Regression estimation using surrogate responses obtained by presmoothing

Musta, E.; Patilea, V.; Van Keilegom, I.

DOI
10.1111/stan.12351
Publication date
2025
Document Version
Final published version
Published in
Statistica Neerlandica

Received: 3 June 2024 | Accepted: 22 June 2024

DOI: 10.1111/stan.12351

ORIGINAL ARTICLE

WILEY

Regression estimation using surrogate responses obtained by presmoothing

Eni Musta[1] | **Valentin Patilea**[2] | **Ingrid Van Keilegom**[3]

[1]Korteweg de Vries Institute of Mathematics, University of Amsterdam, Amsterdam, Netherlands

[2]CREST, Ensai, Bruz, France

[3]Research Centre for Operations Research and Statistics, KU Leuven, Leuven, Belgium

Correspondence
Eni Musta, Korteweg de Vries Institute of Mathematics, University of Amsterdam, Amsterdam, The Netherlands.
Email: e.musta@uva.nl

Presmoothing was initially introduced in the linear regression setting as a method to improve finite sample efficiency by replacing the response variable with a nonparametric estimate of the regression function. Since then, it has found success in various domains, including survival analysis. However, the use of presmoothing with multiple continuous covariates is challenging and undesirable in practice. Inspired by the cure regression setup, we derive a simple estimator for (semi)parametric models with many regressors based on 1-dimensional presmoothing. The method is particularly valuable when the response variable is not directly observed. However, even when the response is available, presmoothing can enhance accuracy for small to moderate sample sizes. We present several applications of the proposed method in different settings and investigate its finite sample behavior through simulations.

KEYWORDS

kernel smoothing, cure model, logistic regression, missing response, (semi)parametric regression, surrogate response

1 | INTRODUCTION

Nonparametric regression models primarily serve as a relaxation of parametric models or for validating parametric assumptions. An unconventional application of nonparametric methods is to facilitate parametric estimation via a presmoothing methodology, which involves the introduction of a preliminary smoothing step before the construction of the final parametric estimates. Initially proposed in linear regression, the idea of presmoothing is to replace the response variable with a nonparametric estimate of the regression function and then compute the parameter estimates via least squares (Cristobal, Roca, & González, 1987). In this setting, the approach has been shown to reduce the second-order term of the asymptotic variance with respect to the least squares estimators, consequently resulting in a smaller mean squared error (Faraldo Roca & González, 1987; Janssen, Swanepoel, & Veraverbeke, 2001). The intuition is that, replacing the response variable by a smooth nonparametric estimator, filters out some of the error, and gets closer to the parametric regression function.

Later on, the idea of matching a parametric model to a nonparametric estimator via the least squares principle, was used in the context of linear or polynomial regression models with incomplete observations due to censoring or truncation. González Manteiga and Cadarso-Suárez (1991) introduced a methodology for random right-censored data, relying on prior nonparametric estimation of the regression function and compared it with alternatives utilizing weighted least squares or least-squares based on a weighted response. Apart from the computational advantage of avoiding an iterative procedure, they show through a simulation study that presmoothing performs generally better in practice in terms of mean squared error. Consistency and asymptotic normality of such estimators for censored and truncated data were further studied in Akritas (1996). The authors argue that the idea of using nonparametric techniques for fitting parametric models is widely applicable and has both computational and theoretical advantages over existing methods for censored and truncated data.

Since then, presmoothing has been successfully applied in different domains. It has been shown to improve model selection by Akaike's Information Criterion for finite samples since, replacing the observed response by the estimated regression curve values, increases stability of the model selection with respect to data perturbations (Aerts, Hens, & Simonoff, 2010). Furthermore, presmoothing has been employed in functional linear regression as a modification of the functional principal component analysis for parameter estimation (Ferraty, González Manteiga, Martínez-Calvo, & Vieu, 2012), leading to improved estimators for small sample sizes, while exhibiting the same behavior as the FPCA estimator for large n. More recently, presmoothing has been applied for instrumental variable estimation of the Cox proportional hazards model (Tedesco, Beyhum, & Van Keilegom, 2023). In this context, the nonparametric estimator is used to generate "proxy" observations for which exogeneity holds and the standard partial likelihood estimator can be applied.

On a different context, the term presmoothing has been used to indicate estimation methods that replace the observed response, such as the censoring indicator in survival analysis, by a nonparametric estimate of its conditional expectation (Cao, López-de Ullibarri, Janssen, & Veraverbeke, 2005; Jácome, Gijbels, & Cao, 2008; Jácome & Iglesias-Pérez, 2008). The difference is that this replacement is done within standard nonparametric estimates, for example, the Kaplan–Meier estimator, and is not used for fitting a parametric model as ultimate objective. Nevertheless, these papers also show that presmoothing leads to gain in mean squared error and would also be useful when censoring indicators are missing at random (MAR).

More recently, estimation methods based on presmoothing have been proposed in the context of semi-parametric mixture cure models in survival analysis, where the population consists of cured (immune to the event of interest) and uncured (susceptible) subjects. Because the cure status is generally unknown for the censored observations, estimation is commonly performed via the Expectation-Maximization algorithm. In this setting, the presmoothing approach proposed in Musta, Patilea, and Van Keilegom (2022) offers a possibility for direct estimation of the parametric incidence component, eliminating the need for iterative procedures. In addition, it avoids reliance on latency estimation, which increases its robustness against potential model misspecifications.

The main challenge the presmoothing method encounters in practice is dealing with many continuous covariates. One could use multi-dimensional smoothing but this faces limitations due to the curse of dimensionality. Theoretically, higher-order kernels might be necessary to ensure consistency and in practice multiple bandwidths or tuning parameters need to be chosen, which can be problematic and computationally intensive. To address these issues, a novel approach in the context of mixture cure models involves using presmoothing as a second-stage estimator, once the covariate dimension is reduced to one through a preliminary available estimator (Musta, Patilea, & Van Keilegom, 2024). This solution retains the advantages of presmoothing, including lower mean squared error and more stable estimators, but relies on a preliminary consistent estimate, which is not always available.

Motivated mainly by the success of presmoothing in the linear regression setting and the mixture cure model, this article introduces a general estimation approach for a parametric component of regression models that relies on one-dimensional presmoothing independently of the covariate dimension. This offers an alternative solution to the dimensionality issue of presmoothing, which differently from the approach in Musta et al. (2024), is a direct one-step estimation procedure not requiring any preliminary estimator. The method is particularly useful in settings where the response variable is not directly available, but one can estimate its conditional expectation nonparametrically. Moreover, even when the response is observed, it might carry a lot of variability and presmoothing might still provide more accurate estimators for small and moderate sample sizes. The drawback in the context of mixture cure models is that the proposed method can only be applied under the independent censoring assumption. The article is organized as follows. We start by introducing the general paradigm in Section 2 together with its theoretical foundation. In Section 3, we present several applications of the proposed methodology with more emphasis on the mixture cure model. Finite sample behavior of the method for three of the application settings is investigated in Section 4, with additional simulation results provided in the Appendix C. We conclude with a discussion on the proposed methodology in Section 5.

2 | THE GENERAL PARADIGM

Consider estimation of the conditional mean of a real-valued random variable Y given a set of covariates X, random vector in $\mathbb{R}^d$. Y can be of any nature, for example, continuous, discrete, binary. We focus on the parametric single index model

$$\mathbb{E}[Y|X] = \phi(\gamma_0^T X), \tag{1}$$

where $\gamma_0 \in G \subset \mathbb{R}^d$ is unknown and $\phi : \mathbb{R} \to \mathbb{R}$ is a known function. We assume the following identifiability condition.

Assumption 1. For any $\gamma \in G$, if $\phi(\gamma^T X) = \phi(\gamma_0^T X)$ almost surely, then $\gamma = \gamma_0$.

Assumption 1 implies the uniqueness of γ_0 in (1). This condition is satisfied if the function ϕ is injective on the codomain of $\gamma^T X$ for $\gamma \in G$ and the covariance matrix of X has full rank. Traditionally, the parameter γ_0 is estimated using directly the response variable Y via the maximum likelihood or the least-squares method. However, there are situations in which one cannot or prefers not to use the response variable Y, for example if it is not directly available or carries a lot of variability. Next we describe how, in such cases, the statistician could use, instead of Y, some transformation of the observations that can often be easily constructed based on one-dimensional kernel smoothing.

Let $T(Q;\eta_0)$ be a transformation of the observed or latent variables Q, where η_0 is a possibly infinite dimensional nuisance parameter that can be estimated in a preliminary step, such that $\mathbb{E}[T(Q;\eta_0)|X] = \mathbb{E}[Y|X]$. In the third example that we discuss below, $T(Q;\eta_0)$ is given by a transformation of the observed data where $\eta_0 : \mathbb{R}^d \times \{0,1\} \mapsto \mathbb{R}$ is a nuisance infinite dimensional parameter (see Section 3.3). In the other examples, $T(Q;\eta_0)$ is the outcome Y itself, which might be observed or not, and there is no need for an additional parameter η_0. For each $\gamma \in \mathbb{R}^d$ and η in the same parameter space as η_0, define

$$r(u;\gamma,\eta) = \mathbb{E}\left[T(Q;\eta)|\gamma^T X = u\right]. \tag{2}$$

Note that, by construction, we have $r(u;\gamma_0,\eta_0) = \mathbb{E}\left[\mathbb{E}[Y|X]|\gamma_0^T X = u\right] = \phi(u)$. We assume that, for each γ, $r(\cdot;\gamma,\eta_0)$ can be estimated consistently from the data by some estimator $\hat{r}_n(\cdot;\gamma,\hat{\eta}_n)$. Then we can estimate the parameter γ_0 by the maximum likelihood or the least squares criterion replacing Y by $\hat{r}_n(\gamma^T X;\gamma,\hat{\eta}_n)$. Several choices for the estimators $\hat{r}_n$ and $\hat{\eta}_n$ might be possible. We discuss some of them in the specific application settings in Section 3. This approach is based on the following fundamental property.

