Journal of Nonparametric Statistics
Vol. 21, No. 2, February 2009, 241–259

Taylor & Francis
Taylor & Francis Group

Nonparametric autocovariance estimation from censored time series by Gaussian imputation

Jung Wook Park[a], Marc G. Genton[b]*, and Sujit K. Ghosh[c]

[a]*Discovery Biometrics, GlaxoSmithKline, NC, USA;* [b]*Department of Statistics, Texas A&M University, TX, USA;* [c]*Department of Statistics, North Carolina State University, Raleigh, NC, USA*

(*Received 26 September 2007; final version received 16 October 2008*)

One of the most frequently used methods to model the autocovariance function of a second-order stationary time series is to use the parametric framework of autoregressive and moving average models developed by Box and Jenkins. However, such parametric models, though very flexible, may not always be adequate to model autocovariance functions with sharp changes. Furthermore, if the data do not follow the parametric model and are censored at a certain value, the estimation results may not be reliable. We develop a Gaussian imputation method to estimate an autocovariance structure via nonparametric estimation of the autocovariance function in order to address both censoring and incorrect model specification. We demonstrate the effectiveness of the technique in terms of bias and efficiency with simulations under various rates of censoring and underlying models. We describe its application to a time series of silicon concentrations in the Arctic.

Keywords: censoring; Gibbs sampling; imputation; nonparametric estimation; truncated multivariate normal distribution

2000 Mathematics Subject Classification: 62G05; 62M10

1. Introduction

Observations collected over time with a monitoring device often have two specific features: a temporal autocorrelation structure and a detection limit. It is common that observations whose values are below the detection limit are recorded as the limit value and flagged to show that the actual values lie somewhere below the detection limit. This is called left censoring and such data should be analysed differently from completely observed cases. The autocorrelation structure is usually handled by time-series models and censoring is often taken care of by a data augmentation method. Data augmentation approaches for time series have been studied by several authors such as Robinson [1], Zeger and Brookmeyer [2], Hopke *et al.* [3], and recently Park *et al.* [4]. All of those methods use a parametric autoregressive and moving average (ARMA) time-series model developed by Box and Jenkins [5]. Thus, a parametric family of models is specified prior to the application of these methods and the estimation is based on these models. The selection of an

*Corresponding author. Email: genton@stat.tamu.edu

ISSN 1048-5252 print/ISSN 1029-0311 online

DOI: 10.1080/10485250802570964
http://www.informaworld.com

'optimal' model within this family is usually based on some information criterion such as the Akaike Information Criterion (AIC) [6] or the Bayesian Information Criterion (BIC) [7]. In practical situations, however, it is not always easy to identify such a model because the ARMA family may not be adequate. Moreover, when the data set is incomplete due to missing or censored values, classical estimation and model selection procedures cannot be used without further modifications. This motivates the need to develop model-free (nonparametric) autocovariance estimation in this context.

Various nonparametric estimators of the covariance function can be found in the literature. For instance, Hall and Patil [8] and Hall *et al.* [9] proposed a nonparametric estimator using kernel smoothing techniques. They chose a symmetric probability density as their kernel. Bjørnstad [10] used the same approach but with a cubic B-spline [11] as an alternative kernel. Wu and Pourahmadi [12] developed a nonparametric estimator of large covariance matrices of longitudinal data based on variograms. Elogne *et al.* [13] proposed the use of an estimator designed by Parzen [14] after spline interpolation. However, all these previous estimation methods are based on completely observed data and may not be useful when part of the data is censored or missing.

This article introduces an approach to analyse data from censored time series based on the imputation method developed by Park *et al.* [4] and the nonparametric estimation of the autocovariance function developed by Hall and Patil [8]. In Park *et al.* [4], an imputation method was introduced to analyse censored time series, the analysis of which is based on simulated observations from the conditional distribution of censored values given the observed ones. The conditional distribution was obtained by using an ARMA model. In this article, as an alternative solution, we estimate the autocovariance function nonparametrically and construct the conditional distribution based on this estimate.

The article is organised as follows. Section 2 describes the nonparametric estimation of the autocovariance function and its implementation within the imputation method. Section 3 reports the results of a simulation study. It compares the proposed nonparametric imputation method with the naive approach under various rates of censoring and underlying models. The naive approach uses the detection limit value as observed data points for the censored observations. In addition, the results of the nonparametric imputation method are compared with those of the parametric imputation method of Park *et al.* [4] based on ARMA models. Section 4 illustrates our new methodology by analysing silicon concentrations in the Arctic. Section 5 presents some discussions on possible extensions of the method.

2. Nonparametric Gaussian imputation

2.1. *Nonparametric autocovariance estimation*

When analysing a time series, it is important to model the autocovariance structure adequately. In case we have a specific second-order stationary model for the data, the autocovariance can easily be derived as a function of the time lags between observations and of the model parameters. For example, the classical stationary autoregressive model of first order, AR(1), has its autocovariance function, $\gamma(h)$, given by

$$\gamma(h) = \frac{\sigma^2}{1-\rho^2}\rho^h, \quad h = 0, 1, 2, \ldots, \tag{1}$$

where h is the time lag between two observations, $|\rho| < 1$ the autoregressive parameter, and σ^2 the error variance. This is a parametric approach because we use the characteristics of the model for the derivation of the autocovariance. However, this approach depends crucially on the correct specification of the autocovariance model. A model-free (nonparametric) approach,

as an alternative solution to a parametric approach, may reduce the risk of using an incorrect autocovariance model.

In this section, we briefly introduce the nonparametric autocovariance estimator of Hall and Patil [8] since it is fairly easy to use and it will be implemented in our simulation study. Hall and Patil [8] and Hall *et al.* [9] proposed a kernel smoothing estimator of a stationary covariance function based on a Nadaraya–Watson kernel estimator on a stationary time series $X_1, \ldots, X_n$ as

$$\tilde{\gamma}(h) = w_n(h)\frac{\sum_{t=1}^{n}\sum_{s=1}^{n} X_{ts}K\{(h-(t-s))/b\}}{\sum_{t=1}^{n}\sum_{s=1}^{n} K\{(h-(t-s))/b\}},$$

where $X_{ts} = (X_t - \bar{X}_n)(X_s - \bar{X}_n)$, $\bar{X}_n$ the sample mean, b a bandwidth, and $K(\cdot)$ a kernel function chosen to be a symmetric probability density. The weight function $w_n(h)$ is chosen to have the property $\sup_{|h|\le c}|w_n(h) - 1| \to 0$ as $n \to \infty$ for each $c > 0$ and $w_n(h) \to 0$ as $|h| \to \infty$ for each $n \ge 1$ [8]. Also, $w_n(h)$ makes $\tilde{\gamma}(h)$ smaller as h increases. However, $\tilde{\gamma}(\cdot)$ is not exactly a covariance function since it does not guarantee the positive semidefiniteness property given by

$$\iint \tilde{\gamma}(s-t)g(s)g(t)\mathrm{d}s\,\mathrm{d}t \ge 0,$$

for all integrable functions $g(\cdot)$. It follows from Bochner's theorem [15] that the non-negativity of the Fourier transformed function of a function guarantees the positive semidefiniteness of that function. To satisfy the positive semidefiniteness of the kernel smoothing estimator, the Fourier transform of $\tilde{\gamma}(h)$ is truncated when below zero. Then, the estimator $\hat{\gamma}(h)$ is given by

$$\hat{\gamma}(h) = (2\pi)^{-1}\int_0^{\infty} \cos(\theta t)\hat{\Psi}(\theta)I(\hat{\Psi}(\theta) > 0)\mathrm{d}\theta, \tag{2}$$

where

$$\hat{\Psi}(\theta) = \int_{-\infty}^{\infty} \cos(\theta h)\tilde{\gamma}(h)\mathrm{d}h, \quad \theta \in \mathbb{R}.$$

Hall and Patil [8] showed that $\hat{\gamma}(h)$ is not only positive semidefinite but also a consistent estimator of $\gamma(h)$.

2.2. *Gaussian imputation based on nonparametrically estimated covariance functions*

The concept of the imputation method is straightforward. Censored observations are imputed by appropriate values in order to construct a pseudo-complete data and a statistical analysis is conducted to analyse these data. The results are used to update the imputation and this process is repeated until the results do not change significantly. Since the observations are correlated, we consider the given data as the realisation of a random vector from a multivariate distribution, typically Gaussian. When a parametric approach is used, the corresponding covariance function can be derived from parameter estimates found by maximum likelihood, least squares, Yule–Walker equations, and so on. Once we have the parameter estimates, we simulate a conditional random sample for the censored observations from a distribution that turns out to be a truncated multivariate distribution. A Gibbs sampling algorithm is then implemented to generate a realisation from a truncated multivariate Gaussian distribution as proposed by Robert [16].

When there is no particular model for the data, a nonparametric approach is used to directly estimate the covariance function. Then the estimated autocovariance function is used to generate a conditional random sample for the censored observations. Whether the imputation method is called parametric or nonparametric depends on how the covariance function is estimated.

