Towards Explainable AI: Significance Tests for Neural Networks

Kay Giesecke

Advanced Financial Technologies Laboratory
Stanford University

people.stanford.edu/giesecke/
fin-tech.stanford.edu

Joint work with Enguerrand Horel (Stanford)
Neural networks underpin many of the best-performing AI 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 finance, economics, operations, marketing, medicine, and many others.

In finance, researchers have developed several promising applications in risk management, asset pricing, and investment management.
First wave: single-layer nets
- Nonlinearity testing: Lee, White & Granger (1993)
- Economic forecasting: Swanson & White (1997)
- Stock market prediction: Brown, Goetzmann & Kumar (1998)
- Pricing kernel modeling: Bansal & Viswanathan (1993)
- Option pricing: Hutchinson, Lo & Poggio (1994)
- Credit scoring: Desai, Crook & Overstreet (1996)

Second wave: multi-layer nets (deep learning)
- Mortgages: Sirignano, Sadhwani & Giesecke (2016)
- Portfolio selection: Heaton, Polson & Witte (2016)
- Real estate: Giesecke, Ohlrogge, Ramos & Wei (2018)
- Treasury markets: Filipovic, Giesecke, Pelger, Ye (2018?)
- Insurance: Wüthrich and Merz (2019)
The success of NNs is largely due to their amazing approximation properties, superior predictive performance, and their scalability.

A major caveat however is **model explainability**: NNs are perceived as black boxes that permit little insight into how predictions are being made.

Key inference questions are difficult to answer:
- Which input variables are statistically significant?
- If significant, how can a variable’s impact be measured?
- What’s the relative importance of the different variables?
This issue is not just academic; it has slowed the implementation of NNs in financial practice, where regulators and other stakeholders often insist on model explainability.

- **Credit and insurance underwriting**
  - Transparency of underwriting decisions

- **Investment management**
  - Transparency of portfolio designs
  - Economic rationale of trading decisions
We develop a pivotal test to assess the statistical significance of the input variables of a NN

- Focus on single-layer feedforward networks
- Focus on regression setting

We propose a gradient-based test statistic and study its asymptotics using nonparametric techniques

- Asymptotic distribution is a mixture of $\chi^2$ laws

The test enables one to address key inference issues:

- Assess statistical significance of variables
- Measure the impact of variables
- Rank order variables according to their influence

Simulation and empirical experiments illustrate the test
Problem formulation

- Regression model $Y = f_0(X) + \epsilon$
  - $X \in \mathcal{X} \subset \mathbb{R}^d$ is a vector of $d$ variables with law $P$
  - $f_0 : \mathcal{X} \to \mathbb{R}$ is an unknown deterministic $C^1$-function
  - $\epsilon$ is an error variable: $\epsilon \mid X, \mathbb{E}(\epsilon) = 0, \mathbb{E}(\epsilon^2) = \sigma^2 < \infty$

- To assess the significance of a variable $X_j$, we propose to test the following hypotheses:

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

  $H_A : \lambda_j \neq 0$

Here, $\mu$ is a positive weight measure

- A typical choice is $\mu = P$ and then $\lambda_j = \mathbb{E}[\left( \frac{\partial f_0(X)}{\partial x_j} \right)^2]$
Suppose the function $f_0$ is linear (multiple linear regression)

$$f_0(x) = \sum_{k=1}^{d} \beta_k x_k$$

Then $\lambda_j \propto \beta_j^2$, the squared regression coefficient for $X_j$, and the null takes the form $H_0 : \beta_j = 0 \ (\rightarrow \ t\text{-test})$

In the general nonlinear case, the derivative $\frac{\partial f_0(x)}{\partial x_j}$ depends on $x$, and $\lambda_j = \int x \left( \frac{\partial f_0(x)}{\partial x_j} \right)^2 d\mu(x)$ is a weighted average
We study the case where the unknown regression function $f_0$ is modeled by a single-layer feedforward NN.

A **single-layer NN** $f$ is specified by a bounded *activation function* $\psi$ on $\mathbb{R}$ and the number of *hidden units* $K$:

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

where $b_0, b_k, a_{0,k} \in \mathbb{R}$ and $a_k \in \mathbb{R}^d$ are to be estimated.

Functions of the form $f$ are dense in $C(\mathcal{X})$ (they are **universal approximators**): choosing $K$ large enough, $f$ can approximate $f_0$ to any given precision.
Neural network with $K = 3$ hidden units
We use $n$ i.i.d. samples $(Y_i, X_i)$ to construct a sieve M-estimator $f_n$ of $f$ for which $K = K_n$ increases with $n$.

We assume $f_0 \in \Theta = \text{class of } C^1 \text{ functions on } d\text{-hypercube } \mathcal{X}$ with uniformly bounded Sobolev norm.

Sieve subsets $\Theta_n \subseteq \Theta$ generated by NNs $f$ with $K_n$ hidden units, bounded $L^1$ norms of weights, and sigmoid $\psi$.