Proposition 1. *If the parametric single index model (1) satisfies the identifiability Assumption 1, then*

$$\gamma_0 = \operatorname{argmax}_{\gamma\in G}\mathbb{E}\left[U(\phi(\gamma^T X)) + r(\gamma^T X;\gamma,\eta_0)V(\phi(\gamma^T X))\right],$$

where the differentiable functions $U, V : \mathbb{R} \to \mathbb{R}$ satisfy the identity $U'(y) + V'(y)y = 0$ for all $y \in \mathbb{R}$ and $V'(\cdot) > 0$.

Proof. First, note that because of the assumptions on U and V the function $g(y) = U(y) + zV(y)$ has a unique maximum at z. Indeed we have $g'(y) = U'(y) + zV'(y)$ which, by the assumptions on U and V, implies that $g'(z) = U'(z) + zV'(z) = 0$, $g'(y) > U'(y) + yV'(y) = 0$ for $y > z$ and $g'(y) < U'(y) + yV'(y) = 0$ for $y < z$. From this and the identifiability Assumption 1, we obtain for any $\gamma \in G$, $\gamma \neq \gamma_0$,

$$\begin{aligned}\mathbb{E}&\left[U(\phi(\gamma^T X)) + T(Q;\eta_0)V(\phi(\gamma^T X))|X\right] = U(\phi(\gamma^T X)) + \phi(\gamma_0^T X)V(\phi(\gamma^T X))\\ &\leq U(\phi(\gamma_0^T X)) + \phi(\gamma_0^T X)V(\phi(\gamma_0^T X)) = \mathbb{E}\left[U(\phi(\gamma_0^T X)) + T(Q;\eta_0)V(\phi(\gamma_0^T X))|X\right] \text{ a.s.,}\end{aligned}$$

and the inequality is strict on a set of positive probability because $\mathbb{P}(\phi(\gamma^T X) = \phi(\gamma_0^T X)) < 1$. From the definition of $r(\gamma^T X;\gamma,\eta_0)$ in (2) and the properties of the conditional expectation, it follows that

$$\mathbb{E}\left[U(\phi(\gamma^T X)) + r(\gamma^T X;\gamma,\eta_0)V(\phi(\gamma^T X))\right] = \mathbb{E}\left[\mathbb{E}\left[U(\phi(\gamma^T X)) + T(Q;\eta_0)V(\phi(\gamma^T X))|\gamma^T X\right]\right]$$

$$= \mathbb{E}\left[\mathbb{E}\left[U(\phi(\gamma^T X)) + T(Q;\eta_0)V(\phi(\gamma^T X))|X\right]\right] < \mathbb{E}\left[\mathbb{E}\left[U(\phi(\gamma_0^T X)) + T(Q;\eta_0)V(\phi(\gamma_0^T X))|X\right]\right]$$
$$= \mathbb{E}\left[\mathbb{E}\left[U(\phi(\gamma_0^T X)) + T(Q;\eta_0)V(\phi(\gamma_0^T X))|\gamma_0^T X\right]\right] = \mathbb{E}\left[U(\phi(\gamma_0^T X)) + r(\gamma_0^T X;\gamma_0,\eta_0)V(\phi(\gamma_0^T X))\right],$$

which means that $\mathbb{E}\left[U(\phi(\gamma^T X)) + r(\gamma^T X;\gamma,\eta_0)V(\phi(\gamma^T X))\right]$ attains its maximum at γ_0. ■

Remark 1. If the distribution of Y given X belongs to the linear exponential family, its conditional density can be written as $f(y|\gamma) = \exp(C(\mu)y + A(\mu) + B(y))$ for $\mu = \phi(\gamma^T X)$ and some univariate real valued functions A, B and C satisfying $A'(s) = C'(s)s$ (see e.g., section 1.3 in Ziegler, 2011). The contribution of one observation to the log-likelihood is $C(\phi(\gamma^T X))y + A(\phi(\gamma^T X))$. Hence, the criterion in Proposition 1 corresponds to the maximum likelihood with Y replaced by $r(\gamma^T X;\gamma,\eta_0)$. Another example would be the Bernoulli likelihood, that is, if $Y \in \{0,1\}$ with $\mathbb{P}(Y=1|X) = \phi(\gamma_0^T X) \in (0,1)$ then the contribution of one observation to the log-likelihood is given by $U(\phi(\gamma^T X)) + YV(\phi(\gamma^T X))$, where $U(y) = \log(1-y)$ and $V(y) = \log y - \log(1-y)$. In addition, Proposition 1 can also be seen as a least-squares criterion with Y replaced by $r(\gamma^T X;\gamma,\eta_0)$. Indeed, we have $\mathbb{E}\left[(Y - \phi(\gamma^T X))^2\right] = \mathbb{E}\left[U(\phi(\gamma^T X)) + YV(\phi(\gamma^T X))\right]$, where $U(y) = -y^2$ and $V(y) = 2y$.

Proposition 1 allows us to estimate γ_0 as we would normally do but using $r(\gamma^T X;\gamma,\eta_0)$ instead of Y. Let $\hat{\eta}_n$ be a consistent estimator of η_0 and $\hat{r}_n(\gamma^T X;\gamma;\hat{\eta}_n)$ a consistent estimator of $r(\gamma^T X;\gamma,\eta_0)$. We propose to estimate γ_0 by

$$\hat{\gamma}_n \in \operatorname{argmax}_{\gamma\in G} \frac{1}{n}\sum_{i=1}^{n}\left\{U(\phi(\gamma^T X_i)) + \hat{r}_n(\gamma^T X_i;\gamma,\hat{\eta}_n)V(\phi(\gamma^T X_i))\right\}, \tag{3}$$

for some functions U and V that satisfy the conditions of Proposition 1. Estimation of $r(\cdot;\gamma,\eta)$ can be done via standard procedures of conditional regression estimation, depending on the problem at hand, for example, based on kernel smoothing. Since, for a fixed γ we are conditioning on $\gamma^T X$, the procedure requires only one-dimensional smoothing independently of the dimension of the covariate vector X.

In terms of asymptotic properties, we expect that one can derive consistency and asymptotic normality under standard assumptions using general results for M-estimators in the presence of a nuisance possibly infinite dimensional parameter as in Delsol and Van Keilegom (2020). We provide some general ideas for the proof of consistency in Appendix B. However derivations are quite technical and depend on the specific setting and the nonparametric estimator $\hat{r}_n(\cdot;\gamma,\hat{\eta}_n)$ at hand. Hence we do not investigate this further in details in the present paper.

3 | APPLICATIONS

In this section, we describe several possible applications of the general approach proposed in Section 2. The first one, the mixture cure model, is of particular interest since the response variable in this setting is not always observed and the one-dimensional presmoothing approach has several advantages over other existing methodologies. The second application is the standard logistic model where the response variable is observed but presmoothing can still lead to enhanced accuracy of the estimators for finite sample sizes by reducing the variance. The third application is about a setting where the response variable might be missing but there is an available surrogate response. Again presmoothing can still help in improving accuracy of the estimators for small/moderate samples. Asymptotically we expect presmoothing to behave the same as the standard methods. Note also that for the first two applications we use Proposition 1 with a maximum likelihood-type criterion, while for the third one we use a least-squares type criterion. Other two applications related to survival analysis are described in the Appendix A since they are similar to the mixture cure model. One of them is about estimating probability of default in credit scoring by fitting a logistic model in presence of censored data. The other is about using a logistic model for censored data in the presence of competing risks, which has been considered in Blanche, Holt, and Scheike (2023).

3.1 | Mixture cure model

We are interested in a lifetime $T \in (0, \infty]$, which can possibly be infinite, indicating that a fraction of the population will never experience the event of interest. We refer to those subjects as "cured." Let $C \in (0, \infty)$ denote a random censoring variable. The observations consist of the observed survival time $Z = \min(T, C)$, the censoring indicator $\Delta = \mathbb{1}_{\{T \leq C\}}$ and two vectors of covariates $X \in \mathbb{R}^d$, $W \in \mathbb{R}^p$, which correspond to the factors that affect the probability of being cured and the survival times of the uncured subjects. X and W can be different or share some of the components. We consider a semi-parametric mixture cure model and assume that the conditional survival function of T is given by

$$S(t|x, w) = \mathbb{P}(T > t|X = x, W = w) = 1 - \phi(\gamma_0^T x) + \phi(\gamma_0^T x)S_u(t|w),$$

where $S_u(\cdot|w)$ is the conditional survival function of the uncured, $\phi(\gamma_0^T x)$ denotes the probability of being noncured. Here $\phi : \mathbb{R} \to (0, 1)$ is a known function, $\gamma_0 \in \mathbb{R}^{d+1}$, the first component of X is taken to be equal to one and the first component of γ_0 corresponds to the intercept. Common choices for the latency distribution S_u are the Cox proportional hazards model or the accelerated failure time model. However, here we focus on estimation of γ_0 and do not need to make assumptions on S_u, hence leave the latency model unspecified. In addition, we also do not need to know the covariates W. For the incidence, the most common model is the logistic (see e.g., Patilea & Van Keilegom, 2020; Amico & Van Keilegom, 2018; Yilmaz, Lawless, Andrulis, & Bull, 2013; Wycinka & Jurkiewicz, 2017)

$$\phi(u) = e^u/(1 + e^u), \tag{4}$$

for which the identifiability Assumption 1 is satisfied when the covariance matrix of X has full rank.