The imputation method via nonparametric estimation of the autocovariance function consists of the following steps.

Step 1. Construct the permutation matrix $\mathbf{P}$ so that the data vector $\boldsymbol{x}$ is partitioned into two parts, observed, $\boldsymbol{x}_{\mathrm{O}}$, and censored, $\boldsymbol{x}_{\mathrm{C}}$.

Step 2. Obtain initial estimates from the naive approach: the sample mean $\hat{\mu}^{(0)}$ and a nonparametric autocovariance estimator $\hat{\gamma}^{(0)}(h)$, $h = 0, \ldots, n-1$, such as the one of Hall and Patil [8]. Then we can construct the mean vector, $\boldsymbol{\mu}^{(0)}$, and covariance matrix, $\hat{\boldsymbol{\Sigma}}^{(0)}$, by

$$\hat{\boldsymbol{\mu}}^{(0)} = \hat{\mu}^{(0)}\mathbf{1}_n,$$

$$\{\hat{\boldsymbol{\Sigma}}^{(0)}\}_{ij} = \hat{\gamma}^{(0)}(|i-j|).$$

Step 3. Calculate the conditional expectation, $\hat{\boldsymbol{v}}^{(0)}$, and covariance, $\hat{\boldsymbol{\Delta}}^{(0)}$, of the censored part given the observed part assuming a Gaussian distribution [17].

Step 4. Generate a random sample vector $\boldsymbol{x}_{\mathrm{C}}^{(1)}$ from the following truncated multivariate Gaussian distribution [16]:

$$\boldsymbol{X}_{\mathrm{C}}^{(1)} \sim \mathrm{TN}_{n_{\mathrm{C}}}(\boldsymbol{v}^{(0)}, \boldsymbol{\Delta}^{(0)}, c, \infty), \tag{3}$$

where $\mathrm{TN}_{n_{\mathrm{C}}}$ represents the truncated multivariate Gaussian distribution with dimension n_{C} and support $(c, \infty)^{n_C} = (c, \infty) \times (c, \infty) \times \cdots \times (c, \infty)$ for all components.

Step 5. Construct the augmented data from the observed part and imputed sample for the censored part using $\boldsymbol{P}^{-1}$.

Step 6. Re-estimate μ and $\gamma(h)$. Let $\hat{\mu}^{(1)}$ and $\hat{\gamma}^{(1)}(h)$ be the estimated mean and nonparametric autocovariance function from $\boldsymbol{x}^{(1)}$. Update the mean vector and covariance matrix in Steps 3 and 4.

Step 7. Repeat *Steps 3–6* until convergence. Since we estimate μ and $\gamma(h)$, we need a convergence criterion for both μ and $\gamma(h)$. In stationary time series, it is known that the autocovariance function $\gamma(h)$ decreases exponentially so we use some of the first lags of the autocovariance function. Let $\boldsymbol{\theta} = \{\mu, \gamma(0), \gamma(1), \ldots, \gamma(L)\}$, $L \leq n-1$. Then, we may use the following convergence rule

$$\frac{(\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{(k)})^{\mathrm{T}}(\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{(k)})}{\{\boldsymbol{\theta}^{(k)}\}^{\mathrm{T}}\boldsymbol{\theta}^{(k)}} < \text{tolerance}. \tag{4}$$

We used tolerance $= 0.001$. The imputation method is, therefore, based on iteratively estimating the mean and covariance matrix of a Gaussian process, approximated by augmented data. From our simulation studies, we observed that the estimates did not change much after only a moderate number of iterations.

3. Monte Carlo simulations

3.1. *Simulation design*

In order to study the finite sample properties of our proposed nonparametric imputation method, we simulated data from an AR(1) process having an autocovariance function given by Equation (1)

with $\rho = 0.7$ and an ARMA(1,1) process having an autocovariance function

$$\gamma(h) = \begin{cases} \dfrac{(1+\psi^2-2\rho\psi)}{1-\rho^2}\sigma^2 & \text{for } h = 0, \\ \dfrac{(\rho-\psi)(1-\rho\psi)}{1-\rho^2}\sigma^2 & \text{for } h = 1, \\ \rho\gamma(h-1) & \text{for } h > 1, \end{cases}$$

with $\rho = 0.7$ and $\psi = 0.3$. We used the method described in Section 2.2 to estimate the autocovariance function when some of the observations are subject to censoring. In particular, with $Y_0 = \epsilon_0 = 0$, we used the following steps to generate data:

(1) Generate $Y_t \sim N(\mu + \rho(Y_{t-1} - \mu), \sigma^2)$, to simulate AR(1) and $Y_t \sim N(\mu + \rho(Y_{t-1} - \mu) - \psi\epsilon_{t-1}, \sigma^2)$, $\epsilon_t \sim N(0, \sigma^2)$ to simulate ARMA(1,1) for $t = 1, \ldots, n$ and
(2) Compute $X_t = Y_t(1 - I(Y_t > c)) + c \cdot I(Y_t > c)$,

where we fixed $\mu = 0$ and $\sigma = 1$ for all data generations. The cut-off point c is determined using $\Pr[Y_t > c] = \alpha$ to maintain a targeted censoring rate of $100\alpha\%$ in order to examine the effect of censoring on statistical estimation of the autocovariance function. It follows that $c = \mu + \Phi^{-1}(1-\alpha)\sqrt{\gamma(0)}$, where $\Phi(\cdot)$ denotes the distribution function of a standard normal distribution. We have used $\alpha = 0.1$ and 0.2 to generate data sets that would correspond to approximate censoring rates of 10% and 20%, respectively. In this article, we denote the censored AR(1) and ARMA(1,1) time series as CENAR(1) and CENARMA(1,1), respectively.

We employed three estimation methods for comparison: (i) estimation based on complete data (*i.e.* using the Y_t's), (ii) a naive method that treats the censored data as observed, and (iii) our proposed imputation method. We did not use the exact parametric form of the autocovariance function for any of the three estimation approaches. Results based on parametric methods have been investigated by Park *et al.* [4] under similar scenarios for various other values of ρ and censoring percentages. In this study, our focus is not on the parametric estimation of ρ, ψ, and σ, but rather on estimating the entire autocovariance function $\gamma(h)$ of a stationary time series when some observations are subject to censoring. For all three estimation approaches, we adopted the method of Hall and Patil [8] which is based on only completely observed data to estimate the autocovariance function. We chose to use the following kernel

$$K\left(\frac{h-|t-s|}{b}\right) = \left[1 - \frac{h-|t-s|}{b}\right] I\left[\frac{h-|t-s|}{b} \geq 0\right], \quad \text{and}$$

$$w_n(h) = I(|h| \leq m), \quad \text{where } t, s = 1, \ldots, n.$$

We chose the bandwidth to be $b = 0.05$ and $m = 10$ for all three estimation approaches. This bandwidth is similar to the one used by Hall *et al.* [9]. We set $m = 10$ because the autocorrelation goes quickly to zero in the models considered and we are mainly interested in the autocovariance at small lags. Other choices of b and m we tried led to qualitatively similar results to those to follow. Alternatively, one could choose the bandwidth via a selection method such as cross-validation, and we implemented this additional step in the data analysis reported in Section 4. The sample size was set to $n = 200$ and the data generation and estimation methods were repeated $N = 200$ times.

In addition to comparing the biases of estimates of the mean, μ, and the pointwise estimates of $\gamma(h)$ at $h = 0$, 1, and 2, we also compared various measures of relative efficiency (RE) of the autocovariance function $\gamma(h)$. Let θ denote the true value of a generic parameter (*e.g.* $\theta = \mu$, $\gamma(0)$

or $\gamma(1)$). In particular, four different types of RE are defined depending on the type of mean squared error (MSE) used:

(a) RE:

$$\mathrm{RE}(\hat{\theta}) = \frac{\mathrm{MSE}(\hat{\theta}_{\mathrm{imp}})}{\mathrm{MSE}(\hat{\theta}_{\mathrm{naive}})}, \tag{5}$$

with

$$\mathrm{MSE}(\hat{\theta}) = \frac{1}{N-1}\sum_{i=1}^{N}(\hat{\theta}_i - \theta)^2, \tag{6}$$

where $\hat{\theta}_{\mathrm{imp}}$ and $\hat{\theta}_{\mathrm{naive}}$ are the parameter estimates of a generic parameter θ by the imputation method and the naive approach, respectively. $\hat{\theta}_i$ represents the estimate based on the data generated at the ith Monte Carlo (MC) replication.

(b) Empirical RE (ERE):

$$\mathrm{ERE}(\hat{\theta}) = \frac{\mathrm{EMSE}(\hat{\theta}_{\mathrm{imp}})}{\mathrm{EMSE}(\hat{\theta}_{\mathrm{naive}})}, \tag{7}$$

with

$$\mathrm{EMSE}(\hat{\theta}) = \frac{1}{N-1}\sum_{i=1}^{N}(\hat{\theta}_i - \hat{\theta}_{\mathrm{raw},i})^2, \tag{8}$$

where $\hat{\theta}_{\mathrm{raw},i}$ represents the parameter estimate of θ based on the complete data set (*i.e.* Y_t's) at the ith MC replication. As some of the observations are subject to censoring, the above measure provides efficiency against the complete case analysis.

