Significance Tests for Neural Networks

Enguerrand Horel   ENGUERRAND.HOREL@GMAIL.COM
Institute for Computational and Mathematical Engineering
Stanford University
Stanford, CA 94305, USA

Kay Giesecke   GIESECKE@STANFORD.EDU
Department of Management Science and Engineering
Stanford University
Stanford, CA 94305, USA

Editor: Moritz Hardt

Abstract

We develop a pivotal test to assess the statistical significance of the feature variables in a single-layer feedforward neural network regression model. We propose a gradient-based test statistic and study its asymptotics using nonparametric techniques. Under technical conditions, the limiting distribution is given by a mixture of chi-square distributions. The tests enable one to discern the impact of individual variables on the prediction of a neural network. The test statistic can be used to rank variables according to their influence. Simulation results illustrate the computational efficiency and the performance of the test. An empirical application to house price valuation highlights the behavior of the test using actual data.

Keywords: significance test, neural network, model interpretability, nonparametric regression, nonlinear regression, feature selection

1. Introduction

Neural networks underpin many of the best-performing artificial-intelligence systems, including speech recognizers on smartphones or Google's latest automatic translator. The tremendous success of these applications has spurred the interest in applying neural networks in a variety of other fields, including medicine, physics, economics, marketing, and finance. However, the difficulty of interpreting a neural network has slowed the practical implementation of neural networks in these fields, where the major stakeholders often insist on the explainability of predictive models.

This paper approaches the neural network explainability problem from the perspective of statistical significance testing. We develop and analyze a pivotal test to assess the statistical significance of the feature variables of a single-layer, fully connected feedforward neural network that models an unknown regression function.[1] We construct a gradient-based test statistic that represents a weighted average of the squared partial derivative of the neural network estimator with respect to a variable. The weights are prescribed by a

---

1. We consider a setting with structured data but the approach presented here could also be potentially used with images or text data.

positive measure that can be chosen freely. We study the large-sample asymptotic behavior of the test statistic. The neural network estimator is regarded as a nonparametric sieve estimator.[2] The dimension of the network (i.e., the number of hidden units) grows with the number of samples. Using an empirical-process approach, we first show that the large-sample asymptotic distribution of the rescaled neural network sieve estimator is given by the argmax of a Gaussian process. The Gaussian process is indexed by the function space that contains the unknown regression function. The rescaling is given by the estimation rate of the neural network identified by Chen and Shen (1998). A second-order functional delta method is then used to obtain the asymptotic distribution of the test statistic as the weighted average of the squared partial derivative of the argmax of the Gaussian process with respect to the variable of interest. The empirical test statistic, which results from a weight measure equal to the empirical law of the feature variables, has the same asymptotics. Moreover, under mild assumptions, the asymptotic distribution can be represented by a mixture of chi-square distributions. We develop several approaches to compute the limiting law; these and the performance of the test are illustrated in a simulation study. An empirical application to house price valuation in the United States complements our theoretical results and highlights the behavior of the test using actual data.

The significance test we propose has several desirable characteristics. First, the test procedure is computationally efficient. The partial derivative underpinning the test statistic is basically a byproduct of the widely used gradient-based fitting algorithms and is provided by standard software packages used for fitting neural networks (e.g., TensorFlow). Furthermore, the test procedure does not require re-fitting the neural network. Second, the test statistic can be used to rank order the variables according to their influence on the regression outcome. Third, the test is not susceptible to the non-identifiability of neural networks.[3] Forth, the test facilitates model-free inference: in the large-sample asymptotic regime, due to the universal approximation property of neural networks (Hornik et al. 1989), we are performing inference on the "ground-truth" regression function. Finally, the test statistic can be extended to higher-order and cross derivatives, facilitating tests of the statistical significance of nonlinear features such as interactions between variables.

The rest of the paper is organized as follows. The remainder of the introduction discusses the related literature. Section 2 discusses the model, hypotheses and test statistic. Section 3 analyzes the large-sample asymptotic distribution of the test statistic. Section 4 provides several approaches to compute the limiting distribution. Section 5 analyzes the performance of the test in a simulation study. Section 6 develops an empirical application to house prices valuation. Section 7 concludes. There are several appendices, one containing the proofs of our technical results.

1.1 Related Literature

There is a growing literature on explaining black-box machine learning models using feature importance analysis (also known as saliency mask or sensitivity analysis). Guidotti et al. (2018) provides a survey. Several papers develop procedures to explain the prediction of a model *locally* at a specific instance. Shrikumar et al. (2017) decomposes the neural network

---

2. See Chen (2007) for an overview.
3. See, for example, Chen et al. (1993) for a discussion of identifiability.

prediction for a specific input by backpropagating the contributions of all neurons in the network to every input feature. Kononenko et al. (2010) explains individual predictions of classification models using the game theoretic concept of Shapley value. Ribeiro et al. (2016) explains the predictions of a classifier by learning an interpretable model locally around the prediction. Baehrens et al. (2010) provides local explanation vectors for classification models by computing class probability gradients. These methods are limited to local explanations and cannot provide a global understanding of model behavior.

In contrast, a second group of methods aim for understanding how the model works overall. Datta et al. (2016) measures the influence of an input on a quantity of interest by computing the difference in the quantity of interest when the data are generated according to the real distribution and when they are generated according to a hypothetical distribution designed to disentangle the effect of different inputs. Sundararajan et al. (2017) attributes the prediction of a deep network to its input features by integrating gradients of the model along the straightline path from a baseline value of the input to its current value. Cortez and Embrechts (2011) assesses the importance of inputs through sensitivity analysis. While similar to our approach in terms of using network gradients, these methods focus on feature importance and do not treat statistical significance. We, in contrast, develop a procedure that establishes the statistical significance of input features. Once significance has been established, the magnitude of the test statistic can be used to judge relative feature importance for the set of significant features. This can then facilitate feature selection.

There is prior work treating the significance of variables in neural network regression models. Regarding neural networks as parametric formulations, Olden and Jackson (2002) propose a randomization test for the significance of the model's parameters. This is combined with the weight-based metric of feature importance introduced by Garson (1991) to test for the statistical significance of input variables. However, this metric is not necessarily identifiable due to the non-identifiability of neural networks.

The large-sample asymptotic results for the estimators of neural network parameters developed by White (1989a) and White (1989b) can be used to develop significance tests for the network weights and the variables. Zapranis and Refenes (1999) construct a test statistic for variable significance whose distribution is estimated by sampling from the asymptotic distribution of the parametric neural network estimator. White and Racine (2001) develop expressions for the asymptotic distributions of two test statistics for variable significance. A bootstrap procedure is used to estimate these distributions.

Likelihood ratio tests could also be used to assess variable significance in a parametric setting. One would fit an unconstrained model incorporating all variables and nested ones with restricted variables sets. If the unconstrained model is assumed to be correctly specified, then we have standard chi-square asymptotics. Theorem 7.2 from Vuong (1989) provides the asymptotic distribution under mis-specification, a weighted sum of chi-square distributions parametrized by the eigenvalues of a $p \times p$ matrix where p is the number of parameters of the network. While appealing, this approach tends to be computationally challenging because p is typically large and one needs to re-fit the neural network multiple times.

Alternatively, neural networks can be regarded as nonparametric models, and this is the approach we pursue. This approach has several advantages over the parametric setting considered by the aforementioned authors. First, viewing a neural network as a mapping

between input variables and regression output, identifiability is no longer an issue. Moreover, by not restricting the dimension of the network (i.e., the number of hidden nodes) a priori, we can exploit the consistency of neural networks (White 1990) which is due to their universal approximation property. Model misspecification is no longer an issue.

Nonparametric variable significance tests fall in two categories: goodness-of-fit tests or conditional moment tests; see Section 6.3 of Henderson and Parmeter (2015) for a review. Goodness-of-fit tests are based on the residuals from an unrestricted estimator fitted using all variables and the residuals from a restricted estimator fitted using all variables except the one being tested. The required distributions are either estimated using the bootstrap, or approximated asymptotically for the specific case of kernel regressions (see, for example, Gozalo 1993, Lavergne and Vuong 1996, Yatchew 1992 and Fan and Jiang 2005). Because the tests require the estimation of a potentially large number of models, they can be computationally challenging, especially if employing the bootstrap. Conditional moment tests rely on the observation that the expectation of the residual of the restricted estimator conditional on all variables is zero under the null hypothesis of insignificance of the variables being tested. Fan and Li (1996), Lavergne and Vuong (2000), Gu et al. (2007), and others study these tests in the context of kernel regressions.

Racine (1997) proposes, again for kernel regressions, a nonparametric variable significance test based on the partial derivative of the fitted regression function with respect to a variable. The distribution of the test statistic is estimated using the bootstrap. Our test is also based on partial derivatives but we analyze a neural network regression rather than a kernel regression. Unlike Racine (1997), we study the asymptotic distribution of the test statistic in order to avoid the use of the bootstrap, which tends to be computationally expensive in the context of neural network models. The large data sets that are often used in practice to estimate neural network regressions justify the use of the asymptotic distribution for the purpose of significance testing.

2. Model, Hypotheses, and Test Statistic

Fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and consider the following regression model:

$$Y = f_0(X) + \epsilon. \tag{1}$$

Here, $Y \in \mathbb{R}$ is the dependent variable, $X \in \mathcal{X} \subset \mathbb{R}^d$ is a vector of regressor or feature variables with distribution P for some $d \geq 3$, and $f_0 : \mathcal{X} \to \mathbb{R}$ is an unknown deterministic regression function that is continuously differentiable. The error variable ϵ satisfies the usual assumptions: $\epsilon \perp\!\!\!\perp X$, $\mathbb{E}(\epsilon) = 0$, and $\mathbb{E}(\epsilon^2) = \sigma^2 < \infty$.