What makes estimation of mixture cure models challenging is the fact that for the censored observations it is not known whether they are cured or they experience the event at a later point. In presence of model assumptions for the latency, estimation is usually performed via the expectation maximization algorithm (Peng and Dear, 2000; Sy and Taylor, 2000; Li and Taylor, 2002; Zhang and Peng, 2007). Problems related to this iterative simultaneous estimation of both incidence and latency are discussed in Musta et al. (2022, 2024). In addition, misspecification of latency model would induce bias even in the estimation of the incidence. On the other hand, the method proposed in Musta et al. (2022) allows for direct estimation of γ_0 independently of the latency model by constructing first a nonparametric estimator for the cure probability and then projecting it to the desired parametric class. Here we extend this idea further by constructing the nonparametric estimator based only on one-dimensional smoothing.

Focusing on the incidence component, the outcome of interest is the cure status $Y = \mathbb{1}_{\{T<\infty\}}$ with $\mathbb{E}[Y|X] = \phi(\gamma_0^T X)$. If Y was observed, one could use the maximum likelihood method to estimate γ_0. Instead of directly using Y, we want to apply the general paradigm described in Section 2 and use Proposition 1 with a Bernoulli-type likelihood. Usually one assumes that the censoring is independent of the survival time conditionally on the covariates. Here, since we construct the nonparametric estimator based on $\gamma^T X$, we need to assume that $C \perp T | \gamma^T X$, that is, C is independent of T given $\gamma^T X$, for any $\gamma \in G$. In addition to Assumption 1, a typical assumption needed to identify the mixture cure model is the sufficiently long follow-up Patilea and Van Keilegom, (2020); Musta et al. (2022), that is,

$$\{T > \tau_0(W)\} = \{T = \infty\} \quad \text{and} \quad \inf_w \mathbb{P}(C > \tau_0(W)|W = w) > 0, \tag{5}$$

where τ_0 is the end point of the support of S_u and might depend on the covariates W. For $u \in \mathbb{R}$ and $\gamma \in G$ define

$$r(u; \gamma, \eta_0) = r(u; \gamma) = \mathbb{E}\left[Y|\gamma^T X = u\right] = 1 - \mathbb{P}(T = \infty|\gamma^T X = u).$$

Note that in this case $T(Q; \eta_0)$ is the outcome Y itself so there is no need for an additional parameter η_0. In particular, we have $r(u; \gamma_0) = \mathbb{E}\left[\mathbb{E}[Y|X]|\gamma_0^T X = u\right] = \phi(u)$. Taking $U(y) = \log(1 - y)$ and $V(y) = \log y - \log(1 - y)$ in Proposition 1, we obtain the following result.

Corollary 1. *Under the identifiability conditions given in Assumption 1 and (5), we have*

$$\gamma_0 = \text{argmax}_{\gamma \in G} \mathbb{E}\left[\{1 - r(\gamma^T X; \gamma)\} \log\{1 - \phi(\gamma^T X)\} + r(\gamma^T X; \gamma) \log \phi(\gamma^T X)\right].$$

Under independent censoring and the identifiability condition (5), as in section 3.1 of Musta et al. (2024), we have

$$r(u; \gamma) = 1 - \prod_{t \in \mathbb{R}} \left\{1 - \frac{H_{1,\gamma}(\mathrm{d}t|u)}{H_\gamma([t, \infty)|u)}\right\},$$

where $H_{l,\gamma}([t, \infty)|u) = \mathbb{P}\left(Z \geq t, \Delta = l|\gamma^T X = u\right)$ for $l = 0, 1$ and $H_\gamma([t, \infty)|u) = H_{1,\gamma}([t, \infty)|u) + H_{0,\gamma}([t, \infty)|u)$. Hence, we can estimate $r(\cdot; \gamma)$ by using a Beran-type estimator of the conditional survival function given the index $\gamma^T X$, computed at the largest observed event time

$$\hat{r}_n(u; \gamma) = 1 - \prod_{t \in \mathbb{R}} \left(1 - \frac{\hat{H}_{1,\gamma}(\mathrm{d}t|u)}{\hat{H}_\gamma([t, \infty)|u)}\right), \tag{6}$$

where $\hat{H}_\gamma([t, \infty)|u) = \hat{H}_{1,\gamma}([t, \infty)|u) + \hat{H}_{0,\gamma}([t, \infty)|u)$, $\hat{H}_{1,\gamma}(\mathrm{d}t|u) = \hat{H}_{1,\gamma}((t - \mathrm{d}t, t]|u)$ for small $\mathrm{d}t$ and

$$\hat{H}_{l,\gamma}([t, \infty)|u) = \sum_{i=1}^n \frac{k_b\left(\gamma^T X_i - u\right)}{\sum_{j=1}^n k_b\left(\gamma^T X_j - u\right)} \mathbb{1}_{\{Y_i \geq t, \Delta_i = l\}}, \quad l = 0, 1.$$

Here k is a one-dimensional kernel function, $b = b_n$ is a bandwidth sequence and $k_b(\cdot) = k(\cdot/b)/b$. Note that the knowledge of $\tau_0(w)$ is not required for estimation of $r(\cdot;\gamma)$ in (6) and the computation of such estimator only consists in a finite product over t in the set of observed event times. Following the proposal in (3) with U and V as in Corollary 1, we define $\hat{\gamma}_n$ as a maximizer over γ of

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\hat{r}_n(\gamma^T X_i,\gamma)\log\phi(\gamma^T X_i) + \left[1-\hat{r}_n(\gamma^T X_i;\gamma)\right]\log\left[1-\phi(\gamma^T X_i)\right]\right\}.$$

3.2 | Logistic regression

Let $X \in \mathbb{R}^{d+1}$ be a vector of covariates with first component equal to one and let $Y \in \{0,1\}$ be the response variable. Assuming the logistic model we have $\mathbb{E}[Y|X=x] = \mathbb{P}(Y=1|X=x) = \phi(\gamma_0^T x)$, for some unknown $\gamma_0 \in \mathbb{R}^{d+1}$ and ϕ as in (4). We are interested in estimation of γ_0 on the basis of n i.i.d. observations $(Y_1, X_1), \ldots, (Y_n, X_n)$. We assume that $d < n$ and the covariance matrix of X has full rank as in the standard logistic regression setting. Hence, the identifiability Assumption 1 is satisfied. This is a very simple model, widely investigated in the literature and the standard estimator is the maximum likelihood estimator

$$\tilde{\gamma}_n = \operatorname{argmax}_\gamma \sum_{i=1}^{n}\left\{Y_i\log\phi(\gamma^T X_i) + (1-Y_i)\log(1-\phi(\gamma^T X_i))\right\}. \tag{7}$$

However, we show in the simulation study that the general method proposed in Section 2 usually leads to better results for small and moderate sample size. Following the idea described in Section 2, we consider $T(Q,\eta_0) = Y$, i.e. again there is no need for an additional parameter η_0, and define for any $\gamma \in \mathbb{R}^p$, $r(u,\gamma) = \mathbb{E}(Y|\gamma^T X = u)$. Considering again $U(y) = \log(1-y)$ and $V(y) = \log y - \log(1-y)$ as in Corollary 1 we obtain

$$\gamma_0 = \operatorname{argmax}_{\gamma\in G}\mathbb{E}\left[r(\gamma^T X;\gamma)\log\phi(\gamma^T X) + \{1-r(\gamma^T X;\gamma)\}\log\{1-\phi(\gamma^T X)\}\right].$$

One possible choice would be to estimate $r(\cdot;\gamma)$ nonparametrically by the Nadaraya–Watson kernel estimator:

$$\hat{r}_n(u;\gamma) = \sum_{i=1}^{n}\frac{k_b(\gamma^T X_i - u)}{\sum_{j=1}^{n}k_b(\gamma^T X_j - u)}Y_i,$$

where k is a one-dimensional kernel function, $b = b_n$ is a bandwidth sequence and $k_b(\cdot) = k(\cdot/b)/b$. Then we define the estimator $\hat{\gamma}_n$ as a maximizer of

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\hat{r}_n(\gamma^T X_i;\gamma)\log\phi(\gamma^T X_i) + \left[1-\hat{r}_n(\gamma^T X_i;\gamma)\right]\log\left[1-\phi(\gamma^T X_i)\right]\right\}.$$

Note that, compared to the maximum likelihood method in (7), the only difference is that we are replacing the outcome Y by a smooth nonparametric estimator of the regression function.

The same idea could be used for other parametric regression models, replacing the maximum likelihood method by least-squares according to Remark 1.

