HAL
open science

Linear regression with stationary errors : the R package slm

Emmanuel Caron, Jérôme Dedecker, Bertrand Michel

To cite this version:

Emmanuel Caron, Jérôme Dedecker, Bertrand Michel. Linear regression with stationary errors : the R package slm. 2019. hal-02157155v2

HAL Id: hal-02157155

https://hal.science/hal-02157155v2

Preprint submitted on 19 Jun 2019 (v2), last revised 22 Oct 2019 (v3)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Linear regression with stationary errors : the R package slm

Emmanuel Caron
LMJL Ecole Centrale Nantes

Jérôme Dedecker
Université Paris Descartes

Bertrand Michel
LMJL Ecole Centrale Nantes

Abstract

This paper introduces the R package **slm** which stands for Stationary Linear Models. The package contains a set of statistical procedures for linear regression in the general context where the error process is strictly stationary with short memory. We work in the setting of Hannan (1973), who proved the asymptotic normality of the (normalized) least squares estimators (LSE) under very mild conditions on the error process. We propose different ways to estimate the asymptotic covariance matrix of the LSE, and then to correct the type I error rates of the usual tests on the parameters (as well as confidence intervals). The procedures are evaluated through different sets of simulations, and two examples of real datasets are studied.

Keywords: Linear model, Least Squares Estimators, Stationary processes, AutoRegressive processes, Spectral density, Hypothesis testing.

1. Introduction

We consider the usual linear regression model

$$Y = X\beta + \varepsilon\,,$$

where Y is the n-dimensional vector of observations, X is a (possibly random) $n \times p$ design matrix, β is a p-dimensional vector of parameters, and $\varepsilon = (\varepsilon_i)_{1\leq i\leq n}$ is the error process (with zero mean and independent of X). The standard assumptions are that the ε_i's are independent and identically distributed (i.i.d.) with zero mean and finite variance.

In this paper, we propose to modify the standard statistical procedures (tests, confidence intervals, ...) of the linear model in the more general context where the ε_i's are obtained from a strictly stationary process $(\varepsilon_i)_{i\in\mathbb{N}}$ with short memory. To be more precise, let $\hat{\beta}$ denote the usual least squares estimator of β. Our approach is based on two papers: the paper by Hannan (1973) who proved the asymptotic normality of the least squares estimator $D(n)(\hat{\beta}-\beta)$ ($D(n)$ being the usual normalization) under very mild conditions on the design and on the error process; and a recent paper by Caron (2019) who showed that, under Hannan's conditions, the asymptotic covariance matrix of $D(n)(\hat{\beta}-\beta)$ can be consistently estimated.

Let us emphasize that Hannan's conditions on the error process are very mild and are satisfied for most of short-memory processes (see the discussion in Section 4.4 of Caron and Dede (2018)). Putting together the two above results, we can develop a general methodology for tests and confidence regions on the parameter β, which should be valid for most of short-

memory processes. This is of course directly useful for time-series regression (we shall present in Section 5 an application to the "Mona Loa" R data-set on CO2 concentration), but also in the more general context where the residuals of the linear model seem to be strongly correlated. More precisely, when checking the residuals of the linear model, if the autocorrelation function of the residuals shows significant correlations, and if the residuals can be suitably modeled by an ARMA process, then our methodology is likely to apply. We shall give an example of such a situation in Section 5 (Shangai pollution data-set).

Hence, the tools presented in the present paper can be seen from two different points of view:

- as appropriate tools for time series regression with short memory error process.
- as a way to robustify the usual statistical procedures when the residuals are correlated.

Let us now describe the organisation of the paper. In Section 2, we recall the mathematical background, the consistent estimator of the asymptotic covariance matrix introduced in Caron (2019) and the modified Z-statistics and χ-square statistics for testing hypothesis on the parameter β. In Section 3 we present the **slm** package, and the different ways to estimate the asymptotic covariance matrix: by fitting an autoregressive process on the residuals (default procedure), by means of the kernel estimator described in Caron (2019) (theoretically valid) with a bootstrap method to choose the bandwidth (Wu and Pourahmadi (2009)), by using an alternative choice of the bandwidth for the rectangular kernel (Efromovich (1998)), by means of an adaptative estimator of the spectral density via Histograms (Comte (2001)). In Section 4, we estimate the level of a χ-square test for a linear model with random design, with different kind of error processes and for different estimation procedures. In Section 5, we present two different data sets "CO2 concentration", "Shangai pollution", and we compare the summary output of `slm` with the usual summary output of `lm`.

2. Linear regression with stationary errors

2.1. Asymptotic results for the kernel estimator

We start this section by giving a short presentation of linear regression with stationary errors, more details can be found for instance in Caron (2019). Let $\hat{\beta}$ be the usual least squares estimator for the unknown vector β. The aim is to provide hypothesis tests and confidence regions for β in the non i.i.d. context.

Let γ be the autocovariance function of the error process ε: for any integers k and m, let $\gamma(k) = \mathrm{Cov}(\varepsilon_m, \varepsilon_{m+k})$. We also introduce the covariance matrix

$$\Gamma_n := [\gamma(j-l)]_{1\leq j,l\leq n}\,.$$

Hannan (1973) has shown a Central Limit Theorem for $\hat{\beta}$ when the error process is strictly stationary, under very mild conditions on the design and the error process. Let us notice that the design can be random or deterministic. We introduce the normalization matrix $D(n)$ which is a diagonal matrix with diagonal term $d_j(n) = \|X_{.,j}\|_2$ for j in $\{1,\ldots,p\}$, where $X_{.,j}$ is the jth column of X. Roughly speaking, Hannan's result says in particular that, given the design X, the vector $D(n)(\hat{\beta}-\beta)$ converges in distribution to a centered Gaussian distribution

with covariance matrix C. As usual, in practice the covariance matrix C is unknown and it has to be estimated. Hannan also showed the convergence of second order moment:[1]

$$\mathbb{E}\left(D(n)(\hat{\beta}-\beta)(\hat{\beta}-\beta)^t D(n)^t \middle| X\right) \xrightarrow[n\to\infty]{} C, \quad a.s.$$

where

$$\mathbb{E}\left(D(n)(\hat{\beta}-\beta)(\hat{\beta}-\beta)^t D(n)^t \middle| X\right) = D(n)(X^tX)^{-1}X^t\Gamma_n X(X^tX)^{-1}D(n).$$

In this paper we propose a general plug-in approach: for some given estimator $\widehat{\Gamma}_n$ of Γ_n, we introduce the plug-in estimator

$$\widehat{C} = \widehat{C}(\widehat{\Gamma}_n) := D(n)(X^tX)^{-1}X^t\widehat{\Gamma}_n X(X^tX)^{-1}D(n),$$

and we use $\widehat{C}$ to standardize the usual statistics considered for the study of linear regression.

Let us illustrate this plug-in approach with a kernel estimator which has been proposed in Caron (2019). For some K and a bandwidth h, the kernel estimator $\widetilde{\Gamma}_{n,h}$ is defined by

$$\widetilde{\Gamma}_{n,h} = \left[K\left(\frac{j-l}{h}\right)\tilde{\gamma}_{j-l}\right]_{1\le j,l\le n}, \tag{1}$$

where the residual based empirical covariance coefficients are defined for $0 \le |k| \le n-1$ by

$$\tilde{\gamma}_k = \frac{1}{n}\sum_{j=1}^{n-|k|}\hat{\varepsilon}_j\hat{\varepsilon}_{j+|k|}. \tag{2}$$

For a well-chosen kernel K and under mild assumptions on the design and the error process, it has been proved in Caron (2019) that

$$\tilde{C}_n^{-1/2}D(n)(\hat{\beta}-\beta) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}_p(0_p, I_p), \tag{3}$$

for the plug-in estimator $\tilde{C}_n := \widehat{C}(\widetilde{\Gamma}_{n,h_n})$, for some suitable sequence of bandwidths (h_n).

More generally, in this paper we say that an estimator $\widehat{\Gamma}_n$ of Γ_n is *consistent for estimating the covariance matrix* C if $\widehat{C}(\widehat{\Gamma}_n)$ is positive definite and if it converges in probability to C. Note that such a property requires assumptions on the design, see Caron (2019). If $\widehat{C}(\widehat{\Gamma}_n)$ is consistent for estimating the covariance matrix C, then $\widehat{C}(\widehat{\Gamma}_n)^{-1/2}D(n)(\hat{\beta}-\beta)$ converges in distribution to a standard Gaussian vector.

2.2. Tests and confidence regions

We now present tests and confidence regions for arbitrary estimators $\widehat{\Gamma}_n$. The complete justifications are available for kernel estimators, see Caron (2019).

[1] The transpose of a matrix X is denoted by X^t.

Z-Statistics. We introduce the following univariate statistics:

$$Z_j = \frac{d_j(n)\hat{\beta}_j}{\sqrt{\widehat{C}_{(j,j)}}} \tag{4}$$

where $\widehat{C} = \widehat{C}(\widehat{\Gamma}_n)$. If $\widehat{\Gamma}_n$ is consistent for estimating the covariance matrix C and if $\beta_j = 0$, the distribution of Z_j converges to a standard normal distribution when n tends to infinity. We directly derive an asymptotic hypothesis test for testing $\beta_j = 0$ against $\beta_j \neq 0$ as well as an asymptotic confidence interval for β_j.

