AN EMPIRICAL DEPENDENCE MESAURES BASED ON RESIDUAL VARIANCE ESTIMATION

Nima Reyhani and *Amaury Lendasse*

Lab of Computer and Information Science, Helsinki University of Technology
Espoo, PL 5400, Helsinki, Finland
{nreyhani, lendasse}@cis.hut.fi

ABSTRACT

In this paper, a solution to empirical dependency measure is proposed. The main idea is to use the notion of predictability as a basis for dependency definition. Considering any nonlinear regression function between two random variables, the power of regression residuals or noise variance defines the desired dependency measure. The residuals variance can be directly computed by estimators without finding the best fit curve. The paper shows the conditions on which two random variables are independent according to the estimated residuals variance. The existence of residual variance, or noise variance estimators make it possible to define such practical measure for dependency. The dependency measure finds wide areas of applications in signal processing and machine learning. In this paper, solutions for Independent Component Analysis and input selection using the proposed dependency measure are discussed.

1. INTRODUCTION

A number of algorithms in signal processing, adaptive filtering and machine learning can be recasted as optimizing the dependence measure between states or redefined input-output pairs. For example, in feature extraction, the goal is to select features which are independent to each other with respect to a given data set. As another example, in multiple step ahead prediction, one might be interested in the dependence between few steps-ahead and previous values. These evidences shows the essence of an empirical measure of dependence.

The dependence measures are studied thoroughly in the paper of Renyi [12], where the dependence measures are listed by *correlation*, *correlation ratio*, *maximal correlation* and *mean square contingency*. Among them, maximal correlation is studied in details in [12]. Further works, including the works of [4] and [1], develop the maximal correlation measure to become tractable by using the reproducing kernel Hilbert space. Mutual Information and Kullback-Leiber divergence are widely used as dependence measure as well [2, 9].

In contrast, in this paper, we propose an approach to dependence estimation grounded on the notion of predictability. Here, the basic idea is to write one of the random variables (r.v.) in terms of nonlinear regression of the other one which results in:

$$y = f(\mathbf{x}) + \varepsilon, \tag{1}$$

f denotes a nonlinear regression function, $\mathbf{x}$ and y are the interesting r.v.s. The amount of dependency between the r.v.s derives variance of ε, i.e. the residuals, to a quantity taking place between two extremes: independence or strict dependence. Two r.v. $\mathbf{x}$ and y are called strictly dependent if $y = f(\mathbf{x})$ for some bounded function f. Therefore, the dependency estimation can be reduced to residual/noise variance estimation.

In the following, first, the possibility of noise variance or residual variance estimation based on the notion of smoothness of regression function is described in section 2. Section 3 draws the contribution of the noise variance in the problem of dependency estimation. To show the efficiency of the proposed approach to dependency estimation, section 4 shows applications in input variable selection and independent component analysis.

In this paper, the goal is not to provide a new technique which might be more efficient in terms of accuracy, computational complexity or robustness compare to other methods like Mutual Information or Kernel Mutual Information [1]. Rather, the goal is to show the existence of another possibility for dependence estimation which is not covered in Renyi works and stems from the notion of nonlinear regression analysis.

2. NOISE VARIANCE ESTIMATION

Consider given data set $\mathcal{L} = \{\mathbf{x}_i, y_i\}_{i=1}^N$ where $\mathbf{x}_i \in \mathbb{R}^M$ for some fixed M and $y_i \in \mathbb{R}$. We assume that y can be estimated by some function f ($f \in L_2$). Thus, we have

$$y = f(\mathbf{x}) + \varepsilon. \tag{2}$$

Furthermore, we assume that ε is independent to $\mathbf{x}$ and $\mathbb{E}\{\varepsilon\} = 0$. ε arises from those parts of y which can not be determined from $\mathbf{x}$. The problem is to find $\mathbb{V}\mathrm{ar}\{\varepsilon\}$. Simply, one can fit a model to the given data set and take the empirical variance of the residuals. Here, instead of fitting a model we employ those estimators which can directly compute the residual variance based on *a priori*. All the proposed estimate can be summarized in terms of smoothness concept, see for example [5, 10, 8, 7, 3].