We are interested in assessing the influence of a variable X_j on Y. To this end, it is natural to consider the partial derivative of $f_0(x)$ with respect to x_j. If $\frac{\partial f_0(x)}{\partial x_j} = 0$ for all $x \in \mathcal{X}$, then the jth variable does not have an impact on Y. The significance test we propose is based on the weighted average of the partial derivative, with weights defined by a positive measure μ. Specifically, we will consider the following hypotheses

$$H_0 : \lambda_j = \int_{\mathcal{X}} \Big(\frac{\partial f_0(x)}{\partial x_j}\Big)^2 d\mu(x) = 0, \tag{2}$$

$$H_A : \lambda_j \neq 0. \tag{3}$$

We take the square of the partial derivative to avoid compensation of positive and negatives values, ensure differentiability, and help discriminate between large and small values. Examples of the weight measure μ include uniform weights, $d\mu(x) = dx$, uniform weights over subsets $I \subset \mathcal{X}$, $d\mu(x) = 1_{\{x \in I\}}dx$, and the choice $\mu = P$, where P is the distribution of X. Under the latter choice, λ_j takes the form $\lambda_j = \mathbb{E}[(\frac{\partial f_0(x)}{\partial x_j})^2]$.

If the regression function f_0 is assumed to be linear, i.e., $f_0(x) = \sum_{k=1}^{d} \beta_k x_k$, then $\lambda_j \propto \beta_j^2$ and the null takes the form $H_0 : \beta_j = 0$. This hypothesis can be tested using a standard t-test. In the general non-linear case, the derivative $\frac{\partial f_0(x)}{\partial x_j}$ is not flat but depends on x. To construct the null in this case, we take a weighted average of the squared values of the derivative, with μ prescribing the weights over the feature space $\mathcal{X}$.

We study the case where the regression function f_0 is modeled by a fully-connected, single-layer feed-forward neural network f, which is specified by a bounded activation function ψ on $\mathbb{R}$ and the number of hidden units K,

$$f(x) = b_0 + \sum_{k=1}^{K} b_k \psi(a_{0,k} + a_k^\top x), \tag{4}$$

where $b_0, b_k, a_{0,k} \in \mathbb{R}$ and $a_k \in \mathbb{R}^d$ are parameters to be estimated. Functions of the form (4) are dense in $C(\mathcal{X})$. That is, they are universal approximators: choosing the dimension K of the network large enough, f can approximate f_0 to any given precision.

Let f_n be an estimator of f based on n i.i.d. samples of (Y, X). The neural network test statistic λ_j^n associated with (2) takes the form

$$\lambda_j^n = \int_{\mathcal{X}} \left(\frac{\partial f_n(x)}{\partial x_j}\right)^2 d\mu(x). \tag{5}$$

Below, we are going to study the asymptotic distribution of λ_j^n as $n \to \infty$ under the assumption that the dimension $K = K_n$ of the neural network grows with n. The asymptotic distribution will be used to construct a test for the null in (2). This approach is typically less expensive computationally than a bootstrap approach. Note that in the asymptotic regime, due to the universal approximation property, we are actually performing inference on the "ground truth" f_0 (i.e., model-free inference).

3. Asymptotic Analysis

The neural network test statistic takes the form $\lambda_j^n = \phi[f_n]$, where ϕ is the functional

$$\phi[h] = \int_{\mathcal{X}} \left(\frac{\partial h(x)}{\partial x_j}\right)^2 d\mu(x). \tag{6}$$

Thus, we first study the asymptotic distribution of f_n, and then infer the asymptotic distribution of λ_j^n under H_0 from that of f_n using a functional delta method.

3.1 Asymptotics of Neural Network Estimator

We consider f_n as a nonparametric estimator. This estimator asymptotically approximates the unknown function f_0 which is assumed to belong to an infinite-dimensional function

space Θ. Since infinite-dimensional estimation does not necessarily lead to a well-posed optimization problem, we construct f_n as a sieve estimator. Intuitively, the method of sieves entails the optimization of an empirical criterion over a sequence of approximating spaces called sieves which are less complex than but dense in the original infinite-dimensional space. This approach renders the resulting optimization problem well-posed. More details can be found in Chen (2007).

The unknown regression function f_0 is assumed to belong to the function space

$$\Theta = \left\{ f \in \mathcal{C}^1, f : \mathcal{X} \subset \mathbb{R}^d \to \mathbb{R}, ||f||_{\lfloor \frac{d}{2} \rfloor + 2} \leq B \right\}, \tag{7}$$

where B is a constant, the feature space $\mathcal{X}$ is a hypercube of dimension d and

$$||f||_k = \sqrt{\sum_{0 \leq |\alpha| \leq k} \mathbb{E}_X \left(\nabla^\alpha f(X)^2 \right)}. \tag{8}$$

We construct the neural network estimator f_n as a sieved M-estimator. Given n i.i.d. samples $(Y_i, X_i)_{i=1}^n$ of the output-input pair (Y, X) and a nested sequence of sieve subsets $\Theta_n \subseteq \Theta$ such that $\cup_{n \in \mathbb{N}} \Theta_n$ is dense in Θ, the neural network estimator f_n is defined as the approximate maximizer of the empirical criterion function

$$L_n(g) = \frac{1}{n} \sum_{i=1}^n l(Y_i, X_i, g), \quad g \in \Theta, \tag{9}$$

with loss function $l : \mathbb{R} \times \mathcal{X} \times \Theta \to \mathbb{R}$ over the sieve space Θ_n, i.e.,

$$L_n(f_n) \geq \sup_{g \in \Theta_n} L_n(g) - o_P(1). \tag{10}$$

In our setting, a sieve subset Θ_n is generated by a neural network whose dimension $K = K_n$ depends on n and whose activation function ψ is a Lipschitz sigmoid function, i.e., a bounded measurable function on $\mathbb{R}$ such that $\psi(x) \to 1$ as $x \to \infty$ and $\psi(x) \to 0$ as $x \to -\infty$,

$$\Theta_n = \left\{ f \in \Theta : f(x) = b_0 + \sum_{k=1}^{K_n} b_j \psi(a_{0,k} + a_k^\top x), \sum_{k=0}^{K_n} |b_k| \leq c, \max_{1 \leq k \leq K_n} \sum_{i=0}^d |a_{i,k}| \leq \tilde{c} \right\}. \tag{11}$$

The upper bound on the norm of the weights b_j's allows us to view Θ_n as the symmetric convex hull of the class of sigmoid functions $\{\psi(a^\top \cdot + a_0), a \in \mathbb{R}^d, a_0 \in \mathbb{R}\}$. It hence ensures that the integral of the metric entropy of this class of functions is bounded for every n due to Theorem 2.6.9 of Van der Vaart and Wellner (1996). This property is necessary for Theorem 1 below and the reason why we restrict our analysis to a single-hidden layer network with sigmoid activation function. Indeed, the most recent bounds on the metric entropy of more general deep neural networks (Barron and Klusowski 2019) violate this property. Cybenko (1989) proves the denseness of $\cup_{n \in \mathbb{N}} \Theta_n$ in Θ.

Theorem 1 (Asymptotic distribution of neural network estimator) *Assume that*

   1. The law P of the feature vector X is absolutely continuous with respect to Lebesgue measure with bounded and strictly positive density;

2. *The loss function* $l(y, x, g) = -\frac{1}{2}(y - g(x))^2$;

3. *The dimension* K_n *of the neural network satisfies* $K_n^{2+1/d} \log K_n = O(n)$;

4. *Letting* $l_f(X, \epsilon) = 2(f - f_0)(X)\epsilon - (f_0 - f)^2(X)$ *where* f_0 *is the regression function and* ϵ *the regression error in* (1), *it holds that*

$$\frac{r_n}{\sqrt{n}} \sum_{i=1}^{n} \left(l_{f_n}(X_i, \epsilon_i) - \mathbb{E}[l_{f_n}(X, \epsilon)] \right)$$

$$\geq \sup_{h \in \Theta} \frac{r_n}{\sqrt{n}} \sum_{i=1}^{n} \left(l_{f_0 + \frac{h}{r_n}}(X_i, \epsilon_i) - \mathbb{E}[l_{f_0 + \frac{h}{r_n}}(X, \epsilon)] \right) - o_P(1)$$

where for all $n \geq 2$

$$r_n = \left(\frac{n}{\log n} \right)^{\frac{d+1}{2(2d+1)}}. \tag{12}$$

Then f_n *converges weakly in the metric space* (Θ, d) *with* $d(f, g) = \mathbb{E}[(f - g)^2]$:

$$r_n(f_n - f_0) \Longrightarrow h^\star, \tag{13}$$

where $h^\star$ *is the argmax of the Gaussian process* $\{\mathbb{G}_f : f \in \Theta\}$ *with mean zero and covariance function* $Cov(\mathbb{G}_s, \mathbb{G}_t) = 4\sigma^2 \mathbb{E}_X(s(X)t(X))$.

Theorem 1 derives the asymptotic distribution of the neural network estimator in a random function sense using the theory of empirical processes (Van der Vaart and Wellner 1996). It complements several other convergence results for neural network estimators, including the consistency result of White (1990) and results on the estimation rate and asymptotic normality of functionals of neural network estimators in Shen and Wong (1994), Shen (1997), Chen and Shen (1998) and Chen and White (1999).

3.2 Asymptotic Distribution of Test Statistic

Using Theorem 1, we can now derive the distribution of the test statistic under the null hypothesis using a functional delta method. A second order approach is needed since a first-order method would lead to a degenerate limiting distribution under the null.

Theorem 2 (Asymptotic distribution of the test statistic) *Under the conditions of Theorem 1 and the null hypothesis* H_0, *we obtain:*

$$r_n^2 \lambda_j^n \Longrightarrow \int_{\mathcal{X}} \left(\frac{\partial h^\star(x)}{\partial x_j} \right)^2 d\mu(x). \tag{14}$$

In the case where the weight measure μ is equal to the law P of the feature vector X, the test statistic takes the form

$$\lambda_j^n = \int_{\mathcal{X}} \left(\frac{\partial f_n(x)}{\partial x_j} \right)^2 dP(x) = \mathbb{E}_X \left[\left(\frac{\partial f_n(X)}{\partial x_j} \right)^2 \right]. \tag{15}$$

In this case it is natural to estimate the expectation in (15) using the empirical measure P_n of the samples $\{X_i\}_{i=1}^n$,

$$\hat{\lambda}_j^n = \int_{\mathcal{X}} \left(\frac{\partial f_n(x)}{\partial x_j}\right)^2 dP_n(x) = n^{-1} \sum_{i=1}^n \left(\frac{\partial f_n(X_i)}{\partial x_j}\right)^2. \tag{16}$$

The following result shows that the asymptotic distribution of the empirical test statistic $\hat{\lambda}_j^n$ is the same as that of λ_j^n (when $\mu = P$).

