Statistics and Probability Letters 174 (2021) 109100

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

A goodness-of-fit test based on neural network sieve estimators

Xiaoxi Shen [a], Chang Jiang [a], Lyudmila Sakhanenko [b], Qing Lu [a,b,*]

[a] Department of Biostatistics, University of Florida, Gainesville, FL, USA
[b] Department of Statistics and Probability, Michigan State University, East Lansing, MI, USA

ARTICLE INFO

Article history:
Received 25 September 2020
Received in revised form 28 February 2021
Accepted 14 March 2021
Available online 26 March 2021

Keywords:
Donsker class
Nonparametric least squares

ABSTRACT

Neural networks have become increasingly popular in the field of machine learning and have been successfully used in many applied fields (e.g., imaging recognition). With more and more research has been conducted on neural networks, we have a better understanding of the statistical proprieties of neural networks. While many studies focus on bounding the prediction error of neural network estimators, limited research has been done on the statistical inference of neural networks. From a statistical point of view, it is of great interest to investigate the statistical inference of neural networks as it could facilitate hypothesis testing in many fields (e.g., genetics, epidemiology, and medical science). In this paper, we propose a goodness-of-fit test statistic based on neural network sieve estimators. The test statistic follows an asymptotic distribution, which makes it easy to use in practice. We have also verified the theoretical asymptotic results via simulation studies and a real data application.

1. Introduction

Deep neural networks have become one of the most popularly used methods in artificial intelligence. Despite its attractive performance in various applications, as a statistical model, it is important to investigate its statistical inference. Due to the unidentifiability of the parameters in the neural network model, which were mentioned in Fukumizu (1996, 2003), classical tests such as Wald test and likelihood ratio test may not work since unidentifiability leads to inconsistency for the parameter estimators (Wu, 1981).

Most existing literature on asymptotic properties of neural networks are based on the nonparametric regression. For example, Chen and Shen (1998) and Shen et al. (2019) developed the rate of convergence of neural network estimators under the random design and the fixed design, respectively. Compared with other commonly used nonparametric estimation methods, such as the Nadaraya–Watson estimator and spline regression, neural networks have advantages in terms of rate of convergence. For instance, it has been shown in Györfi et al. (2006) that the mean integrated squared error (MISE) for Nadaraya–Watson estimator is $\mathbb{E}\left[\left\|\hat{f}_n - f_0\right\|^2\right] = \mathcal{O}\left(n^{-\frac{2}{2+d}}\right)$ when the true function f_0 is L-Lipschitz and the bandwidth $h \asymp n^{-1/(2+d)}$. Györfi et al. (2006) also showed that when the underlying function f_0 has continuous pth order derivative and the degree of B-splines is chosen to be $p-1$, the MISE for spline estimator is $\mathbb{E}\left[\left\|\hat{f}_n - f_0\right\|^2\right] = \mathcal{O}\left((\log n/n)^{\frac{2p}{2p+d}}\right)$. From the order of rate of convergence, both methods suffer from the curse of

[*] Corresponding author at: Department of Biostatistics, University of Florida, Gainesville, FL, USA.
E-mail address: lucienq@ufl.edu (Q. Lu).

https://doi.org/10.1016/j.spl.2021.109100
0167-7152/© 2021 Elsevier B.V. All rights reserved.

dimensionality. On the other hand, it has been shown in Chen and Shen (1998) that for sufficiently smooth true function f_0, neural network estimators have $\left\|\hat{f}_n - f_0\right\| = \mathcal{O}_p\left((n/\log n)^{-\frac{1+1/d}{4(1+1/(2d))}}\right)$. Therefore, neural network estimators can, in some sense, avoid the curse of dimensionality. In the supplementary material, we provide a simple comparison of three types of estimators in estimating a trigonometric function.

While the rate of convergence of neural networks has been studied in previous literature, limited research has been done on the hypothesis testing of neural networks. In Shen et al. (2019), asymptotic normality has been derived for neural network sieve estimators, which can be used to test whether the underlying function has a certain specific form. However, in real data applications, we generally do not know the underlying function, which makes the results hard to apply. Moreover, researchers are often more interested in testing the significant association of multiple covariates with the response of interest. Recently, Horel and Giesecke (2020) proposed a significance test based on neural networks. However, the asymptotic distribution is complicated, which may hinder its use in real data analysis. These issues motive us to develop a new nonparametric significance testing procedure based on neural network sieve estimators. As we demonstrated in the real data application, the new goodness-of-fit test can be used in practical research, such as genetic research. Therefore, one of the importance of our work is to bridge the gap between the theoretical work on neural networks and practical research.

