DISCUSSION

LAWRENCE D. BROWN AND YI LIN

University of Pennsylvania

1. Introduction. We congratulate the authors for a stimulating paper (referred to as HL in the following). As the authors correctly stated, the number of variables does not affect the optimal rate of convergence in a regular parametric model, but it does affect the optimal rate of convergence in nonparametric models. To be more precise, the optimal rate of convergence in a nonparametric function estimation problem depends on the “effective” nonparametric dimension of the model. For example, the “effective” dimension of a nonparametric additive model is 1, no matter how many variables are in the model. Therefore, variable selection in the additive model does not affect the optimal rate of convergence of the model. In light of this, we would like to consider the problem of adaptive estimation in the nonparametric functional ANOVA setup. It is readily seen that the variable selection problem considered in HL is a special case of the problem we consider. Following the example of HL, we will concentrate on the white noise model setting. This can be motivated by the results in Brown and Low [1] and Nussbaum [3].

2. Rates of convergence. Before turning to our main topic we would like to point out a feature of the result in HL that we found surprising. We hope that they will be able to comment on this, and perhaps provide some additional background and a heuristic explanation. The feature that concerns us first appears in the formula for $z_n(i_s)$ near the end of Section 1, and is repeated in various forms later on, including in Theorem 1. To focus on this feature, let us consider the case of one direction, $i_s$, as in Theorem 1 and, for simplicity, we consider only isotropic regression. Thus, assume a fixed smoothness, say $m$, throughout the model. Let the full dimension of the model be $d$, and let the dimension of the “direction” $i_s$ of interest be $s$, say. Then $\beta = m/d$ and $\beta(i_s) = m/s$. 
As HL note, if \( s > d/2 \) one may choose \( \alpha = \alpha_n = n^{-a} \) for suitable \( a > 0 \) and then \( z_n(i_s) \) coincides with the rate of convergence on the set \( \Sigma(i_s) \). This situation seems entirely satisfactory; we are concerned with the opposite situation where \( s \leq d/2 \). In that case, for any \( \alpha_n \) converging to 0,

\[
z_n(i_s) = \left( \frac{\sqrt{\ln(1/\alpha_n)}}{n} \right)^{2\beta/(4\beta+1)} \geq \left( \frac{1}{n} \right)^{\beta(i_s)/(2\beta(i_s)+1)}
\]

\( = \) rate of convergence on \( \Sigma(i_s) \).

Hence, no fully adaptive estimation is possible in the RNF sense of HL. What about the usual sense of adaptive estimation? Is fully adaptive estimation possible here in the usual sense? Or, is the rate \( z_n(i_s) \) the best possible ordinary rate of convergence over \( \Sigma(i_s) \) for an estimator in this situation if the optimal convergence rate is also desired over the full space?

The ordinary sense of adaptation does not involve a choice of \( \alpha \). So if \( z_n(i_s) \) is the best possible ordinary rate of convergence over \( \Sigma(i_s) \), then it is important in constructing an ordinary adaptive estimator to choose \( \alpha_n \) in an optimal way. For ordinary adaptive estimation, what is the optimal choice of \( \alpha_n \), and what is the corresponding optimal adaptive rate result?

In this connection we note for the construction in Section 3.1 it appears that any choice \( \alpha_n \to 0 \) will yield an estimator that converges at the rate with respect to asymptotic risk, defined as

\[
\lim_{B \to \infty} \limsup_{n \to \infty} \sup_{f \in \Sigma(i_s)} E_f^n \left\{ \varphi_n(\alpha_n, i_s)^{-2}(\min(\|\hat{f}_n - f\|, B))^2 \right\} \leq \infty.
\]

On the other hand it appears one must choose \( \alpha_n = n^{-a} \) for suitable \( a > 0 \) in order to attain the appropriately normed limiting rate, as defined via (3) of HL. If this is so, it is of additional technical interest as an instance where the asymptotic and the limiting risks can differ, and also where no optimal asymptotic risk is attained.