(c) Integrated RE (IRE):

$$\mathrm{IRE}(h) = \frac{\mathrm{IMSE}(\hat{\gamma}_{\mathrm{imp}}(h))}{\mathrm{IMSE}(\hat{\gamma}_{\mathrm{naive}}(h))}, \tag{9}$$

with

$$\begin{aligned}\mathrm{IMSE}(\hat{\gamma}(h)) &= \frac{1}{N-1}\sum_{i=1}^{N}\int_0^h (\hat{\gamma}(t) - \gamma(t))^2 \mathrm{d}t \\ &\approx \frac{1}{N-1}\sum_{i=1}^{N}\sum_{j=0}^{h}(\hat{\gamma}_i(j) - \gamma(j))^2,\end{aligned}$$

where $\hat{\gamma}_{\mathrm{imp}}(h)$ and $\hat{\gamma}_{\mathrm{naive}}(h)$ are estimates of $\gamma(h)$ obtained by the imputation method and the naive approach, respectively. For some selected lags, $j = 0, 1, \ldots, h$, $\hat{\gamma}_i(j)$ represents the estimate of $\gamma(j)$ based on data sets generated at the ith MC replication.

(d) Empirical IRE (EIRE):

$$\mathrm{EIRE}(h) = \frac{\mathrm{EIMSE}(\hat{\gamma}_{\mathrm{imp}}(h))}{\mathrm{EIMSE}(\hat{\gamma}_{\mathrm{naive}}(h))}, \tag{10}$$

with

$$\mathrm{EIMSE}(\hat{\gamma}(h)) \approx \frac{1}{N-1}\sum_{i=1}^{N}\sum_{j=0}^{h}(\hat{\gamma}_i(j) - \hat{\gamma}_{\mathrm{raw},i}(j))^2,$$

where again $\hat{\gamma}_{\mathrm{raw},i}(j)$ represents the parameter estimate of $\gamma(j)$ based on the complete data set generated at the ith MC replication.

In the following sections, we compare all three estimation approaches using the above measures of RE and biases.

3.2. *Results based on data simulated from AR(1)*

The main findings from the simulation study are summarised both numerically and graphically. First, we present the results for the mean parameter μ and pointwise values of $\gamma(h)$ at $h = 0$, 1, and 2. Table 1 presents the bias and the MC standard deviation of estimates for μ, $\gamma(0)$, $\gamma(1)$, and $\gamma(2)$. The values corresponding to the row labelled 'naive' are based on the naive approach (that treats censored values as observed) and the results corresponding to the row labelled 'imputed' are based on our proposed imputation method. The results corresponding to the column '0%' represent the estimates based on the complete data, henceforth termed as *complete data analysis*. Thus, the results from 'naive' and 'imputed' are the same for this case. The second and the third columns display the estimated biases and standard deviations of the estimates for the targeted censoring rate of 10% and 20%. The results presented in Table 1 clearly indicate that the proposed imputation method has significantly lower bias compared with the naive method and the estimates are comparable to corresponding estimates from the complete data analysis (*i.e.* 0% censoring).

In order to compare the entire sampling distribution of the estimates, in Figure 1, we display the box plots of the ($N = 200$) estimates of μ, $\gamma(0)$, and $\gamma(1)$ based on the three approaches. The (red) box plots corresponding to 'cen' are based on estimates obtained by the 'naive' approach and the (blue) box plots corresponding to 'imp' are based on the proposed imputation method. Table 1 and Figure 1 strongly indicate that the imputation method improves the performance of the estimation in terms of reducing bias and standard error. In particular, for $\gamma(0)$ and $\gamma(1)$, almost the entire box plot of estimates based on the 'naive' approach lies below the true value when the censoring rate is only about 20%. A similar negative bias can also be observed for the estimate of μ based on the 'naive' approach. The main reason behind this bias is the fact that the 'naive' approach treats a censored value as observed when the true value of that observation is actually greater than the observed value.

When comparing estimates with those that might be significantly biased, instead of using the standard errors of the estimates, we compare the RE of the estimation methods using MSEs. We use the four different measures of RE that were defined in the previous section. The reason we consider several measures of RE is due to the fact that estimates are based on two different types of data, one that is fully observed and the other that contains censored observations. Thus, we would like to compare how well the estimates of the parameters based on the imputation and naive approaches (which uses censored data) compare with those based on the complete data

Table 1. Biases and standard deviations (in parentheses) of the estimates from a CENAR(1) model with $\rho = 0.7$, based on three estimation methods.

		0%	10%	20%
Naive	μ	−0.008 (0.216)	−0.073 (0.203)	−0.167 (0.202)
	$\gamma(0)$	−0.116 (0.327)	−0.392 (0.244)	−0.629 (0.220)
	$\gamma(1)$	−0.114 (0.308)	−0.316 (0.238)	−0.490 (0.203)
	$\gamma(2)$	−0.105 (0.279)	−0.248 (0.221)	−0.355 (0.188)
Impute	μ	−0.008 (0.216)	0.006 (0.215)	−0.056 (0.224)
	$\gamma(0)$	−0.116 (0.327)	−0.001 (0.540)	−0.020 (0.405)
	$\gamma(1)$	−0.114 (0.308)	−0.069 (0.384)	−0.101 (0.322)
	$\gamma(2)$	−0.105 (0.279)	−0.077 (0.324)	−0.087 (0.296)

The rows with 'naive' represent the results based on treating the censored values as observed and those with 'impute' represent the results based on the imputation method.

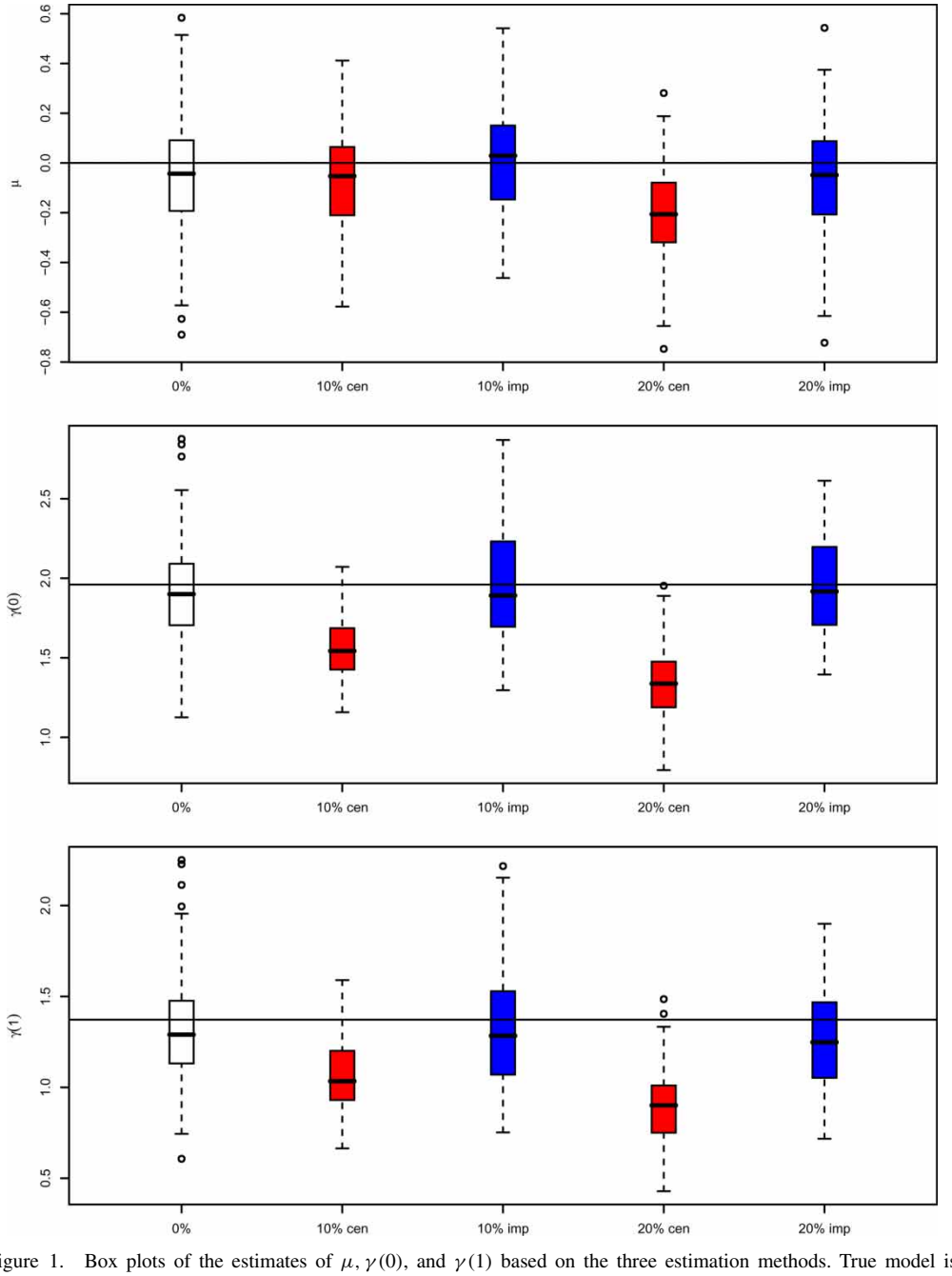

Figure 1. Box plots of the estimates of μ, $\gamma(0)$, and $\gamma(1)$ based on the three estimation methods. True model is CENAR(1) with $\mu = 0$, $\sigma = 1$, and $\rho = 0.7$. White, red, and blue box plots are based on complete data analysis, naive approach, and imputation method, respectively. The solid reference line in each plot represents the true value of the parameter.

analysis (which uses fully observed data). Hence, while RE and IRE measure the closeness of the estimates to the corresponding true values of the parameters, the ERE and EIRE provide a measure of closeness of the estimates to those estimates that are obtained by the complete data analysis.

