A longitudinal study of the influence of air pollutants on children. A robust multivariate approach

Ian Meneghel Danilevicz, Pascal Bondon, Valderio A. Reisen, Faradiba Sarquis

A longitudinal study of the influence of air pollutants on children. A robust multivariate approach

Ian Meneghel Danilevicz*[1, 2], Pascal Bondon[2], Valdério Anselmo Reisen[1, 2, 3, 4] and Faradiba Sarquis Serpa[5]

[1]Universidade Federal de Minas Gerais, Department of Statistics, Brazil.
[2]Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des signaux et systèmes, 91190, Gif-sur-Yvette, France.
[3]Universidade Federal do Espírito Santo, Postgraduate Programs in Environmental Engineering (PPGEA), in Mathematics (PPGMAT) and Economics (PPGEco), UFES, Brazil.
[4]Universidade Federal da Bahia, Department of Statistics, Brazil.
[5]Escola Superior de Ciências da Santa Casa de Misericórdia de Vitória (EMESCAM), Brazil.

Abstract

This paper aims to evaluate the statistical association between exposure to air pollution and forced expiratory volume in the first second (FEV_1) in both asthmatic and non-asthmatic children and teenagers, in which the response variable FEV_1 was repeatedly measured on a monthly basis, characterizing a longitudinal experiment. Due to the nature of the data, a robust linear mixed model (RLMM), combined with a robust principal component analysis (RPCA), is proposed to handle the multicollinearity among the covariates and the impact of extreme observations (high levels of air contaminants) on the estimates. The Huber and Tukey loss functions are considered to obtain robust estimators of the parameters in the linear mixed model (LMM). A finite sample size investigation is conducted under the scenario where the covariates follow linear time series models with and without additive outliers (AO). The impact of the time-correlation and the outliers on the estimates of the fixed effect parameters in the LMM is investigated. In the real data analysis, the robust model strategy evidenced that RPCA exhibits three principal component (PC), mainly related to relative humidity (Hmd), particulate matter with a diameter smaller than 10 μm (PM_{10}) and particulate matter with a diameter smaller than 2.5 μm ($PM_{2.5}$).

Keywords: Linear mixed model, Principal component analysis, M-estimation, robustness, outliers, asthma.

1 Introduction

The development in cities not only benefits the local economy, such as by creating jobs and contributing to urban development, but it also generates various residues which lead to environmental

*Corresponding author, email: iandanilevicz @ gmail.com

and health problems, affecting the quality of life of the population among others. The World Health Organization (WHO) conceptualizes health as a "state of complete physical, mental and social well-being, and not simply the absence of disease or illness" (World Health Organization, 2006). As is widely addressed in the environmental epidemiological studies, children are more vulnerable to air pollution since their respiratory and immunologic systems are immature. Consequently, they suffer more severe mortality and morbidity (Thurston et al., 2017; Abidin et al., 2014; Favarato et al., 2014). In this context, the statistical association between air quality variables and health effects must be computed with caution, independently of the statistical regression and time series models used. In the statistical models in epidemiological studies, a standard strategy is to consider air pollutants and weather as covariate variables. In general, these variables are time-correlated and present multicollinearity. Ignoring these phenomena in the modelling steps can lead to a wrong model selection and have severe consequences on the analysis of the impact of the pollutants on health, like a false-positive conclusion of the population health risk. Additionally, high levels of pollutants frequently appear in the pollutant variables, but their impact is often ignored in the literature. However, these observations can be identified as AO which affect the estimation of some statistical characteristics of the data, like the mean, the variance, and the correlation. Hence, robust estimation methods are needed to get reliable statistical models see, for example, Reisen et al. (2017); Cotta et al. (2020).

Many studies have recently paid attention to these issues in quantifying the association between pollutants and adverse health effects. For example, principal component analysis (PCA) has been proposed to mitigate multicollinearity in the predictor or regressor variables. Wang and Pham (2011) studied the combined effects of pollutants on daily mortality using a generalized additive model (GAM) with a robust PCA. They concluded that the relative risk (RR) estimates were more pronounced when the multivariate robust PCA technique was used. The application of PCA generally requires the data to be obtained through independent replications. However, as addressed in Zamprogno et al. (2020), if the covariates are time-correlated the PC are also autocorrelated. In this context, to handle the multicollinearity among the covariates and the autocorrelation of the PC, Souza et al. (2018) and Ispány et al. (2018) have combined the PCA technique and multivariate time series in the GAM model to quantify the impact of the pollutants on respiratory diseases. They showed that the estimation of the RR was more pronounced than indicated previously in the literature. This corroborates that statistical tools must be used with caution to quantify linear and non-linear statistical associations between response and predictor variables.

The same issues appear in longitudinal studies where repeated measurements are collected over time. General regression models with multiple sources of errors, denoted as LMMs, have been suggested in the literature to analyse this type of data, see for example Verbeke and Molenberghs (2000) for a review. In the early 90s, many researches on robust estimation methods for repeated measurements were conducted, from a theoretical and an applied point of view. Huggins (1993) applied the M-estimation approach proposed by Huber (1964) and since then, significant progress has been made towards proposing robust methodologies for analysing longitudinal data in different areas of application (Richardson and Welsh, 1995; Gill, 2000; Koller and Stahel, 2011), especially because, nowadays, there is a large amount of multivariate data, and also computational and software facilities. An alternative approach to LMM for longitudinal data is to use a functional regression model, see for instance Bauer et al. (2018); Aneiros et al. (2022). Recently, as a robust alternative technique for longitudinal data analysis, special attention has been paid to quantile regression methods see, e.g., Koenker (2004); Galvão et al. (2020); Ji and Shi (2021).

The goal of this paper is to quantify the predictor variables of air pollution, weather, the logarithm of total immunoglobulin E (IgE), among others, with FEV_1, both in asthmatic and non-asthmatic children and teenagers up to 18 years old. FEV_1 values were repeatedly measured monthly in an experiment held at the Public Health Center of Praia do Suá, in Vitória, Capital of Espírito Santo, Brazil. Although all children started the FEV_1 test in the same month, they do not

necessarily take the first measurement on the same day. Each repeated measurement is taken on average 30 days after the previous one, but with minor variations within a week at most. As the days of the measurements are recorded, it is easy to connect this information with the other environmental data as the air pollutants in (10).Biological, social, and economic level variables of each unit (child) were also considered as covariates in the modelling strategy. This data set is presented in Serpa (2019, page 31). Since the response variable is longitudinal, a particular case of the LMM discussed by Huggins (1993) is proposed, but considering the covariates as a multiple time series data with abrupt observations, such as high peaks of the pollutants. As previously mentioned, these observations produce the same effect on the estimates as AO do. Since the estimation of the covariance and correlation matrices are strongly influenced by AO, the estimation of eigenvectors and eigenvalues is also affected, and thus classical PCA is sensitive to AO.

In this context, a RLMM, combined with a RPCA, is proposed to handle simultaneously the multicollinearity among the covariates and the impact of high peaks of the air contaminants on the estimates. The classical least squares (LS) estimator and the standard PCA are also considered for comparison. To the best of our knowledge, this modelling strategy has not been suggested yet for the case of repeated measurement data with multivariate time series covariates and outliers. In addition, to empirically verify the impact of the outliers on the estimates of the fixed effects and the variances of the source errors, a finite sample size investigation is conducted under the scenario where the covariates follow linear time series models, either with or without AO.

The remaining sections of this paper are organized as follows. Section 2 discusses the LMM and its robust estimation. Section 3 presents a Monte Carlo simulation study to evaluate the efficiency of the LS method and M-estimation with Huber and Tukey loss functions, under different scenarios. Section 4 addresses the real data problem, and conclusions are in Section 5.

2 Linear mixed model

We denote by $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{im_i})^T$ the vector of measurements taken at times $t_{i1}, \ldots, t_{im_i}$ of the ith subject for $i = 1, \ldots, n$. Following Diggle (1988) and Verbeke and Molenberghs (2000, page 23), we assume that each measurement Y_{ij} of $\boldsymbol{Y}_i$, $j = 1, \ldots, m_i$, follows the LMM

$$Y_{ij} = \boldsymbol{X}_{ij}\boldsymbol{\beta} + \boldsymbol{Z}_{ij}\boldsymbol{\gamma}_i + U_{ij}, \tag{1}$$

where $\boldsymbol{X}_{ij}$ is a known $(1 \times d)$ design vector, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{d-1})^T$ is a vector of d unknown but fixed parameters to be estimated, $\boldsymbol{Z}_{ij}$ is a known $(1 \times q)$ design vector for the random effect, $\boldsymbol{\gamma_i}$ is a vector of q unobservable random effects assumed to follow a normal distribution with mean 0 and $(q \times q)$ covariance matrix $\boldsymbol{D}$, i.e., $\boldsymbol{\gamma}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D})$, and $\boldsymbol{U}_i = (U_{i1}, \ldots, U_{im_i})^T \sim \mathcal{N}(\boldsymbol{0}, \sigma_e^2\boldsymbol{\Omega}_i)$ models an experimental error. When $q = 1$ and $Z_{ij} = 1$, γ_i represents a random intercept. When $\boldsymbol{\Omega}_i$ is the identity matrix, the components of $\boldsymbol{U}_i$ are independent. However, in real applications, it is usually assumed that these components are time-correlated and follow a first order autoregressive (AR(1)) model, i.e., the (j,k)th entry of $\boldsymbol{\Omega}_i$ is given by $\nu^{|t_{ij}-t_{ik}|}/(1-\nu^2)$ where $|\nu| < 1$. Verbeke and Molenberghs (2000, page 99) suggest alternative correlation structures for the LMM.