We consider a similar setup to the one used in Shen et al. (2019) with the exception of using the random design. Suppose that $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$ are n pairs of i.i.d. samples generated from the following true nonparametric regression model:

$$Y_i = f_0(\boldsymbol{X}_i) + \epsilon, \quad i = 1, \ldots, n,$$

where $\boldsymbol{X}_i \in \mathcal{X} \subset \mathbb{R}^d$, $i = 1, \ldots, n$ and $\mathcal{X}$ is a compact subset in $\mathbb{R}^d$; $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. random errors independent of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ with $\mathbb{E}[\epsilon] = 0$ and $\mathbb{E}[\epsilon^2] = \sigma^2 < \infty$. It is clear that under the quadratic error loss,

$$f_0(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}] = \operatorname*{argmin}_{f\in\mathcal{F}} \mathbb{E}\left[(Y - f(\boldsymbol{X}))^2 | \boldsymbol{X} = \boldsymbol{x}\right],$$

and f_0 is a minimizer of the population criterion function

$$Q(f) = \mathbb{E}\left[(Y - f(\boldsymbol{X}))^2\right] = \sigma^2 + \mathbb{E}\left[(f(\boldsymbol{X}) - f_0(\boldsymbol{X}))^2\right].$$

Suppose that $\mathcal{F}$ is some function space containing f_0. Throughout the paper, we assume that $\mathcal{F} \subset C(\mathcal{X}) \cap L_2(\mathcal{X}, \mu)$ and the pseudo-metric considered on $\mathcal{F}$ is the classical L_2-metric, that is, for any $f \in \mathcal{F}$, $\|f\|^2 = \int_{\mathcal{X}} f^2(\boldsymbol{x}) d\mu(\boldsymbol{x})$. Under the framework of empirical risk minimization (ERM) (Vapnik, 1998; Devroye et al., 2013), an estimator for f_0 is the one that minimizes the empirical loss function $\mathbb{Q}_n(f) = \frac{1}{n}\sum_{i=1}^n (Y_i - f_0(\boldsymbol{X}_i))^2$, that is,

$$\hat{f}_n = \operatorname*{argmin}_{f\in\mathcal{F}} \mathbb{Q}_n(f) = \operatorname*{argmin}_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n (Y_i - f(\boldsymbol{X}_i))^2.$$

As for the sieve extremum estimator based on a neural network, we define

$$\begin{aligned}\mathcal{F}_{r_n} = \Bigg\{ & \alpha_0 + \sum_{j=1}^{r_n} \alpha_j \sigma\left(\boldsymbol{\gamma}_j^T \boldsymbol{x} + \gamma_{0,j}\right) : \boldsymbol{\gamma}_j \in \mathbb{R}^d, \alpha_j, \gamma_{0,j} \in \mathbb{R}, \\ & \sum_{j=0}^{r_n} |\alpha_j| \le V_n \text{ for some } V_n > 4 \text{ and } \max_{1\le j\le r_n} \sum_{i=0}^{d} |\gamma_{i,j}| \le M_n \text{ for some } M_n > 0 \Bigg\}, \end{aligned} \tag{1}$$

where $r_n, V_n, M_n \uparrow \infty$ as $n \to \infty$ and $\sigma(\cdot)$ is the standard sigmoid function ($\sigma(x) = (1 + e^{-x})^{-1}$). The requirement of $V_n > 4$ is necessary for computing the covering number for $\mathcal{F}_{r_n}$, and 4 is the reciprocal of the Lipschitz constant of σ. Due to the Universal Approximation Theorem (Hornik et al., 1989), $\mathcal{F}_{r_n}$ is nondecreasing and $\bigcup_{r_n=1}^{\infty} \mathcal{F}_{r_n}$ is dense in $\mathcal{F}$ under the sup-norm. With some abuse of notation, the sieve extremum estimator $\hat{f}_n$ is defined as

$$\mathbb{Q}_n(\hat{f}_n) \le \inf_{f\in\mathcal{F}_{r_n}} \mathbb{Q}_n(f) + o_p(n^{-1}), \tag{2}$$

where $\eta_n \to 0$ as $n \to \infty$.

