Asymptotic behavior of encompassing test for independent processes: Case of linear and nearest neighbor regressions

Patrick Rakotomarolahy

Abstract: Encompassing test has been well developed for fully parametric modeling. In this study, we are interested on encompassing test for parametric and nonparametric regression methods. We consider linear regression for parametric modeling and nearest neighbor regression for nonparametric methods. We establish asymptotic normality of encompassing statistic associated to the encompassing hypotheses for the linear parametric method and the nonparametric nearest neighbor regression estimate. We also obtain convergence rate depending only on the number of neighbors $k$ while it depends on the number of observation $n$ and the bandwidth $h_n$ for kernel method. We achieve the same convergence rate when $h_n = k/n$. Moreover, asymptotic variance of the encompassing statistic associated to kernel regression depends on the density, this is not the case for nearest neighbor regression estimate.

Keywords: encompassing test; functional parameter; linear regression; nearest neighbor technique; asymptotic normality

1. Introduction

Encompassing tests lie on model selection step. They are used for detection of redundant models among admissible models. In that case an encompassing model is intended to account for the

ABOUT THE AUTHOR

Patrick Rakotomarolahy is the assistant professor in the department of mathematics and their applications at Fianarantsoa University. He has completed his Bsc and Msc in applied mathematics. He received his doctorate at the Panthéon-Sorbonne Paris 1 University. His current researches are in statistical model selection and in modeling macroeconomic and financial variables. He focuses especially on issues about model selection between parametric and nonparametric techniques. This study is in-line with this direction as the findings on asymptotic behavior of encompassing tests allow us to detect redundant models.

PUBLIC INTEREST STATEMENT

Regression techniques are used for quantitative analysis method in many fields such as in economic or financial modeling. They are a useful tool for identification of factors, which may explain the evolution of any variable of interest. In economic modeling for example, when we want to analyze the evolution of Gross Domestic Product or GDP, it might be affected by many variables like interest rate, inflation, exchange rate, sentiment indicators. Researchers or experts may face several admissible models from parametric or/and nonparametric regression methods. Encompassing test can be helpful for detection of redundant models among admissible models. The findings in this study contribute on encompassing test between linear and nearest neighbor regression estimates.
results found by encompassed model. **Theoretical** development on encompassing test can be found in Mizon (1984), Gouriéroux and Monfort (1995) and Florens et al. (1996). For their application, we refer readers to the general to specific computer based model selection procedure, Hendry and Doornik (1994).

Recently, Bontemps et al. (2008) have developed encompassing test for linear parametric against kernel nonparametric regression methods. They provide asymptotic normality of the associated encompassing statistics under the independent and identically distributed hypothesis (i.i.d.). As stated in Hendry et al. (2008) that the work of Bontemps et al. (2008) is the starting treatment of encompassing tests to functional parameter based on nonparametric methods.

We extend this result to nearest neighbor regression method, which has been claimed more flexible compared to kernel. Other motivation would be its interest in application like in Nowman and Saltoglu (2003), Guégan and Huck (2005), Ferrara et al. (2010), Guégan and Rakotomarolahy (2010), and Puspitasari and Rustam (2018), among others.

In the next section, we provide an overview of the encompassing test. After, we establish asymptotic normality for various encompassing statistics associated to linear parametric and nearest neighbor regression methods. Last, we conclude.

### 2. Encompassing test for independent processes

This section introduces the encompassing test and then builds the corresponding encompassing hypothesis. So, given two regression models $\mathcal{M}_1$ and $\mathcal{M}_2$, we are interested in knowing if model $\mathcal{M}_1$ can account the result of model $\mathcal{M}_2$. In fact, we want to know if $\mathcal{M}_1$ encompasses $\mathcal{M}_2$ or, in a short notation $\mathcal{M}_1 \supset \mathcal{M}_2$. Testing such a hypothesis will be done using the notion of encompassing test.

Generally speaking, model $\mathcal{M}_1$ encompasses model $\mathcal{M}_2$, if the parameter $\theta_{\mathcal{M}_1}$ of the latter model can be expressed in function of the parameter $\theta_{\mathcal{M}_2}$ of the former model. In other words, let $\Delta(\theta_{\mathcal{M}_1})$ be the pseudo-true value of $\theta_{\mathcal{M}_2}$ on $\mathcal{M}_1$. In general, the pseudo-true value is defined as the plim of $\hat{\theta}_{\mathcal{M}_2}$ on $\mathcal{M}_1$. For more discussion on pseudo-true value associated with the KLIC, we refer to Sawa (1978) and Govaerts et al. (1994). The encompassing statistic is given by the difference between $\hat{\theta}_{\mathcal{M}_2}$ and $\Delta(\hat{\theta}_{\mathcal{M}_1})$ scaled by a coefficient $a_n$.