Chi-square statistics. Let A be a $n \times k$ matrix with $\text{rank}(A) = k$. Under Hannan (1973)'s conditions, $D(n)(A\hat{\beta} - A\beta)$ converges in distribution to a centered Gaussian distribution with covariance matrix ACA^t. If $\widehat{\Gamma}_n$ is consistent for estimating the covariance matrix C, then $A\widehat{C}(\widehat{\Gamma}_n)$ converges in probability to AC. The matrix $A\widehat{C}(\widehat{\Gamma}_n)A^t$ being symmetric positive definite, this yields

$$W := (A\widehat{C}(\widehat{\Gamma}_n))^{-1/2} D(n) A(\hat{\beta} - \beta) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}_k(0_k, I_k).$$

This last result provides asymptotical confidence regions for the vector $A\beta$. It also provides an asymptotic test for testing the hypothesis $H_0 : A\beta = 0$ against $H_1 : A\beta \neq 0$. Indeed, under the H_0-hypothesis, the distribution of $\|W\|_2^2$ converges to a $\chi^2(k)$-distribution.

The test can be used to simplify a linear model by testing that several linear combinations between the parameters β_j are zero, as we usually do for Anova and regression models. In particular, with $A = I_p$, the test corresponds to the test of overall significance.

3. Introduction to linear regression with the slm package

Using the **slm** package is very intuitive because the arguments and the outputs of `slm` are similar to those of the standard functions `lm`, `glm`, etc. The output of the main function `slm` is an object of class `slm` , a specific class that has been defined for linear regression with stationary processes. The `slm` class has methods `plot`, `summary`, `confint` and `predict`. Moreover, the class `slm` inherits from the `lm` class and thus provides the output of the classical `lm` function.

```
R> library(slm)
```

The statistical tools available in `slm` strongly depend on the choice of the covariance plug-in estimator $\widehat{C}(\widehat{\Gamma}_n)$ we use for estimating C. All the estimators $\widehat{\Gamma}_n$ proposed in `slm` are residual-based estimators, but they rely on different approaches. In this section, we present the main functionality of `slm` together with the different covariance plug-in estimators.

For illustrating the package, we simulate synthetic data according to the linear model:

$$Y_i = \beta_1 + \beta_2(\log(i) + \sin(i) + Z_i) + \beta_3 i + \varepsilon_i,$$

where Z is a gaussian autoregressive process of order 1, and ε is the Nonmixing process described in Section 4.1. We use the functions `generative_model` and `generative_process`

respectively to simulate observations according to this regression design and with this specific stationary process. More details on the designs and the processes available with `generative_model` and `generative_process` are given in Section 4.1.

```
R> n = 500
R> eps = generative_process(n,"Nonmixing")
R> design = generative_model(n,"mod2")
R> design_sim = cbind(rep(1,n), as.matrix(design))
R> beta_vec = c(2,0.001,0.5)
R> Y = design_sim %*% beta_vec + eps
```

3.1. Linear regression via AR fitting on the residuals

A large class of stationary processes with continuous spectral density can be well approximated by AR processes, see for instance Corollary 4.4.2 in Brockwell and Davis (1991). The covariance structure of an AR process having a closed form, it is thus easy to derive an approximation $\widetilde{\Gamma}_{AR(p)}$ of Γ_n by fitting an AR process on the residual process.

The AR-based method for estimating C is the default version of `slm`. This method proceeds in four main steps:

1. Fit an autoregressive process on the residual process $\hat{\varepsilon}$;
2. Compute the theoretical covariances of the fitted AR process ;
3. Plug the covariances in the Toeplitz matrix $\widetilde{\Gamma}_{AR(p)}$;
4. Compute $\widehat{C} = \widehat{C}(\widetilde{\Gamma}_{AR(p)})$.

The `slm` function fits a linear regression of the vector Y on the design X and then fits an AR process on the residual process using the `ar` function from the **stats** package. The output of the `slm` function is an object of class `slm`. The order p of the AR process is set in the argument `model_selec`:

```
R> regslm = slm(Y ~ X1+X2, data = design, method_cov_st = "fitAR",
+          model_selec = 3)
```

The estimated covariance is recorded as a vector in the attribute `cov_st` of `regslm`, which is an object of class `slm`. The estimated covariance matrix can be computed by taking the Toeplitz matrix of `cov_st`, using the `toeplitz` function.

Summary method

As for `lm` objects, a summary of a `slm` object is given by

```
R> summary(regslm)

Call:
"slm(formula = myformula, data = data, x = x, y = y)"
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-13.9086  -3.4586   0.1646   3.5025  13.7488

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 2.936183   0.855214   3.433 0.000596 ***
X1          0.084387   0.082371   1.024 0.305613
X2          0.492590   0.002738 179.938  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.907
Multiple R-squared:  0.9953
chi2-statistic: 3.371e+04 on 2 DF,  p-value: < 2.2e-16
```

The coefficient table output by the summary provides the estimators of the β_j's, which are exactly the classical least squares estimators. The `z value` column provides the values of the Z_j statistics defined by (4). The `Std.Error` column gives an estimation of the standard errors of the $\hat{\beta}_j$'s, which are taken equal to $\frac{\sqrt{\widehat{C}_{(j,j)}}}{d_j(n)}$. As with the `lm` function, the p-value column is the p-value for testing $\beta_j = 0$ against $\beta_j \neq 0$. In this example, the small p-value for the second feature $X2$ is consistent with the value chosen for `beta_vec` at the beginning of the section. The `chi2-statistic` at the end of the summary is the χ^2 statistic for testing the significance of the model (see the end of Section 2.2) For this example, the p-value is very small, indeed the variable $X2$ has a significant effect on Y.

Plot argument and plot method

`method_cov_st=`	plot
`fitAR`	ACF and PACF of the residual process
`kernel`	ACF of the residual process
`kernel` with `model_selec = -1`	Graph of the estimated risk and of the estimated $\gamma(k)$'s
`spectralproj`	Estimated spectral density
`select`	ACF of the estimated $\gamma(k)$'s
`Efromovich`	ACF of the estimated $\gamma(k)$'s

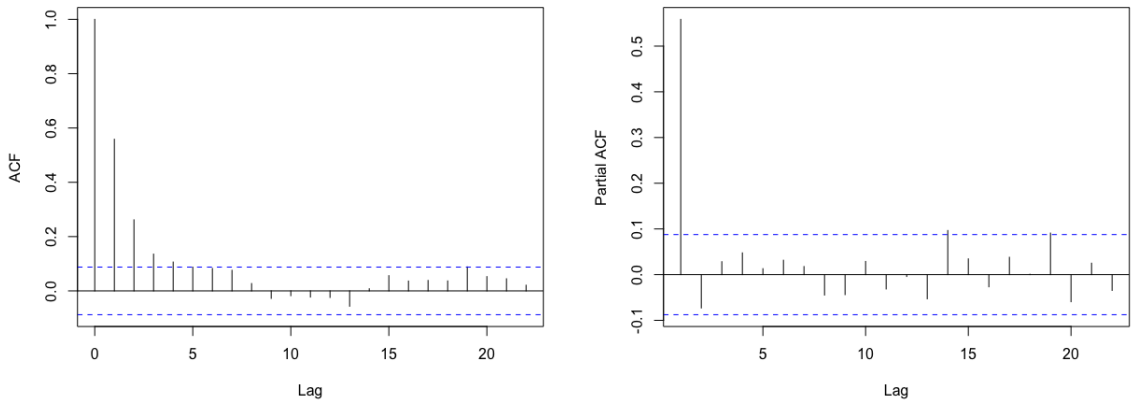
Table 1: Plot output for each method given in the `method_cov_st` of `slm`.

The `slm` function has a `plot` argument: with `plot = TRUE`, the function plots a figure which depends on the method chosen for estimating the covariance matrix C. Table 1 summarizes the plots for each method given in the argument `method_cov_st`. With the AR fitting method, the argument `plot = TRUE` outputs the ACF and the PACF of the residual process. The ACF and PACF are computed with the functions `acf` and `pacf` of the **stats** package. As usual, the ACF and PACF graphs should help the user to choose an appropriate order for the AR

process.

```
R> regslm = slm(Y ~ X1 + X2, data = design, method_cov_st = "fitAR",
+          model_selec = 2, plot = TRUE)
```

The plot output by the `slm` function for this example is given in Figure 1.



(a) ACF of the residual process.

(b) PACF of the residual process.

Figure 1: Plots output by `slm` for the fitAR method.

Since the `slm` class inherits from the `lm` class, the former class comes with a `plot` method which is the same as for the `lm` class, namely the diagnostic analysis of the linear regression. The graphics are displayed using the command

```
R> plot(regslm)
```

Confidence intervals for the coefficients

The `confint` function computes the confidence intervals for the coefficients β_j estimated by `slm`. These intervals are computed according to the distribution of the Z_j statistics defined in (4).

```
R> confint(regslm, level = 0.90)

                    5 %      95 %
(Intercept)  1.56396552 4.3083996
X1          -0.05048587 0.2192597
X2           0.48821351 0.4969666
```

AR order selection

The order p of the AR process can be chosen at hand by setting `model_selec = p`, or automatically with the AIC criterion by setting `model_selec = -1`.

```
R> regslm = slm(Y ~ X1 + X2, data = design, method_cov_st = "fitAR",
+          model_selec = -1)
```

The order of the fitted AR process is recorded in the `model_selec` attribute of `regslm`:

```
R> regslm@model_selec
```

```
[1] 2
```

Here, the AIC criterion suggests to fit an AR(2) process on the residuals.