It follows from (1) that

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\gamma}_i + \boldsymbol{U}_i$$

where $\boldsymbol{X}_i = [\boldsymbol{X}_{i1}^T, \ldots, \boldsymbol{X}_{im_i}^T]^T$ is the $(m_i \times d)$ design matrix of subject i for fixed parameters and $\boldsymbol{Z}_i = [\boldsymbol{Z}_{i1}^T, \ldots, \boldsymbol{Z}_{im_i}^T]^T$ is the $(m_i \times q)$ design matrix for the random effects of subject i. We assume that $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_n, \boldsymbol{U}_1, \ldots, \boldsymbol{U}_n$ are mutually independent. Then, $\boldsymbol{Y}_i$ is a Gaussian vector with mean $\boldsymbol{X}_i\boldsymbol{\beta}$ and covariance matrix

$$\boldsymbol{\Sigma}_i = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T + \sigma_e^2\boldsymbol{\Omega}_i. \tag{2}$$

The standardized residuals $\boldsymbol{R}_i$ is defined as

$$\boldsymbol{R}_i = \boldsymbol{\Sigma}_i^{-1/2}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}). \tag{3}$$

Let $\boldsymbol{\theta} = (\beta_0, \ldots, \beta_{d-1}, D_{11}, \ldots, D_{qq}, \sigma_e^2, \nu)^T$ be the vector parameter with dimension $p = d + q(q+1)/2 + 2$. To estimate $\boldsymbol{\theta}$, we consider the likelihood function proposed by Huggins (1993) in which we introduce the weights ω_{ij} defined below see, for example, Cantoni and Ronchetti (2001). The log-likelihood is written as

$$l(\boldsymbol{\theta}) = -\sum_{i=1}^{n}\left(\frac{\lambda}{2}\log|\boldsymbol{\Sigma}_i| + \sum_{j=1}^{m_i}\omega_{ij}\,\rho(R_{ij})\right), \tag{4}$$

where ρ is a loss function with derivative $\rho' = \psi$, $\lambda = \mathrm{E}(R\psi(R))$ with $R \sim \mathcal{N}(0,1)$, $\omega_{ij} = \sqrt{1 - h_{i^{(j)}}}$, $h_{i^{(j)}}$ is $i^{(j)}$th component of the diagonal of the projection matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ where $\boldsymbol{X} = [\boldsymbol{X}_1^T, \ldots, \boldsymbol{X}_n^T]^T$ is the $(N \times d)$ design matrix , $N = \sum_{i=1}^{n} m_i$, and

$$i^{(j)} = \begin{cases} j & \text{if } i = 1, \\ j + \sum_{k=1}^{i-1} m_k & \text{if } i \geq 2. \end{cases}$$

The role of the weights ω_{ij} and function ρ is to accommodate the outliers not only in the response variable $\boldsymbol{Y}_i$, but also in the covariate $\boldsymbol{X}_i$. When $\omega_{ij} = 1$, (4) becomes the likelihood function given in Huggins (1993). Function ρ is assumed to be convex (Richardson and Welsh, 1995) or ρ is assumed to be twice continuously differentiable and bounded (Huggins, 1993). The M-estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ maximizes (4) where $\boldsymbol{\Sigma}_i$ and R_{ij} are calculated from (2) and (3), respectively, and $\boldsymbol{D}$ is constrained to be a covariance matrix. An alternative robust and weighted method to estimate mixed models was proposed by Koller and Stahel (2011).

Under some regularity conditions (Crowder, 1986; Huggins, 1993), the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}$ may be estimated by

$$\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\theta}}) = G(\hat{\boldsymbol{\theta}})^{-1}\sum_{i=1}^{n}\left(\Lambda_i(\hat{\boldsymbol{\theta}})\Lambda_i(\hat{\boldsymbol{\theta}})^T\right)G(\hat{\boldsymbol{\theta}})^{-1}, \tag{5}$$

where $G(\boldsymbol{\theta}) = \frac{\partial^2 l(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}$ and $\Lambda_i(\boldsymbol{\theta})$ is the vector of derivatives with respect to $\boldsymbol{\theta}$ of the ith summand in (4) (Huggins, 1993; Welsh and Richardson, 1997). An estimated asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ may be obtained as

$$\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}) \approx \left(\frac{\bar{\omega}}{\varrho}\sum_{i=1}^{n}\boldsymbol{X}_i^T\hat{\boldsymbol{\Sigma}}_i^{-1}\boldsymbol{X}_i\right)^{-1}, \tag{6}$$

where $\hat{\boldsymbol{\Sigma}}_i$ is the estimate of $\boldsymbol{\Sigma}_i$ obtained by replacing the unknown parameters by their estimates in (2), $\bar{\omega} = N^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\omega_{ij}$ is a known constant, and

$$\varrho = \frac{\mathrm{E}[\psi(R)^2]}{(\mathrm{E}[\psi'(R)])^2}. \tag{7}$$

The approximation in (6) is based on the inverse Hessian matrix see, for example, Huber (1981, page 173) and Gill (2000) for the classical linear and LMM regression models, respectively.

Remark 1. There are several possible choices of functions ρ which are known in the literature. Here, we consider the classical LS function $\rho_1(r) = \frac{1}{2}r^2$, the Huber function

$$\rho_2(r) = \begin{cases} \frac{1}{2}r^2 & \text{if } |r| \leq \kappa_2, \\ \kappa_2|r| - \frac{1}{2}\kappa_2^2 & \text{if } |r| > \kappa_2, \end{cases}$$

where $\kappa_2 > 0$, and the Tukey bisquared function

$$\rho_3(r) = \begin{cases} \frac{\kappa_3^2}{6}\left(1-(1-(r/\kappa_3)^2)^3\right) & \text{if } |r| \leq \kappa_3, \\ \frac{\kappa_3^2}{6} & \text{if } |r| > \kappa_3, \end{cases}$$

where $\kappa_3 > 0$. We denote by (λ_k, ϱ_k) the value of (λ, ϱ) when $\rho = \rho_k$ for $k = 1, 2, 3$. Obviously, since $\psi_1(r) = r$, $\lambda_1 = \varrho_1 = 1$. The explicit expressions of (λ_2, ϱ_2) and (λ_3, ϱ_3) in terms of κ_2 and κ_3, respectively, are given in the Appendix. These expressions are useful to compute (4) and (6).

Remark 2. We denote by $\hat{\boldsymbol{\beta}}_1$, $\hat{\boldsymbol{\beta}}_2$ and $\hat{\boldsymbol{\beta}}_3$ the LS, Huber and Tukey estimate of $\boldsymbol{\beta}$, respectively. We deduce from (6) that the asymptotic relative efficiency of $\hat{\boldsymbol{\beta}}_2$ and $\hat{\boldsymbol{\beta}}_3$ with respect to $\hat{\boldsymbol{\beta}}_1$ is approximately equal to $1/\varrho_2$ and $1/\varrho_3$, respectively. This can be a guideline to choose κ_2 and κ_3. The standard choice is $\kappa_2 = 1.345$ and $\kappa_3 = 4.685$, which gives a relative efficiency of 95% for $\hat{\boldsymbol{\beta}}_2$ and $\hat{\boldsymbol{\beta}}_3$, respectively, see, e.g., Hampel et al. (1986, page 383), Venables and Ripley (2002, page 123) and Maronna et al. (2006, page 357).

Remark 3. Under some assumptions, the asymptotic properties of the above estimators are well-established in the literature. All estimators are consistent and asymptotically Gaussian (Huggins, 1993; Richardson and Welsh, 1995; Gill, 2000; Cantoni and Ronchetti, 2001). If the loss function is convex, then the estimator converges to the global maximum, otherwise, as in the case of the Tukey loss function, a good initial point is required to ensure that the estimator converges to the true solution (Koller, 2013, page 45).

Remark 4. For each $k = 1, 2, 3$, the maximization of $l(\boldsymbol{\theta})$ where $\rho = \rho_k$ in (4) is achieved numerically by successive iterations. The initial value $\hat{\boldsymbol{\beta}}_{0k}$ of $\boldsymbol{\beta}$ is obtained by minimizing

$$l_{0k}(\boldsymbol{\beta}) = \sum_{i=1}^{n}\sum_{j=1}^{m_i} \rho_k(Y_{ij} - \boldsymbol{X}_{ij}\boldsymbol{\beta}),$$

σ_{01}^2 is the sample variance of the residuals $(Y_{ij} - \boldsymbol{X}_{ij}\hat{\boldsymbol{\beta}}_{01})$ and σ_{0k} is the median absolute deviation of the residuals $(Y_{ij} - \boldsymbol{X}_{ij}\hat{\boldsymbol{\beta}}_{0k})$ for $k = 2, 3$. The initial values of D is D_{0k} where $D_{0k,ii} = \sigma_{0k}^2/(q+1)$ for $i = 1, \ldots, q$ and $D_{0,ij} = 0$ if $i \neq j$, the initial value of σ_e^2 is $\sigma_{0k,e}^2 = \sigma_{0k}^2/(q+1)$ and the initial value of ν is ν_0 where ν_0 is the same for the three loss functions and is sampled from the uniform distribution on the interval $(0, 1)$. From the computational aspect, $l(\boldsymbol{\theta})$ is maximized using the L-BFGS-B method which is a quasi-Newton procedure allowing box constraints, i.e., each variable is restricted inside lower and upper bounds. This ensures that the variances are positive. The gradient of $l(\boldsymbol{\theta})$ is approximated by finite differences. Our codes are developed in R language (R Core Team, 2018) however, we also use the package RcppArmadillo to accelerate the calculations (Eddelbuettel and Sanderson, 2014). This allows to work in a big data context (Byrd et al., 1995). Our R codes, a documentation, and simulated examples are available on GitHub[1]. Other researchers can analyze the method or apply it to their own longitudinal databases if they have analogous problems.