The goal of this paper is to derive a goodness-of-fit statistic to test hypothesis on whether a subset of the covariates is significant, that is, for a given $p > 0$, we test

$$H_0 : f_0 \in C(\mathcal{X}') \cap L_2(\mathcal{X}', \mu) \text{ vs } H_1 : f_0 \in C(\mathcal{X}) \cap L_2(\mathcal{X}, \mu), \tag{3}$$

where $\mathcal{X}'$ is a compact subset in $\mathbb{R}^{d-p}$ and $C(A)$ is the space of continuous function on A. In other words, the statistic is proposed to test is whether the p covariates absent in $\mathcal{X}'$ are significantly associated with the response of interest.

The rest of the paper is organized as follows. Section 2 reviews the necessary concepts and results from the empirical process theory. The main theoretical results and the process of constructing the test statistic are discussed in Section 3. A simulation study is conducted in Section 4 to verify the conditions proposed in the main results followed by a real data application to Alzheimer disease. Additional simulation results and the proofs of the results in the main text are given in the supplementary materials.

Notations: Throughout the paper, bold font alphabetic letters and Greek letters are vectors. We use $\|\cdot\|_{\sup}$ to denote the sup-norm, that is $\|f\|_{\sup} = \sup_{\boldsymbol{x}} |f(\boldsymbol{x})|$. For a pseudo-metric space (T, d), $N(\epsilon, T, d)$, $D(\epsilon, T, d)$ and $N_{[]}(\epsilon, T, d)$ represent the covering number, packing number, and bracketing number, respectively. The natural logarithm of the covering number is denoted by $H(\epsilon, T, d)$, which is also known as the entropy number.

2. Preliminaries

The main tool used in proofs is the Donsker class from the empirical process theory. In this section, we review the definition of Donsker class as well as a way to check whether a function class is Donsker. More details on Donsker class can be found in Dudley (1984) and van der Vaart and Wellner (1996).

Definition 1 (*Donsker Class*)**.**

(i) Let $(\mathcal{X}, \mathcal{A}, P)$ be a probability space and G_P be a Gaussian process indexed by $L_2(\mathcal{X}, \mathcal{A}, P)$ with zero mean and covariance

$$\mathbb{E}\left[G_P(f)G_P(g)\right] = P(fg) - Pf \cdot Pg \text{ for all } f, g \in L_2(\mathcal{X}, \mathcal{A}, P).$$

Define

$$\rho_P(f, g) = \left(\mathbb{E}\left[(G_P(f) - G_P(g))^2\right]\right)^{1/2}.$$

A class $\mathcal{F} \subset L_2(\mathcal{X}, \mathcal{A}, P)$ is called a G_PBUC class (or pre-Gaussian) if and only if the process $G_P(f, \omega)$ can be chosen so that for all ω, the sample functions $f \mapsto G_P(f, \omega)$, restricted to $f \in \mathcal{F}$, are bounded and uniformly continuous for ρ_P.

(ii) A class $\mathcal{F} \subset L_2(\mathcal{X}, \mathcal{A}, P)$ is called a Donsker class (for P) if and only if it is a G_PBUC class and there are processes $Y_j(f, \omega)$, $f \in \mathcal{F}$, $\omega \in \Omega$, where Y_j are independent copies of G_P with $f \mapsto Y_j(f, \omega)$ bounded and ρ_P-uniformly continuous on $\mathcal{F}$ for each j, such that for every $\epsilon > 0$,

$$\mathbb{P}^*\left(n^{-1/2} \max_{m \le n} \sup_{f \in \mathcal{F}} \left|\sum_{j=1}^{m} f(X_j) - Pf - Y_j(f)\right| > \epsilon\right) \to 0 \text{ as } n \to \infty.$$

A common approach to check whether a class $\mathcal{F}$ is a Donsker class is to use the Dudley integral based on the bracketing number.

Theorem 1 (*Theorem 3.1 in Ossiander (1987)*)**.** *If $\mathcal{F}$ is a class of measurable functions with*

$$\int_0^\infty \left(\log N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|_{L_2(P)})\right)^{1/2} d\epsilon < \infty, \tag{4}$$

then $\mathcal{F}$ is a Donsker class.

