2 - \frac{2m(f(x_{\text{new}}) - mc_1(x_{\text{new}}))c_2(x_{\text{new}}) - c_4}{n^{1/2}} + O\left(\frac{1}{n^{3/4}}\right).$$

where c_4 *is some constant and m is a constant defined in the proof. Since the statement for PC and NW are very similar, we omit NW*.

Proof. See Appendix A. □

Remark 1. If we use the usual bandwidth $h_{\text{opt}} = O(n^{-1/5})$ as in $MSE(\widehat{Y}_{\text{PC}}(x_{\text{new}}))$, see Priestley and Chao [24], then plug this bandwidth into (A12) and (A14) in Appendix A. Then from (A11) in Appendix A, it is easy to see

$$MSE(\widehat{Y}_{\text{PC-LM}}(x_{\text{new}})) = [f(x_{\text{new}}) - c_1(x_{\text{new}})]^2 - \frac{2(f(x_{\text{new}}) - c_1(x_{\text{new}}))c_2(x_{\text{new}})}{n^{2/5}} + O\left(\frac{1}{n^{3/5}}\right).$$

This MSE is higher than the MSE in (A16) in Appendix A.

Remark 2. It is seen that our formal results are of a familiar form. For instance, the role of $c_1(x_{new})$ in Clauses 1, 4, and 5 of Theorem 3.1 represents a bias. When it appears, it does so in the leading term of the asymptotic expansions.

Now we asymptotically compare the IMSEs of the six predictors introduced in Section 2.2. The following theorem compares the IMSEs of the four predictors LM, PC, PC-LM, and STK-PC.

Theorem 3.2. *(comparing MSEs and IMSEs). Suppose K has compact support and is Lipschitz-continuous on supp(K), and* $f \in C^2[0, 1]$.

1. *If* $f(x_{\text{new}}) \ne c_1(x_{\text{new}})$, *then, as* $n \to \infty$

$$MSE(\widehat{Y}_{\text{PC}}(x_{\text{new}})) \le MSE(\widehat{Y}_{\text{LM}}(x_{\text{new}})) \le MSE(\widehat{Y}_{\text{PC-LM}}(x_{\text{new}})), \quad (10)$$

and hence

$$IMSE(\widehat{Y}_{\text{PC}}) \le IMSE(\widehat{Y}_{\text{LM}}) \le IMSE(\widehat{Y}_{\text{PC-LM}}).$$

Furthermore, if we use the predictor $\widehat{Y}_{\text{STK,PC}}$ *where the PC predictors generated by bootstrapping are orthonormalized before being stacked, then*

$$IMSE(\widehat{Y}_{\text{STK,PC}}) \le IMSE(\widehat{Y}_{\text{PC}}).$$

2. *If we use the predictor* $\widehat{Y}_{\text{STK,(PC-LM)}}$ *where the PC-LM predictors in the stacking are orthonormal, then, as* $n \to \infty$

$$IMSE(\widehat{Y}_{\text{STK,(PC-LM)}}) \le IMSE(\widehat{Y}_{\text{PC-LM}}),$$

but $IMSE(\widehat{Y}_{\text{STK,(PC-LM)}})$ *could be larger or smaller than the IMSEs of the other three predictors, namely* $IMSE(\widehat{Y}_{\text{LM}})$, $IMSE(\widehat{Y}_{\text{PC}})$, *and* $IMSE(\widehat{Y}_{\text{STK,PC}})$.

3. *As* $n \to \infty$

$$MSE(\widehat{Y}_{\text{PC}}(x_{\text{new}})) \le MSE(\widehat{Y}_{\text{(STK,PC)-LM}}(x_{\text{new}})),$$

but $MSE(\widehat{Y}_{\text{(STK,PC)-LM}}(x_{\text{new}}))$ *could be larger or smaller when compared to* $MSE(\widehat{Y}_{\text{LM}}(x_{\text{new}}))$ *or* $MSE(\widehat{Y}_{\text{PC-LM}}(x_{\text{new}}))$.

Proof: See Appendix B.

Remark 3. For $\widehat{Y}_{\text{NW}}$ we also have

$$MSE(\widehat{Y}_{\text{NW}}(x_{\text{new}})) = O\left(\frac{1}{n^{4/5}}\right),$$

the same rate as $MSE(\widehat{Y}_{\text{PC}}(x_{\text{new}}))$. So, even though $MSE(\widehat{Y}_{\text{NW}}(x_{\text{new}})) \ne MSE(\widehat{Y}_{\text{PC}}(x_{\text{new}}))$, their rates in n are the same. Therefore, if we use NW instead of PC in Theorem 3.2, the statements still hold. Moreover, from Theorem 3.2, we have

$$IMSE(\widehat{Y}_{\text{STK,PC}}) \le IMSE(\widehat{Y}_{\text{PC}}) \le IMSE(\widehat{Y}_{\text{(STK,PC)-LM}}).$$

3.2 | Multivariate case, $d \geq 2$

Our results in Section 3.1 when $d = 1$ are for fixed, equally spaced designs, and it is difficult to extend these results to the case $d \geq 2$ for fixed designs since the calculation involves sorting of the design points, which is not computationally easy in higher dimensional spaces. So, here we consider the multivariate version for random designs, thereby avoiding this problem. Thus, we can make several statements for the $d \geq 2$ case. Specifically, we provide asymptotic expansions for the MSE of the two predictors LM and PC for the case $d \geq 2$ in the following. We also state the corresponding result for NW. Note that these results are quite technical, so we have not stated all the hypotheses; we have only given a reference for where the formal results can be found.

Theorem 3.3. *(asymptotic expansions for MSEs). Let $x_1, \ldots, x_n$ have common density $p(x)$ and let $\mathcal{H}$ be the Hessian matrix of $f(x)$. For a kernel K, let $S(K) = \int K(t)^2\, dt$ be the integral of its square and let $\mu_2(K)$ be defined by the relation*

$$\int tt^T K(t)\, dt = \mu(K) I_d.$$

Then, under further mild conditions (see Liu [16]*) we have the following.*

1. *The asymptotic MSE of the multivariate LM predictor at x_{new} is*

$$MSE(\widehat{Y}_{\mathrm{LM}}(x_{\mathrm{new}})) = \frac{1}{4}(\mu_2(K_0))^2 (tr(H_{0,n}^2 \mathcal{H}(x_{\mathrm{new}})))^2 + \frac{\sigma^2 S(K_0)}{n \mid H_{0,n} \mid} \frac{1}{p(x_{\mathrm{new}})} + \mathcal{O}\left(\lambda_{\max^4}(H_{0,n}) + \frac{1}{\lambda_{\max^d}(H_{0,n})}\right),$$

where K_0 is the uniform kernel, $H_{0,n}$ is defined in the proof, and $\lambda_{\max}(H_{0,n})$ is the maximum eigenvalue of $H_{0,n}$.

2. *(Liu* [16] *Theorem 2.2) The asymptotic MSE of the multivariate PC predictor at x_{new} is*

$$MSE(\widehat{Y}_{PC}(x_{\mathrm{new}})) = \frac{1}{4}(\mu_2(K))^2 (tr(H_n^2 \mathcal{H}(x_{\mathrm{new}})))^2 + \frac{\sigma^2 S(K)}{n \mid H_n \mid / b} + o(tr(H_n^2)^2) + o\left(\frac{1}{n \mid H_n \mid}\right),$$

where b is the volume of the design region $\Omega \subseteq \mathbb{R}^d$. The optimal bandwidth is

$$H_{\mathrm{opt}} = \mathcal{O}\left(\left(\frac{\sigma^2 S(K) b |\widetilde{\mathcal{H}}(x_{\mathrm{new}})|^{1/2}}{n d \mu_2^2(K)}\right)^{1/(d+4)} \widetilde{\mathcal{H}}(x_{\mathrm{new}})^{-1/2}\right),$$

where $\widetilde{\mathcal{H}}(x) = \mathcal{H}(x)$ if $\mathcal{H}(x)$ is positive definite and $\widetilde{\mathcal{H}}(x) = -\mathcal{H}(x)$ if $\mathcal{H}(x)$ is negative definite.