Let $S = (Y, X, Z)$ be a zero mean random process with valued in $\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^q$ where $d, q \in \mathbb{N}$. For $x \in \mathbb{R}^d$ and $z \in \mathbb{R}^q$, we consider the two models $\mathcal{M}_1$ and $\mathcal{M}_2$ defined as follows:

$$\mathcal{M}_1 : \quad m(x) = E[Y|X = x] \quad \text{and} \quad \mathcal{M}_2 : \quad g(z) = E[Y|Z = z].$$  \hspace{1cm} (1)

In addition, the general unrestricted model is given by $r(x, z) = E[Y|X = x, Z = z]$. Following the encompassing test for functional parameter in Bontemps et al. (2008), we have the null hypothesis:

$$\mathcal{H} : E[Y|X = x, Z = z] = E[Y|X = x].$$

This null states that $\mathcal{M}_1$ is the owner model, and $\mathcal{M}_2$ will be served on validating this statement and is called the rivel model. We test this hypothesis $\mathcal{H}$ through the following implicit encompassing hypothesis:

$$\mathcal{H}^* : E[E[Y|X = x]|Z = z] = E[Y|Z = z].$$

The following homoskedasticity condition will be assumed all along this work:

$$\text{Var}[Y|X = x, Z = z] = \sigma^2.$$  \hspace{1cm} (2)
Moreover, a necessary condition for the encompassing test relies on the errors of both models where the intended encompassing model \( \mathcal{M}_1 \) should have smaller standard error than the encompassed model \( \mathcal{M}_2 \).

Given a sample, of size \( n \), \( s_i = (y_i, x_i, z_i) \) for \( i = 1, \ldots, n \) as realization of the random process \( S = (Y, X, Z) \). We suppose that \( s_i, i = 1, \ldots, n \) are i.i.d.. Then, for given functional estimates \( m_n \) and \( g_n \) of the functions \( m \) and \( g \), respectively, we have the following encompassing statistic:

\[
\hat{\delta}_{m, g} = g_n - \hat{G}(m_n),
\]

where \( \hat{G}(m_n) \) is an estimate of the pseudo-true value, associated with \( g_n \) on \( \mathcal{H} \), in the LHS of the hypothesis \( \mathcal{H} \). Bontemps et al. (2008) has provided asymptotic normality of this encompassing statistic \( \hat{\delta} \) by considering kernel regression estimate for nonparametric method. This result can be extended to nearest neighbor regression estimate but of course with different assumptions.

For nearest neighbor regression estimate, we consider the representation in Mack (1981), that is the \textit{k nearest neighbor (or k-NN)} estimate \( g_n \) of \( g \) is given by:

\[
g_n(z) = \frac{1}{nR_n} \sum_{i=1}^{n} w\left(\frac{z - z_i}{R_n}\right) Y_i,
\]

where \( R_n \) will be defined as distance, according to the Euclidean norm in \( \mathbb{R}^d \), from \( z \) to its \( k(n) \)th neighbors, and \( w(u) \) is a bounded, non-negative weight function satisfying

\[
w(u)du = 1 \quad \text{and} \quad w(u) = 0 \quad \text{for} \quad |u| > 1.
\]

To establish an asymptotic distribution of \( \hat{\delta}_{m, g_n} \), we need some assumptions. The following assumptions will be used for insuring the asymptotic normality and are taken from Mack (1981).

Without loss of generality, the function \( f \) will be a marginal density or a conditional density or a joint density according to the variables on its arguments.

The first assumption relies on the density function of the couple \((Y, Z)\).

**Assumption 1.** The function \( \chi_{\beta}(z) = \int y^\beta f(z, y) dy \) is bounded and continuous at \( z \) for \( \beta = 0, 1, 2, \) and continuously differentiable in a neighborhood of \( z \) for \( \beta = 0, 1, 2, \).

The following assumption concerns conditions on the moments up to order three of the variable of interest.

**Assumption 2.** \( E[|Y|^3] < \infty, \ Var[Y|Z = z] > 0 \) and \( f(z) > 0 \).

The last assumption states conditions on the relationship between the number of neighbors \( k \) and the sample size \( n \).

**Assumption 3.** \( k = n^a \) with \( 0 < a < \frac{4}{d+2} \).

When assumptions 1–3 hold and the relation (3) is satisfied, then Mack (1981) has established the asymptotic normality of the centered k-NN regression of \( g_n \). Moreover, **under assumption 3**, the bias of such k-NN regression estimate vanishes to zero.

Without loss of generality, we proceed as previously when model \( \mathcal{M}_1 \) will be estimated by k-NN regression method. In the rest of the paper, \( \mathcal{N}(\mu, v) \) denotes the normal distribution with mean \( \mu \) and variance \( v \). We now present the asymptotic normality of the encompassing statistic.
3. Asymptotic normality of the encompassing statistic

In general, \( M_1 \) or \( M_2 \) can be estimated using nonparametric or parametric regression methods. We can encounter the following four situations: \( M_1 \) and \( M_2 \) are both estimated parametrically, \( M_1 \) and \( M_2 \) are both estimated nonparametrically, \( M_1 \) is estimated nonparametrically and \( M_2 \) parametrically and \( M_1 \) is estimated parametrically and \( M_2 \) nonparametrically.

For development on the asymptotic behavior of the encompassing statistic for fully parametric case, i.e the two models \( M_1 \) and \( M_2 \) have parametric specification, we refer readers to Gouriéroux et al. (1983) and Mizon and Richard (1986) among others. For recent discussion on this encompassing test for fully parametric case, Bontemps et al. (2008) is a good reference.

Next, we will study the completely nonparametric case.

3.1. Nonparametric specification for \( M_1 \) and \( M_2 \)

We consider the case where the two models \( M_1 \) and \( M_2 \) defined in (1) are estimated using the nonparametric nearest neighbor regression method. To test the hypothesis" \( M_1 \) encompasses \( M_2 \)", we establish asymptotic normality of the associated encompassing statistic.

