Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ucgs20

Selective Imputation of Covariates in High Dimensional Censored Data

Caroline Svahn & Oleg Sysoev

To cite this article: Caroline Svahn & Oleg Sysoev (2022): Selective Imputation of Covariates in High Dimensional Censored Data, Journal of Computational and Graphical Statistics, DOI: 10.1080/10618600.2022.2035233

To link to this article: https://doi.org/10.1080/10618600.2022.2035233

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2022, VOL. 00, NO. 0, 1–9
https://doi.org/10.1080/10618600.2022.2035233

Selective Imputation of Covariates in High Dimensional Censored Data

Caroline Svahn and Oleg Sysoev

Department of Computer and Information Science, Linköping University, Linköping, Sweden

ABSTRACT

Efficient modeling of censored data, that is, data which are restricted by some detection limit or truncation, is important for many applications. Ignoring the censoring can be problematic as valuable information may be missing and restoration of these censored values may significantly improve the quality of models. There are many scenarios where one may encounter censored data: survival data, interval-censored data or data with a lower limit of detection. Strategies to handle censored data are plenty, however, little effort has been made to handle censored data of high dimension. In this article, we present a selective multiple imputation approach for predictive modeling when a larger number of covariates are subject to censoring. Our method allows for iterative, subject-wise selection of covariates to impute in order to achieve a fast and accurate predictive model. The algorithm furthermore selects values for imputation which are likely to provide important information if imputed. In contrast to previously proposed methods, our approach is fully nonparametric and therefore, very flexible. We demonstrate that, in comparison to previous work, our model achieves faster execution and often comparable accuracy in a simulated example as well as predicting signal strength in radio network data. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received December 2020
Revised December 2021

KEYWORDS
Censored covariates; Nonparametric model; Random forest; Wireless networks

1. Introduction

Handling censored data is essential for many research fields. Survival models are widely used when the response is subject to censoring, however, less effort has been put into modeling of censored predictors. Censoring due to a lower limit of detection is common in data measured with some instrument not having precision enough to detect small values, such as for instance biomedical data (Paxton et al. 1997; Hughes 1999; Lyles, Lyles, and Taylor 2000), or signal detection (Ryden et al. 2018).

Maximum likelihood is a common approach for handling censored covariate data. Lee et al. (2018) present a maximum likelihood based method using generalized linear models for the case when potentially all covariates are censored. The censoring limits can be set individually for the different covariates. In de Lima Taga and Singer (2018), linear regression is used to obtain maximum likelihood estimators of the parameters in order to handle cases of right- or left-censored data for both the covariates and the response. Gomez, Espinal, and Lagakos (2003) also studies a likelihood approach for the interval censored, single covariate case.

Yue and Wang (2016) consider a Bayesian approach using a Bayesian linear model using auxiliary variables which can handle several types of censoring, including subject specific censoring limits. They argue that even though this model does not perform very well for extensive censoring, it works better than imputing and modeling in two independent steps. Bayesian GLMs are also suggested by Wu et al. (2012) to handle primarily left censored data, although, the method can be extended to right or interval censoring, and offers one lower limit per covariate. The method is however sensitive to the choice of prior distribution. A bridge between the Bayesian and the frequentist approaches is offered by May, Ibrahim, and Chu (2011), where a Monte Carlo version of the EM algorithm is used. The method allows for interval censoring, subject specific and covariate specific limits of detection as well as response censoring. As the method requires solving an extensive integral, they use rejection sampling to approximate the resulting distribution.

Bernhardt, Wang, and Zhang (2015) suggest improper multiple imputation using the Metropolis Hastings algorithms and generalized linear models for imputing all censored values. Lee, Kong, and Weissfeld (2012) focus on variable specific lower limits of detection and use multiple imputation to handle the case where the covariates are correlated and heavily censored. A heavy censoring context may in greater extent eliminate complete cases, which are needed for an initial estimate of the model parameters. Arunajadai and Rauh (2012) also present a multiple imputation method using a generalized Gamma distribution to get the expected value of censored covariates, which allows for varying censoring limits. Tsimikas, Bantis, and Georgiou (2012) also consider a generalized Gamma distribution for the covariates in combination with a simple linear model assuming independence between covariates, response and the censoring limit. Their approach allows for a nonparametric form of the response and is, according to the authors, computationally simple.

CONTACT Caroline Svahn *caroline.svahn@liu.se* Department of Computer and Information Science, Linköping University, Linköping, Sweden.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JCGS*.

Previous work consider cases with one or only a few covariates, and the overview indicates that the field still lacks a strategy for efficient processing of a large number of censored covariates when all covariates are subject to censoring and when the covariates and the response have unknown nonlinear relationships. In this article, we present a selective multiple imputation approach to minimize the mean squared error and execution time when a larger number of covariates are subject to several different types of censoring simultaneously and when there are no complete cases available.

Our selective multiple imputation approach is based on the method proposed by Bernhardt, Wang, and Zhang (2015), motivated by their computationally light framework relative to other approaches and the ability to handle several censored covariates. We propose a number of alterations to the Bernhardt, Wang, and Zhang (2015) approach that, while adapting the approach to fit our predictive scenario, allow for handling high-dimensional data and the absence of complete cases. More specifically, we propose a selective prediction approach for better scaling and to avoid imputations having low impact on predictive quality. To side step the need for complete cases we propose a k nearest neighbor estimation strategy. We furthermore propose using Random Forest to offer a flexible nonparametric estimation of the covariate distribution means. We focus on data with a lower limit of detection, however, our model can be applied to other natures of censoring.

Necessary adjustments of the Bernhardt, Wang, and Zhang (2015) approach will be explained in Section 2.2. In Section 2.3, we introduce an approach for selecting the values that are most likely to be influential for imputation and better estimates for the initial imputations. In Section 3 we run experiments on simple artificial data and in Section 4 we evaluate the models on data simulated to resemble signal strength data in a wireless network. Finally, in Section 5, we discuss the results and make concluding remarks.

2. Methods

The reference algorithm proposed by Bernhardt, Wang, and Zhang (2015) will be presented in Section 2.1. In Section 2.2 we present necessary adjustments to focus on the predictive modeling, for processing data with no complete cases and for increasing the model flexibility. In Section 2.3 we present a selective imputation approach using kNN imputation.

2.1. Improper Multiple Imputation

Bernhardt, Wang, and Zhang (2015) present a model where some, or, with some alterations, all covariates are subject to censoring. They assume a joint truncated normal distribution for the censored values as a result of assuming joint normality for the covariates. Censored values are iteratively imputed by rejection sampling.

