Journal of Machine Learning Research 20 (2019) 1-26 Submitted 11/17; Revised 3/19; Published 5/19

Model Selection via the VC Dimension

Merlin Mpoudeu MERLIN.MPOUDEU@HUSKERS.UNL.EDU
Bank of America
Atlanta, GA, USA

Bertrand Clarke BCLARKE3@UNL.EDU
Department of Statistics
University of Nebraska-Lincoln
Lincoln, NE 68503, USA

Editor: John Shawe-Taylor

Abstract

We derive an objective function that can be optimized to give an estimator for the Vapnik-Chervonenkis dimension for use in model selection in regression problems. We verify our estimator is consistent. Then, we verify it performs well compared to seven other model selection techniques. We do this for a variety of types of data sets.

Keywords: Vapnik-Chervonenkis dimension, model selection, Bayesian information criterion, sparsity methods, empirical risk minimization, multi-type data.

1. Complexity and Model Selection

Model selection is often the first problem that must be addressed when analyzing data. In M-closed problems, see Bernardo and Smith (2000), the analyst posits a list of models and assumes one of them is true. In such cases, model selection is any procedure that uses data to identify one of the models on the model list. There is a vast literature on model selection in this context including information based methods such as the Aikaikie Information Criterion (AIC), the Bayes information criterion (BIC), residual based methods such as Mallows C_p or branch and bound, and code length methods such as the two-stage coding proposed by Barron and Cover (1991). We also have computational search methods such as simulated annealing and genetic algorithms. In addition, cross-validation (CV) is often used with non-parametric methods such as recursive partitioning, neural networks (aka deep learning) and kernel methods. A less well developed approach to model selection is via complexity as assessed by the Vapnik-Chervonenkis (VC) dimension, here denoted by d_{VC}. Its earliest usage seems to be in Vapnik and Chervonenkis (1968). A translation into English was published as Vapnik and Chervonenkis (1971).

Although, the VC dimension goes back to 1968, it wasn't until Vapnik et al. (1994) that a method for estimating d_{VC} was proposed in the classification context. Specifically, given a collection $\mathbb{C}$ of classifiers, Vapnik et al. (1994) tried to estimate the VC dimension of $\mathbb{C}$ by deriving an objective function based on the expected value of the maximum difference between two empirical evaluations of a single loss function, here denoted by Δ. The two empirical values come from dividing a given data set into a first and second part. The

objective function proposed by Vapnik et al. (1994) depends on d_{VC}, the sample size n, and several constants that had to be determined. Using their objective function, they derived an estimator $\hat{d}_{VC}$ for d_{VC} given a class $\mathbb{C}$ of classifiers. This algorithm treated possible sample sizes as design points $n_1, n_2, \cdots, n_L$ and requires one level of bootstrapping. Despite the remarkable contribution of Vapnik et al. (1994), the objective function was over-complex and the algorithm did not give a tight enough bound on Δ. Later, Vapnik and his collaborators suggested a fix to tighten the bound on Δ. We do not use this here; it is unclear if this 'fix' will work in classification, let alone regression.

Choosing the design points is a nontrivial source of variability in the estimate of d_{VC}. So, Shao et al. (2000) proposed an algorithm, based on extensive simulations, to generate optimal values of $n_1, n_2, \cdots, n_L$, given L. They argued that non-uniform values of the n_l's gave better results than the uniform n_l's used in Vapnik et al. (1994).

More recently, in a pioneering paper that deserves more recognition that it has received, McDonald et al. (2011) established the consistency of the Vapnik (1998) estimator $\hat{d}_{VC}$ for d_{VC} in the classification context.

The main reason the estimator for d_{VC} of Vapnik et al. (1994) did not become more widely used, despite the result in McDonald et al. (2011), is, we suggest, that it was too unstable because the objective function did not bound Δ tightly enough in terms of d_{VC}. In addition, the form of the objective function in Vapnik et al. (1994) is more complicated and less well-motivated than our result Theorem 2. The reason is that the derivation in Vapnik et al. (1994) uses conditional probabilities, one of which goes to zero quite quickly (with n). So, it contributes negligibly to the upper bound. Our derivation ignores the conditioning and bounds a CV form of Δ that is typically larger than that used in Vapnik et al. (1994).

Our consistency proof is a simplification of the proof of the main result McDonald et al. (2011). Accordingly, we obtain a slower rate of consistency, but the probability of correct model selection still goes to one.

Our overall strategy is to derive an objective function for estimating d_{VC} in the regression setting that provides, we think, a tighter bound on a modified form of Δ. To convert from classification to regression, we discretize the loss used for regression into m intervals (the case $m = 1$ would then apply to classification). To get a tighter bound, we change the form of Δ from what Vapnik et al. (1994) used and we optimize over the leading factor in our upper bound. To use our estimator, we use an extra layer of bootstrapping so the quantity we empirically optimize represents the quantity we derive theoretically more accurately. The extra layer of bootstrapping stabilizes our estimator of d_{VC} and appears to reduce its dependency on the n_l's. If the models are nested in order of increasing VC dimension, it is straightforward to choose the model with VC dimension closest to our estimate $\hat{d}_{VC}$. Otherwise, we can convert a non-nested problem to the nested case by ordering the inclusion of the covariates using a shrinkage method such as the 'smoothly clipped absolute deviation' (SCAD, Fan and Li (2001)), or correlation (see Fan and Lv (2008)), and use our $\hat{d}_{VC}$ as before. Even when we force a model list to be nested, our model selection method performs well compared to a range of competitors including Vapnik et al. (1994)'s original method, two forms of penalized empirical risk minimization (denoted $\widehat{PERM}_1$ and $\widehat{PERM}_2$), AIC, BIC, CV (10-fold), SCAD, and adaptive LASSO (ALASSO, Zou (2006)). Our general findings indicate that in realistic settings, model selection via estimated VC dimension,

when properly done, is fully competitive with existing methods and, unlike them, rarely gives abberant results.

This manuscript is structured as follows. In Sec. 2 we present the main theory justifying our estimator. In Subsec. 2.1 we discretize bounded loss functions so that upper bounds for the distinct regions involved in the definition of Δ can be derived and in Subsec. 2.2 we define our estimator of the VC dimension and give an algorithm for how to compute it. In Sec. 3 we use McDonald et al. (2011)'s consistency theorem to motivate our consistency theorem for $\hat{d}_{VC}$. In Sec. 4 we present our studies using simulated, benchmark, and real data. We compare our method for model selection to AIC, BIC, CV, $\widehat{PERM}_1$, and $\widehat{PERM}_2$. In this context, we suggest criteria to guide the selection of design points. Our comparisons also include simplifying non-nested model lists by using correlation, SCAD, and ALASSO. In Sec. 5 we discuss our overall findings.

2. Deriving an optimality criterion for estimating VC dimension

This section concerns Δ, the expected supremal difference between two evaluations of a bounded loss function, formally defined in (18) and (19). These bounds will enable us to derive an estimator for the VC dimension. In Sec. 2.1, we present our alternative version of the Vapnik et al. (1994) bounds and in Sec. 2.2, we present our estimator of d_{VC}.

2.1. Extension of the Vapnik et al. (1994) bounds to regression

Let $Z = (X, Y)$ be a random variable with outcomes $z = (x, y)$ assuming values on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. The first entry, $X = x$, is regarded as an explanatory variable leading to $Y = y$. Let $P \in \mathcal{M}(\mathcal{Z})$ be the distribution of Z, where $\mathcal{M}(\mathcal{Z})$ is the collection of probability measures on $\mathcal{Z}$, and let $Z_{1:2n} = (Z_1, \ldots, Z_n, Z_{n+1}, \ldots, \ldots, Z_{2n})$ be a data set of size $2m$ of independently and identically distributed (IID) copies of Z. Write $D^1 = \{Z_1, \ldots, Z_n\}$ for the first half and $D^2 = \{Z_{n+1}, \ldots, Z_{2n}\}$ for the second half. Writing $Z_i = (X_i, Y_i)$ for $i = 1 \ldots 2n$, let

$$Q\left(Z_i, \alpha\right) = L\left(Y_i, f\left(X_i, \alpha\right)\right),$$

for a bounded real valued loss function L and $\alpha \in \Lambda$. We assume that Λ is a compact set in a finite dimensional real space, that the interior of Λ, $\mathrm{Int}(\Lambda)$, is non-void and convex, and that $\Lambda = \overline{\mathrm{Int}(\Lambda)}$. Also, we assume the continuous functions $f(\cdot \mid \alpha)$ are parametrized by α continuously and one-to-one. Thus, in our examples, Λ will be the parameter space for a class of regression functions $f(\cdot \mid \alpha)$. For ease of exposition we assume L, and hence Q, are also continuous.

For a fixed $\alpha \in \Lambda$, discretize $Q(z, \alpha)$ using m disjoint intervals (with union $[0, B)$):

$$Q^m\left(z, \alpha\right) = \sum_{j=0}^{m-1} \frac{(2j+1)B}{2m} \mathbb{I}\left[Q(Z, \alpha) \in I_j^m\right]. \tag{1}$$

The discretization is based on the uniform left-closed, right-open partition of $[0, B)$ into m subintervals, here denoted I_j^m and the numbers $((2j+1)B)/(2m)$ are the midpoints. In (1), $\mathbb{I}[\cdot]$ is an indicator function taking value 1 when its argument is true and value 0 when it is false. We use losses of the form (1) to define a cross-validation form for Δ.

Start by letting $\alpha_1, \alpha_2 \in \Lambda$, with $\alpha_1 \neq \alpha_2$, and let

$$\nu(D^2, \alpha_1) = \frac{1}{n}\sum_{i=1}^{n} Q\left(Z_{n+i}, \alpha_1\right) \quad \text{and} \quad \nu(D^1, \alpha_2) = \frac{1}{n}\sum_{i=1}^{n} Q\left(Z_i, \alpha_2\right). \tag{2}$$

These are the empirical risks of model α_1 on the second half of the data and of model α_2 on the first half of the sample, respectively. Observe that the empirical counts of the data points whose losses land in I_j^m are

$$N_j^m(D^2, \alpha_1) = \sum_{i=1}^{n} \mathbb{I}\left[Q(Z_{n+i}, \alpha_1) \in I_j^m\right] \quad \text{and} \quad N_j^m(D^1, \alpha_2) = \sum_{i=1}^{n} \mathbb{I}\left[Q(Z_i, \alpha_2) \in I_j^m\right]. \tag{3}$$

This means we are counting the errors of the α_1 model on the second half of the data and the erors of the α_2 model on the first half of the data. This begins the set up of the cross-validation form of the error that we use and leads to the following expressions for the empirical losses of the discretized loss functions:

$$\begin{aligned} \nu^m\left(D^2, \alpha_1\right) &= \frac{1}{n}\sum_{j=0}^{m-1} N_j^m(D^2, \alpha_1)\frac{(2j+1)B}{2m} \\ \text{and} \quad \nu^m\left(D^1, \alpha_2\right) &= \frac{1}{n}\sum_{j=0}^{m-1} N_j^m(D^1, \alpha_2)\frac{(2j+1)B}{2m}. \end{aligned} \tag{4}$$