3.3 | Missing outcome context with a surrogate response

Let $X \in \mathbb{R}^{d+1}$ be a vector of covariates with first component equal to one and let Y be the response variable from the following regression model $\mathbb{E}[Y|X] = \phi(\gamma_0^T X)$ for some unknown $\gamma_0 \in \mathbb{R}^{d+1}$ and a known function $\phi : \mathbb{R} \to \mathbb{R}$ satisfying the identifiability Assumption 1. We consider the missing data context in which for some subjects the outcome Y is not observed, but instead we observe a surrogate for Y given by S. S can be any summary of Y, for example $S = \mathbb{1}_{\{Y \geq c\}}$ for some constant $c \in \mathbb{R}$. Let $\Delta = 1$ if Y is observed and $\Delta = 0$ if Y is missing. We consider the challenging missing not at random (MNAR) setup in which, given X, the missingness Δ still depends on Y. However, we assume that, when S is available, the MAR assumption holds, that is, $\Delta \perp Y | X, S$. This means that the dependence of Δ and Y, when X is given, is only through S, which contains observable information about Y. In particular we have $\mathbb{E}[\Delta | Y, X, S] = \mathbb{E}[\Delta | X, S]$ and we denote $\eta_0(x, s) = \mathbb{P}(\Delta = 1 | X = x, S = s)$. We are interested in estimation of γ_0 on the basis of n i.i.d. observations $(X_1, \Delta_1 Y_1, S_1), \ldots, (X_n, \Delta_n Y_n, S_n)$. We have

$$\mathbb{E}\left[\frac{\Delta Y}{\eta_0(X,S)}|X\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{\Delta Y}{\eta_0(X,S)}|X,Y,S\right]|X\right] = \mathbb{E}\left[\frac{Y}{\eta_0(X,S)}\mathbb{E}[\Delta|X,Y,S]|X\right] = \mathbb{E}[Y|X] = \phi(\gamma_0^T X).$$

The function $\eta_0(x, s)$ can be estimated by fitting a regression model on Δ conditionally on X, S. Then one can estimate γ_0 via the least squares principle using as outcome $\Delta Y / \hat{\eta}_n(X, S)$ instead of Y, that is, minimize over $\gamma \in G$ the function

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\Delta_i Y_i}{\hat{\eta}_n(X_i,S_i)} - \phi(\gamma^T X_i)\right)^2. \tag{8}$$

On the other hand, the general method described in Section 2 leads to the following estimate. Let $T(Q; \eta_0) = \Delta Y / \eta_0(X, S)$. Note that here $\eta_0(\cdot, \cdot)$ plays the role of an additional nuisance parameter that needs to be estimated beforehand. For $r(u; \gamma, \eta)$ defined as in (2), $U(y) = -y^2$ and $V(y) = 2y$, from Proposition 1 we obtain

$$\gamma_0 = \mathrm{argmax}_{\gamma \in G}\mathbb{E}\left[-\phi(\gamma^T X)^2 + r(\gamma^T X; \gamma, \eta_0)\phi(\gamma^T X)\right].$$

The function $r(u; \gamma, \eta_0)$ can be estimated by fitting a nonparametric regression model on $\Delta Y / \hat{\eta}_n(X, S)$ conditionally on $\gamma^T X$. For example, using the Nadaraya–Watson kernel estimator we obtain

$$\hat{r}_n(u; \gamma, \hat{\eta}_n) = \sum_{i=1}^{n}\frac{k_b(\gamma^T X_i - u)}{\sum_{j=1}^{n} k_b(\gamma^T X_j - u)}\frac{\Delta_i Y_i}{\hat{\eta}_n(X_i,S_i)},$$

where k is a one-dimensional kernel function, $b = b_n$ is a bandwidth sequence and $k_b(\cdot) = k(\cdot / b)/b$. Then we define

$$\hat{\gamma}_n \in \operatorname{argmax}_{\gamma \in G} \frac{1}{n} \sum_{i=1}^{n} \left\{ -\phi(\gamma^T X_i)^2 + 2\phi(\gamma^T X_i)\hat{r}_n(\gamma^T X_i; \gamma, \hat{\eta}_n) \right\}, \tag{9}$$

which is an estimator that corresponds to minimizing the least squares criterion with $\Delta Y / \hat{\eta}_n(X, S)$ replaced by the smooth nonparametric estimator $\hat{r}_n(\gamma^T X; \gamma, \hat{\eta})$.

4 | SIMULATION STUDY

In this section, we investigate the finite sample performance of the estimation method proposed in Section 2 for the three models described in Section 3. A summary of the main results is presented here while the rest is provided in the Appendix C. In all the settings we compute the nonparametric estimator using the Epanechnikov kernel function $k(u) = (3/4)(1 - u^2)\mathbb{1}_{\{|u| \leq 1\}}$. The standard cross-validation bandwidth implemented in the R package np did not perform very well in this case. Hence, for simplicity, we use a fixed bandwidth $b = cn^{-2/7}$, the order of which is motivated by the theoretical results in Musta et al. (2022), after standardizing the index $\gamma^T X$. We compared several choices of the constant c and found that for the cure model $c = 3$ performed the best overall, while for the other two models $c = 2.5$ seemed to be a good choice. In practice, one could investigate some bootstrap bandwidth selection method but that increases the computational cost and is out of the scope of the current paper. In addition we also encountered problems with local maxima of the criterion functions. Hence, for the optimization procedure we use the Nelder–Mead method with 10 random initializations, which seemed to perform better. One could increase the number of initializations to ensure identification of the global maximum but for computational reasons we restrict to 10.

4.1 | Mixture cure model

We consider the mixture cure model as described in Section 3.1 and the following scenarios.

Model 1. Both incidence and latency component depend on five independent continuous covariates: $X_1 = W_1$ and $X_2 = W_2$ are uniformly distributed on $(-1, 1)$, $X_3 = W_3$ and $X_4 = W_4$ have a standard normal distribution, $X_5 = W_5$ has standard exponential distribution. We generate the cure status B as a Bernoulli random variable with success probability $\phi(\gamma^T X)$ where ϕ is the logistic function in (4) and $\gamma = (0.5, 1, 1, 1, 1, 1)$. The survival times for the uncured observations are generated according to a Weibull law $S_u(t|w) = \exp(-\mu t^{\rho(w)} \exp(\beta^T w))$, for $\rho(w) = 0.75 + \exp(4\beta^T w)$, $\mu = 1.5$ and $\beta = (0.5, 0.4, 0.3, 0.7, 0.4)$ and are truncated at $\tau_0(w)$ being the 97%-quantile of the distribution $F_u(\cdot|w)$. The censoring times are independent from X and T. They are generated from the exponential distribution with parameter 0.15. In this scenario the cure rate is 28% and the censoring rate is 37%.

Model 2. Both incidence and latency component depend on five independent covariates: $X_1 = W_1$ and $X_2 = W_2$ are uniformly distributed on $(-1, 1)$, $X_3 = W_3$ has a standard normal distribution, $X_4 = W_4$ is a binomial random variable with parameters 3 and 0.3, $X_5 = W_5$ is a Bernoully random variable with parameter 0.2. The cure status and the survival times for the uncured observations are generated as in Model 1 with $\gamma = (1, 1, 1, 1, -1, 1)$, $\rho(w) = 0.75 + \exp(4\beta^T w)$, $\mu = 1.5$ and $\beta = (0.5, 0.4, 0.3, -0.7, 0.4)$. The censoring times are generated from the exponential distribution with parameter 0.2 independently of T and X. The cure and censoring rates are around 45% and 55%.

We consider sample sizes $n = 200{,}400$. The proposed method is compared to the maximum likelihood estimator from the `smcure` package, which wrongly assumes that the model is logistic-Cox, and with the multidimensional presmoothing estimator (MPvK) proposed in Musta et al. (2022). Bias, *SD* and root mean-squared-error (MSE) over 1,000 iterations are reported in Table C1. In addition, for Model 1 we also present the boxplots of the estimates in Figure C1. We observe that the `smcure` estimator is usually biased, which is to be expected since the latency model is misspecified. The MPvK estimator has a large bias when there are multiple continuous covariates, while the new approach performs better in terms of bias overall. In terms of MSE, the new approach is better than `smcure` and comparable or better than the MPvK estimator. For settings with lower-dimensional covariates, we see in Table C2 in the Appendix C that the MPvK estimator usually performs better in terms of MSE despite sometimes having larger bias. However, as the sample size increases the bias starts to dominate the MSE indicating that the new approach might be more accurate for larger n.

4.2 | Logistic regression

We consider the standard logistic regression model as described in Section 3.2 and the following scenarios.

Model 1. We have four independent continuous covariates: the first two are generated from the standard normal distribution and the last two are uniformly distributed in $[-1, 1]$. The parameter is $\gamma_0 = (0.5, 1, 1, 1, 1)$.

Model 2. We consider four independent covariates (continuous and discrete): X_1 has a standard normal distribution, X_2 is uniformly distributed in $[-1, 1]$, X_3 has a binomial distribution with parameters 3 and 0.3, X_4 has a Bernoulli distribution with parameter 0.4. The parameter is $\gamma_0 = (0.5, 1, 1, 0.5, -1)$.

Model 3. We consider 10 independent continuous covariates, all with a standard normal distribution. The parameter is $\gamma_0 = (0.5, 1, 0.8, -2, -1, 1.5, 1, -2, -0.6, 1.3, 1.8)$.

Model 4. We have three independent covariates with standard normal distribution and $\gamma_0 = (4, 1.5, -1, -2)$. This scenario corresponds to an unbalanced case with around 90% of the observations with $Y = 1$.

We consider sample sizes $n = 100{,}200$ for the first two models, $n = 200{,}400$ for the last two models. The proposed method is compared to the standard maximum likelihood estimator. For Model 4, we also compare with two methods proposed in Puhr, Heinze, Nold, Lusa, and Geroldinger (2017), which are modifications of Firth's logistic regression for bias correction in the imbalanced case. They are implemented in the `logistf` R-package as FLIC and FLAC methods. Bias, standard deviation and root MSE over 1,000 iterations are reported in Tables C3–C5. We observe that the new approach has a smaller MSE compared to the MLE because of variance reduction, particularly in the higher-dimensional setting. As the sample size increases the two methods become more comparable. Presmoothing seems to work well even in the imbalanced setting, outperforming FLIC and FLAC in terms of mean squared error.