3 Monte Carlo simulations

In this section, a simulation study with finite sample sizes is conducted to verify the performance of the estimators in LMM when the covariates are cross-correlated time series with AO. In the context of time series data, AO are particularly dangerous since they have a strong impact on sample estimates such as sample mean and sample autocorrelations. Additionally, as previously mentioned,

[1]For the printed version, see https://github.com/iandanilevicz/RLMM

the pollution series may present observations with high levels of pollutant concentrations which may produce sample densities with heavy tails, and these observations provoke the same effect on the estimates as AO do. We consider the following particular case of model (1),

$$Y_{ij} = \boldsymbol{X}_{ij}\boldsymbol{\beta} + \gamma_i + U_{ij}, \tag{8}$$

where the measurements are supposed to be taken at times $t_{i,j} = j$, $\boldsymbol{X}_{ij} = (X_{ij}^{(0)}, X_{ij}^{(1)}, X_{ij}^{(2)})$ with $X_{ij}^{(0)} = 1$ and $(X_{ij}^{(1)}, X_{ij}^{(2)})^T$ follows the first order vector autoregressive (VAR(1)) model

$$\begin{bmatrix} X_{i,j}^{(1)} \\ X_{i,j}^{(2)} \end{bmatrix} = \begin{bmatrix} \phi_{1,1} & 0 \\ 0 & \phi_{2,2} \end{bmatrix} \begin{bmatrix} X_{i,j-1}^{(1)} \\ X_{i,j-1}^{(2)} \end{bmatrix} + \begin{bmatrix} W_{i,j}^{(1)} \\ W_{i,j}^{(2)} \end{bmatrix}, \qquad \begin{bmatrix} W_{i,j}^{(1)} \\ W_{i,j}^{(2)} \end{bmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \xi \\ \xi & 1 \end{pmatrix} \right], \tag{9}$$

where $\phi_{1,1} = 0, 0.7$, $\phi_{2,2} = 0, 0.6$ and $\xi = 0, 0.5, 0.9$ see, for example, Brockwell and Davis (1991, pages 417-421). In (8), $\boldsymbol{\beta} = (2, -0.5, 1)^T$, $\gamma_i \sim \mathcal{N}(0, 1)$ is a scalar random intercept, and $U_{ij} \sim \mathcal{N}(0, 1)$. Therefore, the parameter vector is $\boldsymbol{\theta} = (\beta_0, \beta_1, \beta_2, D, \sigma_e^2)^T$ with true value $\boldsymbol{\theta}_0 = (2, -0.5, 1, 1, 1)^T$. We take $n = 20, 40, 80$, $m_i = 5, 50$, and the weights $\omega_{ij} = 1$ and $\omega_{ij} = (1 - h_{i(j)})^{\frac{1}{2}}$ in (4). As previously mentioned, $\kappa_2 = 1.345$ and $\kappa_3 = 4.685$.

To evaluate the robustness of each estimation method, we simulate $\boldsymbol{X}_{ij}$ with (9) and Y_{ij} with (8) where $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Then, when we estimate $\boldsymbol{\theta}$, we replace in (3) $X_{ij}^{(2)}$ by the corrupted version $X_{ij}^{(2)*}$ defined by

$$X_{ij}^{(2)*} = X_{ij}^{(2)} + \mu B_{ij} K_{ij},$$

where B_{ij} is a Bernoulli random variable (RV) with $\mathbb{P}(B_{ij} = 1) = \delta$ where $\delta = 0, 0.01, \ldots, 0.05$ and K_{ij} follows a Student's t-distribution with 3 degrees of freedom. The RVs B_{ij}, K_{ij} for $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$ are mutually independent, and $\mu = 1, 2$ is the magnitude of AO. To compare the estimators, we calculate by Monte Carlo simulations for each component $\hat{\theta}_k$, $k = 1, \ldots, 5$ of the parameter estimate $\hat{\boldsymbol{\theta}}$, the sample mean (SM) $\widehat{\mu}(\hat{\theta}_k)$, mean squared error (MSE) $\widehat{\sigma}^2(\hat{\theta}_k)$, and coverage probability (CP) with 95% of confidence $\widehat{v}(\hat{\theta}_k)$. Let $M = 1000$ be the number of replications and $\hat{\theta}_{k,m}$ be the estimate of θ_k obtained in the mth experiment for $m = 1, \ldots, M$. We have

$$\widehat{\mu}(\hat{\theta}_k) = \frac{1}{M}\sum_{m=1}^{M} \hat{\theta}_{k,m}, \qquad \widehat{\sigma}^2(\hat{\theta}_k) = \frac{1}{M}\sum_{m=1}^{M} (\theta_k - \hat{\theta}_{k,m})^2, \qquad \widehat{v}(\hat{\theta}_k) = \frac{1}{M}\sum_{m=1}^{M} w_{k,m},$$

where $w_{k,m} = 1$ if $\theta_k \in [\hat{\theta}_{k,m} - 1.96\hat{\sigma}(\hat{\theta}_k), \hat{\theta}_{k,m} + 1.96\hat{\sigma}(\hat{\theta}_k)]$ and 0 otherwise. As the nominal level of the CP is 0.95, we say that an estimator is good when its CP value lies in the range 0.95 ± 0.02, and is acceptable when its CP lies in the range 0.95 ± 0.04.

In our simulations, a slight improvement of the performance when using $\omega_{ij} = (1 - h_{i(j)})^{\frac{1}{2}}$ instead of $\omega_{ij} = 1$ in (4) was observed only for the smallest sample size $(n, m_i) = (20, 5)$ and $\xi = 0.9$, while in all other cases both log-likelihood functions led to similar estimates. This shows that in the scenarios considered in our empirical study, the Huber and Tukey loss functions are sufficient to handle some potential outlier effects on the SM, MSE and CP. Based on this empirical evidence, the results presented here are related to $n = 40$, $m_i = 5, 50$ and $\omega_{ij} = 1$.

Figures 1 to 3 plot for $\xi = 0, 0.5, 0.9$, respectively, the SM and CP of each parameter estimate as a function of the percentage of contamination δ using the three estimation methods when $m_i = 5$, $\mu = 1$ and $(\phi_{1,1}, \phi_{2,2}) = (0, 0)$ in (9), which implies that $(X_{ij}^{(1)}, X_{ij}^{(2)})^T$ is not time-correlated. Table 1 display the SM, MSE and CP of each parameter estimate using the three estimation methods when $\xi = 0, 0.5$, $\delta = 0, 0.02$, $m_i = 5$, $\mu = 2$ and $(\phi_{1,1}, \phi_{2,2}) = (0, 0)$ in (9). Tables 2 and 3 present the same results when $\xi = 0, 0.5$, $\delta = 0, 0.002$, $m_i = 50$, $\mu = 2$ and $(\phi_{1,1}, \phi_{2,2}) = (0, 0)$, $(\phi_{1,1}, \phi_{2,2}) = (0.7, 0.6)$, respectively.

In Figure 1, $\xi = 0$ and therefore $X_{ij}^{(1)}$ and $X_{ij}^{(2)}$ are uncorrelated. $\widehat{\mu}(\hat{\beta}_0)$ and $\widehat{\mu}(\hat{\beta}_1)$ fluctuate randomly around the true values. $\widehat{\mu}(\hat{\beta}_2)$ decreases and $\widehat{\mu}(\hat{\sigma}_e^2)$ increases as δ increases, and $\hat{D}$ is slightly underestimated. About CP, the Huber estimation method provides a good estimate of β_0, LS and Tukey provide an acceptable estimate; the Huber estimation method provides a good estimate of β_1 except for $\delta = 0.04$, LS provides a good estimate except for $\delta = 0.05$ and Tukey provides an acceptable estimate; Tukey provides an acceptable estimate of β_2 if $\delta \leq 0.03$, Huber provides an acceptable estimate if $\delta \leq 0.02$, and LS provides an acceptable estimate only if $\delta = 0$; the Tukey estimation method provides a good estimate of D, LS and Huber provide acceptable estimates; Huber estimation method provides a good estimate of σ_e^2, Tukey provides acceptable estimates, and LS is acceptable if $\delta \leq 0.03$. Therefore, the three methods provide acceptable estimates of β_0 and β_1 and Tukey estimate is the most resilient to outliers in β_2.

In Figure 2, $\xi = 0.5$ and therefore the correlation between $X_{ij}^{(1)}$ and $X_{ij}^{(2)}$ is moderate. Again $\widehat{\mu}(\hat{\beta}_0)$ fluctuates randomly around the true value, $\widehat{\mu}(\hat{\sigma}_e^2)$ slowly increases and D is slightly underestimated. However, β_1 is overestimated and β_2 is underestimated and the bias increases as δ increases. About CP, we have a similar situation to $\xi = 0$ for β_0 estimations; Huber and Tukey methods provide acceptable estimates of β_1, and LS also if $\delta \leq 0.02$; Tukey provides an acceptable estimate of β_2 if $\delta \leq 0.03$, Huber provides an acceptable estimate if $\delta \leq 0.01$, and LS provides an acceptable estimate only if $\delta = 0$; the three methods are acceptable to estimate D; Huber and Tukey are acceptable to estimate σ_e^2, but LS is acceptable if $\delta \leq 0.03$. Again, the Tukey estimate method is the most resilient to outliers in β_2.

In Figure 3, $\xi = 0.9$ and then the correlation between $X_{ij}^{(1)}$ and $X_{ij}^{(2)}$ is strong. Concerning the SM, β_0, σ_e^2 and D display similar behaviors as previously. Again, β_1 is overestimated and β_2 is underestimated and the bias increases severely as δ increases. Regarding CP, we have a similar situation to $\xi = 0$ for β_0 estimations; Tukey provides an acceptable estimate of β_1 if $\delta \leq 0.02$, Huber provides an acceptable estimate if $\delta \leq 0.01$, and LS provides an acceptable estimate only if $\delta = 0$; Huber and Tukey provide an acceptable estimate of β_2 if $\delta \leq 0.01$, and LS provides an acceptable estimate only if $\delta = 0$; the three methods provide acceptable estimates of D and σ_e^2 to any δ.

These simulations show that the use of one of the robust methods is strongly recommended when the presence of outliers is suspected in the data. When the covariates present moderate, low or no correlation and the percentage of outliers is less than 2%, Huber estimates are the best choice. If the covariates are strongly correlated or the percentage of outliers is more than 2%, then Tukey estimates are the best choice. Note that if the contamination level of a covariate is 4% or higher, then this covariate should not be included in the model since even Tukey estimation method does not provide an acceptable CP. Also moderate and strong correlations should be avoid when possible.

Table 1 reports the SM, MSE and CP related to the simulations shown in Figures 1 and 2 when $\delta = 0$ (uncontaminated data). This table also displays the performance of the estimators with outlier's magnitude $\mu = 2$ and $\delta = 0.02$. These sample quantities corroborate the previous analysis of the performance of the estimators based on the plots. Now, the increase of the magnitude μ from 1 to 2 affects the preciseness of the estimators. For example, when $\xi = 0$, $\delta = 0.02$ and $\mu = 1$, the estimates of the $\widehat{\sigma}^2(\hat{\beta}_2)$ were 0.0127, 0.0094 and 0.0087 for LS, Huber and Tuckey methods, respectively. Thus, comparing these values to the ones in Table 1 for $\xi = 0$, $\delta = 0.02$ and $\mu = 2$, we see that $\widehat{\sigma}^2(\hat{\beta}_2)$ is almost multiplied by 4 for the LS method. In the same way, the corresponding CP reduces substantially from 89.6% to 59.1%. Similar conclusions are drawn for the other cases. In general, the Tukey loss-function displayed more resistance against the increase of the outlier's magnitude.