It is seen that the expressions in (4) are formed from the counts within each of the intervals. Let these be denoted by

$$\nu_j^m(D^2, \alpha_1) = \frac{1}{n}N_j^m(D^2, \alpha_1)\frac{(2j+1)B}{2m} \quad \text{and} \quad \nu_j^m(D^1, \alpha_2) = \frac{1}{n}N_j^m(D^1, \alpha_2)\frac{(2j+1)B}{2m}. \tag{5}$$

The first step in bounding Δ is to bound the probability of the 'bad set' where $\nu^m(D^2, \alpha_1)$ and $\nu^m(D^1, \alpha_2)$ are not close. Let $\epsilon > 0$ and, using the discretization into m intervals, define the set A_ϵ by the union:

$$A_\epsilon = \bigcup_{m=0}^{m-1} A_{\epsilon,m}, \tag{6}$$

where

$$A_{\epsilon,m} = \left\{ Z_{1:2n} \middle| \sup_{\alpha_1, \alpha_2 \in \Lambda} \left[\nu^m(Z_{n+1:2n}, \alpha_1) - \nu^m(Z_{1:n}, \alpha_2)\right] \geq \epsilon \right\}. \tag{7}$$

The only way a $Z_{1:2n} = z_{1:2n}$ can be in $A_{\epsilon,m}$ is that at least one value of j satisfies

$$\sup_{\alpha_1,\ \alpha_2 \in\ \Lambda} (\nu_j^m(z_{n+1:2n}, \alpha_1) - \nu_j^m(z_{1:n}, \alpha_2)) \geq \frac{\epsilon}{m}.$$

Since A_ϵ is defined on the entire range of our loss function, and we want to partition the range into m disjoint intervals, write

$$A_{\epsilon,m} \subseteq \{Z_{1:2n} \ \Big| \ \exists\, j \sup_{\alpha_1,\ \alpha_2 \in \Lambda} (\nu_j^m(Z_{n+1:2n}, \alpha_1) - \nu_j^m(Z_{1:n}, \alpha_2)) \geq \frac{\epsilon}{m}\} \subseteq \bigcup_{j=0}^{m-1} A_{\epsilon,m,j},$$

where $A_{\epsilon,m,j} = \{Z_{1:2n} \mid \sup_{\alpha_1,\ \alpha_2 \in \Lambda} (\nu_j^m(Z_{n+1:2n}, \alpha_1) - \nu_j^m(Z_{1:n}, \alpha_2)) \geq \frac{\epsilon}{m}\}$.

Next, fix any value $j \in \{0, 1, \ldots, m-1\}$. For any fixed $z_{1:2n}$, and any given $\alpha_1, \alpha_2 \in \Lambda$, define the vector of length $2n$

$$(Q_j^m(z_{n+1}, \alpha_1), \ldots, Q_j^m(z_{2n}, \alpha_1), Q_j^m(z_1, \alpha_2), \ldots, Q_j^m(z_n, \alpha_2)), \tag{8}$$

where $Q_j^m(z, \alpha) = \mathbb{I}(Q(z, \alpha) \in I_j^m)$. Now define $(\alpha_1, \alpha_2) \sim (\alpha_1', \alpha_2')$ when the corresponding $2n$-tuples are equal (for the given $z_{1:2n}$). It is seen that $\sim$ is an equivalence relation on $\Lambda \times \Lambda$ and therefore partitions $\Lambda \times \Lambda$ into disjoint equivalence classes. Let $K_j = K_j(z_{1:2n})$ be the number of equivalence classes for given j and $z_{1:2n}$ and for given $(\alpha_1, \alpha_2) \in \Lambda \times \Lambda$, write $[(\alpha_1, \alpha_2)]$ for the equivalence class that contains it. Now, for $k = 1 \ldots, K_j$ let

$$\left(\alpha_{1jk}^*, \alpha_{2jk}^*\right) = \arg \sup_{\alpha_1, \alpha_2 \in (\Lambda \times \Lambda)_k} \left(\nu_j^m\left(z_{n+1:2n}, \alpha_1\right) - \nu_j^m\left(z_{1:n}, \alpha_2\right)\right), \tag{9}$$

where $(\Lambda \times \Lambda)_k$ is the k-th equivalence class. Now, $\bigcup_{k=1}^{K_j} \left[(\alpha_{1jk}^*, \alpha_{2jk}^*)\right] = \Lambda \times \Lambda$ and $\left[(\alpha_{1jk}^*, \alpha_{2jk}^*)\right] \cap \left[(\alpha_{1jk}^*, \alpha_{2jk'}^*)\right] = \phi$ unless $k = k'$.

Any permutation π of $\{1, \ldots, 2n\}$ induces a permutation map $T_\pi : \mathcal{Z}^{2n} \to \mathcal{Z}^{2n}$ which acts by shuffling coordinates according to the indices permuted by π. There are $(2n)!$ such maps that can be denoted T_i for $i = 1, \ldots, 2n$. The IID assumption implies that the distribution of any $T_i(Z_{1:2n})$ is the same as the distribution of $Z_{1:2n}$. So, if any function $f : \mathcal{Z}^{2n} \to \mathbb{R}$ satisfies the symmetry condition $f(T_i(z_{1:2n})) = f(z_{1:2n})$ and is integrable, its integral satisfies

$$\int_{\mathcal{Z}^{2n}} f\left(Z_{1:2n}\right) \mathrm{d}P^{2n}\left(z_{1:2n}\right) = \int_{\mathcal{Z}^{2n}} f\left(T_i Z_{1:2n}\right) \mathrm{d}P^{2n}\left(z_{1:2n}\right), \tag{10}$$

in which $\mathrm{d}P^{2n}(z_{1:2n}) = \mathrm{d}P(z_1) \cdots \mathrm{d}P(z_{2n})$ and $P \in \mathcal{M}(\mathcal{Z})$.

One of the quantities that will be essential to getting a tight enough bound on $P^{2n}(A_\epsilon)$ is the annealed entropy $H_{ann}^{\Lambda}(2n)$. Given a sample, say $z_{1:2n}$, let $N^{\Lambda}(z_{1:2n})$ be the number of different separations of $z_{1:2n}$ by a given set of functions. In the proof we will choose all the functions in (8) for a given j. Since $N^{\Lambda}(z_{1:2n}) \leq 2^{2n}$ and the $N^{\Lambda}(z_{1:2n})$'s are measurable, $\mathbb{E}N^{\Lambda}(Z_{1:2n})$ exists. The annealed entropy is the natural logarithm (base e) of this, $H_{ann}^{\Lambda}(2n) = \log \mathbb{E}N^{\Lambda}(Z_{1:2n})$. As is customary, $\mathbb{E}$ means expectation in the true distribution, $P \in \mathcal{M}(\mathcal{Z})$.

Our first main result is similar to the corrresponding result in Vapnik et al. (1994). However, there are numerous differences are in the details. For instance, our equivalence class is defined on $\Lambda \times \Lambda$, we use a cross-validation form of the error, we discretized the loss function, and our result leads to Theorem 2 that only has one term, whereas the corresponding result in Vapnik et al. (1994) has three terms.

Theorem 1 : *Let $\epsilon \geq 0$ and $m \in \mathbb{N}$. If $d_{VC} = VC\left(\{Q\left(\cdot, \alpha\right) : \alpha \in \Lambda\}\right)$ is finite, then*

$$P^{2n}\left(A_\epsilon\right) \leq 2m \left(\frac{2ne}{d_{VC}}\right)^{d_{VC}} \exp\left\{-\frac{n\epsilon^2}{m^2}\right\}. \tag{11}$$

Remark: The technique used to prove (11) is similar to the proof of Theorem 4.1 in Vapnik (1998) giving bounds for the uniform convergence of the empirical risk. The hypotheses of Theorem 4.1 in Vapnik (1998) require only the existence of the key quantities e.g the annealed entropy, and the growth function. Our only extra condition is that d_{VC} be finite.
Proof : Let $m \in \mathbb{N}$ and $j \in \{0, 1, \ldots, m-1\}$. For any given $z_{1:2n}$, α_1, and α_2 write $\Delta_j^m(z_{1:2n}, \alpha_1, \alpha_2) = \nu_j^m(z_{n+1:2n}, \alpha_1) - \nu_j^m(z_{1:n}, \alpha_2)$. Also, denote

$$\begin{aligned}\left(\alpha_{1j}^*, \alpha_{2j}^*\right) &= \arg \sup_{(\alpha_1,\alpha_2)\in\Lambda\times\Lambda} \Delta_j^m(z_{1:2n}, \alpha_1, \alpha_2) \\ &= \arg \sup_{(\alpha_1,\alpha_2)\in\Lambda\times\Lambda} \left[\nu_j^m(z_{n+1:2n}, \alpha_1) - \nu_j^m(z_{1:n}, \alpha_2)\right] \qquad (12)\end{aligned}$$

It is seen that $(\alpha_{1j}^*, \alpha_{2j}^*)$ are estimated using D^2 and D^1 respectively; this reversal of the estimators with respect to the data is the essence of the cross-validation form of the error that we use. Using some manipulations, we have by dropping the superscript $2n$ on P:

$$\begin{aligned}P(A_\epsilon) &\leq P\left(\bigcup_{j=0}^{m-1} A_{\epsilon,m,j}\right) \leq \sum_{j=0}^{m-1} P(A_{\epsilon,m,j}) \\ &= \sum_{j=0}^{m-1} P\left(\left\{z_{1:2n} : \sup_{\alpha_1,\ \alpha_2\in\Lambda} \left(\nu_j^m(z_{n+1:2n}, \alpha_1) - \nu_j^m(z_{1:n}, \alpha_2)\right) \geq \frac{\epsilon}{m}\right\}\right). \\ &= \sum_{j=0}^{m-1} P\left(\left\{z_{1:2n} : \sup_{\alpha_1,\ \alpha_2\in\Lambda} \Delta_j^m(z_{1:2n}, \alpha_1, \alpha_2) \geq \frac{\epsilon}{m}\right\}\right) \\ &= \sum_{j=0}^{m-1} P\left(\left\{z_{1:2n} : \Delta_j^m(z_{1:2n}, \alpha_{1j}^*, \alpha_{2j}^*) \geq \frac{\epsilon}{m}\right\}\right).\end{aligned}$$

Now, for each T_i write $T_i(z_{1:2n})$ for the permuted sample and correspondingly write $D^1_{T_i(z_{1:2n})}$ and $D^2_{T_i(z_{1:2n})}$ for the first and second halves of the permuted sample. This implies

$$\Delta_j^m(T_i(z_{1:2n}), \alpha_1, \alpha_2) = \nu_j^m(D^2_{T_i(z_{1:2n})}, \alpha_1) - \nu_j^m(D^1_{T_i(z_{1:2n})}, \alpha_2).$$