4.3 | Missing outcome context with a surrogate response

We consider the model described in Section 3.3 with a three-dimensional covariate: X_1 is a standard normal random variable, X_2 is uniform on $[-1, 1]$ and X_3 is Bernoulli with parameter 0.6. In Model 1 we take $\phi(y) = y$ which corresponds to the linear regression model and $\gamma_0 = (0.4, 0.5, 0.8, 0.3)$. In Model 2, we take $\phi(y) = y^3$ and $\gamma_0 = (0.5, 0.5, 0.5, 0.5)$. The response variable Y is generated using the model $Y = \phi(\gamma_0^T X) + \epsilon$, where ϵ is independent of X and has standard normal distribution. The surrogate response is $S = \mathbb{1}_{\{Y>1\}}$ and the missingness Δ is generated from a logistic model depending, for simplicity, only on S and not on X

$$\mathbb{P}(\Delta = 0|S) = \eta(\beta_0 + \beta_1 S) = \frac{\exp(\beta_0 + \beta_1 S)}{1 + \exp(\beta_0 + \beta_1 S)},$$

with $\beta = (\beta_0, \beta_1) = (1, -3)$ in Model 1 and $\beta = (0, 1)$ in Model 2. This corresponds to around 50% and 60% missing outcomes, respectively, in Models 1 and 2. Note that we could allow the missingness Δ to also depend on X, which would result in more parameters β to be estimated. However, this setting is not very restrictive since it still means that we have MNAR assumption without S and might be realistic in certain applications. We estimate β via maximum likelihood for logistic regression and consider $\hat{\eta}_n(x, s) = \eta(\hat{\beta}_{0,n} + \hat{\beta}_{1,n} s)$. The bias, *SD*, and root MSE of the estimators of γ_0 via the least-squares method and the new approach, over 1,000 generated datasets of size $n = 100$ and $n = 200$ are given in Table C6. We observe that in general presmoothing leads to smaller variance and MSE and the two methods become more comparable as the sample size increases.

5 | DISCUSSION

In this article, we focus on the use of nonparametric techniques for estimation of a parametric component of regression models, known as presmoothing approach. Despite its success in several settings in terms of variance reduction and stability for small and moderate sample sizes, presmoothing faces challenges related to the course-of-dimensionality. Particularly, the estimates might be biased unless higher-order kernels are used and the choice of multiple bandwidths is computationally intensive and undesirable in practice. To solve this, we provide a simple but general estimation procedure that relies only on one-dimensional smoothing independently of the covariate dimension. The method is particularly useful in settings where the response variable is not observed, such as the mixture cure model. In this case, presmoothing allows for estimation of the cure probabilities without making assumptions on the latency model. The price one has to pay for the one-dimensional smoothing is assuming independent censoring, which is reasonable in case of administrative censoring. In other settings, the response variable or a surrogate thereof might be available but presmoothing is still useful for variance reduction particularly with small sample sizes. Hence, our general methodology can have a wide range of applicability beyond the contexts described in this manuscript. Possible future directions could be investigation of the asymptotic properties for specific settings, design of data-driven bandwidth selection methods and extension to single index models where the link function ϕ is also unknown and needs to be estimated, but in that case an iterative procedure would be required which increases computational cost.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

REFERENCES

Aerts, M., Hens, N., & Simonoff, J. S. (2010). Model selection in regression based on pre-smoothing. *Journal of Applied Statistics*, *37*(9), 1455–1472.

Akritas, M. G. (1996). On the use of nonparametric regression techniques for fitting parametric regression models. *Biometrics*, *52*(4), 1342–1362.

Amico, M., & Van Keilegom, I. (2018). Cure models in survival analysis. *Annual Review of Statistics and Its Application*, *5*, 311–342.

Blanche, P. F., Holt, A., & Scheike, T. (2023). On logistic regression with right censored data, with or without competing risks, and its use for estimating treatment effects. *Lifetime Data Analysis*, *29*(2), 441–482.

Cao, R., López-de Ullibarri, I., Janssen, P., & Veraverbeke, N. (2005). Presmoothed Kaplan–Meier and Nelson–Aalen estimators. *Journal of Nonparametric Statistics*, *17*(1), 31–56.

Cristobal, J. C., Roca, P. F., & González, M. W. (1987). A class of linear regression parameter estimators constructed by nonparametric estimation. *The Annals of Statistics*, *15*(2), 603–609.

Delsol, L., & Van Keilegom, I. (2020). Semiparametric M-estimation with non-smooth criterion functions. *Annals of the Institute of Statistical Mathematics*, *72*(2), 577–605.

Faraldo Roca, P., & González, M. W. (1987). *Efficiency of a new class of linear regression estimates obtained by preliminary non-parametric estimation*. In *New perspectives in theoretical and applied statistics (Bilbao, 1986) Wiley series in probability and mathematical statistics* (pp. 229–242). New York: Wiley.

Ferraty, F., González Manteiga, W., Martínez-Calvo, A., & Vieu, P. (2012). Presmoothing in functional linear regression. *Statistica Sinica*, *22*(1), 69–94.

González Manteiga, W., & Cadarso-Suárez, C. M. (1991). *Linear regression with randomly right-censored data using prior nonparametric estimation*. In *Nonparametric functional estimation and related topics (Spetses, 1990) NATO advanced science institutes series C: Mathematical and physical sciences* (Vol. *335*, pp. 315–328). Dordrecht, The Netherlands: Kluwer Academic Publication.

Jácome, M., & Iglesias-Pérez, M. (2008). Presmoothed estimation with left-truncated and right-censored data. *Communications in Statistics—Theory and Methods*, *37*(18), 2964–2983.

Jácome, M. A., Gijbels, I., & Cao, R. (2008). Comparison of presmoothing methods in kernel density estimation under censoring. *Computational Statistics*, *23*, 381–406.

Janssen, P., Swanepoel, J., & Veraverbeke, N. (2001). Efficiency of linear regression estimators based on presmoothing. *Communications in Statistics-Theory and Methods*, *30*(10), 2079–2097.

Li, C. S., & Taylor, J. M. (2002). A semi-parametric accelerated failure time cure model. *Statistics in Medicine*, *21*(21), 3235–3247.

Musta, E., Patilea, V., & Van Keilegom, I. (2022). A presmoothing approach for estimation in the semiparametric cox mixture cure model. *Bernoulli*, *28*(4), 2689–2715.

Musta, E., Patilea, V., & Van Keilegom, I. (2024). A two-step estimation procedure for semiparametric mixture cure models. *Scandinavian Journal of Statistics*, 1–25.

Patilea, V., & Van Keilegom, I. (2020). A general approach for cure models in survival analysis. *The Annals of Statistics*, *48*(4), 2323–2346.

Peláez Suárez, R., Cao Abad, R., & Vilar Fernández, J. M. (2021). Probability of default estimation in credit risk using a nonparametric approach. *Test*, *30*(2), 383–405.

Peláez Suárez, R., Van Keilegom, I., Cao, R., & Vilar, J. M. (2024). Probability of default estimation in credit risk using mixture cure models. *Computational Statistics & Data Analysis*, *189*, 107853.

Peng, Y., & Dear, K. B. (2000). A nonparametric mixture model for cure rate estimation. *Biometrics*, *56*(1), 237–243.

Puhr, R., Heinze, G., Nold, M., Lusa, L., & Geroldinger, A. (2017). Firth's logistic regression with rare events: Accurate effect estimates and predictions? *Statistics in Medicine*, *36*(14), 2302–2317.

Sy, J. P., & Taylor, J. M. (2000). Estimation in a cox proportional hazards cure model. *Biometrics*, *56*(1), 227–236.

Tedesco, L., Beyhum, J., & Van Keilegom, I. (2023). Instrumental variable estimation of the proportional hazards model by presmoothing. *arXiv preprint arXiv:220901429*.

Wycinka, E., & Jurkiewicz, T. (2017). *Mixture cure models in prediction of time to default: Comparison with logit and cox models*. In *Contemporary trends and challenges in finance* (pp. 221–231). Cham, Switzerland: Springer International Publishing.

Yilmaz, Y. E., Lawless, J. F., Andrulis, I. L., & Bull, S. B. (2013). Insights from mixture cure modeling of molecular markers for prognosis in breast cancer. *Journal of Clinical Oncology*, *31*(16), 2047–2054.

Zhang, J., & Peng, Y. (2007). A new estimation method for the semiparametric accelerated failure time mixture cure model. *Statistics in Medicine*, *26*(16), 3157–3171.
Ziegler, A. (2011). *Generalized estimating equations* (Vol. *204*). New York: Springer Science & Business Media.