Figure 2 displays four types of RE where the solid and the broken lines represent the efficiency measures based on 10% and 20% censoring rates, respectively. While RE and ERE are used to measure the RE of finite-dimensional parameters $\mu, \gamma(h)$ at $h = 1, \ldots, 6$ (in the top panels), the IRE and EIRE are used to measure the RE of the function-valued parameter $\gamma(h)$ (in the bottom panels). The reference line at 1 in each of these four plots represents two equally efficient estimates and the graph which lies below 1 represents efficiency at lag values where the imputation method has better efficiency than the naive approach.

In the first plot (in the top left-hand corner), the RE of the naive method seems to be increasing with lag and its value even crosses the reference line at 1 for $h \geq 5$. This is because the RE is based on each lag h, and for the assumed true model (an AR(1) with $\rho = 0.7$), the covariance function $\gamma(h)$ reduces to virtually zero after a certain lag (*e.g.* $h \geq 15$). Thus, even though it seems that the imputation is less efficient than the naive approach at higher lags, the difference is actually negligible. Another reason that our nonparametric imputation method performs slightly poorly is that at higher lags fewer observations are available to obtain efficient nonparametric estimates. We investigate this aspect of estimation from another angle using an integrated version of RE.

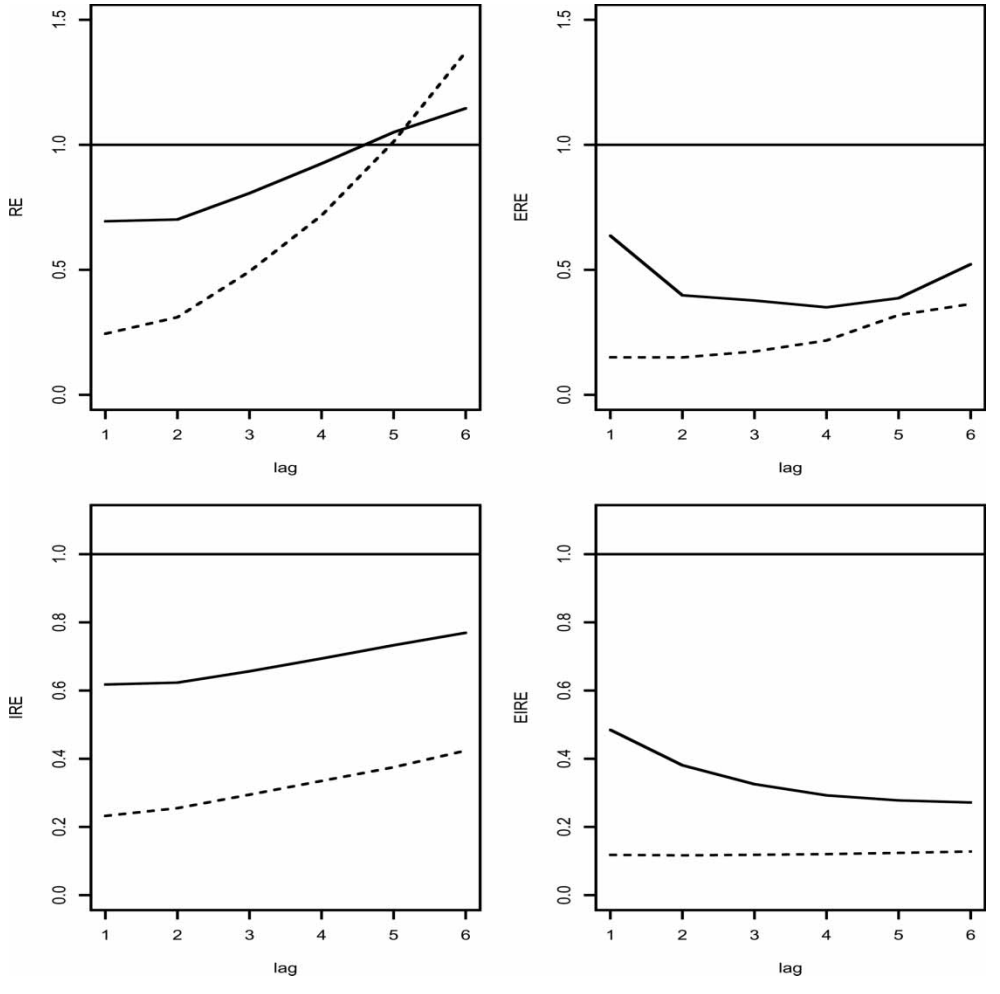

Figure 2. Plots of four types of RE. True model is CENAR(1) with $\mu = 0, \sigma = 1$, and $\rho = 0.7$. The solid and the broken lines represent the efficiency measures based on 10% and 20% censoring rates, respectively. The reference line at 1 in each of these four plots represents two equally efficient estimates.

In the second plot (in the top right-hand corner), instead of a pointwise comparison at each lag with the true value, we compare the efficiency with respect to the complete data analysis. The plot based on ERE now clearly indicates that the imputation method has uniformly better efficiency at all lag values than the naive approach in the sense that the former estimation method produces estimates that are closer to the complete data analysis than those based on the naive approach.

Finally, in the next two plots, both graphs in the lower panels of Figure 2 indicate that the imputation approach has uniformly better efficiency in terms of integrated measure than the naive approach. IRE seems to have a slight increasing pattern but it plateaus around 0.8 and 0.4 for 10% and 20% censoring rates, respectively. Again when an EIRE is used for comparing efficiency, we see significant gain in efficiency by the imputation method over the naive method. This again indicates that the imputation method produces estimates similar to that of the complete data analysis.

3.3. *Results based on data simulated from ARMA(1,1)*

The results presented in Table 2 clearly indicate that the proposed imputation method has significantly lower bias compared with the naive method, particularly for the autocovariances, and the estimates are comparable with corresponding estimates from the complete data analysis (*i.e.* 0% censoring). Figure 3 also indicates that the imputation method improves the performance of the estimation in terms of reducing bias and standard error. In particular, for $\gamma(0)$ and $\gamma(1)$, almost the entire box plot of estimates based on the 'naive' approach lies below the true value when the censoring rate is only about 20%. A similar negative bias can also be observed for the estimate of μ based on the 'naive' approach. The main reason behind this bias is due to the fact that the 'naive' approach treats a censored value as observed when the true value of that observation is actually greater than the observed value.

Figure 4 displays four types of RE. The results are similar to those from the AR(1) model and they indicate that the imputation method produces estimates similar to those of the complete data analysis.

3.4. *Comparison of nonparametric and parametric imputations*

In general, it is well known that a nonparametric estimate loses some efficiency in comparison to a parametric estimate when a correctly specified parametric model is used. So, it would be of interest to calibrate the efficiency loss of our nonparametric method when compared with the parametric method based on the true data generating model. In order to obtain the corresponding

Table 2. Biases and standard deviations (in parentheses) of the estimates from an CENARMA(1,1) model with $\rho = 0.7$ and $\psi = 0.3$, based on three estimation methods.

		0%	10%	20%
Naive	μ	−0.001 (0.171)	−0.057 (0.153)	−0.131 (0.139)
	$\gamma(0)$	−0.007 (0.171)	−0.226 (0.128)	−0.412 (0.125)
	$\gamma(1)$	−0.020 (0.160)	−0.135 (0.120)	−0.231 (0.105)
	$\gamma(2)$	−0.007 (0.152)	−0.091 (0.116)	−0.160 (0.100)
Impute	μ	−0.001 (0.171)	0.000 (0.171)	−0.014 (0.167)
	$\gamma(0)$	−0.007 (0.171)	0.010 (0.197)	−0.033 (0.205)
	$\gamma(1)$	−0.020 (0.160)	−0.032 (0.156)	−0.063 (0.151)
	$\gamma(2)$	−0.007 (0.152)	−0.016 (0.147)	−0.041 (0.140)

The rows with 'naive' represent the results based on treating the censored values as observed and those with 'impute' represent the results based on the imputation method.