By symmetry of this function and (10) we can write

$$P\left(\left\{z_{1:2n} : \Delta_j^m(z_{1:2n}, \alpha_{1j}^*, \alpha_{2j}^*) \geq \frac{\epsilon}{m}\right\}\right) = \frac{1}{(2n)!} \sum_{i=1}^{(2n)!} P\left(\left\{z_{1:2n} : \Delta_j^m(T_i(z_{1:2n}), \alpha_{1j}^*, \alpha_{2j}^*) \geq \frac{\epsilon}{m}\right\}\right)$$

and therefore $P(A_\epsilon)$ is bounded from above by

$$\frac{1}{(2n)!} \sum_{j=0}^{m-1} \sum_{i=1}^{(2n)!} P\left(\left\{z_{1:2n} : \Delta_j^m(T_i(z_{1:2n}), \alpha_{1j}^*, \alpha_{2j}^*) \geq \frac{\epsilon}{m}\right\}\right). \qquad (13)$$

Using the properties of the equivalence relation $\sim$ and letting Z' denote a dummy variable with the same distribution as Z, we have that for each fixed i, j and $z_{1:2n}$

$$\begin{aligned} I_{\left\{Z'_{1:2n}:\Delta_j^m\left(T_iZ'_{1:2n},\alpha^*_{1j},\alpha^*_{2j}\right)\geq\frac{\epsilon}{m}\right\}}(\cdot) &\leq I_{\left\{Z'_{1:2n}:\Delta_j^m\left(T_iZ'_{1:2n},\alpha^*_{1j1},\alpha^*_{2j1}\right)\geq\frac{\epsilon}{m}\right\}}(\cdot) \\ &+ \cdots + I_{\left\{Z'_{2n}:\Delta_j^m\left(T_iZ'_{1:2n},\alpha^*_{1jK_j(z_{1:2n})}\alpha^*_{2jK_j(z_{1:2n})}\right)\geq\frac{\epsilon}{m}\right\}}(\cdot) \\ &= \sum_{k=1}^{K_j(z_{1:2n})} I_{\left\{Z'_{1:2n}:\Delta_j^m\left(T_iZ'_{1:2n},\alpha^*_{1jk}\alpha^*_{2jk}\right)\geq\frac{\epsilon}{m}\right\}}(\cdot)\,. \end{aligned} \tag{14}$$

The inequality in (14) follows because each $z'_{1:2n}$ making the indicator function on the left side 1, must make at least one of the indicators on the right 1. This follows from the fact that $(\alpha^*_{1j},\alpha^*_{2j})$ is a global maximum and each $(\alpha^*_{1jk},\alpha^*_{2jk})$ is a local maximum for n equivalence class, see (12) and (9). Note that in (14) a T_i appears. Formally, this necessitates choosing $(\alpha^*_{1j},\alpha^*_{2j})$ and each $(\alpha^*_{1jk},\alpha^*_{2jk})$ for given k to be dependent on the i in T_i also; this extra step is suppressed in the notation since i has been dropped for ease of exposition.

Now, using (14), (13) is bounded by

$$\begin{aligned} P\left(A_\epsilon\right) &\leq \frac{1}{(2n)!}\sum_{j=0}^{m-1}\sum_{i=1}^{(2n)!}\int\sum_{k=1}^{K_j(z_{1:2n})} I_{\left\{Z'_{1:2n}:\Delta_j^m(T_iZ'_{1:2n},\alpha^*_{1jk}\alpha^*_{2jk})\geq\frac{\epsilon}{m}\right\}}(z_{1:2n})\,\mathrm{d}P\left(z_{1:2n}\right) \\ &= \int\sum_{j=0}^{m-1}\sum_{k=1}^{K_j(z_{1:2n})}\left[\frac{1}{(2n)!}\sum_{i=1}^{(2n)!} I_{\left\{Z'_{1:2n}:\Delta_j^m(T_iZ'_{1:2n},\alpha^*_{1jk},\alpha^*_{2jk})\geq\frac{\epsilon}{m}\right\}}(z_{1:2n})\right]\mathrm{d}P\left(z_{1:2n}\right). \end{aligned}$$

To bound the summation in square brackets, we follow Vapnik (1998), Chap. 4. Let

$$A_{\epsilon,m,j,k} = \left\{Z_{1:2n} : \Delta_j^m\left(T_iZ_{1:2n},\alpha^*_{1jk}\alpha^*_{2jk}\right)\geq\frac{\epsilon}{m}\right\}$$

for fixed j and each k, where $\left[(\alpha^*_{1jk},\alpha^*_{2jk})\right] = (\Lambda\times\Lambda)_k$. Now, the summation in square brackets is the fraction of the number of the $(2n)!$ permutations T_i of $Z_{1:2n}$ for which $A_{\epsilon,m,j,k}$ is closed under T_i for any fixed equivalence class $(\Lambda\times\Lambda)_k$. As proved in Vapnik (1998) Sec. 4.13, it equals

$$\Gamma_k = \sum_\ell \frac{\binom{b}{\ell}\binom{2n-b}{n-\ell}}{\binom{2n}{n}}$$

where $b = b(z_{1:2n})$ is the number of z_i's in $z_{1:2n}$ that satisfy $Q(z_i,\alpha^*_{1jk}) = 1$ (for $i = 1,\ldots,n$) or $Q(z_i,\alpha^*_{2jk}) = 1$ (for $i = n+1,\ldots,2n$), see Vapnik (1998), p. 136 or 143. The summation is over ℓ's in the set

$$\left\{\ell : \left|\frac{\ell}{n} - \frac{b-\ell}{n}\right| \geq \frac{\epsilon}{m}\right\}.$$

From Sec. 4.13 in Vapnik (1998), we have $\Gamma_k \leq 2\exp\left(-\frac{n\epsilon^2}{m^2}\right)$ uniformly in k.

So, using this in the last bound on $P(A_\epsilon)$ gives that $P\left(A_{\epsilon,m}\right)$ is bounded from above by

$$\begin{aligned}
&\int \sum_{j=0}^{m-1} \sum_{k=1}^{K_j(z_{1:2n})} 2\exp\left(-\frac{n\epsilon^2}{m^2}\right) dP(z_{1:2n}) = 2\exp\left(-\frac{n\epsilon^2}{m^2}\right) \int \sum_{j=0}^{m-1} \sum_{k=1}^{K_j(z_{1:2n})} dP(z_{1:2n}) \\
= \;& 2\exp\left(-\frac{n\epsilon^2}{m^2}\right) \sum_{j=0}^{m-1} \int \sum_{k=1}^{K_j(z_{1:2n})} dP(z_{1:2n}) = 2\exp\left(-\frac{n\epsilon^2}{m^2}\right) \sum_{j=0}^{m-1} \int K_j(z_{1:2n}) dP(z_{1:2n}) \\
= \;& 2\exp\left(-\frac{n\epsilon^2}{m^2}\right) \sum_{j=0}^{m-1} E\left(K_j(Z_{1:2n})\right). \qquad (15)
\end{aligned}$$

Since $K_j(z_{1:2n})$ is the number of equivalence classes given α_1, α_2, j, and $z_{1:2n}$ and N^Λ is the number of separations of $z_{1:2n}$ given by the functions in (8) i.e., over all $\alpha_1, \alpha_2 \in \Lambda$, we have that

$$K_j(z_{1:2n}) \leq N^\Lambda(z_{1:2n}).$$

The reasoning is as follows and simply makes the reasoning behind the statement at the top of p. 136 in Vapnik (1998) explicit. Recall $K_j(z_{1:2n})$ is the number of equivalence classes in $\Lambda \times \Lambda$ for fixed j and $z_{1:2n}$. If (α_1, α_2) and (α_1', α_2') are in different equivalence classes then

$$\exists u \quad Q_j^m(z_u, \alpha_1) \neq Q_j^m(z_u, \alpha_1') \quad \text{or} \qquad Q_j^m(z_u, \alpha_2) \neq Q_j^m(z_u, \alpha_2').$$

Without loss of generality, suppose the first inequality holds for some u. Then the two functions $Q_j^m(z_u, \alpha_1), Q_j^m(z_u, \alpha_1')$ must assume values $(0,1)$ or $(1,0)$. Again, without loss of generality suppose the first holds. Then, these two functions can separate $z_{1:2n}$ into two disjoint subsets $\{z_v \big| Q_j^m(z_v, \alpha_1) = 0\}$ and $\{z_v \big| Q_j^m(z_v, \alpha_1) = 0\}$ and this is one of the separations counted by $N^\Lambda(z_{1:2n})$. Taking into account all such separations we have

$$\mathbb{E}K_j(z_{1:2n}) \leq \mathbb{E}N^\Lambda(z_{1:2n}). \qquad (16)$$

The growth function is defined to be

$$G^\Lambda(2n) = \log \sup_{z_1, z_2, \ldots, z_{2n}} N^\Lambda\left(z_1, \ldots, z_{2n}\right) \geq \mathbb{E}N^\Lambda(Z_{1:2n}) = H_{ann}(2n).$$

So it is easy to see that $H_{ann}^\Lambda(2n) \leq G^\Lambda(2n)$. Now, Theorem 4.3 from Vapnik (1998) p.145 gives that

$$G(2n) \leq d_{VC} \log\left(\frac{2ne}{d_{VC}}\right) \Rightarrow \mathbb{E}\left(N^\Lambda(z_{1:2n})\right) \leq \left(\frac{2ne}{d_{VC}}\right)^{d_{VC}}. \qquad (17)$$

Using (17) and (16) m times in (15) gives the theorem. ■

Next, we use use Theorem 1 to identify an objective function that can be minimized to give an estimator for d_{VC}. Formally, let

$$\Delta_m = \mathbb{E}\left(\sup_{\alpha_1, \alpha_2 \in \Lambda} \left|\nu^m(D^2, \alpha_1) - \nu^m(D^1, \alpha_2)\right|\right) \qquad (18)$$

and

$$\Delta = \mathbb{E}\left(\sup_{\alpha_1,\alpha_2\in\Lambda}\left|\nu\left(D^2,\alpha_1\right)-\nu\left(D^1,\alpha_2\right)\right|\right). \tag{19}$$

Obviously, $\Delta_m \approx \Delta$ provided that $m, n,$ and $d_{VC} \to \infty$ at appropriate rates and the argument of Δ_m satisfies appropriate uniform integrability conditions. In fact, we do not use $\Delta_m \to \Delta$. For our purpose, the following bounds are sufficient. They are important to our methodology because they bound the expected maximum difference between two values of the empirical losses by an expression that can be used to estimate the VC dimension.