How to cite this article: Musta, E., Patilea, V., & Van Keilegom, I. (2025). Regression estimation using surrogate responses obtained by presmoothing. *Statistica Neerlandica*, *79*(1), e12351. https://doi.org/10.1111/stan.12351

APPENDIX A. OTHER APPLICATIONS

A.1 Default probability in credit risk

Consider the problem of estimating the probability of default (PD) in credit risk. For a fixed time t and a horizon time b, the PD can be defined as the probability that a credit that has been repaid until time t fails to be repaid not later than time $t+b$ and depends on certain covariates X. Standard methods (and probably still the most used in practice) to estimate the PD include logistic models and other binary response parametric regression models. This means that, for a fixed t, b and $X = x$, the probability of default is modeled by

$$\mathrm{PD}_{t,b}(x) = \phi(\gamma_0^T x), \tag{A1}$$

for some unknown parameter γ_0 and a known function ϕ that can be the logistic one. In order to estimate γ_0 based on a sample of previous customers, for any fixed t and b, the observations are classified into "default" ($Y = 1$) and "non default" ($Y = 0$) and γ_0 is estimated via maximum likelihood. However, this approach cannot handle censoring and the observations for which the default status cannot be determined are marked as undefined and left out of the modeling.

To avoid such problems, survival analysis methods have been proposed for estimating the probability of default. In this setting, let T be the time to default, $Z = \min\{T, C\}$ is the observed maturity and $\Delta = \mathbb{1}_{\{T \le C\}}$ is the censoring indicator. Then we have

$$\mathrm{PD}_{t,b}(x) = \mathbb{P}(T \le t+b | T > t, X = x) = 1 - \frac{S(t+b|x)}{S(t|x)}, \tag{A2}$$

where S denotes the survival function of T. One can estimate $\mathrm{PD}_{t,b}(x)$ by plugging in an estimator $\hat{S}(\cdot|x)$ of the conditional survival function. One can estimate S, for example, by fitting a Cox regression model or completely nonparametrically (Peláez Suárez, Cao Abad, & Vilar Fernández, 2021; Peláez Suárez, Van Keilegom, Cao, & Vilar, 2024). In the case of a semi-parametric model like Cox, it is easy to interpret the effect of covariates on the time to default but not directly on the default probability because of the relation in (A2). On the other hand, the nonparametric estimator, would suffer from the problem of multidimensional smoothing. If one, in practice, would still like to fit the logistic model (A1) for interpretability purposes, then our approach helps to handle independent censoring. For our approach, for a fixed t and b, we define $r(u; \gamma, \eta_0) = r(u; \gamma) = \mathbb{P}(T \le t+b | T > t, \gamma^T X = u)$. Then, for a fixed γ, one can construct a Beran-type estimator of the conditional survival function $S_\gamma(t|u) = \mathbb{P}(T > t | \gamma^T X = u)$ as in (6). Then we obtain the nonparametric estimator

$$\hat{r}(u;\gamma) = 1 - \frac{\hat{S}_\gamma(t+b|u)}{\hat{S}_\gamma(t|u)}, \quad \text{where} \quad \hat{S}_\gamma(t|u) = \prod_{s\le t}\left(1 - \frac{\hat{H}_{1,\gamma}(\mathrm{d}s|u)}{\hat{H}_\gamma([s,\infty)|u)}\right).$$

Finally, we can estimate the coefficients of the logistic model by

$$\hat{\gamma}_n \in \mathrm{argmax}_\gamma \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{r}(\gamma^T x;\gamma)\log\phi(\gamma^T x) + [1-\hat{r}(\gamma^T x;\gamma)]\log[1-\phi(\gamma^T x)]\right\}.$$

A.2 Competing risks

We consider the competing risks setting in survival analysis where in addition to the survival time Z, the censoring indicator Δ and the covariates X as in Section 3.1, for the uncensored subjects we also observe the cause of death or event type $R \in \{1,2\}$, with $R = 1$ being the main event. Standard approaches for dealing with competing risks consist in modeling and estimating cause-specific or subdistribution hazards. Instead, logistic regression for modeling survival up to a given time horizon of interest t can be quite appealing in terms of simplicity and ease of interpretation Blanche et al. (2023). Define the binary indicator of experiencing a main event within time t as $Y = \mathbb{1}_{\{T\le t, R=1\}}$, which is not always observed due to censoring. As in Blanche et al. (2023) we assume a logistic model for the conditional t-risk of event 1, that is,

$$F_1(t|x) = \mathbb{P}(Y=1|X=x) = \phi(\gamma_0^T x),$$

with ϕ as in (4). To fit this model with censored data one can use inverse probability of censoring weighting by weighing the outcome or the estimating equations. These two approaches have been investigated in Blanche et al. (2023). Our general presmoothing methodology provides an alternative estimation procedure which replaces the partially observed outcome Y by a nonparametric estimator of $\mathbb{P}(Y=1|\gamma^T X)$. In this setting $T(Q,\eta_0) = Y$ and $r(u;\gamma,\eta_0) = r(u;\gamma) = \mathbb{E}[Y|\gamma^T X = u]$, that is, there is no need for an additional parameter η_0. We have

$$r(u;\gamma) = \int_0^t S_\gamma(s-|u)\mathrm{d}\Lambda_{1,\gamma}(s|u), \tag{A3}$$

where $S_\gamma(s|u) = \mathbb{P}(T > s|\gamma^T X = u)$ is the overall conditional survival probability and $\Lambda_{j,\gamma}(s|u)$, $j = 1,2$ are the conditional cumulative cause-specific hazards given $\gamma^T X = u$. The cause-specific hazards can be estimated nonparametrically by

$$\hat{\Lambda}_{j,\gamma}(s|u) = \sum_{i=1}^{n}\frac{\mathbb{1}_{\{Z_i\le s,\Delta_i=1,R_i=j\}}w_{i,\gamma}(u)}{\sum_{k=1}^{n}\mathbb{1}_{\{Z_k\ge Z_i\}}w_{k,\gamma}(u)},$$

where $w_{i,\gamma}(u) = k_b(u-\gamma^T X_i)/\sum_{l=1}^{n}k_b(u-\gamma^T X_l)$. Then, under the assumption of independent censoring as in Section 3.1, we can write

$$S_\gamma(t|u) = \prod_{s\le t}\{1 - \mathrm{d}(\Lambda_{1,\gamma}+\Lambda_{2,\gamma})(s|u)\},$$

and estimate it by a plug-in approach which results in computing a finite product as in (6). Consequently, $\hat{r}(u;\gamma)$ is obtained using the formula (A3) with S_γ and $\Lambda_{1,\gamma}$ replaced by their nonparametric estimators. Finally, $\hat{\gamma}_n$ is defined as a maximizer of

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\hat{r}_n(\gamma^T X_i,\gamma)\log\phi(\gamma^T X_i)+\left[1-\hat{r}_n(\gamma^T X_i;\gamma)\right]\log\left[1-\phi(\gamma^T X_i)\right]\right\}.$$

APPENDIX B. ASYMPTOTICS

Here we provide some general ideas of how one can prove consistency of the estimator defined in Section 2. Details depend on the setting at hand and the chosen nonparametric estimator. Hence we do not provide a complete proof but rather a sketch of the steps involved and sufficient high level conditions. The estimator in (3) is an M-estimator and the criterion function $M(\gamma;h)=\mathbb{E}[m(X;\gamma,h)]$, where $m(X;\gamma,h)=U(\phi(\gamma^T X))+h(\gamma^T X)V(\phi(\gamma^T X))$, depends on a nuisance infinite-dimensional parameter $h(\cdot,\gamma)=r(\cdot;\gamma,\eta)$. Consistency of $\hat{\gamma}_n$ can then be derived using theorem 1 in Delsol and Van Keilegom (2020). First we introduce some notation. Let $\mathcal{H}$ be an infinite-dimensional parameter space of univariate real-valued functions $h(\cdot,\gamma)$ that depend on the parameter γ such that $h_0(\cdot,\gamma)=r_0(\cdot;\gamma;\eta_0)\in\mathcal{H}$. Consider the metric $d_{\mathcal{H}}(h_1,h_2)=\sup_{\gamma\in G}d(h_1(\cdot,\gamma),h_2(\cdot,\gamma))$ for some metric d on $\mathcal{H}$. We will work under the following general assumptions :

- (C1) The parameter γ_0 lies in the interior of a compact set $G\subset\mathbb{R}^d$.
- (C2) The covariates are bounded, that is, $\mathbb{P}(\|X\|\le K)=1$ for some $K>0$, where $\|\cdot\|$ denotes the Eucledian distance.
- (C3) There exist constants $a,c>0$ such that function $|\phi(u)-\phi(v)|\le c|u-v|^a$.
- (C4) The functions U and V are uniformly bounded on the codomain of $\phi(\gamma^T X)$ for $\gamma\in G$.
- (C5) The functions $r_0(\cdot;\gamma,\eta_0)$ is continuous for all $\gamma\in G$.
- (C6) The estimator $\hat{h}_n(\cdot,\gamma)=\hat{r}_n(\cdot;\gamma,\hat{\eta}_n)$ belongs to $\mathcal{H}$ with probability converging to 1 and satisfies

$$\sup_{\gamma\in G}d(\hat{h}_n(\cdot,\gamma),h_0(\cdot,\gamma))=o_P(1).$$

Note that (C1), (C2) are standard assumptions. (C3) is satisfied for the logistic model and in particular implies that $\phi(\gamma^T X)$ and $r_0(\cdot;\gamma,\eta_0)$ are uniformly bounded. $(C4)$ holds for U and V as in Remark 1 if $\phi(\gamma^T X)$ is bounded away from 0 and 1. (C5) is also a natural assumption since we consider estimating r_0 nonparametrically via kernel smoothing.