3. Functional ANOVA formulations. A general \( d \)-dimensional nonparametric function estimation problem has an optimal rate of convergence that depends on the magnitude of \( d \). For even moderately large \( d \), the rate of convergence is very slow compared to that of one-dimensional problems. This is one aspect of the so-called curse of dimensionality. To circumvent the curse of dimensionality, we often consider the functional ANOVA decomposition. Consider a \( d \)-dimensional nonparametric function with the following decomposition:

\[
f(x_1, x_2, \ldots, x_d) = \text{constant} + \sum_{i=1}^{d} f_i(x_i) + \sum_{i<j} f_{ij}(x_i, x_j) + \cdots,
\]

where the components satisfy side conditions which generalize the usual side conditions for parametric ANOVA to function spaces, and the series may be truncated in some manner.
There are two different types of functional ANOVA models commonly considered. They differ in the function spaces considered. Let $H^m([0, 1]^d)$ be the $m$th-order Sobolev Hilbert space of functions on $[0, 1]^d$. Stone [4] assumes smoothness conditions like $f_i \in H^m([0, 1])$, $f_{ij} \in H^m([0, 1]^2)$ and so on (Stone actually assumed Hölder spaces, which are similar to the Sobolev Hilbert spaces) and showed that the optimal rate of the model is $n^{-2m/(2m+s)}$, the same as that of $s$-dimensional full function problems, where $s$ is the highest order of interactions considered. Therefore the effective dimension of such functional ANOVA models is $s$. In the following we will refer to such models as partial derivative ANOVA (PD-ANOVA) models. The smoothing spline ANOVA models introduced in Wahba [5] and discussed in detail in Wahba et al. [6] make a different type of assumption on the component functions in the functional ANOVA decomposition. That is, after determining the function space of each main effect, the function space in which an interaction lies is assumed to be the tensor product space of the function spaces of the interacting main effects. Therefore, in the tensor product space ANOVA model, if we assume the main effects are in $H^m([0, 1])$, the $k$th-order interactions lie in $\otimes^k H^m([0, 1])$. We will refer to them as tensor product space ANOVA (TPS-ANOVA) models.

For a Hilbert space $E_1$ of functions of $x_1$ and a Hilbert space $E_2$ of functions of $x_2$, the tensor product space of $E_1$ and $E_2$ is defined as the completion of the class of functions $\{\sum_{i=1}^{k} f_i(x_1)g_i(x_2), f_i \in E_1, g_i \in E_2\}$ under a norm induced by the norms in $E_1$ and $E_2$. It is known (Lin [2]) that the tensor product space of $d$ Sobolev spaces $H^m([0, 1])$ is equivalent to

$$\Omega_m = \left\{ f : \frac{\partial^{|i|} f(x)}{\partial x^i} \in L_2([0, 1]^d), \forall i = \{i_1, i_2, \ldots, i_d\} \in \mathbb{R}^d \right\},$$

where $x = \{x_1, x_2, \ldots, x_d\}$, $|i| = \sum_j i_j$. The norm for any $f \in \Omega_m$ is

$$\|f\|_{\Omega_m}^2 = \sum_{|i|: \max_j i_j \leq m} \int \left[ \frac{\partial^{|i|} f(x)}{\partial x^i} \right]^2 dx.$$ 

Since any function in $\otimes^k H^m([0, 1])$ has one derivative of order $km$ (order $m$ in each direction), and some other derivatives of orders higher than $m$, we can see that TPS-ANOVA models put higher order smoothness conditions on interactions than on main effects, and the order of the smoothness condition imposed on an interaction increases with the order of the interaction. This reveals an intuitively appealing aspect of the tensor product ANOVA model: starting from an additive model, when we make the model more and more complex by throwing in higher and higher order interaction terms, we assume stronger and stronger smoothness conditions on the new terms thrown in to keep the model manageable. This is
consistent with the philosophy of the ANOVA modeling strategy of throwing away higher order interaction terms.

Lin [2] showed that the optimal rate of convergence for the tensor product space ANOVA model is \( [n (\log n)^{1-s}]^{-2m/(2m+1)} \), where \( s \) is the highest order of interactions considered. Notice this implies that the optimal rate of the saturated tensor product space model is \( [n (\log n)^{1-d}]^{-2m/(2m+1)} \). This is only a log factor away from the optimal rate of the one-dimensional nonparametric problems. Therefore the optimal rate of convergence of TPS-ANOVA models depends on the number of variables only through a log term.