Proposition 3 (Asymptotic distribution of the empirical test statistic) *Under the conditions of Theorem 2, we have*

$$r_n^2 n^{-1} \sum_{i=1}^n \left(\frac{\partial f_n(X_i)}{\partial x_j}\right)^2 \implies \mathbb{E}_X\left[\left(\frac{\partial h^\star(X)}{\partial x_j}\right)^2\right] = \int_{\mathcal{X}} \left(\frac{\partial h^\star(x)}{\partial x_j}\right)^2 dP(x).$$

4. Computing the Asymptotic Distribution

The limiting distribution of the test statistic in Theorem 2 is a functional of the argmax of a Gaussian process over a function space. We discuss two approaches to computing this distribution in order to implement a test. One uses a series representation and the other a discretization method. Section 5 numerically evaluates these approaches.

4.1 Series Representation

We can exploit the fact that Θ is a subspace of the Hilbert space $L^2(P) = \{f : \mathcal{X} \to \mathbb{R} : \|f\|_{L^2(P)} < \infty\}$ which admits an orthonormal basis $\{\phi_i\}_{i=0}^\infty$. Assume that this basis is C^1 and is stable under differentiation. This means that $\forall i \in \mathbb{N}, 1 \le j \le d$, there exists an $\alpha_{i,j} \in \mathbb{R}$ and a mapping $k : \mathbb{N} \to \mathbb{N}$ such that

$$\frac{\partial \phi_i}{\partial x_j} = \alpha_{i,j} \phi_{k(i)}. \tag{17}$$

This, in turn, implies that there exists an invertible operator D which is diagonal with respect to the basis such that

$$\|f\|_{k,2}^2 = \|Df\|_{L^2(P)}^2 = \sum_{i=0}^\infty d_i^2 \langle f, \phi_i \rangle_{L^2(P)}^2, \tag{18}$$

where the d_i's are certain functions of the $\alpha_{i,j}$'s that satisfy

$$\sum_{i=0}^\infty \frac{1}{d_i^2} < \infty, \quad \sum_{i=0}^\infty \frac{\alpha_{i,j}^2}{d_i^4} < \infty. \tag{19}$$

Theorem 4 (Series representation of the asymptotic distribution) *Assume the conditions of Theorem 2 hold and that the weight measure $\mu = P$. If the orthonormal basis of $L^2(P)$ satisfies (17), then the asymptotic distribution of the test statistic satisfies*

$$\mathbb{E}_X\left[\left(\frac{\partial h^\star(X)}{\partial x_j}\right)^2\right] = \frac{B^2}{\sum_{i=0}^\infty \frac{\chi_i^2}{d_i^2}} \sum_{i=0}^\infty \frac{\alpha_{i,j}^2}{d_i^4} \chi_i^2,$$

where $\{\chi_i^2\}_{i\in\mathbb{N}}$ are i.i.d. samples from the chi-square distribution.

8

To provide a concrete example, we consider the case where the elements of the feature vector X are mutually independent and distributed uniformly on $[-1, 1]$. The density p of P takes the form

$$p(x) = \frac{1}{2^d}. \tag{20}$$

We now use the Fourier basis, which is an orthonormal basis of C^∞ functions for both $L_2([-1,1]^d)$ and $\mathbb{H}^{\lfloor \frac{d}{2} \rfloor + 2}([-1,1]^d)$ with

$$\|f\|^2_{\mathbb{H}^{\lfloor \frac{d}{2} \rfloor + 2}} = 2^d \sum_{n \in \mathbb{N}^d} \left[\sum_{|\alpha| \leq \lfloor \frac{d}{2} \rfloor + 2} \prod_{k=1}^{d} \gamma_{n_k}^{\alpha_k} \right] |f_n|^2,$$

where $\gamma_n = n^2 \pi^2$, $\phi_n^{[0]} = \cos(n\pi x)$ and $\phi_n^{[1]} = \sin(n\pi x)$. We define the weights d_n^2 as $d_n^2 = \sum_{|\alpha| \leq \lfloor \frac{d}{2} \rfloor + 2} \prod_{k=1}^{d} \gamma_{n_k}^{\alpha_k}$. This implies that

$$h_\omega^\star = \sum_{i \in \{0,1\}^d} \sum_{n \in \mathbb{N}^d} \frac{B}{\|D^{-1}\phi_\xi\|} \frac{\xi_n^{[i]}}{d_n^{[i]2}} \phi_n^{[i]},$$

where $\|D^{-1}\phi_\xi\|^2 = 4\sigma^2 \sum_{i \in \{0,1\}^d} \sum_{n \in \mathbb{N}^d} \frac{\chi_n^{[i]2}}{d_n^2}$ and $\{\chi_n^{[i]2}\}_{n \in \mathbb{N}^d, i \in \{0,1\}^d}$ are i.i.d. samples from a chi-square distribution. It follows that

$$\mathbb{E}_X \left[\left(\frac{\partial h^\star(X)}{\partial x_j} \right)^2 \right] = \left\| \frac{\partial h^\star(X)}{\partial x_j} \right\|^2_{L^2(P)} = \frac{B^2}{\sum_{i \in \{0,1\}^d} \sum_{n \in \mathbb{N}^d} \frac{\chi_n^{[i]2}}{d_n^2}} \sum_{i \in \{0,1\}^d} \sum_{n \in \mathbb{N}^d} \frac{\gamma_{n_j}}{d_n^4} \chi_n^{[i]2}. \tag{21}$$

The implementation of the test requires computing a quantile of the limiting distribution. We can approximate the limiting distribution by truncating the infinite sum in Theorem 4 at some order N, which can be chosen to achieve some given degree of accuracy. By sampling from a chi-squared distribution, we generate m samples $\{Z_i^N\}_{i=1}^m$ from an approximate asymptotic distribution, F^N. If we let m grow along with N, i.e., if we consider a sequence m_N such that $m_N \to \infty$ as $N \to \infty$, we obtain the consistency of the empirical approximate distribution, denoted $\mathbb{F}_{m_N}^N$,

$$\mathbb{F}_{m_N}^N(x) \xrightarrow[N \to \infty]{P} F(x), \quad \forall x \in \mathbb{R}. \tag{22}$$

This follows from a triangular array weak law of large number. Define the empirical approximate quantile function $\mathbb{F}_{m_N, N}^{-1}$ as

$$\mathbb{F}_{m_N, N}^{-1}(\alpha) = Z_{m_N(i)}^N, \quad \text{for } \alpha \in \left(\frac{i-1}{m_N}, \frac{i}{m_N} \right], \tag{23}$$

where $Z_{m_N(1)}^N, ..., Z_{m_N(n)}^N$ are the order statistics of $\{Z_i^N\}_{i=1}^{m_N}$. Then, from the previous result and Lemma 21.2 of Van der Vaart (2000), we get the consistency of the empirical approximate quantile function:

$$\mathbb{F}_{m_N, N}^{-1}(\alpha) \xrightarrow[N \to \infty]{P} F^{-1}(\alpha), \quad \forall 0 \leq \alpha \leq 1. \tag{24}$$

Thanks to this consistent estimator, by letting the truncation order N go to infinity as $n \to \infty$, we can reject the null if the empirical test statistic $\hat{\lambda}_n > \mathbb{F}_{m_N,N}^{-1}(1 - \alpha)$ such that the test will be asymptotically of level $\alpha \in [0, 1]$

$$\mathbb{P}_{H_0}\big(\hat{\lambda}_n > \mathbb{F}_{m_N,N}^{-1}(1 - \alpha)\big) \leq \alpha.$$

The truncation order N can be chosen so as to achieve a given approximation error ϵ. We illustrate this for the choice (20) discussed above. From Fourier approximation theory, we know that given the full index set of level N, $I_N = \{n \in \mathbb{N}^d : \max_{1 \leq j \leq d} n_j \leq N - 1\}$, the approximation rate of the Fourier series of order N for any $f \in \Theta$ is

$$\left\| f - \sum_{i \in \{0,1\}^d} \sum_{n \in I_N} \langle f, \phi_n^{[i]} \rangle \phi_n^{[i]} \right\|_{L^2} \leq O\left(\frac{1}{N^{\lfloor \frac{d}{2} \rfloor + 2}} \right). \tag{25}$$

This entails that $O\left(\frac{1}{\epsilon}\right)^{\frac{d}{\lfloor \frac{d}{2} \rfloor + 2}}$ Fourier basis functions will be necessary to reach a given L^2 approximation error ϵ.

4.2 Discretization Approach

The limiting distribution of the test statistic is a functional of the argmax $h^\star$ of a Gaussian process indexed by the function space Θ. The argmax is defined as a random function such that $\forall \omega \in \Omega$, $h_\omega^\star$ satisfies $\mathbb{G}_{h_\omega^\star}(\omega) \geq \mathbb{G}_f(\omega)$, $\forall f \in \Theta$. Given an ϵ-cover $\{f_1, ..., f_C\}$ of the function space Θ of size C, a discrete approximation of the paths of the process can be obtained by sampling from a multivariate Gaussian variable of dimension C with mean 0 and covariance matrix whose elements are $\mathbb{E}_X(f_i(X)f_j(X))_{i,j=1}^C$. For every ω corresponding to a sample from this multivariate Gaussian, the argmax among the ϵ-cover can be identified and by applying the functional (6) to it, this generates an approximate sample from the limiting distribution. The accuracy is controlled by ϵ.

Computing an explicit ϵ-cover of the function space Θ can be challenging. A simple approach to approximate this cover uses random functions sampled from it. More specifically, we generate random neural networks (4) by sampling the network parameter. By generating a large enough number of networks, we increase the likelihood of covering Θ.

5. Simulation Experiments

This section provides numerical results that illustrate our theoretical results and the properties of the significance test.

5.1 Data-Generating Process

We consider a vector X of eight feature variables satisfying

$$X = (X_1, \ldots, X_8) \sim \mathcal{U}(-1, 1)^8.$$

We consider the following data generating process,

$$Y = 8 + X_1^2 + X_2 X_3 + \cos(X_4) + \exp(X_5 X_6) + 0.1 X_7 + \epsilon, \tag{26}$$

Model	Test MSE
Neural Network	$3.1 \cdot 10^{-4}$
Linear	0.35

Table 1: Test MSE for the neural network and an alternative linear model.