3. *(Liu* [16] *Theorem 2.3) The asymptotic MSE of the multivariate NW predictor at x_{new} is*

$$MSE(\widehat{Y}_{\mathrm{NW}}(x_{\mathrm{new}})) = \left(\frac{1}{2}\mu_2(K) tr(H_n^2 \mathcal{H}(x_{\mathrm{new}})) + \frac{\mu_2(K)}{p(x)} \nabla_f'(x_{\mathrm{new}}) H_n^2 \nabla_p(x_{\mathrm{new}})\right)^2 + \frac{\sigma^2 S(K)}{n \mid H_n \mid} \frac{1}{p(x_{\mathrm{new}})} + o(tr(H_n^2)^2) + o\left(\frac{1}{n \mid H_n \mid}\right),$$

where ∇_f and ∇_p are the gradient vectors of $f(x)$ and $p(x)$, resp. The optimal bandwidth is

$$H_{\mathrm{opt}} = \mathcal{O}\left(\left(\frac{\sigma^2 S(K) |\widetilde{\mathcal{H}}_{\mathrm{NW}}(x_{\mathrm{new}})|^{1/2}}{n d \mu_2^2(K) p(x_{\mathrm{new}})}\right)^{1/(d+4)} \widetilde{\mathcal{H}}_{\mathrm{NW}}(x)^{-1/2}\right),$$

where $\mathcal{H}_{\mathrm{NW}}(x) = \mathcal{H}(x) + (\nabla_p(x)\nabla_f'(x) + \nabla_f(x)\nabla_p'(x))/p(x)$, $\widetilde{\mathcal{H}}_{\mathrm{NW}}(x) = \mathcal{H}_{\mathrm{NW}}(x)$ *if* $\mathcal{H}_{\mathrm{NW}}(x)$ *is positive definite and* $\widetilde{\mathcal{H}}_{\mathrm{NW}}(x) = -\mathcal{H}_{\mathrm{NW}}(x)$ *if* $\mathcal{H}_{\mathrm{NW}}(x)$ *is negative definite.*

Proof. We only give a proof of the Clause 1 (for LM) since the latter two can be found in Liu [16]. Let the kernel K_0 be the uniform kernel and select the bandwidth matrix $H_{0,n}$ such that all data points $(x_1, y_1), \ldots, (x_n, y_n)$ contribute to the least-squares minimization

$$\min_{\beta_0, \beta} \sum_{i=1}^{n} \{Y_i - \beta_0 - \beta'(x_i - x)\}^2 K_{H_0}(x_i - x),$$

where $\beta = (\beta_1, \ldots, \beta_d)'$. Then the corresponding local linear predictor under K_{H_0}, say $\widehat{\beta}$, becomes the usual LM predictor. With this specification of K_0 and $H_{0,n}$ Theorem 2.1 on page 34 Liu [16] gives the asymptotic MSE of the LM predictor as

$$MSE(\widehat{Y}_{\mathrm{LM}}(x_{\mathrm{new}})) = \frac{1}{4}(\mu_2(K_0))^2 (tr(H_{0,n}^2 \mathcal{H}(x_{\mathrm{new}})))^2 + \frac{\sigma^2 S(K_0)}{n \mid H_{0,n} \mid} \frac{1}{p(x_{\mathrm{new}})} + O\left(\lambda_{\max}^4(H_{0,n}) + \frac{1}{\lambda_{\max}^d(H_{0,n})}\right).$$

Remark 1. Observe that the asymptotics for NW are almost the same as for PC. This is reasonable since NW is the random design version of PC. Indeed, the error terms for the MSE in Clause 1 are also very similar to the error terms for the MSE in Clauses 2 and 3 by an argument in Liu [16]. Moreover, if some other linear nonparametric estimator, e.g., k-nearest neighbors for some reasonable k, had been used in place of the PC (or NW) estimators, analogous results would be possible, see Stone [30] or Hardle [10], Chap. 4. We suggest the asymptotic MSEs for spline smoothers would also be similar.

Remark 2. It is easy to see from Theorem 3.3 that, as in Clause 1 of Theorem 3.2,

$$MSE(\widehat{Y}_{\mathrm{PC}}(x_{\mathrm{new}})) \leq MSE(\widehat{Y}_{\mathrm{LM}}(x_{\mathrm{new}})) \quad \text{as} \quad n \to \infty$$

typically, because of the $\mathcal{O}(1/n)$ terms and the big-O and little-o error terms.

The limitation of the results in Liu [16] is that they only apply to the predictors from linear estimators and therefore we do not have statements for $(PC-LM)$, (STK, PC), $(STK, PC)-LM$, and $STK-(PC, LM)$. We conjecture that Clause 1 of Theorem 3.3 can be combined with either Clause 2 or Clause 3 to give an asymptotic forms for the (pointwise) MSEs of (PC, LM) and (NW, LM). However, the results seem too complicated to give clean, useful comparative statements. As to the other three predictors, (STK, PC), $(STK, PC)-LM$, and $STK-(PC, LM)$, the nonlinearity of STK seems to make it impossible to apply existing results to obtain asymptotic forms for their MSEs.

3.3 | Intuition suggested by the formal results

First, consider the asymptotic expressions in Theorem 3.1. It is seen that the parts of the expressions arising from the variance terms are nondecreasing. Unsurprisingly, the ordering of the methods (in terms of asymptotic variance) is LM, PC or NW, PC-LM, and (STK,PC)-LM. That is, the least interpretable method NW has an asymptotic variance in the middle, while more interpretable methods have smaller variances and partially interpretable methods have larger variances. Unfortunately, this ordering neglects bias terms that depend on an unknown function $c_1(x_{\text{new}})$ when the method is not consistent.

Now, consider the inequalities in Theorem 3.2. It is well known that $\widehat{Y}_{\text{PC}}$ smoothes the data and $\widehat{Y}_{\text{LM}}$ linearizes the data, an extreme form of smoothness and hence a stronger interpretation. So $MSE(\widehat{Y}_{\text{PC}}(x_{\text{new}})) \leq MSE(\widehat{Y}_{\text{LM}}(x_{\text{new}}))$ i.e., Clause 1, makes sense. What do we make of

$$MSE(\widehat{Y}_{\text{LM}}(x_{\text{new}})) \leq MSE(\widehat{Y}_{\text{PC-LM}}(x_{\text{new}}))?$$

Obviously, $MSE(\widehat{Y}_{\text{PC-LM}}(x_{\text{new}}))$ is larger because not only has the data been smoothed but it is the smoothed data that is linearized, i.e., there is more "processing" of the data. The fact that $\widehat{Y}_{\text{PC-LM}}$ imposes more processing on the data than $\widehat{Y}_{\text{LM}}$ does suggests that $\widehat{Y}_{\text{PC-LM}}$ also imposes a stronger interpretation on the data. A similar argument can be made for the IMSE. Also, in Clause 1 of Theorem 3.2, $IMSE(\widehat{Y}_{\text{STK,PC}}) \leq IMSE(\widehat{Y}_{\text{PC}})$ makes sense because using the bootstrap to generate $\widehat{Y}_{\text{PC}}$'s to stack undercuts the smoothing in $\widehat{Y}_{\text{PC}}$, thus weakening the interpretability of $\widehat{Y}_{\text{STK,PC}}$ relative to $\widehat{Y}_{\text{PC}}$.

Clause 2 of Theorem 3.2, i.e., $IMSE(\widehat{Y}_{STK,(PC-LM)}) \leq IMSE(\widehat{Y}_{PC-LM})$, makes sense because of the orthonormality. The result is that stacking is essentially optimal because it is combining $(PC-LM)$s that are uncorrelated, and that is the best information to combine.

In Clause 3 of Theorem 3.2, in which the order of operations is reversed relative to Clause 2, the inequality is reversed, i.e., $\widehat{Y}_{PC}$ has lower MSE than $\widehat{Y}_{(STK,PC)-LM}$. At first this seems counterintuitive. It is possible that the linearization in $(STK, PC)-LM$ tends to increase the MSE more than $(STK-PC)$ is able to decrease it, cf. Clauses 1 and 2.

The basic principle is that there is a relationship between processing data (which often de facto forces an interpretation on it) and improving prediction which usually follows from enlarging the collection of predictors under consideration, at least in an asymptotic sense.

Naively, Clause 3 suggests that $(STK, PC)-LM$ should be among the worst in MSE, whereas our computed examples suggest that it is often the best. However, a closer look shows that $(STK, PC)-LM$ is not among the worst according to our theorems. Recall, we showed (i) $IMSE(STK, PC) < IMSE(PC) < IMSE(LM) < IMSE(PC,LM)$, (ii) $MSE(PC) < MSE((STK, PC)-LM)$, and (iii) $MSE((STK, PC)-LM)$ could be larger or smaller than $MSE(LM)$. Taken together, this just means that $(STK, PC)-LM$ is only worse than $STK-PC$ and PC, perhaps because of the extra data processing. Its better performance on finite sample size data may result from $(STK, PC)-LM$ being able to extract linear functions of the data via LM that can be fed into a flexible nonparametric method PC which becomes more flexible by the use of STK.

4 | COMPUTED EXAMPLES

Our first example in Section 4.1 uses a dataset that is representative of a large class of agronomic datasets. It has $n = 2912$, but to match the sample size in Section 2.3, we drew 100 data points at random. (This is not strictly necessary because the datasets are different. However, it can only improve the comparability of our discussion of the two examples.) Even with this smaller n, the number of explanatory variables made the computational running time long. For the sake of making a point about lower dimensional substructures in data, we only compare two models. (This was not possible with the Tour de France data because we only noted a very weak substructure – the wavy line of points in Figure 3.)