3.2. Linear regression via kernel estimation of the error covariance

The second method for estimating the covariance matrix C is the kernel estimation method (1) studied in Caron (2019). In short, this method estimates C via a smooth approximation of the covariance matrix Γ_n of the residuals. This estimation of Γ_n corresponds to the so-called tapered covariance matrix estimator in the literature, see for instance Xiao and Wu (2012), or also to the "lag-window estimator" defined in Brockwell and Davis (1991), page 330. It applies in particular for non negative symmetric kernels with compact support, with an integrable Fourier transform and such that $K(0) = 1$. Table 2 gives the list of the available kernels in the package **slm**.

`kernel_fonc =`	kernel definition
`rectangular`	$K(x) = \mathbb{1}_{\{\lvert x\rvert \leq 1\}}$
`triangle`	$K(x) = (1 - \lvert x\rvert)\mathbb{1}_{\{\lvert x\rvert \leq 1\}}$
`trapeze`	$K(x) = \mathbb{1}_{\{\lvert x\rvert \leq \delta\}} + \frac{1}{1-\delta}(1 - \lvert x\rvert)\mathbb{1}_{\{\delta \leq \lvert x\rvert \leq 1\}}$

Table 2: Available kernel functions in **slm**.

It is also possible for the user to define his own kernel and to use it in the argument `kernel_fonc` of the `slm` function. Below we use the triangle kernel which assures that the covariance matrix is positive definite. The support of the kernel K in Equation (1) being compact, only the terms $\tilde{\gamma}_{j-l}$ for small enough lag $j-l$ are kept and weighted by the kernel in the expression of $\widetilde{\Gamma}_{n,h}$. Rather than setting the bandwidth h, we select the number of $\gamma(k)$'s that should be kept (the lag) with the argument `model_selec` in the `slm` function. Then the bandwidth h is calibrated accordingly, that is equal to `model_selec` $+ 1$.

```
R> regslm = slm(Y ~ X1 + X2, data = design, method_cov_st = "kernel",
+          model_selec = 5, kernel_fonc = triangle, plot = TRUE)
```

The plot output by the `slm` function is given in Figure 2.

Order selection via bootstrap

The order parameter can be chosen at hand as before or automatically by setting `model_selec = -1`. The automatic order selection is based on the bootstrap procedure proposed by Wu

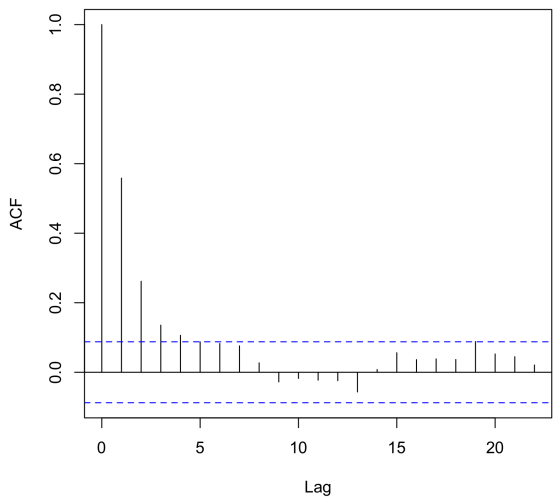


Figure 2: ACF of the residual process.

and Pourahmadi (2009) for banded covariance matrix estimation. The `block_size` argument sets the size of bootstrap blocks and the `block_n` argument sets the number of blocks. The final order is chosen by taking the order which has the minimal risk. Figure 3 gives the plots of the estimated risk for the estimation of Γ_n (left) and the final estimated ACF (right).

```
R> regslm = slm(Y ~ X1 + X2, data = design, method_cov_st ="kernel",
+          model_selec = -1, kernel_fonc = triangle, model_max = 30,
+          block_size = 100, block_n = 100, plot = TRUE)
```

The selected order is recorded in the `model_selec` attribute of the `slm` object output by the `slm` function:

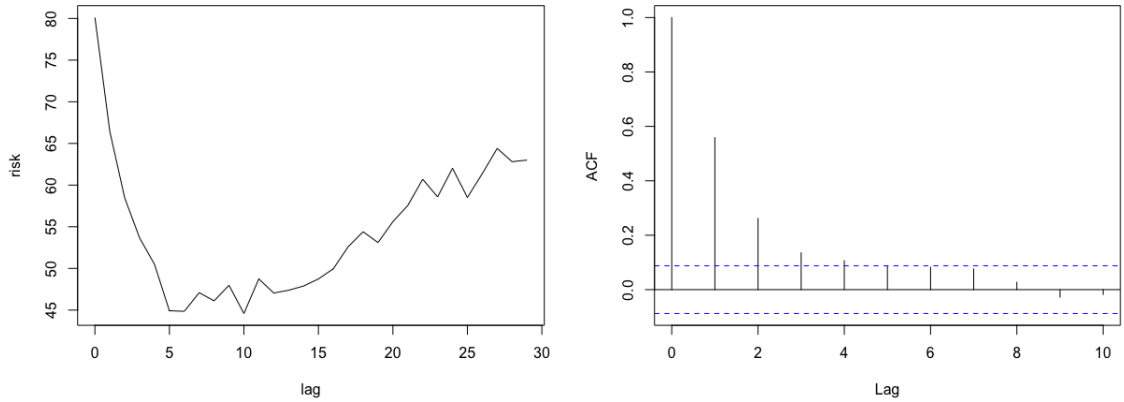
```
R> regslm@model_selec
```

```
[1] 10
```

Order selection by Efromovich's method (rectangular kernel)

An alternative method for choosing the bandwidth in the case of the rectangular kernel has been proposed in Efromovich (1998). For a large class of stationary processes with exponentially decaying autocovariance function (mainly the ARMA processes), Efromovich proved that the rectangular kernel is asymptotically minimax, and he proposed the following estimator:

$$\hat{f}_{J_{nr}}(\lambda) = \frac{1}{2\pi} \sum_{k=-J_{nr}}^{k=J_{nr}} \hat{\gamma}_k e^{ik\lambda},$$



(a) Estimated risk error via bootstrap.

(b) Estimated ACF for the selected order.

Figure 3: Plots output by `slm` for the kernel method with bootstrap selection of the order.

with the lag

$$J_{nr} = \frac{\log(n)}{2r}\left[1 + (\log(n))^{-1/2}\right],$$

where r is a regularity index of the autocovariance index. In practice this parameter is unknown and is estimated thanks to the algorithm proposed in the section 4 of Efromovich (1998). Like for the other methods, we use the residual based empirical covariances $\tilde{\gamma}_k$ to compute $\hat{f}_{J_{nr}}(\lambda)$.

```
R> regslm = slm(Y ~ X1 + X2, data = design, method_cov_st="efromovich",
+          model_selec = -1)
```

Positive definite projection

Depending of the method used, the matrix $\widehat{C}(\widehat{\Gamma}_n)$ may not always be positive definite. It is the case of the kernel method with rectangular or trapeze kernel. To overcome this problem, we make the projection of $\widehat{C}(\widehat{\Gamma}_n)$ into the cone of positive definite matrices by applying a hard thresholding on the spectrum of this matrix: we replace all eigenvalues lower or equal to zero with the smallest positive eigenvalue of $\widehat{C}(\widehat{\Gamma}_n)$.

3.3. Linear regression via projection spectral estimation

The projection method relies on the ideas of Comte (2001), where an adaptive nonparametric method has been proposed for estimating the spectral density of a stationary Gaussian process.

We use the residual process as a proxy for the error process and we compute the projection coefficients with the residual-based empirical covariance coefficients $\tilde{\gamma}_k$, see Equation (2).

For some $d \in \mathbb{N}^*$, the estimator of the spectral density of the error process that we use is defined by computing the projection estimators for the residual process, on the basis of

histogram functions

$$\phi_j^{(d)} = \sqrt{\frac{d}{\pi}} \mathbb{1}_{[\pi j/d, \pi(j+1)/d[}, \qquad j = 0, 1, \ldots, d-1.$$

The estimator is defined by

$$\hat{f}_d(\lambda) = \sum_{j=0}^{d-1} \hat{a}_j^{(d)} \phi_j^{(d)},$$

where the projection coefficients are

$$\hat{a}_j^{(d)} = \sqrt{\frac{d}{\pi}} \left(\frac{\tilde{\gamma}_0}{2d} + \frac{1}{\pi} \sum_{r=1}^{n-1} \frac{\tilde{\gamma}_r}{r} \left[\sin\left(\frac{\pi(j+1)r}{d} \right) - \sin\left(\frac{\pi j r}{d} \right) \right] \right).$$

The Fourier coefficients of the spectral density are equal to the covariance coefficients. Thus, for $k = 1, \ldots, n-1$ it yields

$$\begin{aligned} \gamma_k &= c_k \\ &= \frac{2}{k} \sqrt{\frac{d}{\pi}} \sum_{j=0}^{d-1} \hat{a}_j^{(d)} \left[\sin\left(\frac{k\pi(j+1)}{d} \right) - \sin\left(\frac{k\pi j}{d} \right) \right], \end{aligned}$$

and for $k = 0$:

$$\begin{aligned} \gamma_0 &= c_0 \\ &= 2\sqrt{\frac{\pi}{d}} \sum_{j=0}^{d-1} \hat{a}_j^{(d)}. \end{aligned}$$

This method can be proceeded in the `slm` function by setting `method_cov_st = "spectralproj"`:

```
R> regslm = slm(Y ~ X1 + X2, data = design, method_cov_st = "spectralproj",
+          model_selec = 10, plot = TRUE)
```

The graph of the estimated spectral density can be plotted by setting `plot = TRUE` in the `slm` function, see Figure 4.