Table 2 reports the simulation results when the size of the repeated measurements is $m_i = 50$. For a comparison purpose with the previous scenario, the expected number of outliers was fixed to 4 by choosing $\delta = 0.002$. Similar conclusions to the previous case can be derived. The SM, MSE

and CP are strongly affected by the outliers, especially when $\xi = 0.5$ and $\delta = 0.002$. The MSE estimates decrease as m_i increases, which corroborates the asymptotic result that the estimators are consistent.

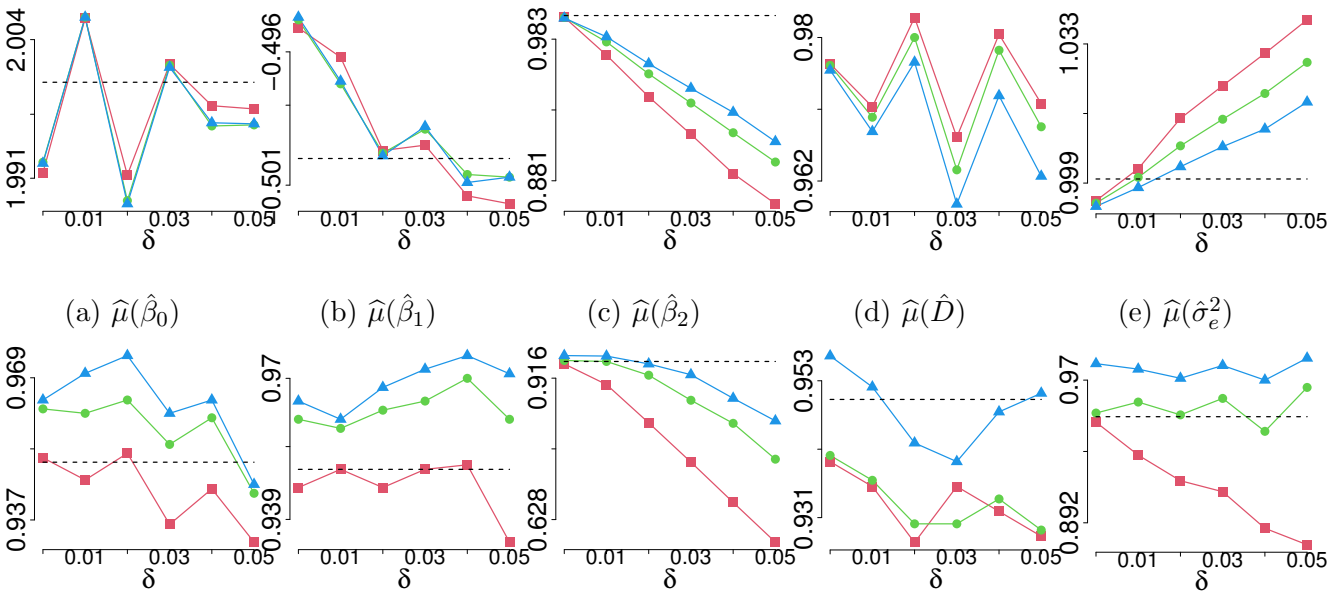
Table 3 displays the empirical quantities when $(\phi_{1,1}, \phi_{2,2}) = (0.7, 0.6)$ in (9), which implies that $(X_{ij}^{(1)}, X_{ij}^{(2)})^T$ is also time-correlated. Although the SM and MSE behave similarly to the case in Table 2, the CP present for the regression coefficients β_1 and β_2 much smaller relative frequency than the previous cases and, again, the Tukey method provides the most resistant estimator. The degradation of the CP is more important when $\xi = 0.5$, i.e. when $X_{ij}^{(1)}$ and $X_{ij}^{(2)}$ are correlated.

We proceed with an additional simulation to verify the computational time. Using model (8) with $\phi_{1,1} = 0$, $\phi_{2,2} = 0$, $\mu = 1$, $\xi = 0$, $\delta = 0.05$ and $(n, m_i) = (40, 5), (80, 5), (40, 50)$. As this model has $p = 5$, and the sample size is equal to 200, 400 and 2000, then the degrees of freedom are 195, 395 and 1995. All the results are obtained using a desktop processor Intel Core i7-6700 CPU 3.40GHz×8. Table 4 displays the time of the methods with $M = 1000$. The three techniques present a similar computational time, slightly over one second for $(n, m_i) = (40, 5), (80, 5)$ and slightly over ten seconds for $(n, m_i) = (40, 50)$.

This empirical study shows that the multivariate time series structure of the covariates, as well as the presence of atypical observations in the data, are phenomena that can not be ignored in the model strategy, otherwise, the conclusions can be totally erroneous and lead to severe consequences in terms of statistical inference.

Table 1: SM, MSE and CP with 95% of confidence when $\xi = 0, 0.5$, $\delta = 0, 0.02$, $(n, m_i) = (40, 5)$, $(\phi_{1,1}, \phi_{2,2}) = (0, 0)$, $\mu = 2$. LMM estimated by LS, Huber (H) and Tukey (T) with $\kappa_2 = 1.345$ and $\kappa_3 = 4.685$, respectively.

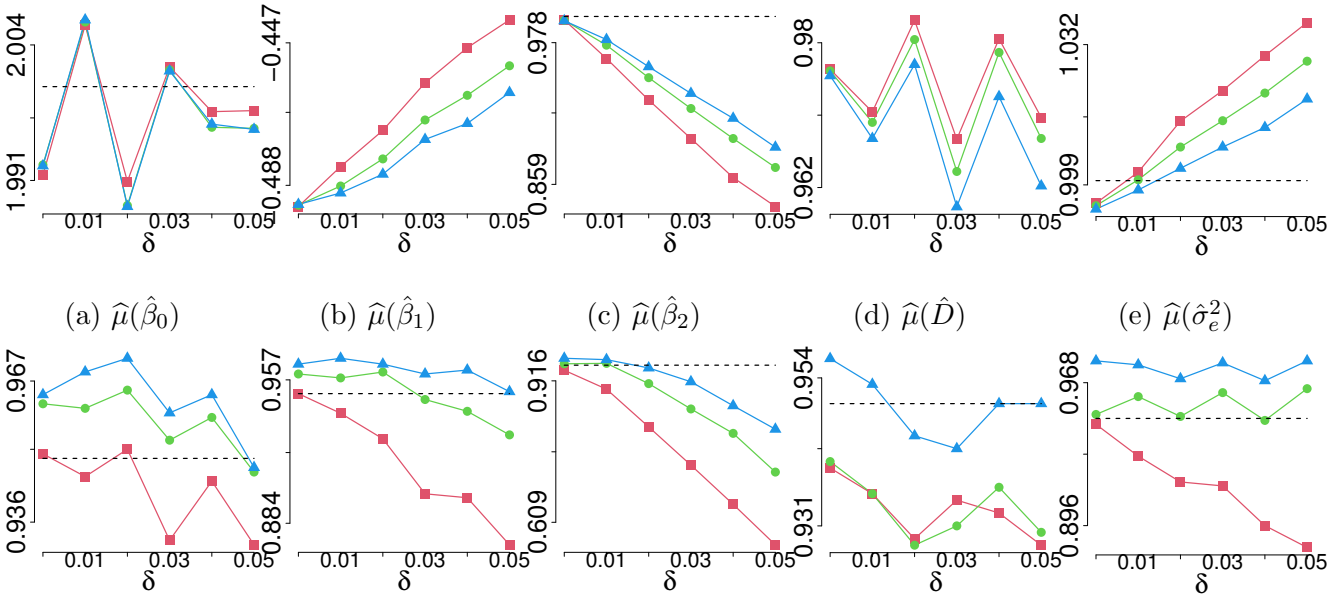
Parameters	SM			MSE			CP		
$\xi = 0, \delta = 0$	LS	H	T	LS	H	T	LS	H	T
β_0	2.002	2.002	2.002	0.0289	0.0306	0.0307	0.939	0.953	0.962
β_1	-0.497	-0.497	-0.497	0.0057	0.0061	0.0061	0.958	0.965	0.972
β_2	0.998	0.998	0.998	0.0067	0.0070	0.0070	0.936	0.955	0.965
D	0.991	0.990	0.990	0.0031	0.0034	0.0035	0.940	0.939	0.972
σ_e^2	0.980	0.980	0.980	0.0188	0.0193	0.0193	0.942	0.944	0.960
$\xi = 0, \delta = 0.02$	LS	H	T	LS	H	T	LS	H	T
β_0	2.006	2.005	2.004	0.0302	0.0319	0.0324	0.947	0.963	0.972
β_1	-0.499	-0.499	-0.500	0.0070	0.0069	0.0069	0.943	0.965	0.971
β_2	0.860	0.920	0.952	0.0467	0.0186	0.0108	0.591	0.815	0.923
D	1.059	1.036	1.016	0.0113	0.0064	0.0043	0.788	0.932	0.983
σ_e^2	0.974	0.968	0.957	0.0192	0.0197	0.0210	0.927	0.927	0.933
$\xi = 0.5, \delta = 0$	LS	H	T	LS	H	T	LS	H	T
β_0	2.002	2.002	2.001	0.0328	0.0346	0.0347	0.926	0.935	0.946
β_1	-0.503	-0.504	-0.504	0.0084	0.0088	0.0088	0.943	0.958	0.969
β_2	1.004	1.003	1.003	0.0086	0.0092	0.0092	0.940	0.957	0.963
D	0.995	0.995	0.994	0.0032	0.0035	0.0036	0.937	0.938	0.972
σ_e^2	0.975	0.975	0.974	0.0195	0.0202	0.0205	0.925	0.933	0.949
$\xi = 0.5, \delta = 0.02$	LS	H	T	LS	H	T	LS	H	T
β_0	1.999	2.000	1.999	0.0291	0.0305	0.0313	0.942	0.957	0.960
β_1	-0.419	-0.449	-0.468	0.0229	0.0146	0.0113	0.802	0.902	0.951
β_2	0.830	0.891	0.930	0.0642	0.0320	0.0197	0.567	0.772	0.892
D	1.049	1.032	1.014	0.0090	0.0061	0.0045	0.819	0.944	0.981
σ_e^2	0.977	0.970	0.957	0.0192	0.0200	0.0214	0.933	0.932	0.948



(a) $\widehat{\mu}(\hat{\beta}_0)$ (b) $\widehat{\mu}(\hat{\beta}_1)$ (c) $\widehat{\mu}(\hat{\beta}_2)$ (d) $\widehat{\mu}(\hat{D})$ (e) $\widehat{\mu}(\hat{\sigma}_e^2)$

(f) $\widehat{v}(\hat{\beta}_0)$ (g) $\widehat{v}(\hat{\beta}_1)$ (h) $\widehat{v}(\hat{\beta}_2)$ (i) $\widehat{v}(\hat{D})$ (j) $\widehat{v}(\hat{\sigma}_e^2)$

Figure 1: SM and CP with 95% of confidence when $\xi = 0$, $(n, m_i) = (40, 5)$, $(\phi_{1,1}, \phi_{2,2}) = (0, 0)$, $\mu = 1$. LS estimates (red square), Huber estimates with $\kappa_2 = 1.345$ (green circle) and Tukey estimates with $\kappa_3 = 4.685$ (blue triangle).