Theorem 2 :

1. *If $d_{VC} < \infty$, we have*

$$\Delta_m \leq m\sqrt{\frac{1}{n}\log\left(2m^3\left(\frac{2ne}{d_{VC}}\right)^{d_{VC}}\right)} + \frac{1}{m\sqrt{n\log\left(2m^3\left(\frac{2ne}{d_{VC}}\right)^{d_{VC}}\right)}} \tag{20}$$

2. *If $d_{VC} < \infty$, and*

$$D_p\left(\alpha\right) = \int_0^\infty \sqrt[p]{P\left\{Q\left(z,\alpha\right)\geq c\right\}}dc \leq \infty$$

where $1 < p \leq 2$ is some fixed parameter, we have

$$\Delta \leq \frac{D_p(\alpha^*)2^{2.5+\frac{1}{p}}\sqrt{d_{VC}\log\left(\frac{ne}{d_{VC}}\right)}}{n^{1-\frac{1}{p}}} + \frac{16D_p(\alpha^*)2^{2.5+\frac{1}{p}}}{n^{1-\frac{1}{p}}\sqrt{d_{VC}\log\left(\frac{ne}{d_{VC}}\right)}}. \tag{21}$$

3. *Assume that $d_{VC} \to \infty$, $\frac{n}{d_{VC}} \to \infty$, $m \to \infty$, $\log(m) = o(n)$, and*

$$D_p\left(\alpha\right) = \int_0^\infty \sqrt[p]{P\left\{Q(z,\alpha)\geq c\right\}}dc \leq \infty$$

where $p = 2$. Then we have that

$$\Delta \leq \min\left(1, 8D_p(\alpha^*)\right)\sqrt{\frac{d_{VC}}{n}\log\left(\frac{2ne}{d_{VC}}\right)}. \tag{22}$$

Proof : Proofs of the three clause of Theorem 2 can be found in Mpoudeu (2017) in Appendices A1–A3. They rest on using the integral of probabilities identity and then bounding the probabilities as in Theorem 1. ■

We can also use Theorem 1 to obtain an upper bound on the unknown true risk via the following propositions. Let $Q\left(\alpha_k\right)$ be the true unknown risk at α_k and $Q_{emp}\left(\alpha_k\right)$ be the empirical risk at α_k. Assume that K values of α_k have been fixed. These correspond to a set of points in the parameter space; in our examples below we use estimates. In effect, we are assuming that given an estimate, there is an α_k so close to it that the approximation error is negligible.

Proposition 3 : *For any $\eta \in (0,1)$, with probability at least $1-\eta$, the inequality*

$$Q(\alpha_k) \leq Q_{emp}(\alpha_k) + m\sqrt{\frac{1}{n}\log\left(\left(\frac{2m}{\eta}\right)\left(\frac{2ne}{d_{VC}}\right)^{d_{VC}}\right)} \tag{23}$$

holds simultaneously for all functions $Q(z,\alpha_k)$, $k = 1, 2, \cdots, K$.

Remark: This inequality follows from the additive Chernoff bounds (see, e.g., Vapnik (1998), formulae (4.4) and (4.5)) and suggests that the best model will be the one that minimizes the RHS of (23). The use of (23) in model selection as a form of risk minimization because as d_{VC} increases the second term on the right increases. This limits the size of d_{VC}; we denoted this technique by $PERM_1$ since a penalized empirical risk is being minimized.
Proof : To obtain inequality (23), we equate the RHS of Theorem 1 to a positive number $0 \leq \eta \leq 1$. Thus:

$$\eta = 2m\left(\frac{2ne}{d_{VC}}\right)^{d_{VC}} \exp\left(-\frac{n\epsilon^2}{m^2}\right).$$

Solving for ϵ gives

$$\epsilon \;=\; m\sqrt{\frac{1}{n}\log\left(\left(\frac{2m}{\eta}\right)\left(\frac{2ne}{d_{VC}}\right)^{d_{VC}}\right)}. \tag{24}$$

Proposition 3 can be obtained from the additive Chernoff bounds, expression 4.4 in Vapnik (1998) as follows

$$Q(\alpha_k) \;\leq\; Q_{emp}(\alpha_k) + \epsilon. \tag{25}$$

Using (24) in inequality (25), completes the proof. ■

Parallel to Prop. 3, we have the following for the multiplicative case.

Proposition 4 : *For any $\eta \in (0,1)$, with probability $1-\eta$, the inequality*

$$Q(\alpha_k) \leq Q_{emp}(\alpha_k) + \frac{m^2}{2n}\log\left(\frac{2m}{\eta}\left(\frac{2ne}{d_{VC}}\right)^{d_{VC}}\right)\left(1+\sqrt{1+\frac{4nQ_{emp}(\alpha_k)}{m^2\log\left(\frac{2m}{\eta}\left(\frac{2ne}{d_{VC}}\right)^{d_{VC}}\right)}}\right) \tag{26}$$

holds simultaneously for all K functions in the set $Q(z,\alpha_k)$, $k = 1, 2, \ldots, K$.

Remark: This follows from the multiplicative Chernoff bounds (see e.g., Vapnik (1998) formulae (4.17) and (4.18)) and suggests that the best model will be the one that minimizes the right hand side (RHS) of (26). Analogous to (23) we refer to the use of (26) in model selection as $PERM_2$.

Proof : Let $\epsilon, \eta > 0$. Then, inequality (4.18) in Vapnik (1998) gives, with probability at least $1 - \eta$, that

$$\frac{Q\left(\alpha_k\right) - Q_{emp}\left(\alpha_k\right)}{\sqrt{Q\left(\alpha_k\right)}} \leq \epsilon.$$

Routine algebraic manipulations and completing the square give

$$\left(Q\left(\alpha_k\right) - 0.5\left(\epsilon^2 + 2Q_{emp}\left(\alpha_k\right)\right)\right)^2 - 0.25\left(\epsilon^2 + 2Q_{emp}\left(\alpha_k\right)\right)^2 \leq -Q_{emp}^2\left(\alpha_k\right).$$

Taking the square root on both sides and re-arranging gives

$$Q\left(\alpha_k\right) \leq Q_{emp}\left(\alpha_k\right) + 0.5\epsilon^2\left(1 + \sqrt{1 + \frac{4Q_{emp}\left(\alpha_k\right)}{\epsilon^2}}\right)$$

Using (24) in the last inequality completes the proof of the Proposition. ■

More details on the use of Propositions 3 and 4 can be found in Vapnik (1998) and Mpoudeu (2017).

2.2. An Estimator of the VC Dimension

The upper bound from Theorem 2 can be written as

$$\Phi_{d_{VC}}(n) = \min\left(1, 8D_p(\alpha^*)\right)\sqrt{\frac{d_{VC}}{n}\log\left(\frac{2ne}{d_{VC}}\right)}. \tag{27}$$

This expression is meaningfully different from the form derived in Vapnik et al. (1994) and studied in McDonald et al. (2011). Moreover, although $\min\left(1, 8D_p(\alpha^*)\right)$ does not affect the optimization, it might not be the best constant for the inequality in (22). So, we replace it with an arbitrary constant c over which we optimize to make our upper bound as tight as possible. In our computations, we let c vary from 0.01 to 100 in steps of size 0.01. However, we have observed in practice that the best value of $\hat{c}$ is usually between 1 and 8. The technique that we use to estimate $\hat{d}_{VC}$ is also different from that in Vapnik et al. (1994). Indeed, our Algorithm #1 below accurately encapsulates the way the LHS of (22) is formed unlike the algorithm in Vapnik et al. (1994).

In particular, we use two bootstrapping procedures, one as a proxy for calculating expectations and the second as a proxy for calculating a maximum. Moreover, we split the data set into two subsets. Using the first data set, we fit model I and using the second we fit model II. To explain how we find our estimate of the RHS of (22) from Theorem 2, we start by replacing the sample size n in (27) with a specified value of design point, so that the only unknown is d_{VC}. Thus, formally, we replace (27) by

$$\Phi^*_{d_{VC}}(n_l) = \widehat{c}\sqrt{\frac{d_{VC}}{n_l}\log\left(\frac{2n_le}{d_{VC}}\right)},$$

where $\hat{c}$ is the optimal data driven constant. If we knew the left hand side (LHS) of (22), even computationally, we could use it to estimate d_{VC}. However, in general we don't know the LHS of (22). Instead, we generate one observation of the form

$$\xi(n_l) = \Phi^*_{d_{VC}}(n_l) + \epsilon(n_l) \tag{28}$$

for each design point n_l by bootstrapping and denoted the realized values by $\hat{\xi}(n_l)$. In (28), we assume $\epsilon(n_l)$ has a mean zero, but an otherwise unknown, distribution. We can therefore obtain a list of values of $\widehat{\xi}(n_l)$ for the elements of N_L. In effect, we are assuming that $\Phi^*_{d_{VC}}(n_l)$ provides a tight bound on Δ, and hence Δ_m as suggested by Theorem 2. Our algorithm is as follows.

Algorithm #1

Inputs:

- A collection of regression models $\mathcal{G} = \{g_\beta\}$,
- A data set,
- Two integers b_1 and b_2 for the number of bootstrap samples,
- An integer m for the number of subintervals to discretize the losses,
- A set of design points $N_L = \{n_1, n_2, \ldots, n_L\}$.

For each $l = 1, 2, \ldots, L$ do:

1. Take a bootstrap sample of size $2n_l$ (with replacement) form the data set;
2. Randomly subdivide the bootstrap data into two groups G^1 and G^2 of size n_l each;
3. Fit two models, one for G^1 and one for G^2;
4. Compute the squared error for each model on the covariates and responses that the other model was trained on. Thus:

$$SE_1 = (\text{predict}(Model_1, x_2) - y_2)^2 \quad \text{and} \quad SE_2 = (\text{predict}(Model_2, x_1) - y_1)^2$$

   where (x_1, y_1) ranges over G^1 and (x_2, y_2) ranges over G^2. So, there are n_l values of SE_1 and n_l values of SE_2.
5. Discretize the loss function, i.e. put each SE_1 and SE_2 in one of the m disjoint intervals;
6. Estimate $\nu^m_j(G^2, \alpha_1)$ and $\nu^m_j(G^1, \alpha_2)$ using the SE_1's and SE_2's respectively in the intervals I^m_j for $j = 0, 1, \ldots, m-1$;
7. Compute the differences $|\nu^m_j(G^2, \alpha_1) - \nu^m_j(G^1, \alpha_2)|$ for $j = 0, 1, \ldots, m-1$;
8. Repeat Steps 1-7 b_1 times, take the mean interval-wise and sum it across all intervals, so we have:

$$r_{b_1}(n_l) = \sum_{j=0}^{m-1} \text{mean}\, |\nu^m_j(G^2, \alpha_1) - \nu^m_j(G^1, \alpha_2)|\,;$$

9. Repeat Steps 1-8 b_2 times to get $r_{b_1,i}$ for $i = 1, 2, \ldots, b_2$ and form

$$\hat{\xi}(n_l) = \frac{1}{b_2} \sum_{i=1}^{b_2} r_{b_1,i}(n_l)\,.$$

It is seen that Step 9 uses a mean even though the definition of Δ_m and Δ (see (18) and (19)) has a supremum inside the expectation. This is intentional because using a supremum within each interval gave a worse estimator. We suggest that summing the mean over the intervals performs well because it is not too far from the supermum and is more stable.

Note that this algorithm is parallelizable because different n_l can be sent to different nodes to speed the process of estimating $\hat{\xi}\,(\cdot)$ for all n_l. After obtaining $\hat{\xi}(n_l)$ for each value of n_l, we estimate d_{VC} by minimizing the squared distance between $\hat{\xi}(n_l)$ and $\Phi^*_{d_{VC}}(n_l)$. Our objective function is

$$f_{n_l}(d_{VC}) = \sum_{l=1}^{L} \left(\hat{\xi}(n_l) - \hat{c}\sqrt{\frac{d_{VC}}{n_l} \log\left(\frac{2n_l e}{d_{VC}}\right)} \right)^2 , \tag{29}$$

where L is the number of design points. Optimizing (29) usually only leads to numerical solutions and in our work below, we set $b_1 = b_2 = W$ for convenience.