Variable	Test Statistic	Leave-One-Out Metric
X_1	1.31	$8.94 \cdot 10^{-2}$
X_2	0.332	$1.12 \cdot 10^{-1}$
X_3	0.331	$1.12 \cdot 10^{-1}$
X_4	0.267	$2.09 \cdot 10^{-2}$
X_5	0.480	$1.30 \cdot 10^{-1}$
X_6	0.479	$1.30 \cdot 10^{-1}$
X_7	$1.01 \cdot 10^{-2}$	$3.50 \cdot 10^{-3}$
X_8	$4.20 \cdot 10^{-6}$	$3.46 \cdot 10^{-6}$

Table 2: Values of empirical test statistic $\hat{\lambda}_j^n$ and leave-one-out metric.

where $\epsilon \sim N(0, \sigma^2)$ and $\sigma = 0.01$. The variable X_8 has no influence on Y and is hence irrelevant. We generate a training set of $n = 100,000$ independent samples and validation and testing sets of $10,000$ independent samples each.

5.2 Fitting a Neural Network

We fit a fully-connected feed-forward neural network with one hidden layer and sigmoid activation function to the training set using the TensorFlow package. We employ the Adam stochastic optimization method with step size 0.001 and exponential decay rates for the moment estimates of 0.9 and 0.999. We use a batch size of 32, a maximum number of 150 epochs and an early stopping criterion that stops the training when the validation error has not decreased by more than 10^{-5} for at least 5 epochs. The number of hidden nodes is chosen so as to minimize the validation loss. A network with 25 hidden units performs best.

Table 1 compares the neural network's mean square error (MSE) for the test set with the test MSE for a linear model fitted to the same training data set. The neural network's test MSE is of the order of the regression noise. The linear model's test MSE is three orders of magnitude larger.

5.3 Test Statistic

For each variable $1 \leq j \leq 8$, we compute the empirical test statistic $\hat{\lambda}_j^n$ in (16) using the gradients function of TensorFlow over the training set. Table 2 reports the results. The statistic properly discriminates the irrelevant variable X_8 from the relevant variables. We also note that the statistic is similar for the variables that have a symmetric influence, namely (X_2, X_3) and (X_5, X_6). Moreover, for the linear variable X_7, the statistic $\hat{\lambda}_7^n = 0.01 = 0.1^2$, showing that the statistic correctly identifies the coefficient of X_7, which is 0.1. We note that the test statistic can be used to rank order variables according to their

influence. The greater the statistic the greater the influence. Thus the quadratic variable X_1 is identified as the most important variable.

For comparison, Table 2 also reports the leave-one-out metric, which is widely used in practice to assess variable influence. The leave-one-out metric for variable X_j is the difference between the loss of the model based on all variables except X_j and the loss of the model based on all variables. It is generally positive; the larger its value the larger one deems the influence of a variable. The results in Table 2 suggest that the leave-one-out metric shares the symmetry properties with our test statistic. There are important differences, however. For example, relative to our test statistic, the leave-one-out metric discounts the influence of the quadratic variable X_1. It identifies the exponential variables X_5 and X_6 as the most important variables. Moreover, the calculation of the leave-one-out metric requires re-fitting the model d times, where d is the dimension of X. This can be computationally expensive for models with many features. The calculation of our test statistic does not require re-fitting the model. Perhaps most importantly, our test statistic is supported by a rigorous test. To our knowledge, no test has been provided for the leave-one-out metric.

5.4 Estimation of Quantile

We choose a confidence level $\alpha = 0.05$ and compute the empirical quantile $\mathbb{F}_{m_N,N}^{-1}(1-\alpha)$ using the series representation approach described in Section 4.1. We use the Fourier basis, choose a truncation order $N = 4$, and sample $m_N = 10,000$ observations from the corresponding approximate asymptotic distribution.

We also need to choose the constant B that uniformly bounds the Sobolev norm of the functions in Θ as defined in (7). The asymptotic distribution in (21) is scaled by B. We need to choose B large enough so that the unknown regression function f_0 belongs to Θ, which translates to $\|f_0\|_{\lfloor \frac{d}{2} \rfloor + 2} \leq B$. But because a B that is too large would result in a test with smaller power, the optimal value of B is the norm of f_0. One could estimate this norm using Monte Carlo simulation or numerical integration, but because this may be computationally expensive, we follow the following alternative approach.

We include several uniformly distributed auxiliary noise variables in X and fit the neural network using both the original variables as well as the noise variables. The test statistics for the noise variables have a limiting distribution that includes the B factor. We also estimate the quantile of the asymptotic distribution as described in Section 4.1 by sampling from the limiting distribution with value $B = 1$, hence these are unscaled samples. An estimator of B can be obtained by computing the ratio of the mean of the test statistics of the noise variables to the mean of the unscaled samples. Because the number of noise variables is relatively small, this estimator can be noisy. Therefore, we prefer to use the maximum value of the test statistics of the noise variables (rather than the average) to increase the likelihood that $\|f_0\|_{\lfloor \frac{d}{2} \rfloor + 2} \leq B$.

5.5 Results

We estimate the power and size of the test by performing it on 250 alternative data sets $(Y_i, X_i)_{i=1}^n$ generated from the model (26). The second column of Table 3 reports the results. The power of the test is ideal (i.e., there is no Type 2 error). The size (Type 1 error) is

Variable	NN Test (S)	NN Test (D)	t-Test
X_1	1	1	0.07
X_2	1	1	0.07
X_3	1	1	0.06
X_4	1	1	0.05
X_5	1	1	0.10
X_6	1	1	0.07
X_7	1	1	1
X_8	0.35	0	0.04

Table 3: Power and size of significance tests at 5% level. The NN Test (S) column reports the values for our test when the series method described in Section 4.1 is used to compute the asymptotic distribution of the test statistic. The NN Test (D) column reports the values for our test when the discretization method described in Section 4.2 is used to compute the asymptotic distribution. The t-Test column reports the values for a standard t-test in an alternative linear regression model.

0.35, larger than the $\alpha = 0.05$ level of the test. The approximate asymptotic distribution on which the test is based might underestimate the variance of the finite sample statistic.

The last column of Table 3 also reports power and size of a t-test for an alternative linear regression model. Only the linear variable X_7 is identified as significant. These results illustrate how model misspecification can hurt inference.

5.6 Correlated Feature Variables

In practice, the elements of the feature vector X may be correlated. We study the robustness of the test in this regard. We generate correlated variables X_i whose marginal distribution is still $\mathcal{U}(-1, 1)$ but whose correlation structure is controlled by a Gaussian copula. We first generate samples from a multivariate normal distribution $(Z_1, \ldots, Z_8) \sim \mathcal{N}(0, \Sigma)$ with covariance matrix Σ given by

$$
\Sigma = \begin{bmatrix}
1 & 0.1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0.3 & 0 \\
0 & 0 & 0 & 0 & 1 & 0.5 & 0 & 0 \\
0 & 0 & 0 & 0 & 0.5 & 1 & 0 & 0 \\
0 & 0 & 0 & 0.3 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix},
$$

implying that X_1 and X_2, X_5 and X_6 as well as X_4 and X_7 are correlated. We then apply the standard normal distribution function on each coordinate to obtain the corresponding correlated uniform variables. We generate new data sets and re-fit the model.

Variable	NN Test (D)
X_1	1
X_2	1
X_3	1
X_4	1
X_5	1
X_6	1
X_7	1
X_8	0.004

Table 4: Power and size of significance tests at 5% level when the feature variables are correlated. The discretization method described in Section 4.2 was used to compute the asymptotic distribution of the test statistic.

Table 4 reports the size and power of the test in this situation (as above, 250 alternative data sets were used). The size and power of the test do not deteriorate, suggesting that the test performs well in the presence of correlated feature variables.

5.7 Discretization Approach

We study the performance of the significance test when using the discretization approach in Section 4.2 to estimate the asymptotic distribution. This approach approximates the argmax of the Gaussian process by using random functions sampled from Θ. More specifically, we randomly sample 500 functions using fully-connected, single-layer feed-forward neural networks with sigmoid activation function and 25 hidden units. The parameters are sampled from a Glorot normal distribution, which is a truncated normal distribution centered at 0 with standard deviation $\sqrt{2/(in + out)}$, where in is the number of input nodes and out is the number of output nodes of the layer for which the parameters are sampled (see Glorot and Bengio (2010)).

To generate a sample from the asymptotic distribution, we first generate a sample from a multivariate normal distribution of dimension 500 with mean vector 0 and covariance matrix that we approximate by the diagonal matrix with diagonal elements equal to $\sigma_j^2 = \frac{1}{n}\sum_{i=1}^{n} f_j(X_i)^2$. The value σ_j^2 is the empirical expectation of the square of the function that we estimate. We extract the index of the maximum value from this multivariate normal, which correspond to the approximate argmax of the Gaussian process. Given this approximate argmax function, we finally generate an approximate sample from the asymptotic distribution of the test statistic. We repeat this process 10,000 times in order to estimate the quantile at the 5% level.

The third column of Table 3 reports the power and size of the test based on this approach (as before, we use 250 data sets to estimate the power and size). Our significance test has perfect power and size.

Variable Name	Test Statistic
Last_Sale_Amount	1.640
Tax_Land_Square_Footage	1.615
Sale_Month_No	1.340
Tax_Amount	0.3834
Last_Mortgage_Amount	0.1040
Tax_Assd_Total_Value	0.0812
Tax_Improvement_Value_Calc	0.0721
Tax_Land_Value_Calc	0.0690
Year_Built	0.0681
SqFt	0.0566

Table 5: Test statistic for the top 10 significant (5%) feature variables (out of 68).

6. Empirical Application: House Price Valuation

We use the significance test to study the variables influencing house prices in the United States. We analyze a data set of 76,247 housing transactions in California's Merced County between 1970 and 2017. The data are obtained from the county's registrar of deed office through the data vendor CoreLogic. The dependent variable Y of the regression is the logarithm of the sale price. The 68 elements of the feature vector X are listed in Appendix B along with descriptions. They include house characteristics, mortgage characteristics, tax characteristics, local and national economic conditions, and other types of variables. Every variable is centered and scaled to unit variance. 70% of the data is used for training, 20% for validation, and the remainder is used for testing.