They assumed the observed data to consist of n observations, continuous censored covariates $\mathbf{x}$ and fully observed continuous covariates $\mathbf{z}$. Let $\boldsymbol{y}$ be the binary response in a generalized linear model where $\mathbf{x}$ and $\mathbf{z}$ are independent variables. Let $\mathbf{x}^o$ be the observed values in the censored covariates, $\mathbf{x}^c$ the censored values of the censored covariates and $\boldsymbol{d}^c$ be a vector containing the lower limit of detection for each observation. For observation i, they assume that the distribution of $\mathbf{x}$, $p(\boldsymbol{x_i}|\boldsymbol{z_i};\boldsymbol{\gamma})$, follows a known q-variate distribution, where $\boldsymbol{\gamma}$ are the distribution parameters and q is the number of censored covariates. Furthermore, for covariate $\boldsymbol{x}^j = (x_1^j, \ldots, x_n^j)$ and a lower limit of detection L_i, we define an indicator for censoring as

$$\delta_i^j = I(x_i^j \geq L_i). \tag{1}$$

Then, their algorithm can be described as follows:

1. Obtain an initial estimate of $\boldsymbol{\gamma}$ using maximum likelihood, where $\boldsymbol{\gamma}$ is the true parameter vector of the candidate distributions for the censored $\mathbf{x}^c$.
2. Using the complete cases, obtain an initial estimate of $\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ are the true parameters of the GLM fit of $p(\boldsymbol{y}|\mathbf{z}, \mathbf{x}, \boldsymbol{\beta})$.
3. For every observation i subject to censoring, generate an imputation vector for $\boldsymbol{x}_i^c$ from the joint distribution $p(\boldsymbol{x}_i^c|y_i, \boldsymbol{z_i}, \boldsymbol{x}_i^o, \boldsymbol{x}_i^c < \boldsymbol{d}_i^c; \hat{\boldsymbol{\theta}})$, where $\boldsymbol{x}_i^c$ are the censored values in observation i for the covariates subject to censoring and $\boldsymbol{x}_i^o$ are the observed values in observation i for the covariates subject to censoring. $\hat{\boldsymbol{\theta}}$ is the entire parameter vector $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})^T$ and d_i^c is the lower detection limit for $\boldsymbol{x}_i$.
4. Using the candidate imputations as well as the observed values, reevaluate $\hat{\boldsymbol{\gamma}}$ using maximum likelihood and $\hat{\boldsymbol{\beta}}$ using a GLM.
5. Repeat Steps 3 and 4 M times, yielding M estimates of $\hat{\boldsymbol{\theta}}$.
6. Obtain the final estimate of each parameter θ_r as the mean of all iterations:

$$\hat{\theta}_r = \frac{\sum_{m=1}^{M} \hat{\theta}_{r,m}}{M}. \tag{2}$$

The imputations in step 3 are generated using the acceptance-rejection method:

1. For $\boldsymbol{x}_i^c$, generate a candidate vector $\tilde{\boldsymbol{x}}_i^c$ from the truncated normal distribution $p(\boldsymbol{x}_i^c|\boldsymbol{z}_i, \boldsymbol{x}_i^o, \boldsymbol{x}_i^c < \boldsymbol{d}_i^c; \hat{\boldsymbol{\gamma}})$ obtained from $p(\boldsymbol{x_i}|\boldsymbol{z_i};\boldsymbol{\gamma})$.
2. Generate u from Unif(0,1).
3. If $u < p(y_i|\boldsymbol{z_i}, \boldsymbol{x_i}^o, \tilde{\boldsymbol{x}}_i^c)$, accept the candidate vector $\tilde{\boldsymbol{x}}_i^c$, otherwise retry with a new candidate vector according to Step 1.

The algorithm results in a dataset where all censored values are imputed.

Note that $u \in [0, 1]$ and therefore, the right hand side of the rejection step inequality must be limited to [0, 1]. Therefore, for a regression scenario some majorizing constant is required. Further note that estimating $\hat{\boldsymbol{\beta}}$ requires complete cases and that data which cannot be considered normally distributed requires an alternative approach for modeling the covariates. Bernhardt, Wang, and Zhang (2015) suggest that the chain rule can be used to model each conditional distribution more flexibly:

$$p(\boldsymbol{x}_i|\boldsymbol{z}_i) = p(x_i^1|x_i^2, \ldots, x_i^q, \boldsymbol{z}_i) \cdot p(x_i^2|x_i^3, \ldots, x_i^q, \boldsymbol{z}_i) \ldots p(x_i^q|\boldsymbol{z}_i). \tag{3}$$

As stated by the authors, assuming the correct distribution for the covariates is crucial to the performance of their method, however, nonparametric approaches were not studied.

2.2. Multiple Imputation

Bernhardt, Wang, and Zhang (2015) offer a promising multiple imputation framework. In our work, we focus on data without complete cases where all covariates are subject to censoring and the distributions cannot be modeled with parametric methods.

Without complete covariates, Equation (3) reduces to

$$p(\boldsymbol{x}_i) = p(x_i^1|x_i^2,\ldots,x_i^q)\cdot p(x_i^2|x_i^3,\ldots,x_i^q)\ldots p(x_i^q), \quad (4)$$

leaving $p(x_i^q)$ to be some distribution over the range of $\boldsymbol{x}^q$. As a flexible, nonparametric alternative to the parametric distribution assumption we propose to model the distribution of the covariates using Random Forests as they are able to model complex nonlinear dependencies (Breiman 2001). Let x_i^j be the j:th feature in an observation vector $\boldsymbol{x}_i$. We model the multivariate probability density $p(\boldsymbol{x}_i)$ by using Equation (4) and computing conditional probability density $p(x_i^j|x_i^{j+1},\ldots,x_i^q)$ by Random Forest regressions, as shown by Algorithms 1 and 2.

Algorithm 1: Inference on $p(\boldsymbol{x})$

given current imputed dataset $\mathbf{x}$; number of trees B in the RF
for $j = 1$ *to* $q-1$ **do**
  fit a RF with B trees, response $\boldsymbol{x}^j$, predictors $\boldsymbol{x}^{j+1},\ldots,\boldsymbol{x}^q$ to $\mathbf{x}$
  obtain point prediction function $\mu_j(\boldsymbol{x}^{j+1},\ldots,\boldsymbol{x}^q)$
  compute σ_j^2 as the residual variance from the RF training
end

Algorithm 2: Sample generation from $p(\boldsymbol{x})$

given functions $\mu_j(\boldsymbol{x}^{j+1},\ldots,\boldsymbol{x}^q)$, scalars $\sigma_j^2, j = 1,\ldots,q-1$ and distribution $p(\boldsymbol{x}^q)$
generate $\hat{\boldsymbol{x}}^q$ from $p(\boldsymbol{x}^q)$
for $j = q-1$ *to* 1 **do**
  generate $\hat{\boldsymbol{x}}^j$ from $\mathcal{N}\left(\mu_j(\boldsymbol{x}^{j+1},\ldots,\boldsymbol{x}^q),\sigma_j^2\right)$
end
output the vector $(\hat{\boldsymbol{x}}^1,\ldots,\hat{\boldsymbol{x}}^q)$

The probability model $p(\boldsymbol{y}|\boldsymbol{x}^1,\ldots,\boldsymbol{x}^q)$ is estimated and generated in the same manner as any of $p(\boldsymbol{x}^j|\boldsymbol{x}^{j+1},\ldots,\boldsymbol{x}^q)$, however, using the full set of features.