Here, we describe the general idea behind these estimators briefly. Suppose f is continuous or smooth, and because of technical reason, $\varepsilon \in L^2$. For continuous functions we have, for all $\varepsilon > 0$ there is a $\delta > 0$ such that $\|f(x) - f(x_0)\| < \delta$ for $\|x - x_0\| < \varepsilon$. In practical situation, the nearest neighbor defines the ε and δ. Therefore, if the maximum distance in mesh of data points goes to zero, we can say that $\|f(x_i) - f(x_{[i,1]})\| < \delta$ or the norm is almost sure zero (again, for a given data set where we do not have a compact space). $[i,1]$ denotes the nearest neighbor of x_i. Very simply, for ordered design, the residuals are $\varepsilon_i = (y_i - y_{i+1})$ and taking expectation over $\frac{1}{2}\varepsilon^2$ results in

$$\mathbb{E}\{\hat{\sigma}^2\} = \frac{1}{2}\frac{1}{N-1}\sum_{i=2}^{N}\mathbb{E}\left\{(y_i - y_{i-1})^2\right\} =$$

$$\sigma^2 + \frac{1}{2}\frac{1}{N-1}\sum_{i=2}^{N}\mathbb{E}\left\{(f(x_i) - f(x_{i-1}))^2\right\}$$

$$= \cdots = \sigma^2 + \frac{1}{N^2}J + o\left(\frac{1}{N^2}\right). \quad (3)$$

where $J = \frac{1}{2}\int_0^1 f'(x)^2\,dx$. In summary, one solution to the empirical residual or noise variance estimation can be written as

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N-1}\frac{1}{2}(y_{i+1} - y_i)^2. \quad (4)$$

Thus, by assuming that a smooth least plausible approximation of a point is its nearest neighbour, we construct an estimate of the noise variable and accordingly, we can estimate the noise variance.

It is possible to extend the above method by taking each three consecutive points. The local residual can be constructed by distance between the response at the middle point and corresponding value at that point of the regression line connecting two outer point. The noise variance can be estimated by computing the empirical variance over the evaluated local residuals. In other words, the local residuals can be casted as

$$\varepsilon_i = \underbrace{\left(\frac{x_{i+1} - x_i}{x_{i+1} - x_{i-1}}\right)}_{a_i} y_{i-1} + \underbrace{\left(\frac{x_{i+1} - x_i}{x_{i+1} - x_{i-1}}\right)}_{b_i} y_{i+1} - y_i.$$

A similar approach shown in expression (3) gives

$$\mathbb{E}\left\{\varepsilon_i^2\right\} = \left(a_i^2 + b_i^2 + 1\right)\sigma^2 + O\left(\frac{1}{N^2}\right). \quad (5)$$

In [3, 7], an extension to higher degrees is described with the general name *polynomial estimator*. In order to extend the above estimator to $\mathbb{R}^M$, the differences should be taken in all M directions, for more details, see [3].

More sophisticated approaches have been proposed for noise variance estimation that use kernels to provide consistent estimate, for example the method proposed by Hall [8], Tong approach [13], Delta test by Pi and Peterson [11], and etc. Among them, Muller [14] proposed a $\sqrt{N}$ consistent estimate based on U-statistic which is defined by

$$\mathcal{U} = \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N}\underbrace{\frac{1}{2}(y_i - y_j)^2}_{\text{Similar to 4}}\underbrace{\frac{1}{2}\left(\frac{1}{\widehat{\pi}_i} + \frac{1}{\widehat{\pi}_j}\right)\mathcal{K}_h(\mathbf{x}_i - \mathbf{x}_j)}_{\text{Weights}} \quad (6)$$

where $\hat{\pi}$ is defined by

$$\widehat{\pi}_i = \frac{1}{N-1}\sum_{\substack{j=1\\ i\neq j}}^{N}\mathcal{K}_h(\mathbf{x}_i - \mathbf{x}_j), \quad i = 1, 2, \ldots, N.$$

$\mathcal{K}$ is a kernel function and h is the kernel width. The weights in expression (6) are chosen in such a way that the estimator becomes $\sqrt{N}$ consistent [1], for more details, see [14]. The main difficulty in employing the Muller approach is choosing the kernel width. To check the sensitivity of Muller approach to the kernel width, a toy data set is made on $y = sinc(5\pi x) + \varepsilon$, where $x \in [0,1]$ and ε denotes the noise with different distribution (Normal, t-student and χ^2). The true variance of ε and the proper value of kernel width which results in an estimate with error less than ≈ 0.00001 are given in table 1.