We fit a fully-connected, single-layer feed-forward neural network with sigmoid activation function and l-1 regularization term using TensorFlow. We employ the Adam stochastic optimization method with step size of 0.001 and exponential decay rates for the moment estimates of 0.9 and 0.999. We use a batch size of 32, a maximum number of 100 epochs and an early stopping criterion that stops the training when the validation error has not decreased by more than 10^{-3} for at least 10 epochs. The number of hidden nodes and the regularization weight are chosen so as to minimize the validation loss. The optimal network architecture has 150 hidden units and the optimal regularization weight is 10^{-5}. The mean squared error (MSE) for the test set is 0.45.

We use the discretization method to compute the asymptotic distribution of the test statistic (see Section 4.2 above). We sample 500 random functions using neural networks with the same architecture as the fitted one but with weights sampled from a Glorot normal distribution (see Section 5.7 above). 10,000 samples are generated to estimate the quantile of the asymptotic distribution.

Table 5 reports the values of the empirical test statistic (15) for the top 10 variables that are significant at the 5% level; the variables are ordered according to the magnitude of the test statistic. The last sale price is the most influential variable, closely followed by the square footage of the land. The month of sale is also very influential. This reflects the seasonality of sales: the spring months usually bring the most listings and the best sales

Variable Name	Test Statistic
Median_NonInst_PriceChange_pct	4.430e-05
Median_SAF_PriceChange	3.096e-05
Median_SAF_PriceChange_Model	2.795e-05
Median_NonInst_SAF_PriceChange_Model_pct	1.405e-05

Table 6: Variables that are found to be insignificant at the 5% level.

prices, as many people try to move in the summer during summer holidays. The square footage of the home turns out to be less influential than perhaps expected.

Table 6 reports the variables that are found to be insignificant at the 5% level.

7. Conclusion

We develop and analyze a pivotal test for assessing the statistical significance of the feature variables of a single-layer, fully connected feedforward neural network that models an unknown regression function. The gradient-based test statistic is a weighted average of the squared partial derivative of the neural network estimator with respect to a variable. We show that under technical conditions, the large-sample asymptotic distribution of the test statistic is given by a mixture of chi-square distributions. A simulation study illustrates the performance of the test and an empirical application to house price valuation highlights its behavior given real data.

Acknowledgments

Appendix A. Proofs

The following result is used repeatedly in the proofs.

Theorem 5 (Theorem 2.11.23 of Van der Vaart and Wellner (1996)) *For each n, let $\mathcal{F}_n = \{f_{n,t} : t \in T\}$ be a class of measurable functions indexed by a totally bounded semimetric space (T, ρ). Given envelope functions F_n, assume that*

$$\mathbb{E}(F_n^2) = O(1),$$

$$\mathbb{E}(F_n^2 1_{\{F_n > \eta\sqrt{n}\}}) \to 0, \text{ for every } \eta > 0,$$

16

$$\sup_{\rho(s,t)<\delta_n} \mathbb{E}[(f_{s,t} - f_{n,t})^2] \to 0, \text{ for every } \delta_n \to 0.$$

If

$$\int_0^{\delta_n} \sqrt{\log N_{[]}(\epsilon\|F_n\|_{P,2}, \mathcal{F}_n, L^2(P))} d\epsilon \to 0, \text{ for every } \delta_n \to 0,$$

then the sequence $\{\mathbb{G}_n f_{n,t} : t \in T\}$ is asymptotically tight in $l^\infty(T)$ and converges in distribution to a tight Gaussian process provided the sequence of covariance functions $\mathbb{E}(f_{n,s}f_{n,t}) - \mathbb{E}(f_{n,s})\mathbb{E}(f_{n,t})$ converges pointwise on $T \times T$.

Proof [Proof of Theorem 1]

The proof proceeds in three steps. First we derive the estimation rate r_n of f_n. Then we show that a rescaled and shifted version of the empirical criterion function converges in distribution to a Gaussian process. Finally we apply the argmax continuous mapping theorem.

First we consider the estimation rate r_n of f_n which implies the tightness of the random sequence $h_n = r_n(f_n - f_0)$. Chen and Shen (1998) derive an estimation rate with respect to the $L^2(P)$-metric on Θ, $\|f - g\|^2 = \mathbb{E}_X[(f(X) - g(X))^2]$ of the neural network sieved M-estimator, i.e. $\|f_n - f_0\| = O_P\left(\frac{1}{r_n}\right)$ with K_n such that $K_n^{2+1/d} \log K_n = O(n)$.

The second step is to show that a rescaled and shifted version of the empirical criterion function converges in distribution to a Gaussian process. We use a CLT for empirical processes on classes of functions changing with n, see Theorem 2.11.23 of Van der Vaart and Wellner (1996).

Using the definition of the regression problem in (1), let us redefine the criterion function as $l_g(X, \epsilon) = 2(g - f_0)(X)\epsilon - (f_0 - g)^2(X)$. We now prove a CLT for the empirical processes defined on the class of function $\{r_n(l_{f_0+h/r_n} - l_{f_0}) : h \in K\}$ for every compact subset $K \subset \Theta$ by using (Van der Vaart and Wellner, 1996, Theorem 2.11.23). K is a compact metric space with associated metric $\rho(s,t) = \|s - t\|_{L^2(P)}$ where P denotes the joint law of X and ϵ.

We define

$$f_{n,h} = r_n(l_{f_0+h/r_n} - l_{f_0}) = 2h(X)\epsilon - \frac{h(X)^2}{r_n},$$

and the associated class of functions

$$\mathcal{F}_n = \{f_{n,h} : h \in K\}.$$

The envelope functions F_n of $\mathcal{F}_n$ must be such that $\forall h \in K$, $f_{n,h}(x, \epsilon) \leq F_n(x, \epsilon)$, hence F_n can be chosen as $(x, \epsilon) \to \sup_{h \in K} f_{n,h}(x, \epsilon)$. Using (Chen and Shen, 1998, Lemma 2), we have that $\|h\|_\infty \lesssim \|h\|_2^c$ where $c = \frac{2}{2+d}$. Since K is a compact metric space in the L^2-norm, it is bounded which means that $\exists 1 < M < \infty$ such that $\forall h \in K$, $\|h\|_2 < M$.

Hence

$$|f_{n,h}| \leq 2|\epsilon||h(x)| + \frac{1}{r_n}|h(x)|^2 \leq 2|\epsilon|M + \frac{1}{r_n}M^2,$$

and therefore, we can chose the envelope function as

$$F_n(x, \epsilon) = 2|\epsilon|M + \frac{2}{r_n}M^2,$$

where the additional factor 2 is added to simplify the computation of the metric entropy.

17

By definition of the envelope function and because we assume finite second moment for the regression error ϵ, we have
$$\mathbb{E}[F_n^2(X, \epsilon)] = O(1),$$

and
$$\mathbb{E}(F_n^2 1_{\{F_n > \eta\sqrt{n}\}}) \to 0, \forall \eta > 0.$$

In addition,
$$(f_{n,h_1} - f_{n,h_2})^2 = \left[\left(2\epsilon - \frac{1}{r_n}(h_1 + h_2)\right)(h_1 - h_2)\right]^2 \leq \left[\left(2|\epsilon| + \frac{2}{r_n}M\right)(h_1 - h_2)\right]^2.$$

By taking expectations, we obtain
$$\mathbb{E}[(f_{n,h_1} - f_{n,h_2})^2] \lesssim \mathbb{E}[(h_1 - h_2)^2],$$

which implies that
$$\sup_{\rho(h_1,h_2)<\delta_n} \mathbb{E}(f_{n,h_1} - f_{n,h_2})^2 \leq O(\delta_n^2),$$

$$\sup_{\rho(h_1,h_2)<\delta_n} \mathbb{E}(f_{n,h_1} - f_{n,h_2})^2 \to 0, \forall \delta_n \downarrow 0.$$

Now, we need to find an upper bound on the metric entropy with bracketing of $\mathcal{F}_n$, $N_{[]}(\epsilon||F_n||_{P,2}, \mathcal{F}_n, L^2(P))$, to be able to prove that
$$\int_0^{\delta_n} \sqrt{\log N_{[]}(\epsilon||F_n||_{P,2}, \mathcal{F}_n, L^2(P))} d\epsilon \to 0, \forall \delta_n \downarrow 0.$$

To do so, we will upper bound the metric entropy with bracketing associated with $\mathcal{F}_n$ with the metric entropy without bracketing of the initial function space Θ. Indeed,
$$|f_{n,h_1} - f_{n,h_2}| = |h_1 - h_2|\left|2\epsilon - \frac{1}{r_n}(h_1 + h_2)\right|$$
$$\leq |h_1 - h_2|\left(2|\epsilon| + \frac{2}{r_n}M\right) = |h_1 - h_2|F_n(x, \epsilon)$$
$$\leq ||h_1 - h_2||_\infty F_n(x, \epsilon).$$

The last inequality implies that
$$N_{[]}(\epsilon||F_n||_{P,2}, \mathcal{F}_n, L^2(P)) \leq N\left(\frac{\epsilon}{2}, \Theta, ||\cdot||_\infty\right). \tag{27}$$

Thanks to (Nickl and Pötscher, 2007, Corollary 2, Corollary 4) and because Θ as defined in (7) is a subset of the weighted Sobolev space $H_2^s(\mathbb{R}^d, \langle x \rangle^3)|\mathcal{X}$ restricted to the input space with $s = \lfloor \frac{d}{2} \rfloor + 2$, we have that
$$\log N\left(\frac{\epsilon}{2}, \Theta, ||\cdot||_\infty\right) \lesssim \left(\frac{1}{\epsilon}\right)^{\frac{d}{s}}. \tag{28}$$

18

By combining (27) and (28), we have that

$$\int_0^1 \sqrt{\log N_{[]}(\epsilon||F_n||_{P,2}, \mathcal{F}_n, L^2(P))} d\epsilon < \infty,$$

for all n which implies that

$$\int_0^{\delta_n} \sqrt{\log N_{[]}(\epsilon||F_n||_{P,2}, \mathcal{F}_n, L^2(P))} d\epsilon \to 0, \forall \delta_n \downarrow 0.$$

Now, we want to show that the sequence of covariance functions $Pf_{n,s}f_{n,t} - Pf_{n,s}Pf_{n,t}$ converges pointwise on $\Theta \times \Theta$.

$$Pf_{n,s}Pf_{n,t} = \mathbb{E}\left[2s(X)\epsilon - \frac{s(X)^2}{r_n}\right]\mathbb{E}\left[2t(X)\epsilon - \frac{t(X)^2}{r_n}\right] \to 0,$$

as $n \to \infty$ and

$$Pf_{n,s}f_{n,t} = \mathbb{E}\left[\left(2s(X)\epsilon - \frac{s(X)^2}{r_n}\right)\left(2t(X)\epsilon - \frac{t(X)^2}{r_n}\right)\right] \to 4\sigma^2\mathbb{E}(s(X)t(X)),$$

hence the sequence of covariance functions converges pointwise to $4\sigma^2\mathbb{E}(s(X)t(X))$.

