A Simultaneous Equation Approach to Estimating HIV Prevalence with Non-Ignorable Missing Responses*

Giampiero Marra† Rosalba Radice‡ Till Bärnighausen§

Simon N. Wood¶ Mark E. McGovern‖

Abstract

Estimates of HIV prevalence are important for policy in order to establish the health status of a country's population and to evaluate the effectiveness of population-based interventions and campaigns. However, participation rates in testing for surveillance conducted as part of household surveys, on which many of these estimates are based, can be low. HIV positive individuals may be less likely to participate because they fear disclosure, in which case estimates obtained using conventional approaches to deal with missing data, such as imputation-based methods, will be biased. We develop a Heckman-type simultaneous equation approach which accounts for non-ignorable selection, but unlike previous implementations, allows for spatial dependence and does not impose a homogeneous selection process on all respondents. In addition, our framework addresses the issue of separation, where for instance some factors are severely unbalanced and highly predictive of the response, which would ordinarily prevent model convergence. Estimation is carried out within a penalized likelihood framework where smoothing is achieved using a parametrization of the smoothing criterion which makes estimation more stable and efficient. We provide the software for straightforward implementation of the proposed approach, and apply our methodology to estimating national and sub-national HIV prevalence in Swaziland, Zimbabwe and Zambia.

Key Words: Heckman-Type Selection Model, HIV, Penalized Regression Spline, Selection Bias, Simultaneous Equation Model, Spatial Dependence.

*E-mail for correspondence: giampiero.marra@ucl.ac.uk.

†Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, UK.

‡Department of Economics, Mathematics and Statistics, Birkbeck, University of London, Malet Street, London WC1E 7HX, UK.

§Department of Global Health and Population, Harvard T.H. Chan School of Public Health, Boston, MA, USA. Wellcome Trust Africa Centre for Population Health, University of KwaZulu-Natal, Mtubatuba, South Africa.

¶Department of Mathematical Sciences, University of Bristol, University Walk, Clifton, Bristol BS8 1TW, UK.

‖Queen's Management School, Queen's University Belfast, Belfast, Northern Ireland.

1 Missing data in HIV research

Interventions targeted to control the HIV epidemic, improve population health, and reduce HIV-related health disparities, are often motivated by prevalence data obtained from HIV testing (Beyrer et al., 1999; De Cock et al., 2006). In many countries, estimates of HIV prevalence obtained from home-based testing during nationally representative household surveys are now considered the gold standard (Boerma et al., 2003). However, these data can be affected by non-participation because some of those who are eligible opt out of HIV testing. In general, the treatment of missing information in survey data has the potential to have a substantial impact on both the model's parameter estimates and the resulting policy recommendations (Nicoletti, 2006). Because we cannot observe the true outcome for those who do not participate, and because of the role these data have in informing policy, modeling non-participation in testing and developing a framework for accounting for missing data in a manner which imposes as few assumptions as possible is particularly relevant for the field of HIV research.

Non-participation can occur through a variety of mechanisms, including directly declining to test for HIV when a respondent is approached to test after their interview, or being an eligible respondent for HIV testing but not being present when the interviewers seek to contact the person for interview (Marston et al., 2008). This means that ex post the surveyed group who consent to HIV testing may not be representative of the population of interest. Selection bias occurs if HIV prevalence among those who participate in testing differs from those who do not. In many contexts the extent of non-participation is substantial; for example, 37% of eligible male respondents failed to participate in testing in the 2004 Malawi Demographic and Health Survey (Hogan et al., 2012).

There are several options for dealing with missingness caused by non-participation (Donders et al., 2006). Standard approaches include multiple imputation, inverse probability re-weighting and propensity score methods, which all require that missing data are missing at random. However, due to stigma, HIV positive individuals may be less likely to participate in testing because they fear disclosure of their status. Longitudinal evidence from demographic surveillance sites supports the hypothesis that HIV positive individuals are less likely to consent to test (Arpino et al., 2014; Floyd et al., 2013; Reniers & Eaton, 2009; Bärnighausen et al., 2012; Obare, 2010). Participation in testing is also lower in communities with higher knowledge of HIV status (Reniers & Eaton, 2009). If data are missing because HIV positive individuals are more likely to decline to test (con-

ditional on observed characteristics), then the assumption of missing at random is violated and hence conventional methods, including imputation or analysis based only on non-missing observations, will generate biased results (e.g., Heckman, 1990; Puhani, 2000; Vella, 1998; Janssens et al., 2014). In addition, because imputation-based models do not acknowledge that there is uncertainty surrounding the relationship between participation in testing and HIV status, confidence intervals based on this approach are likely to be too narrow when non-participation is common (Hogan et al., 2012).

1.1 Towards a more flexible framework for estimating HIV prevalence

Although the simultaneous equation modeling approach, such as that proposed by Heckman (1979), has the advantage of not requiring the assumption of missing at random, previous techniques implementing this model are limited by a number of methodological drawbacks. This article makes four methodological contributions to the literature, and for each of these we outline the relevant problem and illustrate how our framework is designed to correct for the issue.

First, we introduce a linear predictor equation for the parameter modeling the association between consent to HIV testing and HIV status; this allows us to capture potential heterogeneity in the selection process. Moreover, we include spatial information in the model to reflect the manner in which HIV is spread through social interaction (Klovdahl, 1985) using a Markov random field approach (Rue & Held, 2005). To the best of our knowledge, this is the first time that spatial information and heterogeneity in the selection process have been incorporated into Heckman-type models. In this way, we are able to provide better calibrated region-specific HIV prevalence estimates. Networks and proximity propagate the transmission and spread of infectious disease, and therefore HIV status and other outcomes which are determined by proximal interaction will be affected by geographic clustering (Tanser et al., 2009), with likely spill-over effects and spatial dependence among communities (e.g., Larmarange & Bendaud, 2014; Aral et al., 2005). Also, there may be some groups among the population for whom the stigma associated with being HIV positive is particularly strong, hence inducing more selection bias. The association between the decision to consent to HIV testing and HIV status may vary between these communities as a result (e.g., Kranzer et al., 2008). Therefore, as well as being inefficient, the imposition of a common selection process across all sub-groups could bias sub-national HIV prevalence estimates. The best

that could be done with previous implementations is to stratify according to the group of interest, however given the resulting inefficiency and that sample sizes can be low across groups, this is not a realistic solution. Our proposal has potentially important applications beyond HIV research and will likely be of interest in situations in which there is spatial dependence and missing data.

Second, we extend the selection framework to allow for the utilization of ridge penalties to deal with problematic parameters (associated with categorical regressors, for instance) which would ordinarily lead to convergence failure. It is known that, with binary responses the problem of separation, where for instance some factor variables are severely unbalanced and highly predictive of the response, often prevents algorithms from converging (e.g., Heinze & Schemper, 2002). In practice, the bivariate probit models which have been used to implement selection models when the outcome is binary are not very stable and fail to converge relatively frequently (Butler, 1996; Clark & Houle, 2014). Therefore, there is a danger that such models are only employed in cases with specific data configurations. In our case study, we apply a ridge penalty on the parameters of the selection variable, interviewer identity. As we describe further in the next section, the selection variable can be thought of as an instrumental variable in that it predicts participation but is assumed not to predict directly the outcome of interest (Madden, 2008). In all three countries considered in our analysis, we were unable to implement the traditional selection model. The interviewers in these surveys are often matched to participants on the basis of some group-level characteristics (e.g., language). Moreover, some interviewers obtain participation in testing from all their interviewees, while for some other interviewers all their interviewees may decline to participate. This means that some interviewer effects will not be estimable due to lack of within-interviewer variation in testing participation. Solutions involving pooling very successful interviewers with very unsuccessful interviewers, dropping problematic interviewers, or estimating interviewer persuasiveness in a two-stage process are clearly not desirable (McGovern et al., 2015a). To the best of our knowledge, there is no alternative implementation of selection models which would allow us to deal with the above mentioned issue in a theoretically founded way. Given that, in practical applications, selection models often suffers from these types of convergence failures, it is likely that our proposed development will be of use beyond the HIV study considered in this article.

Third, we make use of a parametrization of the smoothing criterion that is different from the one discussed in the previous literature on bivariate equation models (Marra & Radice, 2013;

Radice et al., 2015). This has the advantage of making smoothing parameter estimation more stable and efficient. Our derivations also show that the proposed approach can in principle be applied to any situation in which a model is fitted by penalized maximum likelihood, thereby appealing to a wider audience of researchers.

Fourth, all the developments discussed in this article have been made available through the freely distributed and easy to use `R` package `SemiParBIVProbit` (Marra & Radice, 2016), which can allow researchers and policy-makers to apply a flexible selection approach to account for systematic non-participation in their data.

Our methodology incorporates each of these developments in a flexible simultaneous equation framework for adjusting for systematic non-participation in HIV surveys. We outline further details of this methodology in the rest of the paper as follows. Section 2 introduces the approach in more detail by describing its main statistical components, including estimation and inference. Sections 3 and 4 describe the data and apply the proposed approach to three Sub-Saharan African countries (Swaziland, Zambia, and Zimbabwe). In Section 5, we outline two approaches for evaluating the sensitivity of results to model assumptions. The final section provides a discussion and directions for future research.

2 Extending Heckman-type selection models

Heckman-type selection models can be used to correct for selection bias due to unobserved characteristics of respondents, as would be the case if HIV positive individuals were systematically opting out of HIV testing because of fear of disclosure. This simultaneous equation approach acknowledges the sequential decision making process involved in survey participation; respondents first decide whether to participate in testing, and it is only conditional on consenting to test that we observe their HIV status. Heckman (1979) originally proposed explicitly modeling the selection mechanism (whether respondents test or not) and outcome of interest (the HIV status of respondents) as a function of the observed characteristics of respondents, and linking the selection and outcome equations through a bivariate normal distribution. In this approach, parameters are typically estimated under a maximum likelihood framework. When the outcome is binary, the conventional Heckman selection model is a bivariate probit (Dubin & Rivers, 1989; Van de Ven & Van Praag, 1981). Common criticisms of this approach include, however,

the reliance on the assumption of bivariate normality and the lack of flexibility in modeling covariate effects. Subsequent developments have addressed these issues (Marra & Radice, 2013; McGovern et al., 2015b) and have represented the starting point for our proposed extensions.

Selection models require a valid instrument for identification. As mentioned in the previous section, interviewer identity will serve as an exclusion restriction in our study. The identity of the interviewer who contacts the respondent to seek consent for an HIV test is often recorded in survey data as an anonymized code. The allocation of interviewers to eligible survey respondents is typically highly correlated with whether the respondents consent to test. Such allocation is also based on survey design features, as opposed to the characteristics of the respondents themselves. Therefore, interviewer identity is plausibly exogenous and should be unrelated to the HIV status of survey respondents. In other words, interviewer identity satisfies potentially the condition of exclusion restriction (Bärnighausen et al., 2011). The validity of this assumption is discussed further in Section 5.2.

2.1 Model representation

Let us assume that there are two random variables (Y_{1i}, Y_{2i}), for $i = 1, \ldots, n$, where $Y_{1i}, Y_{2i} \in \{0, 1\}$ and n represents the sample size. Variable Y_{1i} indicates whether an individual takes part in the study whereas Y_{2i} denotes the outcome. The probability of event $(Y_{1i} = 1, Y_{2i} = 1)$, conditional on the sets of covariates $\mathbf{z}_{1i}$ and $\mathbf{z}_{2i}$, can be defined as (Kolev & Paiva, 2009; Sklar, 1959, 1973; Zimmer & Trivedi, 2006)

$$p_{11i} = \mathbb{P}(Y_{1i} = 1, Y_{2i} = 1 | \mathbf{z}_{1i}, \mathbf{z}_{2i}) = \mathcal{C}(\mathbb{P}(Y_{1i} = 1 | \mathbf{z}_{1i}), \mathbb{P}(Y_{2i} = 1 | \mathbf{z}_{2i}); \theta_i),$$

where $\mathbb{P}(Y_{vi} = 1 | \mathbf{z}_{vi}) = \Phi(\eta_{vi})$ for $v = 1, 2$, $\Phi(\cdot)$ is the cumulative distribution function (cdf) of the standard univariate Gaussian distribution, $\eta_{vi} \in \mathbb{R}$ is a linear predictor made up of regression coefficients and covariates (defined in generic terms in the next section), $\mathcal{C}$ is a two-place copula function and θ_i is an association parameter measuring the dependence between the two random variables. Since the strength of the association between the selection and outcome equations may vary across groups of observations (specifically, across regions in our case), in our framework we allow the copula dependence parameter to be specified as a function of a linear predictor. That is, $\theta_i = m(\eta_{3i})$ where m is a one-to-one transformation which ensures that the dependence parameter

lies in its range, and η_{3i} is the linear predictor associated with the copula parameter. For the list of transformations and copulae (as well as the counter-clockwise rotated versions of some of them) see Radice et al. (2015). In this context, Y_{2i} is available only if $Y_{1i} = 1$, hence the only additional events are $(Y_{1i} = 1, Y_{2i} = 0)$ and $(Y_{1i} = 0)$, with probabilities $p_{10i} = \Phi(\eta_{1i}) - p_{11i}$ and $p_{0i} = \Phi(-\eta_{1i})$. Therefore, the log-likelihood function of the sample is expressed as

$$\ell = \sum_{i=1}^{n} \left\{ y_{1i} y_{2i} \log(p_{11i}) + y_{1i}(1 - y_{2i}) \log(p_{10i}) + (1 - y_{1i}) \log(p_{0i}) \right\},$$

where y_{1i} and y_{2i} are realizations of Y_{1i} and Y_{2i}, respectively.