The sieve M-estimator $f_n$ is the approximate maximizer of the empirical criterion function $L_n(g) = \frac{1}{n} \sum_{i=1}^{n} l(Y_i, X_i, g)$, where $l : \mathbb{R} \times \mathcal{X} \times \Theta \to \mathbb{R}$, over $\Theta_n$:

$$L_n(f_n) \geq \sup_{g \in \Theta_n} L_n(g) - o_P(1)$$
The NN test statistic is given by

\[ \lambda_n^j = \int_{\mathcal{X}} \left( \frac{\partial f_n(x)}{\partial x_j} \right)^2 d\mu(x) = \phi[f_n] \]

We will use the asymptotic \((n \to \infty)\) distribution of \(\lambda_n^j\) for testing the null since a bootstrap approach would typically be too computationally expensive.

In the large-\(n\) regime, due to the universal approximation property, we are actually performing inference on the “ground truth” \(f_0\) (model-free inference).
Asymptotic distribution of NN estimator

**Theorem**

Assume that

- \(dP = \nu d\lambda\) for bounded and strictly positive \(\nu\)
- The dimension \(K_n\) of the NN satisfies \(K_n^{2+1/d} \log K_n = O(n)\),
- The loss function \(l(Y_i, X_i, g) = -\frac{1}{2}(Y_i - g(X_i))^2\).

Then

\[
r_n(f_n - f_0) \Rightarrow h^*
\]

in \((\Theta, L^2(P))\) where

\[
r_n = \left( \frac{n}{\log n} \right)^{\frac{d+1}{2(2d+1)}}
\]

and \(h^*\) is the argmax of the Gaussian process \(\{G_f : f \in \Theta\}\) with mean zero and \(\text{Cov}(G_s, G_t) = 4\sigma^2 \mathbb{E}(s(X)t(X))\).
- $r_n$ is the estimation rate of the NN (Chen and Shen (1998)):

$$\mathbb{E}_X[(f_n(X) - f_0(X))^2] = O_P(r_n^{-1})$$

assuming the number of hidden units $K_n$ is chosen such that

$$K_n^{2+1/d} \log K_n = O(n)$$

- Outline of proof
  - Estimation rate implies tightness of $h_n = r_n(f_n - f_0)$
  - Rescaled and shifted criterion function converges weakly to Gaussian process
  - Gaussian process has a unique maximum at $h^*$
  - Argmax continuous mapping theorem
Theorem

Under the conditions of Theorem 1 and the null hypothesis,

\[ r_n^2 \lambda_j \rightarrow \int_X \left( \frac{\partial h^*(x)}{\partial x_j} \right)^2 d\mu(x) \]
Assume $\mu = P$ so that the test statistic

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

Under the conditions of Theorem 1 and the null hypothesis, the empirical test statistic satisfies

$$r_n^2 n^{-1} \sum_{i=1}^{n} \left( \frac{\partial f_n(X_i)}{\partial x_j} \right)^2 \Rightarrow \mathbb{E}_X \left[ \left( \frac{\partial h^*(X)}{\partial x_j} \right)^2 \right]$$
Identifying the asymptotic distribution

**Theorem**

Take $\mu = P$. If $\Theta$ admits an orthonormal basis $\{\phi_i\}$ that is $C^1$ and stable under differentiation, then

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

where $\{\chi_i^2\}$ are i.i.d. samples from the chi-square distribution, and where $\alpha_{i,j} \in \mathbb{R}$ satisfies $\frac{\partial \phi_i}{\partial x_j} = \alpha_{i,j} \phi_{k(i)}$ for some $k : \mathbb{N} \rightarrow \mathbb{N}$, and the $d_i$’s are certain functions of the $\alpha_{i,j}$’s.
Implementing the test

- Truncate the infinite sum at some finite order \( N \)
- Draw samples from the \( \chi^2 \) distribution to construct a sample of the approximate limiting law
- Repeat \( m \) times and compute the empirical quantile \( Q_{N,m} \) at level \( \alpha \in (0, 1) \) of the corresponding samples
  - If \( m = m_N \rightarrow \infty \) as \( N \rightarrow \infty \), then \( Q_{N,m_N} \) is a consistent estimator of the true quantile of interest
- Reject \( H_0 \) if \( \lambda_j^n > Q_{N,m_N}(1 - \alpha) \) such that the test will be asymptotically of level \( \alpha \):

\[
P_{H_0}(\lambda_j^n > Q_{N,m_N}(1 - \alpha)) \leq \alpha
\]
Simulation study

- 8 variables:

\[ X = (X_1, \ldots, X_8) \sim U(-1, 1)^8 \]

- Ground truth:

\[ Y = 8 + X_1^2 + X_2 X_3 + \cos(X_4) + \exp(X_5 X_6) + 0.1X_7 + \epsilon \]

where \( \epsilon \sim N(0, 0.01^2) \) and \( X_8 \) has no influence on \( Y \)

- Training (via TensorFlow): 100,000 samples \((Y_i, X_i)\)

- Testing: 10,000 samples

- Out-of-sample MSE:

<table>
<thead>
<tr>
<th>Model</th>
<th>Mean Squared Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear Regression</td>
<td>0.35</td>
</tr>
<tr>
<td>NN with ( K = 25 )</td>
<td>( 3.1 \cdot 10^{-4} \sim \text{Var}(\epsilon) )</td>
</tr>
</tbody>
</table>
### Linear model fails to identify significant variables

| Variable | coef  | std err | t       | $P > |t|$ |
|----------|-------|---------|---------|------|
| const    | 10.2297 | 0.002   | 5459.250 | 0.000 |
| 1        | -0.0031 | 0.003   | -0.964  | 0.335 |
| 2        | 0.0051  | 0.003   | 1.561   | 0.118 |
| 3        | -0.0026 | 0.003   | -0.800  | 0.424 |
| 4        | 0.0003  | 0.003   | 0.085   | 0.932 |
| 5        | 0.0016  | 0.003   | 0.493   | 0.622 |
| 6        | -0.0033 | 0.003   | -1.035  | 0.300 |
| 7        | 0.0976  | 0.003   | 30.059  | 0.000 |
| 8        | -0.0018 | 0.003   | -0.563  | 0.573 |

Only the intercept and the linear term $0.1X_7$ are identified as significant. The irrelevant $X_8$ is correctly identified as insignificant.
<table>
<thead>
<tr>
<th>Input Variable</th>
<th>Test Statistic</th>
<th>Power / Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.310</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0.332</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0.331</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>0.267</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>0.480</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>0.479</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>$1.010 \cdot 10^{-2} (= 0.1^2)$</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>$4.200 \cdot 10^{-6}$</td>
<td>0.13</td>
</tr>
</tbody>
</table>

The asymptotic distribution tends to underestimate the variance of the finite sample distribution of the test statistic.
Application: House price valuation

- **Data:** 120+ million housing sales from county registrar of deed offices across the US (source: CoreLogic)
- **Sample period:** 1970 to 2017
- **Geographical area:** Merced County, CA; 76,247 samples
- **Prediction** of $Y = \log$ sale price
- **Variables** $X$: Bedrooms, Full_Baths, Last_Sale_Amount, N_Originations, N_Past_Sales, Sale_Month, SqFt, Stories, Tax_Amount, Time_Since_Prior_Sale, Year_Built
- Training and gradients via TensorFlow
- Cross validation (80-20 split) suggests $K = 50$ nodes
- Test MSE is 0.60 vs. 0.85 for linear baseline model
Application: House price valuation

<table>
<thead>
<tr>
<th>Variable</th>
<th>Test Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sale Month</td>
<td>2.660</td>
</tr>
<tr>
<td>Last_Sale_Amount</td>
<td>0.768</td>
</tr>
<tr>
<td>N_Past_Sales</td>
<td>0.705</td>
</tr>
<tr>
<td>Year_Built</td>
<td>0.197</td>
</tr>
<tr>
<td>Tax_Amount</td>
<td>0.182</td>
</tr>
<tr>
<td>SqFt</td>
<td>0.088</td>
</tr>
<tr>
<td>Time_Since_Prior_Sale</td>
<td>0.061</td>
</tr>
<tr>
<td>Bedrooms</td>
<td>0.047</td>
</tr>
<tr>
<td>Full_Baths</td>
<td>0.043</td>
</tr>
<tr>
<td>Stories</td>
<td>0.028</td>
</tr>
<tr>
<td>N_Originations</td>
<td>0.0003</td>
</tr>
</tbody>
</table>

All variables but N_Originations are significant at the 5% level.
Conclusion

- We develop a statistical significance test for neural networks
- The test enables one to assess the impact of feature variables on the network’s prediction, and to rank variables according to their predictive importance
- We believe this is a significant step towards making neural nets explainable, and hope that it enables a broader range of applications in (financial) practice

Ongoing work
- Treatment of NN classifiers
- Treatment of deep networks
- Treatment of more complex network architectures
- Cross derivatives for testing interactions between variables
Example

- Suppose the elements of $X$ are i.i.d. uniform on $[-1, 1]$.
- Using the Fourier basis, the limiting distribution takes the form:

$$\frac{B^2}{\sum_{n \in \mathbb{N}^d} \frac{\chi_n^2}{d_n^2}} \sum_{n \in \mathbb{N}^d} \frac{n_j^2 \pi^2}{d_n^4} \chi_n^2,$$

- $n = (n_1, n_2 \ldots, n_j, \ldots, n_d)$
- $d_n^2 = \sum_{|\alpha| \leq \lfloor \frac{d}{2} \rfloor + 2} \prod_{k=1}^d (n_{n_k} \pi)^{2\alpha_k}$
- $\{\chi_n^2\}_{n \in \mathbb{N}^d}$ are i.i.d. chi-square variables.
Computing the asymptotic distribution

- We note that $\Theta$ is a subspace of the Hilbert space $L^2(P)$ which admits an orthonormal basis $\{\phi_i\}_{i=0}^{\infty}$.

- If this basis is $C^1$ and stable under differentiation, i.e. if there are a real $\alpha_{i,j}$ and a mapping $k : \mathbb{N} \rightarrow \mathbb{N}$ such that

$$\frac{\partial \phi_i}{\partial x_j} = \alpha_{i,j} \phi_k(i),$$

then there exists an invertible operator $D$ such that

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

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