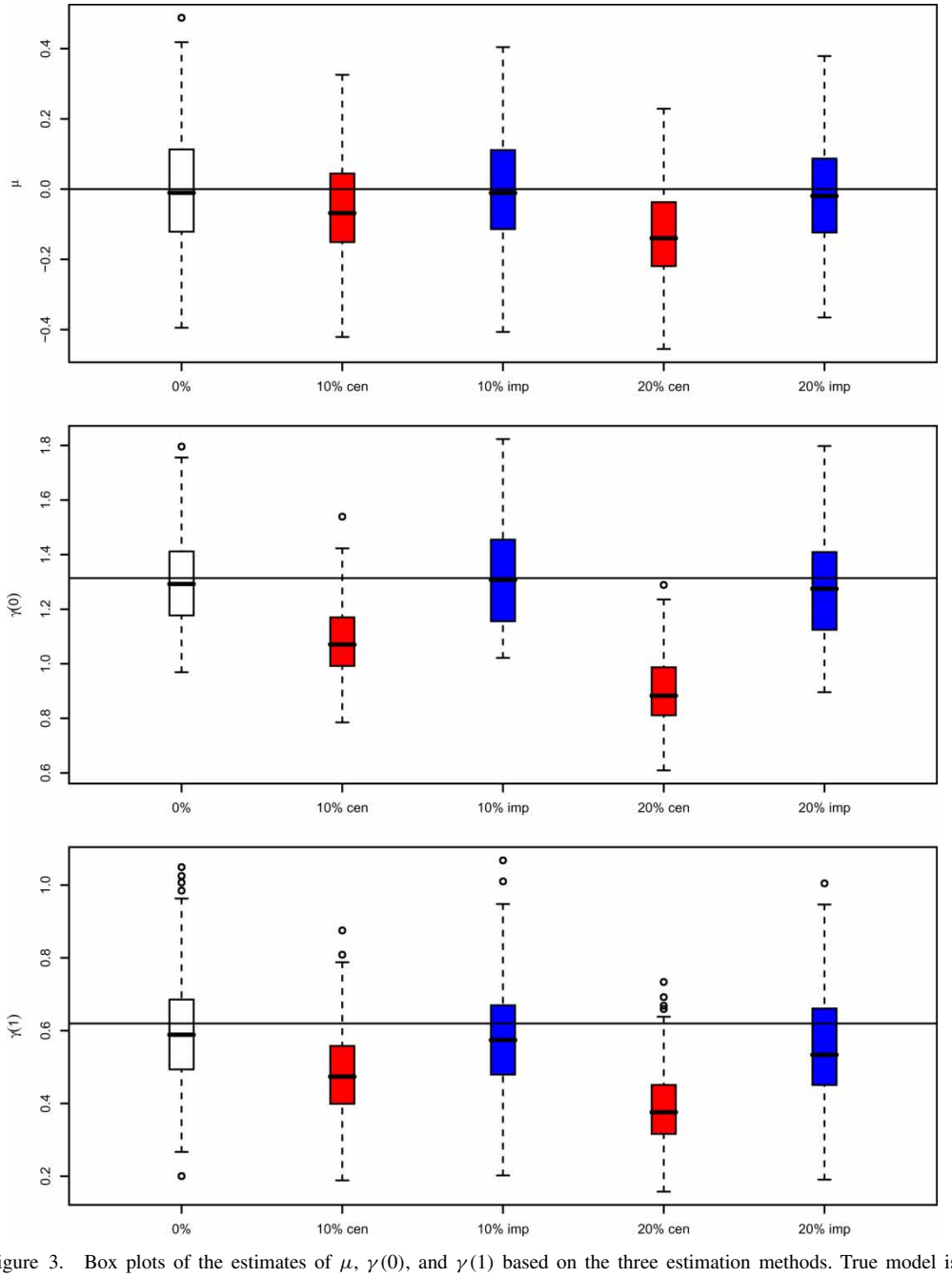

Figure 3. Box plots of the estimates of μ, $\gamma(0)$, and $\gamma(1)$ based on the three estimation methods. True model is CENARMA(1,1) with $\mu = 0$, $\sigma = 1$, $\rho = 0.7$, and $\psi = 0.3$. White, red, and blue box-plots are based on complete data analysis, naive approach, and imputation method, respectively. The solid reference line in each plot represents the true value of the parameter. This figure is available in colour online.

parametric estimates, we borrow some of the results obtained by Park *et al.* [4] for an AR(1) based on a similar simulation study. As the parametric method based on the correct model specification provides estimates of ρ and σ^2, the estimate of $\gamma(h)$ was derived by plugging their estimates in Equation (1).

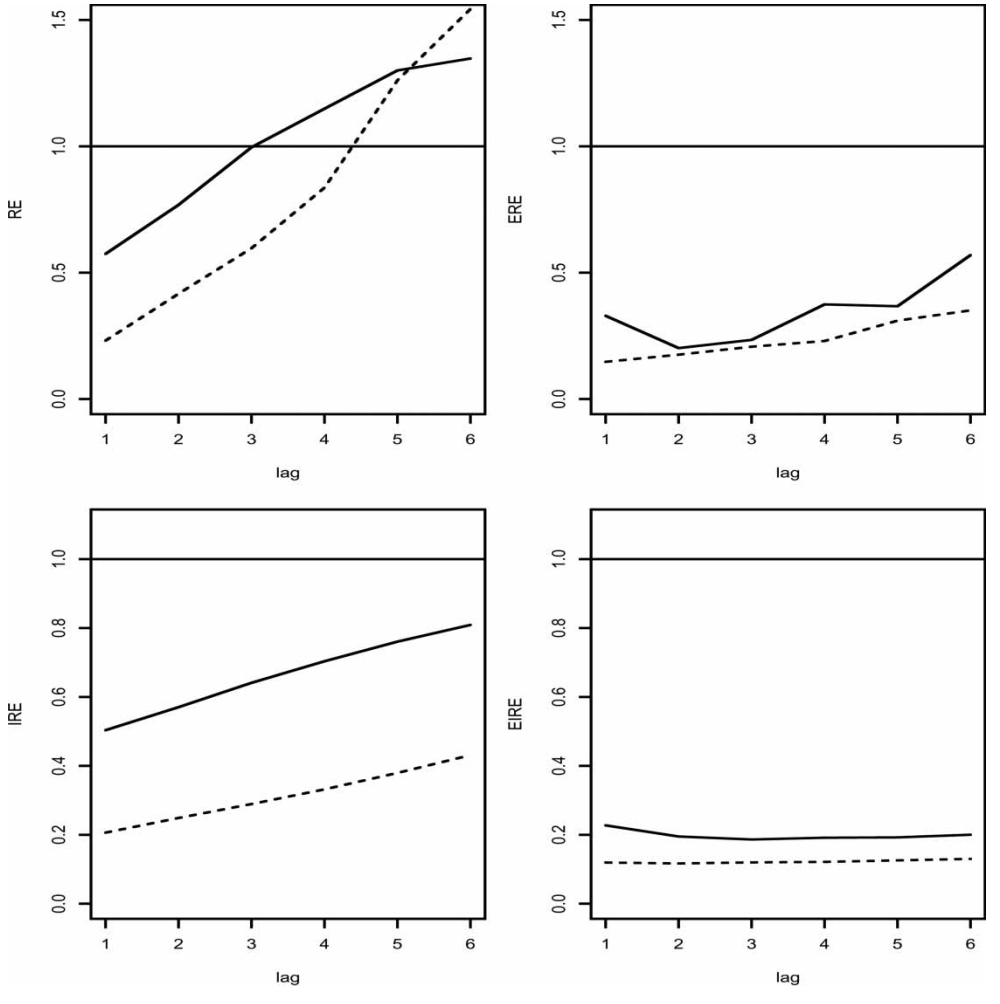

Figure 4. Plots of four types of RE. True model is CENARMA(1,1) with $\mu = 0$, $\sigma = 1$, $\rho = 0.7$, and $\psi = 0.3$. The solid and the broken lines represent the efficiency measures based on 10% and 20% censoring rates, respectively. The reference line at 1 in each of these four plots represents two equally efficient estimates.

In Figure 5, we plot the histograms of the estimates of the parameter μ based on the nonparametric method (on the left) and the parametric method (on the right), when the observed censoring rate is about 20%. Although the biases of both methods are comparable, clearly the nonparametric method produces estimates with more variability having a standard deviation of 0.224 when compared with that of 0.082 based on the parametric method.