As there are no complete cases available for estimation, a naïve imputation approach considered by previous research can be used, such as for example imputing censored values with a lower limit of detection vector $\boldsymbol{L} = (L_1,\ldots,L_n)$ (Hornung and Reed 1990).

In order to compare the likelihood of the candidate vector to u in the acceptance-rejection step, we introduce a majorizing constant, C, as the highest point of the density for each y prediction in order to have an appropriate majorizing density for the generator distribution (Gentle 2002). For all observations, we set C as

$$C = \frac{1}{\sigma\sqrt{2\pi}}, \quad (5)$$

where σ is the standard deviation of the residuals of the current imputed dataset. This ensures that the value to compare to u is between 0 and 1. An algorithmic overview of the multiple imputation process can be found in the supplementary materials.

2.3. Selective Multiple Imputation Using kNN

Due to the lack of information available when many predictors and many values are censored the algorithm proposed by Bernhardt, Wang, and Zhang (2015) may lead to low predictive accuracies and large computational times needed to predict a large amount of censored values. We therefore propose a modification which we call *selective multiple imputation*. This approach is *selective* as it imputes only some portion of the censored values that the approach considers to be useful to aid the prediction of the response while the remaining censored values are set to a constant. More specifically, our approach skips imputations for which the observed part of the subject lacks resemblance to other observations, which also speeds up execution. Our approach is *multiple* as it is based on multiple improper imputation techniques.

For an observation $\boldsymbol{x}_i$ with one or more censored values, we investigate whether it is feasible to make realistic imputations. Let n_o be the number of fully observed values in observation i and n_c be the number of censored values in observation i. We then investigate the feasibility of imputation by checking that the user set ratio of fully observed values requirement in observation i, $o_{\min} \in [0,1]$, is met by comparing it to the ratio of observed values, $n_o/(n_o+n_c)$. Note that $o_{\min} = 1$ results in no imputation, as this means that we require all values of an observation to be noncensored. If the ratio of fully observed values is lower than the set minimum, the entire observation is skipped and all censored values for observation i are set to a fixed value S_i.

As observations that are similar have a potential to offer a more informative starting imputation than imputing with S_i, with S_i being for instance equal to the lower limit of detection, we propose to use k nearest neighbor estimation introduced by Cover and Hart (1967) for finding suitable initial imputations for selected values. The kNN algorithm computes the distances between all observations and thereby finds the observations that are the most similar. The neighborhood of an observation is defined in the space of the obtained distances, and the size of the neighborhood is decided by a user set integer k. Then the k nearest neighborhood, that is, the k closest observations in terms of distance, can be used to make decisions or predictions for the observation by getting majority votes or an average estimation of said neighborhood (Bishop 2006). Since the traditional Euclidean distance measure would yield erroneous distances for censored data, we suggest a version of kNN which computes the distances modified to handle the censored values, according to (Jonsson and Wohlin 2004).

In order to explain our kNN strategy, we first introduce some notations and provide an illustration, see Figure 1. Let I_i be a set of all indices of the fully observed values in $\boldsymbol{x_i}$, that is,

$$I_i = \{j|\delta_i^j = 1\}. \quad (6)$$

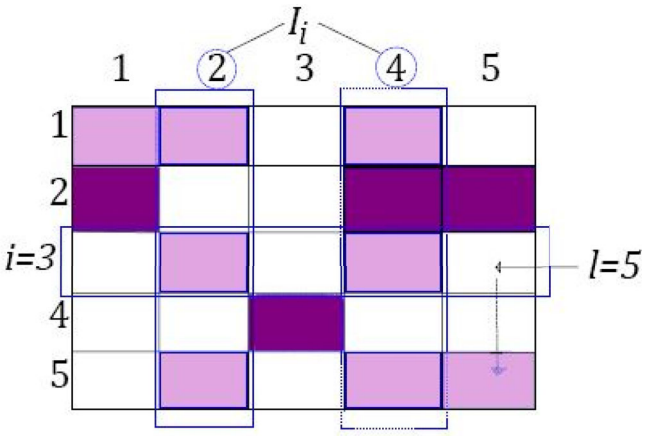

Figure 1. An illustration of searching for neighbors of observation i that can potentially be used to impute the feature indexed by l. Rows with light colored cells can potentially be used to impute some cells in row i as they have noncensored values in the columns indexed by I_i; rows with dark colored cells cannot be used for imputing row i.

In Figure 1, these are illustrated by highlighted cells in row 3, while censored values are illustrated by empty cells. Our method aims to impute the empty cells.

Let l be an index of one of the censored values in observation i. In Figure 1, we consider $l = 5$. Let I_l be the set of indices of all observations which are noncensored in the features indexed by I_i and in the feature l, that is,

$$I_l = \{s | \delta_s^j = 1 \ \& \ \delta_s^l = 1 \text{ for each } j \in I_i\}. \tag{7}$$

Thus, I_l are the indices of all potential neighbors to observation i. For the example in Figure 1, only observation 5 is noncensored in feature 2 and 4 as well as 5, therefore, this is the only potential neighbor.

From the pool of potential neighbors, we extract those that satisfy the minimum observation ratio criteria and the minimum k criteria set by the user, and we find the distances between observation x_i and neighbor x_s as

$$d(\boldsymbol{x}_i, \boldsymbol{x}_s) = \sqrt{\sum_{p \in I_i} (x_i^p - x_s^p)^2}. \tag{8}$$

The k nearest neighbors are then chosen as the k observations with the smallest d values.

If the minimum ratio criterion or minimum k criterion is not satisfied, the value of the l:th feature is not imputed. In the case where at least k neighbors are found, we obtain the initial imputation for the l:th missing value as

$$(\tilde{x}^c)_i^l \leftarrow \begin{cases} \frac{\sum_{a=1}^k (x^c)_a^l}{k}, & \text{if } \frac{\sum_{a=1}^k (x^c)_a^l}{k} \leq L_i^d \\ S_i, & \text{otherwise} \end{cases} \tag{9}$$

for all censored values chosen for iterative imputation. The selective initial imputation algorithm is then defined as Algorithm 3.