Dudley (1984) provides a relationship between the bracketing number and packing number:

$$N_{[]}(2\epsilon, \mathcal{F}, \|\cdot\|_\infty) \le 2D(\epsilon, \mathcal{F}, \|\cdot\|_{\sup}).$$

It then follows from the duality of packing number and covering number (see Lemma 5.5 in Wainwright (2019)) that

$$N_{[]}(2\epsilon, \mathcal{F}, \|\cdot\|_\infty) \le 2D(\epsilon, \mathcal{F}, \|\cdot\|_{\sup}) \le 2N\left(\frac{\epsilon}{2}, \mathcal{F}, \|\cdot\|_{\sup}\right). \tag{5}$$

Based on these facts, we verify in the following proposition that $\mathcal{F}_{r_n}$, $\mathcal{G}_{r_n} = \{y - f(\boldsymbol{x}) : f \in \mathcal{F}_{r_n}\}$ and $\mathcal{H}_{r_n} = \{(y - f(\boldsymbol{x})^2 : f \in \mathcal{F}_{r_n})\}$ are all Donsker classes for each fixed n.

Proposition 2. *For each fixed n, the following classes of functions are Donsker classes:*

(i) $\mathcal{F}_{r_n}$,
(ii) $\mathcal{G}_{r_n} = \{y - f(\boldsymbol{x}) : f \in \mathcal{F}_{r_n}\}$, *and*
(iii) $\mathcal{H}_{r_n} = \{(y - f(\boldsymbol{x})^2 : f \in \mathcal{F}_{r_n})\}$.

3. Main results

The theory that derives the asymptotic distribution of a goodness-of-fit test statistic depends heavily on Lemma 2 in Yatchew (1992). The lemma, however, requires strong uniform consistency for nonparametric least squares estimators. Therefore, we modify the lemma under a weaker consistency assumption.

Lemma 3. *Let $\mathcal{H}$ be a Donsker class. Let h_0 and $\hat{h}_n$, $n = 1, 2, \ldots$ be in $\mathcal{H}$, where $\|\hat{h}_n - h_0\|_{L_2(P)} \xrightarrow{P} 0$, then*

$$n^{1/2}\left[\frac{1}{n}\sum_{i=1}^{n}\hat{h}_n(Z_i) - P\hat{h}_n(Z)\right] - n^{1/2}\left[\frac{1}{n}\sum_{i=1}^{n}h_0(Z_i) - Ph_0(Z)\right] \xrightarrow{P} 0.$$

As $\mathcal{H}_{r_n}$ is a Donsker class from Proposition 2. Based on Lemma 3, we can construct a goodness-of-fit statistic.

Theorem 4. *Let $\mathbb{E}[\epsilon^4] < \infty$, then if $\left\|\pi_{r_n} f_0 - f_0\right\| = o(n^{-1/4})$ and*

$$[r_n(d+2)+1]V_n^4 \log\left(V_n[r_n(d+2)+1]\right) = o(n),$$

we have

$$\frac{n^{1/2}}{\kappa^{1/2}}\left[\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{f}_n(\boldsymbol{X}_i)\right)^2 - \sigma^2\right] \xrightarrow{d} \mathcal{N}(0,1),$$

where $\kappa = Var[\epsilon^2]$.

Remark 1. If the underlying function f_0 is real analytic, it follows from a combination of the results in Goulaouic (1971) and Lemma 3.2 in Mhaskar (1996) that the approximation rate can decay exponentially so that the condition $\left\|\pi_{r_n} f_0 - f_0\right\| = o(n^{-1/4})$ satisfies easily.

Theorem 4 provides a theoretical justification for constructing the following goodness-of-fit test statistic for hypothesis testing. The test statistic is formed based on the following steps.

Step 1. Partition the sample $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$ into two equal parts. For simplicity, we assume that $n = 2m$ for some $m > 0$.

Step 2. From Theorem 4, under the null hypothesis $H_0 : f_0 \in C(\mathcal{X}') \cap L_2(\mathcal{X}', \mu)$, we have

$$T_0 = \frac{m^{1/2}}{\kappa^{1/2}}\left[\frac{1}{m}\sum_{i=1}^{m}\left(Y_i - \hat{f}_{n,H_0}(\boldsymbol{X}_i)\right)^2 - \sigma^2\right] \xrightarrow{d} \mathcal{N}(0,1),$$

$$T_1 = \frac{m^{1/2}}{\kappa^{1/2}}\left[\frac{1}{m}\sum_{i=m+1}^{n}\left(Y_i - \hat{f}_{n,H_1}(\boldsymbol{X}_i)\right)^2 - \sigma^2\right] \xrightarrow{d} \mathcal{N}(0,1),$$

where $\hat{f}_{n,H_0}$ is the neural network sieve extremum estimator obtained by using the first m samples and the covariates excluding the p covariates, and f_{n,H_1} is the neural network sieve extremum estimator calculated based on the remaining samples and all of the covariates.