Our second example in Section 4.2 uses a dataset, Online News Popularity, that is qualitatively different from those used in Sections 4.1 and 2.3. The sample size is 39 797, but again we randomly selected $n = 100$ data points. The results of the analysis are qualitatively similar to those in Section 2.3. However, with Tour de France the central

tendency of the data points are of interest in characterizing SPEED, whereas with Online News Popularity interest includes, indeed may center on, the outliers since they are the news stories that are "shared" most.

4.1 | Wheat data

In this subsection, the behavior of the MSEs is examined on a real agronomic dataset collected in Nebraska from 1999 to 2001, see Campbell et al. [2] for the original presentation of the data and a standard analysis. See Dhungana et al. [8] and Xiaojuan et al. [35] for more elaborate analyses based on structural equation models. The data comes from a randomized blocks with repetition designed experiment to evaluate which varieties of wheat give the best yield (YLD) under various conditions. All varieties are highly inbred so the genetic difference from variety to variety is small. There are actually 36 variables in addition to YLD, and since NW does not effectively scale up to many explanatory variables, we had to do severe variable selection. Of the 36 variables, 10 are related to the design of the experiment and 19 are single-nucleotide polymorphism (SNP) data. We ignore this portion of the data since its effect is small. This leaves seven variables. We also ignore the date of planting variable since it has too many missing values and we ignore the average height (HT) of the plants in a region because its relation to YLD is ambiguous. Two other variables were testweight (the weight of a specified volume of wheat) and thousand kernel weight (TKWT), which were seen as being very similar, so only TKWT was retained. Likewise, kernels per square meter is essentially the product of kernels per stalk (KPS) with stalks per square meter (SPSM) and was therefore dropped. Finally, the models we fit were

$$YLD \sim TKWT + KPS \cdot SPSM \tag{11}$$

$$YLD \sim TKWT + (KPS + SPSM), \tag{12}$$

and we used the same six techniques as before, i.e., we used the first two Legendre polynomials for each term in (11) and (12) as our explanatory variables. We comment that the model $YLD \sim TKWT * KPS * SPSM$ is actually the model one would propose on highly oversimplified physical grounds; we have used (11) and (12) because, as will be seen, the lower dimensional structure is two-dimensional and easily visualized. In fact, it is known that SNPs, at least in aggregate, affect YLD, so our models are not correct to state-of-the-art experimental precision let alone infinite precision. Our models are, however, close enough to a correct model to make our points.

Let us start with some scatterplots. Figure 4 shows the scatterplots of YLD versus the three explanatory variables we have retained. None of panels (a), (b), or (c) shows strong patterns, but the first two suggest YLD is increasing with TKWT and SPSM; this is not a surprise since both explanatory variables have an obvious physical relationship to YLD. Figure 4c physically suggests that there is little relationship between KPS and YLD. This is surprising since one expects YLD to increase with KPS. Thus, a superficial graphical analysis suggests that only TKWT and SPSM are important to YLD.

Next, we compare the six regression methods using the models in (11) and (12). In both cases, Figure 5 shows the perspective plots of the best and worst MSE surfaces. The best method is (STK, NW)-LM – the same as for the Tour de France data. The second best method is STK-NW. The worst method was LMs – again the same as for the Tour de France (and other datasets not shown here). The other three methods are essentially indistinguishable. This is qualitatively the same as was seen in Figure 2.

Figure 5 also suggests that model (11) is a little better than the model (12) because the surfaces for (STK, NW)-LM and LM for (12) are higher than the corresponding surfaces for (11), respectively. However, in both panels there is a (small) region on which LMs outperform (STK, NW)-LM, as was the case for the Tour de France data. This suggests that model (11) with (STK, NW)-LM is preferred in an MSE sense even though it is not as interpretable as LMs. The issue is how large the region is on which LM outperforms (STK, NW)-LM and the amount by which it does. This is a physical relevance question, not one that can be answered by statistical analysis alone.

Why does (STK, NW)-LM with (11) perform overall better than LM with (11) or (12) or (STK, NW)-LM with (12)? To investigate this, let us assess the interpretability of KPS*SPSM versus KPS + SPSM. After all, the explanatory variables are physically meaningful and LMs, whatever their other failings, are interpretable and have convergence rate $O(1/n)$ rather than $O(1/n^{1/2})$ for (STK, NW)-LMs. One answer is bias. However, the real answer is where that bias comes from.

Figure 6 shows a scatter plot of YLD as a function of TKWT and $KPS \cdot SPSM$. It is seen that the dots form a sheet that is roughly triangular so that, as the product $KPS \cdot SPSM$ increases, the triangle becomes more pointed at the top. The fact that the structure is almost a plane means that model (11) is taking advantage of lower dimensional structure that (12) is missing. Indeed, KPS*SPSM is more or less the number of kernels per meter squared, a natural, if coarse, measure of yield. To the extent that the structure is not a plane – in fact it bows outward from the page – it means that we have left out important explanatory variables such as SNPs, varieties, and the various design variables. Despite this, it is the modeling information represented by the product term in (11) that gives an improvement in MSE for (STK,NW)-LM even though for

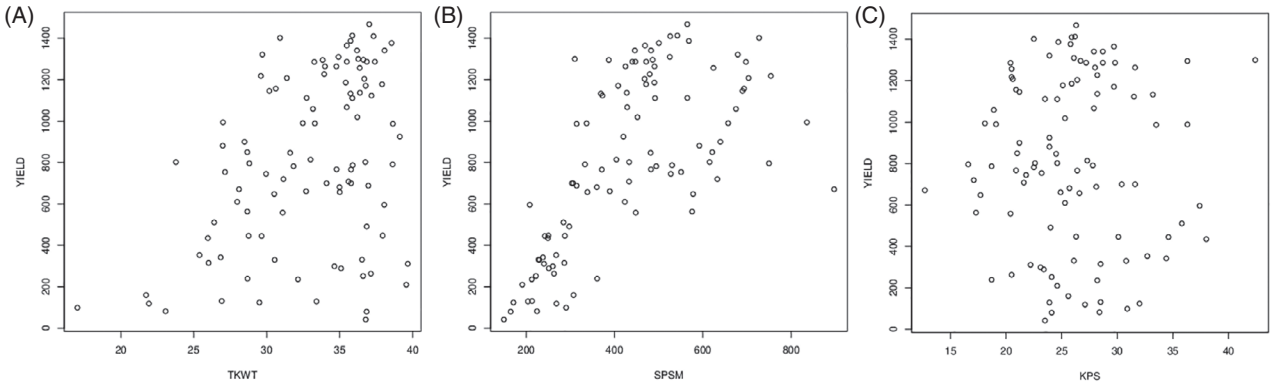

FIGURE 4 From left to right: Scatter plots of YIELD vs. TKWT, YIELD vs. SPSM, and YIELD vs. KPS for the Wheat data.

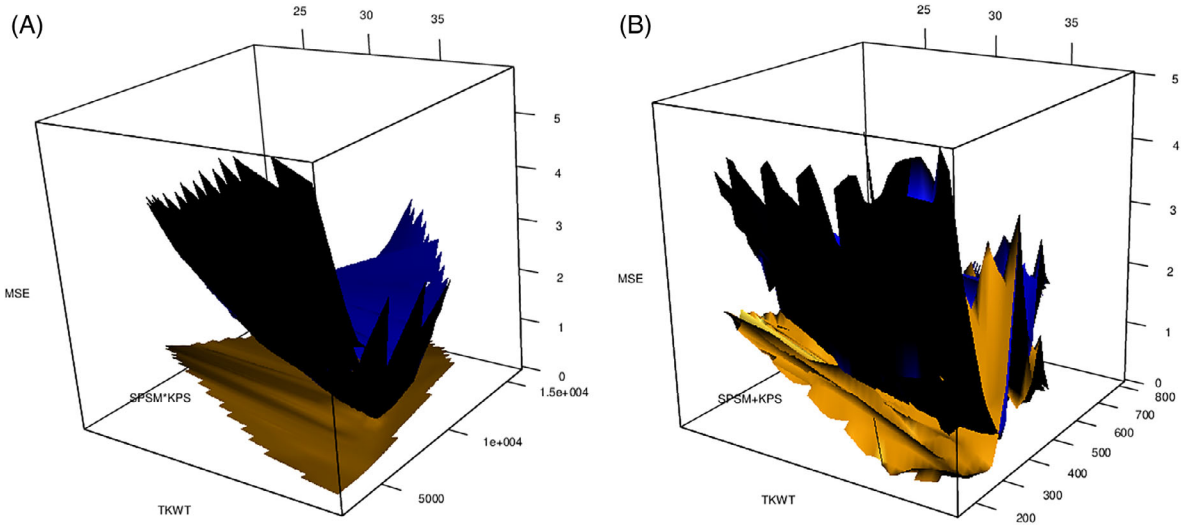

FIGURE 5 MSEs of (STK,NW)-LM (lower sheets) and LM (upper sheets) for models (11) and (12).

both (11) and (12) LMs are poor. The implication is that LMs should not be used by themselves and that uninterpretable methods like (STK, NW)-LM may give better performance than other methods when they take advantage of valid substructure.