2.2 Linear predictor specification

For simplicity, and without loss of generality, we suppress subscript v and define the generic linear predictor as

$$\eta_i = \beta_0 + \sum_{k=1}^{K} s_k(\mathbf{z}_{ki}), \; i = 1, \ldots, n, \tag{1}$$

where $\beta_0 \in \mathbb{R}$ is an overall intercept, $\mathbf{z}_{ki}$ denotes the k^{th} sub-vector of the complete covariate vector $\mathbf{z}_i$ (which contains, e.g., binary, categorical, continuous and spatial variables), and the K functions $s_k(\mathbf{z}_{ki})$ represent generic effects which are chosen according to the type of covariate(s) considered. Each $s_k(\mathbf{z}_{ki})$ can be approximated as a linear combination of J_k basis functions $b_{kj_k}(\mathbf{z}_{ki})$ and regression coefficients $\beta_{kj_k} \in \mathbb{R}$, i.e.

$$\sum_{j_k=1}^{J_k} \beta_{kj_k} b_{kj_k}(\mathbf{z}_{ki}). \tag{2}$$

Equation (2) implies that the vector of evaluations $\{s_k(\mathbf{z}_{k1}), \ldots, s_k(\mathbf{z}_{kn})\}^{\mathsf{T}}$ can be written as $\mathbf{Z}_k \boldsymbol{\beta}_k$, with coefficient vector $\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{kJ_k})^{\mathsf{T}}$ and design matrix $Z_k[i, j_k] = b_{kj_k}(\mathbf{z}_{ki})$. This allows us to write the linear predictor in equation (1) as

$$\boldsymbol{\eta} = \beta_0 \mathbf{1}_n + \mathbf{Z}_1 \boldsymbol{\beta}_1 + \ldots + \mathbf{Z}_K \boldsymbol{\beta}_K, \tag{3}$$

where $\mathbf{1}_n$ is an n-dimensional vector made up of ones. Equation (3) can also be written in a more compact way as $\boldsymbol{\eta} = \mathbf{Z}\boldsymbol{\beta}$, where $\mathbf{Z} = (\mathbf{1}_n, \mathbf{Z}_1, \ldots, \mathbf{Z}_K)$ and $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1^{\mathsf{T}}, \ldots, \boldsymbol{\beta}_K^{\mathsf{T}})^{\mathsf{T}}$. The smooth functions may represent linear, non-linear, random and spatial effects, to name but a few. Moreover, each $\boldsymbol{\beta}_k$ has an associated quadratic penalty $\lambda_k \boldsymbol{\beta}_k^{\mathsf{T}} \mathbf{D}_k \boldsymbol{\beta}_k$ whose role is to enforce specific properties on the k^{th} function, such as smoothness. $\mathbf{D}_k$ is defined in the next paragraphs for several cases. Smoothing parameter $\lambda_k \in [0, \infty)$ controls the trade-off between fit and smoothness,

and plays a crucial role in determining the shape of $\hat{s}_k(\mathbf{z}_{ki})$; a large value for λ_k means that the corresponding penalty has a large influence on the parameters of the function during fitting, and viceversa. The overall penalty can be defined as $\boldsymbol{\beta}^{\mathsf{T}}\mathbf{D}_{\boldsymbol{\lambda}}\boldsymbol{\beta}$, where $\mathbf{D}_{\boldsymbol{\lambda}} = \mathrm{diag}(0, \lambda_1\mathbf{D}_1, \ldots, \lambda_K\mathbf{D}_K)$. Note also that smooth functions are subject to centering (identifiability) constraints and we employ the parsimonious approach detailed in Wood (2006) to deal with this issue. In the following paragraphs, we outline the rationale for adopting the specific model components relevant to our case study.

Spatial effects To model the spatial information based on the geographic location of survey respondents, we employ a Markov random field smoother. This approach is popular when the geographic area (or country) of interest is split up into discrete contiguous geographic units (or regions), and allows us to take advantage of the information contained in neighboring observations which are located in the same country. In this case, equation (2) becomes $\mathbf{z}_{ki}^{\mathsf{T}}\boldsymbol{\beta}_k$, where $\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{kR})^{\mathsf{T}}$ represents the vector of spatial effects, R denotes the total number of regions, and $\mathbf{z}_{ki}$ is made up of a set of area labels. The design matrix linking an observation i to the corresponding spatial effect is therefore defined as

$$\mathbf{Z}_k[i, r] = \begin{cases} 1 \text{ if the observation belongs to region } r \\ 0 \text{ otherwise} \end{cases},$$

where $r = 1, \ldots, R$. The smoothing penalty is based on the neighborhood structure of the geographic units, so that spatially adjacent regions share similar effects. That is

$$\mathbf{D}_k[r, q] = \begin{cases} -1 \text{ if } r \neq q \wedge \ r \text{ and } q \text{ are adjacent neighbors} \\ 0 \quad \text{if } r \neq q \wedge \ r \text{ and } q \text{ are not adjacent neighbors} \\ N_r \text{ if } r = q \end{cases},$$

where N_r is the total number of neighbors for region r. In a stochastic interpretation, this penalty is equivalent to the assumption that $\boldsymbol{\beta}_k$ follows a Gaussian Markov random field (e.g., Rue & Held, 2005). This approach is also used to allow for a heterogeneous selection process where the copula parameter (measuring the conditional association between HIV status and participation in testing) varies according to region.

Linear and random effects For parametric, linear effects, equation (2) becomes $\mathbf{z}_{ki}^{\mathsf{T}}\boldsymbol{\beta}_k$, and the design matrix is obtained by stacking all covariate vectors $\mathbf{z}_{ki}$ into $\mathbf{Z}_k$. In general, no penalty is assigned to linear effects ($\mathbf{D}_k = \mathbf{0}$). This would be the case for variables such as ever tested for HIV and condom use at last sexual activity. However, sometimes the parameters of factor variables such as interviewer identity may be weakly or not identified by the data (see Section 1.1). In such cases, we recommend using a ridge penalty (i.e., $\mathbf{D}_k = \mathbf{I}$, where $\mathbf{I}$ is an identity matrix) to make the model parameters estimable. This is equivalent to the assumption that the coefficients are $i.i.d.$ normal random effects with unknown variance (e.g., Ruppert et al., 2003; Wood, 2006).

Non-linear effects For continuous variables such as age and years of education the smooth functions are represented using the regression spline approach popularized by Eilers & Marx (1996). Specifically, for each continuous variable z_{ki} we use equation (2), where the $b_{kj_k}(z_{ki})$ are known spline basis functions. The design matrix $\mathbf{Z}_k$ comprises the basis function evaluations for each i, and describe the J_k curves which have varying degrees of complexity. We employ low rank thin plate regression splines (Wood, 2003) which are numerically stable and have convenient mathematical properties, although other spline definitions (including B-splines and cubic regression splines) and corresponding penalties are supported in our implementation. To enforce smoothness, a conventional integrated square second derivative spline penalty is typically employed. That is, $\mathbf{D}_k = \int \mathbf{d}_k(z_k)\mathbf{d}_k(z_k)^{\mathsf{T}}dz_k$, where the j_k^{th} element of $\mathbf{d}_k(z_k)$ is given by $\partial^2 b_{kj_k}(z_k)/\partial z_k^2$ and integration is over the range of z_k. The formulae used to compute the basis functions and penalties for many spline definitions are provided in Ruppert et al. (2003) and Wood (2006). This flexible spline approach allows us to avoid arbitrary modeling decisions, such as choosing the appropriate degree of a polynomial or specifying cut-points, which could induce misspecification.

In the context of our study, the linear predictors for the selection (η_1) and outcome equations (η_2) and for the copula parameter (η_3) are specified as

$$\eta_{1i} = \beta_{10} + \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}_{11} + s_{11}(age_i) + s_{12}(education_i) + s_{13}(wealth_i) + s_{1spatial}(region_i) + \beta_{interviewerID_i},$$

$$\eta_{2i} = \beta_{20} + \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}_{21} + s_{21}(age_i) + s_{22}(education_i) + s_{23}(wealth_i) + s_{2spatial}(region_i),$$

$$\eta_{3i} = \beta_{30} + s_{3spatial}(region_i),$$

where parameters β_{10}, β_{20}, β_{30} are constants comprising the overall levels of the predictors, vector $\mathbf{x}_i$ contains discrete and binary variables that are associated with the selection and outcome equa-

tions, $\boldsymbol{\beta}_{11}$ and $\boldsymbol{\beta}_{21}$ are the respective parameter vectors, the s_{vk} for $v = 1, 2$ and $k = 1, 2, 3$ are smooth functions of age, education and wealth represented using penalized thin plate regression splines, and the $s_{vspatial}$ for $v = 1, 2, 3$ model spatial regional effects using a Markov random field approach. Finally, $\beta_{interviewerID_i}$ denotes the random effects for the set of binary variables defined by interviewer identity. The variables included in $\mathbf{x}_i$ are: type of location (urban or rural), marital status, had a sexually transmitted disease, age at first intercourse, had high risk sex, number of partners, condom use, would care for an HIV-infected relative, knows someone who died of AIDS, previously tested for HIV, smokes, drinks alcohol, language, region, ethnicity and religion. The choice of variables followed previous studies which examined the predictors of testing and HIV status in detail (Bärnighausen et al., 2011; Hogan et al., 2012). Linear predictor η_3 models the presence of unobserved confounders and therefore specifying the predictor equation as a function of observed characteristics only makes sense from an estimation perspective if there are groups for which there is a clear rationale for expecting heterogeneity in the selection process. While in theory we could include additional group-level identifier variables, we opt to specify the copula parameter as depending on region. This parametrization is motivated by the evidence on the spatial clustering of HIV prevalence (Larmarange & Bendaud, 2014; Tanser et al., 2009). There are other types of smooth functions that could be incorporated in our framework, should they be required. These include varying coefficient models obtained, for instance, by multiplying one or more smooth components by some predictor(s), and smooth functions of two or more continuous covariates; see Hastie & Tibshirani (1993), Ruppert et al. (2003) and Wood (2006) for more details.

2.3 Parameter estimation

Let us define the overall quantities $\boldsymbol{\delta}^\mathsf{T} = (\boldsymbol{\beta}_1^\mathsf{T}, \boldsymbol{\beta}_2^\mathsf{T}, \boldsymbol{\beta}_3^\mathsf{T})$ and $\mathbf{S}_{\boldsymbol{\lambda}} = \mathrm{diag}(\boldsymbol{\lambda}_1\mathbf{D}_1, \boldsymbol{\lambda}_2\mathbf{D}_2, \boldsymbol{\lambda}_3\mathbf{D}_3)$, where $\boldsymbol{\lambda}_v^\mathsf{T} = (\lambda_{vk_v}, \ldots, \lambda_{vK_v})$ for $v = 1, 2, 3$. The parameter vectors and matrices that make up $\boldsymbol{\delta}$ and $\mathbf{S}_{\boldsymbol{\lambda}}$ are related to η_{1i}, η_{2i} and η_{3i}. Because of the flexible linear predictor specifications employed here, the use of a classic (unpenalized) optimization algorithm is likely to result in function estimates that may not reflect the true underlying trends in the data (e.g., Ruppert et al., 2003; Wood, 2006). Therefore, we maximize

$$\ell_p(\boldsymbol{\delta}) = \ell(\boldsymbol{\delta}) - \frac{1}{2}\boldsymbol{\delta}^\mathsf{T}\mathbf{S}_{\boldsymbol{\lambda}}\boldsymbol{\delta}. \tag{4}$$

Estimation of $\boldsymbol{\delta}$ and $\boldsymbol{\lambda}$ is carried out in two steps. Given $\hat{\boldsymbol{\lambda}}^{\mathsf{T}} = (\hat{\boldsymbol{\lambda}}_1^{\mathsf{T}}, \hat{\boldsymbol{\lambda}}_2^{\mathsf{T}}, \hat{\boldsymbol{\lambda}}_3^{\mathsf{T}})$, we seek to maximize (4). As in Radice et al. (2015), we use a trust region approach which is generally more stable and faster than its line-search counterparts (such as Newton-Raphson), particularly for functions that are, for example, non-concave and/or exhibit regions that are close to flat (Nocedal & Wright, 2006, Chapter 4). Let us define the penalized gradient and Hessian at iteration a as $\mathbf{g}_p^{[a]} = \mathbf{g}^{[a]} - \mathbf{S}_{\boldsymbol{\lambda}^{[a]}}\boldsymbol{\delta}^{[a]}$ and $\boldsymbol{\mathcal{H}}_p^{[a]} = \boldsymbol{\mathcal{H}}^{[a]} - \mathbf{S}_{\boldsymbol{\lambda}^{[a]}}$, where $\mathbf{g}^{[a]}$ consists of $\mathbf{g}_1^{[a]} = \partial\ell(\boldsymbol{\delta})/\partial\boldsymbol{\beta}_1\big|_{\boldsymbol{\beta}_1=\boldsymbol{\beta}_1^{[a]}}$, $\mathbf{g}_2^{[a]} = \partial\ell(\boldsymbol{\delta})/\partial\boldsymbol{\beta}_2\big|_{\boldsymbol{\beta}_2=\boldsymbol{\beta}_2^{[a]}}$ and $\mathbf{g}_3^{[a]} = \partial\ell(\boldsymbol{\delta})/\partial\boldsymbol{\beta}_3\big|_{\boldsymbol{\beta}_3=\boldsymbol{\beta}_3^{[a]}}$, and the Hessian matrix has elements $\boldsymbol{\mathcal{H}}_{o,h}^{[a]} = \partial^2\ell(\boldsymbol{\delta})/\partial\boldsymbol{\beta}_o\partial\boldsymbol{\beta}_h^{\mathsf{T}}\big|_{\boldsymbol{\beta}_o=\boldsymbol{\beta}_o^{[a]},\boldsymbol{\beta}_h=\boldsymbol{\beta}_h^{[a]}}$ with $o, h = 1, \ldots, 3$. For a given $\boldsymbol{\lambda}^{[a]}$, the trust region algorithm solves the problem

$$\min_{\mathbf{p}} \breve{\ell}_p(\boldsymbol{\delta}^{[a]}) \stackrel{\text{def}}{=} -\left\{\ell_p(\boldsymbol{\delta}^{[a]}) + \mathbf{p}^{\mathsf{T}}\mathbf{g}_p^{[a]} + \frac{1}{2}\mathbf{p}^{\mathsf{T}}\boldsymbol{\mathcal{H}}_p^{[a]}\mathbf{p}\right\} \text{ such that } \|\mathbf{p}\| \leq ra^{[a]},$$

$$\boldsymbol{\delta}^{[a+1]} = \arg\min_{\mathbf{p}} \breve{\ell}_p(\boldsymbol{\delta}^{[a]}) + \boldsymbol{\delta}^{[a]},$$

where $\|\cdot\|$ denotes the Euclidean norm and $ra^{[a]}$ is the radius of the trust region; full details can be found, e.g., in Geyer (2015). Note that, near the solution, the trust region method typically behaves as a classic unconstrained algorithm (e.g., Nocedal & Wright, 2006). Our implementation provides the possibility of using $\mathbb{E}(\boldsymbol{\mathcal{H}}^{[a]})$ instead of the default option $\boldsymbol{\mathcal{H}}^{[a]}$. However, as in Wood (2011), we generally found observed information to be superior in terms of speed, stability and accuracy of results (Efron & Hinkley, 1978).