HL consider a particular family of submodels related to the problem of variable selection, but their formulation can be applied to other subfamilies. For example, we can apply the framework to the model selection problem in the functional ANOVA framework. In principle we can take the function space \( \Sigma \) of HL to be the space corresponding to any functional ANOVA model and consider adaptive estimation with respect to smaller ANOVA submodels. For simplicity, we assume \( \Sigma \) to be a function space corresponding to the saturated ANOVA model, and we consider adaptive estimation with respect to the functional ANOVA submodels.

3.1. PD-ANOVA model. The saturated function space considered in PD-ANOVA is the same as the one considered in HL with \( \beta_1 = \beta_2 = \cdots = \beta_d = m \). Here \( \Sigma \) is the Sobolev space \( H^m([0, 1]^d) \). It is anticipated that the same results on rate of convergence in HL should also be valid PD-ANOVA. This can actually be proved by the same line of proof in HL. We only provide a brief description of how HL’s proofs can be modified to give results in PD-ANOVA setting. First consider a given functional ANOVA submodel \( M_1 \) of the form (1), with the highest order of interaction to be \( s \). Let \( S \) be the set of index sets corresponding to the generators of \( M_1 \). For example, for the model

\[
f(x_1, x_2, x_3) = f_0 + f_1 + f_2 + f_3 + f_{12} + f_{13} + f_{23},
\]

we have \( S = \{1, 2\}, \{1, 3\}, \{2, 3\} \}. The index set \( I \) in HL corresponds to the set of \( k \) such that \( \theta_k \) [defined in (22) in HL] are not zero for the functions in the submodel. In the ANOVA setting, the index set \( I \) can be defined correspondingly as

\[
I = \bigcup_{i \in S} \{ (k_1, k_2, \ldots, k_d) \in \mathbb{N}^d : k_j = 0, \ \forall j \not\in i \}.
\]

To avoid technicalities, we concentrate on the rate of convergence and ignore the constants that do not depend on \( \varepsilon \). We now show that, ignoring the constants, Theorem 1 of HL applies to PD-ANOVA. The \( \beta \) in HL should now be \( m/d \), and \( \beta(i_s) \) should now be replaced by \( m/s \). All the other quantities in HL can then be defined similarly ignoring the constants. The proof of the upper bound part of Theorem 1 follows the same line of argument as that of the proof of upper
bound in HL. The lower bound for the PD-ANOVA model is a corollary of the lower bound part of Theorem 1 of HL, since the function space in the PD-ANOVA model contains the function space in the variable selection model. Once Theorem 1 is established, the generalization to adaptation to multiple subspaces follows the development in HL.

3.2. Tensor product space model. Tensor product space framework is relevant to both variable selection and TPS-ANOVA model choice. In such a framework the saturated space Σ is \( \bigotimes^d H^m([0, 1]) \), which is different from the function space \( H^m([0, 1]^d) \) considered in HL. However, the framework in HL can still apply. For simplicity in this discussion we will consider only the TPS variable selection model. That is equivalent to a TPS-ANOVA model with only one term of dimension \( s \). It should be clear that the rates we obtain are valid for the general TPS-ANOVA model.

We basically follow the notation of HL. For notational simplicity, we concentrate on the case \( p = 2 \). We concentrate on the rate of convergence and ignore all the constants that do not affect rates. These include \( C’s \), \( Z’s \) and \( \lambda \) in HL. Also, we do not need \( \beta \) and \( \beta(i_s) \) in our tensor product space case. We also assume \( \alpha_\varepsilon \) to be a small fixed number independent of \( \varepsilon \), though we keep \( \alpha_\varepsilon \) in the notation just so that it is easy to see to which term in HL it corresponds. The full space \( \Sigma \) corresponds to

\[
\Sigma(m, L) = \left\{ \theta : \sum_{k \in \mathbb{N}^d} \left( \theta_k^2 \prod_{i=1}^d (1 + k_i)^{2m} \right) \leq L^2 \right\}.
\]

\( I \) and \( J \) are defined the same way as in HL. That is,

\[
I = \{(k_1, k_2, \ldots, k_d) \in \mathbb{N}^d : k_j = 0, \forall j \notin i_s\},
\]

\[
J = \mathbb{N}^d \setminus I.
\]