Since our general point is that the right level of partial interpretability results in the best MSE, why do not the results for the Wheat data contradict this? For instance, it is reasonable to argue that (STK, NW)-LM is more interpretable than STK-NW. The answer is again the presence of low-dimensional structure in the data, as seen in Figure 6. The model (11) does not capture this low-dimensional structure perfectly, but the low-dimensional structure is strong enough that (STK,NW)-LM can outperform STK-NW slightly. That is, the interpretability of KPS*SPSM as an important term for yield makes model (11) preferable to (12) and Figure 6 effectively validates (11). So the question is: Do LMs with (11) have enough interpretability to make up for their worse MSE performance compared to (STK, NW)-LM? This is a physical question, not purely statistical.

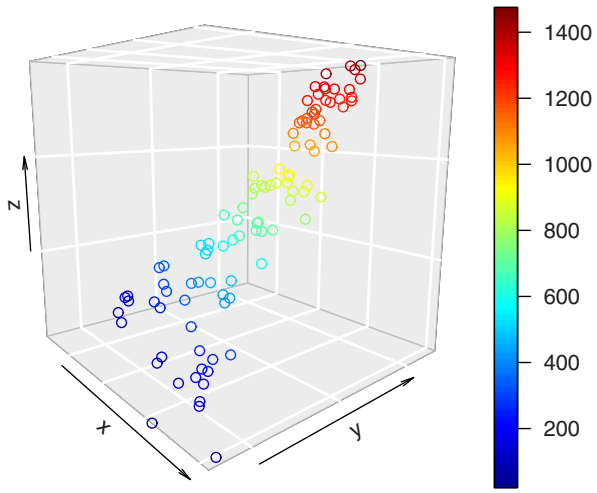

FIGURE 6 Scatter plot of YIELD vs. TKWT and KPS*SPSM.

We comment that in other examples not shown here, STK-NW sometimes does best, e.g., on the mid-range of the Tour de France data in Section 2.3 where there was enough regularity in the data that partially interpretable

methods could detect structure yet not be tied too closely to it.

In the absence of detectable structure, we can combine Clauses 1 and 3 of Theorem 3.2 to obtain

$$IMSE(\hat{Y}_{\text{STK,NW}}) \leq IMSE(\hat{Y}_{\text{(STK,NW)-LM}}), \tag{13}$$

as $n \to \infty$, in general, in terms of IMSEs not MSEs. However, for the Wheat data, Figure 5 shows that (STK,NW)-LM with model (11) is best while STK-NW is not as good. Again, the presence of substructure, i.e., something we might interpret, reverses the general inequalities in Section 3.

In the most extreme case, the difference is seen from Theorems 3.1, Clauses 1 and 5, that LMs have smaller variance than (STK,NW)-LM, so their poor performance comes from the bias. The bias comes from the fact that the model (12) encapsulates very little of the planar structure seen in Figure 6. By contrast, STK-NW picks up so much of the low-dimensional structure that applying LM afterward is helpful because it strengthens the planar interpretation built into the STK-NW part of the predictor.

4.2 | Online News Popularity

As a second example, consider the Online News Popularity dataset publicly available from the UC Irvine Machine Learning Repository. There are 58 nontrivial explanatory variables related to the dependent variable, namely the number of shares in social networks (popularity). For simplicity, we chose the two explanatory variables that had the largest correlations to the response variable *shares*: the "maximum of the average keyword shares" (*kw_max_avg*) and the "average of the average keyword shares" (*kw_avg_avg*); see http://archive.ics.uci.edu/ml/datasets/Online+News+Popularity for details and references. In Le and Clarke [13], these two variables were enough to distinguish the performance of predictors. We will see again that a fully interpretable LM has a limited range of validity, while (STK,NW)-LM is almost always an improvement.

Let us start with some scatterplots. Figure 7 shows the plot of *shares* as a function of *kw_max_avg* and *kw_avg_avg* zeroing in on the main data cloud, i.e., omitting outliers. Figure 8 shows the univariate scatterplots, analogous to two of the plots in Figure 4. No obvious patterns or substructures can be seen in Figure 7 or 8.

We compare the six regression methods and, as before, use the first two Legendre polynomials for *kw_max_avg* and *kw_avg_avg* as our explanatory variables. The best method is (STK, NW)-LM (the same as for the Tour de France and Wheat data), and, interestingly, the second best method is LM, while the other four methods are

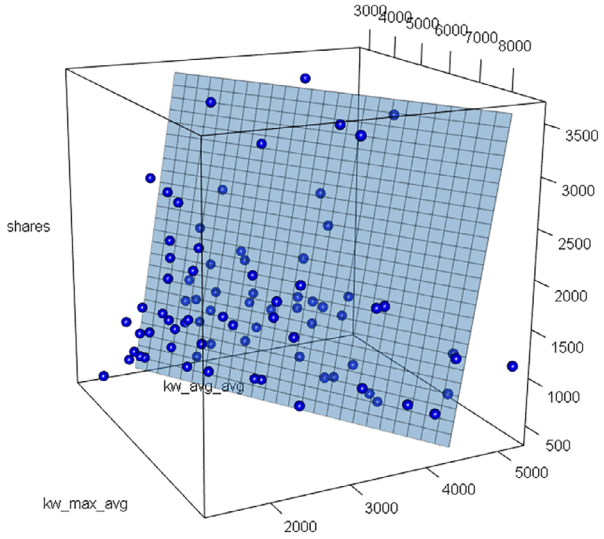

FIGURE 7 Scatter plot of *shares* vs. *kw_max_avg* and *kw_avg_avg* with outliers removed. A regression plane for *shares* as a function of *kw_max_avg* and *kw_avg_avg* is shown to help visualize the data.

essentially indistinguishable and worse than LM. Figure 9 shows a perspective plot of MSE surfaces corresponding to the two best methods (STK, NW)-LM and LM, i.e., the MSE is approximated by a function at each value of (*kw_max_avg*, *kw_avg_avg*). The overall lower surface (the yellow one) shows the MSE for (STK,NW)-LM and the overall upper surface (the blue one) shows the higher MSE for LM by itself. In this case, we have, for all practical purposes, discredited the LM. This shows that seeking too much interpretability via this LM-based predictor for the Online News Popularity data is suboptimal.

Although impossible to see in the perspective used in Figure 9, there is a small region in which the MSE for the LM predictor is a little smaller than for the (STK, PC)-LM predictor, i.e., a region where the interpretable estimator is a little bit better than the partially interpretable one. As with the Tour de France data in Section 2.3, we regard this as the region where LM provides a good approximation to whatever the true model is. Note also that in this sort of dataset, the outliers provide the test for whether a predictor is good. The spike in the blue surface of Figure 9 represents an outlier that the LM predictor could not predict well.

5 | DISCUSSION

The central point of this paper is that enlarging an interpretable model to include uninterpretable deviations from the model class frequently leads to predictors that are better than those simply based on the putative model. We have shown this through a series of examples and argued

