# Feature Selection via Detecting Ineffective Features

Kris De Brabanter
Dep. Electrical Engineering
Kasteelpark Arenberg 10, 3001 Leuven
Katholieke Universiteit Leuven
kris.debrabanter@esat.kuleuven.be

László Györfi
Dep. Computer Science & Information Theory
Magyar Tudósok körútja 2., Budapest
Budapest University of Technology and Economics
gyorfi@szit.bme.hu

**Abstract**: Consider the regression problem with a response variable $Y$ and with a feature vector $\mathbf{X}$. For the regression function $m(\mathbf{x}) = \mathbf{E}\{Y \mid \mathbf{X} = \mathbf{x}\}$, we introduce a new and simple estimator of the minimum mean squared error $L^* = \mathbf{E}\{(Y - m(\mathbf{X}))^2\}$. Let $\mathbf{X}^{(-k)}$ be the feature vector, in which the $k$-th component of $\mathbf{X}$ is missing. In this paper we analyze a nonparametric test for the hypothesis that the $k$-th component is ineffective, i.e., $\mathbf{E}\{Y \mid \mathbf{X}\} = \mathbf{E}\{Y \mid \mathbf{X}^{(-k)}\}$ a.s.

**Keywords**: feature selection, minimum mean squared error, hypothesis test

## 1 Introduction

Let the label $Y$ be a real valued random variable and let the feature vector $\mathbf{X} = (X_1, \ldots, X_d)$ be a $d$-dimensional random vector. The regression function $m$ is defined by

$$m(\mathbf{x}) = \mathbf{E}\{Y \mid \mathbf{X} = \mathbf{x}\}.$$

The minimum mean squared error, called also variance of the residual $Y - m(\mathbf{X})$, is denoted by

$$L^* := \mathbf{E}\{(Y - m(\mathbf{X}))^2\} = \min_f \mathbf{E}\{(Y - f(\mathbf{X}))^2\}.$$

The regression function $m$ and the minimum mean squared error $L^*$ cannot be calculated when the distribution of $(\mathbf{X}, Y)$ is unknown. Assume, however, that we observe data

$$D_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$$

consisting of independent and identically distributed copies of $(\mathbf{X}, Y)$. $D_n$ can be used to produce an estimate of $L^*$. Nonparametric estimates of the minimum mean squared error are given in [2, 4].

## 2 New estimate of the minimum mean squared error

One can derive a new and simple estimator of $L^*$ by considering the definition

$$L^* = \mathbf{E}\{(Y - m(\mathbf{X}))^2\} = \mathbf{E}\{Y^2\} - \mathbf{E}\{m(\mathbf{X})^2\}. \quad (1)$$

The first and second term on the right-hand-side of (1) can be estimated by $\frac{1}{n}\sum_{i=1}^n Y_i^2$ and $\frac{1}{n}\sum_{i=1}^n Y_i Y_{n,i,1}$ respectively where $Y_{n,i,1}$ denotes the labels of the first nearest neighbors of $\mathbf{X}_i$ among $\mathbf{X}_1, \ldots, \mathbf{X}_{i-1}, \mathbf{X}_{i+1}, \ldots, \mathbf{X}_n$. Therefore, the minimum mean squared error $L^*$ can be estimated by

$$\tilde{L}_n := \frac{1}{n}\sum_{i=1}^n Y_i^2 - \frac{1}{n}\sum_{i=1}^n Y_i Y_{n,i,1}. \quad (2)$$