We define

\[
\Sigma(i_s) = \left\{ \theta : \sum_{k \in I} \left( \theta_k^2 \prod_{i=1}^d (1 + k_i)^{2m} \right) \leq L^2 \right\}.
\]

We further define

\[
I_\varepsilon = \left\{ (k_1, k_2, \ldots, k_d) \in I : \prod_{j=1}^s (1 + k_{i_j})^{2m} \leq K(\varepsilon) \right\},
\]

\[
J_\varepsilon = \left\{ (k_1, k_2, \ldots, k_d) \in J : \prod_{i=1}^d (1 + k_i)^{2m} \leq N(\varepsilon) \right\},
\]

\[
Q_\varepsilon = \left\{ (k_1, k_2, \ldots, k_d) \in \mathbb{N}^d : \prod_{i=1}^d (1 + k_i)^{2m} \leq M(\varepsilon) \right\}.
\]
with

\[ K(\varepsilon) = \left[ \varepsilon^2 \left( \log \frac{1}{\varepsilon} \right)^{s-1} \right]^{-2m/(2m+1)}, \]

\[ M(\varepsilon) = \left[ \varepsilon^2 \left( \log \frac{1}{\varepsilon} \right)^{d-1} \right]^{-2m/(2m+1)}, \]

\[ N(\varepsilon) = \left[ \varepsilon^4 \left( \log \frac{1}{\varepsilon} \right)^{d-1} \right]^{-2m/(4m+1)}. \]

Note that the way the above six quantities are defined is different from that in HL. Notice also for our definition it is always true that \( N(\varepsilon) > K(\varepsilon) \). This is a consequence of the fact that the rates of convergence in TPS models differ by only a log term.

Notations \( \hat{\theta}_e, \hat{\theta}_e^{(0)}(i_s) \) and \( T_e(i_s) \) have the same definitions as those in HL, but with our new \( I_e, J_e, Q_e \). We define

\[ \varphi_e(\Sigma) = \left[ \varepsilon^2 \left( \log \frac{1}{\varepsilon} \right)^{m/(2m+1)} \right]. \]

\[ \varphi_e(\alpha_e, i_s) = \left[ \varepsilon^2 \left( \log \frac{1}{\varepsilon} \right)^{s-1} \right]^{-m/(2m+1)}. \]

Notations \( A_e(i_s), \rho_e^{*}(i_s) \) and \( \theta_e^{*}(i_s) \) have the same definitions as those in HL, but with our new \( T_e(i_s), \varphi_e(\alpha_e, i_s), \varphi_e(\Sigma), \hat{\theta}_e \) and \( \hat{\theta}_e^{(0)}(i_s) \). We will show that Theorem 1 in HL holds in our case with our definitions (ignoring the constants). To do that, we first prove Lemma 1. This is a different lemma since all the quantities involved are defined differently now. Therefore we provide a detailed proof:

\[ E_0 \| \hat{\theta}_e - \theta \|^2 = \sum_{Q_e} \varepsilon^2 + \sum_{Q_e^c} \theta_k^2 = \varepsilon^2 |Q_e| + \sum_{Q_e^c} \theta_k^2. \]

We have the following approximation:

\[ |Q_e| \sim \int_{\prod_{i=1}^d (1+x_i)^{2m} \leq M(\varepsilon)} 1 \, dx_1 \, dx_2 \cdots \, dx_d \]

\[ = \int_{\prod_{i=1}^d (1+x_i)^{1/2m(\varepsilon)}} 1 \, dx_1 \, dx_2 \cdots \, dx_d. \]

Changing the variable in the integral, \( z_i = \prod_{j \leq i} (1+x_j), i = 1, 2, \ldots, d \), the above quantity becomes
\[
\int_1^{M^{1/2m}(\varepsilon)} \left[ \int_1^{z_d} \cdots \int_1^{z_2} \frac{1}{z_1 \cdots z_{d-1}} \, dz_1 \cdots \, dz_{d-1} \right] \, dz_d \\
= \int_1^{M^{1/2m}(\varepsilon)} \left[ (\log z_d)^{d-1} \right] \, dz_d \\
= z_d (\log z_d)^{d-1} \left| M^{1/2m}(\varepsilon) \right| -(d-1) \int_1^{M^{1/2m}(\varepsilon)} \left[ (\log z_d)^{d-2} \right] \, dz_d \\
\sim M^{1/2m}(\varepsilon)(\log M)^{d-1}.
\]