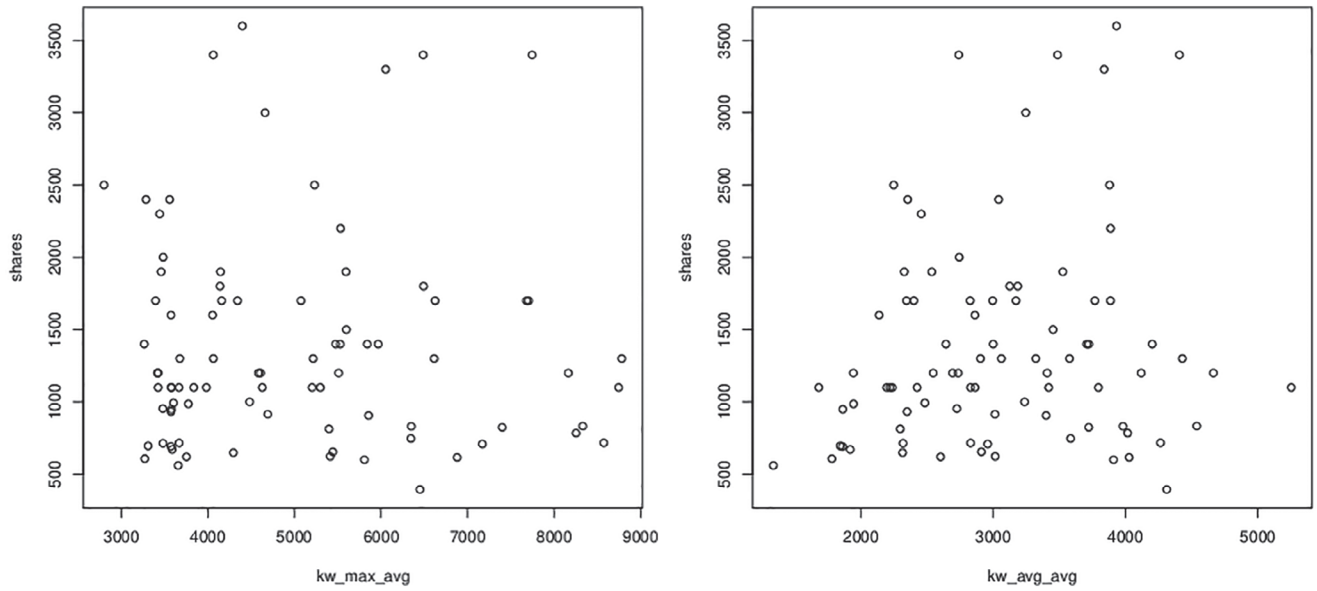

FIGURE 8 *Left*: Scatterplot of *shares* vs. *kw_max_avg*. *Right*: Scatterplot of *shares* vs. *kw_avg_avg*.

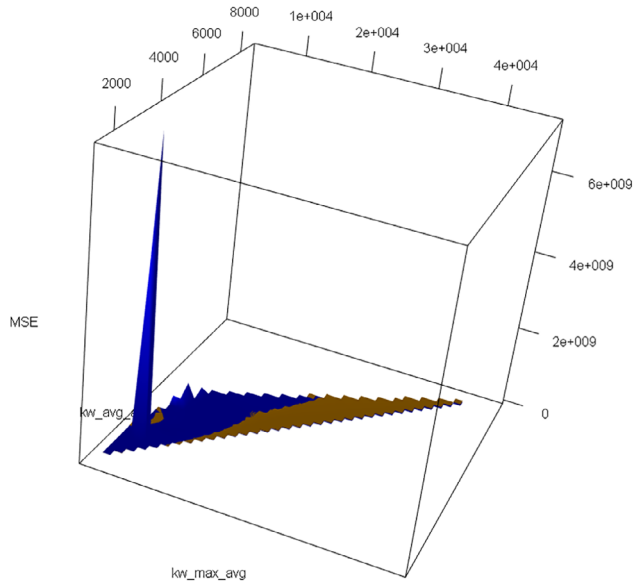

FIGURE 9 Graph of the MSEs of the two predictors in three dimensions using the `persp3d()` function in R. Away from the outliers, the surface for (STK,NW)-LM is lower than the surface for LM.

more generally, see Section 3.3, that this is not an anomaly. Indeed, the asymptotic MSE can increase or decrease as a result of data processing, i.e., using additional techniques, model-based or not, for generating predictions. For this reason we argue that, under appropriate caveats, there is a sort of convex function representing the tradeoff between interpretability and MSE: Asking for high interpretability leads to an elevated MSE. Permitting essentially no interpretability likewise leads to elevated MSE. It is on the intermediate range of interpretability that MSE is often the smallest. The caveats on this intuition include that the system under study be sufficiently complex, the modeling be not too misleading, and the sample size be in the right range relative to the complexity of the predictors and data generator, among others.

Indeed, another example of this phenomenon can be found in Weng et al. [33]. They compared Cox models (highly interpretable) with both a deep learning technique and random forests (not as interpretable) for predicting human lifespan. They found that the less interpretable methods were better by a substantial margin. The weaknesses of the Cox models for predictive purposes have been known to specialists for some years, see Henderson [11] and Henderson and Keiding [12] among others, but they are not widely appreciated. So, with hindsight, its poorer performance was to be expected.

There are methodological implications of this point of view. First, even though it is not always possible, one pragmatic test for which of a collection of predictive strategies is most successful is to ensure that more and less interpretable strategies give higher MSEs than a blended strategy. A special case of this is to check whether the six methods used here are implemented for a particular problem, and whether their MSEs are the reverse of any of the inequalities shown in Section 3. In such cases, one is led to suspect that something more is going on: perhaps lower dimensional substructures or highly accurate or inaccurate modeling. Essentially, the possibility of useful interpretability, or more precisely, the availability of more pre-experimental information, is indicated anytime the theoretical inequalities in Section 3 fail to hold.

Of course, if there is a large gap between the two sides of one of our inequalities, the noninterpretable parts of the predictor may be more effective than the interpretable parts.

Our proposal is an alternative to choosing terms for a physically plausible model based on MSE, e.g., adding terms assumed to be important until MSE begins to increase. The point is to choose plausible model ingredients based strictly on knowledge at hand and not going beyond what is genuinely known. Then, rather than trying to make this initial model perform better using further physically interpretable terms, admit that this is likely to be ineffective and therefore incorporate statistically useful approaches to prediction ignoring whether they have physical correlates. Indeed, the absence of physical correlates may be regarded as a plus: if detailed modeling is infeasible, a statistical method of prediction that ignores physical considerations in favor of improved prediction may well be the better approach. In this case, finding a passable physical model that performs middling well without assuming dubious real-world information can be augmented by nonparametric model averaging or other techniques that lead to good prediction rather than physical meaning. In a sense, this is akin to semi-parametrics where part of the "model" is taken as real and the rest is purely statistical.

The stance taken here is that good prediction trumps physical modeling and should be the key goal, at least initially, rather than defining and solving a physical research/modeling problem. This actually follows from the falsification paradigm mentioned in Section 1 in the sense that we are using predictors rather than models to achieve good prediction. The novelty is the observation that modeling per se is often not an effective way to achieve good prediction unless sample size, data complexity, and other criteria are met. These points seem particularly strong when no true model can be assumed or, if it exists, is too complex to be useful. This is often the case with streaming data.

Another implication of our stance is that it is important to be much more tentative about model selection, regarding model uncertainty and mis-specification as frequently the dominant factors in statistical analysis. Models are often less useful than imagined, and relying on them, as is commonly done, can often be little more than a confidence game outside of the simplest "toy" examples. For instance, it is one thing to say X affects Y; this is often noncontroversial. It is another thing to say Y depends on X through βX when $\sqrt{X}$ or $X^{1+\alpha}$ is just as reasonable. Arguing for βX is fine if the model is used for data summarization, but it is important to distinguish between the use of a model for data summarization and its use as a model for reality.

We conclude with two points. First, the difference between model selection and the kind of ensembling we are advocating is intended to stabilize predictions from models that are not too far wrong at the same time as providing a wider range of predictors from which to select. Second, we should regard a predictor as *validated* if it has a satisfactorily low MSE (or IMSE) and provides at least a partial interpretation, e.g., via LMs, or other functions that can be precisely expressed, and can be convincingly exhibited in a form such as (1). This can in principle be tested on future data. This is a stronger criterion than more conventional notions of validation which often just rely on a model as not having been easily discredited. It is an effort to propose a definition that will apply to prediction with complex data where model uncertainty or model mis-specification is nontrivial. Finding the best tradeoff between high interpretability and low predictive error by optimally enlarging the predictor class, given the sample size, is the main task of statistical prediction.

ACKNOWLEDGMENTS

This study was supported in part by the NSF grant DMS-1419754.

ORCID

Tri Le https://orcid.org/0000-0001-8454-0220

REFERENCES

1. L. Breiman, *Stacked regressions*, Mach. Learn. 24 (1996), 49–64.
2. B. T. Campbell et al., *Identification of qtls and environmental interactions associated with agronomic traits on chromosome 3a of wheat*, Crop Sci. 43 (2003), 1493–1505.
3. B. Clarke, *Bayes model averaging and stacking when model approximation error cannot be ignored*, J. Mach. Learn. Res. 4 (2003), 683–712.
4. Clarke, B. and J. Clarke, 2018: *Predictive statistics*, volume 46 of *Cambridge series in statistical and probabilistic mathematics*. Cambridge University Press, Cambridge, UK.
5. J. Clarke, C. W. Yu, and B. Clarke, *Prediction in M-complete problems with limited sample size*, Bayesian Analysis 8 (2017), 647–690.
6. J. L. Clarke, B. Clarke, and C.-W. Yu, *Prediction in M-complete problems with limited sample size*, Bayesian Anal. 8 (2013), 647–690.
7. M. Clyde and E. Iversen, *Bayesian model averaging in the M-open framework*, in *Bayesian theory and applications*, P. Damien et al., Eds., Oxford University Press, Oxford, 2013, 484–498.
8. P. Dhungana et al., *Analysis of genotype-by-environment interaction in wheat using a structural equation model and chromosome substitution lines*, J. Crop Sci. 47 (2007), 477–484.
9. T. Gasser and H.-G. Müller, *Estimating regression functions and their derivatives by the kernel method*, Scand. J. Statist. 11 (1984), 171–185.
10. W. Härdle, *Applied nonparametric regression*, Humboldt-Universität zu Berlin, Berlin, 1994.