Now, since all hypotheses of Theorem (Van der Vaart and Wellner, 1996, Theorem 2.11.23) are satisfied, we have that the sequence of empirical processes $\{\mathbb{G}_n f_{n,h} : h \in K\}$ is asymptotically tight in $l^\infty(K)$ and converges in distribution to a tight mean zero Gaussian process $\mathbb{G}$ with covariance $4\sigma^2\mathbb{E}(s(X)t(X))$.

We now need to show that all sample paths $h \to \mathbb{G}(h)$ are upper semi continuous and that the limiting Gaussian process $\{\mathbb{G}_f : f \in \Theta\}$ has a unique maximum at a random function $h^\star$.

Let's first note that the space Θ is $||\cdot||_{L_2(P)}$-compact as shown in Freyberger and Masten (2015).

Concerning continuity of the sample paths of the Gaussian process, as stated in (Adler, 1990, Corollary 4.15), a sufficient condition of continuity of a centered Gaussian process with an index space Θ totally bounded with respect to the canonical metric d induced by the process, $d(s,t) = \sqrt{\mathbb{E}[(\mathbb{G}_s - \mathbb{G}_t)^2]}$, is that

$$\int_0^\infty \sqrt{\log N(\epsilon, \Theta, d)} d\epsilon < \infty.$$

Let's note that in our case the canonical metric induced by the process d is proportional to the regular $||\cdot||_{L^2(P)}$ metric,

$$\begin{aligned}
d(s,t) &= \sqrt{\mathbb{E}[(\mathbb{G}_s - \mathbb{G}_t)^2]} \\
&= \sqrt{\mathbb{E}[\mathbb{G}_s^2] + \mathbb{E}[\mathbb{G}_t^2] - 2\mathbb{E}[\mathbb{G}_s\mathbb{G}_t]} \\
&= \sqrt{4\sigma^2\mathbb{E}[s^2] + 4\sigma^2\mathbb{E}[t^2] - 8\sigma^2\mathbb{E}[st]} \\
&= 2\sigma||s - t||_{L^2(P)}.
\end{aligned}$$

19

Therefore we need to show that

$$\log N(\epsilon, \Theta, L^2(P)) \lesssim \left(\frac{1}{\epsilon}\right)^\alpha,$$

with $\alpha < 2$. This can also be proved thanks to (Nickl and Pötscher, 2007, Corollary 2, Corollary 4) and by definition of Θ in (7).

This proves the boundedness of the integral of the metric entropy defined under the d metric. The total boundedness of the index space Θ is induced by its compactness. Hence $\{\mathbb{G}_f : f \in \Theta\}$ is continuous over a compact index space which implies that $\{\mathbb{G}_f : f \in \Theta\}$ reaches its maximum.

Uniqueness of the maximum of a Gaussian process is given in (Kim and Pollard, 1990, Lemma 2.6) and requires continuity of sample paths, a compact index space and for every $s \neq t$, $\mathsf{Var}(\mathbb{G}_s - \mathbb{G}_t) \neq 0$. This is satisfied in our case, indeed, $\mathsf{Var}(\mathbb{G}_s - \mathbb{G}_t) = 4\sigma^2\mathbb{E}(s^2 + t^2) - 8\sigma^2\mathbb{E}(st) = 4\sigma^2\mathbb{E}((s-t)^2) \neq 0$.

In order to apply the argmax continuous mapping theorem, as stated in (Van der Vaart and Wellner, 1996, Theorem 3.2.2), we finally need to show that

$$\mathbb{G}_n f_{n,h_n} \geq \sup_{h \in \Theta} \mathbb{G}_n f_{n,h} - o_P(1). \tag{29}$$

Given the definition of $f_{n,h}$ and $\mathbb{G}_n f_{n,h}$, (29) is equivalent to

$$\frac{r_n}{\sqrt{n}} \sum_{i=1}^n \left(l_{f_n}(X_i, \epsilon_i) - \mathbb{E}[l_{f_n}(X, \epsilon)]\right) \geq \sup_{h \in \Theta} \frac{r_n}{\sqrt{n}} \sum_{i=1}^n \left(l_{f_0 + \frac{h}{r_n}}(X_i, \epsilon_i) - \mathbb{E}[l_{f_0 + \frac{h}{r_n}}(X, \epsilon)]\right) - o_P(1). \tag{30}$$

This condition is satisfied given Condition 4 of Theorem 1.

Finally, from the convergence of the empirical criterion function and the tightness of the random sequence h_n previously shown, one can apply the argmax continuous mapping theorem as stated in (Van der Vaart and Wellner, 1996, Theorem 3.2.2).

This shows the convergence of h_n to the argmax of the Gaussian process $h^\star$. ■

Proof [Proof of Theorem 2] The result follows from Theorem 1 and the second order functional delta method (Römisch, 2005, Theorem 2). We need to compute the Hadamard derivative of our functional of interest ϕ as defined in (6). The second-order Hadamard directional derivative of a mapping between two normed space $\mathbb{D}$ and $\mathbb{F}$, $\phi : \mathbb{D}_0 \subseteq \mathbb{D} \to \mathbb{F}$ at θ_0 in the direction h tangential to $\mathbb{D}_0$ is defined as

$$\phi''_{\theta_0}(h) = \lim_{t \to 0, n \to \infty} \frac{\phi(\theta_0 + th_n) - \phi(\theta_0) - t\phi'_{\theta_0}(h_n)}{\frac{1}{2}t^2}, \tag{31}$$

for every sequence $h_n \to h$ such that $\theta_0 + th_n$ is contained in the domain of ϕ for all small t.

In our case, the second derivative of ϕ at θ_0 under the null hypothesis is equal to

$$\phi''_{\theta_0}(h) = 2 \int_{\mathcal{X}} \left(\frac{\partial h(x)}{\partial x_j}\right)^2 d\mu(x).$$

This completes the proof. ■

Proof [Proof of Proposition 3] We simplify the notation as

$$\hat{\lambda}_j^n = \mathbb{P}_n \eta(f_n), \qquad \lambda_j^n = P\eta(f_n),$$

where P and $\mathbb{P}_n$ denote the expectation and the empirical expectation respectively, and $\eta(f) = \left(\frac{\partial f}{\partial x_j}\right)^2$. We have

$$r_n^2\Big(\mathbb{P}_n\eta(f_n) - P\eta(f_0)\Big) = \frac{r_n^2}{\sqrt{n}}\mathbb{G}_n\Big(\eta(f_n) - \eta(f_0)\Big) + \frac{r_n^2}{\sqrt{n}}\mathbb{G}_n\Big(\eta(f_0)\Big) + r_n^2\Big(P\eta(f_n) - P\eta(f_0)\Big),$$

where $\mathbb{G}_n$ denotes the empirical process operator defined as $\mathbb{G}_n f = \sqrt{n}(\mathbb{P}_n - Pf)$.

Concerning the second term on the right-hand side, given the definition of the null hypothesis H_0, we have that $\mathbb{E}[\eta(f_0)] = 0$ which implies that $\frac{\partial f_0}{\partial x_j} = 0$ P-almost surely. Thanks to that, we have that $\frac{r_n^2}{\sqrt{n}}\mathbb{G}_n\Big(\eta(f_0)\Big) = 0$ P-almost surely.

The distribution of the last term $r_n^2\Big(P\eta(f_n) - P\eta(f_0)\Big)$ is given by Theorem 2.

For the remaining term, we show that the empirical process $\{\mathbb{G}_n f_{n,h} : h \in K\}$ based on the family of functions $\{f_{n,h} = \eta(\theta_0 + h/r_n) - \eta(\theta_0), h \in K\}$ converges to a tight Gaussian process using (Van der Vaart and Wellner, 1996, Theorem 2.11.23). We have

$$f_{n,h} = \eta(\theta_0 + h/r_n) - \eta(\theta_0) = \frac{1}{r_n^2}\left(\frac{\partial h}{\partial x_j}\right)^2,$$

under the null hypothesis. By definition of Θ, $\exists M > 0$ such that $\forall h \in \Theta$, $\left\|\frac{\partial h}{\partial x_j}\right\|_2 \leq M$. Because $\frac{\partial h}{\partial x_j}$ is C^1 and by (Chen and Shen, 1998, Lemma 2), we have a uniform boundedness with the infinity norm, i.e. $\exists M > 0$ such that $\forall h \in K$, $\|\frac{\partial h}{\partial x_j}\|_\infty \leq M$.

We can then define the envelope function F_n as

$$F_n(x) = 2\frac{M^2}{r_n^2},$$

which satisfies

$$\mathbb{E}(F_n^2) = O(1),$$

$$\mathbb{E}(F_n^2 1_{\{F_n > \eta\sqrt{n}\}}) \to 0,$$

for every $\eta > 0$.

Since $f_{n,s} - f_{n,t} = \frac{1}{r_n^2}\left[\left(\frac{\partial s}{\partial x_j}\right)^2 - \left(\frac{\partial t}{\partial x_j}\right)^2\right] \to 0$ for all $s, t \in \Theta \times \Theta$ we immediately have that

$$\sup_{\rho(s,t)<\delta_n} \mathbb{E}[(f_{n,s} - f_{n,t})^2] \to 0,$$

for every $\delta_n \to 0$. Further,

$$|f_{n,s} - f_{n,t}| \leq \frac{1}{r_n^2}\left[\left(\frac{\partial s}{\partial x_j}\right)^2 - \left(\frac{\partial t}{\partial x_j}\right)^2\right] \leq |F_n(x)|\left|\frac{\partial s}{\partial x_j} - \frac{\partial t}{\partial x_j}\right|.$$

Hence by the same argument used in the proof of Theorem 1, we obtain

$$N_{[]}(\epsilon||F_n||_{P,2}, \mathcal{F}_n, L^2(P)) \leq N\left(\frac{\epsilon}{2}, \Theta', ||\cdot||_\infty\right),$$

where $\Theta' = \{\frac{\partial f}{\partial x_j} : f \in \Theta\}$.