Model selection

The Gaussian model selection method proposed in Comte (2001) follows the ideas of Birgé and Massart, see for instance Massart (2007). It consists in minimizing the l_2 penalized criterion, see Section 5 in Comte (2001):

$$\text{crit}(d) := -\sum_{j=0}^{d-1} \left[\hat{a}_j^{(d)} \right]^2 + c\frac{d}{n}$$

where c is a multiplicative constant that in practice can be calibrated using the slope heuristic method, see Birgé and Massart (2007); Baudry, Maugis, and Michel (2012) and the R package **capushe**.

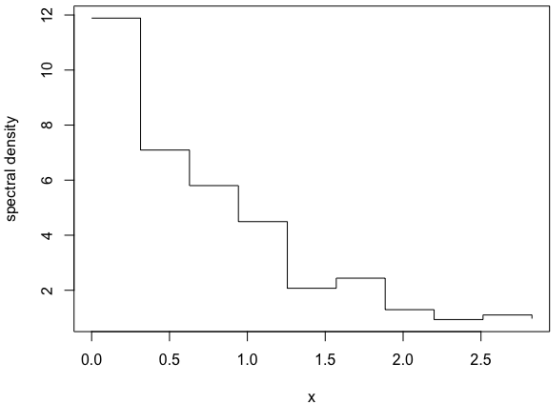


Figure 4: Spectral density estimator by projection on the histogram basis.

```
R> regslm = slm(Y ~ X1 + X2, data = design, method_cov_st = "spectralproj",
+          model_selec = -1, model_max = 50, plot = TRUE)
```

The selected dimension is recorded in the `model_selec` attribute of the `slm` object output by the `slm` function:

```
R> regslm@model_selec
```

```
[1] 8
```

The slope heuristic algorithm here selects an Histogram on a regular partition of size 8 over the interval $[0, \pi]$ to estimate the spectral density.

3.4. Linear regression via masked covariance estimation

This method is a full-manual method for estimating the covariance matrix C by only selecting covariance terms from the residual covariances $\tilde{\gamma}_k$ defined by (2). Let I be a set of positive integers, then we consider

$$\hat{\gamma}_I(k) := \tilde{\gamma}_k \mathbb{1}_{k \in I \cup \{0\}} \qquad 0 \leq |k| \leq n-1$$

and then we define the estimated covariance marix $\widehat{\Gamma}_I$ by taking the Toeplitz matrix of the vector $\hat{\gamma}_I$. This estimator is a particular example of masked sample covariance estimator, as introduced by Chen, Gittens, and Tropp (2012), see also Levina and Vershynin (2012). Finally we derive from $\widehat{\Gamma}_I$ an estimator $\widehat{C}(\widehat{\Gamma}_I)$ for C.

The next instruction selects the coefficients 0, 1, 2 and 4 from the residual covariance terms:

```
R> regslm = slm(Y ~ X1 + X2, data = design, method_cov_st = "select",
+          model_selec = c(1,2,4))
```

The positive lags of the selected covariances are recordered in the `model_selec` argument. Let us notice that the variance γ_0 is automatically selected.

Like for the kernel method, the resulting covariance matrix may not be positive definite. If it is the case, the positive definite projection method, described at the end of the section 3.2, is used.

3.5. Linear regression via manual plugged covariance matrix

This last method is a direct plug-in method. The user proposes his own vector estimator $\hat{\gamma}$ of γ and then the Toeplitz matrix $\widehat{\Gamma}_n$ of the vector $\hat{\gamma}$ is used for estimating C with $\widehat{C}(\widehat{\Gamma}_n)$.

```
R> v = rep(0,n)
R> v[1:10] = acf(epsilon, type = "covariance", lag.max = 9)$acf
R> regslm = slm(Y ~ X1 + X2, data = design, cov_st = v)
```

The user can also propose his own covariance matrix $\widehat{\Gamma}_n$ for estimating C.

```
R> v = rep(0,n)
R> v[1:10] = acf(epsilon, type = "covariance", lag.max = 9)$acf
R> V = toeplitz(v)
R> regslm = slm(Y ~ X1 + X2, data = design, Cov_ST = V)
```

Let us notice that the user must verify that the resulting covariance matrix is positive definite. The positive definite projection algorithm is not used with this method.

4. Numerical experiments and method comparisons

This section summarizes an extensive study which has been carried out to compare the performances of the different approaches presented before in the context of linear model with short range dependent stationary errors.

4.1. Description of the generative models

We first present the five generative models for the errors that we consider in the paper. We choose different kinds of processes to reflect the diversity of short-memory processes.

- **AR1 process.** The AR1 process is a gaussian AR(1) process defined by:

$$\varepsilon_i - 0.7\varepsilon_{i-1} = W_i,$$

  where W_i is a standard gaussian distribution $\mathcal{N}(0,1)$.

- **AR12 process.** The AR12 process is a seasonal AR(12) process defined by:

$$\varepsilon_i - 0.5\varepsilon_{i-1} - 0.2\varepsilon_{i-12} = W_i,$$

  where W_i is a standard gaussian distribution $\mathcal{N}(0,1)$. When studying monthly datasets, one usually observes a seasonality of order 12. For example, when looking at

climate data (such as the "CO2 concentration" dataset of Section 5), the data are often collected per month, and the same phenomenon tends to repeat every year. Even if the design integrates the deterministic part of the seasonality, a correlation of order 12 remains usually present in the residual process.

- **MA12 process.** The MA12 is also a seasonal process defined by:

$$\varepsilon_i = W_i + 0.5W_{i-2} + 0.3W_{i-3} + 0.2W_{i-12},$$

where the (W_i)'s are i.i.d. random variables following Student's distribution with 10 degrees of freedom.

- **Nonmixing process.** The three processes described above are basic ARMA processes, whose innovations have absolutely continuous distributions; in particular, they are strongly mixing in the sense of Rosenblatt (1956), with a geometric decay of the mixing coefficients (in fact the MA12 process is even 12-dependent, which means that the mixing coefficient $\alpha(k) = 0$ if $k > 12$). Let us now describe a more complicated process: let $(Z_1, \ldots, Z_n)$ satisfying the AR(1) equation

$$Z_{i+1} = \frac{1}{2}(Z_i + \eta_{i+1}),$$

where Z_1 is uniformly distributed over $[0, 1]$ and the η_i's are i.i.d. random variables with distribution $\mathcal{B}(1/2)$, independent of Z_1. The process $(Z_i)_{i\geq 1}$ is a strictly stationary Markov chain, but it is not α-mixing in the sense of Rosenblatt (see Bradley (1986)). Let now Q_{0,σ^2} be the inverse of the cumulative distribution function of a centered Gaussian distribution with variance σ^2 (for the simulations below, we choose $\sigma^2 = 25$). The Nonmixing process is then defined by

$$\varepsilon_i = Q_{0,\sigma^2}(Z_i).$$

The sequence $(\varepsilon_i)_{i\geq 1}$ is also a stationary Markov chain (as an invertible function of a stationary Markov chain). By construction, ε_i is $\mathcal{N}(0, \sigma^2)$-distributed, but the sequence $(\varepsilon_i)_{i\geq 1}$ is not a Gaussian process (otherwise it would be mixing in the sense of Rosenblatt). Although it is not obvious, one can prove that the process $(\varepsilon_i)_{i\geq 1}$ satisfies Hannan's condition (see Caron (2019), Section 4.2).

- **Sysdyn process.** The four processes described above have the property of "geometric decay of correlations", which means that the $\gamma(k)$'s tend to 0 at an exponential rate. However, as already pointed out in the introduction, Hannan's condition is valid for most of short memory processes, even for processes with slow decay of correlations (provided that the $\gamma(k)$'s are summable). Hence, our last example will be a non-mixing process (in the sense of Rosenblatt), with an arithmetic decay of the correlations.

  For $\gamma \in]0, 1[$, the intermittent map $\theta_\gamma : [0, 1] \mapsto [0, 1]$ introduced in Liverani, Saussol, and Vaienti (1999) is defined by

$$\theta_\gamma(x) = \begin{cases} x(1 + 2^\gamma x^\gamma) & \text{if } x \in [0, 1/2[\\ 2x - 1 & \text{if } x \in [1/2, 1]. \end{cases}$$

It follows from Liverani et al. (1999) that there exists a unique θ_γ-invariant probability measure ν_γ. The Sysdyn process is then defined by

$$\varepsilon_i = \theta_\gamma^i.$$

From Liverani et al. (1999), we know that, on the probability space $([0,1], \nu_\gamma)$, the autocorrelations $\gamma(k)$ of the stationary process $(\varepsilon_i)_{i\geq 1}$ are exactly of order $k^{-(1-\gamma)/\gamma}$. Hence $(\varepsilon_i)_{i\geq 1}$ is a short memory process provided $\gamma \in]0, 1/2[$. Moreover, it has been proved in Section 4.4 of Caron and Dede (2018) that $(\varepsilon_i)_{i\geq 1}$ satisfies Hannan's condition in the whole short-memory range, that is for $\gamma \in]0, 1/2[$. For the simulations below, we took $\gamma = 1/4$, which give autocorrelations $\gamma(k)$ of order k^{-3}.