11. R. Henderson, *Problems and prediction in survival data analysis*, Stat. Med. 14 (1995), 161–184.
12. R. Henderson and N. Keiding, *Individual survival time prediction using statistical models*, J. Med. Ethics 31 (2005), 703–706.
13. T. Le and B. Clarke, *Using the Bayesian Shtarkov solution for predictions*, Comp. Stat. Data Anal. 104 (2016), 183–196.
14. T. Le and B. Clarke, *A Bayes interpretation of stacking for M-complete and M-open settings*, Bayesian Anal. 12 (2017), 807–829.
15. T. Le and B. Clarke, *On the interpretation of ensemble classifiers in terms of bayes classifiers*, J. Classification 35 (2018), 1–32.
16. X.-H. Liu, Kernel smoothing for spatially correlated data. Ph.D., Iowa State University, Ames, IA, 2001.
17. J. Mays, Model robust regression: combining parametric, nonparametric and semiparametric methods. Ph.D. Thesis, Virginia Polytechnic Institute, 1995.
18. J. Mays and J. Birch, *Smoothing for small samples with model misspecification: Nonparametric and semiparametric concerns*, J. Appl. Statist. 29 (2002), 1023–1045.
19. J. Mays, J. Birch, and R. Einsporn, *An overview of model-robust regression*, J. Statist. Comput. Simulation 66 (2000), 79–100.
20. J. Mays, J. Birch, and A. Starnes, *Model robust regression: Combining parametric, nonparametric and semiparametric methods*, J. Nonparametr. Statist. 13 (2001), 245–277.
21. M. Milkowski, W. Hensel, and M. Hotol, *Replicability or reproducibility? on the replication crisis in computational neuroscience and sharing only relevant detail*, J. Comp. Neuro. 45 (2018), 163–172.
22. E. A. Nadaraya, *On estimating regression*, Theory Probab. Appl. 9 (1964), 141–142.
23. M. Ozay, F.T.Y. Vural, A new fuzzy stacked generalization technique and analysis of its performance. *arXiv:1204.0171*, 2012.
24. M. B. Priestley and M. T. Chao, *Nonparametric function fitting*, J. R. Stat. Soc. Ser. B 34 (1972), 385–392.
25. A. Raftery and Y. Zheng, *Performance of Bayesian model averaging*, J. Amer. Statist. Assoc. 98 (2003), 931–938.
26. D. F. Ransohoff, *Bias as a threat to the validity of cancer molecular-marker research*, Nat. Rev. Cancer 5 (2005), 142–149.
27. J. Sill, G. Takacs, L. Mackey, and D. Lin, *Feature-weighted linear stacking*. 2009, available at arxiv.org/pdf/0911.0460.
28. K. Skouras and P. Dawid, *On efficient point prediction systems*, J. R. Stat. Soc. Ser. B 60 (2002), 765–780.
29. P. Smyth and D. Wolpert, *Linearly combining density estimators via stacking*, Mach. Learn. 36 (1999), 59–83.
30. C. Stone, *Optimal global rates of convergence for nonparametric regression*, Ann. Statist. 10 (1982), 1040–1053.
31. A. B. Tsybakov, *Introduction to nonparametric estimation*, Springer Series in Statistics, New York, 2009.
32. G. S. Watson, *Smooth regression analysis*, Ind. J. Statist., Ser. A 26 (1964), 359–372.
33. S. Weng et al., *Prediction of premature all-cause mortality: A prospective general population cohort study comparing machine-learning and standard epidemiological approaches*, PLoS One 14 (2019), 1–22.
34. D. Wolpert, *Stacked generalization*, Neur. Netw. 5 (1992), 241–259.
35. X. Mi, K. Eskridge, D. Wang, et al., *Regression-Based Multi-Trait QTL Mapping Using a Structural Equation Model*. Statistical Applications in Genetics and Molecular Biology, 9(1) (2010), doi:10.2202/1544-6115.1552.

How to cite this article: Le T, Clarke B. In praise of partially interpretable predictors. *Stat Anal Data Min: The ASA Data Sci Journal*. 2020;13:113–133. https://doi.org/10.1002/sam.11450

APPENDIX A. PROOF OF THEOREM 3.1

A.1 Asymptotic MSE of LM

We have

$$MSE(\widehat{Y}_{\mathrm{LM}}(x_{\mathrm{new}})) = Var(\widehat{Y}_{\mathrm{LM}}(x_{\mathrm{new}})) + [f(x_{\mathrm{new}}) - E(\widehat{Y}_{\mathrm{LM}}(x_{\mathrm{new}}))]^2, \tag{A1}$$

where $Y_i = f(x_i) + \varepsilon_i, i = 1, \ldots, n, \varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. $(0, \sigma^2)$, and the design points are equidistant in [0, 1], i.e.,

$$x_i = \frac{i-1}{n-1}, \quad i = 1, \ldots, n. \tag{A2}$$

Since $\widehat{Y}_{\mathrm{LM}}(x_{\mathrm{new}}) = x'_{\mathrm{new}}\widehat{\beta}_{\mathrm{LM}} = x'_{\mathrm{new}}(X'X)^{-1}X'Y$ where $Y = (Y_1, \ldots, Y_n)'$ and

$$X' = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ x_1 & x_2 & \ldots & x_n \end{pmatrix}, \tag{A3}$$

we have

$$\begin{aligned} E(\widehat{Y}_{\mathrm{LM}}(x_{\mathrm{new}})) &= x'_{\mathrm{new}}(X'X)^{-1}X'E(Y) \\ &= x'_{\mathrm{new}}(X'X)^{-1}X'(f(x_1) \ \ldots \ f(x_n))', \end{aligned} \tag{A4}$$

and

$$\begin{aligned} Var(\widehat{Y}_{\mathrm{LM}}(x_{\mathrm{new}})) &= x'_{\mathrm{new}}(X'X)^{-1}X'Var(Y)X(X'X)^{-1}x_{\mathrm{new}} \\ &= \sigma^2 x'_{\mathrm{new}}(X'X)^{-1}x_{\mathrm{new}}. \end{aligned} \tag{A5}$$

With x_i and X as in (A2) and (A3), closed-form expression for $(X'X)^{-1}$ is

$$(X'X)^{-1} = \frac{1}{\frac{n^2(2n-1)}{6(n-1)} - \frac{n^2}{4}} \begin{pmatrix} \frac{n(2n-1)}{6(n-1)} & -\frac{n}{2} \\ -\frac{n}{2} & n \end{pmatrix}.$$

So it is easy to see

$$(X'X)^{-1} = \begin{pmatrix} \frac{4}{n} - \frac{6}{n^2} + O\left(\frac{1}{n^3}\right) & -\frac{6}{n} + \frac{12}{n^2} + O\left(\frac{1}{n^3}\right) \\ -\frac{6}{n} + \frac{12}{n^2} + O\left(\frac{1}{n^3}\right) & \frac{12}{n} - \frac{24}{n^2} + O\left(\frac{1}{n^3}\right) \end{pmatrix}, \tag{A6}$$

and hence from (A4) and (A5)

$$E(\widehat{Y}_{\text{LM}}(x_{\text{new}})) = \frac{\sum_{i=1}^{n}[4-6x_i+6(2x_i-1)x_{\text{new}}]f(x_i)}{n} + \frac{\sum_{i=1}^{n}[12x_i-6+12(1-2x_i)x_{\text{new}}]f(x_i)}{n^2} + O\left(\frac{1}{n^3}\right)\sum_{i=1}^{n} f(x_i), \tag{A7}$$

and

$$Var(\widehat{Y}_{LM}(x_{\text{new}})) = \frac{4(1-3x_{\text{new}}+3x_{\text{new}}^2)\sigma^2}{n} + O\left(\frac{1}{n^2}\right). \tag{A8}$$

Since $\sum_{i=1}^{n}[12x_i-6+12(1-2x_i)x_{\text{new}}] = n$, $\sum_{i=1}^{n}[12x_i-6+12(1-2x_i)x_{\text{new}}] = 0$, and $f(x)$ is bounded, say $a \le f(x) \le b$, from (A7), we have

$$E(\widehat{Y}_{\text{LM}}(x_{\text{new}})) = c_1(x_{new}) + O\left(\frac{1}{n^3}\right)\sum_{i=1}^{n} f(x_i),$$

where $c_1(x_{\text{new}}) \in [a, b]$. Therefore

$$\begin{aligned}&[f(x_{\text{new}}) - E(\widehat{Y}_{\text{LM}}(x_{\text{new}}))]^2\\ &= \left[f(x_{\text{new}}) - c_1(x_{\text{new}}) + O\left(\frac{1}{n^3}\right)\sum_{i=1}^{n} f(x_i)\right]^2\\ &= \left[f(x_{\text{new}}) - c_1(x_{\text{new}}) + O\left(\frac{1}{n^2}\right)\left(\frac{1}{n}\sum_{i=1}^{n} f(x_i)\right)\right]^2.\end{aligned}$$