The parameters in $p(\boldsymbol{x}^j | \boldsymbol{x}^1, \ldots, \boldsymbol{x}^{j-1})$ and $p(\boldsymbol{y} | \boldsymbol{x}^1, \ldots, \boldsymbol{x}^q)$ are the parameters of the Random Forest models. The values for the censored values not chosen for iterative imputation ($\mathbf{x}^{c^{-*}}$) are fixed to $\mathbf{S}$. An overview of the process can be found in Algorithm 4.

Algorithm 3: Initial imputation using kNN

```
for i = 1 to n do
    if o_min < n_o/(n_o+n_c) < 1 then
        compute I_i according to Equation (6)
        for l ∈ {1, ..., q} \ I_i do
            compute I_l according to Equation (7)
            let I be the indices of the k smallest d(x_i, x_s) such
            that s ∈ I_l
            if |I| ≥ k then
                set initial imputation (x^{c*})_i^l according to
                Equation (9)
            end
        end
    end
end
```

Algorithm 4: Selective Multiple Imputation using kNN

```
choose x^{c*} and impute using kNN and impute x^{c^{-*}} with S
for m = 1 to M do
    for j = 1 to q do
        estimate all parameters in p(x^j | x^1, ..., x^{j-1})
    end
    estimate all parameters in p(y | x^1, ..., x^q)
    for i = 1 to n do
        repeat
            generate x̃_i^c ~ p(x^c | x^1, ..., x^q; L_i)
            generate u ~ Unif(0, 1)
        until u < p(y_i | x^1, ..., x^q) / C;
        impute x_i^{c*} with x̃_i^{c*}
    end
end
```

3. Simulation Study

We have evaluated the performance of the algorithms described in Sections 2.2 and 2.3 in terms of accuracy and execution time for simple artificially generated data. Let $\boldsymbol{\omega}$ be a range of deterministic values, and σ_j and μ_j the parameters of a normal distribution. To enable easy visual illustrations, we choose $\boldsymbol{\alpha}$ as a grid of integer values. We furthermore choose σ_j and μ_j so that we can generate data in which most observations are subject to censoring. The value for observation index i and covariate index j is then generated using a normal density (scaled with some constant G) according to:

$$x_i^j = \phi(\alpha_i | \mu_j, \sigma_j) + \varepsilon_i^j, \tag{10}$$

where

$$\phi(\alpha_i | \mu_j, \sigma_j) = \frac{G}{\sigma_j \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\alpha_i - \mu_j}{\sigma_j}\right)^2}, \tag{11}$$

and the noise, ε_i^j, is proportional to the maximum of each observation:

$$\varepsilon_i^j \sim \mathcal{N}\left(0, 0.05 \cdot \max(x_i^1, \ldots, x_i^q, y_i)\right). \tag{12}$$

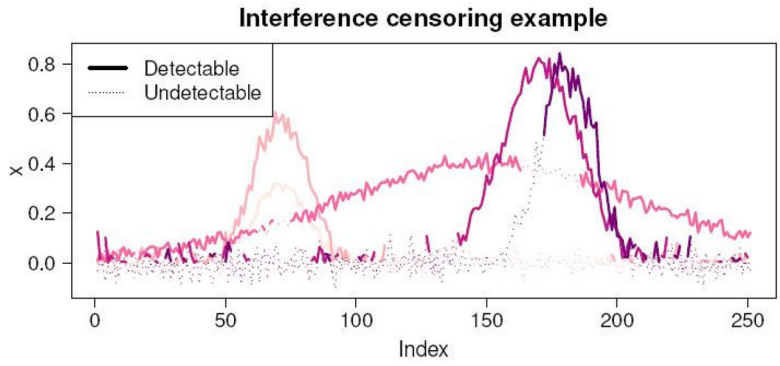

Figure 2. A visual example of censoring dependent on the maximum of each observation. The plot shows the covariate values for a range of indices. The different lines represent different covariates, all subject to censoring. The solid lines are the detectable parts of the covariate, while the dotted lines are undetectable. The figure illustrates that the high valued covariates drenches the low valued covariates. For instance, there is only one detectable covariate around indices 110–120, as it has drenched all other covariates due to its magnitude.

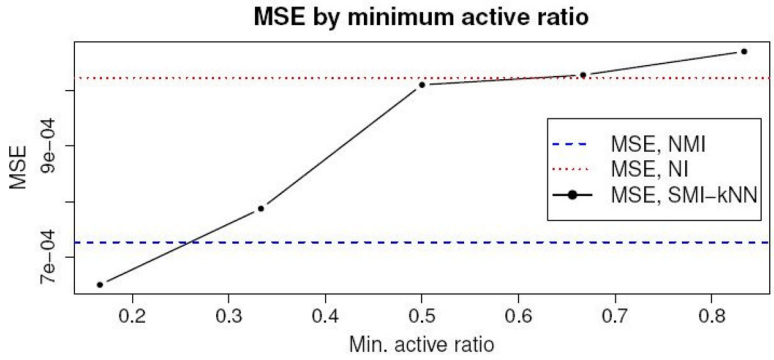

Figure 3. MSE by minimum active. The plot shows the MSE dependence of o_{min} for an artificially generated set of 250 × 6 modeled with SMI-*k*NN where $k = 5$. The red line shows the MSE when no imputations are made and the blue line shows the MSE when all censored values are imputed.

We set the standard deviation to 5% of the maximum as this introduces some dynamic random variation in the data without having a big impact on the relationship between the covariates and the response. The response in our study, $\boldsymbol{y}$, is computed as follows:

$$y_i = \max(\boldsymbol{\omega_i}) + \varepsilon_i^j, \tag{13}$$

where each element in $\boldsymbol{\omega_i}$ follows Equation (11) with parameter values μ_i^y and σ_i^y.

As we want to evaluate if our method can handle complex censoring, all covariates are censored according to the following principle:

$$x_i^j \leftarrow \begin{cases} x_i^j, & \text{if } x_i^j \geq L_i \,\&\, x_i^j \geq \max(\boldsymbol{x}_i) - \Delta \\ L_i, & \text{otherwise,} \end{cases} \tag{14}$$

where Δ is a known threshold representing the maximum difference between the highest valued covariate and every other covariate, and L_i is a known censoring limit. Thus, a value in observation i can be censored either by being below a physical lower limit of detection, L_i, or by being too small in comparison to the maximum value in observation i. From this, we define a second, dynamic, lower limit of detection in addition to L_i as

$$L_i^d = \max(\boldsymbol{x}_i) - \Delta. \tag{15}$$

This aims to mimic a type of interference censoring, where dominant values in observations "drenches" less dominant values. Interference is a common problem in signal processing, for instance in localization problems (Dovis 2015). Scaling the noise with the maximum therefore yields a dynamic fluctuation, sensitive to the subject specific magnitudes of the data. For this example, we let $\boldsymbol{L} = 0$ to limit to positive values and $\Delta = 0.15$ to achieve censoring which will censor the majority of the values in the data. We will elaborate this statement and explain the reasoning behind this type of censoring in the scenario presented in the next section. We perform simulations for two different data sizes (the parameter settings can be found in the supplementary materials). See a visual example of the covariates described above as well as this censoring nature in Figure 2.