Step 3. Since $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$ are independent, T_0 and T_1 are also independent. Then

$$T_2 = T_0 - T_1 = \frac{m^{1/2}}{\kappa^{1/2}}\left[\frac{1}{m}\sum_{i=1}^{m}\left(Y_i - \hat{f}_{n,H_0}(\boldsymbol{X}_i)\right)^2 - \frac{1}{m}\sum_{i=m+1}^{n}\left(Y_i - \hat{f}_{n,H_1}(\boldsymbol{X}_i)\right)^2\right]$$
$$\xrightarrow{d} \mathcal{N}(0,2).$$

Step 4. For any consistent estimator $\hat{\kappa}_n$ of κ, it then follows from the Slutsky's Theorem that

$$T = \frac{m^{1/2}}{\hat{\kappa}_n^{1/2}}\left[\frac{1}{m}\sum_{i=1}^{m}\left(Y_i - \hat{f}_{n,H_0}(\boldsymbol{X}_i)\right)^2 - \frac{1}{m}\sum_{i=m+1}^{n}\left(Y_i - \hat{f}_{n,H_1}(\boldsymbol{X}_i)\right)^2\right]$$
$$\xrightarrow{d} \mathcal{N}(0,2).$$

As mentioned in Yatchew (1992), a possible choice for $\hat{\kappa}_n$ is

$$\hat{\kappa}_n = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}_{n,H_0}(\boldsymbol{X}_i))^4 - \hat{\sigma}^4,$$

where $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}\left(Y_i - \hat{f}_{n,H_0}(\boldsymbol{X}_i)\right)^2$.

Table 1
Summary of results from the normality tests.

Sample size	50	100	200	500	1500	3000
Shapiro–Wilks	0.412	0.808	0.065	0.521	0.704	0.498
Anderson–Darling	0.980	0.820	0.098	0.492	0.837	0.950

Table 2
Empirical type I error rates under different sample sizes.

Sample size	50	100	200	500	1500	3000
Empirical type I error	0.044	0.056	0.052	0.054	0.046	0.046

4. Simulation

We have conducted a simulation study to verify the main result established in the previous section. Suppose that the response variables $Y_1, \dots, Y_n$ are generated from the following model,

$$Y_i = f_0(\boldsymbol{X}_i) + \epsilon_i, \quad i = 1, \dots, n,$$

where $\epsilon_1, \dots, \epsilon_n \sim$ i.i.d. $\mathcal{N}(0, 1)$. The covariates $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n \in \mathbb{R}^2$ are i.i.d. samples from $\mathcal{N}_2(\boldsymbol{0}, \boldsymbol{I}_2)$, where $\boldsymbol{I}_2$ is the 2-dimensional identity matrix. The hypothesis of interest is whether $\boldsymbol{X}^{(2)}$, the second element in the vector $\boldsymbol{X}$, is significant. Under the null hypothesis H_0, the true function is chosen to be a trigonometric function:

$$f_0(\boldsymbol{x}) = \sin\left(\frac{\pi}{3}\boldsymbol{x}^{(1)}\right) + \frac{1}{3}\cos\left(\frac{\pi}{4}\boldsymbol{x}^{(1)} + 1\right).$$

Based on the testing procedure we established the previous section, the test statistic can be written as

$$T_n = \frac{\sqrt{n/2}}{(2\hat{\kappa})^{1/2}}\left[\frac{2}{n}\sum_{i=1}^{n/2}\left(Y_i - \hat{f}_{n,H_0}(\boldsymbol{X}_i)\right)^2 - \frac{2}{n}\sum_{i=n/2+1}^{n}\left(Y_i - \hat{f}_{n,H_1}(\boldsymbol{X}_i)\right)^2\right],$$

where

$$\hat{\kappa} = \widehat{\mathrm{Var}}[\epsilon^2] = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{f}_{n,H_0}(\boldsymbol{X}_i)\right)^4 - \left[\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{f}_{n,H_0}(\boldsymbol{X}_i)\right)^2\right]^2.$$