The linear regression models simulated in the experiments all have the following form:

$$Y_i = \beta_1 + \beta_2(\log(i) + \sin(i) + Z_i) + \beta_3 i + \varepsilon_i, \qquad \text{for all } i \text{ in } \{1, \ldots, n\}, \tag{5}$$

where Z is a gaussian autoregressive process of order 1 and ε is one of the stationary processes defined above. For the simulations, β_1 is always equal to 3. All the error processes presented above can be simulated with the **slm** package with the `generative_process` function. The design can be simulated with the `generative_model` function.

4.2. Automatic calibration of the tests

It is of course of first importance to provide hypothesis tests with correct significance levels or at least with correct asymptotical significance levels, which is possible if the estimator $\widehat{\Gamma}_n$ of the covariance matrix Γ_n is consistent for estimating C. For instance, the results of Caron (2019) show that it is possible to construct statistical tests with correct asymptotical significance levels. However in practice such asymptotical results are not sufficient since they do not indicate how to tune the bandwidth on a given dataset. This situation makes the practice of linear regression with dependent errors really more difficult than linear regression with i.i.d. errors. This problem happens for several methods given before : order choice for the `fitAR` method, bandwidth choice for the `kernel` method, dimension selection for the `spectralproj` method.

It is a tricky issue to design a data driven procedure for choosing test parameters in order to have to correct Type I Error. Note that unlike with supervised problems and density estimation, it is not possible to calibrate hypothesis tests in practice using cross validation approaches. We thus propose to calibrate the tests using well founded statistical procedures for risk minimization : AIC criterion for the `fitAR` method, bootstrap procedures for the `kernel` method and slope heuristics for the `spectralproj` method. These procedures are implemented in the `slm` function with the `model_selec = -1` argument, as detailed in the previous section.

Let us first illustrate the calibration problem with the AR12 process. For $T = 1000$ simulations, we generate an error process of size n under the null hypothesis: $H_0 : \beta_2 = \beta_3 = 0$. Then we use the `fitAR` method of the `slm` function with orders between 1 and 50 and we perform the model significance test. The procedure is repeated 1000 times and we estimate the true level of the test by taking the average of the estimated levels on the 1000 simulations

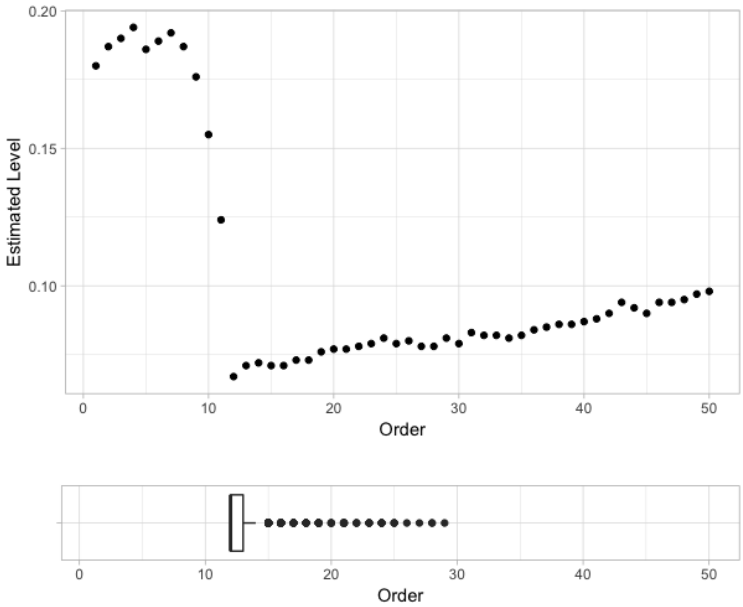


Figure 5: Estimated level of the test according to the order of the fitted AR process on the residuals (top) and boxplot of the order selected by AIC, over 1000 simulations. The data has been simulated according to Model (5) with $\beta_1 = 3$ and $\beta_2 = \beta_3 = 0$, with $n = 1000$.

for each order. The results are given on Figure 5 for $n = 1000$. A boxplot is also displayed to visualize the distribution of the order selected by the automatic criterion (AIC).

4.3. Non-Seasonal errors

We first study the case of non-Seasonal error processes. We simulate a n-error process according to the AR1, the Nonmixing or the Sysdyn processes. We simulate realizations of the linear regression model (5) under the null hypothesis: $H_0 : \beta_2 = \beta_3 = 0$. We use the automatic selection procedures for each method (`model_selec = -1`). The simulations are repeated 1000 times in order to estimate the true level of the model significance for each test procedure. We simulate either small samples ($n = 200$) or larger samples ($n = 1000, 2000, 5000$). The results of this experiments are summarized in Table 3.

For n large enough ($n \geq 1000$), all methods work well and the estimated level is around 0.05. However, for small samples ($n = 200$), it is more complicated. We can observe that the fitAR method works better than the others. The kernel method is slightly less effective. With this method, we must choose the size of the bootstrap blocks as well as the number of blocks and the test results are really sensitive to these parameters. In these simulations, we have chosen 100 blocks with a size of $n/2$. The results are expected to improve with a larger number of blocks.

Let us notice that for all methods and for all sample sizes, the estimated level is much better than if no correction is made (usual Fisher tests).

n	Method / Process	Fisher i.i.d.	fitAR	spectralproj	efromovich	kernel
200	AR1 process	0.465	**0.109**	0.154	0.149	0.164
	NonMixing	0.298	**0.075**	0.106	0.093	0.13
	Sysdyn process	0.385	**0.108**	0.123	0.132	0.168
1000	AR1 process	0.418	0.04	0.049	**0.05**	0.084
	NonMixing	0.298	0.043	0.048	**0.051**	0.078
	Sysdyn process	0.393	**0.073**	0.078	0.082	0.075
2000	AR1 process	0.454	**0.068**	0.075	0.072	0.071
	NonMixing	0.313	**0.05**	0.057	0.059	0.068
	Sysdyn process	0.355	**0.065**	0.067	**0.065**	0.075
5000	AR1 process	0.439	0.044	0.046	**0.049**	0.046
	NonMixing	0.315	**0.051**	0.056	0.058	0.068
	Sysdyn process	0.381	**0.059**	0.062	**0.059**	0.066

Table 3: Estimated levels for the non-seasonal processes.

4.4. Seasonal errors

We now study the case of linear regression with seasonal errors. The experiment is exactly the same as before, except that we simulate AR12 or MA12 processes. The results of these experiments are summarized in Table 4.

n	Method / Process	Fisher i.i.d.	fitAR	spectralproj	efromovich	kernel
200	AR12 process	0.436	**0.189**	0.222	0.241	0.241
	MA12 process	0.228	0.124	**0.12**	0.125	0.152
1000	AR12 process	0.468	**0.069**	0.187	0.188	0.127
	MA12 process	0.209	**0.064**	0.072	0.074	**0.064**
2000	AR12 process	0.507	**0.075**	0.197	0.158	0.112
	MA12 process	0.237	0.064	0.065	**0.056**	0.067
5000	AR12 process	0.47	**0.054**	0.188	0.1	0.086
	MA12 process	0.242	**0.05**	0.052	0.045	0.057

Table 4: Estimated levels for the seasonal processes.

We directly see that the case of seasonal processes is more complicated than for the non-seasonal processes especially for the AR12 process. For small sample size, the estimated level is around 0.2, which is clearly too large. It is however much better than the estimated level of the usual Fisher test, which is around 0.45. The fitAR method is the best method here for the AR12 process, because, for $n = 1000$, the estimated level is close to 0.05. For the other methods (except `spectralproj`), a level close to 5% is reached but for large samples only. The `spectralproj` method does not seem to work well for the AR12 process, although it remains much better than the usual Fisher tests (around 19% of rejection instead of 45%). The case of the MA12 process is better. For n large enough ($n \geq 1000$), the estimated level

is around 0.05, whatever the method. It is less effective for small sample size ($n = 200$) but no method is really better than another, with an estimated level around 0.12.

5. Application to real data

5.1. Data CO2

Let us introduce the first dataset that we want to study. It concerns the well-known dataset "co2", available in the package **datasets** of R:

```
R> data("co2")
```

This dataset is provided by the observatory of Mona Loa (Hawaii). It contains average monthly measurements of CO2 (parts per million: ppmv) in the atmosphere of the Hawaiian coast. Surveys were produced monthly between 1959 and 1998, giving a total of 468 measurements. The graph of the data is displayed in Figure 6. More information on this dataset is available in the R documentation.

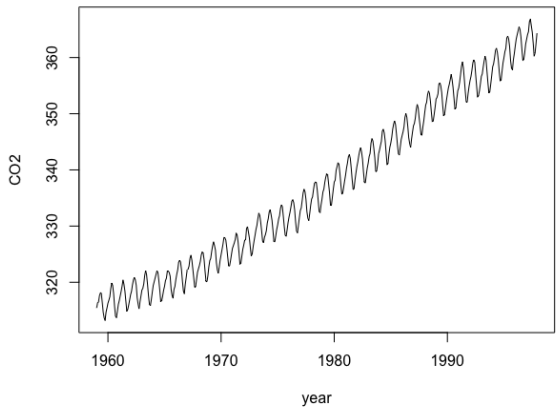


Figure 6: CO2 rate as a function of time.