Since $\sum_{i=1}^{n} f(x_i)/n = \int_0^1 f(x)dx + O_{a.s.}(n^{-2})$ if $f \in C^2[0, 1]$,

$$\begin{aligned}&[f(x_{\text{new}}) - E(\widehat{Y}_{\text{LM}}(x_{\text{new}}))]^2\\ &= \left[f(x_{\text{new}}) - c_1(x_{\text{new}}) + O\left(\frac{1}{n^2}\right)\int_0^1 f(x)dx\right]^2\\ &= [f(x_{\text{new}}) - c_1(x_{\text{new}})]^2 + O\left(\frac{1}{n^2}\right).\end{aligned} \tag{A9}$$

So, from (A1), (A8), and (A9) we get

$$MSE(\widehat{Y}_{\text{LM}}(x_{\text{new}})) = [f(x_{\text{new}}) - c_1(x_{\text{new}})]^2 + \frac{4(1-3x_{\text{new}}+3x_{\text{new}}^2)\sigma^2}{n} + O\left(\frac{1}{n^2}\right). \tag{A10}$$

A.2 Asymptotic MSE of PC-LM

We have

$$MSE(\widehat{Y}_{\text{PC-LM}}(x_{\text{new}})) = Var(\widehat{Y}_{\text{PC-LM}}(x_{\text{new}})) + [f(x_{\text{new}}) - E(\widehat{Y}_{\text{PC-LM}}(x_{\text{new}}))]^2, \tag{A11}$$

where $\widehat{Y}_{\text{PC-LM}}(x_{\text{new}}) = x_{\text{new}}\widehat{\beta}_{\text{PC-LM}} = x_{\text{new}}(X'X)^{-1}X'\widehat{Y}_{\text{PC}}$.

Consider $E(\widehat{Y}_{\text{PC-LM}}(x_{\text{new}}))$. Using the well-known formula

$$E(\widehat{Y}_{\text{PC}}(x_i)) = f(x_i) + 1/2h_n^2 f''(x_i)\int t^2K(t)dt + O(h_n^3),$$

and similar arguments as above we have

$$E(\widehat{Y}_{\text{PC-LM}}(x_{\text{new}})) = x'_{\text{new}}(X'X)^{-1}X'E(\widehat{Y}_{\text{PC}}),$$

and hence equals

$$\begin{aligned}&x'_{\text{new}}(X'X)^{-1}X'\\ &\quad\times\left(f(x_1) + \frac{1}{2}h_n^2 f''(x_1)\int t^2K(t)dt + O(h_n^3), \ldots,\right.\\ &\qquad\left.\times f(x_n) + \frac{1}{2}h^2 f''(x_n)\int t^2K(t)dt + O(h_n^3)\right)'\\ &= c_1(x_{\text{new}}) + c_2(x_{\text{new}})h_n^2 + O(h_n^3) + O\left(\frac{1}{n^2}\right),\end{aligned}$$

where $c_1(x_{\text{new}}) \in [a, b]$ and $c_2(x_{\text{new}}) \in [1/2c\int t^2K(t)dt, 1/2d\int t^2K(t)dt]$ if $c \le f''(x) \le d$. Therefore

$$\begin{aligned}&[f(x_{\text{new}}) - E(\widehat{Y}_{\text{PC-LM}}(x_{\text{new}}))]^2\\ &= \left[f(x_{\text{new}}) - c_1(x_{\text{new}}) - c_2(x_{\text{new}})h^2 + O(h_n^3) + O\left(\frac{1}{n^2}\right)\right]^2\\ &= [f(x_{\text{new}}) - c_1(x_{\text{new}})]^2 - 2(f(x_{\text{new}}) - c_1(x_{new}))c_2(x_{\text{new}})h_n^2\\ &\quad + O(h_n^3) + O\left(\frac{1}{n^2}\right).\end{aligned} \tag{A12}$$

So, the asymptotic squared bias is

$$\begin{aligned}&[f(x_{\text{new}}) - E(\widehat{Y}_{\text{PC-LM}}(x_{\text{new}}))]^2\\ &= [f(x_{\text{new}}) - c_1(x_{\text{new}})]^2 - 2(f(x_{\text{new}}) - c_1(x_{\text{new}}))c_2(x_{\text{new}})h_n^2.\end{aligned} \tag{A13}$$

Consider $Var(\widehat{Y}_{\text{PC-LM}}(x_{\text{new}}))$. Using $\nu = \lambda = 0$, $\delta = \gamma_\nu = \gamma_\lambda = 1$, equation labeled (7) in Gasser and Müller [9] yields $Var(\widehat{Y}_{\text{PC}}) = [c_3/(nh_n^2) + O((nh_n)^{-1})]\mathbb{1}$ where $\mathbb{1}$ is the $n \times n$ matrix of 1's. Then, with X and $(X'X)^{-1}$ as in (A3) and (A6), some algebraic manipulations give

$$\begin{aligned}&Var(\widehat{Y}_{\text{PC-LM}}(x_{\text{new}}))\\ &= x'_{\text{new}}(X'X)^{-1}X'Var(\widehat{Y}_{PC})X(X'X)^{-1}x_{\text{new}}\\ &= \frac{c_3}{nh_n^2} + O\left(\frac{1}{nh_n}\right).\end{aligned} \tag{A14}$$

So, the asymptotic variance is

$$Var(\widehat{Y}_{\text{PC-LM}}(x_{\text{new}})) = \frac{c_3}{nh_n^2}. \tag{A15}$$

Therefore, from (A13) and (A15), the asymptotic MSE is

$$AMSE(\widehat{Y}_{\text{PC-LM}}(x_{\text{new}})) = \frac{c_3}{nh_n^2} + [f(x_{\text{new}}) - c_1(x_{\text{new}})]^2 - 2(f(x_{\text{new}}) - c_1(x_{\text{new}}))c_2(x_{\text{new}})h_n^2.$$

By solving $\partial AMSE(\widehat{Y}_{\text{PC-LM}}(x_{\text{new}}))/\partial h = 0$, it is straightforward to see that the bandwidth that minimizes the *AMSE* above is $h_{\text{opt}} = O(n^{-1/4})$. Plugging this optimal bandwidth into (A12) and (A14), (A11) becomes

$$MSE(\widehat{Y}_{\text{PC-LM}}(x_{\text{new}})) = [f(x_{\text{new}}) - c_1(x_{\text{new}})]^2 - \frac{2(f(x_{\text{new}}) - c_1(x_{\text{new}}))c_2(x_{\text{new}}) - c_3}{n^{1/2}} + O\left(\frac{1}{n^{3/4}}\right). \tag{A16}$$

A.3 Asymptotic MSE of (STK,PC)-LM

The proof is similar to the arguments for dealing with $MSE(\widehat{Y}_{\text{PC-LM}}(x_{\text{new}}))$ in Theorem 3.1, Clause 4. Note that

$$E(\widehat{Y}_{\text{(STK,PC)-LM}}(x_{\text{new}})) = x'_{\text{new}}(X'X)^{-1}X'E(\widehat{Y}_{\text{STK,PC}}),$$

and hence equals

$$\begin{aligned}
&x'_{\text{new}}(X'X)^{-1}X'(E\widehat{Y}_{\text{STK,PC}}(x_1), \dots, E\widehat{Y}_{\text{STK,PC}}(x_n))' \\
&= x'_{\text{new}}(X'X)^{-1}X'\left(E\sum_{j=1}^{J} w_j\widehat{Y}_{j,\text{PC}}(x_1), \dots, E\sum_{j=1}^{J} w_j\widehat{Y}_{j,\text{PC}}(x_n)\right)' \\
&= x'_{\text{new}}(X'X)^{-1}X'\left(mf(x_1) + \frac{m}{2}h_n^2 f''(x_1)\int t^2K(t)dt \right. \\
&\quad \left. + O(h_n^3), \dots, mf(x_n) + \frac{m}{2}h_n^2 f''(x_n)\int t^2K(t)dt + O(h_n^3)\right)' \\
&= mc_1(x_{\text{new}}) + mc_2(x_{\text{new}})h_n^2 + O(h_n^3) + O\left(\frac{1}{n^2}\right),
\end{aligned}$$

where $m = \sum_{j=1}^{J} w_j$.