Due to the constraints in the neural network sieve $\mathcal{F}_{r_n}$, we used a subgradient method discussed in Section 7 in Boyd and Mutapcic (2008) to estimate the parameters and obtain the fitted function. As mentioned in Boyd and Mutapcic (2008), the algorithm converges when the step size δ_k is diminishing nonsummable ($\delta_k \downarrow 0$ and $\sum_{k=1}^{\infty}\delta_k = \infty$). We thus chose the step size for the kth iteration in the subgradient to be $\delta_k = 0.1/\log(e + k)$ and the number of iterations was set as 3e4. We specified $r_m = \lfloor m^{1/6}\rfloor$ and $V_m = 20m^{1/6}$, where $m = n/2$ so that the assumption in Theorem 4 is fulfilled. For each sample size, 500 Monte Carlo iterations were performed to obtained the normal QQ-plot, which is shown in Fig. 1.

The QQ-plots indicate that T_n does not deviate from the standard normal distribution, which is consistent with the results from the Shapiro–Wilks test and the Anderson–Darling test as summarized in Table 1.

Empirical type I error rates were also calculated based on the test statistics obtained from the 500 Monte Carlo iterations. The results are shown in Table 2.

In the supplementary materials, we evaluated the method's performance for different choices of V_n and r_n. The results are consistent with the findings provided above. Moreover, additional simulations were conducted on multiple covariates. Based on these empirical studies, we found that as long as the choice of r_n and V_n satisfied the required condition $[r_n(d+2)+1]V_n^4\log(V_n[r_n(d+2)+1]) = o(n)$, the asymptotic distribution and the type I error were guaranteed. Moreover, we conducted a simulation study when the null model contains multiple covariates. The empirical type I error can also be controlled when the sample size is large. Details of the simulations can be found in the supplementary materials.

5. A real data application

We applied our method to the sequencing data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) and performed a genetic association analysis. The best known gene related to the Alzheimer's disease (AD) is the *APOE* on chromosome 19 (Strittmatter et al., 1993). The covariates in our analysis are therefore the single-nucleotide polymorphisms (SNPs) in *APOE*. After quality control, a total sample of 780 individuals with 169 SNPs remained for the analysis. Studies have shown that for patients with AD, the whole brain volume decreases significantly (Thambisetty et al., 2011). We therefore chose the logarithm of the whole brain volume as the response.

For each SNP in *APOE* gene, we conducted the goodness-of-fit test as we discussed in Section 3. Table 3 summarizes the top 10 associated SNPs in the *APOE* gene detected by our method. SNPs with P-value less than 0.05 are shown in bold font.