$\sigma^2(A)$	Width	$\sigma^2(B)$	Width	$\sigma^2(C)$	Width
8.3747e-04	0.0033	0.0106	0.0033	0.1119	0.0051
0.0034	0.0033	0.0423	0.0033	0.4476	0.0051
0.0077	0.0033	0.0952	0.0033	1.0070	0.0051
0.0130	0.0065	0.1693	0.0033	1.7903	0.0047
0.0203	0.0065	0.2645	0.0033	2.7973	0.0047
0.0308	0.0065	0.3809	0.0033	4.0281	0.0047
0.0410	0.0065	0.5184	0.0033	5.4827	0.0047
0.0543	0.0065	0.6771	0.0033	7.1611	0.0047
0.0678	0.0065	0.8569	0.0033	9.0633	0.0047
0.0855	0.0065	1.0580	0.0033	11.1892	0.0047

Table 1. This table shows the sensitivity of Muller approach to the kernel width in a Toy Example: $y = sinc(5\pi x) + \varepsilon$. (A) column denote the result with gaussian noise, (B) for t-student noise and (C) for χ^2-distribution noise

The same experiments with other methods for noise variance estimation shows that Muller approach is reliable and less sensitive to the kernel width. So, for experimental parts, we applied the Muller approach.

3. DEPENDENCE MEASURE BASED ON PREDICTABILITY

Let's *reconsider* or *redefine* the problem of dependence estimation as the question of "how likely the random variable y can be nonlinearly predicted or approximated from the random variable x". Intuitively, when the residual variance is high compare to the variance of the variable taken as the response, then they are likely independent.

Referring to brief discussion in section 1, we assume that it is possible to estimate r.v. y for a given value of r.v. x, so we have

$$y_i = f(\mathbf{x}_i) + \varepsilon(i).$$

[1] $\lim_{N\to\infty}\mathcal{U} = \frac{1}{N}\sum_{i=1}^{N}\varepsilon_i^2 + \mathcal{O}\left(N^{-\frac{1}{2}}\right)$

Here, $\varepsilon(.)$ is an unknown function, which in principle, represents that part of signal which can not be determined and it is referred by noise. By this notation, we mean that r.v. y is span of basis $\{\varepsilon(.), f(.)\}$, where $f \in L^2$. Thus, we have:

$$\mathbb{E}\left\{\|y-\varepsilon\|^2\right\} = \mathbb{E}\left\{\|y\|^2\right\} + \mathbb{E}\left\{\|\varepsilon\|^2\right\} - \mathbb{E}\left\{2y^T\varepsilon\right\} \quad (7)$$

For independent r.v.s $\mathbf{x}$ and y, $\mathbb{E}\{\|y-\varepsilon\|^2\}$ approaches to zero, which implies that $\mathbb{E}\left\{\|f(.)\|^2\right\} \to 0$. Also, $\mathbb{E}\left\{\|\varepsilon\|^2\right\} \to 0$, implying that $\varepsilon \to 0$ and the r.v. y is strictly dependent to r.v. $\mathbf{x}$. For values of $\mathbb{E}\left\{\|\varepsilon\|^2\right\}$ which are closer to zero, the conclusion is that the r.v.s $\mathbf{x}$ and y are more dependent, i.e. it is most likely that one can find a smooth nonlinear function which can approximate y based on $\mathbf{x}$ almost sure.

Theorem 3.1 *Let $\mathbb{V}ar\{\varepsilon\}$ denotes the residual power or the noise variance in regression analysis between r.v. $\mathbf{x}$ as design and y as response variable. Two random variables $\mathbf{x}$ and y are independent if and only if $\mathbb{V}ar\{\varepsilon\} = \mathbb{V}ar(y)$.*

Proof: Suppose $\mathbb{V}\text{ar}\{\varepsilon\} = \mathbb{V}\text{ar}(y)$ holds. since ε and x are independent, we have $\mathbb{V}\text{ar}\left\{f(\mathbf{x})\right\} = 0$. $\mathbb{V}\text{ar}\{y\} = \mathbb{V}\text{ar}\{\varepsilon\} + \mathbb{V}\text{ar}\{f(x)\} + 2\mathbb{C}\text{ov}\{f(x), \varepsilon\}$. Therefore, $f(x) = C$