One can show without any conditions that

$$\tilde{L}_n \to L^*$$

a.s. Moreover, for bounded $|Y|$ and $\|X\|$, and for Lipschitz continuous $m$, and for $d \geq 2$, we have (cf. [3])

$$\mathbf{E}\{|\tilde{L}_n - L^*|\} \leq c_1 n^{-1/2} + c_2 n^{-2/d}.$$

## 3 Feature Selection and Hypothesis Test

One way of feature selection would be to detect ineffective components of the feature vector. Let $\mathbf{X}^{(-k)} = (X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_d)$ be the $d-1$ dimensional feature vector such that we leave out the $k$-th component from $\mathbf{X}$. Then the corresponding minimum error is

$$L^{*(-k)} := \mathbf{E}\big\{Y - \mathbf{E}\{Y|\mathbf{X}^{(-k)}\}\big\}^2.$$

We want to test the following (null) hypothesis:

$$\mathcal{H}_k : L^{*(-k)} = L^*,$$

which means that leaving out the $k$-th component the minimum mean squared error does not increase. The hypothesis $\mathcal{H}_k$ means that

$$m(\mathbf{X}) = \mathbf{E}\{Y \mid \mathbf{X}\} = \mathbf{E}\{Y \mid \mathbf{X}^{(-k)}\} =: m^{(-k)}(\mathbf{X}^{(-k)})$$

a.s.

By using the data

$$D_n^{(-k)} = \{(\mathbf{X}_1^{(-k)}, Y_1), \ldots, (\mathbf{X}_n^{(-k)}, Y_n)\},$$

$L^{*(-k)}$ can be estimated by

$$\tilde{L}_n^{(-k)} := \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \frac{1}{n}\sum_{i=1}^{n} Y_i Y_{n,i,1}^{(-k)},$$

so the corresponding test statistic is

$$\tilde{L}_n^{(-k)} - \tilde{L}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i (Y_{n,i,1} - Y_{n,i,1}^{(-k)}).$$

We can accept the hypothesis $\mathcal{H}_k$ if

$$\tilde{L}_n^{(-k)} - \tilde{L}_n$$

is "close" to zero. Since with large probability the first nearest neighbors of $\mathbf{X}_i$ and of $\mathbf{X}_i^{(-k)}$ are the same, $Y_{n,i,1} - Y_{n,i,1}^{(-k)} = 0$ in the test statistic. We know that $\mathbf{P}(Y_{n,i,1} = Y_{n,i,1}^{(-k)})$ is decreasing as $n$ increases (and $d$ remains fixed) and vice versa, this probability is increasing as $d$ increases (while $n$ remains fixed). Hence, this test statistic is small even when the hypothesis $\mathcal{H}_k$ is not true.

To correct for this problem we modify the test statistic such that

$$(\hat{Y}_{n,i,1}, \hat{Y}_{n,i,1}^{(-k)}) = (Y_{n,i,1}, Y_{n,i,1}^{(-k)}) \quad \text{if} \quad Y_{n,i,1} \neq Y_{n,i,1}^{(-k)}$$

and

$$(\hat{Y}_{n,i,1}, \hat{Y}_{n,i,1}^{(-k)}) = I_i(Y_{n,i,2}, Y_{n,i,1}^{(-k)}) + (1-I_i)(Y_{n,i,1}, Y_{n,i,2}^{(-k)}),$$

otherwise (where $Y_{n,i,2}$ denotes the labels of the second nearest neighbors of $\mathbf{X}_i$ among $\mathbf{X}_1, \ldots, \mathbf{X}_{i-1}, \mathbf{X}_{i+1}, \ldots, \mathbf{X}_n$), with

$$I_i = \begin{cases} 0 & \text{with probability } 1/2, \\ 1 & \text{with probability } 1/2, \end{cases}$$

yielding

$$\hat{L}_n^{(-k)} - \hat{L}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i (\hat{Y}_{n,i,1} - \hat{Y}_{n,i,1}^{(-k)}).$$

As in classical hypothesis testing, we need to find the limit distribution of the test statistic. The main difficulty here is that $\hat{L}_n^{(-k)} - \hat{L}_n$ is an average of dependent random variables. However, this dependence has a special property, called exchangeable. Based on a central limit theorem for exchangeable arrays [1], we can show the following result.