One can then take $\mathcal{H}$ to be the space of continuous and uniformly bounded functions, equipped with the sup norm $d(g_1,g_2)=\|g_1-g_2\|_\infty$. The first condition (A1) of theorem 1 in Delsol and Van Keilegom (2020) is automatically satisfied since $\hat{\gamma}_n$ is the maximizer of the empirical criterion function

$$M_n(\gamma;\hat{h}_n)=\frac{1}{n}\sum_{i=1}^{n}m(X_i;\gamma,\hat{r}_n(\cdot;\gamma,\hat{\eta}_n)).$$

Condition (A2) follows from Proposition 1 and the continuity of m with respect to γ. Condition (A4) is satisfied if

$$\sup_{\gamma\in G}\sup_{h\in\mathcal{H}}|M_n(\gamma,h)-M(\gamma,h)|=o_P(1),$$

which holds if the family $\mathcal{F} = \{m(\cdot;\gamma,h), \gamma \in G, h \in \mathcal{H}\}$ is Glivenko–Cantelli. That can be shown using standard arguments for bracketing numbers and assumptions (C1)–(C5). Finally condition (A5) requires that $\sup_{\gamma\in G} |M(\gamma,h) - M(\gamma,h_0)|$ converges to zero as $d_{\mathcal{H}}(h,h_0) \to 0$. Indeed we have

$$\sup_{\gamma\in G}|M(\gamma,h) - M(\gamma,h_0)| = \sup_{\gamma\in G}|\{h(\gamma^TX,\gamma) - h_0(\gamma^TX,\gamma)\}V(\phi(\gamma^TX))|$$
$$\leq k \sup_{\gamma\in G}\|h(\cdot,\gamma) - h_0(\cdot,\gamma)\|_\infty = kd_{\mathcal{H}}(h,h_0),$$

for some constant $k > 0$. Condition (A3) corresponds to our assumption (C6) and depends on the estimator at hand. Note that, in the context of mixture cure models, this assumption is different from the requirements in Musta et al. (2022) and Musta et al. (2024) because there the nonparametric estimator is constructed beforehand using multivariate smoothing or a preliminary estimator and does not depend on γ. Instead, here for any γ, we have an estimator of the conditional mean given the one-dimensional covariate γ^TX and such family of estimators needs to be uniformly consistent for the corresponding conditional means. The proof of such uniform consistency is usually contingent upon the specific context.

APPENDIX C. SIMULATION RESULTS

C.1 Mixture cure model

In addition to the two models described in Section 4, we consider also the following two models to see how the new method performs when the covariate dimension is lower, particularly also for only one continuous covariate.

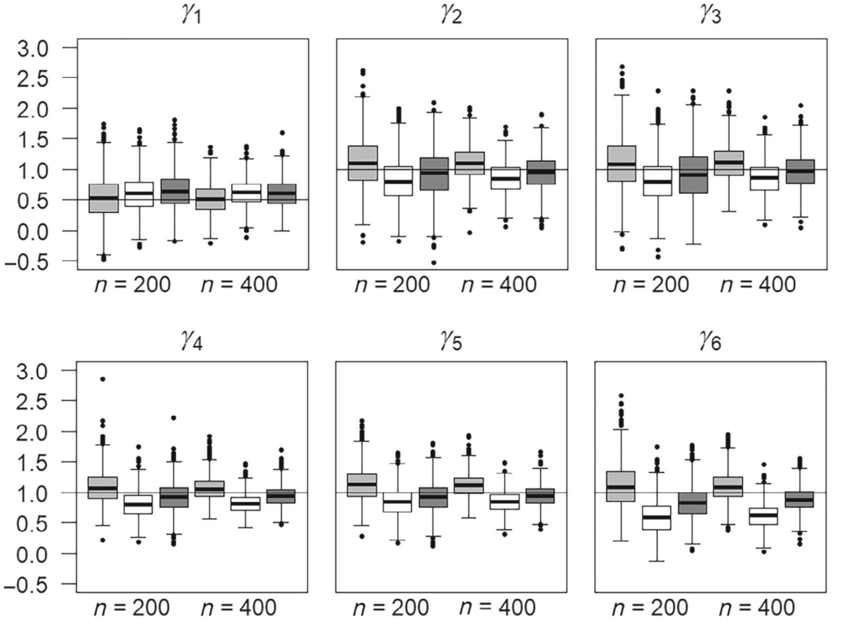

FIGURE C1 Boxplots of parameter estimates in Model 1 for the mixture cure model with sample size 200 and 400. Gray: maximum likelihood estimator. White: the multivariate presmoothing estimator (MPvK). Dark gray: the new approach.

TABLE C1 Bias, *SD* and root- mean-square-error (MSE) of $\hat{\gamma}_n$ for the maximum likelihood estimator (MLE), the presmoothing estimator (MPvK) and the new approach in Models 1 and 2 for the mixture cure model.

			MLE			Presmoothing (MPvK)			New approach		
Model	***n***	**Par.**	**Bias**	***SD***	$\sqrt{\text{MSE}}$	**Bias**	***SD***	$\sqrt{\text{MSE}}$	**Bias**	***SD***	$\sqrt{\text{MSE}}$
1	200	γ_1	−0.013	0.327	0.327	0.063	0.285	0.292	0.121	0.285	0.310
		γ_2	0.113	0.418	0.434	−0.185	0.339	0.387	−0.086	0.379	0.390
		γ_3	0.099	0.392	0.405	−0.210	0.326	0.387	−0.078	0.352	0.361
		γ_4	0.092	0.268	0.285	−0.200	0.217	0.295	−0.071	0.228	0.239
		γ_5	0.153	0.276	0.315	−0.149	0.226	0.270	−0.062	0.230	0.239
		γ_6	0.167	0.332	0.371	−0.368	0.259	0.451	−0.138	0.241	0.277
1	400	γ_1	0.009	0.237	0.237	0.109	0.224	0.249	0.104	0.219	0.243
		γ_2	0.073	0.300	0.308	−0.177	0.257	0.311	−0.062	0.286	0.293
		γ_3	0.079	0.286	0.298	−0.164	0.257	0.303	−0.056	0.266	0.272
		γ_4	0.078	0.190	0.205	−0.174	0.164	0.239	−0.045	0.167	0.173
		γ_5	0.102	0.197	0.224	−0.152	0.173	0.230	−0.061	0.179	0.187
		γ_6	0.103	0.253	0.274	−0.377	0.210	0.431	−0.118	0.207	0.239
2	200	γ_1	0.178	0.341	0.385	−0.077	0.305	0.315	0.010	0.315	0.315
		γ_2	0.121	0.378	0.396	−0.142	0.333	0.363	−0.075	0.356	0.363
		γ_3	0.146	0.383	0.411	−0.119	0.330	0.351	−0.059	0.358	0.362
		γ_4	0.114	0.259	0.283	−0.140	0.210	0.253	−0.067	0.221	0.232
		γ_5	−0.176	0.319	0.356	−0.036	0.261	0.263	0.030	0.272	0.274
		γ_6	0.138	0.584	0.600	−0.320	0.474	0.573	−0.086	0.536	0.543
2	400	γ_1	0.094	0.235	0.253	−0.079	0.214	0.230	−0.031	0.221	0.223
		γ_2	0.114	0.261	0.286	−0.097	0.234	0.253	−0.042	0.249	0.253
		γ_3	0.097	0.253	0.270	−0.120	0.230	0.259	−0.049	0.247	0.251
		γ_4	0.089	0.170	0.192	−0.128	0.148	0.197	−0.045	0.158	0.164
		γ_5	−0.116	0.205	0.237	−0.012	0.176	0.176	0.043	0.187	0.190
		γ_6	0.130	0.383	0.405	−0.257	0.342	0.428	−0.044	0.366	0.369

Model 3. For the incidence we consider four independent covariates: X_1 has a uniform distribution on $(-1, 1)$, X_2, X_3 and X_4 are Bernoulli random variables with parameters 0.2, 0.8 and 0.6, respectively. The latency depends on four covariates: $W_1 = X_1$, $W_2 = X_2$, $W_3 = X_3$ and W_4 is an independent Bernoulli variable with parameter 0.4. The cure status and the survival times for the uncured observations are generated as in Model 1 with $\gamma = (1.4, 2, 1, -1, 0.5)$, $\rho(w) = 0.75 + \exp(4\beta^T w)$, $\mu = 1.5$ and $\beta = (0.5, 0.4, -0.3, 0.2)$. The censoring times are independent from X and T and generated from the exponential distribution with parameter 0.2. In this scenario, the cure rate is around 30% and the censoring rate around 40%.

TABLE C2 Bias, *SD*, and root-mean-square-error (MSE) of $\hat{\gamma}_n$ for the maximum likelihood estimator (MLE), the presmoothing estimator (MPvK) and the new approach in Models 3 and 4 for the mixture cure model.