3. Proof of Consistency

In this section, we provide a proof of consistency for the estimator $\hat{d}_{VC}$ for d_{VC} that we presented in Subsec. 2.2. In many respects, the structure of this proof should be credited to McDonald et al. (2011). Our contribution is to adapt McDonald et al. (2011) to our stable estimator for the regression context. We begin with some notation and definitions.

Let $\Phi = \{\phi_{d_{VC},c}\}$ be a collection of real valued functions parametrized by $d_{VC} \in \mathbb{H} = [1, M]$ and $c \in I = [a, b] \subset \mathbb{R}$ with $M \in \mathbb{N}$ large enough and $0 < a < b < \infty$ so that $b - a > 0$ is also large enough. Elements of this collection are of the form

$$\phi_{d_{VC},c}\,(n_l) = c\sqrt{\frac{d_{VC}}{n_l} \log\left(\frac{2n_l e}{d_{VC}}\right)} \tag{30}$$

as derived in Subsec. 2.2 (see expression (27)). In expression (30), we assume L values $n_1, \ldots, n_L$ have been pre-specified. Fix a value of c and let $\Phi_c \subset \Phi$ be the section of elements corresponding to the fixed c. The proof holds for each fixed c and if we optimize over c to obtain $\hat{c}$ as explained in Subsec. 2.2, the convergence of $\hat{d}_{VC}$ to the true value d_{VC} will only be faster.

The collection of functions Φ is the continuous image of a compact set and hence is compact. Now, without loss of generality, we can choose $R > \sup_{d_{VC}} \|\phi_{d_{VC}}\|_L$ where the norm $\|\cdot\|_L$ is derived from the inner product

$$\langle f, g \rangle_L = \frac{1}{L} \sum_{l=1}^{L} f\,(n_l)\, g\,(n_l)$$

for real valued functions of a real variable. Thus $\phi_{d_{VC}} = (\phi_{d_{VC}}(n_1), \ldots, \phi_{d_{VC}}(n_L))$ (where the subscript c on the $\phi_{d_{VC},c}(n_L)$'s in expression (30) have been dropped for ease of notation). Fix a value of c and consider the compact subclass of Φ given by

$$\Phi_c(R) = \left\{\phi \in \Phi_c : \|\phi - \phi_{d_{VC}}\|_L < R\right\}, \tag{31}$$

where $\phi_{d_{VC}}$ is the element of Φ_c corresponding to the correct value of d_{VC}. For a given n_l, we have

$$\hat{\xi}(n_l) = \frac{1}{b_2}\sum_{i=1}^{b_2} r_{b_1,i}(n_l) \tag{32}$$

where $r_{b_1,i}(n_l)$ is the i bootstrapped value of the integrand of Δ_m for each n_l, $i = 1, \ldots, W$ and $l = 1, \ldots, L$. In vector form, write $\hat{\xi} = \left(\hat{\xi}(n_1), \ldots, \hat{\xi}(n_L)\right)$. Using (28), each $\hat{\xi}(n_l)$ can be represented as

$$\hat{\xi}(n_l) = \phi_{d_{VC}}(n_l) + \epsilon(n_l). \tag{33}$$

We have the following result.

Theorem 5 : *Suppose the true* $d_{VC} \in [1, M]$ *and that* $\forall i = 1, \ldots, W$, $\forall l = 1, \ldots, L$, $r_{b_1,i}(n_l) \sim N\left(\phi_{d_{VC}}(n_l), \sigma^2\right)$ *and independent,* $\mathbb{E}(\epsilon(n_l)) = 0$, $Var(\epsilon(n_l)) = \sigma^2$. *Then, on* $\Phi_c(R)$, *as* $n \to \infty$, $m \to \infty$ *and* $W = W(n) \to \infty$ *at suitable rates we have that*

$$P\left(\left\|\phi_{\hat{d}_{VC}} - \phi_{d_{VC}}\right\|_L \geq \delta\right) = \mathcal{O}\left(\frac{1}{W}\right). \tag{34}$$

Remark: In fact, the $r_{b_1,i}(n_l)$'s are only approximately independent $N\left(\phi_{d_{VC}}(n_l), \sigma^2\right)$. However, as n increases they become closer and closer to being independent $N\left(\phi_{d_{VC}}(n_l), \sigma^2\right)$, assuming $\phi_{d_{VC}}(n_l)$ is a tight enough upper bound, as $n, m \to \infty$ at appropriate rates. Also, it is seen that if $L = L(n)$ is increasing then $\|\cdot\|_L$ averages the evaluations of more and more components of, say, $\phi_{\hat{d}_{VC}}$. In the limit, this can be exihibited as an integral, i.e. as a quadratic norm. So, $\|\cdot\|_L$ can be regarded as an approximation of a L^2-space norm that strengthens as a norm (or inner product) as $n \to \infty$. In Theorem 5, if we controlled the distance between $\|\cdot\|_L$ and its limit, we could get a stronger mode of consistency.

Proof : By definition of $\phi_{d_{VC}}$, we have

$$\sum_{l=1}^{L}\left(\hat{\xi}(n_l) - c\sqrt{\frac{\hat{d}_{VC}}{n_l}\log\left(\frac{2n_l e}{\hat{d}_{VC}}\right)}\right)^2 \leq \sum_{l=1}^{L}\left(\hat{\xi}(n_l) - c\sqrt{\frac{d_{VC}}{n_l}\log\left(\frac{2n_l e}{d_{VC}}\right)}\right)^2 \tag{35}$$

or more compactly $\left\|\hat{\xi} - \phi_{\hat{d}_{VC}}\right\|_L^2 \leq \left\|\hat{\xi} - \phi_{d_{VC}}\right\|_L^2$. Expanding both sides of (35) gives

$$\sum_{l=1}^{L}\left(\phi^2_{\hat{d}_{VC}}(n_l) - \phi^2_{d_{VC}}(n_l)\right) \leq 2\sum_{l=1}^{L}\hat{\xi}(n_l)\left(\phi_{\hat{d}_{VC}} - \phi_{d_{VC}}(n_l)\right)$$

and hence

$$\begin{aligned}\left\|\phi_{\hat{d}_{VC}}\right\|_L^2 - \|\phi_{d_{VC}}\|_L^2 &\leq \frac{2}{L}\sum_{l=1}^{L}\left(\phi_{d_{VC}}(n_l)+\epsilon(n_l)\right)\left(\phi_{\hat{d}_{VC}}(n_l)-\phi_{d_{VC}}(n_l)\right)\\ &= \frac{2}{L}\sum_{l=1}^{L}\Big(\phi_{d_{VC}}(n_l)\,\phi_{\hat{d}_{VC}}(n_l)-\phi_{d_{VC}}^2(n_l)\\ &\qquad +\epsilon(n_l)\left(\phi_{\hat{d}_{VC}}(n_l)-\phi_{d_{VC}}(n_l)\right)\Big).\end{aligned}$$

Rearranging gives

$$\left\|\phi_{\hat{d}_{VC}}\right\|_L^2 - 2\langle\epsilon,\phi_{\hat{d}_{VC}}\rangle + \|\phi_{d_{VC}}\|_L^2 \leq 2\langle\epsilon,\phi_{d_{VC}}-\phi_{\hat{d}_{VC}}\rangle_L,$$

where $\epsilon=(\epsilon(n_1),\ldots,\epsilon(n_L))$, i.e.

$$\left\|\phi_{\hat{d}_{VC}}-\phi_{d_{VC}}\right\|_L^2 \leq 2\langle\epsilon,\phi_{\hat{d}_{VC}}-\phi_{d_{VC}}\rangle_L. \tag{36}$$

It is seen that the LHS is the main quantity we want to control. We have

$$\begin{aligned}P\left(\left\|\phi_{\hat{d}_{VC}}-\phi_{d_{VC}}\right\|_L>\delta\right) &\leq P\left(\langle\epsilon,\phi_{d_{VC}}-\phi_{\hat{d}_{VC}}\rangle\geq\frac{\delta^2}{2}\right)\\ &\leq P\left(\|\epsilon\|_L\left\|\phi_{d_{VC}}-\phi_{\hat{d}_{VC}}\right\|_L>\frac{\delta^2}{2}\right)\\ &\leq \frac{2R^2}{\delta^2}\mathbb{E}\|\epsilon\|_L^2,\end{aligned} \tag{37}$$

using the Cauchy-Schwarz inequality, the bound in (31), and Markov's inequality.

By construction, we have that

$$\begin{aligned}\mathbb{E}\|\epsilon\|_L^2 &= \frac{1}{L}\sum_{l=1}^{L}\mathbb{E}\left(\epsilon^2(n_l)\right)\\ &= \frac{1}{L}\sum_{l=1}^{L}\mathbb{E}\left[\left(\frac{1}{W}\sum_{i=1}^{W}r_{b_1,i}(n_l)\right)-\phi(n_l)\right]^2\\ &= \frac{1}{L}\sum_{l=1}^{L}\mathbb{E}\left[\frac{1}{W}\sum_{i=1}^{W}\left(r_{b_1,i}(n_l)-\phi(n_l)\right)\right]^2\\ &= \frac{1}{LW^2}\sum_{l=1}^{L}\sum_{i=1}^{W}\mathbb{E}\left(r_{b_1,1}(n_l)-\phi(n_l)\right)^2\\ &= \frac{1}{LW^2}\sum_{l=1}^{L}\sum_{i=1}^{W}Var\left(r_{b_1,1}(n_l)\right)=\frac{\sigma^2}{W}.\end{aligned} \tag{38}$$

Using (38) in (37) gives

$$P\left(\left\|\phi_{\hat{d}_{VC}}-\phi_{d_{VC}}\right\|_L\geq\delta\right)\leq\frac{2R^2\sigma^2}{\delta^2W} \tag{39}$$

in which the upper bound decreases as n increases because $W(n)$ is increasing, thereby giving (34). ■

A notable difference between (34) and the corresponding theorem in McDonald et al. (2011) is that our simplified result effectively only gives

$$P\left(\left\|\phi_{\hat{d}_{VC}} - \phi_{d_{VC}}\right\|_L \geq \delta\right) = \mathcal{O}\left(\frac{1}{W}\right) \tag{40}$$

rather than $\mathcal{O}\left(e^{-\gamma W}\right)$ for some $\gamma > 0$, a much faster rate. We conjecture that the more sophisticated techniques used in McDonald et al. (2011) could be adapted to our setting and thereby give an exponentially fast rate of convergence of $\hat{d}_{VC}$ to d_{VC} in probability. However, as yet, we have not been able to show this. Also, although it is suppressed in the notation, our result implicitly requires $m \to \infty$ to justify the use of $\phi_{d_{VC}}$.