We model the CO2 measurements with a time series. Typically, a time series can be decomposed into three parts: a trend m and a seasonality s, which are deterministic components, and the errors ε, which constitute the random part of the model. The trend represents the overall behavior of the series and seasonality its periodic behavior. Formally, we have:

$$Y_t = m_t + s_t + \varepsilon_t,$$

where Y_t represents the CO2 rate at time t, with the usual constraints $s_t = s_{t+12}$ and $\sum_{t=1}^{12} s_t = 0$. The two deterministic components can be grouped into a matrix X and the model can be rewritten into a linear regression model:

$$Y = X\beta + \varepsilon.$$

For this example, we fit a 3-degree polynomial for the trend and a trigonometric polynomial with well-chosen frequencies for the seasonality. Here the time t represents a month and t goes from 1 to 40 by step of length 1/12. Let us perform a linear regression to fit the trend and the seasonality on the CO2 time series, using the `lm` function:

```
R> y = as.vector(co2)
R> t = as.vector(time(co2)) - 1958
R> regtrigo = lm(y ~ t + I(t^2) + I(t^3) + sin(2*pi*t) + cos(2*pi*t)
+          + sin(4*pi*t) + cos(4*pi*t) + sin(6*pi*t) + cos(6*pi*t)
+          + sin(8*pi*t) + cos(8*pi*t))
```

We obtain the following output:

```
R> summary.lm(regtrigo)

Call:
lm(formula = y ~ t + I(t^2) + I(t^3) + sin(2 * pi * t) + cos(2 *
    pi * t) + sin(4 * pi * t) + cos(4 * pi * t) + sin(6 * pi *
    t) + cos(6 * pi * t) + sin(8 * pi * t) + cos(8 * pi * t))

Residuals:
     Min       1Q   Median       3Q      Max
-1.54750 -0.32688  0.00233  0.28100  1.50295

Coefficients:
                  Estimate Std. Error  t value Pr(>|t|)
(Intercept)      3.157e+02  1.118e-01 2823.332  < 2e-16 ***
t                3.194e-01  2.306e-02   13.847  < 2e-16 ***
I(t^2)           4.077e-02  1.293e-03   31.523  < 2e-16 ***
I(t^3)          -4.562e-04  2.080e-05  -21.930  < 2e-16 ***
sin(2 * pi * t)  2.751e+00  3.298e-02   83.426  < 2e-16 ***
cos(2 * pi * t) -3.960e-01  3.296e-02  -12.015  < 2e-16 ***
sin(4 * pi * t) -6.743e-01  3.296e-02  -20.459  < 2e-16 ***
cos(4 * pi * t)  3.785e-01  3.296e-02   11.484  < 2e-16 ***
sin(6 * pi * t) -1.042e-01  3.296e-02   -3.161  0.00168 **
cos(6 * pi * t) -4.389e-02  3.296e-02   -1.332  0.18362
sin(8 * pi * t)  8.733e-02  3.296e-02    2.650  0.00833 **
cos(8 * pi * t)  2.559e-03  3.296e-02    0.078  0.93814
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5041 on 456 degrees of freedom
Multiple R-squared:  0.9989,        Adjusted R-squared:  0.9989
F-statistic: 3.738e+04 on 11 and 456 DF,  p-value: < 2.2e-16
```

We see in the summary that two variables have no significant effect on the CO2 rate. Next, we perform a backward selection method with a p-value threshold equal to 0.05. This selects the following model:

```
R> regtrigo = lm(y ~ t + I(t^2) + I(t^3) + sin(2*pi*t) + cos(2*pi*t)
+          + sin(4*pi*t) + cos(4*pi*t) + sin(6*pi*t) + sin(8*pi*t))
```

with the corresponding summary

```
R> summary.lm(regtrigo)

Call:
lm(formula = y ~ t + I(t^2) + I(t^3) + sin(2 * pi * t) + cos(2 *
    pi * t) + sin(4 * pi * t) + cos(4 * pi * t) + sin(6 * pi *
    t) + sin(8 * pi * t))

Residuals:
     Min       1Q   Median       3Q      Max
-1.59287 -0.32364  0.00226  0.29884  1.50154

Coefficients:
                  Estimate Std. Error  t value Pr(>|t|)
(Intercept)      3.157e+02  1.118e-01 2824.174  < 2e-16 ***
t                3.196e-01  2.306e-02   13.861  < 2e-16 ***
I(t^2)           4.075e-02  1.293e-03   31.522  < 2e-16 ***
I(t^3)          -4.560e-04  2.080e-05  -21.927  < 2e-16 ***
sin(2 * pi * t)  2.751e+00  3.297e-02   83.446  < 2e-16 ***
cos(2 * pi * t) -3.960e-01  3.295e-02  -12.018  < 2e-16 ***
sin(4 * pi * t) -6.743e-01  3.295e-02  -20.464  < 2e-16 ***
cos(4 * pi * t)  3.785e-01  3.295e-02   11.487  < 2e-16 ***
sin(6 * pi * t) -1.042e-01  3.295e-02   -3.162  0.00167 **
sin(8 * pi * t)  8.734e-02  3.295e-02    2.651  0.00831 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.504 on 458 degrees of freedom
Multiple R-squared:  0.9989,        Adjusted R-squared:  0.9989
F-statistic: 4.57e+04 on 9 and 458 DF,  p-value: < 2.2e-16
```

The sum of the estimated trend and estimated tendency is displayed on the left plot of Figure 7, and the residuals are displayed on the right plot. The `lm` procedure assumes that the errors are independent, but if we look at the autocorrelation function of the residual process we clearly observe that the residuals are strongly correlated, see Figure 8. Consequently, the `lm` procedure may be unreliable in this context.

The autocorrelation function of the residuals decreases rather fast. Looking at the partial autocorrelation function, it seems reasonable to fit an AR process on the residuals. The automatic `fitAR` method selects an AR of order 14 and the residuals look like a white noise, see Figure 9.

We now use the `slm` function with the `fitAR` method with the following complete model

```
R> regtrigo = slm(y ~ t + I(t^2) + I(t^3) + sin(2*pi*t) + cos(2*pi*t)
+          + sin(4*pi*t) + cos(4*pi*t) + sin(6*pi*t) + cos(6*pi*t)
```

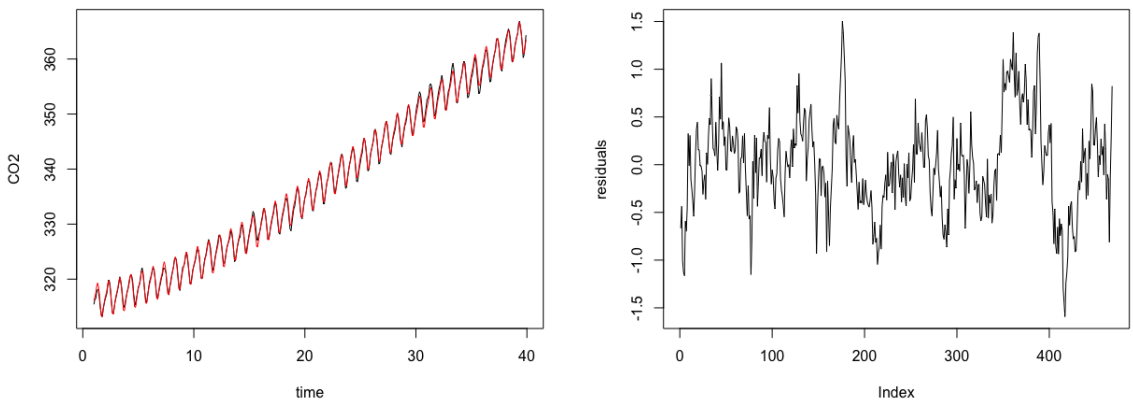


Figure 7: CO2 adjustment (left) and residuals (right)

```
+           + sin(8*pi*t) + cos(8*pi*t), method_cov_st = "fitAR",
+           model_selec = -1)
```