(a) $\widehat{\mu}(\hat{\beta}_0)$ (b) $\widehat{\mu}(\hat{\beta}_1)$ (c) $\widehat{\mu}(\hat{\beta}_2)$ (d) $\widehat{\mu}(\hat{D})$ (e) $\widehat{\mu}(\hat{\sigma}_e^2)$

(f) $\widehat{v}(\hat{\beta}_0)$ (g) $\widehat{v}(\hat{\beta}_1)$ (h) $\widehat{v}(\hat{\beta}_2)$ (i) $\widehat{v}(\hat{D})$ (j) $\widehat{v}(\hat{\sigma}_e^2)$

Figure 2: SM and CP with 95% of confidence when $\xi = 0.5$, $(n, m_i) = (40, 5)$, $(\phi_{1,1}, \phi_{2,2}) = (0, 0)$, $\mu = 1$. LS estimates (red square), Huber estimates with $\kappa_2 = 1.345$ (green circle) and Tukey estimates with $\kappa_3 = 4.685$ (blue triangle).

Table 2: SM, MSE and CP with 95% of confidence when $\xi = 0, 0.5$, $\delta = 0, 0.002$, $(n, m_i) = (40, 50)$, $(\phi_{1,1}, \phi_{2,2}) = (0, 0)$, $\mu = 2$. LMM estimated by LS, Huber (H) and Tukey (T) with $\kappa_2 = 1.345$ and $\kappa_3 = 4.685$, respectively.

Parameters	SM			MSE			CP		
$\xi = 0, \delta = 0$	LS	H	T	LS	H	T	LS	H	T
β_0	1.999	1.999	2.000	0.0256	0.0270	0.0269	0.934	0.957	0.963
β_1	-0.501	-0.502	-0.502	0.0005	0.0006	0.0006	0.948	0.955	0.966
β_2	1.000	0.999	0.999	0.0005	0.0005	0.0005	0.937	0.956	0.969
D	1.000	1.000	1.000	0.0003	0.0003	0.0003	0.950	0.963	0.983
σ_e^2	0.976	0.977	0.977	0.0130	0.0132	0.0132	0.927	0.928	0.950
$\xi = 0, \delta = 0.002$	LS	H	T	LS	H	T	LS	H	T
β_0	2.000	2.000	2.000	0.0255	0.0269	0.0270	0.948	0.970	0.974
β_1	-0.499	-0.500	-0.500	0.0005	0.0005	0.0005	0.957	0.971	0.977
β_2	0.980	0.994	0.997	0.0027	0.0006	0.0006	0.819	0.966	0.974
D	1.009	1.004	1.001	0.0007	0.0003	0.0003	0.870	0.971	0.981
σ_e^2	0.983	0.982	0.981	0.0123	0.0128	0.0128	0.932	0.931	0.949
$\xi = 0.5, \delta = 0$	LS	H	T	LS	H	T	LS	H	T
β_0	1.993	1.993	1.993	0.0250	0.0262	0.0262	0.947	0.956	0.966
β_1	-0.501	-0.500	-0.500	0.0008	0.0008	0.0008	0.936	0.955	0.966
β_2	1.001	1.001	1.001	0.0007	0.0008	0.0008	0.948	0.961	0.971
D	1.000	1.000	1.000	0.0003	0.0003	0.0003	0.951	0.958	0.982
σ_e^2	0.979	0.979	0.979	0.0128	0.0131	0.0131	0.929	0.927	0.943
$\xi = 0.5, \delta = 0.002$	LS	H	T	LS	H	T	LS	H	T
β_0	2.002	2.003	2.003	0.0246	0.0254	0.0254	0.935	0.957	0.966
β_1	-0.488	-0.497	-0.499	0.0014	0.0008	0.0007	0.899	0.962	0.972
β_2	0.975	0.993	0.997	0.0035	0.0009	0.0008	0.800	0.943	0.966
D	1.009	1.003	1.001	0.0006	0.0003	0.0003	0.880	0.959	0.987
σ_e^2	0.988	0.987	0.986	0.0138	0.0142	0.0143	0.922	0.925	0.945

4 Real data analysis

In this section, we apply the model and estimation methods discussed before in order to quantify the statistical association between FEV_1, in both asthmatic and non-asthmatic children and teenagers, aged 7 to 18 years, with PM_{10}, $PM_{2.5}$ and sulfur dioxide (SO_2) pollutants, Hmd, temperature in degrees Celsius (Tmp), IgE and passive smoking (PS). The pollutants and the weather variables were measured at the air quality automatic monitoring network (AQAMN) of the great Vitória region (GVR), which is a densely populated region with approximately 1,900,000 inhabitants in an area of 2319 km^2 located on the east coast in the State of Espírito Santo (ES), Brazil (latitude 20°19S, longitude 40°20W) and has a humid tropical climate, with average temperatures ranging from 24°C to 30°C. The GVR is a port region and industrialized area. The AQAMN of the GVR consists of eight monitoring stations: two in Serra (Laranjeiras and Carapina), three in Vitória city (Jardim Camburi, Praia do Suá and Vix-Centro), two in Vila Velha (VV-Centro and Ibes), and one in Cariacica. The response variable FEV_1 was measured monthly, from July to December 2017, at the Public Health Center of Praia do Suá, Vitória, ES, Brazil. All children live and study in the same neighborhood, which is close to the air monitoring station Praia do Suá. For this reason, although atmospheric information from several meteorological stations in the GVR are available, only the information from one station is used. All environmental covariates refer to the daily averages of the day before the spirometry test. This one-day lag for the environmental covariates is

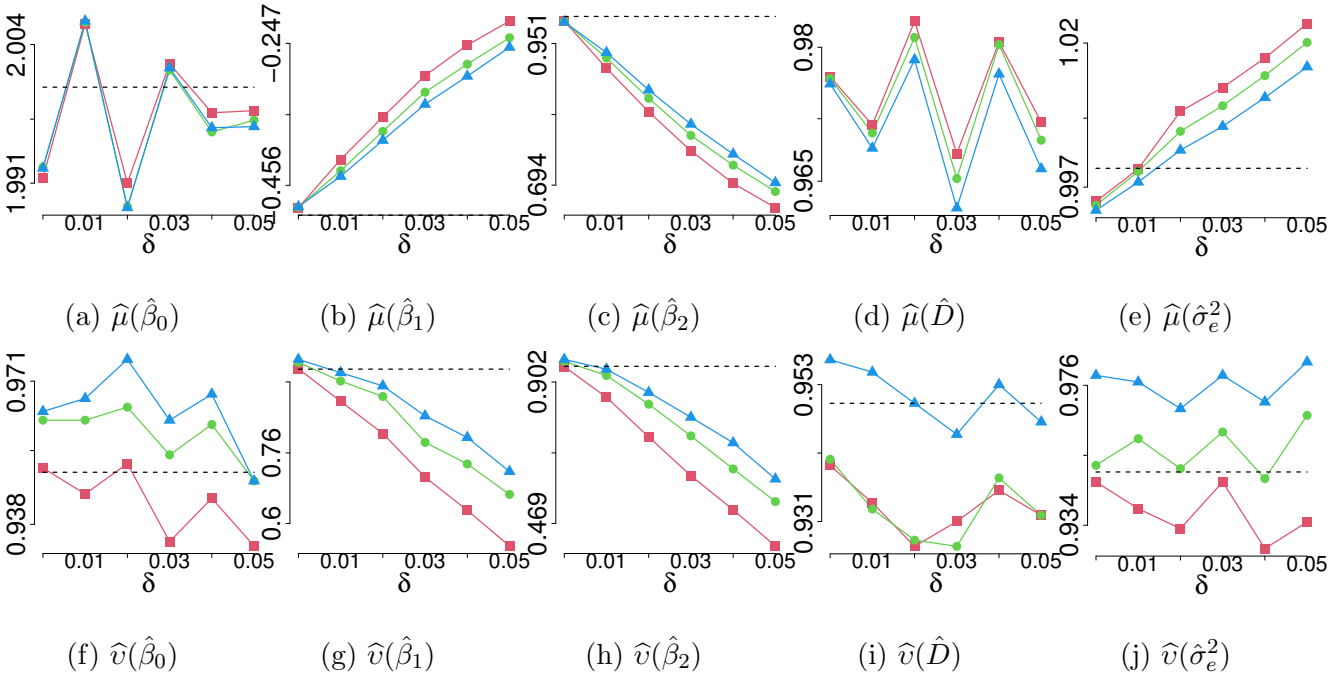
Table 3: SM, MSE and CP with 95% of confidence when $\xi = 0, 0.5$, $\delta = 0, 0.002$, $(n, m_i) = (40, 50)$, $(\phi_{1,1}, \phi_{2,2}) = (0.7, 0.6)$, $\mu = 2$. LMM estimated by LS, Huber (H) and Tukey (T) with $\kappa_2 = 1.345$ and $\kappa_3 = 4.685$, respectively.