			MLE			Presmoothing (MPvK)			New approach		
Model	***n***	**Par.**	**Bias**	***SD***	$\sqrt{\text{MSE}}$	**Bias**	***SD***	$\sqrt{\text{MSE}}$	**Bias**	***SD***	$\sqrt{\text{MSE}}$
3	200	γ_1	0.186	0.625	0.653	−0.112	0.573	0.584	0.060	0.593	0.597
		γ_2	0.180	0.442	0.476	−0.057	0.405	0.409	−0.079	0.392	0.401
		γ_3	0.175	0.602	0.628	−0.102	0.563	0.573	−0.034	0.550	0.551
		γ_4	−0.171	0.599	0.623	0.111	0.555	0.566	0.007	0.574	0.574
		γ_5	0.010	0.416	0.416	0.006	0.401	0.401	−0.058	0.424	0.429
3	400	γ_1	0.111	0.401	0.416	−0.065	0.378	0.383	0.019	0.407	0.407
		γ_2	0.105	0.276	0.295	−0.057	0.263	0.270	−0.087	0.257	0.270
		γ_3	0.114	0.399	0.415	−0.072	0.390	0.396	−0.029	0.401	0.402
		γ_4	−0.093	0.401	0.411	0.081	0.377	0.385	0.038	0.395	0.397
		γ_5	−0.002	0.286	0.286	−0.008	0.285	0.285	−0.054	0.307	0.311
4	200	γ_1	0.021	0.465	0.465	−0.035	0.418	0.421	0.162	0.473	0.500
		γ_2	0.113	0.518	0.530	−0.284	0.395	0.486	−0.058	0.495	0.499
		γ_3	0.238	0.344	0.418	−0.123	0.300	0.324	−0.035	0.319	0.321
		γ_4	0.144	0.358	0.386	−0.114	0.324	0.344	−0.054	0.340	0.345
4	400	γ_1	0.005	0.324	0.324	−0.015	0.302	0.303	0.095	0.324	0.338
		γ_2	0.089	0.326	0.338	−0.216	0.285	0.358	−0.036	0.339	0.340
		γ_3	0.184	0.241	0.303	−0.102	0.207	0.230	−0.045	0.224	0.228
		γ_4	0.107	0.239	0.263	−0.082	0.217	0.232	−0.042	0.230	0.232

Model 4. Both the incidence and the latency component depend on three independent covariates: $X_1 = W_1$ is uniformly distributed on $(-1, 1)$, $X_2 = W_2$ has a standard normal distribution and $X_3 = W_3$ is a Binomial random variable with parameters 3 and 0.5. The cure status and the survival times for the uncured observations are generated as in Model 1 for $\gamma = (0.5, 1, 1, 1)$, $\rho(w) = 0.75 + 2\exp(\beta^T w)$, $\mu = 1.5$ and $\beta = (0.3, 1, 0.5)$. The censoring times are independent from X and T. They are generated from the exponential distribution with parameter 0.4. In this scenario the cure rate is around 20% and the censoring rate around 40%.

C.2 Logistic regression

TABLE C3 Bias, *SD*, and root-mean-square-error (MSE) of $\hat{\gamma}_n$ for the maximum likelihood estimator (MLE) and the new approach in Models 1 and 2 for logistic regression.

			MLE			New approach		
Model	***n***	**Par.**	**Bias**	***SD***	$\sqrt{\text{MSE}}$	**Bias**	***SD***	$\sqrt{\text{MSE}}$
1	100	γ_1	0.045	0.281	0.285	0.018	0.266	0.266
		γ_2	0.092	0.339	0.352	−0.042	0.292	0.295
		γ_3	0.114	0.330	0.349	−0.022	0.279	0.279
		γ_4	0.112	0.495	0.508	−0.021	0.454	0.455
		γ_5	0.118	0.524	0.537	−0.023	0.475	0.475
1	200	γ_1	0.022	0.187	0.187	0.006	0.182	0.182
		γ_2	0.052	0.224	0.230	−0.035	0.207	0.210
		γ_3	0.045	0.221	0.226	−0.044	0.205	0.207
		γ_4	0.053	0.330	0.333	−0.032	0.315	0.316
		γ_5	0.042	0.346	0.349	−0.049	0.336	0.341
2	100	γ_1	0.025	0.434	0.435	0.022	0.431	0.432
		γ_2	0.083	0.329	0.339	−0.027	0.285	0.286
		γ_3	0.069	0.488	0.493	−0.043	0.457	0.459
		γ_4	0.052	0.352	0.355	−0.011	0.333	0.333
		γ_5	−0.070	0.528	0.533	0.039	0.499	0.501
2	200	γ_1	0.015	0.298	0.298	0.008	0.291	0.291
		γ_2	0.037	0.207	0.210	−0.035	0.190	0.195
		γ_3	0.045	0.300	0.303	−0.030	0.298	0.300
		γ_4	0.019	0.237	0.237	−0.019	0.231	0.232
		γ_5	−0.049	0.365	0.369	0.027	0.358	0.359

TABLE C4 Bias, *SD*, and root-mean-square-error (MSE) of $\hat{\gamma}_n$ for the maximum likelihood estimator (MLE) and the new approach in Model 3 for logistic regression.

		MLE			New approach		
n	Par.	Bias	*SD*	$\sqrt{\text{MSE}}$	Bias	*SD*	$\sqrt{\text{MSE}}$
200	γ_1	0.097	0.335	0.348	−0.019	0.263	0.263
	γ_2	0.173	0.390	0.427	−0.108	0.270	0.290
	γ_3	0.141	0.361	0.386	−0.076	0.261	0.272
	γ_4	−0.333	0.526	0.623	0.224	0.307	0.379
	γ_5	−0.157	0.370	0.402	0.121	0.261	0.288
	γ_6	0.246	0.466	0.527	−0.162	0.297	0.338
	γ_7	0.168	0.373	0.409	−0.106	0.261	0.281
	γ_8	−0.346	0.539	0.641	0.215	0.315	0.382
	γ_9	−0.085	0.336	0.346	0.077	0.245	0.257
	γ_{10}	0.225	0.414	0.471	−0.135	0.276	0.307
	γ_{11}	0.311	0.497	0.587	−0.191	0.292	0.349
400	γ_1	0.031	0.195	0.195	−0.030	0.172	0.173
	γ_2	0.058	0.230	0.237	−0.102	0.190	0.214
	γ_3	0.062	0.221	0.230	−0.067	0.182	0.195
	γ_4	−0.135	0.305	0.335	0.192	0.230	0.300
	γ_5	−0.060	0.226	0.232	0.102	0.184	0.212
	γ_6	0.103	0.266	0.285	−0.139	0.207	0.249
	γ_7	0.073	0.221	0.232	−0.088	0.182	0.202
	γ_8	−0.138	0.318	0.346	0.192	0.235	0.303
	γ_9	−0.043	0.210	0.214	0.058	0.179	0.187
	γ_{10}	0.089	0.243	0.259	−0.123	0.192	0.230
	γ_{11}	0.112	0.285	0.307	−0.180	0.219	0.283

TABLE C5 Bias, *SD*, and root-mean-square-error (MSE) of $\hat{\gamma}_n$ for the maximum likelihood estimator (MLE), the new approach and the two modifications of Firth's logistic regression (FLIC, FLAC) in Model 4 for logistic regression.

		$n = 200$			$n = 400$		
Method	**Par.**	**Bias**	***SD***	$\sqrt{\text{MSE}}$	**Bias**	***SD***	$\sqrt{\text{MSE}}$
MLE	γ_1	0.352	0.854	0.924	0.141	0.480	0.500
	γ_2	0.143	0.486	0.506	0.055	0.295	0.300
	γ_3	−0.080	0.390	0.399	−0.036	0.261	0.265
	γ_4	−0.179	0.582	0.609	−0.080	0.339	0.349
New	γ_1	0.043	0.633	0.635	−0.039	0.412	0.415
approach	γ_2	−0.057	0.378	0.382	−0.067	0.263	0.270
	γ_3	0.055	0.339	0.344	0.047	0.241	0.247
	γ_4	0.079	0.439	0.446	0.082	0.292	0.302
FLIC	γ_1	0.074	0.694	0.698	0.021	0.442	0.443
	γ_2	0.001	0.416	0.416	−0.008	0.277	0.277
	γ_3	0.013	0.343	0.343	0.006	0.247	0.247
	γ_4	0.009	0.492	0.492	0.003	0.318	0.318
FLAC	γ_1	0.048	0.690	0.691	0.007	0.440	0.440
	γ_2	−0.012	0.415	0.415	−0.015	0.277	0.277
	γ_3	0.022	0.342	0.342	0.011	0.247	0.247
	γ_4	0.027	0.491	0.492	0.013	0.318	0.318

C.3 Missing outcome context with a surrogate response

TABLE C6 Bias, *SD*, and root-mean-square-error (MSE) of $\hat{\gamma}_n$ for the least squares estimator (LSE) and the new approach in Models 1 and 2 for the missing data context.

			LSE			New approach		
Model	***n***	**Par.**	**Bias**	***SD***	**$\sqrt{\text{MSE}}$**	**Bias**	***SD***	**$\sqrt{\text{MSE}}$**
1	100	γ_1	−0.001	0.267	0.267	0.013	0.255	0.255
		γ_2	0.000	0.163	0.163	−0.042	0.149	0.155
		γ_3	−0.010	0.280	0.280	−0.076	0.257	0.268
		γ_4	−0.007	0.337	0.337	−0.031	0.310	0.311
1	200	γ_1	−0.004	0.185	0.185	0.009	0.188	0.188
		γ_2	0.015	0.116	0.116	−0.019	0.118	0.120
		γ_3	−0.001	0.197	0.197	−0.050	0.202	0.208
		γ_4	0.007	0.242	0.242	−0.014	0.250	0.250
2	100	γ_1	−0.126	0.361	0.383	−0.001	0.294	0.294
		γ_2	−0.007	0.262	0.262	−0.003	0.250	0.250
		γ_3	0.057	0.449	0.453	−0.031	0.396	0.397
		γ_4	0.011	0.476	0.476	−0.114	0.382	0.399
2	200	γ_1	−0.007	0.318	0.326	0.002	0.272	0.272
		γ_2	−0.010	0.213	0.213	−0.007	0.210	0.210
		γ_3	0.006	0.335	0.335	−0.026	0.302	0.303
		γ_4	0.012	0.358	0.359	0.082	0.302	0.313