Now, suppose that the random variables $\mathbf{x}$ and y are independent. So, we have

$$\begin{aligned}\mathbb{E}\left\{g(\mathbf{x})\right\}\mathbb{E}\left\{y\right\} &= \mathbb{E}\left\{g(\mathbf{x})\right\}\mathbb{E}\left\{f(\mathbf{x})+\varepsilon\right\} = \\ &\mathbb{E}\left\{g(\mathbf{x})\right\}\mathbb{E}\left\{f(\mathbf{x})\right\} + \mathbb{E}\left\{g(\mathbf{x})\right\}\mathbb{E}\left\{\varepsilon\right\} \\ &= \mathbb{E}\left\{g(\mathbf{x})\right\}\mathbb{E}\left\{f(\mathbf{x})\right\} = \mathbb{E}\left\{g(\mathbf{x})f(\mathbf{x})\right\}.\end{aligned}$$

This hold if and only if $f(\mathbf{x}) = C$. □

The discussion above is true for the case where $\mathbf{x} \in \mathbb{R}^M$, for fixed M and $y \in \mathbb{R}$. To extend the proposed dependence measure, we put a strong condition when $\mathbf{y} \in \mathbb{R}^C$ for some fixed C:

Theorem 3.2 *Two random variables $\mathbf{x} \in \mathbb{R}^M$ and $\mathbf{y} \in \mathbb{R}^C$ are independent if and only if*

$$\forall l = 1, \ldots, C: \quad \mathbb{V}ar\left\{\varepsilon_l\right\} = \mathbb{V}ar\left\{y_l\right\},$$

where $\mathbb{V}ar\left\{\varepsilon_l\right\}$ is the residual variance obtained in regression analysis between $\mathbf{x}$ and y_l. □

4. APPLICATIONS

The idea of employing the dependence estimation can be implemented in *Infomax*-like framework [9], where the dependence estimation plays the role of objective function. By minimizing or maximizing the dependency measure between two random variables, a variety of applications appears. Two realization of such a framework are described as following.

4.1. Input Variable Selection or Feature Selection

Feature selection or input variable selection involves finding input variables or features having most dependency with interesting variable, e.g. response variable. The problem of the input variable selection can be addressed through weighting the input variables according to their contribution to response variable y. The weights can be chosen from the set $\{0, 1\}$, for input variable selection, or from the interval $[0, 1]$, for input variable weighting. Then problem can be formulated as

$$\min_W \hat{\sigma}^2\left\{W \odot \mathbf{x}, y\right\} = \min_{w_1,\ldots,w_M} \hat{\sigma}^2\left\{\underbrace{\{w_1x_1, w_2x_2, \ldots, w_Mx_M\}}_{\text{weighted input variables}}, \underbrace{y}_{\text{output variable}}\right\}. \quad (8)$$

Within such framework, input variables with more contribution to the response variable will be assigned to higher weights. These weights also cancel out the noise in the data set. Since the input variables with small weights indicates their poor dependency with the response variables, they can be filtered out by thresholding.

4.2. Application of Dependence Estimation in ICA

The independent component analysis problem consists of finding the source signals or components and the mixing coefficients given the mixed signals, provided that the original components are independent to each other. Let's suppose the sources $\mathbf{s} = \{s_1(t), s_2(t), \ldots, s_M(t)\}$ mixed by some matrix A. The observation is represented by $\mathbf{x} = \{x_1(t), x_2(t), \ldots, x_M(t)\} = A\mathbf{s}$. The goal is to find the matrix B which is equal to A^{-1} together with the source vectors $\{s_i\}_{i=1}^M$ [2].

Given that the source signals are independent to each other, one solution can be obtained by weighting the observations in such a way that the estimated components become independent to each other. The weight matrix is called demixing matrix. The traditional approaches to ICA usually apply Mutual Information to measure the independency between the estimated components [2].