**Theorem 3.1.** Assume that assumptions 1–3 and relations (2) and (3) hold. Then under \( \mathcal{H}_1 \), we have:

\[
\sqrt{k - 1} \hat{\delta}_{m,g_n}(z) \rightarrow \mathcal{N}(0, c \text{Var}(\epsilon/Z = z) \int w^2(u)du) \quad \text{in distribution as} \quad n \to \infty,
\]

where \( \epsilon_i = Y_i - m(x_i) \) for \( i = 1, \ldots, n \) are the residuals from model \( M_1 \) and \( c = \frac{e^{\frac{1}{2}}}{\Gamma\left(\frac{1}{2}\right)} \) is the volume of unit ball in \( \mathbb{R}^n \) with \( \Gamma(.) \) the gamma function.

**Proof of Theorem 3.1**

The proof will be based on the decomposition of the encompassing statistic into two parts as an expression of nearest neighbor regression and a kind of bias. Before all, let’s denote by:

\[
W\left(\frac{z - Z_i}{R_n}\right) = \frac{1}{mR_n} \tilde{w}\left(\frac{z - Z_i}{R_n}\right).
\]

We write down our encompassing statistic by replacing our estimates \( g_n \) and \( \tilde{g}(m_n) \) at a given point \( z \), and we have:

\[
\sqrt{k - 1} \hat{\delta}_{m,g_n}(z) = \sqrt{k - 1} \sum_{i=1}^{n} W\left(\frac{z - Z_i}{R_n}\right) Y_i - \sqrt{k - 1} \sum_{i=1}^{n} W\left(\frac{z - Z_i}{R_n}\right) m_n(x_i)
= \sqrt{k - 1} \sum_{i=1}^{n} W\left(\frac{z - Z_i}{R_n}\right) \epsilon_i + \sqrt{k - 1} \sum_{i=1}^{n} W\left(\frac{z - Z_i}{R_n}\right) \left(m(x_i) - m_n(x_i)\right) = A + B,
\]

where \( A \) is the first expression in RHS of the equality. This involves a \( k \)-NN regression of \( \epsilon_i = Y_i - m(x_i) \) on \( Z_i \) scaled by the coefficient \( \sqrt{k - 1} \) seeing as convergence speed rate when \( n \) goes to infinity. Using Mack (1981), under assumptions 1–3 and when relation (3) holds, we have:

\[
A \rightarrow \mathcal{N}(0, c \text{Var}(\epsilon/Z = z) \int w^2(u)du) \quad \text{in distribution as} \quad n \to \infty.
\]

Next, for the second expression \( B \), we can bound by taking its supremum with respect to \( x_i \) and then we get:
\[ |B| \leq \sup_{x} \sqrt{k - 1} |m_n(x) - m(x)| \leq \sup_{x} \sqrt{k - 1} |m_n(x) - E[m_n(x)]| + \sup_{x} \sqrt{k - 1} |E[m_n(x)] - m(x)| = B_1 + B_2. \]  

(4)

When using the expression of the bias, Theorem 1 in [2], \( B_2 \) becomes:

\[
B_2 = (\sup_{x} A(x_i)) \left( \frac{k^{\frac{1}{2}}}{n} \sqrt{k - 1} + o \left( \frac{k^{\frac{1}{2}}}{n} \sqrt{k - 1} + O \left( \frac{1}{n} \right) \right) \right) \sqrt{k - 1},
\]

where \( A(.) \) is a function depending only on \( x_i \) and its expression can be found in Mack (1981). Then from Assumption 3, \( B_2 \) vanishes to zero when \( n \to \infty \). It remains on showing that \( B_1 \) goes to zero also. This can be achieved using result of Mukerjee (1993) extension of Cheng’s work (Cheng, 1984). Therefore, we remark that when the number of neighbors \( k \) increases more the weights given to neighbors decrease, then rewriting \( m_n(x_i) \) and we have the following equivalence:

\[
m_n(x_i) = \frac{\sum_{j=1}^{n} K \left( \frac{x_i - x_j}{\sqrt{n}} \right) y_j}{\sum_{j=1}^{n} K \left( \frac{x_i - x_j}{\sqrt{n}} \right)} \approx \sum_{j=1}^{\infty} c_j y_j,
\]

where \( K(.) \) is a given weight function which satisfies condition (3), \( c_j \) is a bounded weight equal to zero when \( j \) is larger than the number of neighbors and \( R_i \) is the distance between \( x_i \) and its \( k^{th} \) neighbor. When we denote by \( m_n(x_i) = \sum_{j=1}^{n} \frac{c_j}{R_i} y_j \), then from Theorem 2.1 in Mukerjee (1993), we have:

\[
B_1 = \sup_{x} |m_n(x_i) - E[m_n(x_i)]| = O \left( \frac{1}{n^r} \right) + O \left( n^{-\theta} \right),
\]

with \( r > 1 \) and \( \theta \), a positive sequence which tends to zero as \( n \to \infty \). So we get \( |B| \) converges to zero in probability as \( B_1 \). This completes the proof of theorem.

Next, we will consider the mixed situation where the owner model has parametric specification and the rival is from nonparametric method.

### 3.2. Parametric modelling for \( M_1 \) vs nonparametric specification for \( M_2 \)

In this section, we consider the case that model \( M_1 \) is a linear parametric model and \( M_2 \) is estimated by nearest neighbor regression technique. Therefore, the hypothesis \( \mathcal{H} \) will have linear parametric specification. The encompassing statistic associated to the null \( M_1 \) can be rewritten as follows:

\[
\hat{g}(\beta), g_n(z) - \hat{G}_L(\hat{\beta})(z),
\]

where \( \hat{g}(\hat{\beta}) \) is an estimate of the pseudo-true value \( G_\beta(\beta)(z) \) associated with \( g_n \) on \( \mathcal{H} \), and is defined as \( G_\beta(\beta)(z) = \beta \mathbb{E}[Z | Z = z] \).