On the other hand, by the definition of \( Q_\varepsilon \) and \( \Sigma \), we have

\[
\sum_{Q_\varepsilon} \theta_k^2 M(\varepsilon) \leq \sum_{Q_\varepsilon} \theta_k^2 \prod_{i=1}^{d} (1+k_i)^{2m} \leq L^2.
\]

So we have

\[
E_{\theta} \| \hat{\theta}_\varepsilon - \theta \|^2 \sim \varepsilon^2 M^{1/2m}(\varepsilon)(\log M)^{d-1} + M^{-1}(\varepsilon)L^2 \sim \varphi_{\varepsilon}^2(\Sigma).
\]

Hence Lemma 1 is proved. Lemma 2 stays the same since it does not depend on our notation.

Similar to the derivation of \(|Q_\varepsilon|\), we have

\[
|I_\varepsilon| \sim K^{1/2m}(\varepsilon)(\log K)^{s-1}.
\]

With these in mind, the proof of the upper bound in HL goes through with our definition of the notation. In particular, the rate \( \varphi_{\varepsilon}(\alpha_\varepsilon, i_s) \) is \( \varepsilon^2(\log(\frac{1}{\varepsilon}))^{s-1} \). The proof of the lower bound is not needed. It follows directly from the fact that \( \varphi_{\varepsilon}(\alpha_\varepsilon, i_s) = \varepsilon^2(\log(\frac{1}{\varepsilon}))^{s-1} \) coincides with the minimax rate of convergence on \( \Sigma(i_s) \).

It is interesting to note that since the rate of convergence \( \varphi_{\varepsilon}(\alpha_\varepsilon, i_s) \) in the tensor product space framework is \( \varepsilon^2(\log(\frac{1}{\varepsilon}))^{s-1} \), the same as the minimax rate of convergence on \( \Sigma(i_s) \), the estimation is fully adaptive in the RNF sense of HL. The generalization to adaptation to multiple subspaces can be made following the development in HL. We do not pursue that here.

REFERENCES


Department of Statistics
University of Pennsylvania
Philadelphia, Pennsylvania 19104-6302
E-mail: lbrown@wharton.upenn.edu

Discussion

Sam Efromovich
University of New Mexico

This is a nice and very stimulating article. The article suggests estimating an underlying function together with its normalization factor. An estimate of the normalization factor is referred to as a random rate.

The article is also thought-provoking. Actually after reading this article I realized that I had many unanswered questions so I decided to use this discussion primarily for clarification of some of the issues. In what follows I denote by HL a procedure or a notion introduced in the article.

1. Can an HL estimator be data-driven? As I understand it, the HL setting is defined as follows. There are given an underlying (principal) function space $\Sigma$ and its $N$ subspaces $\Sigma_1, \ldots, \Sigma_N$. A statistician (and thus an estimator) knows everything about these spaces. The space $\Sigma$ is the one traditionally studied in the nonparametric minimax literature, and the subspaces describe a belief that an estimated signal may have a simpler structure and thus can be estimated more accurately. Then, according to the HL paradigm, an HL estimator must be minimax over the largest space $\Sigma$ but it is allowed to be not minimax over the subspaces where, instead of being necessarily minimax, an HL random rate (a new notion introduced by the authors) should satisfy some restrictions.

As a result, there are $N+1$ candidates (the underlying space and its $N$ subspaces) to choose from and this is the reason the authors refer to an HL procedure of estimation as an adaptive one. However, let us stress that all the HL procedures suggested require complete information about these $N+1$ candidates.

Can this requirement about knowing the candidates be dropped? In other words, can an HL procedure be data-driven? (To avoid possible confusion with notions of adaptive estimates used in the article, I use the notion of a data-driven estimation to stress that an estimator does not a priori depend on the spaces, in particular on the underlying $\Sigma$.)

At first glance the answer is “yes” because a similar setting was considered in the original articles by Lepski [3, 4] where a Lepski (L) adaptive procedure