For this simple example we let $q = 6$ and the chosen δ results in 62% of the data to be censored. The starting values for the censored values not chosen for imputation are set to $\boldsymbol{L}$, as they are reasonably below that limit.

In Figure 3, different levels of minimum observed ratio by observation is plotted against the mean squared error (MSE) of the regression predictions of the model $p(y_i|\boldsymbol{x^1}, \ldots, \boldsymbol{x^q})$ relative to the mean squared error of the complete data model, computed as

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(\max(\boldsymbol{\omega}_i) - \hat{y}_i\right)^2, \tag{16}$$

where $\hat{y}_i$ is the ith prediction of y_i. Note that $\max(\boldsymbol{\omega}_i)$ is the true noiseless response, which follows Equation (11). Therefore, the MSE in these results is a measure for how well the model estimates the true response and not the training error of the model. This shows the resilience against noise for each approach.

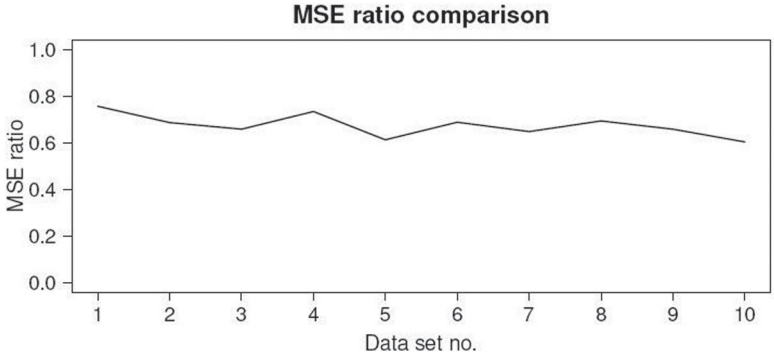

Figure 4. A comparison of the predictive performance of *k*NN selected imputation versus selecting the same ratio of values to impute at random. The plot shows that for 10 randomized datasets, the MSE is reduced to between 60% and 80% when using our selective imputation approach.

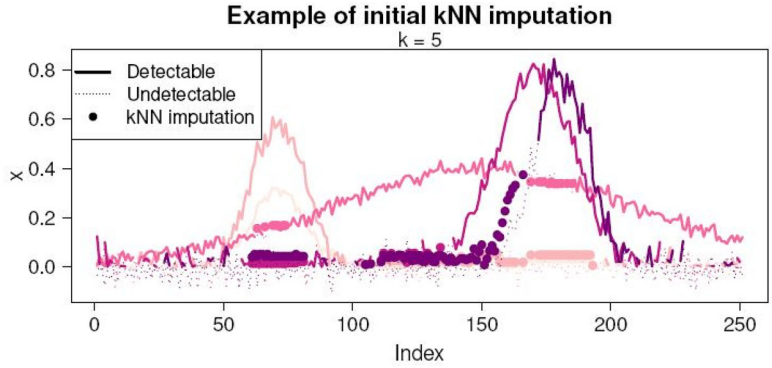

Figure 5. Example of *k*NN starting values. The lines represent the different censored covariates in one dataset, and the points represent the *k*NN initial values for the censored values chosen for iterative imputation. For the indices where no points are present, the algorithm has chosen to skip iterative imputation due to lack of information.

In the figure and throughout this article, the Nonselective Multiple Imputation will be referred to as NMI, the Selective Multiple Imputation using *k*NN starting values as SMI-*k*NN and the approach where all imputations are kept set to $\boldsymbol{L}$ as *NI*. One can see that imputing all censored values does not, in fact, necessarily yield the lowest MSE, supporting the approach to avoid imputations with few neighbors according to our custom *k*NN selection or skip heavily censored observations.

To investigate the impact of the modeling order in Equation (3), a comparison between three strategies was performed; modeling the features by decreasing and increasing level of censoring as well as in a random order. The comparison showed no significant difference in MSE, therefore, random order has been used in the proceeding analyses for execution speed purposes. An investigation of the impact of the choice of distribution for $p(\boldsymbol{x}^q)$ was also conducted. The analyses did not show significant difference in predictive performance between using $p(\boldsymbol{x}^q) = \delta\left(E[\boldsymbol{x}^q]\right)$, an empirical distribution over the observed $\boldsymbol{x}^q$ and $p(\boldsymbol{x}^q) \sim U\left(\min(\boldsymbol{x}^q), \max(\boldsymbol{x}^q)\right)$. Therefore, the fastest strategy, $p(\boldsymbol{x}^q) = \delta\left(E[\boldsymbol{x}^q]\right)$, has been used in the further analyses. Both analyses can be found in the supplementary materials.

In Figure 4, the predictive performance of our *k*NN approach is compared to imputation of missing values selected at random. The plot confirms that our *k*NN strategy manages to choose and impute censored values which aid prediction as it achieves an MSE which is between 60% and 80% lower than if the values to impute are chosen at random.

Results from the first imputation using Equation (9) with an appropriate number of neighbors ($k = 5$) are demonstrated in Figure 5. For clarity purposes only five covariates are plotted. The lines again represent the censored covariates and the points are the starting values for the values chosen by *k*NN. It can be observed that most starting values found by our *k*NN approach are better than using the lower limit of detection. For example, all points for indices after 150 offer reasonable approximations of the underlying functions. One can also further note that our algorithm skips values where there are very few active features, and manages to focus more on imputations for values where there are more similar observations available. For instance, between indices 1 and 50, where there is heavy censoring, the algorithm skips the imputation, while between indices 150 and 160, where more noncensored features are available, the algorithm provides a reasonable imputation. This makes sense since imputations with little available information may result in predictions that are far from the true response.

As it is not possible to compare our approach to Bernhardt, Wang, and Zhang (2015) directly due to the absence of complete cases in our data and the fact that the parametric assumption of covariate distributions does not hold, we choose our baseline comparison models as NMI, NI and *Complete*, a Random Forest model for the complete (uncensored) data. We also present results for the Selective Multiple Imputation approach using $\boldsymbol{L}^d$, which will be referred to as SMI-LD. Results for various n and k can be found in Table 1. We limit the table to these k as values outside this range did not yield better results. The table gives an average of 100 different datasets per data size for which

Table 1. Results for various k for two different sizes of datasets.