We estimate the rival model \( M_2 \) using k-NN regression method where the owner model \( M_1 \) is still with linear parametric specification. The following theorem provides the asymptotic normality of the encompassing statistic introduced in relation (5).

**Theorem 3.2.** Assume that assumptions 1–3, relations (2) and (3) hold. Then under \( \mathcal{H} \), we get:

\[
\sqrt{k - 1} \hat{g}(\beta) \rightarrow N(0, \Sigma) \quad \text{in distribution as} \quad n \to \infty,
\]

where \( \Sigma = \text{cov} \left[ w_u^2(u) du \right] \) with \( c \) is the volume of unit ball in \( \mathbb{R}^q \).

**Proof of Theorem 3.2.**
When the owner model $\mathcal{M}_1$ is the linear regression parametric and the rival model $\mathcal{M}_2$ is the k-NN regression, we write the encompassing statistic as follows:

$$
\sqrt{n} \hat{\delta}_{\hat{\gamma}, \hat{\mu}}(z) = \sqrt{k - 1} \left( g_n(z) - \hat{G}_n(\hat{\mu}(z)) \right) = \sqrt{k - 1} \left( \sum_{i=1}^{n} \tilde{W} \left( \frac{Z_i - \hat{Z}_i}{R_n} \right) Y_i - \sum_{i=1}^{n} \tilde{W} \left( \frac{Z_i - \hat{Z}_i}{R_n} \right) \beta \hat{X}_i \right) 
$$

$$
= \sqrt{k - 1} \sum_{i=1}^{n} \tilde{W} \left( \frac{Z_i - \hat{Z}_i}{R_n} \right) (Y_i - \beta \hat{X}_i) + \sqrt{k - 1} \sum_{i=1}^{n} \tilde{W} \left( \frac{Z_i - \hat{Z}_i}{R_n} \right) (\beta - \tilde{\beta}) \hat{X}_i 
$$

$$
\quad + \sqrt{k - 1} \sum_{i=1}^{n} \tilde{W} \left( \frac{Z_i - \hat{Z}_i}{R_n} \right) \beta \hat{X}_i - \sqrt{k - 1} \sum_{i=1}^{n} \tilde{W} \left( \frac{Z_i - \hat{Z}_i}{R_n} \right) \beta \hat{X}_i = N_1 + N_2 + N_3 - N_4.
$$

(6)

where $\tilde{W} \left( \frac{Z_i - \hat{Z}_i}{R_n} \right)$ is the weight associated to the nearest neighbor regression of $\hat{\beta} \hat{X}_i$ on $Z_i$ and $R_n$ is the distance from $z$ to its $k$th neighbor.

We remark that $Y_i$ and $\beta \hat{X}_i$ as fitted values of $Y_i$, would have the same $Z_i$ nearest to $z$. We then have $N_3 - N_4 = 0$. Otherwise, this can happen asymptotically, that is $Y_i$ and $\hat{\beta} \hat{X}_i$ as fitted values of $Y_i$ have the same $Z_i$ nearest to $z$ when $k$ and $k$ tend to infinity. Thus, $N_3$ is asymptotically equivalent to $N_4$.

For the first expression $N_1 = \sqrt{k - 1} \sum_{i=1}^{n} \tilde{W} \left( \frac{Z_i - \hat{Z}_i}{R_n} \right) \epsilon_i$, with $\epsilon_i = Y_i - \beta \hat{X}_i$. Under assumptions in Theorem 3.2, then using result in Mack (1981), we have:

$$
N_1 \to N(0, \Sigma) \text{ in distribution as } n \to \infty,
$$

where $\Sigma = c \sigma^2 \int \tilde{W}(u) du$.

For $N_2 = (\beta - \tilde{\beta}) \sqrt{k - 1} \sum_{i=1}^{n} \tilde{W} \left( \frac{Z_i - \hat{Z}_i}{R_n} \right) X_i$, under assumptions in Theorem 3.2, we know that the estimate $\sqrt{n}(\beta - \tilde{\beta})$ converges in distribution to a normal law $Z$ with mean zero. The remaining part of $N_2$ has the following expression $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{W} \left( \frac{Z_i - \hat{Z}_i}{R_n} \right) X_i$, which converges in distribution to zero. Thus, from Slutsky’s theorem, $N_2$ tends to zero in distribution.

We will consider the last case where the owner model $\mathcal{M}_1$ is a nonparametric method and the rival model $\mathcal{M}_2$ is a linear parametric model.

3.3. Nonparametric specification for $\mathcal{M}_1$ vs parametric modelling for $\mathcal{M}_2$

We now consider the owner model $\mathcal{M}_1$ to be estimated using a k-NN nonparametric regression and the rival model $\mathcal{M}_2$ to be a linear parametric method. Therefore, the encompassing statistic associated to the null $\mathcal{M}_1 \in \mathcal{M}_2$ is given by:

$$
\hat{\delta}_{\gamma, \mu}(m_n) = \hat{\gamma} - \hat{\gamma}(m_n),
$$

(7)

where $\hat{\gamma}(m_n)$ is an estimate of the pseudo-true value $\gamma(m)$ associated with $\hat{\gamma}$ on $\mathcal{M}_1$, which is defined by $\gamma(m) = \mathbb{E}[Z|Y]^{-1} \mathbb{E}[Z|m]$. We estimate the unknown conditional mean $m$ associated to the model $\mathcal{M}_1$ using k-NN regression estimate. We state in the following theorem the asymptotic normality of the encompassing statistic in relation (7). For precision, we use the assumptions introduced in previous section for k-NN regression estimate $m_n$ instead of $g_n$.