By applying the noise variance as a measure of independency, we can propose a solution to the ICA as following

$$\min_{B_i} \left\|\hat{\sigma}^2\left\{\{\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_{i-1}, \hat{s}_{i+1}, \ldots, \hat{s}_M\}, \hat{s}_i\right\} - \mathbb{V}\text{ar}\left\{\hat{s}_i\right\}\right\|_2^2 \quad \forall i = 1, 2, \ldots, M \quad (9)$$

Here, $\hat{s}_i$ denotes the estimated source which is obtained by $\hat{s}_i = B_i\mathbf{x}$, where B_i is the i-th row vector of demixing matrix B. Among all the methods proposed for noise variance estimation, there are some which are not differentiable with respect to their arguments, for example the proposed methods in [11, 6]. In this paper, the proposed estimator by Muller (in expression 6) is employed. To make a simpler optimization, the given data set is whiten, in a sense that the $\mathbb{E}\left\{\mathbf{z}\mathbf{z}^T\right\} = \mathbf{I}$, where $\mathbf{I}$ is the unit matrix and $\mathbf{z}$ is the whited matrix of $\mathbf{x}$[2]. The experimental results

[2]Let C denotes the covariance matrix $\mathbb{E}\left\{\mathbf{x}\mathbf{x}^T\right\}$. Then the whiten version of $\mathbf{x}$, which is denoted by $\mathbf{z}$ can be obtained by $\mathbf{z} = D^{-\frac{1}{2}}E$,

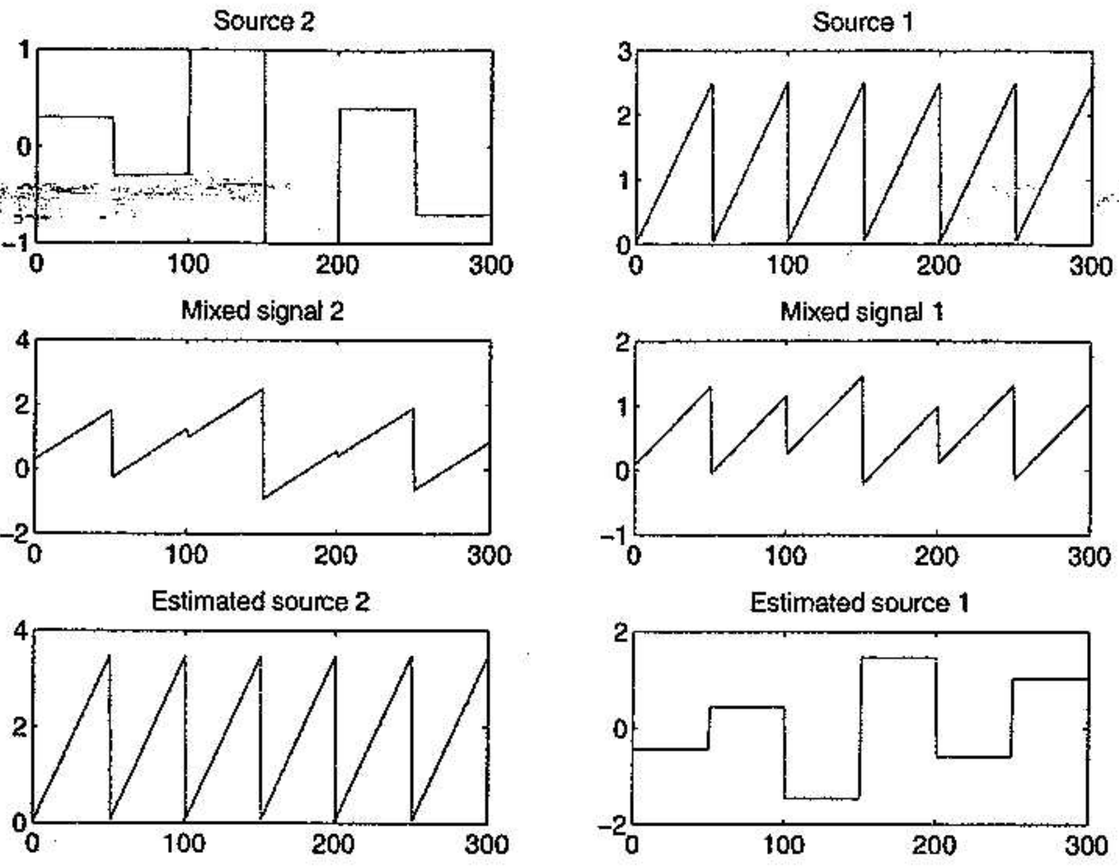

Fig. 1. The results for ICA obtained by optimization in (9) on the Toy example. First row shows the original components, second row, shows the whited mixtures, and the third row shows the estimated independent components.

are shown in figure 1. The figures in the first row show the independent sources applied in this experiment. For finding the optimum value, Matlab optimization toolbox is used. The gradient vector is not provided in the experiments, but still the optimization algorithm is able to find optimal solution.