Data size	Model	k	Chosen	MSE	s_{MSE}	Speed-up	$s_{speed\text{-}up}$
250 × 6	Complete	–	–	1.000	0.110	–	–
	NI	–	0.000	2.728	0.288	–	–
	NMI	–	1.000	1.447	0.132	1x	0.130x
	SMI-LD	1	0.436	1.598	0.170	3x	0.232x
		5	0.434	1.602	0.172	3x	0.226x
		10	0.391	1.593	0.166	4x	0.220x
		20	0.286	1.646	0.190	5x	0.306x
	SMI-kNN	1	0.436	1.561	0.158	4x	0.239x
		5	0.434	1.577	0.172	4x	0.226x
		10	0.391	1.593	0.174	5x	0.226x
		20	0.286	1.680	0.199	6x	0.289x
500 × 6	Complete	–	–	1.000	0.093	–	–
	NI	–	0.000	3.196	0.264	–	–
	NMI	–	1.000	1.890	0.142	1x	0.098x
	SMI-LD	1	0.395	2.084	0.155	6x	0.146x
		5	0.394	2.084	0.156	6x	0.149x
		10	0.393	2.087	0.157	6x	0.167x
		20	0.360	2.079	0.158	6x	0.175x
	SMI-kNN	1	0.395	2.047	0.151	6x	0.151x
		5	0.394	2.054	0.149	6x	0.149x
		10	0.393	2.043	0.151	6x	0.149x
		20	0.360	2.025	0.158	7x	0.171x

NOTE: Chosen refers to the ratio of censored values chosen for imputation and Speed-up refers to the speed-up relative to NMI. s_{MSE} and $s_{speed\text{-}up}$ refer to the standard deviations of MSE and Speed-up, respectively.

each *iterative* imputation model has been iterated $M = 8$ times and the first three iterations have been removed to account for a burn-in period. Appendix F in the supplementary materials shows that, for four different noise levels, the impact of setting a high M on the predictive performance for these data is very limited. The minimum active ratio per observation required has been set as $o_{\min} = 0.2$. We choose $o_{\min} = 0.2$ as a low $o_{\min}$ enables greater impact from our selective kNN.

In Table 1, *Chosen* refers to the ratio of all censored values chosen for imputation. *Speed-up* refers to the mean execution time speed-up relative to the computational time of NMI. s_{MSE} is the standard deviation for the 100 separate evaluations and $s_{speed\text{-}up}$ is the standard deviation of the speedup relative to the mean speed-up. Table 1 shows that the NI approach yields about three times the MSE as for the complete model, and our suggested approach only 1.5–2 times as high. We can see that imputing all censored observations yields the lowest MSE for both sizes of datasets. Choosing merely between 29% and 44% of the missing values for imputation in the smaller set with kNN gives an MSE that is almost as low as NMI, yet 4–6 times faster. For the SMI-LD, the MSE can get almost as low as NMI for the small set, yet slightly slower for the lowest MSE. For the larger data sizes, the SMI-kNN MSE is very close to the MSE of NMI, for which $k = 20$ gives seven times faster execution time. The SMI-LD also gives a significantly lower MSE than for NI, however, not as low as for SMI-kNN. We can note that the mean MSE for SMI-LD and SMI-kNN differ from the NI mean MSE by more than two standard deviations, and the mean speed-up differ from the NMI execution time with more than two standard deviations, thus, indicating a statistically significant difference between our approaches and the NI for all experiments.

4. Application to Signal Strength Prediction

Wireless network applications demonstrate many use cases where the underlying data are censored. At each geographical location, there may be several signal frequency options for connection. Each frequency does in turn consist of a network of smaller geographical areas, called *cells*, which are available for connection. As users in the network move, they are assigned to the cell in the frequency which gives the best, or most reliable, connection. See a simplified illustration in Figure 6. Within the frequency that a user is connected to, the signal strengths of the surrounding cells are accessible for the user device, allowing for easy assessment of which cell to connect to for optimal reception. Evaluation of the signal strengths of cells on another frequency does, however, require disconnecting from the current frequency and connecting to the alternative frequency to measure the performance, leaving the user without connection for a small window of time. The signals in the network have a lower limit of detection, as weak signals are inaudible. As cells which are located far from the user will naturally fall below the limit of detection, this nature of detectability of the signals results in the absence of complete cases. Furthermore, due to interference between the cells within each frequency, a strong signal can drown out weaker signals, making them inaudible despite being above the lower limit of detection (3GPP 2018).

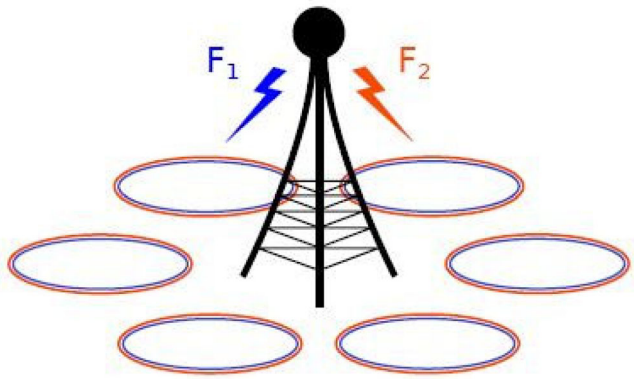

Figure 6. A simple example of a wireless network with a base station transmitting two frequencies to surrounding cells.

As our approach attempts to target the censored values likely to aid prediction, these data constitute an interesting problem as the signals censored due to interference are likely to be of more interest to impute than signals censored due to being out of range.

The data considered in this section are simulated by Ericsson AB, a multinational Swedish networking and telecommunications company, to mimic a real network. They represent simultaneous cell-wise signal strength data of two different frequencies in a geographic area modeled to resemble the wireless network in a typical urban area. One observation consists of q signal strength values for one frequency and the maximum of all available signal strengths for an alternative frequency. All datasets presented are censored in the covariates to around 47% (the censoring level varies somewhat due to the nature of the censoring and the random effects in data).

We consider a regression problem where the aim is, given a connection to a specific frequency, to predict the maximum signal strength on the alternative frequency. We use the cell-wise signal strengths of the current frequency as covariates and the

maximum of an alternative frequency as the dependent variable. We use the maximum in this way as we imagine a scenario where we are interested in the potential gain of switching to the alternative frequency without having to measure if not needed, as in Ryden et al. (2018) and Svahn et al. (2019). All covariates in the scenario are subject to censoring. Due to the assumption that our approach is likely to find values censored by interference more helpful in prediction, we set the initial values for the nonchosen censored values to $\boldsymbol{L} = \boldsymbol{0}$ as the underlying values are likely to fall below that limit. We have set the minimum observed per observation to $o_{\min} = 0.5$ and have limited the iterations to each model to $M = 4$ times to provide faster execution times. Thus, removing the first three runs as a burn-in period, the table aims to show the potential of the method even with small M. Furthermore, the results provided refer to using the same three datasets 10 times, resulting in different outcomes each time since SMI and NMI are randomized algorithms. Thus, the standard deviations aim to show how much the predictive results vary with randomness. Note that since NI and the complete model are deterministic algorithms, all runs give the same predictions and therefore, no standard deviations are reported. The results for three sizes of datasets and various k can be found in Table 2.