**Theorem 3.3.** Assume that relations (2) and (3), assumptions 1–3, and the regularity conditions in linear regression are satisfied.

Then under $\mathcal{H}_0$, we get:

$$
\sqrt{n} \hat{\delta}_{\gamma, \mu} \to N(0, \Omega) \text{ in distribution as } n \to \infty,
$$
where $\Omega = \sigma^2 (E(ZZ^\top))^{-1}$.

**Proof of Theorem 3.3.**

When the functional parameters $m_n$ is from $k$-NN regression estimate, we rewrite the associated encompassing statistic as follows:

$$
\sqrt{n} \hat{m}_n = \sqrt{n}(\hat{y} - \hat{y}(m_n)) = \sqrt{n}\left( \left( \frac{1}{n} \sum_{i=1}^{n} Z_i Z_i^\top \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^{n} Z_i Y_i \right) - \left( \frac{1}{n} \sum_{i=1}^{n} Z_i Z_i^\top \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^{n} Z_i m_n(x_i) \right) \right)
= \sqrt{n}\left( \frac{1}{n} \sum_{i=1}^{n} Z_i Z_i^\top \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^{n} Z_i (Y_i - m(x_i)) \right) + \sqrt{n}\left( \frac{1}{n} \sum_{i=1}^{n} Z_i Z_i^\top \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^{n} Z_i (m(x_i) - m_n(x_i)) \right)
= L_1 + L_2,
$$

where $L_1$ corresponds to the first expression in the RHS of the equality (8). It coincides to the linear regression of the error $\epsilon$ (with $\epsilon_i = Y_i - m(x_i)$) on $Z$.

**Under i.i.d. assumption in Theorem 3.3, $L_1$ converges in distribution to $Z$ where $Z$ is normally distributed with mean zero and variance $\Omega = \sigma^2 (E(ZZ^\top))^{-1}$. For the second expression $L_2$, we bound it by taking the maximum with respect to $x_i$ and then we get $|L_2| \leq \sqrt{n}S_nD_n\sup_{x_i \in \mathbb{R}^\delta} |m(x) - m_n(x_i)|$, $x_i \in \mathbb{R}^\delta$ where $S_n = \frac{1}{n} \sum_{i=1}^{n} Z_i|Z_i|$ and $D_n = (\frac{1}{n} \sum_{i=1}^{n} Z_i Z_i^\top)^{-1}$. We remark that $\sqrt{n}S_nD_n\sup_{x_i \in \mathbb{R}^\delta} |m(x) - m_n(x_i)|$ is asymptotically equivalent to the bound of [8] in Equation 4 which converges to zero in probability. Thus, the product vanishes to zero also from Slutsky’s theorem. This completes the proof.

**4. Illustration**

In this section, we illustrate our theoretical results on real data. We focus on socio-economic factor determinants of Life expectancy. As explanatory variables for Life expectancy at birth, we consider the Gross National Income per capita in US $\$, the Gross Domestic Product per capita in US $\$, and the government health expenditure per capita in US $\$. Impact of these variables on Life expectancy at birth has been analyzed a long way in the literature, for regression analysis we may look at Hussain (2002) and Ali and Ahmad (2014). We use cross sectional data for 169 countries in 2017, which have been collected from the United Nation and the World Health Organization websites. To start our empirical study, we compute some basic statistics.

The highest life expectancy hits 84 years and belongs to Japan. While, the lowest is around 52 years belonging to Central African Republic. The best life expectancy of 84 years would be remarkable. Besides, life expectancy mean 72 years seems interesting. Moreover, the median value 73.69 indicates that around 84 countries have life expectancy above 73 years, largely beyond the retirement ages.

For socio-economic variables; Luxembourg, Switzerland and USA have the highest GDP, Income and health expenditure per capita, respectively. Burundi has the lowest GDP and Income per capita. Congo Democratic Republic registers the lowest government spending on

| Table 1. Summary statistics |
|-----------------------------|
|                             | Life expectancy | GDP     | Income  | Health expenditure |
| Min.                        | 52.24           | 293     | 280     | 19.4               |
| Median                      | 73.69           | 5395    | 5110    | 307.2              |
| Mean                        | 72.10           | 13,769  | 12,809  | 1086.0             |
| Sd                          | 7.664666        | 18,967.33 | 17,387.52 | 1829.39 |
| Max.                        | 84.10           | 107,361 | 81,120  | 10,246.1           |
health care. These variables exhibit some common behaviors such as the median of each variable is around fifteen times of its minimum and one over fifteen times of its maximum. They also have high dispersion. We now proceed on analysis of their relationship with life expectancy.

Let compute the correlation coefficients between life expectancy and the predictor variables.

From Table 2, Life expectancy has positive and high correlation with each explanatory variables. Such correlations indicate that higher GDP, income or expenditure on health will link with longer life expectancy. This preliminary analysis could motivate us on exploring other statistic and econometric analysis of the relationship between life expectancy and the three socio-economic variables. We will use the linear and the nearest neighbor regression methods. In sequel, we will work on demeaned and scaled (by a factor \( \frac{1}{\max - \min} \)) variables.