Using Theorem 5, we can show that our $\hat{d}_{VC}$ is consistent. Suppose that $\phi_{d_{VC}}(\cdot)$ is κ-expansive, or simply expansive when κ is understood, i.e. $\forall n_l, \exists \kappa = \kappa(n_l)$ so that $\kappa(n_l)\left|d_{VC} - d'_{VC}\right| \leq \left|\phi_{d_{VC}}(n_l) - \phi_{d'_{VC}}(n_l)\right|$, where $\kappa(n)$, the expansion factor, is bounded on compact sets. Since the form of $\phi_{d_{VC}}(n_l)$ is known from (27), it is clear that the uniform expansivity condition we have assumed below actually holds, at least for appropriately chosen compact sets. We also observe that for $c \in I$ there exists a neighborhood $B(c, \epsilon_l), \eta > 0$, on which (34) is true. Cover $I \times \mathbb{H}$ by sets of the form $B(c, \eta) \times \{d_{VC}\}$; finitely many will be enough since $I \times \mathbb{H}$ is compact.

Theorem 6 : *Given that the assumptions of Theorem 5 hold and that $\phi_{d_{VC}}(\cdot)$ is expansive, we have, as $n \to \infty$, that*

$$P\left(\left|\hat{d_{VC}} - d_{VC}\right| \geq \delta\right) \leq \frac{2R^2\sigma^2}{\delta^2 \kappa W} = \mathcal{O}\left(\frac{1}{W}\right), \text{ as } W_n \to \infty, \tag{41}$$

where $\kappa = \sqrt{\frac{1}{L}\sum_{l=1}^{L}\kappa(n_l)}$ is the overall expansion factor.

Proof : Since all L of the $\phi_{d_{VC}}(n_l)$'s are at least locally expansive, their local expansivity inequalities can be summarized by an inequality of the form

$$\begin{aligned}\left|d_{VC} - d'_{VC}\right| \sqrt{\frac{1}{L}\sum_{l=1}^{L}\kappa(n_l)} &\leq \sqrt{\frac{1}{L}\sum_{l=1}^{L}\left(\phi_{d_{VC}}(n_l) - \phi_{d'_{VC}}(n_l)\right)^2} \\ &= \|\phi_{d_{VC}}(n_l) - \phi_{d'_{VC}}(n_l)\|_L,\end{aligned} \tag{42}$$

where d_{VC} is the true value and d'_{VC} is any other value in $\mathbb{H}$, and any extra constant from the local expansion factors are assumed to have been absorbed into the $\kappa(n_l)$'s as needed. Let $\kappa = \sqrt{\frac{1}{L}\sum_{l=1}^{L}\kappa(n_l)}$. Using Theorem 5, and (42) we have

$$\begin{aligned}P\left(\left|\hat{d}_{VC} - d_{VC}\right| \geq \delta\right) &\leq P\left[\|\phi_{\hat{d}_{VC}}(n_l) - \phi_{d_{VC}}(n_l)\|_L \geq \delta\kappa\right] \\ &\leq \frac{2R^2\sigma^2}{\kappa\delta^2 W},\end{aligned} \tag{43}$$

where the last upper bound decreases as $W = W_n \to \infty$ as $n \to \infty$, giving (41). ■

4. Numerical Comparisons

For any model, we can estimate the LHS of (22) from Theorem 2 by Algorithm #1 in Sec. 2.2. Then, we can use nonlinear regression in (29) to find $\hat{d}_{VC}$. So, it is seen that $\hat{d}_{VC}$ is a function of the conjectured model. In principle, for any given model class, the VC dimension can be found, so our method can be applied.

Since our goal is to estimate the true VC dimension, when a conjectured model $P(\cdot \mid \beta)$ is linear and correct, we expect $VC(P(\cdot \mid \beta)) \cong \hat{d}_{VC}$. By the same logic, if $P(\cdot \mid \beta)$ is far from the true model, we expect $VC(P(\cdot \mid \beta)) \gg \hat{d}_{VC}$ or $VC(P(\cdot \mid \beta)) \ll \hat{d}_{VC}$. This suggests we estimate d_{VC} by seeking

$$\hat{d}_{VC} = \arg\min_k \left| VC(P_k(\cdot \mid \beta)) - \hat{d}_{VC,k} \right|, \tag{44}$$

where $\{P_k(\cdot \mid \beta) \,|k = 1, 2, \cdots, K\}$ is some set of models and $\hat{d}_{VC,k}$ is calculated using model k, t is a positive and usually small number that such that $t \leq 2$. In the case of linear models, with $q = 1, 2, \cdots, Q$ explanatory variables, we get

$$\hat{d}_{VC} = \arg\min_q \left| q - \hat{d}_{VC,q} \right|, \tag{45}$$

where $\hat{d}_{VC,q}$ is the estimated VC dimension for model of size q. Note that (44) can identify a good model even when consistency fails. The reason is that (44) only requires a minimum at the VC dimension not convergence to the true VC dimension which may be any model under consideration. Here, to achieve uniqueness we use (45) and choose the smalest q achieving the smallest value of $|q - \hat{d}_{VC,q}|$, provided this makes sense in context.

Our numerical work uses linear models, since for these we know the VC dimension equals the number of explanatory variables, see Anthony and Bartlett (2009). To establish notation, we write the regression function as a linear combination of the fixed effect covariates x_j, $j = 0, 1, \cdots, p$,

$$y = f(x, \beta) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p = \sum_{j=0}^{p} \beta_j x_j.$$

Given a data set, $\{(x_i, y_i),\ i = 1, 2, \ldots, n\}$, the matrix representation is $Y = X\beta + \epsilon$ where Y is the $n \times 1$ vector of response values, X is the $n \times (1 + p)$ matrix with rows $(1, x_{1,1}, x_{1,2}, \ldots, x_{1,p})$, $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ is the vector of model parameters, and $\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^T$ is a $n \times 1$ mean zero Gaussian random vector. The least squares estimator $\widehat{\beta}$, assuming it exists, is given by $\widehat{\beta} = \left(X'X\right)^{-1} X'Y$.

4.1. Simulated data

We simulate data from

$$Y = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon,$$

where $\epsilon \sim N(0, \sigma_\epsilon = 0.4)$ and

$$x_0 = 1,\ \beta_j \sim N(\mu = 5, \sigma_\beta = 3), x_j \sim N(\mu = 5, \sigma_x = 2),\ \text{for } j = 1, 2, \cdots p,$$

in which all the β's, x's and ϵ's are independently generated. We center and scale all our variables, including the response. For convenience, we use a nested sequence of model lists. If our covariates were highly correlated, before applying our method we could de-correlate them by sphering, i.e. transforming the covariates using their covariance matrix so they become approximately uncorrelated with variance one, see Murphy (2012) p. 144.

In Subsec. 4.1.1, we present a typical simulation result to verify our estimator for VC dimension is consistent for the VC dimension of the true model. In Subsec. 4.1.2, we discuss simulations we have done where results do not initially appear to be consistent with the theory. First, large values of n are needed to get good performance with large values of p. Second as p increases, we must choose n_l's that are properly spread out over $[0, n]$.

4.1.1. A FIRST EXAMPLE

We implemented simulations for a range of model sizes $p = 15, 30, 40, 50, 60$ and 70 and applied six model selection techniques AIC, BIC, CV, $\widehat{PERM}_1$, $\widehat{PERM}_2$, and VC dimension (VCD), see Mpoudeu and Clarke (2018) for details. We tended to use larger sample sizes with larger values of p and spaced the design points uniformly over $[0, n]$, even though this may be suboptimal. We arbitrarily set $m = 10$ and $W = 50$. Our models were nested, including models that were too small and some that were too large, so that the estimate of d_{VC} would uniquely specify a model.

As a typical example, Fig. 1 shows the results for $n = 700$ and $p = 70$ with $L = 7$. When the size of the conjectured model is strictly less than the size of the true model, $\hat{d}_{VC}$ is equal to the smallest design point. However, when the conjectured model exactly matches the true model, $\hat{d}_{VC} \approx 61$, underestimating d_{VC}. Interestingly, if we simply look at the minimal VCD value it occurs at the conjectured model of size 70, the true values of p. When the conjectured model is more complex than the true model, the VCD value is visibly higher than the VCD value for the true model. Thus, using VCD favors parsimony more than the other methods do. In results not shown here, we increased n to 2000 and used good design points (as discussed in Subsec. 4.1.2) and found $\hat{d}_{VC} \approx 70$. Our observation for the other five model selection methods is that they are less affected by the small sample size, but decrease and 'flatline' in the sense that the AIC, BIC and CV values decrease very slowly (making it unclear which model to choose) while $\widehat{PERM}_1$ and $\widehat{PERM}_2$ routinely give models that are too small if one follows the usual rule of choosing the smallest local minimum. Overall, using VCD penalizes models that are too large more than other methods do, thereby giving a clearer statement about which model is true, at least in the limit with intelligently chosen design points. Other choices of n, p, the design points, and other inputs, gave results compatible with this interpretation.

Although the diagrams are not shown here (but see Mpoudeu and Clarke (2018), Sec. 4) the smallest discrepancy between the size d_{VC} of the model and $\hat{d}_{VC}$ usually occurs at the true model. This indicates that $\hat{d}_{VC}$ is consistent. In addition, even though the VCD values generally increase as the size of the conjectured model exceeds the size of the true model, in some cases, past a certain value d_{VC}, the VCD value may flatline as well. The