The table shows that both SMI approaches achieve statistically significant faster execution, up to 50 times faster than the benchmark algorithm, while still managing to achieve a significantly lower MSE compared to the naïve imputation approach NI. We can see that as k is increased, the execution time goes down. It can also be observed that, for the smallest dataset, selectively imputing with $k = 40$ offers a speedup of 48 times versus 7 times for SMI-kNN, yet, the MSE does not appear to be notably affected. While the MSE for our suggested approach are about 20%–30% higher than for the complete data model, all MSE are significantly lower than for the NI approach except for the values in italic. We can furthermore see that the SMI approaches reach an MSE rather close to the NMI approach for the 500×36 data for suitable k values. For the two larger data sizes, the mean MSE decreases as k increases, hitting the lowest MSE for high k, with both faster execution time and lower share of values chosen for imputation. One can further note that SMI-kNN and SMI-LD appear to outperform NMI in terms of MSE for high k on the $n = 750$ datasets since the SMI-LD and SMI-kNN achieve significantly lower MSE than NI while the difference between NMI and NI cannot be statistically established. The differences in MSE and speed-up for SMI-LD and SMI-kNN are not substantial for these data, as they offer similar results. For the smallest sets, SMI-kNN appear to yield a slightly lower MSE for most k, however, the difference is not statistically significant.

Table 2. Results for various k for three different sizes of datasets.

Data	Model	k	Chosen	MSE	s_{MSE}	Speed-up	$s_{\text{speed-up}}$
500×36	Complete	—	–	1.000	–	–	–
	NI	—	0.000	1.407	–	–	–
	NMI	—	1.000	1.282	0.015	1x	0.052x
	SMI-LD	1	0.333	1.318	0.011	8x	0.041x
		5	0.204	1.317	0.009	11x	0.073x
		10	0.098	1.326	0.007	18x	0.044x
		20	0.046	1.332	0.010	28x	0.061x
		40	0.015	1.320	0.006	50x	0.029x
	SMI-kNN	1	0.333	1.313	0.010	7x	0.064x
		5	0.204	1.308	0.008	11x	0.069x
		10	0.098	1.319	0.009	17x	0.035x
		20	0.046	1.330	0.008	29x	0.054x
		40	0.015	1.323	0.007	48x	0.111x
750×36	Complete	—	–	1.000	–	–	–
	NI	—	0.000	1.333	–	–	–
	NMI	—	1.000	*1.306*	0.015	1x	0.026x
	SMI-LD	5	0.327	*1.339*	0.008	6x	0.045x
		10	0.234	*1.328*	0.006	7x	0.031x
		20	0.122	1.313	0.006	11x	0.024x
		40	0.059	1.292	0.005	17x	0.036x
		60	0.035	1.299	0.007	22x	0.042x
	SMI-kNN	5	0.327	*1.344*	0.006	6x	0.041x
		10	0.234	*1.334*	0.006	7x	0.031x
		20	0.122	1.315	0.009	10x	0.027x
		40	0.059	1.292	0.005	17x	0.027x
		60	0.035	1.297	0.005	22x	0.038x
1000×37	Complete	—	–	1.000	–	–	–
	NI	—	0.000	1.288	–	–	–
	NMI	—	1.000	1.196	0.020	1x	0.046x
	SMI-LD	20	0.151	*1.279*	0.007	7x	0.072x
		40	0.055	1.262	0.005	11x	0.056x
		60	0.031	1.252	0.005	16x	0.077x
		80	0.021	1.245	0.003	19x	0.108x
		100	0.017	1.246	0.004	23x	0.105x
	SMI-kNN	20	0.151	*1.279*	0.006	7x	0.107x
		40	0.055	1.267	0.009	11x	0.061x
		60	0.031	1.249	0.005	16x	0.048x
		80	0.022	1.245	0.006	19x	0.119x
		100	0.017	1.248	0.005	24x	0.151x

NOTE: *Chosen* refers to the ratio of censored values chosen for iterative imputation and *Speed-up* refers to the speed-up relative to NMI. s_{MSE} and $s_{\text{speed-up}}$ refer to the standard deviations of MSE and speed-up, respectively for 10 runs of the same dataset.

5. Discussion and Conclusion

We introduced a new selective imputation approach to speed up imputation compared to the strategy of imputing all censored values, that is, NMI. We have showed that while iteratively imputing all censored values typically yields a statistically significant lower MSE than imputing with the lower limit of detection, our selective approaches drastically reduce the CPU time required while maintaining an MSE quite close to, or even lower, than imputing all censored values. We have showed that the predictive performance is valid for several different noise levels in data. We have demonstrated that, for data simulated to resemble a wireless network, our strategy can reduce the naïve imputation model (NI) MSE from 12.59 to 11.81 while being up to 50 times faster than imputing all censored values which reduce the MSE to 11.45. We have also showed that, for a high number of k, our approach tangent or may even outperform NMI in terms of predictive accuracy.

We have adapted the Bernhardt, Wang, and Zhang (2015) method for multiple imputation and made necessary alterations to accommodate a scenario where there are no complete cases in data and when the parametric assumption for the covariates does not hold.

We have, in addition, demonstrated that for simple artificial data using kNN estimations for the initial imputations can be beneficial. For more complex data, however, the performance of this approach is similar to selecting the values to impute with kNN, yet using $\boldsymbol{L}^d$ as the initial imputations.

We have showed that, since the SMI-kNN outperforms random selection of values to impute, our approach focuses on the most important values to impute, omitting censored values with little information. We have further provided an example

that demonstrates the ability to choose such values and that our selective multiple imputation method using *k*NN can generate better initial imputations for the selected values than other, more naïve methods such as using the lower limit of detection for the initial imputations, speeding up the process additionally. We have showed that while the approach of Bernhardt, Wang, and Zhang (2015) modified for the case of high dimensional data with incomplete covariates may take an extensive amount of time, our approach can handle a high number of covariates even with few iterations. The speed-up of our algorithm was shown to be up to 50 times faster in our studies.