**Theorem 1** *Under the conditions of [1, Theorem 2], we have that*

$$\sqrt{n}(\hat{L}_n^{(-k)} - \hat{L}_n) \xrightarrow{d} N(0, 2L^* \mathbf{E}\{Y^2\})$$

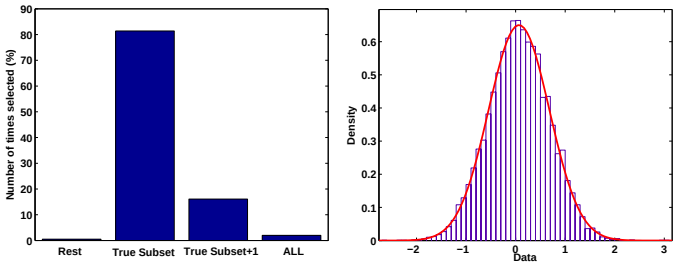
*under the null hypothesis* $\mathcal{H}_k$.

In the above theorem, $L^*$ and $\mathbf{E}\{Y^2\}$ can be estimated by (2) and $\frac{1}{n}\sum_{i=1}^{n} Y_i^2$ respectively. Note that such a results is quite surprising, since under $\mathcal{H}_k$ the smoothness of the regression function $m$ and the dimension $d$ do not count.

## 4 Simulations

First, consider the following nonlinear function with 4 uniformly distributed inputs on $[0,1]^4$ with $n = 1,000$: $Y = \sin(\pi X^{(1)})\cos(\pi X^{(4)}) + \varepsilon$, with $\varepsilon \sim N(0, 0.1^2)$. Figure 1(a) illustrates the frequency of the true selected subset, true subset with additional component and full subset selected by the proposed test procedure during 1,000 runs. The significance level is set to 0.05.

Second, we experimentally verify Theorem 1 by means of bootstrap (10,000 replications). Consider the following five dimensional function with additive noise: $Y = \sum_{i=1}^{5} c_i X^{(i)} + \varepsilon$, where $c_1 = 0$ and $c_i = 1$ for $i = 2, \ldots, 5$. Let $\mathbf{X}$ be uniform on $[0,1]^5$ and $\varepsilon \sim N(0, 0.05^2)$. Figure 1(b) shows the histogram of $\sqrt{n}(\hat{L}_n^{(-k)} - \hat{L}_n)$ under the null hypothesis i.e., $\mathcal{H}_k$ for $k = 1$. A Kolmogorov-Smirnov test confirms this result.




(a) (b)

**Fig. 1:** (a) Illustration of the frequency of the true selected subset, true subset with additional component and full subset. Rest denotes at least one component of the true subset is selected; (b) Density histogram of $\sqrt{n}(\hat{L}_n^{(-k)} - \hat{L}_n)$ under the null hypothesis with corresponding Normal fit.

## 5 Conclusion

We have presented a simple nonparametric hypothesis test for detecting ineffective features. The simulation shows the capability of the proposed methodology.

**Acknowledgments**

Kris De Brabanter is supported by an FWO fellowship grant. László Györfi was partially supported by the European Union and the European Social Fund through project FuturICT.hu (grant no.: TAMOP-4.2.2.C-11/1/KONV-2012-0013).

## References

[1] J.R. Blum, H. Chernoff, M. Rosenblatt & H. Teicher. Central limit theorems for interexchangeable processes. *Canad. J. Math.*, 10:222–2229, 1958.

[2] L. Devroye, D. Schäfer, L. Györfi & H. Walk. The estimation problem of minimum mean squared error. *Statistics and Decisions*, 21(1): 15-28, 2003.

[3] L. Devroye, P. Ferrario, L. Györfi & H. Walk. Strong universal consistent estimate of the minimum mean squared error. *Submitted*, 2013.

[4] E. Liitiäinen, F. Corona & A. Lendasse. Residual variance estimation using a nearest neighbor statistic. *J. Multivariate Anal.*, 101(4): 811-823, 2010.