Since

$$\begin{aligned}
&Cov(\widehat{Y}_{\text{STK,PC}}(x_i), \widehat{Y}_{\text{STK,PC}}(x_k)) \\
&= Cov\left(\sum_{j=1}^{J} w_j\widehat{Y}_{j,\text{PC}}(x_i), \sum_{j=1}^{J} w_j\widehat{Y}_{j,\text{PC}}(x_k)\right) \\
&= \sum_{j,l=1}^{J} w_j w_l Cov(\widehat{Y}_{j,\text{PC}}(x_i), \widehat{Y}_{l,\text{PC}}(x_k)) = \frac{c_4}{nh_n^2} + O\left(\frac{1}{nh_n}\right),
\end{aligned}$$

and hence, again, $Var(\widehat{Y}_{\text{STK,PC}}) = [c_4/(nh_n^2) + O((nh_n)^{-1})]\mathbb{1}$ where $\mathbb{1}$ is the $n \times n$ matrix of 1's. So,

$$\begin{aligned}
&Var(\widehat{Y}_{\text{(STK,PC)-LM}}(x_{\text{new}})) \\
&= x'_{\text{new}}(X'X)^{-1}X'Var(\widehat{Y}_{\text{STK,PC}})X(X'X)^{-1}x_{\text{new}} \\
&= \frac{c_4}{nh_n^2} + O\left(\frac{1}{nh_n}\right).
\end{aligned}$$

Therefore, it is easy to see that the optimal bandwidth is $h_{\text{opt}} = O(n^{-1/4})$ as in Theorem 3.1 (Clause 4) and

$$\begin{aligned}
&MSE(\widehat{Y}_{\text{(STK,PC)-LM}}(x_{\text{new}})) = Var(\widehat{Y}_{\text{(STK,PC)-LM}}(x_{\text{new}})) \\
&+ [f(x_{\text{new}}) - E(\widehat{Y}_{\text{(STK,PC)-LM}}(x_{\text{new}}))]^2 = [f(x_{\text{new}}) - mc_1(x_{\text{new}})]^2 \\
&- \frac{2m(f(x_{\text{new}}) - mc_1(x_{\text{new}}))c_2(x_{\text{new}}) - c_4}{n^{1/2}} + O\left(\frac{1}{n^{3/4}}\right).
\end{aligned} \tag{A17}$$

APPENDIX B. PROOF OF THEOREM 3.2

1. First, for the predictor $\widehat{Y}_{\text{PC}}$, Gasser and Müller [9] gives

$$MSE(\widehat{Y}_{\text{PC}}(x_{\text{new}})) = O\left(\frac{1}{n^{4/5}}\right), \tag{B18}$$

using the optimal bandwidth $h_{\text{opt}} = O(n^{-1/5})$. Therefore, from (A18) and Theorem 3.1 (Clauses 1 and 4), we have, as $n \to \infty$,

$$MSE(\widehat{Y}_{\text{PC}}(x_{\text{new}})) \le MSE(\widehat{Y}_{\text{PC-LM}}(x_{\text{new}})),$$

and

$$MSE(\widehat{Y}_{\text{LM}}(x_{\text{new}})) \le MSE(\widehat{Y}_{\text{PC-LM}}(x_{\text{new}})).$$

In addition, if $f(x_{\text{new}}) - c_1(x_{\text{new}}) \neq 0$, then, again from (A18) and Theorem 3.1 (Clause 1),

$$MSE(\widehat{Y}_{\text{PC}}(x_{\text{new}})) \le MSE(\widehat{Y}_{\text{LM}}(x_{\text{new}})),$$

and hence

$$\begin{aligned}
&MSE(\widehat{Y}_{\text{PC}}(x_{\text{new}})) \le MSE(\widehat{Y}_{\text{LM}}(x_{\text{new}})) \\
&\le MSE(\widehat{Y}_{\text{PC-LM}}(x_{\text{new}})) \text{ as } n \to \infty.
\end{aligned}$$

Next, the inequality $IMSE(\widehat{Y}_{\text{STK,PC}}) \le IMSE(\widehat{Y}_{\text{PC}})$, as $n \to \infty$, is seen from the general result Theorem 2.2 in Le and Clarke [15], which proved

$$IMSE(\widehat{Y}_{\text{STK,PC}}) = IMSE(\widehat{Y}_{\text{PC}}) - \sum_{k=1}^{J} (w_k - 1_{k,j})^2,$$

where the stacking estimate of the jth model $\widehat{w}_j \xrightarrow{P} w_j$, $j = 1, \dots, J$, and $1_{k,j} = 1$ if $k = j$ and 0 otherwise.

2. The inequality $IMSE(\widehat{Y}_{\text{STK,(PC-LM)}}) \le IMSE(\widehat{Y}_{\text{PC-LM}})$, as $n \to \infty$, can be seen from the general result Theorem 2.2 in Le and Clarke [15] which proved

$$IMSE(\widehat{Y}_{\text{STK,(PC-LM)}}) = IMSE(\widehat{Y}_{\text{PC-LM}}) - \sum_{k=1}^{J} (w_k - 1_{k,j})^2,$$

where the stacking estimate of the jth model $\widehat{w}_j \xrightarrow{P} w_j$, $j = 1, \dots, J$, and $1_{k,j} = 1$ if $k = j$ and 0 otherwise.

The comparisons of $IMSE(\widehat{Y}_{\text{STK,(PC-LM)}})$ to $IMSE(\widehat{Y}_{\text{LM}})$, $IMSE(\widehat{Y}_{\text{PC}})$, or $IMSE(\widehat{Y}_{\text{STK,PC}})$ depend on the value of the sum $\sum_{k=1}^{J}(w_k - 1_{k,j})^2$. If the sum $\sum_{k=1}^{J}(w_k - 1_{k,j})^2 = 0$ when $w_j = 1$ and $w_k = 0$ $(k \neq j)$, then

$$IMSE(\widehat{Y}_{\text{STK,(PC-LM)}}) = IMSE(\widehat{Y}_{\text{PC-LM}}),$$

and hence, by Theorem 3.2 (Clause 1), $IMSE(\widehat{Y}_{\text{STK,(PC-LM)}})$ is larger than $IMSE(\widehat{Y}_{\text{LM}})$, $IMSE(\widehat{Y}_{\text{PC}})$, and $IMSE(\widehat{Y}_{\text{STK,PC}})$, as $n \to \infty$. On the other hand, if one of the weights $w_k, k = 1, \ldots, J$, is large enough, then the sum $\sum_{k=1}^{J}(w_k - 1_{k,j})^2$ will be large enough to make $IMSE(\widehat{Y}_{\text{STK,(PC-LM)}})$ smaller than $IMSE(\widehat{Y}_{\text{LM}})$, $IMSE(\widehat{Y}_{\text{PC}})$, or $IMSE(\widehat{Y}_{\text{STK,PC}})$, as $n \to \infty$.

3. The inequality $MSE(\widehat{Y}_{\text{PC}}(x_{\text{new}})) \leq MSE(\widehat{Y}_{\text{(STK,PC)-LM}}(x_{\text{new}}))$ as $n \to \infty$ could be seen from (A18) and Theorem 3.1 (Clause 5)

Next, from Theorem 3.1 (Clause 5), the leading term of $MSE(\widehat{Y}_{\text{(STK,PC)-LM}}(x_{\text{new}}))$ is $[f(x_{\text{new}}) - mc_1(x_{\text{new}})]^2$ while, from Clauses 1 and 4, the leading term of $MSE(\widehat{Y}_{\text{LM}}(x_{\text{new}}))$ or $MSE(\widehat{Y}_{\text{PC-LM}}(x_{\text{new}}))$ is $[f(x_{new}) - c_1(x_{new})]^2$. So, in order to compare $MSE(\widehat{Y}_{\text{(STK,PC)-LM}}(x_{\text{new}}))$ to $MSE(\widehat{Y}_{\text{LM}}(x_{\text{new}}))$ or $MSE(\widehat{Y}_{\text{PC-LM}}(x_{\text{new}}))$ we just compare the two leading terms $[f(x_{\text{new}}) - mc_1(x_{\text{new}})]^2$ and $[f(x_{\text{new}}) - c_1(x_{\text{new}})]^2$. However, the difference

$$\begin{aligned}&[f(x_{\text{new}}) - mc_1(x_{\text{new}})]^2 - [f(x_{\text{new}}) - c_1(x_{\text{new}})]^2\\ &\quad = (1-m)c_1(x_{\text{new}})[2f(x_{\text{new}}) - (m+1)c_1(x_{\text{new}})]\end{aligned}$$

can be positive or negative depending on the value x_{new}. This means that $MSE(\widehat{Y}_{\text{(STK,PC)-LM}}(x_{\text{new}}))$ could be larger or smaller when it is compared to $MSE(\widehat{Y}_{\text{LM}}(x_{\text{new}}))$ or $MSE(\widehat{Y}_{\text{PC-LM}}(x_{\text{new}}))$.