We have investigated and concluded that the modeling order of the covariates in Equation (3) has little to no effect on the predictive potential of the model and that we thereby can benefit, in terms of execution time, from using a random modeling order. We have also showed, for three different strategies for modeling the last covariate in Equation (3), that there is no considerable difference. Furthermore, according to Appendix G, supplementary materials, our method is relatively robust to inclusion of variables unrelated to the response (a known property of random forests). However, the quality of prediction may be affected if these unrelated variables have very different degree of censoring compared to the rest of the data.

As the cases with high dimensional data requires extensive CPU time, the simulation studies have been limited thereof, as our approach is then difficult to compare in a statistical way to the benchmark. While the influence of $o_{\min}$ on the predictive *MSE* have been presented for one dataset, this article does not cover an extensive analysis of this parameter. This article has evaluated a regression scenario, however, the approach can be applied to a classification scenario by adjusting the scaling majorizing constant C in Equation (5) accordingly.

As the presented approaches do not extend to the case when the response is censored, further research for this scenario is needed. Furthermore, the censoring threshold for the nature of censoring in this article have been assumed known, which may be interesting to consider unknown for better generalization.

Supplementary Materials

Additional details: A collection of information regarding the data generation process, diagnostic plots and detailed algorithm descriptions. (pdf)

R-code and data for the SMI algorithms: One of the simple artificial datasets used in the article and R-code to perform the diagnostic methods. (zip)

Acknowledgments

We extend our gratitude to the Associate Editor, the Editor, and the reviewers for their insightful comments, which have contributed to considerable value added to our article.

Funding

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

3GPP. (2018), *Evolved Universal Terrestrial Radio Access (E-UTRA), "Physical layer; Measurements" in Technical Specification (TS) 36.214, 3rd Generation Partnership Project (3GPP)*. Available at *http://www.3gpp.org/ftp//Specs/archive/36_series/36.214/* [7]

Arunajadai, S. G., and Rauh, V. A. (2012), "Handling Covariates Subject to Limits of Detection in Regression," *Environmental and Ecological Statistics*, 19, 369–391. [1]

Bernhardt, P. W., Wang, H. J., and Zhang, D. (2015), "Statistical Methods for Generalized Linear Models with Covariates Subject to Detection Limits," *Statistics in Biosciences*, 7, 68–89. [1,2,3,6,8,9]

Bishop, C. M. (2006), *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Berlin: Springer. [3]

Breiman, L. (2001), "Random Forests," *Machine Learning*, 45, 5–32. [3]

Cover, T., and Hart, P. (1967), "Nearest Neighbor Pattern Classification," *IEEE Transactions on Information Theory*, 13, 21–27. [3]

de Lima Taga, M. F., and Singer, J. M. (2018), "Simple Linear Regression with Interval Censored Dependent and Independent Variables," *Statistical Methods in Medical Research*, 27, 198–207. [1]

Dovis, F. (2015), *GNSS Interference Threats and Countermeasures*, Boston: Artech House. [5]

Gentle, J. (2002), *Elements of Computational Statistics*, Statistics and Computing, New York: Springer. [3]

Gomez, G., Espinal, A., and Lagakos, S. (2003), "Inference for a Linear Regression Model with an Interval-Censored Covariate," *Statistics in medicine*, 22, 409–425. [1]

Hornung, R. W., and Reed, L. D. (1990), "Estimation of Average Concentration in the Presence of Nondetectable Values," *Applied Occupational and Environmental Hygiene*, 5, 46–51. [3]

Hughes, J. P. (1999), "Mixed Effects Models with Censored Data with Application to HIV RNA Levels," *Biometrics*, 55, 625–629. [1]

Jonsson, P., and Wohlin, C. (2004), "An Evaluation of k-Nearest Neighbour Imputation Using Likert Data, in *10th International Symposium on Software Metrics: Proceedings*, pp. 108–118. [3]

Lee, M., Kong, L., and Weissfeld, L. (2012), "Multiple Imputation for Left-Censored Biomarker Data Based on Gibbs Sampling Method," *Statistics in Medicine*, 31, 1838–1848. [1]

Lee, W.-C., Sinha, S. K., Arbuckle, T. E., and Fisher, M. (2018), "Estimation in Generalized Linear Models Under Censored Covariates with an Application to Mirec Data," *Statistics in Medicine*, 37, 4539–4556. [1]

Lyles, R. H., Lyles, C. M., and Taylor, D. J. (2000), "Random Regression Models for Human Immunodeficiency Virus Ribonucleic Acid Data Subject to Left Censoring and Informative Drop-Outs," *Journal of the Royal Statistical Society*, Series C, 49, 485–497. [1]

May, R. C., Ibrahim, J. G., and Chu, H. (2011), "Maximum Likelihood Estimation in Generalized Linear Models with Multiple Covariates Subject to Detection Limits," *Statistics in Medicine*, 30, 2551–2561. [1]

Paxton, W. B., Coombs, R. W., McElrath, M. J., Keefer, M. C., Hughes, J., Sinangil, F., Chernoff, D., Demeter, L., Williams, B., and Corey, L. (1997), "Longitudinal Analysis of Quantitative Virologic Measures in Human Immunodeficiency Virus-Infected Subjects with $\geq$ 400 CD4 Lymphocytes: Implications for Applying Measurements to Individual Patients," *The Journal of Infectious Diseases*, 175, 247–254. [1]

Ryden, H., Berglund, J., Isaksson, M., Cöster, R., and Gunnarsson, F. (2018), "Predicting Strongest Cell on Secondary Carrier Using Primary Carrier Data," in *2018 IEEE Wireless Communications and Networking Conference Workshops*, WCNC 2018 Workshops, Barcelona, Spain, April 15–18, 2018, pp. 137–142, IEEE. [1,8]

Svahn, C., Sysoev, O., Cirkic, M., Gunnarsson, F. and Berglund, J. (2019), "Inter-Frequency Radio Signal Quality Prediction for Handover, Evaluated in 3GPP LTE," in *Proceedings of VTC2019-Spring*, pp. 1–5. [8]

Tsimikas, J. V., Bantis, L. E., and Georgiou, S. D. (2012), "Inference in Generalized Linear Regression Models with a Censored Covariate," *Computational Statistics & Data Analysis*, 56, 1854–1868. [1]

Wu, H., Chen, Q., Ware, L., and Koyama, T. (2012), "A Bayesian Approach for Generalized Linear Models with Explanatory Biomarker Measurement Variables Subject to Detection Limit – An Application to Acute Lung Injury," *Journal of Applied Statistics*, 39, 1733–1747. [1]

Yue, Y. R., and Wang, X.-F. (2016), "Bayesian Inference for Generalized Linear Mixed Models with Predictors Subject to Detection Limits: An Approach that Leverages Information from Auxiliary Variables," *Statistics in Medicine*, 35, 1689–1705. [1]