In Figure 6, we present pointwise estimates and corresponding 95% confidence intervals of the $\gamma(h)$ function evaluated at the lags, $h = 1, \ldots, 10$. It is easy to see that both methods provide a very good estimate of the true $\gamma(h)$ function (Equation (1)), but the nonparametric method looses some efficiency at higher lag values (*e.g.* when $h \geq 5$) when compared with the parametric method which is based on the correct model specification. Thus, we observe that although the nonparametric method may loose some efficiency in estimating $\gamma(h)$, especially at higher lag values, the point estimates are still reliable in the sense of having a bias of the same magnitude as the parametric method (which is based on a correct specification). On the other hand, if a wrong parametric autocovariance model is used, the biases (and standard errors) of the parametric method

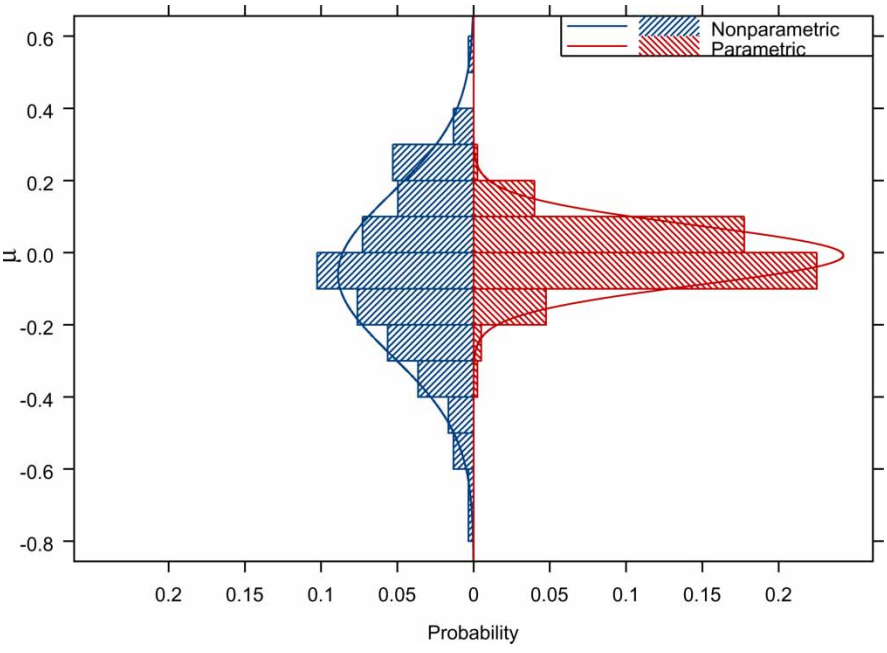

Figure 5. Histograms of the estimates of μ using nonparametric (on the left) and parametric (on the right) methods, when the censoring rate is 20%.

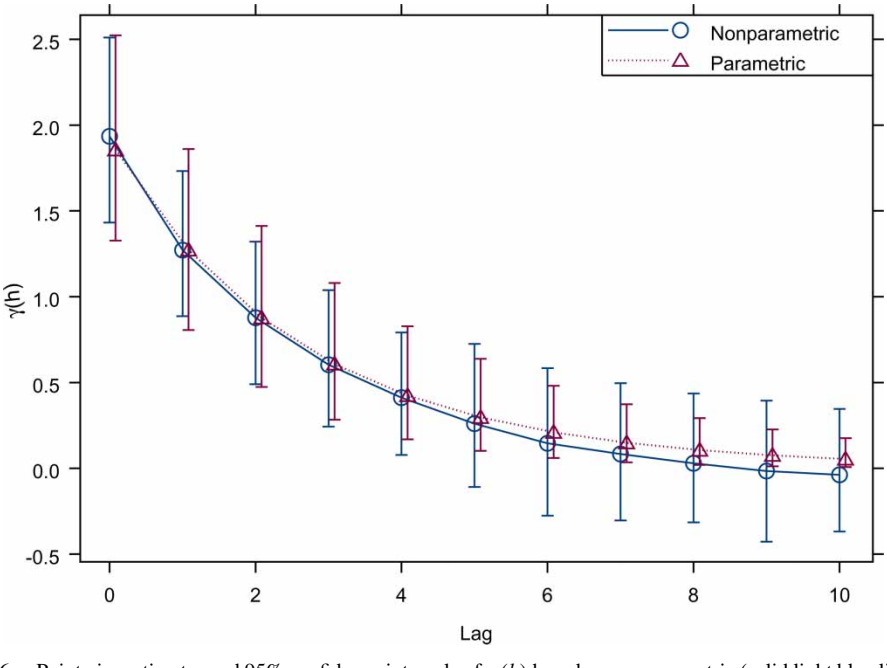

Figure 6. Pointwise estimates and 95% confidence intervals of $\gamma(h)$ based on nonparametric (solid light blue line) and parametric (dashed pink line) methods, when the censoring rate is 20%. This figure is available in colour online.

Table 3. Biases and standard deviations (in parentheses) of the estimates from a CENAR(1) model with errors having a t-distribution with d.f. $= 5$, based on the three estimation methods ($\rho = 0.7$ and d.f. $= 5$).

		0%	10%	20%
Naive	μ	0.003 (0.277)	−0.056 (0.259)	−0.181 (0.243)
	$\gamma(0)$	−0.193 (0.599)	−0.627 (0.454)	−1.087 (0.43)
	$\gamma(1)$	−0.172 (0.515)	−0.477 (0.4)	−0.825 (0.357)
	$\gamma(2)$	−0.147 (0.474)	−0.359 (0.372)	−0.618 (0.303)
Impute	μ	0.003 (0.277)	−0.004 (0.266)	−0.046 (0.265)
	$\gamma(0)$	−0.193 (0.599)	−0.21 (0.601)	−0.284 (0.663)
	$\gamma(1)$	−0.172 (0.515)	−0.221 (0.484)	−0.316 (0.509)
	$\gamma(2)$	−0.147 (0.474)	−0.189 (0.445)	−0.273 (0.421)

The rows with 'naive' represent the results based on treating the censored values as observed and those with 'impute' represent the results based on the imputation method.

would be of a much greater magnitude than the corresponding nonparametric method, making them very unreliable.

3.5. *Gaussian imputation for heavy-tailed time series*

One possible limitation of our proposed imputation method is that it relies on imputing the censored observations based on a conditional Gaussian distribution and hence might not be fully efficient when the error process of a time series follows a heavy-tailed distribution (e.g. a t-distribution or a double-exponential distribution). In order to evaluate the performance of our proposed Gaussian-based imputation method even when the white noise follows a heavy-tailed distribution, we perform a simulation study of an AR(1) model with errors having a t-distribution. The parameters of the model were set to $\mu = 0, \sigma = 1, \rho = 0.7$, and degrees of freedom d.f. $= 5$ of the t-distribution. Following essentially a very similar format as in the previous simulation studies in earlier sections, the results are presented in Table 3 and Figures 7 and 8.

The results from our simulation study with a heavy-tailed error process indicate that the estimates based on our Gaussian imputation method are quite similar to those obtained from the complete data analysis case, and hence the Gaussian imputation method seems to be robust against heavy-tailed error processes. However, the naive method seems to produce highly biased autocovariance estimates, and overall, the biases and standard deviations are significantly larger than those for Gaussian error series. Thus, it appears that the naive method is quite sensitive to the underlying error distribution, while the proposed Gaussian-based imputation method seems to be relatively insensitive to the error process of the time series that we have tested in our simulation studies. One possible reason for this apparent insensitivity of our Gaussian-based imputation method is that the kernel estimates that we use are based on the second-order moments of the time series and hence are approximately unbiased even when the error process is heavy-tailed. On the other hand, the naive method leads to biased estimates even when the error process is Gaussian and hence the biases get even inflated when the error process is heavy-tailed.

4. Silicon concentrations in the Arctic

Time-series data on silicon (Si) concentrations were collected at Alert, Northwest Territories, Canada, by the Atmospheric Environment Service of Canada starting July 1980 for 572 weeks [3,18]. The purpose of collecting these data is to study the effects of Arctic air pollution on the