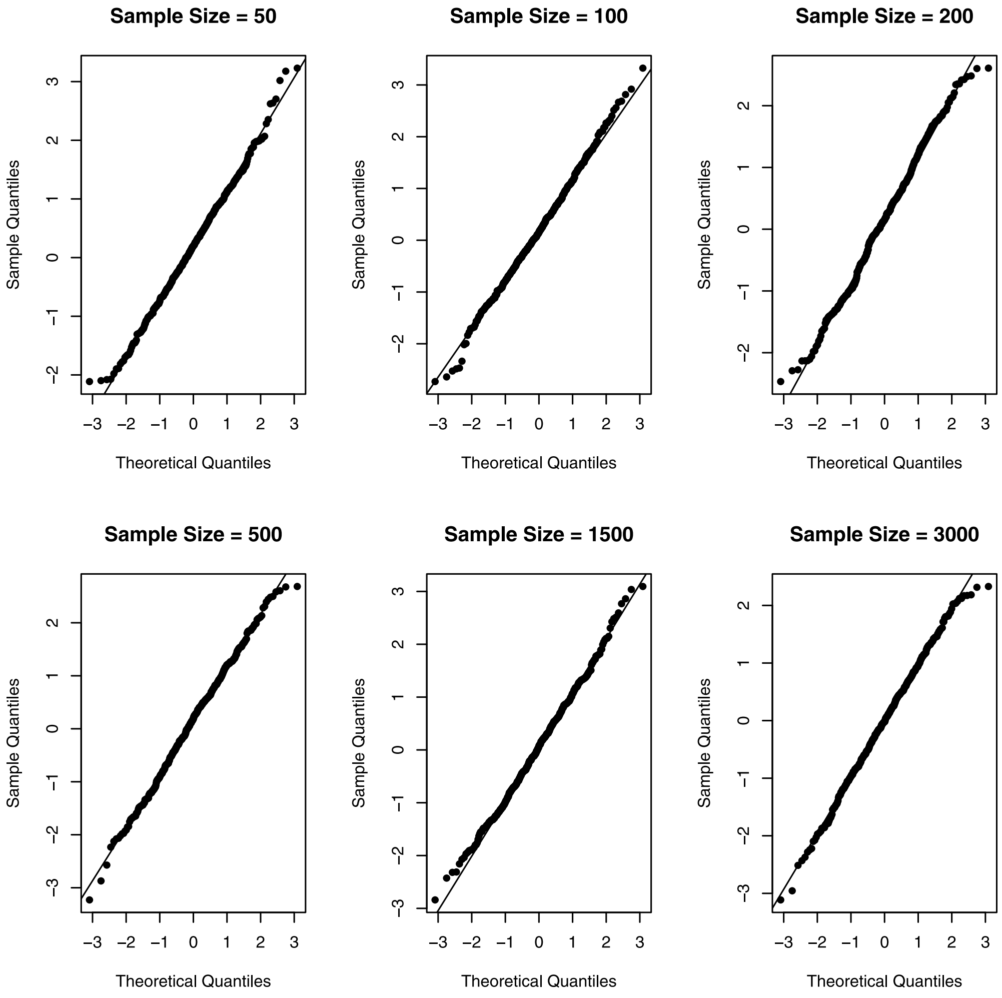

Fig. 1. Normal QQ-plot for the test statistic T_n under various sample sizes $n = 50$, 100, 200, 500, 1500, 3000.

Table 3
Top 10 SNPs in the *APOE* gene associated with the whole brain volume by using the goodness of fit test. Significant SNPs ($P < 0.05$) are highlighted in bold font.

SNP name	P-value
rs_x94	5.046E−3
rs_x21	5.918E−3
rs_x42	8.754E−3
rs72654471	1.595E−2
rs112757453	2.781E−2
rs_x127	2.822E−2
rs59325138	5.029E−2
rs_x131	5.057E−2
rs_x132	5.188E−2
rs_x52	5.398E−2

6. Conclusion

Deep neural networks have been increasingly used in areas such as computer vision and speech recognition. While numerous studies have shown that neural networks attained high performance in terms of prediction accuracy, few studies have investigated the statistical inference of neural networks. Many biomedical studies are hypothesis-driven studies. For instance, in a typical genetic study, investigators are interested in testing the association of genetic variants with a disease of interest. While neural networks hold great promise to reveal the complex relationship between genetic variants and the disease, the lack of established statistical inference limits the use of neural networks in genetic research and other biomedical research.

To fill this gap, we proposed a goodness-of-fit test statistic based on neural network sieve estimators. The proposed test statistic has a simple limiting distribution, which facilitates its use in practice. The idea is to split the sample into two portions, one of which is used to fit the reduced model while the remaining is used to fit the full model. The test statistic is then built on the difference between the mean squared error of the reduced model and that of the full model. As pointed out in Yatchew (1992), sample splitting is necessary. Otherwise, the asymptotic distribution of the test statistic may be degenerate when the mean squared error of the reduced model and that of the full model are computed from the same sample. In fact, the idea of sample splitting has also been applied to construct important quantities in other theoretical studies. For example, Bartlett et al. (2002) defined the maximum discrepancy of a function class $\mathcal{F}$ as

$$\hat{D}_n(\mathcal{F}) = \sup_{f\in\mathcal{F}} \left(\frac{2}{n}\sum_{i=1}^{n/2} f(X_i) - \frac{2}{n}\sum_{i=n/2+1}^{n} f(X_i) \right),$$

which quantifies how much the behavior on half of the sample can be unrepresentative of the behavior on the other half. Due to the independence between the two sample portions, the asymptotic normality of the test statistic can be easily established. Nevertheless, the tradeoff is that we have to split the samples to calculate the test statistic, which may result in power loss due to the reduced sample size.

We have conducted a simulation study to confirm the theoretical results. By using the QQ-plots and normality tests, we showed that the type I error of the proposed test was well controlled. This paper can be viewed as our initial effort on building test statistics based on neural networks. Additional topics, such as developing a neural-network-based test using the entire sample, could also be further investigated in future studies.