By definition of Θ and thanks to (Nickl and Pötscher, 2007, Corollary 2 and 4) we have that

$$\log N_{[]}(\epsilon||F_n||_{P,2}, \mathcal{F}_n, L^2(P)) \lesssim \left(\frac{1}{\epsilon}\right)^\gamma,$$

with $\gamma < 2$.

This last condition ensures that

$$\int_0^{\delta_n} \sqrt{\log N_{[]}(\epsilon||F_n||_{P,2}, \mathcal{F}_n, L^2(P))} d\epsilon \to 0, \forall \delta_n \downarrow 0.$$

By (Van der Vaart and Wellner, 1996, Theorem 2.11.23), we obtain that $\mathbb{G}_n\left(\eta(\hat{\theta}_n) - \eta(\theta_0)\right)$ is asymptotically tight.

Then, $\frac{r_n^2}{\sqrt{n}}\mathbb{G}_n\left(\eta(\hat{\theta}_n) - \eta(\theta_0)\right)$ converges to 0 in probability. By Slutsky's theorem, we finally get that the limiting distribution of $\hat{\lambda}_j^n$ is the same as that of λ_j^n. ■

Proof [Proof of Theorem 4]

The new expression of the limiting distribution is obtained by considering an orthonormal decomposition of the Gaussian process. This is then used to express the argmax of the Gaussian process as the maximizer of a dot product over a ball.

Recall that in our case, the limiting Gaussian process is centered and has covariance $Cov(\mathbb{G}_f, \mathbb{G}_g) = 4\sigma^2\mathbb{E}(fg)$, which is the scaled inner product $\langle f, g \rangle = \mathbb{E}(fg)$ in the Hilbert space $L_2(P)$. If we consider an orthonormal basis $\{\phi_i\}_{i=1}^\infty$ of the Hilbert space, then $f = \sum_{i=1}^\infty \langle f, \phi_i \rangle \phi_i, \forall f \in \Theta$.

By linearity of the Gaussian process, we have that

$$\mathbb{G}_f = \sum_{i=1}^\infty \langle f, \phi_i \rangle \mathbb{G}_{\phi_i}. \tag{32}$$

The linearity of $\mathbb{G}$ is proved as follows: $\mathbb{G}_{f+\lambda g} \sim \mathcal{N}(0, 4\sigma^2\mathbb{E}[(f + \lambda g)^2])$ by definition of $\mathbb{G}$ and $\mathbb{G}_f + \lambda\mathbb{G}_g \sim \mathcal{N}(0, 4\sigma^2\mathbb{E}[f^2] + \lambda^2 4\sigma^2\mathbb{E}[g^2] + \lambda 8\sigma^2\mathbb{E}(fg))$, hence $\mathbb{G}_{f+\lambda g}$ and $\mathbb{G}_f + \lambda\mathbb{G}_g$ are equal in distribution.

Because $\{\phi_i\}_{i=1}^\infty$ is an orthonormal basis,

$$\mathbb{G}_{\phi_i} \sim \mathcal{N}(0, 4\sigma^2), \tag{33}$$

and $\forall i \neq j$, we have $Cov(\mathbb{G}_{\phi_i}, \mathbb{G}_{\phi_j}) = 0$. This, combined with the fact that $\mathbb{G}_{\phi_i}$ and $\mathbb{G}_{\phi_j}$ are jointly normal, implies that $\mathbb{G}_{\phi_i}$ is independent of $\mathbb{G}_{\phi_j}$.

Hence, by combining (32) and (33), we obtain the following decomposition of the Gaussian process $\forall f \in \Theta$,

$$\sum_{i=1}^{\infty} \langle f, \phi_i \rangle \xi_i, \tag{34}$$

where $\{\xi_i\}_{i=1}^{\infty}$ is an infinite sequence of independent normal variables such that $\xi_i \sim \mathcal{N}(0, 4\sigma^2)$.

Our objective is to estimate the distribution of the argmax $h^\star$ of the previous Gaussian process. The argmax is defined as a random function $\Omega \to \Theta$ such that $\forall \omega \in \Omega$, $h_\omega^\star$ satisfies $\mathbb{G}_{h_\omega^\star}(\omega) \geq \mathbb{G}_f(\omega)$, $\forall f \in \Theta$. The previous expansion of the Gaussian process gives us one way to compute the argmax for each ω. Indeed, for every $\omega \in \Omega$ which corresponds to a sample from the infinite sequence of independent normal variables $\{\xi_i\}_{i=1}^{\infty}$, let's define $\Phi_\xi : \Theta \to \mathbb{R}$ as $\Phi_\xi(f) = \mathbb{G}_f(\omega) = \sum_{i=1}^{\infty} \langle f, \phi_i \rangle \xi_i$. Φ_ξ is a bounded linear functional on the Hilbert space $L^2(P)$, hence by Riesz representation theorem, $\exists \phi_\xi \in L^2(P)$ such that $G_f(\omega) = \langle f, \phi_\xi \rangle_{L_2(P)}$. Hence, finding $f \in \Theta$ that maximizes $\mathbb{G}_f(\omega)$ is equivalent of finding $f \in \Theta$ that maximizes the inner product $\langle f, \phi_\xi \rangle_{L^2(P)}$.

By associating to any $f \in L^2(P)$ its corresponding infinite sequence $\{\langle f, \phi_i \rangle\}_{i \in \mathbb{N}}$, one can identify $L^2(P)$ with the l^2 space of infinite sequences with bounded 2-norm. Hence, (18) shows that the Sobolev norm $\|\cdot\|_{\mathbb{H}^k(P)}$ is a weighted version of the l^2 norm with weights d_i^2.

Thanks to the previous observation, finding $f \in \Theta$ that maximizes $\mathbb{G}_f(\omega)$ can be re-expressed as solving the following optimization problem,

$$\max_{\{\langle f, \phi_i \rangle\}_{i \in \mathbb{N}}} \langle f, \phi_\xi \rangle_{l^2} = \sum_{i=0}^{\infty} \langle f, \phi_i \rangle \langle \phi_\xi, \phi_i \rangle,$$

subject to $\|Df\|_{l^2}^2 = \sum_{i=0}^{\infty} d_i^2 \langle f, \phi_i \rangle^2 \leq B$.

By defining $g = Df$, the optimization problem above can be reformulated as

$$\max \langle g, D^{-1} \phi_\xi \rangle_{l^2},$$

subject to $\|g\|_{l^2} \leq B$ whose solution by Cauchy-Schwartz inequality obviously is

$$g^\star = \frac{B}{\|D^{-1} \phi_\xi\|_{L^2(P)}} D^{-1} \phi_\xi.$$

Because $f = D^{-1} g$ and $\phi_\xi = \sum_{i=0}^{\infty} \xi_i \phi_i$, the argmax can be computed explicitly as

$$h_\omega^\star = \sum_{i=0}^{\infty} \frac{B}{\|D^{-1} \phi_\xi\|} \frac{\xi_i}{d_i^2} \phi_i,$$

where $\|D^{-1} \phi_\xi\|^2 = 4\sigma^2 \sum_{i=0}^{\infty} \frac{\chi_i^2}{d_i^2}$ and $\{\chi_i^2\}_{i \in \mathbb{N}}$ are i.i.d. samples from a chi-square distribution.

Given the argmax $h^\star$ and because of (17), we can compute the associated statistic

$$\mathbb{E}_X \left(\frac{\partial h^\star(X)^2}{\partial x_j} \right) = \left\| \frac{\partial h^\star(X)}{\partial x_j} \right\|_{L^2(P)}^2 = \frac{B^2}{\sum_{i=0}^{\infty} \frac{\chi_i^2}{d_i^2}} \sum_{i=0}^{\infty} \frac{\alpha_{i,j}^2}{d_i^4} \chi_i^2.$$

This completes the proof. ∎

Appendix B. Variables for House Price Valuation

Variable Name	Variable Description
House Variables	
Bedrooms	Number of bedrooms
Full_Baths	Number of bathrooms
Last_Sale_Amount	Previous sale price
SqFt	Square footage
Stories_Number	Number of stories
Time_Since_Prior_Sale	Time since prior sale
Year_Built	Year of construction
Mortgage Variables	
Last_Mortgage_Amount	Amount of the mortgage of the last sale
Last_Mortgage_Interest_Rate	Interest rate of the mortgage
Tax Variables	
Tax_Amount	The Total Tax amount provided by the county or local taxing/assessment authority.
Tax_Assd_Improvement_Value	The Assessed Improvement Values as provided by the county or local taxing/assessment authority.
Tax_Assd_Land_Value	The Assessed Land Values as provided by the county or local taxing/assessment authority.
Tax_Assd_Total_Value	The Total Assessed Value of the Parcel's Land and Improvement values as provided by the county or local taxing/assessment authority.
Tax_Improvement_Value_Calc	The "IMPROVEMENT" Value closest to current market value used for assessment by county or local taxing authorities.
Tax_Land_Square_Footage	Total land mass in Square Feet.
Tax_Land_Value_Calc	The "LAND" Value closest to current market value used for assessment by county or local taxing authorities.
Tax_Total_Value_Calc	The "TOTAL" (i.e., Land + Improvement) Value closest to current market value used for assessment by county or local taxing authorities.
Foreclosure Variables	
N_Foreclosures	How many foreclosures occurred in that given month in that given CBSA
N_Outstanding_Foreclosures_SAF	Number of properties that have already been foreclosed upon but have not yet been sold to a party other than the original lender for the given CBSA in the given month.
N_Outstanding_Foreclosures_NonInst_SAF	Same as above but for non-institutional / non-corporate owner SAF.
Foreclosures_To_Loans_Ratio	Ratio of N_Outstanding_Foreclosures_SAF and N_Outstanding_Loans
Other Variables	
Sale_Month_No	Month of the sale