For the linear regression, we explain life expectancy at birth \( Y \) by health expenditure per capita \( X \), gross income per capita \( Z \) and GDP per capita \( W \). Considering several combination of these explanatory variables, following we summarize regression coefficient estimates with their standard errors in parenthesis.

\[
\begin{align*}
M_1 : Y_i &= 0.8X_i + \hat{u}_1^i \\
M_2 : Y_i &= 0.76Z_i + \hat{u}_2^i \\
M_3 : Y_i &= 0.89W_i + \hat{u}_3^i \\
M_4 : Y_i &= -0.26X_i + 0.96Z_i + \hat{u}_4^i \\
M_5 : Y_i &= 0.02X_i + 0.86W_i + \hat{u}_5^i \\
M_6 : Y_i &= -0.28X_i + 1.24Z_i - 0.32W_i + \hat{u}_6^i,
\end{align*}
\]

where \( \hat{u}_j^i \) an estimate of the error term of model \( M_j, j = 1, \ldots, 6 \).

Coefficients of models \( M_1, M_2 \) and \( M_3 \) are all significant. In contrast, models \( M_4 \) and \( M_6 \) nest to model \( M_2 \) due to non-significance of \( X \) and \( Z \) coefficient estimates. Besides, \( M_5 \) nests to \( M_3 \) as \( X \)'s coefficient estimate is not significant. We then focus our analysis on models \( M_1, M_2 \) and \( M_3 \) and proceed on their diagnostics. Results are reported in Table 3.

| Table 2. Correlation between life expectancy and the socio-economic variables |
|--------------------------------|----------------|----------------|
|                               | GDP          | Income        | Health expenditure |
| Life expectancy               | 0.6556506    | 0.6705147     | 0.6017740          |

| Table 3. Regression diagnostics |
|---------------------------------|----------------|----------------|----------------|
|                                 | Models         |                |                |
|                                 | \( M_1 \)      | \( M_2 \)      | \( M_3 \)      |
| Breusch-Pagan heteroscedastic test | 0.077          | 3.761          | 0.681          |
|                                  | (0.781)        | (0.052)        | (0.409)        |
| Pearson’s correlation test       | 1.67e-16       | 3.4e-16        | -1.27e-16      |
| for \( u \) and \( X \)          | (1)            | (1)            | (1)            |
| Residuals mean                   | 4.03e-18       | 7.33e-18       | 1.49e-17       |
| Residuals standard error         | 0.192          | 0.179          | 0.182          |

Notes: we report associated statistics with their p-values in parenthesis.
From Table 3, we accept the homoscedasticity property of residuals and their non-correlation with predictors. In addition, residuals of the three models have zero mean. Thus, our three models meet standard assumptions on linear regression.

$M_1$, $M_2$ and $M_3$ are non-nested models. Thus, the decision on choosing one model will be based on encompassing test. A necessary condition is that the encompassing model should fit better than encompassed model. Therefore, encompassing model is expected to have smaller error variance than its rival. The standard errors of models $M_i$, $i = 1, 2, 3$ are $\sigma_1 = 0.192$, $\sigma_2 = 0.179$ and $\sigma_3 = 0.182$, respectively. Then, among the three models, $M_1$ has the worst fit and $M_2$ has the best fit. We report in Table 4 various encompassing tests associated to models $M_1$, $M_2$ and $M_3$.

From Table 4, we accept the null $M_2 \not\subseteq M_3$ and $M_3 \not\subseteq M_1$ that is, $M_2$ encompasses $M_3$ and $M_3$ encompasses $M_1$. In contrast, we reject $M_3 \not\subseteq M_2$ and $M_1 \not\subseteq M_3$, there are no mutual encompassing. Thus, we retain model $M_2$ as it also has the smallest standard error. We will re-examine the link between life expectancy and explanatory variables using nearest neighbor regression.

For $k$-NN regression of life expectancy, we need the specification of the weighting function $w(\cdot)$ and the estimation of the parameter $k$. Two weighting functions have been mostly used in the literature: the exponential function $\frac{\exp(-|z-Z_i|^2)}{\sum_{i=1}^{k} \exp(-|z-Z_i'|^2)}$ with $(Z_i)_i = 1,...,k$ the $k$ nearest to $z$, and the uniform function $\frac{1}{k}$. We also consider these two weighting functions.

Assumption 3 states that the number $k$ should satisfy $1 < k = n^{\alpha} < \frac{n}{d^{\beta}}$ for $n$ observations and $d$ explanatory variables. Then, as $n = 169$, we have maximum values for $k$ which are 60, 30, and 18 for $d = 1$, $d = 2$ and $d = 3$ respectively. We estimate this parameter $k$ by minimizing the root mean squared error (RMSE). Results are summarized in Table 5 where we keep the following notation already used in linear regression: $X$ for health expenditure per capita, $Z$ for gross income per capita and $W$ for GDP per capita.

For models $M_7$ to $M_{11}$, model $M_{10}$ has the lowest standard error. We also remark that standard errors of models $M_9$ and $M_{10}$ are very close. We will check if model $M_{10}$ can account results of other models and if there are mutual encompassing between $M_{10}$ and $M_9$. We now compute the following standardized encompassing statistics using result developed in Theorem 3.1:

### Table 4. Encompassing tests for models $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$

| $\mathcal{M}_1 \not\subseteq \mathcal{M}_3$ | $\mathcal{M}_2 \not\subseteq \mathcal{M}_3$ | $\mathcal{M}_1 \not\subseteq \mathcal{M}_2$ | $\mathcal{M}_2 \not\subseteq \mathcal{M}_1$ |
|-------------------------------------------|-------------------------------------------|-------------------------------------------|-------------------------------------------|
| 19.74 (0.00)                              | 0.021 (0.883)                             | 6.238 (0.013)                             | 0.282 (0.595)                             |

Notes: Statistics of the test with their $p$-values in parenthesis.