For future studies, we think it is important to develop similar asymptotic theories for modern convolutional neural networks (CNN) and long-short term memory (LSTM) networks. Most theories nowadays on statistical learning are derived under the framework of Vapnik (1998), which depends on the entropy argument and the Dudley integer. Due to the complex network structures of CNN and LSTM, it may be difficult to obtain upper bounds for the covering numbers of the CNN or LSTM classes. Meanwhile, the rate of convergence of the sieve estimator also depends on the approximation rate of the network to the underlying function, which is difficult to derive. Nevertheless, this is an interesting topic worth further studying in the future.

CRediT authorship contribution statement

Xiaoxi Shen: Conceptualization, Formal analysis, Writing - original draft. **Chang Jiang:** Software, Formal analysis, Writing - original draft. **Lyudmila Sakhanenko:** Validation, Writing - review & editing. **Qing Lu:** Supervision, Funding acquisition, Writing - review & editing.

Acknowledgments

This work was supported by the National Institute on Drug Abuse (Award No. R01DA043501) and the National Library of Medicine (Award No. R01LM012848).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.spl.2021.109100.

References

Bartlett, P.L., Boucheron, S., Lugosi, G., 2002. Model selection and error estimation. Mach. Learn. 48 (1–3), 85–113.
Boyd, S., Mutapcic, A., 2008. Subgradient Methods (notes for EE364B Winter 2006-07. Stanford University).
Chen, X., Shen, X., 1998. Sieve extremum estimates for weakly dependent data. Econometrica 289–314.
Devroye, L., Györfi, L., Lugosi, G., 2013. A Probabilistic Theory of Pattern Recognition, Vol. 31. Springer Science & Business Media.
Dudley, R.M., 1984. A course on empirical processes. In: Ecole D'ÉTÉ de ProbabilitÉS de Saint-Flour XII-1982. Springer, pp. 1–142.
Fukumizu, K., 1996. A regularity condition of the information matrix of a multilayer perceptron network. Neural Netw. 9 (5), 871–879.
Fukumizu, K., 2003. Likelihood ratio of unidentifiable models and multilayer neural networks. Ann. Statist. 31 (3), 833–851.
Goulaouic, C., 1971. Approximation polynômiale de fonctions C^∞ et analytiques. Ann. Inst. Fourier Grenoble 21, 149–173.
Györfi, L., Kohler, M., Krzyzak, A., Walk, H., 2006. A Distribution-Free Theory of Nonparametric Regression. Springer Science & Business Media.
Horel, E., Giesecke, K., 2020. Significance tests for neural networks. Journal of Machine Learning Research 21 (227), 1–29.
Hornik, K., Stinchcombe, M., White, H., 1989. Multilayer feedforward networks are universal approximators. Neural Netw. 2 (5), 359–366.
Mhaskar, H.N., 1996. Neural networks for optimal approximation of smooth and analytic functions. Neural Comput. 8 (1), 164–177.
Ossiander, M., 1987. A central limit theorem under metric entropy with L2 bracketing. Ann. Probab. 897–919.
Shen, X., Jiang, C., Sakhanenko, L., Lu, Q., 2019. Asymptotic properties of neural network sieve estimators. arXiv preprint arXiv:1906.00875.
Strittmatter, W.J., Saunders, A.M., Schmechel, D., Pericak-Vance, M., Enghild, J., Salvesen, G.S., Roses, A.D., 1993. Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. Proc. Natl. Acad. Sci. 90 (5), 1977–1981.
Thambisetty, M., Simmons, A., Hye, A., Campbell, J., Westman, E., Zhang, Y., Wahlund, L.-O., Kinsey, A., Causevic, M., Killick, R., et al., 2011. Plasma biomarkers of brain atrophy in Alzheimer's disease. PLoS One 6 (12), e28527.
van der Vaart, A.W., Wellner, J.A., 1996. Weak Convergence and Empirical Processes. Springer.
Vapnik, V., 1998. Statistical Learning Theory. 1998, Vol. 3. Wiley, New York.
Wainwright, M.J., 2019. High-Dimensional Statistics: a Non-Asymptotic Viewpoint, Vol. 48. Cambridge University Press.
Wu, C.-F., 1981. Asymptotic theory of nonlinear least squares estimation. Ann. Statist. 501–513.
Yatchew, A.J., 1992. Nonparametric regression tests based on least squares. Econometric Theory 8 (4), 435–451.