Let us display the summary of the procedure:

```
R> summary(regtrigo)

Call:
"slm(formula = myformula, data = data, x = x, y = y)"

Residuals:
     Min       1Q   Median       3Q      Max
-1.54750 -0.32688  0.00233  0.28100  1.50295

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      3.157e+02  3.968e-01 795.646  < 2e-16 ***
t                3.194e-01  8.222e-02   3.884 0.000103 ***
I(t^2)           4.077e-02  4.619e-03   8.825  < 2e-16 ***
I(t^3)          -4.562e-04  7.430e-05  -6.140 8.23e-10 ***
sin(2 * pi * t)  2.751e+00  4.739e-02  58.054  < 2e-16 ***
cos(2 * pi * t) -3.960e-01  4.716e-02  -8.396  < 2e-16 ***
sin(4 * pi * t) -6.743e-01  2.051e-02 -32.875  < 2e-16 ***
cos(4 * pi * t)  3.785e-01  2.041e-02  18.548  < 2e-16 ***
sin(6 * pi * t) -1.042e-01  1.359e-02  -7.663 1.82e-14 ***
cos(6 * pi * t) -4.389e-02  1.359e-02  -3.228 0.001245 **
sin(8 * pi * t)  8.733e-02  1.246e-02   7.009 2.41e-12 ***
cos(8 * pi * t)  2.559e-03  1.252e-02   0.204 0.838038
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

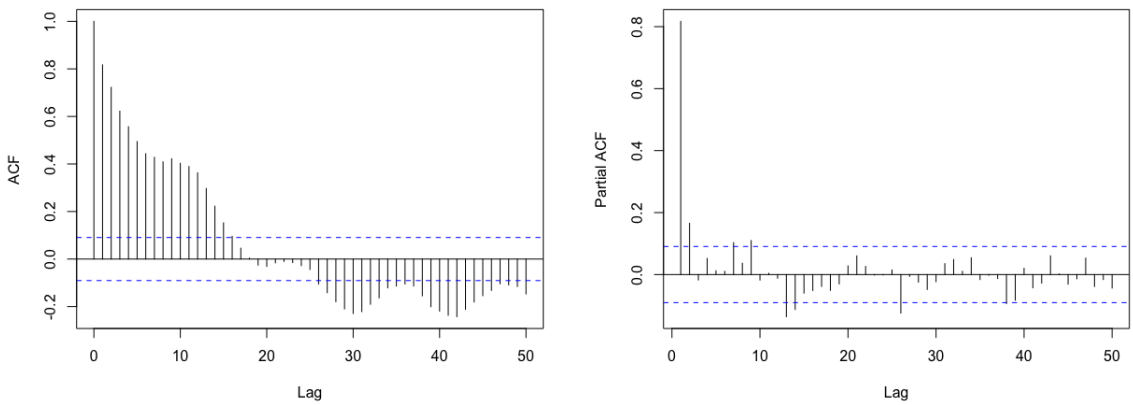


Figure 8: Autocorrelation function (left) and partial autocorrelation function (right) of the residuals.

```
Residual standard error: 0.5041
Multiple R-squared:  0.9989
chi2-statistic: 1.724e+04 on 11 DF,  p-value: < 2.2e-16
```

The last variable has no significant effect on the CO2. After performing a backward selection method with a p-value threshold equal to 0.05, we obtain the following model

```
R> regtrigo = slm(y ~ t + I(t^2) + I(t^3) + sin(2*pi*t) + cos(2*pi*t)
+           + sin(4*pi*t) + cos(4*pi*t) + sin(6*pi*t) + cos(6*pi*t)
+           + sin(8*pi*t), method_cov_st = "fitAR", model_selec = -1)
```

and the associated summary

```
R> summary(regtrigo)

Call:
"slm(formula = myformula, data = data, x = x, y = y)"

Residuals:
     Min       1Q   Median       3Q      Max
-1.54877 -0.32432  0.00187  0.28069  1.50168

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      3.157e+02  3.969e-01 795.627  < 2e-16 ***
t                3.194e-01  8.223e-02   3.884 0.000103 ***
I(t^2)           4.077e-02  4.619e-03   8.825  < 2e-16 ***
I(t^3)          -4.562e-04  7.430e-05  -6.140 8.23e-10 ***
```

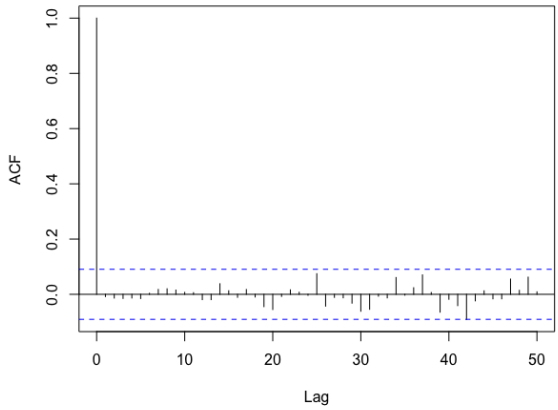


Figure 9: Autocorrelation function of the residuals for the AR fitting.

```
sin(2 * pi * t)  2.751e+00  4.738e-02  58.061  < 2e-16 ***
cos(2 * pi * t) -3.960e-01  4.716e-02  -8.397  < 2e-16 ***
sin(4 * pi * t) -6.743e-01  2.051e-02 -32.874  < 2e-16 ***
cos(4 * pi * t)  3.785e-01  2.041e-02  18.547  < 2e-16 ***
sin(6 * pi * t) -1.042e-01  1.359e-02  -7.664 1.80e-14 ***
cos(6 * pi * t) -4.389e-02  1.359e-02  -3.229 0.001244 **
sin(8 * pi * t)  8.733e-02  1.248e-02   6.998 2.60e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5036
Multiple R-squared:  0.9989
chi2-statistic: 1.724e+04 on 10 DF,  p-value: < 2.2e-16
```

There is a clear difference between the two backward procedures: `slm` keeps the variable $\cos(6\pi x)$, while `lm` does not. Given the obvious dependency of the error process, we recommend using `slm` instead of `lm` in this context.

5.2. PM2.5 Data of Shanghai

This dataset comes from a study about fine particle pollution in five Chinese cities. The data are available on the following website https://archive.ics.uci.edu/ml/datasets/PM2.5+Data+of+Five+Chinese+Cities#. We are interested here by the city of Shanghai. We study the regression of PM2.5 pollution in Xuhui District by other measurements of pollution in neighboring districts and also by meteorological variables. The dataset contains hourly observations between January 2010 and December 2015. More precisely it contains 52584 records of 17 variables: date, time of measurement, pollution and meteorological variables. More information on these data is available in the paper of Liang, Li, Zhang, Huang, and Chen (2016).

We remove the lines that contain NA observations and we then extract the first 5000 observations. For simplicity, we will only consider pollution variables and weather variables. We start the study with the following 10 variables:

- PM_Xuhui: PM2.5 concentration in the Xuhui district (ug/m^3)
- PM_Jingan: PM2.5 concentration in the Jing'an district (ug/m^3)
- PM_US.Post: PM2.5 concentration in the U.S diplomatic post (ug/m^3)
- DEWP: Dew Point (Celsius Degree)
- TEMP: Temperature (Celsius Degree)
- HUMI: Humidity (%)
- PRES: Pressure (hPa)
- Iws: Cumulated wind speed (m/s)
- precipitation: hourly precipitation (mm)
- Iprec: Cumulated precipitation (mm)

```
R> shan = read.csv("ShanghaiPM20100101_20151231.csv", header = TRUE,
+          sep = ",")
R> shan = na.omit(shan)
R> shan_complete = shan[1:5000,c(7,8,9,10,11,12,13,15,16,17)]
R> shan_complete[1:5,]
```

```
      PM_Jingan PM_US.Post PM_Xuhui DEWP  HUMI PRES TEMP Iws
26305        66         70       71   -5 69.00 1023    0  60
26306        67         76       72   -5 69.00 1023    0  62
26308        73         78       74   -4 74.41 1023    0  65
26309        75         77       77   -4 80.04 1023   -1  68
26310        73         78       80   -4 80.04 1023   -1  70
      precipitation Iprec
26305             0     0
26306             0     0
26308             0     0
26309             0     0
26310             0     0
```

The aim is to study the concentration of particles in Xuhui District according to the other variables. We first fit a linear regression with the `lm` function:

```
R> reglm = lm(shan_complete$PM_Xuhui ~ . ,data = shan_complete)
R> summary.lm(reglm)
```

```
Call:
lm(formula = shan_complete$PM_Xuhui ~ ., data = shan_complete)

Residuals:
     Min       1Q   Median       3Q      Max
-132.139   -4.256   -0.195    4.279  176.450

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -54.859483  40.975948  -1.339 0.180690
PM_Jingan       0.596490   0.014024  42.533  < 2e-16 ***
PM_US.Post      0.375636   0.015492  24.246  < 2e-16 ***
DEWP           -1.038941   0.170144  -6.106 1.10e-09 ***
HUMI            0.291713   0.045799   6.369 2.07e-10 ***
PRES            0.025287   0.038915   0.650 0.515852
TEMP            1.305543   0.168754   7.736 1.23e-14 ***
Iws            -0.007650   0.002027  -3.774 0.000163 ***
precipitation   0.462885   0.132139   3.503 0.000464 ***
Iprec          -0.125456   0.039025  -3.215 0.001314 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.68 on 4990 degrees of freedom
Multiple R-squared:  0.9409,        Adjusted R-squared:  0.9408
F-statistic:  8828 on 9 and 4990 DF,  p-value: < 2.2e-16
```

The variable PRES has no significant effect on the PM_Xuhui variable. We then perform a backward selection procedure, which leads to select 9 significant variables:

```
R> shan_lm = shan[1:5000,c(7,8,9,10,11,13,15,16,17)]
R> reglm = lm(shan_lm$PM_Xuhui ~ . ,data = shan_lm)
R> summary.lm(reglm)

Call:
lm(formula = shan_lm$PM_Xuhui ~ ., data = shan_lm)

Residuals:
     Min       1Q   Median       3Q      Max
-132.122   -4.265   -0.168    4.283  176.560

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -28.365506   4.077590  -6.956 3.94e-12 ***
PM_Jingan       0.595564   0.013951  42.690  < 2e-16 ***
PM_US.Post      0.376486   0.015436  24.390  < 2e-16 ***
DEWP           -1.029188   0.169471  -6.073 1.35e-09 ***
HUMI            0.285759   0.044870   6.369 2.08e-10 ***
```