The second step concerns smoothing parameter selection. There are a number of methods for automatically estimating smoothing parameters within a penalized likelihood framework, and in the context of bivariate equation models the approach discussed in Radice et al. (2015) and Marra & Radice (2013) has proven successful. However, for the models considered in this paper such a scheme may be unstable and inefficient when the linear predictors are highly flexible and the copula parameter is specified as a function of covariates (see Supplementary Material (SM)-A for a through explanation of this). We therefore perform smoothing parameter estimation using an alternative (more stable and efficient) parametrization of the smoothing criterion. After some manipulation, the model's parameter estimator can be expressed as

$$\boldsymbol{\delta}^{[a+1]} = \left(\boldsymbol{\mathcal{I}}^{[a]} + \mathbf{S}_{\boldsymbol{\lambda}^{[a]}}\right)^{-1}\sqrt{\boldsymbol{\mathcal{I}}^{[a]}}\mathbf{z}^{[a]}, \tag{5}$$

where $\boldsymbol{\mathcal{I}}^{[a]} = -\boldsymbol{\mathcal{H}}^{[a]}$, $\mathbf{z}^{[a]} = \sqrt{\boldsymbol{\mathcal{I}}^{[a]}}\boldsymbol{\delta}^{[a]} + \boldsymbol{\epsilon}^{[a]}$ and $\boldsymbol{\epsilon}^{[a]} = \sqrt{\boldsymbol{\mathcal{I}}^{[a]}}^{-1}\mathbf{g}^{[a]}$. From likelihood theory, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{z}}, \mathbf{I})$, where $\mathbf{I}$ is an identity matrix, $\boldsymbol{\mu}_{\mathbf{z}} = \sqrt{\boldsymbol{\mathcal{I}}}\boldsymbol{\delta}^0$ and $\boldsymbol{\delta}^0$ is the true parameter

vector. The predicted value vector for $\mathbf{z}$ is $\hat{\boldsymbol{\mu}}_{\mathbf{z}} = \sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\delta}} = \mathbf{A}_{\hat{\boldsymbol{\lambda}}}\mathbf{z}$, where $\mathbf{A}_{\hat{\boldsymbol{\lambda}}} = \sqrt{\boldsymbol{\mathcal{I}}}\left(\boldsymbol{\mathcal{I}} + \mathbf{S}_{\hat{\boldsymbol{\lambda}}}\right)^{-1}\sqrt{\boldsymbol{\mathcal{I}}}$. Representation (5) allows us to base smoothing parameter estimation on a parametrization of $\mathbf{z}$ that uses $\boldsymbol{\mathcal{H}}$ and $\mathbf{g}$ as a whole instead of the n components that make them up. As elaborated in SM-A, this is advantageous in our context. Since our goal is to estimate $\boldsymbol{\lambda}$ so that the smooth terms' complexity which is not supported by the data is suppressed, the smoothing parameter vector is estimated so that $\hat{\boldsymbol{\mu}}_{\mathbf{z}}$ is as close as possible to $\boldsymbol{\mu}_{\mathbf{z}}$. Using this, for a given $\boldsymbol{\delta}^{[a+1]}$, the problem to minimize becomes

$$\boldsymbol{\lambda}^{[a+1]} = \arg\min_{\boldsymbol{\lambda}} \ \mathcal{V}(\boldsymbol{\lambda}) \stackrel{\text{def}}{=} \|\mathbf{z}^{[a+1]} - \mathbf{A}^{[a+1]}_{\boldsymbol{\lambda}^{[a]}}\mathbf{z}^{[a+1]}\|^2 - \check{n} + 2\text{tr}(\mathbf{A}^{[a+1]}_{\boldsymbol{\lambda}^{[a]}}),$$

where $\check{n} = 3n$, which is solved using the automatic stable and efficient computational routine by Wood (2004). Details on the derivation of the results stated above are provided in SM-A.

2.4 Further considerations

Estimation of $\boldsymbol{\lambda}$ is achieved using $\mathbf{g}$ and $\boldsymbol{\mathcal{I}}$ which are obtained as a byproduct of the estimation step for $\boldsymbol{\delta}$, hence little computational effort is required to set up the quantities needed for the smoothing step. The additional key benefit of using $\mathbf{z}$ and $\mathbf{A}$ as defined in the previous section is that the proposed smoothing approach is in principle suitable for any model fitted by penalized maximum likelihood. Consistency of the proposed estimator can be proved along the lines of Wojtys & Marra (2015), but this is beyond the scope of this paper.

At convergence, reliable point-wise confidence intervals for linear and non-linear functions of the model coefficients (e.g., smooth components, prevalence estimates, copula parameter) can be obtained using $\mathcal{N}(\hat{\boldsymbol{\delta}}, -\hat{\boldsymbol{\mathcal{H}}}_p^{-1})$. The rationale for using this result is provided in Marra & Wood (2012), and references therein, and some examples of interval construction are given in Radice et al. (2015). We can also test smooth components for equality to zero using the results discussed in Wood (2013a) and Wood (2013b). However, we do not deem this necessary as we followed the previous literature for variable selection (Bärnighausen et al., 2011; Hogan et al., 2012). HIV prevalence estimates are obtained using $\sum_{i=1}^{n} w_i\Phi(\hat{\eta}_{2i})/\sum_{i=1}^{n} w_i$, where the w_i are survey weights, while confidence intervals are derived using the delta method or posterior simulation using the above mentioned distributional result (e.g., McGovern et al., 2015b).

All the developments discussed in this paper have been implemented in `SemiParBIVProbit`. See SM-B for a brief description of the software.

For the reader's convenience, we have summarized the letters and symbols used in the paper and their corresponding meanings in the table reported in the last section of the Supplementary Material (SM-F).

3 Data

We implement the extended simultaneous equation model framework to estimate HIV prevalence in three sub-Saharan African countries. Data are obtained from the Demographic and Health Surveys (DHS) conducted in Zambia in 2007, Zimbabwe in 2005-2006, and Swaziland in 2006-2007. For further details on the DHS and HIV testing procedures, see Corsi et al. (2012), Fabic et al. (2012) and Mishra et al. (2006). Regional identifiers for respondents are used in this analysis, along with information on spatial boundaries at the sub-national level from http://gadm.org/. DHS are not designed to be representative below the regional level, and sampling within regions can be sparse. Therefore, in this analysis we focus on regional level heterogeneity in estimating HIV prevalence.

We follow the previous literature by including in $\mathbf{x}_i$ the variables described at the end of Section 2.2. Unlike the previous literature, we specify smooth functions of age, years of education, and wealth index (based on household assets) and employ Markov random field smoothers to model spatial variation. All these components enter into the linear predictors for participation (η_1) and HIV status (η_2). Exclusion restriction is achieved by including interviewer identity into η_1 only. We apply a ridge penalty to the coefficients of this variable in order to account for the difficulties associated with its use which we outlined in Section 1.1. Linear predictor η_3 only depends on a Markov random field term and allows for the copula association parameter to vary by region. All of our models are stratified by sex to reflect potentially sex-specific consent and HIV related factors. All our prevalence estimates are weighted to be nationally representative. We do not weight during model fitting as the variables on which the DHS weights are based are already included in the model (Hogan et al., 2012). Nevertheless, we have conducted a sensitivity analysis where we use the weights as part of the model fitting procedure and have found very similar results. Table 1 illustrates the sample size, number of regions, number of respondents who participate in testing, and the number of respondents who are HIV positive (among those who participate in testing) in each survey.

There are between 4 and 8 thousand observations in each country, with the percentage of eligi-

	Zambia		Zimbabwe		Swaziland	
	Men	Women	Men	Women	Men	Women
No. HIV$^-$	4,457	4,689	4,773	5,941	2,898	3,146
No. HIV$^+$	641	936	782	1,553	704	1,438
% HIV$^+$ (95% CI)	12% (11% - 13%)	16% (14% - 18%)	14% (13% - 16%)	21% (20% - 23%)	19% (18% - 21%)	31% (29% - 33%)
No. Declined to Test	1,318	1,400	1,620	1,413	554	403
No. Consented to Test	5,098	5,625	5,555	7,494	3,602	4,584
% Consented to Test (95% CI)	78% (76% - 80%)	79% (78% - 81%)	78% (76% - 80%)	84% (83% - 85%)	87% (86% - 89%)	92% (91% - 93%)
No. of Regions	9		10		4	

Table 1: Descriptive Statistics for Demographic and Health Survey HIV Data. HIV prevalence (%) and consent to test (%) estimates are weighted, and confidence intervals are clustered to account for survey design. HIV status is only available for those who consent to test. Individuals who were eligible but not contacted to test for HIV are not included in the analysis.

ble respondents consenting to test for HIV ranging from 78% for men in Zambia and Zimbabwe, to 92% for women in Swaziland. The percentage of HIV positive individuals (among those who consent to test) is high in all countries, and ranges from 12% for men in Zambia to 31% among women in Swaziland. Confidence intervals for the HIV prevalence estimates which do not account for non-participation are between 3 and 4 percentage points wide in each country.

In this paper, we focus on non-participation due to eligible respondents declining to test for HIV after interview. The amount of missing data due to this type of non-participation is typically more substantial than non-participation due to eligible respondents not being available for interview (Hogan et al., 2012). In addition, previous analysis of the Zambia data found little evidence of selection bias among this second group (Bärnighausen et al., 2011). The HIV datasets used for the analysis are freely available from http://www.measuredhs.com after registration, and the code for preparing the data can be obtained from http://hdl.handle.net/1902.1/17657 (Bärnighausen et al., 2011; Hogan et al., 2012). In the following section, we present new sex-specific national HIV prevalence point estimates and confidence intervals, and illustrate the regional heterogeneity in HIV prevalence and dependence parameter in each country.

4 Results

Table 2 presents national estimates of HIV prevalence (and associated confidence intervals) obtained from the simultaneous equation framework introduced in this paper. These are compared to imputation-based estimates shown in column 1, which only use the single linear predictor equation for HIV status (η_2). The reason we compare selection model results with imputation

		Imputation model	Selection model	
	Country	HIV Prevalance (95% CI)	HIV Prevalance (95% CI)	$\hat{\theta}$ (95% CI)
Men	Swaziland	$19.4\ (18.2, 20.6)$	$26.5\ (23.9, 29.2)$	$-4.09\ (-10.4, -1.82)$
	Zambia	$12.1\ (11.2, 12.9)$	$22.9\ (19.8, 26.0)$	$-8.45\ (-16.4, -4.25)$
	Zimbabwe	$14.4\ (13.5, 15.3)$	$14.5\ (12.4, 16.5)$	$-1.03\ (-22.7, -1.00)$
Women	Swaziland	$30.7\ (29.4, 31.9)$	$35.1\ (33.5, 36.8)$	$-9.83\ (-30.9, -3.91)$
	Zambia	$16.1\ (15.2, 17.1)$	$19.3\ (13.8, 24.8)$	$-1.40\ (-2.39, -1.07)$
	Zimbabwe	$20.5\ (19.6, 21.3)$	$21.7\ (19.2, 24.1)$	$-1.45\ (-3.79, -1.05)$

Table 2: National estimates of HIV prevalence (and associated confidence intervals) obtained from the single imputation and proposed simultaneous equation approaches. The estimates shown in column 1 do not account for potentially systematic non-participation whereas those in column 2 do. The dependence structure used for estimating the sample selection models is based on the Joe copula rotated by 90 degrees. Because we specify the dependence parameter in terms of a linear predictor, the values shown in column 3 are the average values in each country. Intervals are calculated using the inferential result mentioned in Section 2.4. The range of θ is $(-\infty, -1)$, with higher values (in absolute terms) indicating greater association; Figure 1 in SM-C shows three dependence scenarios.

estimates is that the latter is the recommended approach for dealing with missing data in HIV research by UNAIDS/WHO, and is also very popular in the applied literature for dealing with data affected by missingness. As was found in previous research, imputation estimates are almost identical to those in Table 1 which were based only on observations without missing data (Mishra et al., 2008; Marston et al., 2008; Hogan et al., 2012; Bärnighausen et al., 2011). Moreover, the imputation-based confidence intervals are, similarly, between 3 and 4 percentage points wide. Column 2, which shows our selection model estimates, which account for potentially systematic non-participation, indicate evidence of selection bias for men (Swaziland and Zambia) and women (Swaziland). In each of these cases, we can reject that the selection model point estimates are the same as the imputation-based approach, or analysis of observations without missing data.

In the final column of Table 2, we present estimates of the copula association parameter which measures the degree of association between participation in testing and HIV status (conditional on observed covariates). The values shown in column 3 are the average values in each country. The range of this parameter is $(-\infty, -1)$, with higher values (in absolute terms) indicating greater association. Three dependence scenarios for the 90° rotated Joe copula are illustrated in Figure 1 in SM-C. The precise definition of this parameter will vary according to the copula of interest.

If the copula parameter is close to -1 then there is lack of noticeable association between participation in testing and HIV status once observed characteristics have been adjusted for and hence no selection bias due to unobserved characteristics. This is the case for men in Zimbabwe, and women in Zambia and Zimbabwe, where in fact the selection model HIV prevalence estimates

are close to those of the imputation-based approach. However, even if point estimates are similar, we find that the imputation method substantially understates the amount of uncertainty associated with estimating HIV prevalence when survey testing data are affected by non-participation; confidence intervals obtained from the selection model are generally twice as wide as those from the single-equation approach.

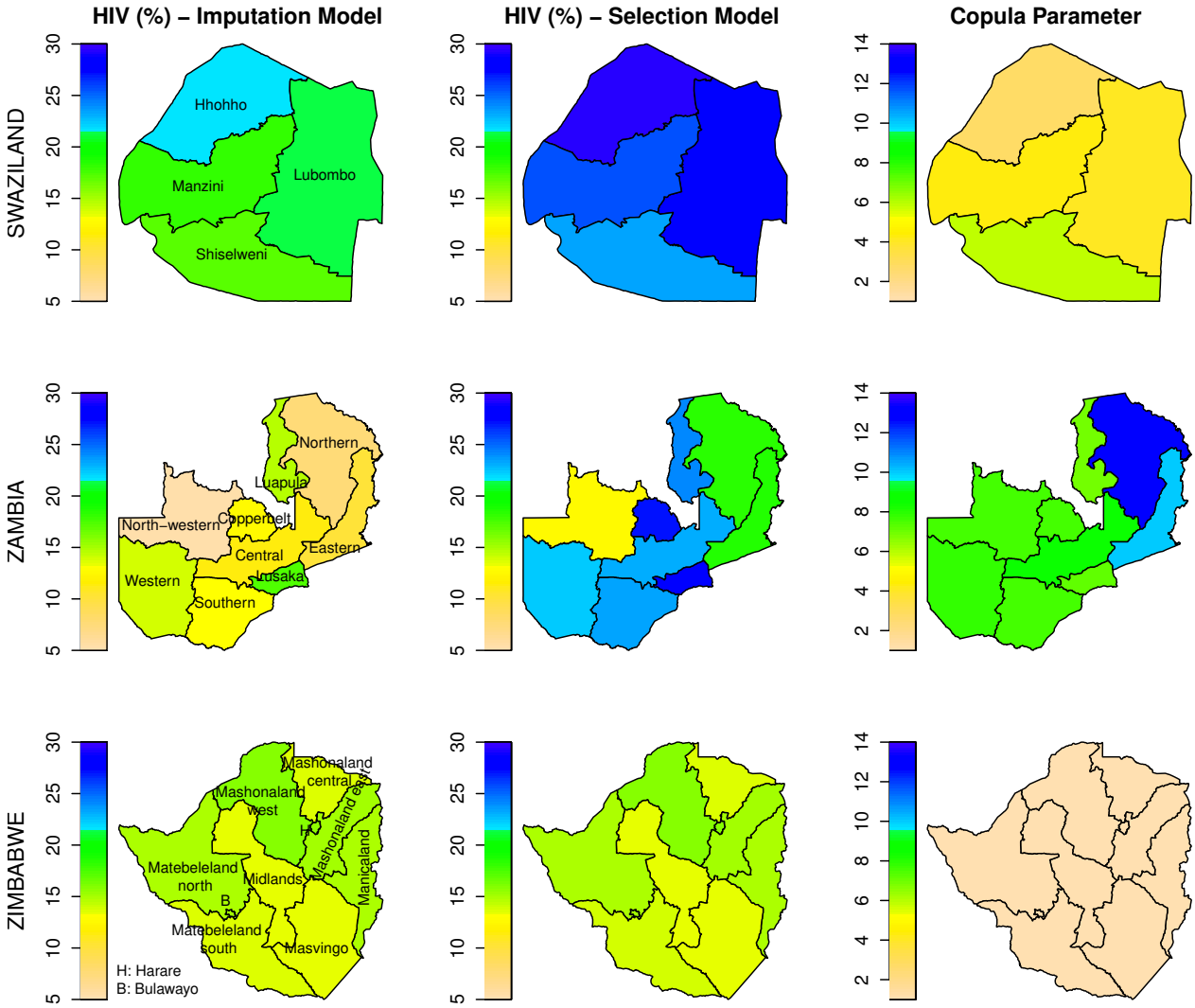


Figure 1: Sub-national HIV prevalence estimates for men obtained by applying the single imputation and proposed simultaneous equation approaches. The copula dependence parameter plot reports the estimated absolute values of θ with range $(1, \infty)$ in a Joe copula rotated by 90 degrees. The higher the value, the stronger the association between the selection and outcome equations.