Parameters	SM			MSE			CP		
$\xi = 0, \delta = 0$	LS	H	T	LS	H	T	LS	H	T
β_0	1.999	1.999	2.000	0.0252	0.0263	0.0261	0.939	0.962	0.969
β_1	-0.501	-0.501	-0.501	0.0004	0.0004	0.0004	0.907	0.928	0.938
β_2	0.999	0.999	0.999	0.0004	0.0004	0.0004	0.916	0.938	0.947
D	1.000	0.999	0.999	0.0002	0.0003	0.0003	0.965	0.962	0.987
σ_e^2	0.976	0.975	0.975	0.0133	0.0137	0.0137	0.924	0.923	0.935
$\xi = 0, \delta = 0.002$	LS	H	T	LS	H	T	LS	H	T
β_0	2.005	2.004	2.005	0.0262	0.0268	0.0269	0.938	0.961	0.965
β_1	-0.500	-0.500	-0.500	0.0004	0.0004	0.0004	0.919	0.939	0.945
β_2	0.989	0.997	0.999	0.0013	0.0005	0.0004	0.811	0.916	0.942
D	1.010	1.004	1.001	0.0008	0.0003	0.0003	0.876	0.968	0.985
σ_e^2	0.980	0.979	0.978	0.0127	0.0130	0.0131	0.938	0.934	0.952
$\xi = 0.5, \delta = 0$	LS	H	T	LS	H	T	LS	H	T
β_0	1.999	1.999	1.999	0.0252	0.0263	0.0261	0.939	0.962	0.969
β_1	-0.501	-0.501	-0.501	0.0006	0.0006	0.0006	0.899	0.920	0.927
β_2	1.000	1.000	1.000	0.0005	0.0006	0.0006	0.908	0.932	0.939
D	1.000	0.999	0.999	0.0002	0.0003	0.0003	0.965	0.962	0.986
σ_e^2	0.976	0.975	0.975	0.0133	0.0137	0.0137	0.924	0.923	0.935
$\xi = 0.5, \delta = 0.002$	LS	H	T	LS	H	T	LS	H	T
β_0	2.005	2.004	2.005	0.0262	0.0268	0.0269	0.938	0.961	0.965
β_1	-0.493	-0.498	-0.499	0.0009	0.0006	0.0006	0.871	0.935	0.940
β_2	0.985	0.996	0.998	0.0019	0.0007	0.0006	0.776	0.905	0.924
D	1.010	1.004	1.001	0.0007	0.0003	0.0003	0.877	0.968	0.985
σ_e^2	0.980	0.979	0.978	0.0127	0.0131	0.0131	0.939	0.934	0.952

motivated by the following review of literature Strickland et al. (2010); Rice et al. (2013); Qu et al. (2018). Furthermore, a sensitivity analysis with two-days lag is realized, and the results are very similar. Other noise covariates, such as body mass index (BMI), age and gender were dropped from the final model since, in the first steps of the model strategy, these variables were not statistically significant.

The longitudinal study involved 82 children, and for each child, 6 measurements of FEV_1 were obtained. The age range of the study is justified by the fact that children of these ages are more susceptible to the effects of air pollution and have a higher prevalence of respiratory diseases. This data set is presented in Serpa (2019) and its main characteristics are summarized in Table 5 with the mean, standard deviation (s.d.), minimum, and maximum of each variable. Furthermore, the dichotomous variable PS shows 26.5% of passive smokers and 73.5% of non-passive smokers.

Before analysing the statistical association between pollution and FEV_1, we verify if the high peaks of concentrations present in the data have some impact on the correlation matrix of the environmental covariates. Let $\boldsymbol{M}_i$ be the (6×5) matrix of the 6 measurements of the 5 environmental covariates, Hmd, Tmp, SO_2, PM_{10} and $PM_{2.5}$, corresponding to the ith child, and $\boldsymbol{M} = [\boldsymbol{M}_1^T, \boldsymbol{M}_2^T, \ldots, \boldsymbol{M}_{82}^T]^T$ be the (492×5) matrix of all environmental covariates. Table 6 displays respectively below and above the diagonal, the classical and robust correlations of the columns of $\boldsymbol{M}$. Shevlyakov and Smirnov (2011) propose a robust correlation estimation based on S_n, a robust variance defined by $S_n = c\,\mathrm{med}_i\{\mathrm{med}_j\,|r_i - r_j|\}$, where $\boldsymbol{r} = (r_1, \ldots, r_n)^T$, $r_i \in \mathbb{R}$ and for each



(a) $\widehat{\mu}(\hat{\beta}_0)$ (b) $\widehat{\mu}(\hat{\beta}_1)$ (c) $\widehat{\mu}(\hat{\beta}_2)$ (d) $\widehat{\mu}(\hat{D})$ (e) $\widehat{\mu}(\hat{\sigma}_e^2)$

(f) $\widehat{v}(\hat{\beta}_0)$ (g) $\widehat{v}(\hat{\beta}_1)$ (h) $\widehat{v}(\hat{\beta}_2)$ (i) $\widehat{v}(\hat{D})$ (j) $\widehat{v}(\hat{\sigma}_e^2)$

Figure 3: SM and CP with 95% of confidence when $\xi = 0.9$, $(n, m_i) = (40, 5)$, $(\phi_{1,1}, \phi_{2,2}) = (0, 0)$, $\mu = 1$. LS estimates (red square), Huber estimates with $\kappa_2 = 1.345$ (green circle) and Tukey estimates with $\kappa_3 = 4.685$ (blue triangle).

Table 4: Average time (in seconds) of each method when $(n, m_i) = (40, 5), (80, 5), (40, 50)$, $\phi_{1,1} = 0$, $\phi_{2,2} = 0$, $\mu = 1$, $\delta = 0.05$ and $\xi = 0$

	Methods		
(n, m_i)	LS	Huber	Tukey
$(40, 5)$	1.078	1.061	1.073
$(80, 5)$	1.135	1.123	1.140
$(40, 50)$	11.799	12.920	13.688

Table 5: Exploratory data analysis

Variable	mean	s.d.	min	max
FEV_1	93.63	15.83	35.00	141.00
Hmd	83.44	6.48	70.89	95.46
Tmp	23.13	2.03	19.47	26.77
PM_{10}	18.99	4.99	10.25	29.54
$PM_{2.5}$	10.36	2.68	6.75	19.00
SO_2	7.13	5.17	2.27	26.31
IgE	5.01	1.62	0.76	8.52

i we compute the median of $\{|r_i - r_j|; j = 1, \ldots, n\}$. This builds n numbers, the median of which provides the final value of S_n. The constant c is by default equal to 1.1926 (Rousseeuw and Croux, 1993).

As can be seen, the correlations are very different, and in general, the robust ones are numerically larger. This may indicate that there are high levels of concentration that are causing the same effect as AO. To clarify this issue, the original data were modified by replacing the 10% smallest and greatest peaks of the pollutants by their respective means. The sample correlations of the modified data are shown in Table 7, and differently from the results in Table 6, both sample correlation functions display close values. Thus, this empirical result illustrates the effect of the large peaks of the concentrations on the sample correlations and the occurrence of the multicollinearity phenomenon.

Table 6: Correlations between environmental covariates.

		Robust correlation				
		Hmd	Tmp	SO_2	PM_{10}	$PM_{2.5}$
	Hmd		-0.572	-0.324	-0.044	0.273
	Tmp	-0.469		0.378	0.231	0.594
Classical	SO_2	0.033	0.433		0.601	0.101
correlation	PM_{10}	-0.245	0.149	0.417		0.526
	$PM_{2.5}$	-0.074	0.389	0.264	0.648	

Table 7: Correlation between environmental covariates after replacing the 10% smallest and greatest values by the average of each covariate.

		Robust correlation				
		Hmd	Tmp	SO_2	PM_{10}	$PM_{2.5}$
	Hmd		-0.445	0.094	0.451	0.231
	Tmp	-0.286		0.045	-0.153	0.447
Classical	SO_2	0.131	0.053		0.487	-0.054
correlation	PM_{10}	0.398	0.090	0.448		0.218
	$PM_{2.5}$	0.237	0.385	-0.048	0.268	

To minimize the effects of outliers, we use a RPCA method before the regression analysis in the LMM in order to quantify the statistical association between FEV_1 and the covariates. Table 8 displays the loadings $\boldsymbol{L}^{(k)}$ for $k = 1, 2, 3$ of the three first PC of the classical PCA (Venables and Ripley, 2002, page 304), and RPCA (Hubert et al., 2005) of $\boldsymbol{M}$. The cumulative proportion of

the total variation (CPV) of the classical PCA is smaller than the RPCA, suggesting a number of regressors larger than three in the model to achieve a CPV larger than 90%. The CPV of the first k principal component is

$$\text{CPV}(k) = \sum_{i=1}^{k} \boldsymbol{L}^{(i)^T} \boldsymbol{V} \boldsymbol{L}^{(i)} / \sum_{i=1}^{5} \boldsymbol{L}^{(i)^T} \boldsymbol{V} \boldsymbol{L}^{(i)},$$

where $\boldsymbol{V} = \text{Var}(\boldsymbol{M})$ (Johnson and Wichern, 2002, pages 440-444). In the case of the classical PCA, the first PC is essentially an average of all variables except Hmd, whereas the most important contributions to the second and third PC are from Hmd and SO_2, respectively. The situation is very different in the RPCA where the loadings values differ from classical PCA and each PC is mainly explained by only one variable: the main contributions to the first, second and third PC are from Hmd, PM_{10} and $PM_{2.5}$, respectively. Note that the variables PM_{10} and SO_2 are highly correlated (see Table 6).

Table 8: Loadings of the three first PC of the classical PCA and RPCA, and CPV.