```
TEMP            1.275880   0.162453   7.854 4.90e-15 ***
Iws            -0.007734   0.002023  -3.824 0.000133 ***
precipitation   0.462137   0.132127   3.498 0.000473 ***
Iprec          -0.127162   0.038934  -3.266 0.001098 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.68 on 4991 degrees of freedom
Multiple R-squared:  0.9409,        Adjusted R-squared:  0.9408
F-statistic:  9933 on 8 and 4991 DF,  p-value: < 2.2e-16
```

The autocorrelation of the residual process shows that the errors are clearly not i.i.d., see Figure 10. We thus suspect the `lm` procedure to be unreliable in this context.

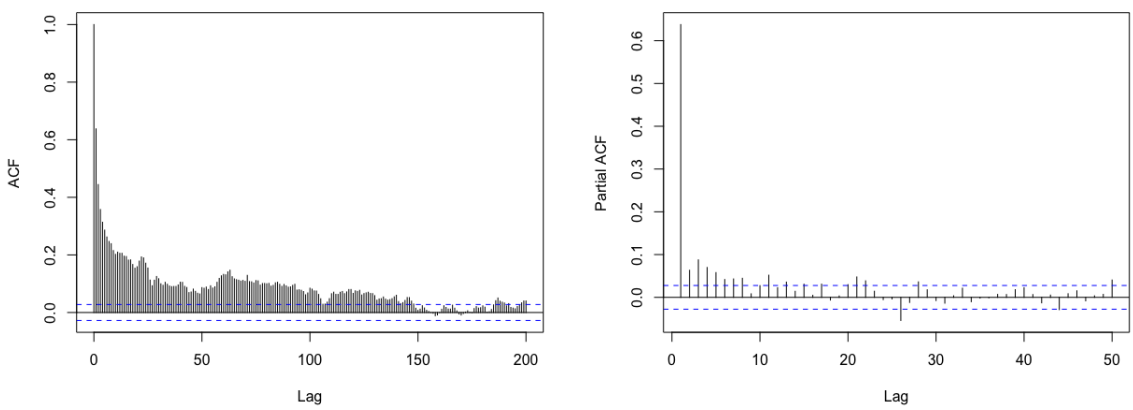


Figure 10: Autocorrelation function (left) and partial autocorrelation function (right) of the residuals.

The autocorrelation function decreases pretty fast, and the partial autocorrelation function suggests that fitting an AR process on the residuals should be an appropriate method in this case. The automatic `fitAR` method of `slm` selects an AR process of order 28. The residuals of this AR fitting look like a white noise, as shown in Figure 11. Consequently, we propose to perform a linear regression with `slm` function, using the fitAR method on the complete model

```
R> regslm = slm(shan_complete$PM_Xuhui ~ . ,data = shan_complete,
+          method_cov_st = "fitAR", model_selec = -1)
R> summary(regslm)

Call:
"slm(formula = myformula, data = data, x = x, y = y)"

Residuals:
     Min       1Q   Median       3Q      Max
```

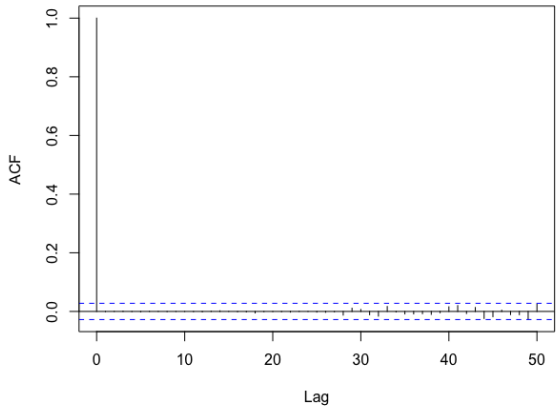


Figure 11: Autocorrelation function of the residuals for the AR fitting.

```
-132.139   -4.256   -0.195    4.279  176.450

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)   -54.859483 143.268399  -0.383 0.701783
PM_Jingan       0.596490   0.028467  20.953  < 2e-16 ***
PM_US.Post      0.375636   0.030869  12.169  < 2e-16 ***
DEWP           -1.038941   0.335909  -3.093 0.001982 **
HUMI            0.291713   0.093122   3.133 0.001733 **
PRES            0.025287   0.137533   0.184 0.854123
TEMP            1.305543   0.340999   3.829 0.000129 ***
Iws            -0.007650   0.005698  -1.343 0.179399
precipitation   0.462885   0.125641   3.684 0.000229 ***
Iprec          -0.125456   0.064652  -1.940 0.052323 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.68
Multiple R-squared:  0.9409
chi2-statistic:  3183 on 9 DF,  p-value: < 2.2e-16
```

Note that the variables show globally larger p-values than with the `lm` procedure, and more variables have no significant effect than with `lm`. After performing a backward selection we obtain the following results

```
R> shan_slm = shan[1:5000,c(7,8,9,10,11,13)]
R> regslm = slm(shan_slm$PM_Xuhui ~ . , data = shan_slm,
+          method_cov_st = "fitAR", model_selec = -1)
R> summary(regslm)
```

```
Call:
"slm(formula = myformula, data = data, x = x, y = y)"

Residuals:
     Min       1Q   Median       3Q      Max
-132.263   -4.341   -0.192    4.315  176.501

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -29.44924    8.38036  -3.514 0.000441 ***
PM_Jingan     0.60063    0.02911  20.636  < 2e-16 ***
PM_US.Post    0.37552    0.03172  11.840  < 2e-16 ***
DEWP         -1.05252    0.34131  -3.084 0.002044 **
HUMI          0.28890    0.09191   3.143 0.001671 **
TEMP          1.30069    0.32435   4.010 6.07e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.71
Multiple R-squared:  0.9406
chi2-statistic:  3751 on 5 DF,  p-value: < 2.2e-16
```

The backward selection with `slm` only keeps 5 variables.

Acknowledgements

The authors are grateful to Anne Philippe and Aymeric Stamm for valuable discussions.

References

Baudry JP, Maugis C, Michel B (2012). "Slope heuristics: overview and implementation." Statistics and Computing, **22**(2), 455–470.

Birgé L, Massart P (2007). "Minimal penalties for Gaussian model selection." Probability theory and related fields, **138**(1-2), 33–73.

Bradley RC (1986). "Basic properties of strong mixing conditions." In Dependence in probability and statistics (Oberwolfach, 1985), volume 11 of Progr. Probab. Statist., pp. 165–192. Birkhäuser Boston, Boston, MA.

Brockwell PJ, Davis RA (1991). Time Series: Theory and Methods. Springer Science & Business Media.

Caron E (2019). "Asymptotic distribution of least square estimators for linear models with dependent errors." Statistics, **53**(4), 885–902. doi:10.1080/02331888.2019.1593987. https://doi.org/10.1080/02331888.2019.1593987.

Caron E, Dede S (2018). "Asymptotic Distribution of Least Squares Estimators for Linear Models with Dependent Errors: Regular Designs." Mathematical Methods of Statistics, **27**(4), 268–293.

Chen RY, Gittens A, Tropp JA (2012). "The masked sample covariance estimator: an analysis using matrix concentration inequalities." Information and Inference: A Journal of the IMA, **1**(1), 2–20.

Comte F (2001). "Adaptive estimation of the spectrum of a stationary Gaussian sequence." Bernoulli, **7**(2), 267–298.

Efromovich S (1998). "Data-driven efficient estimation of the spectral density." Journal of the American Statistical Association, **93**(442), 762–769.

Hannan EJ (1973). "Central limit theorems for time series regression." Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete, **26**(2), 157–170.

Levina E, Vershynin R (2012). "Partial estimation of covariance matrices." Probability theory and related fields, **153**(3-4), 405–419.

Liang X, Li S, Zhang S, Huang H, Chen SX (2016). "PM2.5 data reliability, consistency, and air quality assessment in five Chinese cities." Journal of Geophysical Research: Atmospheres, **121**(17), 10–220.

Liverani C, Saussol B, Vaienti S (1999). "A probabilistic approach to intermittency." Ergodic theory and dynamical systems, **19**(3), 671–685.

Massart P (2007). Concentration inequalities and model selection, volume 1896 of Lecture Notes in Mathematics. Springer-Verlag Berlin Heidelberg.

Rosenblatt M (1956). "A central limit theorem and a strong mixing condition." Proceedings of the National Academy of Sciences, **42**(1), 43–47.

Wu WB, Pourahmadi M (2009). "Banding sample autocovariance matrices of stationary processes." Statistica Sinica, pp. 1755–1768.

Xiao H, Wu WB (2012). "Covariance matrix estimation for stationary time series." The Annals of Statistics, **40**(1), 466–493.

Affiliation:

Emmanuel Caron, Bertrand Michel
Laboratoire de Mathématiques Jean Leray UMR 6629
Ecole Centrale Nantes
44300 Nantes, France
E-mail: emmanuel.caron@ec-nantes.fr, bertrand.michel@ec-nantes.fr
URL: http://ecaron.perso.math.cnrs.fr, http://bertrand.michel.perso.math.cnrs.fr

Jérôme Dedecker
Laboratoire MAP5 UMR 8145
Université Paris Descartes
75006 Paris, France
E-mail: jerome.dedecker@parisdescartes.fr
URL: http://w3.mi.parisdescartes.fr/~jdedecke/