We have considered a number of different dependence structures for estimating these models, the majority of which do not rely on the assumption of bivariate normality. Using the Akaike information criterion (AIC), we found that the Joe copula rotated by 90 degrees was the preferred choice for most cases, and therefore all estimates in Table 2 use this dependence structure.

Our sub-national HIV prevalence estimates, which are based on the region-specific copula de-

pendence parameters, are presented in Figures 1 (men) and 2 (women). There is clear variation in HIV prevalence within some countries, most notably for men in Zambia and women in Zambia and Zimbabwe, either on the basis of the imputation-based model, or the selection model estimates.

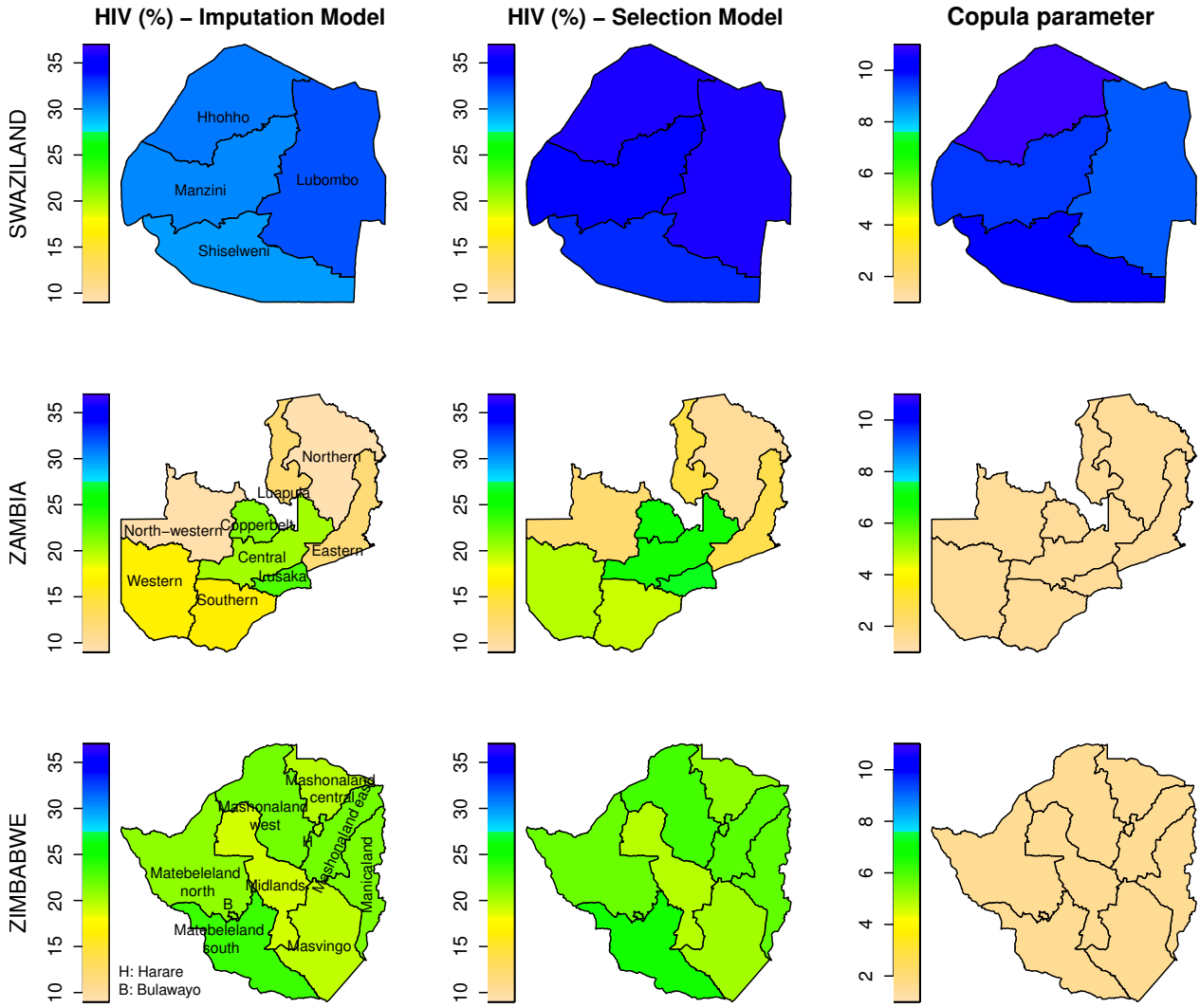


Figure 2: Sub-national HIV prevalence estimates for women obtained by applying the single imputation and proposed simultaneous equation approaches. The copula dependence parameter plot reports the estimated absolute values of θ with range $(1, \infty)$ in a Joe copula rotated by 90 degrees. The higher the value, the stronger the association between the selection and outcome equations.

For men in Zambia, the selection model HIV prevalence estimates range from 28% $(24\%, 32\%)$ in Lusaka to 13% $(7\%, 18\%)$ in Northwestern. For women in Zambia, the selection model HIV prevalences range from 26% $(19\%, 33\%)$ in Lusaka to 10% $(6\%, 14\%)$ in Northern. For women in Zimbabwe, the selection model HIV prevalences range from 25% $(21\%, 28\%)$ in Matebeleland South to 20% $(18\%, 22\%)$ in Midlands. Although the sample size is reduced when conducting sub-national analyses and confidence intervals are enlarged compared to the national prevalence estimates, most of these differences between highest and lowest prevalence regions show non-

overlapping intervals. In Swaziland, which is relatively more homogeneous, the selection model HIV prevalence estimates differ by 6 percentage points between the region with the highest prevalence (29% (26%, 32%) in Hhohho) and lowest prevalence (23% (21%, 26%) in Shiselweni) for men, and 3 percentage points between the region with the highest prevalence (36% (34%, 38%) in Hhohho) and lowest prevalence (33% (31%, 35%) in Shiselweni) for women. However, these estimates have overlapping intervals.

There is also support for a heterogeneous selection process across regions within some of these countries, as we find the copula dependence parameter varies according to location. For example, for men in Zambia, the selection model HIV prevalence for Northwestern is 8 percentage points greater than the imputation-based model (13% compared to 5%), while for Luapula, the difference is 9 percentage points (16% to 25%). In addition to this heterogeneity at the regional level, compared to a model which imposed homogeneity on the copula parameter, we found that this approach of allowing the dependence to reflect spatial variation was more efficient for estimating national HIV prevalence.

There are important non-linearities and functional form differences across sex and country in the association between observed characteristics of survey respondents and testing participation and HIV status outcomes, which highlights the relevance of our spline and penalized smoothing framework (see SM-E).

5 Sensitivity of results to violations of model assumptions

When dealing with data which are affected by non-participation or other missing information, it is necessary to make assumptions about the missing data mechanism. This is because we cannot observe the outcomes of interest for those individuals who do not consent to test for HIV, and therefore we can not simply test certain assumptions empirically (Nicoletti, 2006). To relax the assumption of missing at random conditional on observed covariates using the selection model framework, we require an alternative set of assumptions which describe the missing data mechanism. We argue that missing at random is not reasonable in the context of HIV surveys because those who are HIV positive have an incentive not to participate, and that the proposed framework is therefore much more realistic. However, it is important to critically assess the likely validity of these alternative assumptions. In this section, we discuss two approaches to evaluating the

parametric assumptions and exclusion restriction.

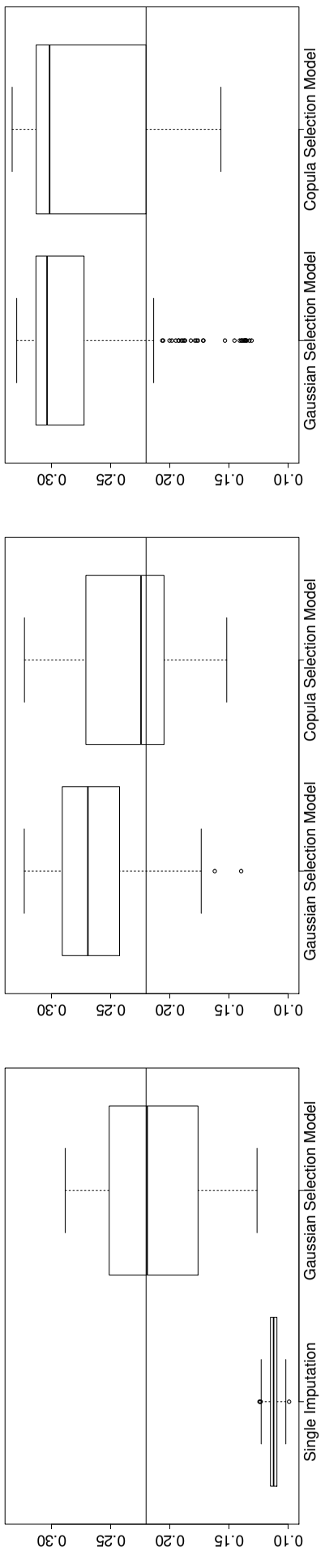
5.1 Simulation study

We assess the empirical effectiveness of the proposed sample selection modeling framework through a simulation study, in which we use the results presented in the previous section and employ parsimonious model settings to maintain feasibility. We constructed responses for consent and HIV status using several unobserved confounding variable distributions (normal, uniform and log-normal) and link functions (derived from the Gaussian, logistic and Weibull cumulative distribution functions). Imposition of assumptions about the model's link functions has been a criticism of selection models with continuous response (Kenward, 1998). For each of these nine combinations, we considered the situation in which the exclusion restriction assumption holds (i.e., interviewer identity predicts participation in HIV testing but not HIV status), and the cases where the assumption is mildly and strongly violated. Interest was in prevalence estimates. Exact simulation settings of the resulting 27 scenarios are given in SM-D.

We present results for the best-case and worst-case scenarios (called S0PG, S0WL and S1WL in Table 3 of SM-D). Figure 3 compares the results from the single imputation model, classic Heckman model (assuming bivariate normality) and the preferred copula selection model (as determined by the AIC). Figure 3a confirms that, when the Gaussian assumption holds and the exclusion restriction is valid, the traditional selection model is appropriate for correcting for systematic non-participation (bias in absolute value $= 1.6\%$ and root mean squared error (RMSE) $= 0.04$) and that the single imputation model performs poorly (bias $= 49\%$ and RMSE $= 0.107$). The wider variability of the selection model estimates as compared to those of single imputation is not surprising; imputation-based models do not acknowledge the uncertainty surrounding the relationship between participation in testing and HIV status. Figure 3b shows the results under model misspecification (Weibull link function and log-normal unobserved confounder) when the exclusion restriction holds. The performance of the Gaussian selection model worsens (bias $= 19.2\%$). In contrast, the preferred copula model (in this case the 90° rotated Joe, although the 90° rotated Gumbel and 270° rotated Clayton copulae unsurprisingly produced similar results) gives a bias of 6.5%. The corresponding RMSE is equal to 0.044 and is lower than that for the Gaussian selection model (RMSE $= 0.054$). In the absence of a valid instrument (in this case the independence of HIV status of the instrumental variable conditional on observed and unobserved covariates was

violated) and under model misspecification, the Gaussian and Joe 90° copula selection models perform poorly (see Figure 3c), although the latter is less biased and has lower RMSE as compared to the former (bias of 22% and 17% and RMSE of 0.072 and 0.067, respectively).

In summary, the simulations indicate that estimates obtained from the classic selection model can be biased when the model assumptions are not met, and that the proposed copula approach performs better. In particular, the copula selection model seems to be robust to situations in which the link functions are not Gaussian. It is worth pointing out that it is difficult to simulate the highly complex processes that likely underlie the relation between consent to HIV testing and HIV status. Nevertheless, our findings suggest that the copula approach has merit in dealing with non-random sample selection. In the absence of a valid exclusion restriction all models considered essentially deliver biased estimates. In general, it is not possible to determine a priori how the model assumptions will affect prevalence estimates as the data generating process is unknown. However, we believe that the proposed approach is a useful addition to the statistical toolbox as it can allow researchers to gain a better understanding of the sensitivity of estimation results to non-Gaussian specifications, for instance. These simulations clearly point to the validity of the exclusion restriction as a determinant of the performance of the selection approach; this is discussed in detail in the next section.



(a) Scenario S0PG (b) Scenario S0WL (c) Scenario S1WL

Figure 3: Simulation results of prevalence estimates obtained under the best-case (S0PG) and worst-case scenarios (S0WL and S1WL) considered in this paper. In S0PG the unobserved confounder distribution and cumulative distribution function (used to derive the link function) were both Gaussian. In S0WL and S1WL the unobserved confounder distribution and cumuative function were Log-normal and Weibull. In S0PG and S0WL a valid exclusion restriction was employed, whereas in S1WL the assumption that interviewer identity predicts participation in HIV testing but not HIV status was violated. The number of replicates was 250 and the horizontal lines represent the true prevalence. Prevalence estimates were obtained using the single imputation, classic Gaussian selection and preferred copula selection models. Exact simulation settings are given in SM-D.