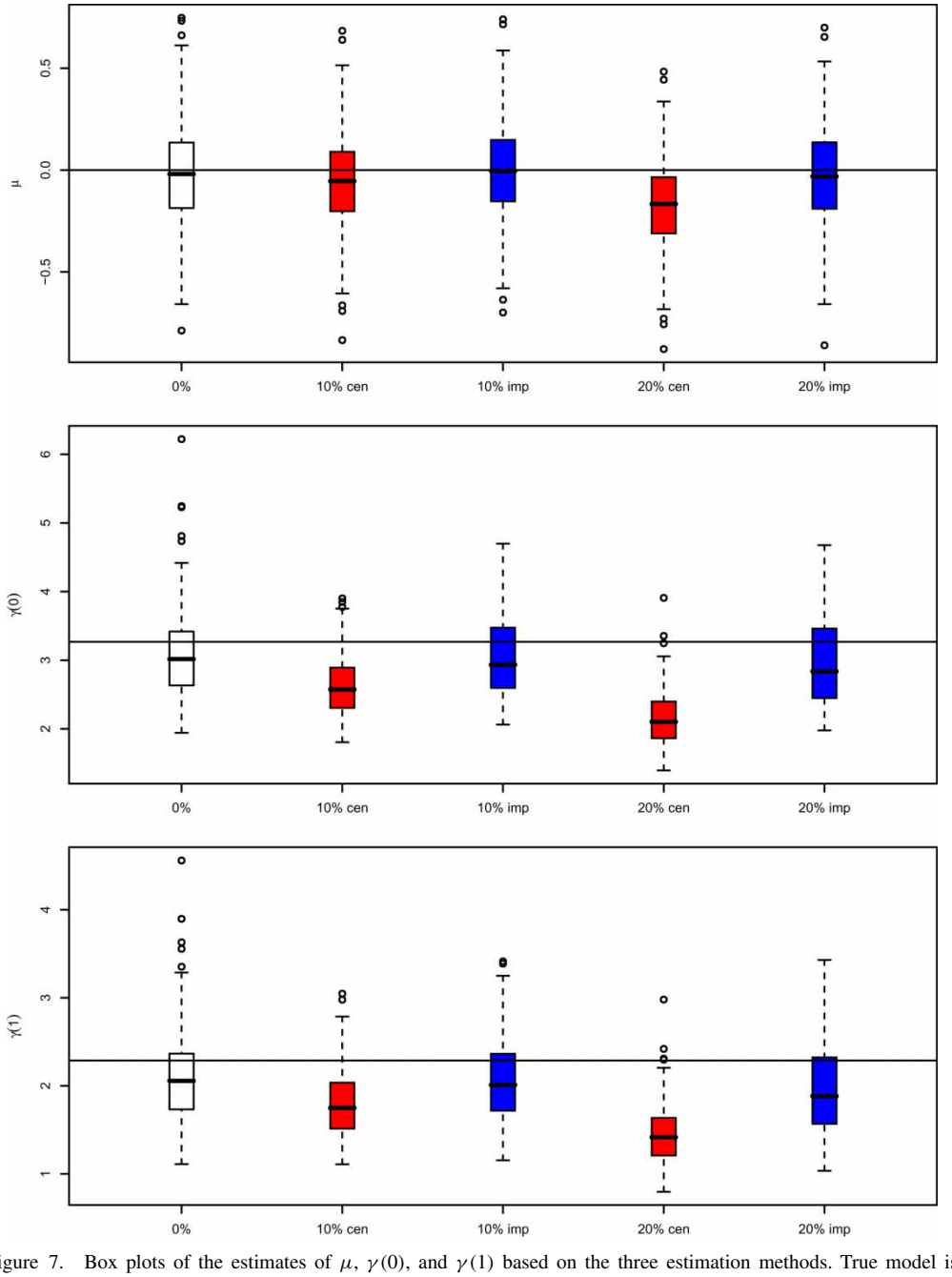

Figure 7. Box plots of the estimates of μ, $\gamma(0)$, and $\gamma(1)$ based on the three estimation methods. True model is CENAR(1) with $\mu = 0$, $\sigma = 1$, $\rho = 0.7$, and errors having a t-distribution with d.f. $= 5$. White, red, and blue box plots are based on complete data analysis, naive approach, and imputation method, respectively. The solid reference line in each plot represents the true value of the parameter.

polar ecosystem. Part of the data is censored or missing, hence it is an incomplete data set and we need to handle it properly. Because the monitoring machine does not detect concentrations below a certain value, the data are regarded as left censored. About 20% of the data are missing and 45% are censored.

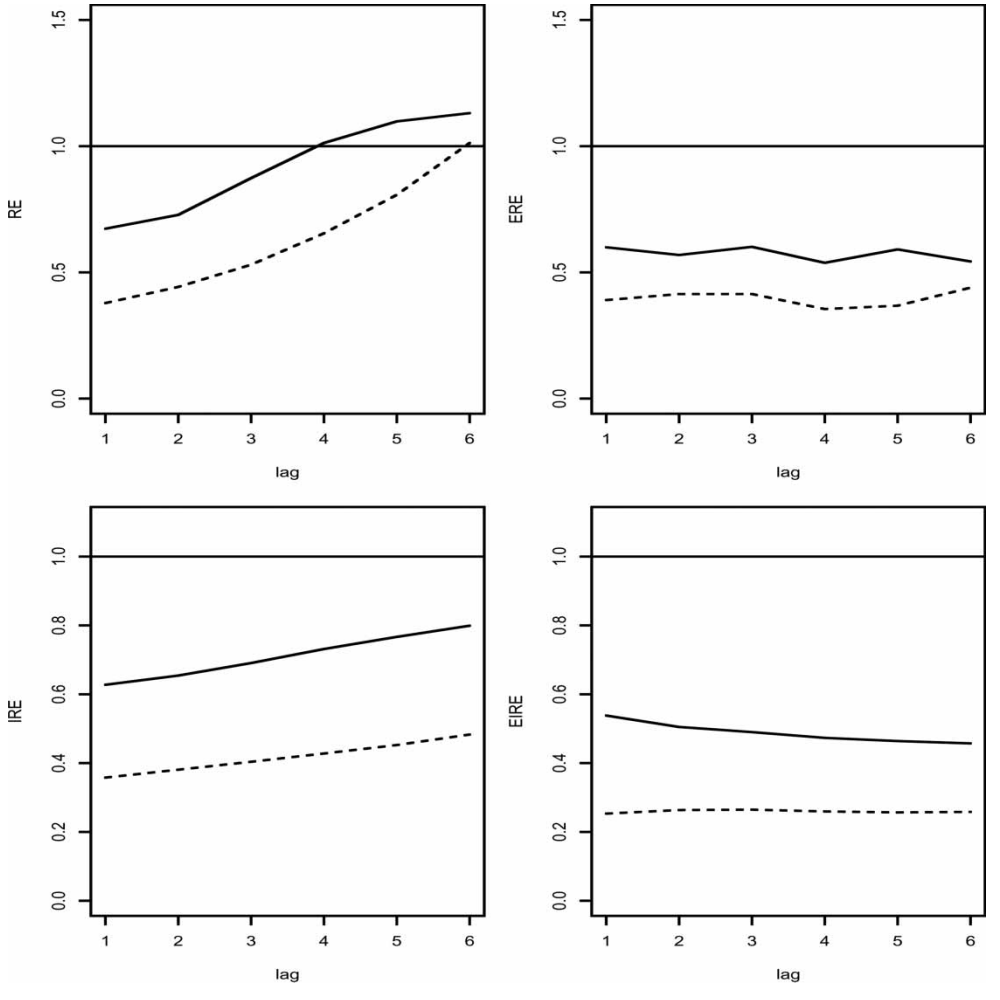

Figure 8. Plots of four types of RE. True model is CENAR(1) with $\mu = 0$, $\sigma = 1$, $\rho = 0.7$, and errors having a t-distribution with d.f. $= 5$. The solid and the broken lines represent the efficiency measures based on 10% and 20% censoring rates, respectively. The reference line at 1 in each of these four plots represents two equally efficient estimates.

First, we apply the naive approach and the imputation method based on parametric time-series models. We use the Box and Jenkins [5] methodology to identify an appropriate ARMA model. In terms of the AIC selection criterion, AR(2) and ARMA(1,1) models gave very similar results displayed in Table 4. The AR parameters are ρ_i, $i = 1, 2$, and the MA parameter is ψ. Because the data are left censored, the mean, μ, obtained from the imputation method is deflated in comparison to the naive approach. The two approaches give rather different results for the fitting of the ARMA(1,1) model. For the imputation method, the moving average parameter is somewhat smaller than that for the naive approach, although both are significantly different from 0. Also, the AIC value for the imputation method is much bigger than that for the naive approach due to the inflation of the estimate of the standard deviation σ.

Next, we apply the nonparametric imputation method to the Si time-series data. We use a cross-validation method on the nonparametric autocovariance estimator within our imputation procedure in order to choose the bandwidth and obtained $b = 0.9$. As in the simulation study, we use $m = 10$. Because no specific autocovariance model is used, we estimate the mean and the autocovariance

Table 4. Parameter estimates for Si.

	$\hat{\mu}$	$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\psi}$	$\hat{\sigma}$	AIC
Results on Si from the naive approach						
AR(2)	5.686 (0.010)	0.384 (0.006)	0.179 (0.006)	NA	0.648 (0.007)	912.7
ARMA(1,1)	5.686 (0.010)	0.716 (0.008)	NA	0.325 (0.007)	0.649 (0.007)	912.5
Results on Si from the imputation method						
AR(2)	5.394 (0.046)	0.400 (0.027)	0.142 (0.027)	NA	0.881 (0.024)	1491.6
ARMA(1,1)	5.255 (0.054)	0.641 (0.053)	NA	0.261 (0.073)	1.037 (0.029)	1652.3

Two candidate models were selected by the AIC selection criterion.

Table 5. Estimates of the mean, μ, and autocovariance, $\gamma(h)$, for Si.