### Table 5. Specification of $k$-NN regression estimates

| Explanatory variables and associated models | $W$ | $X$ | $(X, W)$ | $(X, Z)$ | $(X, W, Z)$ |
|-------------------------------------------|-----|-----|----------|----------|-------------|
| $M_7$                                      | Exp | Exp | Exp      | Exp      | Exp         |
| $M_8$                                      | Exp | Exp | Exp      | Exp      | Exp         |
| $M_9$                                      | Exp | Exp | Exp      | Exp      | Exp         |
| $M_{10}$                                   | Exp | Exp | Exp      | Exp      | Exp         |
| $M_{11}$                                   | Exp | Exp | Exp      | Exp      | Exp         |
| Optimal $k$                                | 20  | 11  | 13       | 12       | 16          |
| Minimal RMSE                               | 0.1350 | 0.1301 | 0.1292 | 0.1290 | 0.1352 |
| Maximum value for $k$                      | 60  | 60  | 30       | 30       | 18          |
| Weight                                     | Exp | Exp | Exp      | Exp      | Exp         |

Notes: $W$, $X$, and $(X, W)$ are the explanatory variables for models $M_7$, $M_8$, and $M_9$, respectively. $W$, $X$, and $(X, W, Z)$ are the explanatory variables for models $M_{10}$ and $M_{11}$. The optimal $k$ is the $k$ that minimizes the root mean squared error (RMSE). The minimum RMSE is the smallest RMSE across all models. Maximum value for $k$ is the maximum value of $k$ that satisfies Assumption 3.
Table 6. Encompassing tests for models $M_r$ to $M_{11}$

|          | $M_0 \subseteq M_r$ | $M_2 \subseteq M_3$ | $M_{10} \subseteq M_9$ | $M_2 \subseteq M_{10}$ | $M_{10} \subseteq M_{11}$ |
|----------|----------------------|----------------------|------------------------|------------------------|--------------------------|
| $\delta_3$ | 1.39                 | 0.56                 | 1.76                   | 2.07                   | -0.75                    |

\[
\delta_i = \sqrt{\frac{k - 1}{\text{Var}(\epsilon/Z)} \cdot \text{Var}(\epsilon/Z) \cdot \frac{1}{\varphi(n)}} \cdot \frac{\sqrt{n}}{\sqrt{\frac{1}{\varphi(n)} \sum_{i=1}^{169} Z_i^2}} 
\]

where $\hat{\delta}$ is the k-NN regression of the residuals $\epsilon$ of owner model on explanatory variables $Z$ of rival model. Results are reported in Table 6.

Values in Table 6 are all less than 1.96 in absolute value, except for $M_2 \subseteq M_{10}$. We accept null hypotheses $M_{10} \subseteq M_8$, $M_8 \subseteq M_7$, $M_{10} \subseteq M_9$ and $M_{10} \subseteq M_{11}$. In other word, $M_{10}$ can account information content in other models. As $M_9$ does not encompass $M_{10}$, there is no mutual encompassing. Thus, we can retain model $M_{10}$ from all k-NN regression models.

Next illustration concerns encompassing test on nonparametric and parametric regression techniques in Theorem 3.3, having as null hypothesis: the nearest neighbor regression $M_{10}$ encompasses the linear regression $M_2$. Under this null, we have the following statistic from Theorem 3.3:

\[
\delta_5 = \frac{\sqrt{n}}{\sqrt{\Omega}} \cdot \frac{\sqrt{n}}{\sqrt{\frac{1}{\varphi(n)} \sum_{i=1}^{169} Z_i^2}} 
\]

where $\hat{\Omega}$ is an estimate of the asymptotic variance $\Omega$, $\hat{\epsilon}$, residuals of model $M_{10}$ and $\hat{\sigma}^2$ is a k-NN regression estimate of the conditional variance $\sigma^2 = \text{var}(Y/X = x, Z = z)$.

Absolute value of the standardized encompassing statistic $\delta_5 = -0.01$ is less than 1.96. Therefore, we accept the null hypothesis at a risk level 5% i.e the nearest neighbor regression $M_{10}$ encompasses the linear regression $M_2$. We conclude that we may retain k-NN regression of life expectancy on health expenditure and income.

5. Conclusion

We know that different approaches of encompassing tests present in the literature provide different results. We have considered encompassing test in asymptotic way which is inline with the encompassing principle announced in the introduction. The work has been conducted for parametric and nonparametric regression techniques.

As stated in Hendry et al. (2008) that the work of Bontemps et al. (2008) is the starting treatment of encompassing tests to functional parameter based on nonparametric methods. We have extended that work to nearest neighbor functional parameter estimate under the i.i.d. assumption. When using linear and nearest neighbor regressions as estimators for conditional expectations, we have established asymptotic normality of the associated encompassing statistics for independent processes.

Comparing the convergence rate of the asymptotic encompassing statistic of k-NN regression estimate to kernel regression obtained by Bontemps et al. (2008), it depends only on the number of neighbors $k$ for k-NN while for kernel ones depends on the number of observation $n$ and the bandwidth $h_n$. We have the same convergence rate when $h_n = k/n$. 