5.2 Plausibility of the exclusion restriction

Since it is generally not possible to empirically test whether an exclusion restriction holds, we provide some arguments as to why interviewer assignment satisfies potentially this assumption. In particular, we explain how interviewers are allocated to respondents in the DHS and use an empirical approach which helps us gain some insights into the plausibility of this assumption.

The sampling procedure for the DHS is designed in two stages. First, a random sample of primary sampling units (PSU) is drawn where geographic locations are usually defined by a preceding census; PSU sampling is often stratified by urban/rural location, and/or region. Then, a random sample of households is chosen within each PSU, and all eligible residents of these households are sought for interview. There are several aspects of the DHS procedure which support the assumptions that interviewer identity satisfies the exclusion restriction assumption. Two-stage sampling is designed to provide a systematic way of selecting households to participate in the survey, and the DHS procedure recommends that households be pre-selected in the central office rather than by teams in the field. Therefore, the opportunity for interviewers to select who they interview is limited, and interviewer allocation by field supervisors is recommended to be made on the basis of equally distributing workload and linguistic capability (ICF International, 2015). If DHS guidelines are followed, the risk of bias associated with violation of the exclusion restriction seems low.

However, we do not expect pure random assignment of interviewers because of the presumption that interviewers be matched on language, and the fact that travel times per team may tend to be minimized (ICF International, 2012). We addressed this issue by controlling for language and region of the respondent in the model; this accounted for the fact that respondents with different languages and from different regions may have differential risk of being HIV positive. Nevertheless, there may be some small scale variation in HIV risk due to the fact that interviewers may tend to work in proximal areas within regions. Unfortunately, we cannot include an indicator variable for the PSU in the model as this would involve too many parameters. Nor can we include interviewer fixed effects in the HIV status equation as then we would no longer have an exclusion restriction. A potential solution is to follow Chamberlain (1980), Dustmann & Rochina-Barrachina (2007) and Mundlak (1978) who proposed approximating fixed effects using the mean of the observed characteristics of the group of interest. We have, therefore, conducted a sensitivity anal-

ysis where we included the mean characteristics of each interviewer's interviewees, say $\bar{\mathbf{x}}_i$, as additional predictors in the HIV status outcome equation. Because we expect interviewers and respondents to be matched on region and language, we could not include these in $\bar{\mathbf{x}}_i$ but these controls remained in $\mathbf{x}_i$. Using this empirical approach led to very similar results to those in the main analysis, hence suggesting that the assumption of exclusion restriction is reasonable.

An alternative way of approaching the validity of the exclusion restriction is provided by Angrist et al. (1996). These authors theoretically derive the bias associated with violation of the exclusion restriction in their application. Due to the complexity of our model, it is not clear how to derive the relevant bias theoretically. However, the simulation results provide us with an indication of the potential consequences for our estimates if the exclusion restriction is violated. We focus on the worst case scenario (S1WL) where the model was misspecified and the exclusion restriction was not valid. In this case, a bias of 17% was found for the best performing selection model. If we assume that violation of the exclusion restriction arises because good interviewers are more likely to be assigned to respondents who are more likely to be HIV positive, then our selection model estimates will be upward biased. Considering the cases in which there is substantial difference between the selection and imputation estimates, we have that, for men in Swaziland, if the selection estimates are biased upwards by 17%, then the true HIV prevalence is $26.5 - 3.9 = 22.6\%$ (compared to the imputation estimate of 19.4%). For men in Zambia, the bias-corrected estimate would be $22.9 - 3.3 = 19.6\%$ (compared to the imputation estimate of 12.1%). Therefore, the results presented in this article indicate substantial concern about the validity of the assumption of missing at random, at a minimum for the surveys among men in Swaziland and Zambia, even if some or all of the selection model assumptions do not fully hold.

6 Discussion

Nationally representative datasets containing information on HIV status conducted through home-based testing have made an important contribution to our understanding of the evolution of the HIV epidemic. However, non-participation in testing as part of these surveys can lead to substantial amounts of missing data, and missing at random may not be a realistic assumption. In this article, we have developed a simultaneous equation framework which extends the capabilities of Heckman-type selection models. Our results for Zambia, Zimbabwe, and Swaziland indicate that some DHS HIV surveys are likely to be affected by selection bias. Using our modeling

framework, we find that HIV prevalence estimates are substantially higher than, and statistically different from, those found by either the imputation-based single equation approach or the analysis of cases without missing data for men in Swaziland and Zambia, and women in Swaziland. We also find that not accounting for the relationship between participation in testing and HIV status yields confidence intervals that are too narrow as they do not reflect the true uncertainty associated with surveys which are affected by systematic non-participation.

Our sub-national estimates indicate that there is clear variation in HIV prevalence within some countries and that the dependence parameter varies according to location, hence supporting the developed framework. Because the copula parameter models unobserved characteristics, it is difficult to concretely assess what could be driving these regional differences. It seems reasonable that the incentive to participate in testing for HIV positive individuals, hence the unmeasured dependence, would vary by location. For example, we would expect areas where the stigma associated with HIV was greatest to exhibit the greatest negative unmeasured dependence because of the greater consequences of disclosure. We have attempted to find comparable data on HIV stigma to assess which countries and regions were most likely to be affected, however we were unable to locate such data; investigating the reasons underlying this heterogeneity is an important direction for future research. We cannot conclusively rule out that the exclusion restriction is less likely to be valid in some locations. However, given that the imputation-based model also implies substantial heterogeneity in HIV prevalence, it seems implausible that these differences could be largely attributed to violation of the exclusion restriction in certain areas.

In this paper, we have focused on HIV testing, however there are many other contexts in which biomarker data collection is affected by non-participation. More broadly, there are many instances of missing data in medical and social science surveys in which the assumption of missing at random may be unrealistic due to the existence of plausible behavioral mechanisms leading to selection bias. The methodology we introduced in this paper, therefore, has wide range of potential applications outside of HIV research. The approach proposed in this paper is flexible and can easily be applied to other countries and contexts, and the software for doing so has been designed specifically with this in mind. The main requirement to adopt this approach is a valid selection variable in the survey of interest, and in many contexts interviewer identity is a plausible choice as a selection variable and is often available. However, future surveys could be designed with

this methodology in mind, for example by providing additional meta-data to act as selection variables, documenting survey procedure, or implementing specific randomized interventions aimed at increasing participation.

From a methodological point of view, it would be interesting to explore the use of semi/non-parametric copula approaches. These would allow the margins and/or the copula to be estimated non-parametrically using, for instance, smoothing methods such as kernels, wavelets and orthogonal polynomials. If the specification of the model for the margins and copula is correct, then the parametric approach will outperform semi/non-parametric methods; however, the reverse will be true under misspecification. Without any plausible prior information, semi/non-parametric techniques should be favored as they will be more flexible in determining the shape of the underlying distribution. However, in practice, such techniques are typically limited with regard to the inclusion of covariates and very flexible linear predictor structures, may require the imposition of restrictions on the functions approximating the underlying distribution and may be computationally demanding (e.g., Kauermann et al., 2013; Segers et al., 2014; Shen et al., 2008). Future research will determine the feasibility of such developments.

Acknowledgement

We are indebted to the Editor, Associate Editor and two anonymous reviewers for many detailed and well thought out suggestions which helped to clarify the contribution and the presentation of the paper.

References

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434), 444–455.

Aral, S. O., Padian, N. S., & Holmes, K. K. (2005). Advances in multilevel approaches to understanding the epidemiology and prevention of sexually transmitted infections and HIV: an overview. *Journal of Infectious Diseases*, 191(Supplement 1), S1–S6.

Arpino, B., Cao, E. D., & Peracchi, F. (2014). Using panel data for partial identification of Human Immunodeficiency Virus prevalence when infection status is missing not at random. *Journal of the Royal Statistical Society: Series A*, 177(3), 587–606.

Bärnighausen, T., Bor, J., Wandira-Kazibwe, S., & Canning, D. (2011). Correcting HIV prevalence estimates for survey nonparticipation using Heckman-type selection models. *Epidemiology*, 22(1), 27–35.

Bärnighausen, T., Bor, J., Wandira-Kazibwe, S., & Canning, D. (2011). Interviewer identity as exclusion restriction in epidemiology. *Epidemiology*, 22(3), 446.

Bärnighausen, T., Tanser, F., Malaza, A., Herbst, K., & Newell, M.-L. (2012). HIV status and participation in HIV surveillance in the era of antiretroviral treatment: a study of linked population-based and clinical data in rural South Africa. *Tropical Medicine and International Health*, 17(8), e103–e110.

Beyrer, C., Baral, S., Kerrigan, D., El-Bassel, N., Bekker, L.-G., & Celentano, D. D. (1999). Expanding the space: Inclusion of most-at-risk populations in HIV prevention, treatment, and care services. *Journal of Acquired Immune Deficiency Syndromes*, 57(Suppl 2), S96.

Boerma, J. T., Ghys, P. D., & Walker, N. (2003). Estimates of HIV-1 prevalence from national population-based surveys as a new gold standard. *Lancet*, 362(9399), 1929–1931.

Butler, J. S. (1996). Estimating the correlation in censored probit models. *Review of Economics and Statistics*, 78(2), 356–358.

Chamberlain, G. (1980). Analysis of covariance with qualitative data. *Review of Economic Studies*, 47(1), 225–238.

Clark, S. J. & Houle, B. (2014). Validation, replication, and sensitivity testing of heckman-type selection models to adjust estimates of HIV prevalence. *PloS one*, 9, e112563.

Corsi, D. J., Neuman, M., Finlay, J. E., & Subramanian, S. (2012). Demographic and Health Surveys: a profile. *International Journal of Epidemiology*, 41(6), 1602–1613.

De Cock, K. M., Bunnell, R., & Mermin, J. (2006). Unfinished business: expanding HIV testing in developing countries. *New England Journal of Medicine*, 354(5), 440–442.

Donders, A. R. T., van der Heijden, G. J., Stijnen, T., & Moons, K. G. (2006). Review: a gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, 59(10), 1087–1091.

Dubin, J. A. & Rivers, D. (1989). Selection bias in linear regression, logit and probit models. *Sociological Methods & Research*, 18(2-3), 360–390.

Dustmann, C. & Rochina-Barrachina, M. E. (2007). Selection correction in panel data models: An application to the estimation of females' wage equations. *Econometrics Journal*, 10(2), 263–293.

Efron, B. & Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected fisher information. *Biometrika*, 65(3), 457–483.

Eilers, P. H. C. & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2), 89–121.

Fabic, M. S., Choi, Y., & Bird, S. (2012). A systematic review of Demographic and Health Surveys: data availability and utilization for research. *Bulletin of the World Health Organization*, 90(8), 604–612.

Floyd, S., Molesworth, A., Dube, A., Crampin, A. C., Houben, R., Chihana, M., Price, A., Kayuni, N., Saul, J., & French, N. (2013). Underestimation of HIV prevalence in surveys when some people already know their status, and ways to reduce the bias. *AIDS*, 27(2), 233–242.

Geyer, C. J. (2015). *trust: Trust Region Optimization*. R package version 0.1-6.

Hastie, T. & Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society Series B*, 55(4), 757–796.

Heckman, J. (1990). Varieties of selection bias. *American Economic Review*, 80(2), 313–318.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1), 153–161.

Heinze, G. & Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine*, 21(16), 2409–2419.

Hogan, D. R., Salomon, J. A., Canning, D., Hammitt, J. K., Zaslavsky, A. M., & Bärnighausen, T. (2012). National HIV prevalence estimates for sub-Saharan Africa: controlling selection bias with Heckman-type selection models. *Sexually Transmitted Infections*, 88(Suppl 2), i17–i23.

ICF International (2012). *Survey Organization Manual for Demographic and Health Surveys*. Technical report, MEASURE DHS, Calverton. Maryland: ICF International.

ICF International (2015). *Demographic and Health Survey Supervisor's and Editor's manual*. Technical report, The Demographic and Health Survey Program, Rockville, Maryland, U.S.A.: ICF International.

Janssens, W., van der Gaag, J., de Wit, T., & Tanović, Z. (2014). Refusal bias in the estimation of HIV prevalence. *Demography*, 51(3), 1131–1157.

Kauermann, G., Schellhase, C., & Ruppert, D. (2013). Flexible copula density estimation with penalized hierarchical b-splines. *Scandinavian Journal of Statistics*, 40(4), 685–705.

Kenward, M. G. (1998). Selection models for repeated measurements with non-random dropout: an illustration of sensitivity. *Statistics in Medicine*, 17(23), 2723–2732.

Klovdahl, A. S. (1985). Social networks and the spread of infectious diseases: the AIDS example. *Social Science & Medicine*, 21(11), 1203–1216.

Kolev, N. & Paiva, D. (2009). Copula-based regression models: A survey. *Journal of Statistical Planning and Inference*, 139, 3847–3856.

Kranzer, K., McGrath, N., Saul, J., Crampin, A. C., Jahn, A., Malema, S., Mulawa, D., Fine, P. E., Zaba, B., & Glynn, J. R. (2008). Individual, household and community factors associated with HIV test refusal in rural Malawi. *Tropical Medicine & International Health*, 13(11), 1341–1350.

Larmarange, J. & Bendaud, V. (2014). HIV estimates at second subnational level from national population-based surveys. *AIDS*, 28(Supp), S469–S476.

Madden, D. (2008). Sample selection versus two-part models revisited: the case of female smoking and drinking. *Journal of Health Economics*, 27(2), 300–307.

Marra, G. & Radice, R. (2013). A penalized likelihood estimation approach to semiparametric sample selection binary response modeling. *Electronic Journal of Statistics*, 7, 1432–1455.

Marra, G. & Radice, R. (2016). *SemiParBIVProbit: Semiparametric Bivariate Probit Modelling*. R package version 3.6.1.

Marra, G. & Wood, S. (2012). Coverage properties of confidence intervals for generalized additive model components. *Scandinavian Journal of Statistics*, 39(1), 53–74.

Marston, M., Harriss, K., & Slaymaker, E. (2008). Non-response bias in estimates of HIV prevalence due to the mobility of absentees in national population-based surveys: a study of nine national surveys. *Sexually Transmitted Infections*, 84(Suppl 1), i71–i77.

McGovern, M., Bärnighausen, T., Salomon, J., & Canning, D. (2015a). Using interviewer random effects to remove selection bias from HIV prevalence estimates. *BMC Medical Research Methodology*, 15, 8.

McGovern, M. E., Bärnighausen, T., Marra, G., & Radice, R. (2015b). On the assumption of bivariate normality in selection models: A copula approach applied to estimating HIV prevalence. *Epidemiology*, 26(2), 229–237.

Mishra, V., Barrere, B., Hong, R., & Khan, S. (2008). Evaluation of bias in HIV seroprevalence estimates from national household surveys. *Sexually Transmitted Infections*, 84(Suppl 1), i63–i70.