5. DISCUSSION

In this paper, the idea of using the noise variance as the dependency measure is proposed. There, two random variables are independent iff the residual variance of regression analysis between two random variables (one as design and the other as response) reaches the variance of response variable. The noise variance can be computed by taking the difference between the best fit curve and the response variable. One can, also, directly estimate the additional noise variance based on the smoothness assumption. Since each noise variance estimator requires *priors* on the smoothness of the underlying function, the estimation is biased on this information. Recall that the prior on the smoothness is implemented as the kernel width or the order of smoothing filter. Accordingly, the drawn way to detect independency is somehow biased on the smoothness of the possible nonlinear function between the random variables. The study in section (2) on the Muller estimator reveals that the bias of estimation is not so sensitive to the smoothness prior. In a way, it shows that the resolution of the estimation is much more than the resolution of the kernel width. In addition, the convergence or other statistical properties of the residual variance estimator transforms to the proposed dependency measure.

To check the reliability of using the noise variance as dependency measure, it is applied in ICA problem. The ICA problem is used as benchmark problem for some of other dependency measures, for example see [4]. Experimental results support the reliability and accuracy of the proposed method.

In side of applications, a number of problems can be addressed by minimizing or maximizing the dependence measure between, in general, two group of variables. Therefore, finding an efficient optimization techniques for the proposed dependence measure is of great importance. Extension of this work for more general domains like trees, string or graphs draws line for future works.

6. REFERENCES

[1] A. GRETTON, R. HERBRICHM A. SMOLA, O. B., AND SCHOLKOPF, B. Kernel methods for measuring independence. *Journal of Machine Learning Research 6* (2006), 2075–2129.

[2] A. HYVARINEN, J. K., AND OJA, E. *Independent Component Analysis*. John Wiley and Sons, 2001.

[3] A. MUNK, N. BISSANTZ, T. W., AND FREITAG, G. On difference-based variance estimation in nonparamteric regression when the covariate is high dimensional. *Journal of the Royal Statistical Society B(Methodological) 67* (2005), 901–919.

[4] BACH, F. R., AND JORDAN, M. I. Kernel independent component analysis. *Journal of Machine Learning Research 3* (2002), 1–48.

[5] CARTER, C. K., AND EAGLESON, G. K. A comparision of variance estimators in nonparametric regerssion. *Journal of Royal Statistical Society B (Methodological) 54*.

[6] EVANS, D., AND JONES, A. J. A proof of the gamma test. *Journal of the Royal Statistical Society A 458* (2002), 2759–2799.

[7] H. DETTE, A. MUNK, T. W. Estimating the variance in nonparamteric regression-what is a reasonable choice. *Journal of the Royal Statistical Society B (Methodological) 60* (1998), 751–764.

[8] HALL, P., AND MARRON, J. S. On variance estimation in nonparamteric regrssion. *Biometrika 77* (1990), 415–419.

[9] HAYKIN, S. *Neural Networks: A Comprehensive Foundation*. Prentice Hall, NJ, 2 edition, 1999.

[10] M. J. BUCKLEY, G. K. E., AND SILVERMAN, B. W. The estimation of residual variance in nonparamteric regression. *Biometrika 75*.

[11] PI, H., AND PETERSON, C. Finding the embedding dimnesion and variable dependencies in time series. *Neural Computation 6* (1994), 509–520.

[12] RENYI, A. On measures of dependence. *Acta Math. Acad. Sci. Hungar 10* (1959), 441–451.

[13] TONG, T., AND WANG, Y. Estimating the residual variance in nonparametric regression using least squares. *Biometrika 92* (2005), 821–830.

[14] U. MULLER, A. S., AND WEFELMEYER, W. Estimating the error variance in nonparametric regression by a covariate-matched u-statistic. *Statistics 37* (2003), 179–188.

where E is the eigen matrix of C and $D = \text{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_M\}$.