N_Past_Sales	Number of past sales
Mortgage Rate Variables	
Freddie_30yrFRM_5/1ARM_Spread_PriorSale	Spread between Freddie Mac 30-year fixed mortgage rate and 5/1 hybrid amortizing adjustable-rate mortgage (ARM)
Freddie_5/1_ARM_Rate_PriorSale	Freddie Mac 5/1 hybrid amortizing ARM initial coupon rate
Freddie_5/1_ARM_margin_PriorSale	Freddie Mac 5/1 hybrid amortizing ARM margin
Freddie_5/1_feespts_PriorSale	Freddie Mac 5/1 hybrid amortizing ARM fees and points
Freddie_Regional_NC_1yrARM_PriorSale	Freddie Mac 1-Year ARM North Central
Freddie_Regional_NE_1yrARM_PriorSale	Freddie Mac 1-Year ARM North East
Freddie_Regional_SE_1yrARM_PriorSale	Freddie Mac 1-Year ARM South East
Freddie_Regional_SW_1yrARM_PriorSale	Freddie Mac 1-Year ARM South West
Freddie_Regional_W_1yrARM_PriorSale	Freddie Mac 1-Year ARM West
Freddie_US_15yr_FRM_PriorSale	Freddie Mac 15-Year fixed mortgage rate
Freddie_US_15yr_feespts_PriorSale	Freddie Mac 15-Year fixed mortgage rate fees and points
Freddie_US_1yrARM_Fees_PriorSale	Freddie Mac 1-Year ARM fees and points
Freddie_US_1yrARM_Margin_PriorSale	Freddie Mac 1-Year ARM margin
Freddie_US_1yrARM_PriorSale	Freddie Mac 1-Year ARM
Freddie_US_30yr_FRM_PriorSale	Freddie Mac 30-Year fixed mortgage rate
Freddie_US_30yr_feespts_PriorSale	Freddie Mac 30-Year fixed mortgage rate fees and points
Zip-code Variables	
N_ShortSales	Number of short sales that occurred within the same zipcode as the property within the past 12 months.
N_PrePayments	Number of loans that have been prepaid
N_Outstanding_Loans	How many loans are currently outstanding
N_PIF	Number of loans that were paid in full by reaching their term maturity
N_Originations	Number of loans originated at same location and same time period of sale
Median_Time_to_SAF	Median time between the original foreclosure month and the SAF month of all SAF that occurred in the given month.
Median_Time_to_NonInst_SAF	Same as above but for non-institutional / non-corporate owner SAF.
Median_Time_to_SAF_l12	Same as above except that computation rolls over the 12 preceding months.
Median_Time_to_NonInst_SAF_l12	Same as above except that computation rolls over the 12 preceding months.
Median_time_since_foreclosure_SAF	Median time between the current month and when the foreclosure originally occurred of all properties that have already been foreclosed upon but have not yet been sold.

Median_time_since_foreclosure_NonInst_SAF	Same as above but for non-institutional / non-corporate owner sales after foreclosure.
Median_SAF_PriceChange	Median difference between price of the SAF and last recorded sale price of the property for all SAF that occurred in current month.
Median_NonInst_PriceChange	Same as above but for non-institutional / non-corporate owner SAF.
Median_SAF_PriceChange_l12	Same as above except that computation rolls over the 12 preceding months.
Median_NonInst_PriceChange_l12	Same as above except that computation rolls over the 12 preceding months.
Median_SAF_PriceChange_pct	Same as above except that the price change is expressed as a percentage, i.e. the price change divided by the previous sale price.
Median_NonInst_PriceChange_pct	Same as above except that the price change is expressed as a percentage.
Median_SAF_PriceChange_pct_l12	Same as above except that the price change is expressed as a percentage.
Median_NonInst_PriceChange_pct_l12	Same as above except that the price change is expressed as a percentage.
Median_SAF_PriceChange_Model	Median difference between price of the SAF and a modeled sale price of the property (last sale price $*[1+$ zip code level House Price Index precent change since last sale]) for all SAF that occurred in current month.
Median_NonInst_PriceChange_Model	Same as above but for non-institutional / non-corporate owner SAF.
Median_SAF_PriceChange_Model_l12	Same as above except that computation rolls over the 12 preceding months.
Median_NonInst_PriceChange_Model_l12	Same as above except that computation rolls over the 12 preceding months.
Median_SAF_PriceChange_Model_pct	Same as above except that the price change is expressed as a percentage, i.e. the price change divided by the previous sale price.
Median_SAF_PriceChange_Model_pct_l12	Same as above except that the price change is expressed as a percentage.
Median_NonInst_SAF PriceChange_Model_pct	Same as above except that the price change is expressed as a percentage.
Median_NonInst_SAF PriceChange_Model_pct_l12	Same as above except that the price change is expressed as a percentage.
Time_to_NonInst_to_Outstanding_Loans	Ratio of Median_Time_to_NonInst_SAF_l12 variable over N_Outstanding_Loans

References

Robert J Adler. *An Introduction to Continuity, Extrema, and Related Topics for General Gaussian Processes*, volume 12. IMS, 1990.

David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. How to explain individual classification decisions. *Journal of Machine Learning Research*, 11(Jun):1803–1831, 2010.

Andrew R Barron and Jason M Klusowski. Complexity, statistical risk, and metric entropy of deep nets using total path variation. *arXiv preprint arXiv:1902.00800*, 2019.

An Mei Chen, Haw-minn Lu, and Robert Hecht-Nielsen. On the geometry of feedforward neural network error surfaces. *Neural Computation*, 5(6):910–927, 1993.

Xiaohong Chen. Large sample sieve estimation of semi-nonparametric models. In J.J. Heckman and E.E. Leamer, editors, *Handbook of Econometrics*, volume 6B, chapter 76. Elsevier, 1 edition, 2007.

Xiaohong Chen and Xiaotong Shen. Sieve extremum estimates for weakly dependent data. *Econometrica*, 66(2):289–314, 1998.

Xiaohong Chen and Halbert White. Improved rates and asymptotic normality for nonparametric neural network estimators. *IEEE Transactions on Information Theory*, 45(2):682–691, 1999.

Paulo Cortez and Mark J Embrechts. Opening black box data mining models using sensitivity analysis. In *2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, pages 341–348. IEEE, 2011.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989.

Anupam Datta, Shayak Sen, and Yair Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE symposium on security and privacy (SP)*, pages 598–617. IEEE, 2016.

Jianqing Fan and Jiancheng Jiang. Nonparametric inferences for additive models. *Journal of the American Statistical Association*, 100(471):890–907, 2005.

Yanqin Fan and Qi Li. Consistent model specification tests: Omitted variables and semiparametric functional forms. *Econometrica*, 64(4):865–890, 1996.

Joachim Freyberger and Matthew Masten. Compactness of infinite dimensional parameter spaces. Technical report, cemmap working paper, Centre for Microdata Methods and Practice, 2015.

G. David Garson. Interpreting neural-network connection weights. *AI Expert*, 6(4):46–51, April 1991. ISSN 0888-3785.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR.

Pedro L Gozalo. A consistent model specification test for nonparametric estimation of regression function models. *Econometric Theory*, 9(03):451–477, 1993.

Jingping Gu, Dingding Li, and Dandan Liu. Bootstrap non-parametric significance test. *Journal of Nonparametric Statistics*, 19(6-8):215–230, 2007.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

Daniel J Henderson and Christopher F Parmeter. *Applied nonparametric econometrics*. Cambridge University Press, 2015.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

Jeankyung Kim and David Pollard. Cube root asymptotics. *Ann. Statist.*, 18(1):191–219, 03 1990. doi: 10.1214/aos/1176347498.

Igor Kononenko et al. An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research*, 11(Jan):1–18, 2010.

Pascal Lavergne and Quang Vuong. Nonparametric significance testing. *Econometric Theory*, 16 (4):576–601, 2000.

Pascal Lavergne and Quang H. Vuong. Nonparametric selection of regressors: The nonnested case. *Econometrica*, 64(1):207–219, 1996.

Richard Nickl and Benedikt M Pötscher. Bracketing metric entropy rates and empirical central limit theorems for function classes of besov-and sobolev-type. *Journal of Theoretical Probability*, 20(2): 177–199, 2007.

Julian D Olden and Donald A Jackson. Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. *Ecological Modelling*, 154 (1-2):135–150, 2002.

Jeff Racine. Consistent significance testing for nonparametric regression. *Journal of Business & Economic Statistics*, 15(3):369–378, 1997.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144. Association for Computing Machinery, 2016.

Werner Römisch. Delta method, infinite dimensional. *Encyclopedia of Statistical Sciences*, 2005.

Xiaotong Shen. On methods of sieves and penalization. *The Annals of Statistics*, pages 2555–2591, 1997.

Xiaotong Shen and Wing Hung Wong. Convergence rate of sieve estimates. *The Annals of Statistics*, 22(2):580–615, 06 1994. doi: 10.1214/aos/1176325486.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3145–3153. JMLR. org, 2017.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3319–3328. JMLR. org, 2017.

Aad W Van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge University Press, 2000.

Aad W Van der Vaart and Jon A Wellner. *Weak convergence and empirical processes: with applications to statistics.* Springer, 1996.

Quang H. Vuong. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57(2):307–333, 1989. ISSN 00129682, 14680262.

Halbert White. Learning in artificial neural networks: A statistical perspective. *Neural Computation*, 1(4):425–464, 1989a.

Halbert White. Some asymptotic results for learning in single hidden-layer feedforward network models. *Journal of the American Statistical Association*, 84(408):1003–1013, 1989b.

Halbert White. Connectionist nonparametric regression: Multilayer feedforward networks can learn arbitrary mappings. *Neural Networks*, 3(5):535–549, 1990.

Halbert White and Jeffrey Racine. Statistical inference, the bootstrap, and neural-network modeling with application to foreign exchange rates. *IEEE Transactions on Neural Networks*, 12(4):657–673, 2001.

Adonis John Yatchew. Nonparametric regression tests based on least squares. *Econometric Theory*, 8(4):435–451, 1992.

Achilleas D. Zapranis and Apostolos-Paul N. Refenes. Neural model identification, variable selection and model adequacy. *Journal of Forecasting*, 18(5):299–332, 1999.