Mishra, V., Vaessen, M., Boerma, J., Arnold, F., Way, A., Barrere, B., Cross, A., Hong, R., & Sangha, J. (2006). HIV testing in national population-based surveys: experience from the Demographic and Health Surveys. *Bulletin of the World Health Organization*, 84(7), 537–545.

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica*, 46(1), 69–85.

Nicoletti, C. (2006). Nonresponse in dynamic panel data models. *Journal of Econometrics*, 132(2), 461–489.

Nocedal, J. & Wright, S. J. (2006). *Numerical Optimization*. New York: Springer-Verlag.

Obare, F. (2010). Nonresponse in repeat population-based voluntary counseling and testing for HIV in rural Malawi. *Demography*, 47(3), 651–665.

Puhani, P. (2000). The Heckman correction for sample selection and its critique. *Journal of Economic Surveys*, 14(1), 53–68.

Radice, R., Marra, G., & Wojtys, M. (2015). Copula regression spline models for binary outcomes. *Statistics and Computing*, Forthcoming.

Reniers, G. & Eaton, J. (2009). Refusal bias in HIV prevalence estimates from nationally representative seroprevalence surveys. *AIDS*, 23(5), 621–629.

Rue, H. & Held, L. (2005). *Gaussian Markov Random Fields*. New Haven: Chapman & Hall/CRC, Boca Raton, FL.

Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press, New York.

Segers, J., van den Akker, R., & Werker, B. J. M. (2014). Linear b-spline copulas with applications to nonparametric estimation of copulas. *Annals of Statistics*, 42, 1911–1940.

Shen, X., Zhu, Y., & Song, L. (2008). Linear b-spline copulas with applications to nonparametric estimation of copulas. *Computational Statistics and Data Analysis*, 52(7), 3806–3819.

Sklar, A. (1959). Fonctions de répartition é n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8, 229–231.

Sklar, A. (1973). Random variables, joint distributions, and copulas. *Kybernetica*, 9, 449–460.

Tanser, F., Bärnighausen, T., Cooke, G. S., & Newell, M.-L. (2009). Localized spatial clustering of HIV infections in a widely disseminated rural South African epidemic. *International Journal of Epidemiology*, 38(4), 1008–1016.

Van de Ven, W. P. & Van Praag, B. (1981). The demand for deductibles in private health insurance: A probit model with sample selection. *Journal of Econometrics*, 17(2), 229–252.

Vella, F. (1998). Estimating models with sample selection bias: a survey. *Journal of Human Resources*, 33(1), 127–169.

Wojtys, M. & Marra, G. (2015). Copula based generalized additive models with non-random sample selection. *arXiv:1508.04070.*

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society Series B*, 65(1), 95–114.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, 99(467), 673–686.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction With R*. Chapman & Hall/CRC, London.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B*, 73(1), 3–36.

Wood, S. N. (2013a). On p-values for smooth components of an extended generalized additive model. *Biometrika*, 100(1), 221–228.

Wood, S. N. (2013b). A simple test for random effects in regression models. *Biometrika*, 100(4), 1005–1010.

Zimmer, D. M. & Trivedi, P. K. (2006). Using trivariate copulas to model sample selection and treatment effects: application to family health care demand. *Journal of Business & Economic Statistics*, 24, 63–76.

Supplementary Material A: Justification of smoothing approach

Data-driven and automatic estimation of smoothing parameters is pivotal for practical modeling, especially when each model equation contains more than one smooth component (as in our case study). Estimating the effects of individual-level predictors may not be straightforward and in HIV studies continuous variables are typically entered into the equations as parametric components, polynomials of various degrees, or else categorized according to a series of cut-points. This approach runs the risk of under/over-fitting, may be inefficient, and can be arbitrary. Because some portion of the data are missing, often a substantial percentage, it can be difficult to reliably specify these choices ex ante. Moreover, the degrees of the relevant polynomial or the effective cut-points can be difficult to set in general because they may vary according to the context. For example, years of education in one country could have a different meaning to years of education in another, and specifying education groups according to some common threshold could be inappropriate. This is an important issue because identifying the relevant associations requires an appropriate flexible specification of the covariate effects. In addition, in the absence of a strong selection variable which is sufficiently predictive of the selection outcome, model identification can in theory be achieved through non-linearities and hence misspecification of the model component effects could introduce bias into the results (Madden, 2008). Misspecification of the linear predictor equations could also result in inducing a violation of the assumed model's bivariate distribution typically required for identification, even if this assumption holds under the correct model specification.

To this end, we employ a penalized regression spline approach which allows us to estimate flexibly non-linear effects and does not depend on arbitrary modeling decisions by the researcher (e.g., Marra & Radice, 2013; Ruppert et al., 2003; Wood, 2006). For example, modeling the association of age with HIV status is crucial for understanding when peak incidence occurs, and such evidence can be used for appropriate targeting of efforts to reduce risky behavior (Gouws et al., 2008). The role of education in the evolution of the HIV epidemic is another question of fundamental importance to policy makers due to its potential for affecting population health, behavior and knowledge. However the literature has found its impact as protective or to be changing over time (Hargreaves et al., 2008). Finally, the literature has debated the association of poverty with HIV risk (Gillespie et al., 2007). If any of these factors (age, education and poverty, which we measure with household wealth defined by an asset index) are systematically associated with the

outcomes of interest, and such relationships are not modeled flexibly and reliably, then results could be misleading.

Radice et al. (2015) and Marra & Radice (2013) discussed a smoothing approach for bivariate equation models with penalized regression splines which is based on $\mathbf{z} = \sqrt{\mathbf{W}}\left(\mathbf{W}^{-1}\mathbf{d} + \mathbf{Z}\boldsymbol{\delta}\right)$. Loosely speaking, $\mathbf{W}$ is of dimensions $\check{n} \times \check{n}$, where $\check{n} = 3n$, and represents a block diagonal weight matrix containing minus the second derivatives of the log-likelihood with respect to η_1, η_2 and η_3, $\mathbf{d}$ is a vector of length $\check{n}$ containing the first derivatives of the log-likelihood with respect to η_1, η_2 and η_3, and $\mathbf{Z}$ is an overall design matrix of dimensions $\check{n} \times m$, where m is the total number of columns, which has a block diagonal structure and contains the design matrices associated with η_1, η_2 and η_3. Pseudodata vector $\mathbf{z}$ requires $\mathbf{W}$ be positive definite. Unfortunately, when the copula parameter is specified as a function of covariates and/or the model is highly flexible, the n weight matrices contained in $\mathbf{W} = \operatorname{diag}\left(\mathbf{W}_1, \ldots, \mathbf{W}_n\right)$ need not all be positive definite, and in practice a non-negligible number of non-positive definite $\mathbf{W}_i$ may be encountered for perfectly reasonable models (see, e.g., Wood (2011) for an example in a related context). Therefore, positive definiteness can only guaranteed if $\mathbb{E}\left(\mathbf{W}\right)$ is used in place of $\mathbf{W}$. However, as in Wood (2011), we generally found observed information to be superior in terms of speed, stability and accuracy of results (Efron & Hinkley, 1978). All this suggests employing observed information and basing smoothing parameter estimation on a parametrization of $\mathbf{z}$ that uses $\boldsymbol{\mathcal{H}}$ and $\mathbf{g}$ as a whole instead of the n components that make them up. There will clearly be situations in which $\boldsymbol{\mathcal{H}}$ is not positive definite but these would occur considerably less frequently than when working with the n weight matrices that make it up, and can be addressed by perturbing $\boldsymbol{\mathcal{H}}$ to positive definiteness (e.g., Wood, 2015, Chapter 5). The additional advantage of such an approach is that $\boldsymbol{\mathcal{H}}$ and $\mathbf{g}$ would be obtained as a byproduct of the estimation step for $\boldsymbol{\delta}$, hence little computational effort will be required to set up the pseudodata vector needed for the smoothing step.

Using the quantities and notation defined in Section 3, recall that a first order Taylor expansion of $\mathbf{g}_p^{[a+1]}$ about $\boldsymbol{\delta}^{[a]}$ yields $\mathbf{0} = \mathbf{g}_p^{[a+1]} \approx \mathbf{g}_p^{[a]} + \left(\boldsymbol{\delta}^{[a+1]} - \boldsymbol{\delta}^{[a]}\right)\boldsymbol{\mathcal{H}}_p^{[a]}$, where $\mathbf{g}_p^{[a]} = \mathbf{g}^{[a]} - \mathbf{S}_{\boldsymbol{\lambda}^{[a]}}\boldsymbol{\delta}^{[a]}$ and $\boldsymbol{\mathcal{H}}_p^{[a]} = \boldsymbol{\mathcal{H}}^{[a]} - \mathbf{S}_{\boldsymbol{\lambda}^{[a]}}$. As explained above, finding an expression for $\boldsymbol{\delta}^{[a+1]}$ that is based on $\mathbf{g}^{[a]}$ and $\boldsymbol{\mathcal{H}}^{[a]}$ is crucial to our developments and it can be obtained as follows. Let us define $\boldsymbol{\mathcal{I}}^{[a]} = -\boldsymbol{\mathcal{H}}^{[a]}$,

we then have

$$\begin{aligned}
\mathbf{0} &= \mathbf{g}_p^{[a]} + \left(\boldsymbol{\delta}^{[a+1]} - \boldsymbol{\delta}^{[a]}\right)\left(-\boldsymbol{\mathcal{I}}^{[a]} - \mathbf{S}_{\boldsymbol{\lambda}^{[a]}}\right), \\
\mathbf{g}_p^{[a]} &= \left(\boldsymbol{\delta}^{[a+1]} - \boldsymbol{\delta}^{[a]}\right)\left(\boldsymbol{\mathcal{I}}^{[a]} + \mathbf{S}_{\boldsymbol{\lambda}^{[a]}}\right), \\
\mathbf{g}^{[a]} - \mathbf{S}_{\boldsymbol{\lambda}^{[a]}}\boldsymbol{\delta}^{[a]} &= \boldsymbol{\delta}^{[a+1]}\left(\boldsymbol{\mathcal{I}}^{[a]} + \mathbf{S}_{\boldsymbol{\lambda}^{[a]}}\right) - \boldsymbol{\delta}^{[a]}\boldsymbol{\mathcal{I}}^{[a]} - \boldsymbol{\delta}^{[a]}\mathbf{S}_{\boldsymbol{\lambda}^{[a]}}, \\
\boldsymbol{\delta}^{[a+1]}\left(\boldsymbol{\mathcal{I}}^{[a]} + \mathbf{S}_{\boldsymbol{\lambda}^{[a]}}\right) &= \mathbf{g}^{[a]} + \boldsymbol{\delta}^{[a]}\boldsymbol{\mathcal{I}}^{[a]}, \\
\boldsymbol{\delta}^{[a+1]} &= \left(\boldsymbol{\mathcal{I}}^{[a]} + \mathbf{S}_{\boldsymbol{\lambda}^{[a]}}\right)^{-1}\sqrt{\boldsymbol{\mathcal{I}}^{[a]}}\left(\sqrt{\boldsymbol{\mathcal{I}}^{[a]}}\boldsymbol{\delta}^{[a]} + \sqrt{\boldsymbol{\mathcal{I}}^{[a]}}^{-1}\mathbf{g}^{[a]}\right).
\end{aligned}$$

Therefore, the parameter estimator can be expressed as

$$\boldsymbol{\delta}^{[a+1]} = \left(\boldsymbol{\mathcal{I}}^{[a]} + \mathbf{S}_{\boldsymbol{\lambda}^{[a]}}\right)^{-1}\sqrt{\boldsymbol{\mathcal{I}}^{[a]}}\mathbf{z}^{[a]},$$

where $\mathbf{z}^{[a]} = \boldsymbol{\mu}_{\mathbf{z}}^{[a]} + \boldsymbol{\epsilon}^{[a]}$ with $\boldsymbol{\mu}_{\mathbf{z}}^{[a]} = \sqrt{\boldsymbol{\mathcal{I}}^{[a]}}\boldsymbol{\delta}^{[a]}$ and $\boldsymbol{\epsilon}^{[a]} = \sqrt{\boldsymbol{\mathcal{I}}^{[a]}}^{-1}\mathbf{g}^{[a]}$. The square root of $\boldsymbol{\mathcal{I}}$ and its inverse are obtained by eigen-value decomposition. Note that, to within an additive constant, pseudodata vector $\mathbf{z}$ is also a quadratic approximation to the model log-likelihood in the vicinity of the converged parameter vector, since they share first and expected second derivatives with respect to $\boldsymbol{\delta}$. From likelihood theory, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{z}}, \mathbf{I})$, where $\mathbf{I}$ is an identity matrix, $\boldsymbol{\mu}_{\mathbf{z}} = \sqrt{\boldsymbol{\mathcal{I}}}\boldsymbol{\delta}^0$ and $\boldsymbol{\delta}^0$ is the true parameter vector. The predicted value vector for $\mathbf{z}$ is $\hat{\boldsymbol{\mu}}_{\mathbf{z}} = \sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\delta}} = \mathbf{A}_{\hat{\boldsymbol{\lambda}}}\mathbf{z}$, where $\mathbf{A}_{\hat{\boldsymbol{\lambda}}} = \sqrt{\boldsymbol{\mathcal{I}}}\left(\boldsymbol{\mathcal{I}} + \mathbf{S}_{\hat{\boldsymbol{\lambda}}}\right)^{-1}\sqrt{\boldsymbol{\mathcal{I}}}$. Since our goal is to estimate $\boldsymbol{\lambda}$ so that the smooth terms' complexity which is not supported by the data is suppressed, the smoothing parameter vector is estimated so that $\hat{\boldsymbol{\mu}}_{\mathbf{z}}$ is as close as possible to $\boldsymbol{\mu}_{\mathbf{z}}$. Therefore, we use

$$\begin{aligned}
\mathbb{E}\left(\|\boldsymbol{\mu}_{\mathbf{z}} - \hat{\boldsymbol{\mu}}_{\mathbf{z}}\|^2\right) &= \mathbb{E}\left(\|\left(\mathbf{z} - \boldsymbol{\epsilon}\right) - \mathbf{A}_{\boldsymbol{\lambda}}\mathbf{z}\|^2\right) = \mathbb{E}\left(\|\mathbf{z} - \mathbf{A}_{\boldsymbol{\lambda}}\mathbf{z} - \boldsymbol{\epsilon}\|^2\right) \\
&= \mathbb{E}\left(\|\mathbf{z} - \mathbf{A}_{\boldsymbol{\lambda}}\mathbf{z}\|^2\right) + \mathbb{E}\left(-\boldsymbol{\epsilon}^{\mathsf{T}}\boldsymbol{\epsilon} - 2\boldsymbol{\epsilon}^{\mathsf{T}}\boldsymbol{\mu}_{\mathbf{z}} + 2\boldsymbol{\epsilon}^{\mathsf{T}}\mathbf{A}_{\boldsymbol{\lambda}}\boldsymbol{\mu}_{\mathbf{z}} + 2\boldsymbol{\epsilon}^{\mathsf{T}}\mathbf{A}_{\boldsymbol{\lambda}}\boldsymbol{\epsilon}\right) \qquad (1) \\
&= \mathbb{E}\left(\|\mathbf{z} - \mathbf{A}_{\boldsymbol{\lambda}}\mathbf{z}\|^2\right) - \check{n} + 2\mathrm{tr}(\mathbf{A}_{\boldsymbol{\lambda}}),
\end{aligned}$$

where $\check{n} = 3n$ and $\mathrm{tr}(\mathbf{A}_{\boldsymbol{\lambda}})$ is the number of effective degrees of freedom of the penalized model. Line 2 is obtained by expanding the square in line 1. The last line follows from line 2 by recalling the properties of $\boldsymbol{\epsilon}$ and that a scalar is its own trace. In practice, $\boldsymbol{\lambda}$ is estimated by minimizing an

estimate of (1), i.e.