	PCA			RPCA		
	$\boldsymbol{L}^{(1)}$	$\boldsymbol{L}^{(2)}$	$\boldsymbol{L}^{(3)}$	$\boldsymbol{L}^{(1)}$	$\boldsymbol{L}^{(2)}$	$\boldsymbol{L}^{(3)}$
Hmd	-0.436	**-0.795**	0.297	**-0.972**	0.011	-0.048
Tmp	**0.704**	0.481	0.373	0.184	-0.116	0.358
SO_2	**0.639**	-0.279	**0.635**	0.134	0.194	-0.469
PM_{10}	**0.767**	-0.344	-0.393	0.021	**0.945**	-0.059
$PM_{2.5}$	**0.765**	-0.317	-0.311	-0.060	0.235	**0.803**
CPV	0.454	0.686	0.862	0.826	0.913	0.967

We adjust the following LMM to our data

$$Y_{ij} = \beta_0 + \beta_1 P_{ij}^{(1)} + \beta_2 P_{ij}^{(2)} + \beta_3 P_{ij}^{(3)} + \beta_4 \,\text{IgE}_i + \beta_5 \,\text{PS}_i + \gamma_i + U_{ij}, \tag{10}$$

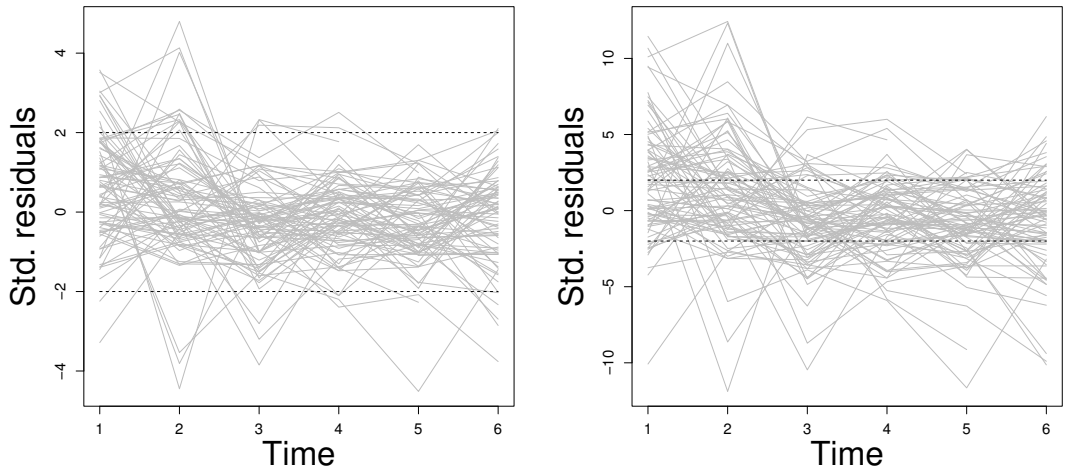
where, for $i = 1, \ldots, 82$ and $j = 1, \ldots, 6$, Y_{ij} is the FEV_1 measurement of ith child at jth trial, $P_{ij}^{(k)}$ is the jth element of $\boldsymbol{P}_i^{(k)} = \boldsymbol{M}_i \boldsymbol{L}^{(k)}$ for $k = 1, 2, 3$, IgE_i is the allergic condition of ith child, PS_i is a binary variable indicating if the ith child is a passive smoker or not, $\gamma_i \sim \mathcal{N}(0, D)$ is a scalar random intercept and U_{ij} is the measurement error of the model. We assume that $\boldsymbol{U}_i = (U_{i,1}, \ldots, U_{i,6})^T \sim \mathcal{N}(\boldsymbol{0}, \sigma_e^2 \boldsymbol{\Omega})$ where the (j,k)th entry of $\boldsymbol{\Omega}$ is $\boldsymbol{\Omega}_{jk} = \nu^{|j-k|}/(1-\nu^2)$ with $|\nu| < 1$.

Let $\boldsymbol{\theta} = (\beta_0, \ldots, \beta_5, D, \sigma_e^2, \nu)^T$. Since $n = 82$ is large enough, the empirical results in Section 3 suggest to take $\omega_{ij} = 1$ in (4). Six different models can be fitted according to the choice of ρ in (4) and whether the loadings $\boldsymbol{L}^{(k)}$ of the PC correspond to the classical PCA or the RPCA. For instance, LS-PCA means that we use $\rho = \rho_1$ and the loadings of the classical PCA, and Tukey-RPCA means that we use $\rho = \rho_3$ and the loadings of the RPCA. Table 9 displays the estimates of $\boldsymbol{\theta}$, their standard errors (SE) and p-values in the fitted models LS-PCA, LS-RPCA, Huber-RPCA and Tukey-RPCA. For the four models, the intercept β_0 and the parameter ν are significant and the parameters β_3 and β_4 are not significant with a p-value larger than 18%. The SE of the estimates of β_1 and β_2 obtained by the LS-PCA are much larger compared to those of the robust approaches, which display very close estimates and standard deviation values. In general, the Huber-RPCA and Tukey-RPCA fitted models are quite similar by presenting very close parameter estimates and p-values. The LS-RPCA also shows similar estimates to these two robust approaches, except the estimate of β_1 which is not significant with a p-value larger than 10%. All methods find that the PS (parameter β_5) give a significant contribution to the response variable with a p-value smaller

Table 9: Estimated coefficients, SE and p-values in the fitted LMM LS-PCA, LS-RPCA, Huber-RPCA and Tukey-RPCA.

	LS-PCA			LS-RPCA			Huber-RPCA			Tukey-RPCA		
	coeff.	SE	p	coeff.	SE	p	coeff.	SE	p	coeff.	SE	p
β_0	103.62	5.00	*	93.74	7.78	*	92.59	8.21	*	92.84	8.18	*
β_1	-1.34	0.30	*	-0.11	0.07	0.115	-0.13	0.07	0.058	-0.14	0.07	0.052
β_2	-0.54	0.41	0.188	-0.22	0.08	0.003	-0.21	0.08	0.008	-0.21	0.08	0.008
β_3	-0.96	0.48	0.046	0.20	0.15	0.180	0.18	0.15	0.225	0.17	0.14	0.222
β_4	-0.85	0.86	0.324	-0.66	0.86	0.440	-0.92	0.94	0.328	-1.11	0.96	0.247
β_5	-6.89	3.07	0.025	-6.49	3.06	0.034	-6.61	3.23	0.041	-5.86	3.29	0.075
D	10.69	1.21	*	9.86	1.44	*	10.17	1.37	*	9.77	1.38	*
σ_e^2	9.89	0.37	*	10.16	0.39	*	9.15	0.45	*	8.49	0.51	*
ν	0.40	0.08	*	0.51	0.08	*	0.48	0.09	*	0.49	0.10	*

* p-value smaller than 0.001.



(a) AR(1) (b) Independent

Figure 4: Estimated standardized residuals of Huber-RPCA model with AR(1) and independent errors.

than $7,5\%$. For the four models, the estimates of variances σ_e^2 and D are significant even at 0.001 (Wellek, 2017).

In Figure 4a, we plot for each $i = 1, \ldots, 82$, the estimated standardized residuals $\widehat{R}_{ij}$ for $j = 1, \ldots, 6$ obtained by replacing $\boldsymbol{\Sigma}_i$ and $\boldsymbol{\beta}$ by their estimates in (3). For all estimation methods, these residuals behave similarly and we display only the residuals obtained with the Huber-RPCA model. To see the influence of the AR(1) parameter ν in model (10), we have also fitted a Huber-RPCA model with independent errors $U_{ij} \sim N(0, \sigma_e^2)$. The corresponding estimated standardized residuals are plotted in Figure 4b. There is a clear difference between Figures 4a and 4b; the variability on the left side is much smaller than on the right, which is a crucial point in favor of the AR(1) structure, see Verbeke and Molenberghs (2000, pages 125-127).

5 Conclusion

This paper proposes to combine RPCA and robust estimation in a LMM in the context where the covariates present outliers and are time-correlated. The high levels of the pollutants create a similar effect as AO on the estimates. The robust methods provide a model which is simple to interpret. They display three predictor factors for the respiratory ability which are the first PC (mainly Hmd), the second PC (mainly PM_{10}) and PS. Monte Carlo simulations show that Huber and Tukey estimation methods are very competitive with the classical LS when there are no outliers in the data. On the other hand, the performances of the robust methods

Funding

The authors are grateful for the support of MESRI (France).

References

Abidin, E. Z., Semple, S., Rasdi, I., Ismail, S. N. S., and Ayres, J. G. (2014). The relationship between air pollution and asthma in Malaysian schoolchildren. *Air Qual. Atmos. Health*, 7(4):421–432.

Aneiros, G., Novo, S., and Vieu, P. (2022). Variable selection in functional regression models: A review. *J. Multivar. Anal.*, 188:104871. 50th Anniversary Jubilee Edition.

Bauer, A., Scheipl, F., Küchenhoff, H., and Gabriel, A.-A. (2018). An introduction to semiparametric function-on-scalar regression. *Stat. Model.*, 18(3-4):346–364.

Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods*. Springer-Verlag, New York, USA, 2 edition.

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.*, 16(5):1190–1208.

Cantoni, E. and Ronchetti, E. (2001). Robust Inference for Generalized Linear Models. *J. Am. Stat. Assoc.*, 96(455):1022–1030.

Cotta, H. H. C., Reisen, V. A., Bondon, P., and Prezotti, F. P. (2020). Identification of redundant air quality monitoring stations using robust principal component analysis. *Environ. Model. Assess.*, 25:521–530.

Crowder, M. (1986). On Consistency and Inconsistency of Estimating Equations. *Econom. Theory*, 2(3):305–330.

Diggle, P. (1988). An Approach to the Analysis of Repeated Measurements. *Biometrics*, 44(4):959–971.

Eddelbuettel, D. and Sanderson, C. (2014). RcppArmadillo: Accelerating R with high-performance C++ linear algebra. *Comput. Stat. Data Anal.*, 71:1054–1063.

Favarato, G., Anderson, H. R., Atkinson, R., Fuller, G., Mills, I., and Walton, H. (2014). Traffic-related pollution and asthma prevalence in children. Quantification of associations with nitrogen dioxide. *Air Qual. Atmos. Health*, 7(4):459–466.

Galvão, A. F., Gu, J., and Volgushev, S. (2020). On the unbiased asymptotic normality of quantile regression with fixed effects. *J. Econom.*, 218(1):178–215.

Gill, P. S. (2000). A robust mixed linear model analysis for longitudinal data. *Stat. Med.*, 19(7):975–987.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. John Wiley and Sons, New York, USA.

Huber, P. J. (1964). Robust Estimation of a Location Parameter. *Ann. Stat.*, 1(53):73–101.

Huber, P. J. (1981). *Robust Statistics*. Wiley and Sons, New York, USA.

Hubert, M., Rousseeuw, P. J., and Branden, K. V. (2005). ROBPCA: A New Approach to Robust Principal Component Analysis. *Technometrics*, 47(1):64–79.

Huggins, R. M. (1993). A Robust Approach to the Analysis of Repeated Measures. *Biometrics*, 49(3):715–720.