	$\hat{\mu}$	$\hat{\gamma}(0)$	$\hat{\gamma}(1)$	$\hat{\gamma}(2)$
Results on Si from the naive approach				
AR(2)	5.686 (0.010)	0.555 (0.009)	0.260 (0.004)	0.199 (0.004)
ARMA(1,1)	5.686 (0.010)	0.553 (0.009)	0.260 (0.004)	0.186 (0.003)
Nonparametric	5.686 (0.010)	0.430 (0.005)	0.169 (0.002)	0.120 (0.002)
Results on Si from the imputation method				
AR(2)	5.394 (0.046)	1.017 (0.061)	0.475 (0.048)	0.336 (0.046)
ARMA(1,1)	5.255 (0.054)	1.343 (0.069)	0.580 (0.053)	0.372 (0.045)
Nonparametric	5.381 (0.052)	1.038 (0.104)	0.468 (0.064)	0.318 (0.045)

function of the data. In order to compare the nonparametric imputation method with the parametric imputation method, we calculate the autocovariance function from the parametric imputation method based on the AR(2) and ARMA(1,1) models. The results are displayed in Table 5. The standard errors in Tables 4 and 5 were computed using a block bootstrap method [19].

The autocovariance estimated from the naive approach seems much smaller than that obtained from the imputation method. This shrinkage is a consequence of the lack of information below the detection limit. The estimates of $\gamma(0)$ from the imputation method are similar for the AR(2) and the nonparametric approach, while the ARMA(1,1) model yields a larger value. This may be due to the fact that the underlying autocovariance function of the Si data more closely resembles that of an AR(2) process than an ARMA(1,1) process.

5. Discussion

In many longitudinal and time-series studies, repeated observations are collected over time on the same subject or event, and hence it is more reasonable to assume that such observations, per subject or event are autocorrelated. Also often some of the intermittent observations may be missing or censored (e.g. the silicon concentration example of Section 4). The autocovariance function is often assumed to belong to a class of parametric models (e.g. compound symmetry or ARMA(p, q)) but such an assumption can be misleading as the true autocovariance function may not belong to such a class of models and will necessarily lead to biased estimates (even asymptotically). In this article, we proposed a nonparametric estimation of the underlying autocovariance function in the presence of censored and missing data.

Our proposed method can be easily extended to adapt to a regression model with correlated errors and one may modify the popular iteratively reweighted least squares method [20] to estimate the parameters in the model. Instead of assuming a specific form of the covariance function

that is required for a generalized least squares procedure, we can use our proposed nonparametric method to estimate the underlying covariance function and the regression coefficients simultaneously. Even when the primary objective of such regression analysis is the estimation of the regression coefficients, the standard errors of the coefficient estimates are affected by the choice of the underlying covariance structure of the errors. It would be a rather time-consuming work to use a trial-and-error method to find the optimal correlation model for the error terms. Our proposed nonparametric procedure provides an adaptive estimate that depends only on the assumption that the underlying covariance function is stationary and the fact that the error process is Gaussian. Thus, the proposed imputation method combined with the nonparametric estimation of a covariance function provides a fairly flexible method to analyse correlated data involving censored observations. The results based on simulated data (in Section 3) and a real data set (in Section 4) illustrate the usefulness and benefits of our method in terms of reducing biases and increasing the RE of the estimates when compared with a naive approach.

The application of the proposed method is not limited to time series or longitudinal data with missing or censored observations. Notice that the imputation method is based on the assumption of representing the observed process by a multivariate Gaussian distribution with a specified mean and covariance structure. We can extend the use of the imputation method to a data set that may consist of spatially autocorrelated observations. There are many approaches to estimate the so-called variogram/semivariogram function nonparametrically [21,22] under a weak stationarity assumption and we may incorporate our imputation method within such nonparametric estimation procedures.

There are, however, some limitations of the method. First, it assumes a multivariate Gaussian distribution for the observed data and hence the proposed method of imputation may not be reliable if such an assumption is violated. However, using data simulated from a heavy-tailed error process (e.g., t_5 distribution), we have demonstrated that our Gaussian-based imputation method still produces approximately unbiased estimates of the autocovariance function and is hence relatively robust against error model misspecifications. It is to be noted that we do not make any parametric assumption about the form of the mean and the covariance functions. Wei [23] discussed the benefits of the Box–Cox transformation and one may explore extensions of our imputation method to such trans-Gaussian models. Alternatively, one could use multivariate skew-elliptical distributions [24] to relax the Gaussian assumption. Second, during our simulation studies, we have observed the possibility of divergent sequence of estimates in some rare cases when the lag, L, for the covariance function is large (e.g. $L \approx n - 1$). This computational problem may be solved by finding good starting values for the mean and covariance function, but this numerical limitation requires further studies.

Acknowledgements

The authors would like to thank the Editor, the Associate Editor, and two referees for comments that improved this manuscript. This research was partially supported by NSF grants DMS-0504896 and CMG ATM-0620624.

References

[1] P.M. Robinson, *Estimation and forecasting for time series containing censored or missing observations*, in *Time Series: Proceedings of the International Meeting held at Nottingham University, Nottingham, March 26–30, 1979*, O.D. Anderson, ed., North-Holland, Amsterdam, 1980, pp. 167–182.

[2] S.L. Zeger and R. Brookmeyer, *Regression analysis with censored autocorrelated data*, J. Amer. Statist. Assoc. 81 (1986), pp. 722–729.

[3] P.K. Hopke, C. Liu, and D.B. Rubin, *Multiple imputation for multivariate data with missing and below-threshold measurements: time-series concentrations of pollutants in the Arctic*, Biometrics 57 (2001), pp. 22–33.

[4] J. Park, M.G. Genton, and S. Ghosh, *Censored time series analysis with autoregressive moving average models*, Canadian J. Statist. 35 (2007), pp. 151–168.

[5] G.E.P. Box and G.M. Jenkins, *Time Series Analysis: Forecasting and Control*, Holden-Day, San Francisco, 1976.
[6] H. Akaike, *Maximum likelihood identification of Gaussian autoregressive moving average models*, Biometrika 60 (1973), pp. 255–265.
[7] G. Schwartz, *Estimating the dimension of a model*, Ann. Statis. 6 (1978), pp. 461–464.
[8] P. Hall and P. Patil, *Properties of nonparametric estimators of autocovariance for stationary random fields*, Probab. Theory Related Fields 99 (1994), pp. 399–424.
[9] P. Hall, N.I. Fisher, and B. Hoffmann, *On the nonparametric estimation of covariance functions*, Ann. Statist. 22 (1994), pp. 2115–2134.
[10] O. Bjørnstad, *Nonparametric spatial covariance function: Estimation and testing*, Environ. Ecol. Statist. 8 (2001), pp. 53–70.
[11] P.J. Green and B.W. Silverman, *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, Chapman and Hall, London, 1994.
[12] W.B. Wu and M. Pourahmadi, *Nonparametric estimation of large covariance matrices of longitudinal data*, Biometrika 90 (2003), pp. 831–844.
[13] S.N. Elogne, O. Perrin, and C. Thomas-Agnan, *Non parametric estimation of smooth stationary covariance functions by interpolation methods*, Statist. Inference Stoch. Process. 11 (2008), pp. 177–205.
[14] E. Parzen, *Mathematical considerations in the estimation of spectra*, Technometrics 3 (1961), pp. 167–190.
[15] S. Bochner, *Harmonic Analysis and the Theory of Probability*, University of California Press, Los Angeles, 1955.
[16] C.P. Robert, *Simulation of truncated normal variables*, Statist. Computing 5 (1995), pp. 121–125.
[17] T.W. Anderson, *An Introduction to Multivariate Statistical Analysis*, Wiley, Hoboken, NJ, 1984.
[18] Y.-L. Xie, P.K. Hopke, P. Paatero, L.A. Barrie, and Li, S.-M. *Identification of source nature and seasonal variations of Arctic aerosol by positive matrix factorization*, J. Atmos. Sci. 56 (1999), pp. 249–260.
[19] S.N. Lahiri, *Resampling Methods for Dependent Data*, Springer-Verlag, New York, NY, 2003.
[20] B. Kedem and K. Fokianos, *Regression Models for Time Series Analysis*, Wiley, NY, 2002.
[21] M.G. Genton and D.J. Gorsich, *Nonparametric variogram and covariogram estimation with Fourier-Bessel matrices*, Comput. Statist. Data Anal. 41 (2002), pp. 47–57.
[22] P.H. García-Soidán, M. Febrero-Bande, and W. González-Manteiga, *Nonparametric kernel estimation of an isotropic variogram*, J. Statist. Plan. Inference 121 (2004), pp. 65–92.
[23] W.S. Wei, *Time Series Analysis: Univariate and Multivariate Methods*, Addison-Wesley, Reading, MA, 1990.
[24] M.G. Genton, *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality*, Edited Volume, Chapman & Hall/CRC Press, Boca Raton, Florida, 2004.