$$\mathcal{V}(\boldsymbol{\lambda}) = \|\widehat{\boldsymbol{\mu}_{\mathbf{z}} - \hat{\boldsymbol{\mu}}_{\mathbf{z}}}\|^2 = \|\mathbf{z} - \mathbf{A}_{\boldsymbol{\lambda}}\mathbf{z}\|^2 - \check{n} + 2\text{tr}(\mathbf{A}_{\boldsymbol{\lambda}}). \tag{2}$$

Given $\boldsymbol{\delta}^{[a+1]}$, the problem becomes

$$\boldsymbol{\lambda}^{[a+1]} = \arg\min_{\boldsymbol{\lambda}} \ \mathcal{V}(\boldsymbol{\lambda}) \stackrel{\text{def}}{=} \|\mathbf{z}^{[a+1]} - \mathbf{A}^{[a+1]}_{\boldsymbol{\lambda}^{[a]}}\mathbf{z}^{[a+1]}\|^2 - \check{n} + 2\text{tr}(\mathbf{A}^{[a+1]}_{\boldsymbol{\lambda}^{[a]}}),$$

which is solved using the automatic stable and efficient computational routine by Wood (2004). This approach is based on Newton's method and can evaluate in an efficient and stable way the components in $\mathcal{V}(\boldsymbol{\lambda})$ and their first and second derivatives with respect to $\log(\boldsymbol{\lambda})$ (since the smoothing parameters can only take positive values). Note that, to within an additive constant, the first term on the right hand side of (2) is a quadratic approximation to $-2\ell(\hat{\boldsymbol{\delta}})$. Therefore, dropping irrelevant constants yields $\mathcal{V}(\boldsymbol{\lambda}) \propto -2\ell(\hat{\boldsymbol{\delta}}) + 2\text{tr}(\mathbf{A}_{\boldsymbol{\lambda}})$. This means that smoothing parameters would be estimated to minimize what is effectively the Akaike information criterion with effective degrees of freedom instead of number of parameters. Finally, it is worth stressing that another key benefit of using $\mathbf{z}$ and $\mathbf{A}$ as defined above is that the proposed smoothing approach can in principle be applied to any situation in which a model is fitted by penalized maximum likelihood.

Supplementary Material B: Software implementation

The framework this paper provides allows researchers and policy-makers to apply a transparent approach to account for systematic non-participation in their data. The features of this software have been designed specifically with transparent and straightforward dissemination of results in mind. First, the choice of optimization algorithm and confidence interval procedure allow for results to be obtained relatively quickly without the need for bootstrapping or complex simulation methods. Second, model fitting is designed to avoid arbitrary decisions by the researcher (e.g., pooling of interviewers, polynomial or cut-point specification for the effects of continuous variables) to the maximum extent possible. Finally, national HIV prevalence estimates and adjusted confidence intervals (which account for the uncertainty inherent in estimating the relationship between testing participation and HIV status) can be obtained directly as the primary output of the

model, along with sub-national spatial maps for HIV prevalence and associational graph for the relevant covariates of interest, as shown for instance in SS-E.

We have implemented the proposed approach in `R` (R Development Core Team, 2016), by extending the package `SemiParBIVProbit` (Marra & Radice, 2016) so that the main function `SemiParBIVProbit()` can estimate all the models mentioned in this paper. The function should be easy to use for anyone familiar with (generalized) linear and additive models in `R`. For the copula selection models, the user simply supplies one of the bivariate distributions `F`, `C0`, `C180`, `C180`, `C270`, `J0`, `J180`, `J180`, `J270`, `G0`, `G180`, `G180` or `G270` to `SemiParBIVProbit` as the `BivD` argument, in place of the default Gaussian (`N`) copula. For example, the call to fit a rotated 90° Clayton copula selection model is:

```
f.list <- list(sel ~ x1 + s(x2, bs = "tp") + x3,
               HIV ~ x1 + s(x2),
                   ~ s(x4, bs = "mrf"))
SemiParBIVProbit(f.list, data, Model = "BSS", BivD = "C90",
                 weights = NULL)
```

where `f.list` specifies, in the following order, the equation for consent to HIV testing (selection) and for HIV status (outcome), and the third equation allows the user to model the copula association parameter as a function of covariates. The `s` terms represent smooth functions of the continuous predictor `x2` and factor variable `x4`. `Model = "BSS"` denotes a bivariate model with non-random sample selection. Argument `bs` specifies the type of spline basis; possible choices are `cr` (cubic regression spline), `cs` (shrinkage version of `cr`), `tp` (thin plate regression spline, the default), `ts` (shrinkage version of tp), `re` (random effect smoother, used in this paper for the interviewer variable), and `mrf` (Markov random field smoother, used for the regional variable). Argument `weights` allows the user to employ a vector of prior weights in fitting. Model `summary()` and `plot()` functions work in a similar fashion as those of generalized linear and additive models. The prevalence, with corresponding interval, can be obtained using the `prev()` function. More details and options can be found in the documentation of `SemiParBIVProbit`.

Supplementary Material C: Some dependence scenarios

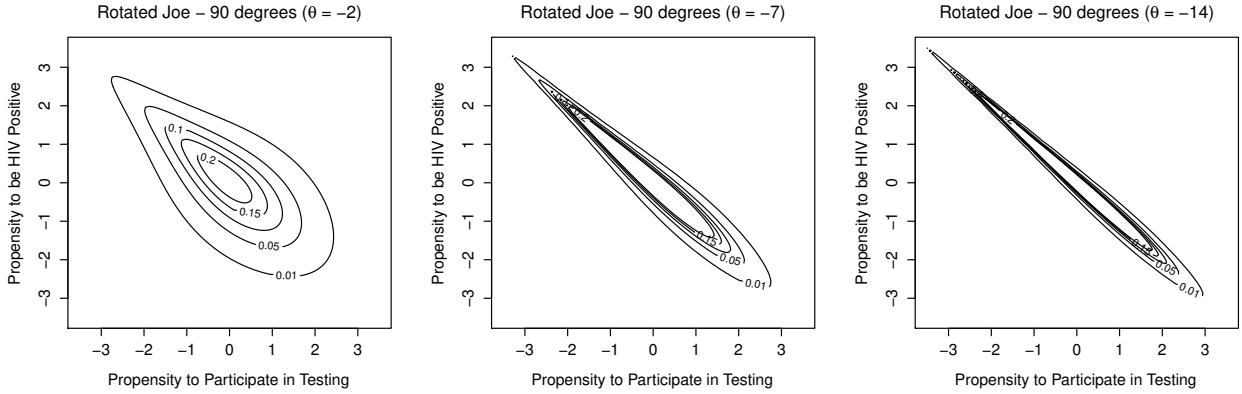


Figure 1: Three dependence scenarios for the counter-clockwise 90-degree rotated Joe copula: $\theta = -2$, minimal dependence, $\theta = -7$, moderate dependence, $\theta = -14$, high dependence. The range of θ is $(-\infty, -1)$. If this parameter is close to -1 then there is lack of noticeable association between participation in testing and HIV status once observed characteristics have been adjusted for. Note that dependence structure implied by the Joe copula rotated by 90 degrees is consistent with the interpretation that those who are most likely to be HIV positive are those who are also most likely to decline to participate in testing.

Supplementary Material D: Simulation settings

This section provide details of the simulation study used for evaluating the performance of the classic and proposed selection models. We constructed responses for consent to HIV testing and HIV status using several unobserved confounding variable distributions and link functions. For each scenario, we considered the situation in which the exclusion restriction assumption holds (i.e., interviewer identity predicts participation in HIV testing but not HIV status), and the cases where the assumption is mildly and strongly violated. Interest was in prevalence estimates.

We simulated an HIV survey with missing data in which the assumption of missing at random does not hold. We followed the approach implemented in Clark & Houle (2014) by generating a dataset based on a real HIV survey which in this case was the 2007 Zambia Demographic and Health Survey (DHS) for men. Therefore, our simulations closely match the overall observed consent rates in the actual data and HIV prevalence estimated by fitting a selection model on the real data used in the empirical part of this paper (the HIV prevalence was around 22% and consent rate around 80%). For each individual in the simulated dataset, we constructed variables for consent and HIV status based on two observed covariates (age and urban or rural place of residence) and an unobserved confounder. We used place of residence as our second covariate rather than sex as all our empirical models are stratified by sex and thus could not be included as a regressor. The

Age Category	N	%	Urban/Rural Place of Residence	N	%
15-19	1257	21	Urban	2540	42
20-24	1008	17	Rural	3460	58
25-29	921	16	Total	6000	
30-34	862	14			
35-39	745	12			
40-44	423	7			
45-49	350	6			
50-54	244	4			
55-59	190	3			
Total	6000	100			

Table 1: Summary statistics of simulated covariate information. N represents the number of observations (within each category and total).

dependence of consent to HIV testing and HIV status on a common unobserved predictor induced an association between the two variables and hence created a problem of systematic selection. The distributions of the two observed covariates were drawn to match those in the data (see Table 1 for a description of these characteristics), whereas the distribution of the unobserved confounder was allowed to be a standard normal, a uniform over the interval $[0, 1]$, or a standard log normal.

Consent to HIV testing and HIV status were based on linear predictors η_{1i} and η_{2i} which were determined by age and place of residence, contained in $\mathbf{x}_i$, an unobserved confounder u_i and interviewer identity:

$$\eta_{1i} = \beta_{10} + \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}_{11} + \gamma_1 u_i + \beta_{interviewerID_{1i}},$$
$$\eta_{2i} = \beta_{20} + \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}_{21} + \gamma_2 u_i + \delta\beta_{interviewerID_{2i}}.$$

Individuals were matched to one of 30 interviewers, whose persuasiveness ($\beta_{interviewerID_{1i}}$ and $\beta_{interviewerID_{2i}}$) were drawn from two uniform distributions over the interval $[-0.3, 0.4]$. We considered the case in which interviewer persuasiveness was always included in the consent equation but excluded from the HIV equation ($\delta = 0$), and the situations in which it was included in the HIV status equation with mild ($\delta = 0.5$) and strong effects ($\delta = 1$). The parameter vectors $\boldsymbol{\beta}_{11}$ and $\boldsymbol{\beta}_{21}$ were chosen by fitting a bivariate sample selection model on the 2007 Zambia DHS for men and are summarized in Table 2. All remaining parameters (β_{10}, γ_1, β_{20}, γ_2) were selected so that the consent rate and HIV prevalence were around 80% and 22%, respectively. Probabilities of consent to HIV testing and HIV status were obtained by transforming η_{1i} and η_{2i} using the cumulative distribution functions of the Gaussian, logistic and Weibull. Finally, binary outcomes were

	Consent equation $\boldsymbol{\beta}_{11}$	HIV status equation $\boldsymbol{\beta}_{21}$
Age 15-19	(Omitted Category)	(Omitted Category)
Age 20-24	-0.039	0.229
Age 25-29	-0.036	0.703
Age 30-34	0.017	1.036
Age 35-39	0.081	1.147
Age 40-44	0.134	1.203
Age 45-49	0.053	1.063
Age 50-54	0.028	0.834
Age 55-59	0.166	0.661
Rural	0.123	-0.396

Table 2: Regression parameters for age and place of residence in the consent to HIV testing and HIV status equations.

generated by using a random generator of the Bernoulli distribution with probabilities determined as just explained.

In the simulated data we observe HIV status for all individuals. In practice, we only observe the HIV status of those who consent to test. Therefore, when comparing the performance of the models, we censored the HIV outcome for individuals who did not consent to HIV testing. This allowed us to compare the true HIV prevalence (which we know) to that which would actually be observed in practice when there is missing data for HIV status, because of refusal to test (or other mechanisms for missing data). We compared the results obtained from the single imputation and selection models to the known true value. By varying the distribution of the unobserved covariate, link function and strength of interviewer persuasiveness in the HIV status equation, we evaluated the extent to which the standard selection model is sensitive to the assumption of normality and valid exclusion restriction, and whether the copula model could improve on the performance of the standard approach.

We considered nine different scenarios resulting from choosing several unobserved covariate distributions (normal, uniform and log-normal) and link functions (derived from the Gaussian, logistic and Weibull cumulative distribution functions). For each of the nine scenarios, we considered the case in which the assumption of exclusion restriction holds ($\delta = 0$) and the situations in which the assumption was mildly and strongly violated (δ equal to 0.5 and 1, respectively). A total of 27 scenarios were explored; these are summarized in Table 3. Each scenario was replicated 250 times.

cdf	Gaussian			Logistic			Weibull		
Unobservable	G	U	L	G	U	L	G	U	L
$\delta = 0$	S0PG	S0PU	S0PL	S0LG	S0LU	S0LL	S0WG	S0WU	S0WL
$\delta = 0.5$	S5PG	S5PU	S5PL	S5LG	S5LU	S5LL	S5WG	S5WU	S5WL
$\delta = 1$	S1PG	S1PU	S1PL	S1LG	S1LU	S1LL	S1WG	S1WU	S1WL

Table 3: Summary of the 27 scenarios explored in the simulation study. G, U and L stand for Gaussian, uniform and Log-normal unobserved confounder distributions. cdf stands for cumulative distribution function.

For each of the 27 scenarios we estimated the HIV prevalence, and compared the percent bias and root mean squared error (RMSE) for each of the following models: Gaussian, which is equivalent to the standard bivariate normal probit model; 90 and 270 degrees rotated Clayton; 90 and 270 degrees rotated Joe; 90 and 270 degrees rotated Gumbel; imputation-based estimate from univariate regression. Using more complex imputation approaches did not lead to significantly different results as compared to those obtained from the single imputation model.