Moreover, Bontemps et al. (2008) obtained asymptotic variance of the encompassing statistic associated to kernel regression depending on the density, which is not the case for nearest neighbor regression estimate.

Development of encompassing test to nonparametric methods opens new research direction in theory as well as in practice.

Acknowledgments
The author thanks the anonymous referees and the Editor Professor Hiroshi Shiraishi.

Funding
The author received no direct funding for this research.

Author details
Patrick Rakotomarolahy
E-mail: rakotopapa@gmail.com
ORCID ID: http://orcid.org/0000-0002-7897-5356

1 Mention Mathematics and Their Applications, Faculty of Sciences, University of Fianarantsoa, Andraijana, Fianarantsoa; BP 1164, Madagascar.

Citation information
Cite this article as: Asymptotic behavior of encompassing test for independent processes: Case of linear and nearest neighbor regressions, Patrick Rakotomarolahy, Cogent Mathematics & Statistics (2020), 7: 1805092.

Note
1. Kullback-Leibler Information Criterion

References
Ali, A., & Ahmad, K. (2014). The impact of socio-economic factors on life expectancy in sultanate of Oman: An empirical analysis. Middle-East Journal of Scientific Research, 22(2), 218–224. https://www.idosi.org/mejsr/mejsr22(2)14-2.pdf

Bontemps, C., Florens, J. P., & Richard, J. F. (2008). Parametric and non-parametric encompassing procedures. Oxford Bulletin of Economics and Statistics, 70(1), 751–780. https://doi.org/10.1111/j.1468-0084.2008.00529.x

Cheng, P. E. (1984). Strong consistency of nearest neighbor regression function estimators. Journal of Multivariate Analysis, 15(1), 63–72. https://doi.org/10.1016/0047-259X(84)90067-8

Ferrara, L., Guégan, D., & Rakotomarolahy, P. (2010). GDP nowcasting with ragged-edge data: A semi-parametric modeling. Journal of Forecasting, 29(1–2), 186–199. https://doi.org/10.1002/for.1159

Florens, J. P., Hendry, D. F., & Richard, J. F. (1996). Encompassing and specificity. Econometric Theory, 12(4), 620–656. https://doi.org/10.1017/S0266466600006964

Gourieroux, C., & Monfort, A. (1995). Testing, encompassing, and simulating dynamic econometric models. Econometric Theory, 11(2), 195–228. https://doi.org/10.1017/S0266466600009142

Gourieroux, C., Monfort, A., & Trognon, A. (1983). Testing nested or non-nested hypotheses. Journal of Econometrics, 21(1), 83–115. https://doi.org/10.1016/0304-4076(83)90121-5

Govaerts, A., Hendry, D. F., & Richard, J. F. (1994). Encompassing in stationary linear dynamic models. Journal of Econometrics, 63(1), 245–270. https://doi.org/10.1016/0304-4076(93)01567-6

Guégan, D., & Huck, N. (2009). On the use of nearest neighbors in finance. Revue De Finance, 26(2), 67–86. https://www.cairn.info/revue-finance-2005-2-page-67.htm

Guégan, D., & Rakotomarolahy, P. (2010). A short note on the nowcasting and the forecasting of Euro-area GDP using non-parametric techniques. Economics Bulletin, 30(1), 508–518. http://www.accessecon.com/Pubs/EB/2010/Volume30/EB-10-V30-11-P46.pdf

Hendry, D. F., & Doornik, J. A. (1994). Modelling linear dynamic econometric systems. Scottish Journal of Political Economy, 41(1), 1–33. https://doi.org/10.1111/j.1467-9485.1994.tb01107.x

Hendry, D. F., Marcellino, M., & Mizon, G. E. (2008). Encompassing. Oxford Bulletin of Economics and Statistics, Guest Editor Introduction. Special Issue

Hussain, A. R. (2002). Life expectancy in developing countries: A cross-section analysis. The Bangladesh Development Studies, 28(12), 161–178.

Mack, Y. P. (1981). Local Properties of k-NN Regression Estimates. SIAM Journal on Algebraic and Discrete Methods, 2(3), 311–323. https://doi.org/10.1137/0602035

Mizon, G. E. (1984). The encompassing approach in econometrics. In D. F. Hendry & K. F. Wallis (Eds.), Econometrics and quantitative economics (pp. 135–172). Blackwell Publishers.

Mizon, G. E., & Richard, J. F. (1986). The encompassing principle and its application to non-nested hypothesis tests. Econometrica, 54(3), 657–678. https://doi.org/10.2307/1911313

Mukherjee, H. (1991). Nearest neighbor regression with heavy-tailed errors. Annals of Statistics, 21(2), 681–693. https://doi.org/10.1214/aos/1176349144

Nowman, B., & Soltogu, B. (2003). Continuous time and nonparametric modelling of U.S. interest rate models. International Review of Financial Analysis, 12(1), 25–34. https://doi.org/10.1016/S1057-5219(02)00123-0

Puspatitasari, D. A., & Rustam, Z. (2018). Application of SVM-KNN using SVR as feature selection on stock analysis for Indonesia stock exchange. AIP Conference Proceedings 2023, Bali, Indonesia, 020207. https://doi.org/10.1063/1.5064204

Sowa, T. (1979). Information criteria for discriminating among alternative regression models. Econometrica, 46(6), 1273–1292. https://doi.org/10.2307/1913828