Ispány, M., Reisen, V. A., Franco, G. C., Bondon, P., Cotta, H. H. A., Filho, P. R. P., and Serpa, F. S. (2018). On generalized additive models with dependent time series covariates. In Rojas, I., Pomares, H., Valenzuela, and O., editors, *Time Series and Forecasting: Contributions to Statistics*, Contributions to statistics, pages 289–308. Springer series.

Ji, Y. and Shi, H. (2021). Shrinkage estimation of fixed and random effects in linear quantile mixed models. *J. Appl. Stat.*, 0(0):1–24.

Johnson, R. A. and Wichern, D. W. (2002). *Applied multivariate statistical analysis*. Prentice Hall, 6 edition.

Koenker, R. (2004). Quantile regression for longitudinal data. *J. Multivar. Anal.*, 91(1):74–89.

Koller, M. (2013). *Robust Estimation of Linear Mixed Models*. Ph.d., ETH, Zurich, Switzerland.

Koller, M. and Stahel, W. A. (2011). Sharpening Wald-type inference in robust regression for small samples. *Comput. Stat. Data Anal.*, 55(8):2504–2515.

Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006). *Robust Statistics: Theory and Methods*. John Wiley & Sons, New Jersey, USA.

Qu, Y., Pan, Y., Niu, H., He, Y., Li, M., Li, L., Liu, J., and Li, B. (2018). Short-term effects of fine particulate matter on non-accidental and circulatory diseases mortality: A time series study among the elder in Changchun. *PLoS One*, 31(13):e0209793.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reisen, V. A., Lévy-Leduc, C., Cotta, H. A. C., Albuquerque, T. T. A., and Stummer, W. (2017). Long-memory model under outliers: An application to air pollution levels. In Kumar, P., Gurjar, B. R., and Govil, J., editors, *Air and Noise Pollution*, volume 3 of *Environmental Science and Engineering*, pages 211–243. Springer, Studium Press LLC, USA.

Rice, M. B., Ljungman, P. L., Wilker, E. H., Gold, D. R., Schwartz, J. D., Koutrakis, P., Washko, G. R., O'Connor, G. T., and Mittleman, M. A. (2013). Short-term exposure to air pollution and lung function in the Framingham Heart Study. *Am. J. Respir. Crit. Care Med.*, 188(11):1351–1357.

Richardson, A. M. and Welsh, A. H. (1995). Robust Restricted Maximum Likelihood in Mixed Linear Models. *Biometrics*, 51(4):1429–1439.

Rousseeuw, P. and Croux, C. (1993). Alternatives to Median Absolute Deviation. *J. Am. Stat. Assoc.*, 88:1273–1283.

Serpa, F. S. (2019). *Modelo Linear Misto com Interações e Componentes Principais para Avaliar o Efeito Múltiplo de Poluentes e Variáveis Climáticas na Saúde Respiratória*. Ph.d., Universidade Federal do Espírito Santo, Vitória, Brazil.

Shevlyakov, G. and Smirnov, P. (2011). Robust Estimation of the Correlation Coefficient: An Attempt of Survey. *Austrian J. Stat.*, 40:147–156.

Souza, J., Reisen, V., Franco, G., Ispány, M., Bondon, P., and Santos, J. M. (2018). Generalized additive model with principal component analysis: An application to time series of respiratory disease and air pollution data. *J. R. Stat. Soc. Ser. C Appl. Stat.*, 67(2):453–480.

Strickland, M. J., Darrow, L. A., Klein, M., Flanders, W. D., Sarnat, J. A., Waller, L. A., Sarnat, S. E., Mulholland, J. A., and Tolbert, P. E. (2010). Short-term associations between ambient air pollutants and pediatric asthma emergency department visits. *Am. J. Respir. Crit. Care Med.*, 182(3):307–316.

Thurston, G. D., Kipen, H., Annesi-Maesano, I., Balmes, J., Brook, R. D., Cromar, K., Matteis, S. D., Forastiere, F., Forsberg, B., Frampton, M. W., Grigg, J., Heederik, D., Kelly, F. J., Kuenzli, N., Laumbach, R., Peters, A., Rajagopalan, S. T., Rich, D., Ritz, B., Samet, J. M., Sandstrom, T., Sigsgaard, T., Sunyer, J., and Brunekreef, B. (2017). A joint ERS/ATS policy statement: what constitutes an adverse health effect of air pollution? An analytical framework. *Eur. Respir. J.*, 49:1600419.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, USA, 4 edition.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer-Verlag, New York, USA.

Wang, Y. and Pham, H. (2011). Analyzing the effects of air pollution and mortality by generalized additive models with robust principal components. *Int. J. Syst. Assur. Eng. Manag.*, 2:253–259.

Wellek, S. (2017). A critical evaluation of the current "p-value controversy". *Biom. J.*, 59:1–19.

Welsh, A. H. and Richardson, A. M. (1997). Approaches to the robust estimation of mixed models. In *Handbook of Statistics*, volume 15, chapter 13, pages 343–384. Elsevier, Amsterdam, Netherlands.

World Health Organization (2006). *WHO Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide: global update 2005: summary of risk assessment*. Genève, Switzerland.

Zamprogno, B., Reisen, V. A., Bondon, P., Cotta, H. H. A., and Reis Jr., N. C. (2020). Principal component analysis with autocorrelated data. *J. Stat. Comput. Simul.*, 90(12):2117–2135.

6 Appendix

We establish explicit expressions of λ and ϱ in terms of κ_2 and κ_3 when $\rho = \rho_2$ and $\rho = \rho_3$. Let $R \sim \mathcal{N}(0,1)$ and f denote the probability density of R. Let $\kappa > 0$, $n \in \mathbb{N}$ and

$$I_n = \int_{-\kappa}^{\kappa} r^{2n} f(r) dr.$$

Then, $I_0 = \mathbb{P}(|R| \leq \kappa)$ and since $f'(r) = -rf(r)$ under Gaussianity, and using integration by parts, we have for $n \geq 1$,

$$I_n = -\int_{-\kappa}^{\kappa} r^{n-1} f'(r) dr = u_n + (2n-1) I_{n-1},$$

where $u_n = -2\kappa^{2n-1} f(\kappa)$. By iteration, we get

$$I_n = u_n + (2n-1)u_{n-1} + (2n-1)(2n-3)u_{n-2} + \cdots + 3 \cdot 5 \cdots (2n-1)(u_1 + I_0).$$

It was shown by Gill (2000) that

$$\lambda_2 = \mathrm{E}(R\psi_2(R)) = I_0.$$

Let

$$\lambda_3 = \mathrm{E}[R\psi_3(R)] = \int_{-\kappa_3}^{\kappa_3} r^2 \left(1 - \left(\frac{r}{\kappa_3}\right)^2\right)^2 f(r) dr = \int_{-\kappa_3}^{\kappa_3} \left(r^2 - \frac{2r^4}{\kappa_3^2} + \frac{r^6}{\kappa_3^4}\right) f(r) dr$$
$$= I_1 - \frac{2I_2}{\kappa_3^2} + \frac{I_3}{\kappa_3^4} = -\frac{2f(\kappa_3)}{\kappa_3}\left(\frac{15}{\kappa_3^2} - 1\right) + \left(1 - \frac{6}{\kappa_3^2} + \frac{15}{\kappa_3^4}\right) I_0.$$

We have

$$\psi_2(r) = \begin{cases} r & \text{if } |r| \leq \kappa_2, \\ \kappa_2 \operatorname{sgn}(r) & \text{if } |r| > \kappa_2, \end{cases} \qquad \psi_3(r) = \begin{cases} r\left(1 - \frac{r^2}{\kappa_3^2}\right)^2 & \text{if } |r| \leq \kappa_3, \\ 0 & \text{if } |r| > \kappa_3, \end{cases}$$

and

$$\psi_2'(r) = \begin{cases} 1 & \text{if } |r| < \kappa_2, \\ 0 & \text{if } |r| > \kappa_2, \end{cases} \qquad \psi_3'(r) = \begin{cases} \frac{5r^4}{\kappa_3^4} - \frac{6r^2}{\kappa_3^2} + 1 & \text{if } |r| < \kappa_3, \\ 0 & \text{if } |r| > \kappa_3. \end{cases}$$

Then

$$\mathrm{E}(\psi_2'(R)) = I_0,$$
$$\mathrm{E}(\psi_2(R)^2) = \int_{-\kappa_2}^{\kappa_2} r^2 f(r) dr + \kappa_2^2 \left[\int_{-\infty}^{-\kappa_2} f(r) dr + \int_{\kappa_2}^{\infty} f(r) dr\right]$$
$$= I_1 + \kappa_2^2(1 - I_0) = \kappa_2^2 - 2\kappa_2 f(\kappa_2) + (1 - \kappa_2^2) I_0,$$
$$\mathrm{E}(\psi_3'(R)) = \int_{-\kappa_3}^{\kappa_3} \left(\frac{5r^4}{\kappa_3^4} - \frac{6r^2}{\kappa_3^2} + 1\right) f(r) dr = \frac{5I_2}{\kappa_3^4} - \frac{6I_1}{\kappa_3^2} + I_0$$
$$= \frac{2f(\kappa_3)}{\kappa_3}\left(1 - \frac{15}{\kappa_3^2}\right) + \left(1 - \frac{6}{\kappa_3^2} + \frac{15}{\kappa_3^4}\right) I_0,$$

and

$$\mathrm{E}(\psi_3(R)^2) = \int_{-\kappa_3}^{\kappa_3} \left(\frac{r^{10}}{\kappa_3^8} - \frac{4r^8}{\kappa_3^6} + \frac{6r^6}{\kappa_3^4} - \frac{4r^4}{\kappa_3^2} + r^2\right) f(r) dr = \frac{I_5}{\kappa_3^8} - \frac{4I_4}{\kappa_3^6} + \frac{6I_3}{\kappa_3^4} - \frac{4I_2}{\kappa_3^2} + I_1$$
$$= \frac{2f(\kappa_3)}{\kappa_3}\left(1 - \frac{13}{\kappa_3^2} + \frac{105}{\kappa_3^4} - \frac{945}{\kappa_3^6}\right) + \left(1 - \frac{12}{\kappa_3^2} + \frac{90}{\kappa_3^4} - \frac{420}{\kappa_3^6} + \frac{945}{\kappa_3^8}\right) I_0.$$

The explicit expressions of ϱ when $\psi = \psi_2$ and $\psi = \psi_3$ follow easily from (7) and the above calculations.