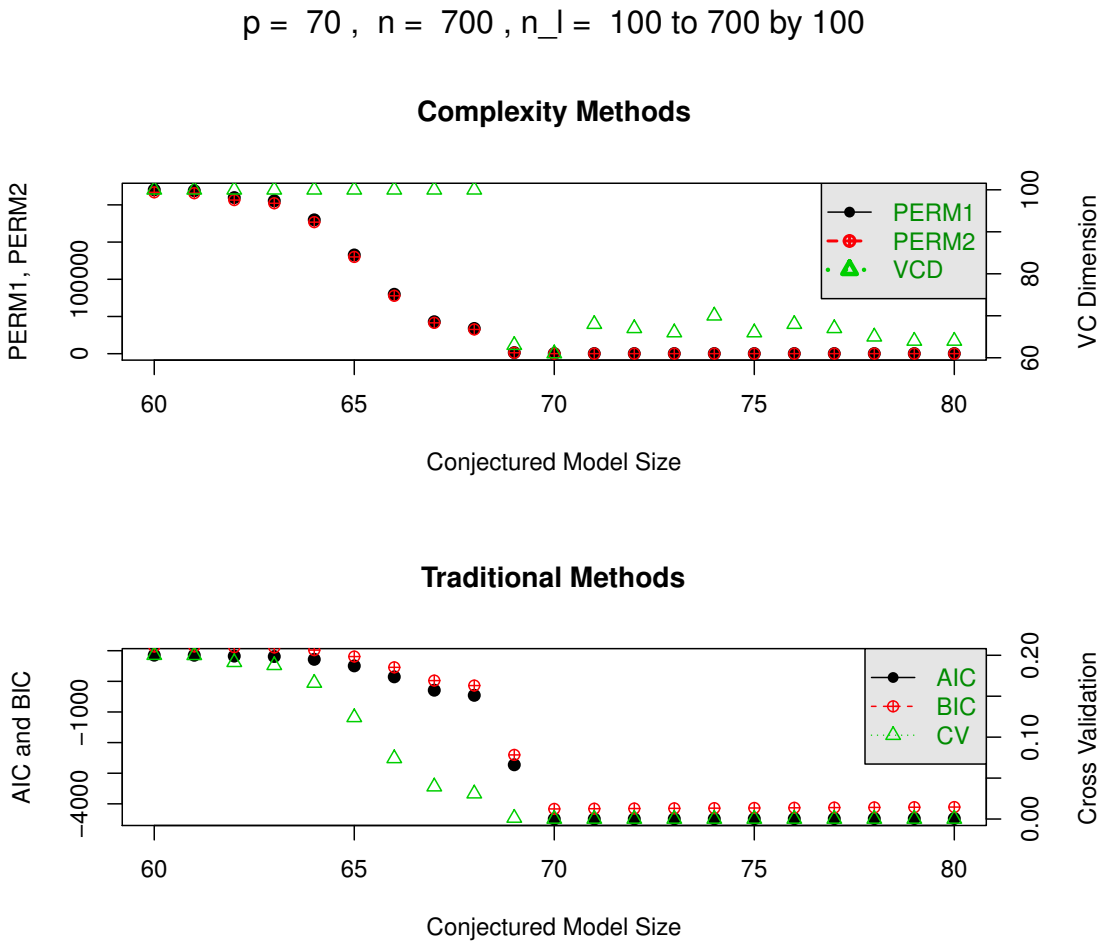

Figure 1: Upper: Values of $\widehat{PERM}_1$, $\widehat{PERM}_2$, and VC dimension. Lower: Values of AIC, BIC and CV for $p = 70$, $\sigma_\epsilon = 0.4$, $\sigma_\beta = 3$, $\sigma_x = 2$.

problem with $\hat{d}_{VC}$ flatlining (or decreasing) past a certain value of d_{VC} occurs mostly due to instability, e.g., when n is not large enough relative to p.

We argue that estimating VC dimension directly is better than using $\widehat{PERM}_1$ or $\widehat{PERM}_2$. There are several reasons. First, the computation of $\widehat{PERM}_1$ and $\widehat{PERM}_2$ requires $\hat{d}_{VC}$. It also requires a threshold η be chosen (see Propositions 3 and 4) and is more dependent on m than $\hat{d}_{VC}$ is. Being more complicated than $\hat{d}_{VC}$, $\widehat{PERM}_1$, $\widehat{PERM}_2$ will break down faster than $\hat{d}_{VC}$. This is seen, for instance in tables of Mpoudeu (2017) Chap. 3 and the discussion there. More generally, we argue that $\widehat{PERM}_1$, and $\widehat{PERM}_2$ break down faster than $\hat{d}_{VC}$ with increasing p, if the sample size is held constant. That is, $\widehat{PERM}_1$ and $\widehat{PERM}_2$ are less efficient than $\hat{d}_{VC}$.

Finally for this subsection, we reiterate our observation that in practice, when our VC dimension technique is used properly i.e., n is large enough relative to p, the σ used in Theorem 6 and the design points are adequately chosen, $\left|\hat{d}_VC - d_{VC}\right|$ has a well defined minimum at the true value of d_{VC}. Also, in contrast with other methods (including using shrinkage methods to nest models) our technique is generally more sensitive to over fit, thereby giving better parsimony.

4.1.2. Dependence of $\hat{d}_{CV}$ on n and N_L

It is no surprise that the higher n/p is the better the discrimination of $\hat{d}_{VC}$ over models is. However, our findings are more complicated because of the design points. The informal rule is that one wants $n \approx 10p$ for good parametric inference. However, this does not take into account model selection that often requires $n > 10p$. Indeed, for good model selection with d_{VC}, we find that $n \geq 15$ is usually sufficient, provided the design points are not badly chosen. In our examples choosing the n_l's uniformly over $[0, n]$ generally gives decent but not optimal performance and for typical ranges of model sizes $L \geq 5$ will suffice even though larger values of L are usually better, say $7 \leq L \leq 10$. Although we are unable at this point to chracterize the tradeoff between design points and sample size, we have noted that in some cases, good choice of design points can compensate for insufficient sample size. Indeed, relatively small changes in the n_l's can have a large numerical effect when n is small; possible due to instability in the nonlinear regression step, (29).

We leave the question of optimally choosing L and the n_l's as future work even though we make the following recommendations: 1) Good choices of n_l's are spread over $[0, n]$. 2. More n_l's should be in $[n/2, n]$ than in $[0, \frac{n}{2}]$, but neither should be empty. 3. Good choices for n_l's tend to remain good while poor choices of n_l's may not be as damaging to inference, as $n \to \infty$. 4. It is better to use fewer design points over a larger range than more design points over a smaller range. 5. As n increases the n_l's should shift upward, i.e., more in $[n/2, n]$. 6. If enough n_l's are large relative to p, $\hat{d}_{VC}$ may be accurate for smaller n, perhaps $n \geq 12p$ will suffice.

4.2. Analysis of a Benchmark Data Set

The goal of this section is to evaluate our method on a 'benchmark' data set Tour obtained from the Tour de France[1]. We start by giving some information about Tour, then in Sec.

1. Tour de France Data was compiled by B. Clarke. More information can be found at http://www.letour.fr/.

4.2.1 we analyze it using a model list based on two of the explanatory variables. The list is a sequence of models nested by SCAD. We evaluate our method by comparing $\hat{d}_{VC}$ to AIC, BIC, CV, $\widehat{PERM}_1$, and $\widehat{PERM}_2$. In Subsec. 4.2.2, we look at the effect of outliers in the estimates $\hat{d}_{VC}$, $\widehat{PERM}_1$, and $\widehat{PERM}_2$.

The full data set Tour has $n = 103$ data points. The data points are dependent (associated) because many cyclists competed in the Tour de France for more than one year. Here we ignore the dependence structure because the dependencies are small enough that the complication of accounting for them is not worthwhile. Each data point has a value of the response variable (Speed), the average speed in kilometers per hour (km/h) of the winner of the Tour. The explanatory variables are the Year (Y) of the Tour and its distance (D) in kilometers. Our data is from 1903 to 2016. However, during World Wars 1 and 2 there was no Tour, so we do not have data points for those years. The effect of World War I on the speed of the winner of the tour can be seen in a scatterplot of Speed vs. Y – the lowest winning speeds were in the years just after World War I, probably due to casualties. After World War II, there was also a decrease in average winning speed, but the decrease was less than after World War I. There is a curvilinear relationship between Speed and Year and a roughly linear relationship between Speed and D; the variability of Speed increases with D.

4.2.1. Analysis of Tour Using a Nested Collection of Models

We identify a nested model list using Y, D, Y^2, D^2 and the interaction between Year and Distance denoted $Y : D$ as covariates. Because the size of the data set is not large, we can only use a small model list.

We order the variables using SCAD because as a shrinkage method it perturbs parameter estimates the least and satisfies an oracle property. Under SCAD, the order of inclusion of variables is Y, D, D^2, Y^2, and $Y : D$. We therefore fit five different models. We use the six model selection techniques from Sec. 4.1 and include $\hat{d}_{VC}$ the original estimator in Vapnik et al. (1994) for the sake of comparison. It is seen that Vapnik et al. (1994)'s original

Model	$\hat{d}_V$	$\hat{d}_{VC}$	$\widehat{PERM}_1$	$\widehat{PERM}_2$	AIC	BIC	CV
Y	20	4	16.42	44.95	84	79.67	0.1294
$Y,\ D$	20	4	15.10	42.83	77	71.66	0.1209
$Y,\ D,\ D^2$	20	4	11.21	36.37	24	**26.21**	0.0727
$Y, D,\ D^2,\ Y^2$	20	4	11.09	36.16	**17**	28.77	**0.0681**
$Y, D,\ D^2,\ Y^2,\ Y : D$	20	4	**11.06**	**36.11**	19	32.96	0.0691

Table 1: Model selection for the Tour de France data using seven methods. The design points for $\hat{d}_V$ and $\hat{d}_{VC}$ are 20, 30, 40, 50, 60, 70, 80, 90, and 100 and $W = 50$. So $\hat{d}_V$ equals the smallest design point in all cases. This problem frequently occurs for $\hat{d}_V$.

method is helpful only if it is reasonable to surmise that there are 15 missing variables whereas our method uniquely identifies one of the models on the list. Even though there is likely no model for Tour de France data set that is accurate to infinite precision, our method is giving a useful result. Indeed, our method is choosing the fourth model list the same as

indicated by AIC and CV. The BIC drops the Y^2 term which is not unreasonable because the curvilinearity is less than quadratic. $\widehat{PERM}_1$ and $\widehat{PERM}_2$ include the interaction term (which can be seen to be zero by a simple t-test). It may also be the case that the derivation of the BIC rests heavily on the independent of data which is not the case here.

4.2.2. Analysis of The Tour de France data set with outliers removed

The observations just after World War I may be outliers. So, we consider the data set formed by deleting the points from 1919 to 1926. Let us see how the six model selection techniques now behave.

The process of analyzing this reduced data set is the same: We identify the nested model lists by SCAD and then find the models corresponding to $\hat{d}_{VC}$, AIC, BIC, CV, $\widehat{PERM}_1$, and $\widehat{PERM}_2$. The results are given in Table 2.

Model Size	$\widehat{d}_{VC}$	$\widehat{PERM}_1$	$\widehat{PERM}_2$	AIC	BIC	CV
Y	4	12.87	40.72	67	75	0.1181
Y, D^2	4	12.01	39.26	69	79	0.1336
Y, D^2, D	4	11.66	38.36	46	59	0.0919
Y, D^2, D, Y^2	**4**	11.48	38.34	**28**	**43**	0.0742
Y, D^2, D, Y^2, $Y:D$	4	**11.35**	**38.13**	29	46	**0.0735**

Table 2: Model selection for the Tour de France data set with outliers removed.

Under SCAD, the order of inclusion of our covariates is: Y, D^2, D, Y^2 and $Y:D$. This order is different from when we used all data points. The outliers suggest $Y:D$ was more important than it probably is. Note also that when we used all the data points, D was included before D^2 and D^2 was included after Y^2. Again, we fit 5 nested models.

From Table 2, if we choose a model using $\hat{d}_{VC}$, we get the same answer as in Sec. 4.2.1, the model with four variables: Y, D^2, D, Y^2. AIC and BIC choose the same model probably because $(Y:D)$ has low correlation with Speed (-0.08). $\widehat{PERM}_1$, $\widehat{PERM}_2$ and CV choose the model of size 5, which we discount as before because $Y:D$ is only slightly correlated with Speed. That is, the reasoning in Subsec. 4.2.1 for why we think that the model chosen by $\hat{d}_{VC}$ is best continues to hold.