Supplementary Material E: Smooth estimates

Smoothed estimates obtained from our flexible spline approach for modeling the effects of age, years of education and wealth in Swaziland are shown in Figures 2 and 3. There is clear evidence of non-linearity for most of these variables in both consent to test for HIV and HIV status. Some of these relationships are consistent across sex, for example, the impact of education on participation in testing and on HIV status. Other associations differ by sex, for example, wealth exhibits a very different association with HIV status among men as compared to that among women. Among women, higher wealth is linearly associated with an increasing risk of being HIV positive, while there seems not be a statistically significant association between household wealth and HIV status among men. We can use these results to identify peak prevalence (which has been adjusted for selective non-participation) according to the predictor of interest, for instance, age. Highest HIV prevalence occurs at age 25 in women in Swaziland, compared to age 35 among men in Swaziland. The functional form for these relationships also differs across models, which supports our data-driven approach to model specification and the avoidance of imposing a common specification across models. Smooth function estimates for Zambia and Zimbabwe are available upon request.

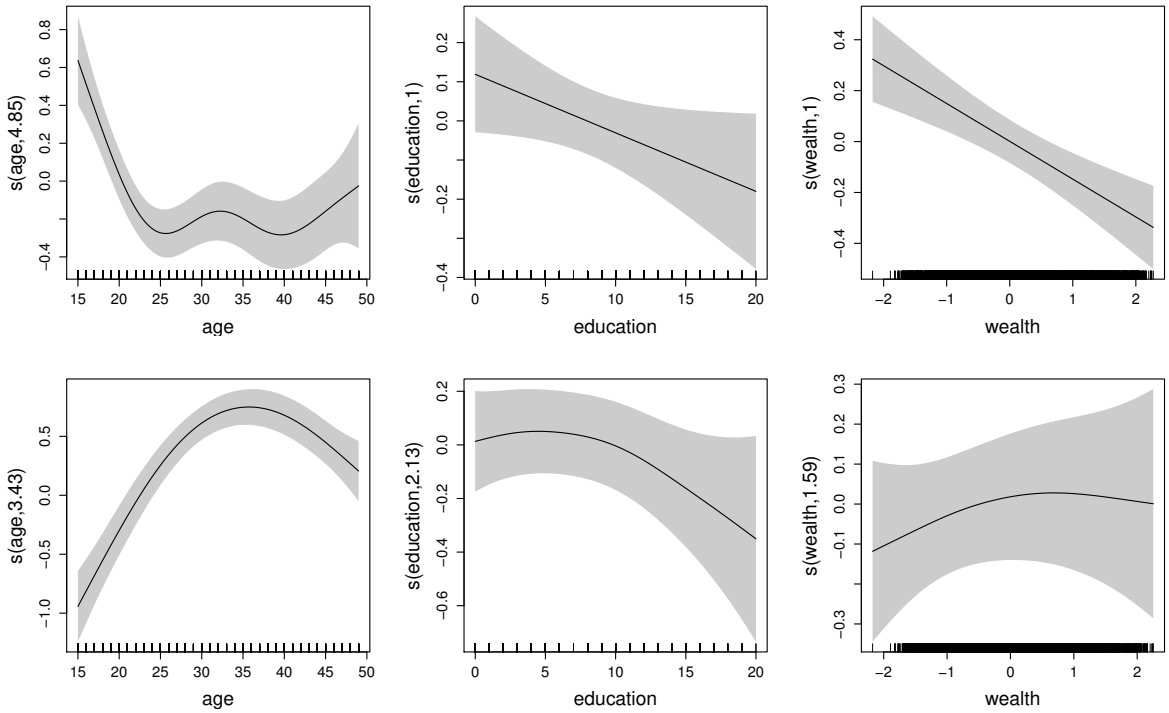


Figure 2: Swaziland (men). Smooth function estimates and associated 95% point-wise confidence intervals in the selection (first row) and outcome (second row) equations obtained from the proposed sample selection model based on the Joe copula rotated by 90 degrees. Results are plotted on the scale of respective linear predictors. The jittered rug plot, at the bottom of each graph, shows the covariate values. The numbers in brackets in the y-axis captions are the effective degrees of freedom of the smooth curves; the higher the value, the more complex the estimated curve.

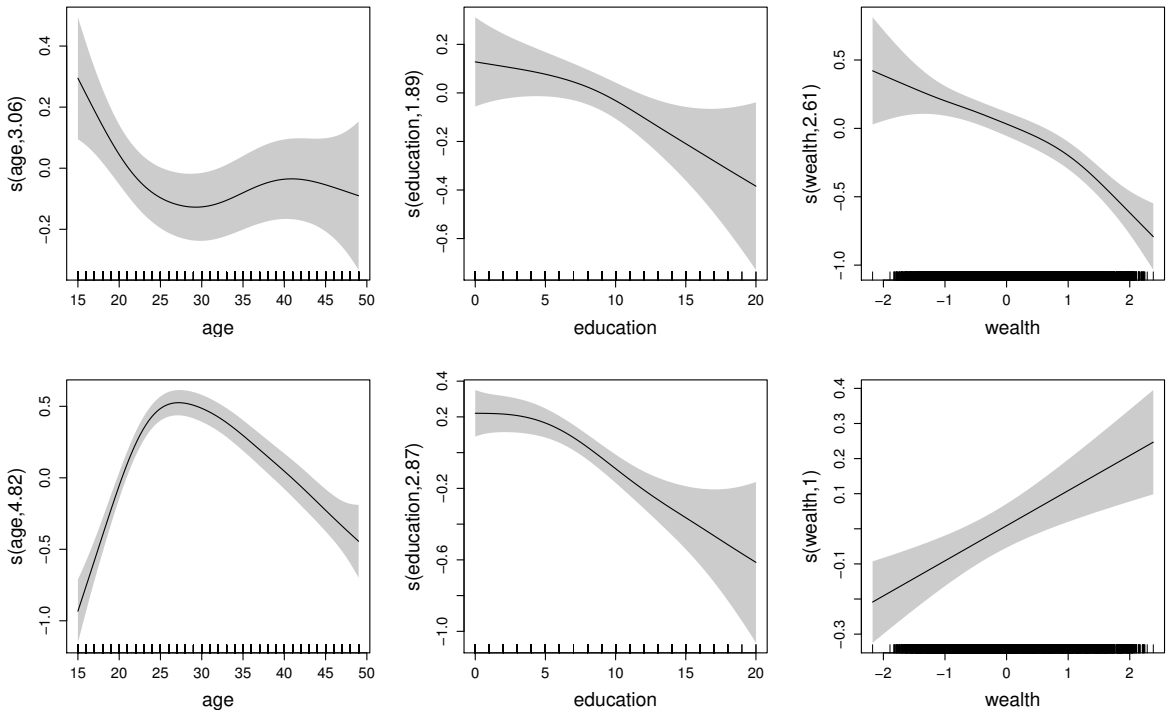


Figure 3: Swaziland (women). Smooth function estimates and associated 95% point-wise confidence intervals in the selection (first row) and outcome (second row) equations obtained from the proposed sample selection model based on the Joe copula rotated by 90 degrees. Results are plotted on the scale of respective linear predictors. The jittered rug plot, at the bottom of each graph, shows the covariate values. The numbers in brackets in the y-axis captions are the effective degrees of freedom of the smooth curves; the higher the value, the more complex the estimated curve.

Supplementary Material F: Table of letters and symbols used in the paper

Letter/Symbol	Definition
Y_{1i}	Binary selection random variable for i^{th} individual
Y_{2i}	Binary outcome random variable for i^{th} individual
y_{1i}	Observed value for Y_{1i}
y_{2i}	Observed value for Y_{2i}
n	Sample size
$\mathbf{z}_i$	Generic vector of covariates for i^{th} individual
$\mathbb{P}$	Probability function
$\mathcal{C}$	Two-place copula function
η_i	Generic linear predictor for i^{th} individual containing parameters and covariates
Φ	Cumulative distribution function of standard univariate Gaussian distribution
θ_i	Copula parameter measuring the dependence between Y_{1i} and Y_{2i}
m	One-to-one transformation mapping the linear predictor to the copula parameter
ℓ	Log-likelihood function
$\mathbb{R}$	Set of real numbers
β_0	Overall intercept of generic linear predictor
$\mathbf{z}_{ki}$	k^{th} sub-vector of $\mathbf{z}_i$
$s_k(\mathbf{z}_{ki})$	Smooth function of $\mathbf{z}_{ki}$
K	Generic number of smooth functions
J_k	Generic number of basis functions
$Z_k[i, j_k], b_{kj_k}(\mathbf{z}_{ki})$	j_k^{th} basis function of $\mathbf{z}_{ki}$
β_{kj_k}	Regression coefficient associated with $b_{kj_k}(\mathbf{z}_{ki})$
$\boldsymbol{\eta}$	Vector containing the η_i values for all individuals
$\mathbf{Z}_k$	k^{th} design matrix containing the J_k basis functions for k^{th} covariate
$\boldsymbol{\beta}_k$	Coefficient vector associated with $\mathbf{Z}_k$
$\mathbf{1}_n$	n-dimensional vector made up of ones
$\mathbf{Z}$	Overall design matrix made up of $\mathbf{1}_n$ and $\mathbf{Z}_k$ for $k = 1, \ldots, K$
$\boldsymbol{\beta}$	Overall coefficient vector associated with $\mathbf{Z}$
λ_k	k^{th} smoothing parameter controlling the trade-off between fit and smoothness
$\mathbf{D}_k$	k^{th} smoothing penalty whose structure depends on the type smooth employed
$\mathbf{D}_{\boldsymbol{\lambda}}$	Overall smoothing penalty for one equation made up of 0 and $\lambda_k\mathbf{D}_k$ for $k = 1, \ldots, K$
$\mathbf{I}$	Identity matrix
r	Region r
q	Region q
$\wedge$	AND operator
N_r	Total number of neighbors for region r
R	Number of regions
$\mathbf{d}_k(z_k)$	Vector with j_k^{th} element given by $\partial^2 b_{kj_k}(z_k)/\partial z_k^2$

Table 4: Definition of letters and symbols used in the paper and their corresponding meanings.

Letter/Symbol	Definition
$\beta_{10}, \beta_{20}, \beta_{30}$	Overall levels of the linear predictors η_{1i}, η_{2i}, η_{3i}
$\mathbf{x}_i$	Vector of discrete and binary variables associated with the selection and the outcome
$\boldsymbol{\beta}_{11}, \boldsymbol{\beta}_{21}$	Vectors of parameters for the selection and outcome equations associated with $\mathbf{x}_i$
s_{vk}	Smooth functions of continuous covariates for $v = 1, 2$ and $k = 1, 2, 3$
$s_{vspatial}$	Spatial regional effects for $v = 1, 2, 3$
$\beta_{interviewerID_i}$	Random effects for the set of binary variables defined by interviewer identity
$\hat{}$	Estimate or estimator of argument
$\intercal$	Transpose
$\boldsymbol{\delta}$	Overall parameter vector of the model
$\mathbf{S}_{\boldsymbol{\lambda}}$	Overall penalty matrix of the model's parameters
$\boldsymbol{\lambda}$	Overall smoothing parameter vector
ℓ_p	Penalized log-likelihood function
a	Iteration index
$\mathbf{g}$	Gradient vector
$\mathbf{g}_p$	Penalized gradient vector
$\boldsymbol{\mathcal{H}}$	Hessian matrix
$\boldsymbol{\mathcal{H}}_p$	Penalized Hessian matrix
$\breve{\ell}_p$	Quadratic approximation of ℓ_p
$\mathbf{p}$	Step update
$\|\cdot\|$	Euclidean norm
ra	Radius of the trust region
$\mathbb{E}$	Expected value
$\boldsymbol{\mathcal{I}}$	$-\boldsymbol{\mathcal{H}}$
$\boldsymbol{\epsilon}$	$\sqrt{\boldsymbol{\mathcal{I}}}^{-1}\mathbf{g}$
$\mathbf{z}$	$\sqrt{\boldsymbol{\mathcal{I}}}\boldsymbol{\delta} + \boldsymbol{\epsilon}$
$\sim$	Distribution
$\boldsymbol{\delta}^0$	True overall parameter vector
$\mathcal{N}$	Multivariate Gaussian distribution
$\boldsymbol{\mu}_{\mathbf{z}}$	$\sqrt{\boldsymbol{\mathcal{I}}}\boldsymbol{\delta}^0$
$\mathbf{A}_{\hat{\boldsymbol{\lambda}}}$	Hat matrix $\sqrt{\boldsymbol{\mathcal{I}}}\left(\boldsymbol{\mathcal{I}} + \mathbf{S}_{\hat{\boldsymbol{\lambda}}}\right)^{-1}\sqrt{\boldsymbol{\mathcal{I}}}$
$\mathcal{V}(\boldsymbol{\lambda})$	Smoothing criterion
$\arg\min$	Argument of the minimum
$\check{n}$	$3n$
tr	Trace operator
w_i	Survey weight for i^{th} individual
$\bar{\mathbf{x}}_i$	Mean characteristics of each interviewer's interviewees

Table 5: Definition of letters and symbols used in the paper and their corresponding meanings.

References

Clark, S. J. & Houle, B. (2014). Validation, replication, and sensitivity testing of heckman-type selection models to adjust estimates of HIV prevalence. *PloS one*, 9, e112563.

Efron, B. & Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected fisher information. *Biometrika*, 65(3), 457–483.

Gillespie, S., Greener, R., Whiteside, A., & Whitworth, J. (2007). Investigating the empirical evidence for understanding vulnerability and the associations between poverty, HIV infection and AIDS impact. *AIDS*, 21(Supp), S1–S4.

Gouws, E., Stanecki, K. A., Lyerla, R., & Ghys, P. D. (2008). The epidemiology of HIV infection among young people aged 15–24 years in southern africa. *AIDS*, 22(Supp), S5–S16.

Hargreaves, J. R., Bonell, C. P., Boler, T., Boccia, D., Birdthistle, I., Fletcher, A., Pronyk, P. M., & Glynn, J. R. (2008). Systematic review exploring time trends in the association between educational attainment and risk of HIV infection in sub-Saharan Africa. *AIDS*, 22(3), 403–414.

Madden, D. (2008). Sample selection versus two-part models revisited: the case of female smoking and drinking. *Journal of Health Economics*, 27(2), 300–307.

Marra, G. & Radice, R. (2013). A penalized likelihood estimation approach to semiparametric sample selection binary response modeling. *Electronic Journal of Statistics*, 7, 1432–1455.

Marra, G. & Radice, R. (2016). *SemiParBIVProbit: Semiparametric Bivariate Probit Modelling*. R package version 3.6.1.

R Development Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Radice, R., Marra, G., & Wojtys, M. (2015). Copula regression spline models for binary outcomes. *Statistics and Computing*, Forthcoming.

Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press, New York.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, 99(467), 673–686.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction With R*. Chapman & Hall/CRC, London.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B*, 73(1), 3–36.

Wood, S. N. (2015). *Core Statistics*. ims Textbooks, Cambridge.