4.3. Application to a Real Data Set

To demonstrate the use of our technique, we re-analyze the Wheat data set presented and studied in Campbell et al. (2003), Dilbirligi et al. (2006), and Dhungana et al. (2007) from a non-complexity based standpoint. The Wheat data set has 2912 observations. More information concerning the data set and the design structure can be found in Campbell et al. (2003). The response variable is YIELD (MG/ha), the covariates that we used are 1000 kernel weight (TKWT), kernels per spike (KPS), Spikes per square meter (SPSM), height of the plant (HT), test weight (TSTWT(KG/hl)), and kernels per square meter (KPSM). Often, in agronomic data sets, there are several classes of explanatory variables, here we have phenotypic, single nucleotide polymorphisms (SNP's) and the variables defining the design. We compare VC dimension based model selection to the methods used in the last subsection and verify our method gives good results.

Our collection of explanatory variables can be grouped into three categories: phenotype, SNP, and design variables. For brevity, we fit only the phenotype variables in Subsec. 4.3.1 and phenotype plus design variables in Subsec. 4.3.2. Comparing these will show that the model selection is unaffected by the design. Also, for brevity, our only analysis is 'multilocation' because we pooleded the data over location-year pairs. A full analysis is in Mpoudeu and Clarke (2018).

4.3.1. Estimation of VC Dimension Using Phenotypic Covariates Only

Intuition suggests

$$YIELD = \beta_0 + \beta_1 \cdot TKWT \cdot KPSM + \epsilon \tag{46}$$

will be a good model because YIELD is essentially the product of the number of kernels and their average weight. Likewise,

$$YIELD = \beta_0 + \beta_1 \cdot TKWT \cdot KPS \cdot SPSM + \epsilon \tag{47}$$

should also be a good model. So, using only phenotpic variables does not lead to a *unique* good model. Both are over simplifications and we can be confident that other influences on YIELD must be considered. Indeed, a 3-dimensional plot of the vectors (YIELD, TKWT, KPSM) looks like a triangle that is bowed out to one side. The bowing means that (46) is only an approximation; other terms are required to explain YIELD. Henceforth, we focus on (46) rather than (47) because we have limited ourselves to second order models.

To implement our multilocation analysis, we first find the order of inclusion of the phenotypic variables in the model, using correlation with YIELD. Then, we find values for $\hat{d}_{VC}$, AIC, BIC, CV, $\widehat{PERM}_1$, $\widehat{PERM}_2$, and the models given by SCAD and ALASSO. Under absolute value of correlation with YIELD, the order of inclusion of the explanatory variables is: TKWT · KPSM, TSTWT · KPSM, KPSM, SPSM · KPS, KPSM^2, KPSM · HT, TKWT · SPSM, TSTWT^2, TSTWT, SPSM · KPSM, KPS · KPSM, TSTWT · SPSM, SPSM SPSM · HT, SPSM^2, TKWT · TSTWT, TKWT, TSTWT · HT, TKWT^2 TKWT · KPS, TSTWT · KPS, TKWT · HT, KPS · HT, HT, HT^2 KPS, KPS^2. So, we consider 27 nested models. This leads to Table 3.

First, we note there is no variability in $\widehat{PERM}_1$ and $\widehat{PERM}_2$ so following standard usage, they select a model with TKWT · KPSM as the only explanatory variable. Likewise, AIC, BIC, and CV suggest the one term model. However, $\hat{d}_{VC} = 14$ means the VC dimension chooses the model with the first 14 terms from the ordered list. SCAD and ALASSO both give the one term model

$$\widehat{YIELD} = 3.43 + 1.12 \cdot TKWT \cdot KPSM, \tag{48}$$

the same as $\widehat{PERM}_1$, $\widehat{PERM}_2$, AIC, BIC and CV. Thus the only reasonable model is the one chosen by $\hat{d}_{VC}$.

4.3.2. Analysis of Wheat Using Phenotypic Data and the Design Structure

Our objective in this subsection is to take the design structure into account and see its impact on the values of the VC dimension and hence on the chosen model. As before, we implement a multilocation analysis. Including design variables forces us to use a more

Size	$\hat{d}_{VC}$	$\widehat{PERM}_1$	$\widehat{PERM}_2$	AIC	BIC	CV
1	13	**7**	**11**	**-9160**	**-9142**	**0.001801809**
2	14	7	11	-9159	-9136	0.001802107
3	13	7	11	-9159	-9130	0.001802470
4	13	7	11	-9159	-9124	0.001803912
5	13	7	11	-9159	-9118	0.001803933
6	14	7	11	-9157	-9110	0.001805230
7	14	7	11	-9156	-9103	0.001805349
8	14	7	11	-9158	-9099	0.001806966
9	14	7	11	-9156	-9091	0.001807176
10	14	7	11	-9154	-9084	0.001808455
11	13	7	11	-9153	-9076	0.001808248
12	13	7	11	-9151	-9069	0.001808966
13	14	7	11	-9150	-9061	0.001810670
14	**14**	7	11	-9148	-9054	0.001812884
15	13	7	11	-9147	-9047	0.001813791

Table 3: The column labeled size gives the number of coefficients for each linear model. The 2^{nd} through 7^{th} columns give the corresponding estimates for $\hat{d}_{VC}$, $\widehat{ERM}_1$, $\widehat{ERM}_2$, AIC, BIC, and CV for Wheat.

complicated bootstrap procedure that would otherwise be sufficient. Thus, to implement our method here, we perform a *restricted* bootstrap. Specifically, we bootstrap in each level of the design variable (incomplete block) so that each half data set has all levels of the design structure. We do this to maintain the design structure and its effects. To include phenotypic variables in the models, we use the same order of inclusion as in Subsec. 4.3.1.

The natural comparison is between Tables 3 and 4. Apart from random variation, they are identical. Moreover, the sparsity methods give exactly the same results in both settings. Thus the conclusions here are the same as in Subsec. 4.3.1: The design variables have no impact on model selection and $\hat{d}_{VC}$ gives the only plausible model.

5. Conclusions

A concise summary of the contributions in this paper is as follows. Sec. 2.1 presents the derivation of the objective function we used to estimate the VC dimension. It is essentially an upper bound on the expected difference between two losses that we have defined as Δ_m or Δ, where the m indicates the discretization of the loss function for a regression problem. In Subsec. 2.2 we give an estimator for d_{VC} that uses our upper bound, nonlinear regression treating sample sizes as design points, a data driven estimator of Δ_m, and an optimization over an arbitrary constant. While this sounds complex, in practice the computations can usually be done in minutes on a regular laptop. Even though we only have an upper bound on Δ_m, in Sec. 3 we are able to give conditions under which our estimator is consistent. This is circumstantial evidence that our upper bound is tight.

Size	$\hat{d}_{VC}$	$\widehat{PERM}_1$	$\widehat{PERM}_2$	AIC	BIC	CV
1	13	**7**	**11**	**-9160**	**-9142**	**0.001801809**
2	13	7	11	-9159	-9136	0.001802107
3	13	7	11	-9159	-9130	0.001802470
4	13	7	11	-9159	-9124	0.001803912
5	13	7	11	-9159	-9118	0.001803933
6	13	7	11	-9157	-9110	0.001805230
7	13	7	11	-9156	-9103	0.001805349
8	13	7	11	-9158	-9099	0.001806966
9	13	7	11	-9156	-9091	0.001807176
10	13	7	11	-9154	-9084	0.001808455
11	13	7	11	-9153	-9076	0.001808248
12	13	7	11	-9151	-9069	0.001808966
13	**13**	7	11	-9150	-9061	0.001810670
14	13	7	11	-9148	-9054	0.001812884

Table 4: The column labeled size gives the number of coefficients for each linear model. The 2^{nd} through 7^{th} columns give the corresponding estimates for $\hat{d}_{VC}$, $\widehat{PERM}_1$, $\widehat{PERM}_2$, AIC, BIC, and CV for multi-location analysis with design structure.

We have done an extensive comparison of our estimator of VC dimension as a model selection method with seven established model selection methods, namely two forms of empirical risk minimization, AIC, BIC, CV, and two sparsity criteria. Other examples can be found in Mpoudeu and Clarke (2018). We did this for the special case of linear models but the same reasoning can be used for any class of nonlinear models e.g., trees, for which the VC dimension can be identified. We also gave one example of how our estimator for VC dimension performs better than the original estimator in Vapnik et al. (1994). As a generality, our method equals or outperforms these other methods.

Acknowledgments

The authors gratefully acknowledge support from NSF grant # DMS-1419754 and invaluable computational support from the Holland Computing Center. Data, code, and results for the full versions of the analyses presented here can be found at `https://github.com/poudas1981/Wheat_data_set.` We also express our gratitude to a sophisticated and hard-working referee and Editor who helped us improve our work immensely.

References

M. Anthony and P. Bartlett. *Neural network learning: Theoretical foundations*. Cambridge University Press, 2009.

A. Barron and T. Cover. Minimum complexity density estimation. *IEEE Transactions on Information theory*, 37(4):1034–1054, 1991.

J. Bernardo and A. F. M. Smith. *Bayesian Theory*. Wiley Series in Probability and Statistics, 2000.

B. Campbell, P. Baenziger, P. Stephen, K. Gill, K. Eskridge, H. Budak, M. Erayman, I. Dweikat, and Y. Yen. Identification of qtls and environmental interactions associated with agronomic traits on chromosome 3a of wheat. *Crop Science*, 43:1493–1505, 2003.

P. Dhungana, K. Eskridge, P. Baenziger, B. Campbell, K. Gill, and I. Dweikat. Analysis of genotype-by-environment interaction in wheat using a structural equation model and chromosome substitution lines. *Crop Science*, 47:477–484, 2007.

M. Dilbirligi, M. Erayman, B. Campbell, H. Randhawa, P. Baenziger, I. Dweikat, and K. Gill. High-density mapping and comparative analysis of agronomically important traits on wheat chromosome 3a. *Genomics*, 88:74 – 87, 2006. ISSN 0888-7543. doi: https://doi.org/10.1016/j.ygeno.2006.02.001. URL `http://www.sciencedirect.com/science/article/pii/S0888754306000437`.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001. doi: 10.1198/016214501753382273. URL `http://dx.doi.org/10.1198/016214501753382273`.

J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

D. McDonald, C. Shalizi, and M. Schervish. Estimated VC dimension for risk bounds. preprint arXiv:1111.3404, 2011.

M. Mpoudeu. *Use of Vapnik-Chervonenkis Dimension in Model Selection*. PhD thesis, University of Nebraska-Lincoln, 2017. See: `https://arxiv.org/pdf/1808.06684.pdf`.

M. Mpoudeu and B. Clarke. Model selection via the VC dimension. *See: https://arxiv.org/abs/1808.05296*, 2018.

K. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

X. Shao, V. Cherkassky, and W. Li. Measuring the VC dimension using optimized experimental design. *Neural computation*, 12(8):1969–1986, 2000.

V. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Soviet Math. Dokl*, volume 9, pages 915–918, 1968.

V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies to their probabilities. In *Soviet Math. Dokl*, volume 9, pages 915–918, 1971.

V. Vapnik, E. Levin, and Y. LeCun. Measuring the VC dimension of a learning machine. *Neural Computation*, 6(5):851–876, 1994.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.