https://doi.org/10.1080/01621459.2016.1224713
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. (2016) and Marra & Radice (2013) discussed a smoothing approach for bivariate equation models with penalized regression splines which is based on \( z = \sqrt{W} (W^{-1} d + Z\delta) \). Loosely speaking, \( W \) is of dimensions \( \tilde{n} \times \tilde{n} \), where \( \tilde{n} = 3n \), and represents a block diagonal weight matrix containing minus the second derivatives of the log-likelihood with respect to \( \eta_1, \eta_2 \) and \( \eta_3 \), \( d \) is a vector of length \( \tilde{n} \) containing the first derivatives of the log-likelihood with respect to \( \eta_1, \eta_2 \) and \( \eta_3 \), and \( Z \) is an overall design matrix of dimensions \( \tilde{n} \times m \), where \( m \) is the total number of columns, which has a block diagonal structure and contains the design matrices associated with \( \eta_1, \eta_2 \) and \( \eta_3 \). Pseudodata vector \( z \) requires \( W \) 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 \( W = \text{diag}(W_1, \ldots, W_n) \) need not all be positive definite, and in practice a non-negligible number of non-positive definite \( W_i \) 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 \( E(W) \) is used in place of \( W \). 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 \( z \) that uses \( \mathcal{H} \) and \( g \) as a whole instead of the \( n \) components that make them up. There will clearly be situations in which \( \mathcal{H} \) 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 \( \mathcal{H} \) to positive definiteness (e.g., Wood, 2015, Chapter 5). The additional advantage of such an approach is that \( \mathcal{H} \) and \( g \) would be obtained as a byproduct of the estimation step for \( \delta \), 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 \( g_{[a+1]}^{[a]} \) about \( \delta^{[a]} \) yields \( 0 = g_{[a+1]}^{[a]} \approx g_{[a]}^{[a]} + (\delta_{[a+1]} - \delta^{[a]}) \mathcal{H}_{[a]}^{[a]} \), where \( g_{[a]}^{[a]} = g^{[a]} - S_\lambda \delta^{[a]} \) and \( \mathcal{H}_{[a]}^{[a]} = \mathcal{H}^{[a]} - S_\lambda \). As explained above, finding an expression for \( \delta_{[a+1]} \) that is based on \( g^{[a]} \) and \( \mathcal{H}^{[a]} \) is crucial to our developments and it can be obtained as follows. Let us define \( \mathcal{I}^{[a]} = -\mathcal{H}^{[a]} \),
we then have

\[
0 = g_p^{[a]} + (\delta^{[a+1]} - \delta^{[a]}) \left( -I^{[a]} - S_\lambda \right),
\]

\[
g_p^{[a]} = (\delta^{[a+1]} - \delta^{[a]}) \left( I^{[a]} + S_\lambda \right),
\]

\[
g^{[a]} - S_\lambda \delta^{[a]} = \delta^{[a+1]} \left( I^{[a]} + S_\lambda \right) - \delta^{[a]} I^{[a]} - \delta^{[a]} S_\lambda,
\]

\[
\delta^{[a+1]} \left( I^{[a]} + S_\lambda \right) = g^{[a]} + \delta^{[a]} I^{[a]},
\]

\[
\delta^{[a+1]} = \left( I^{[a]} + S_\lambda \right)^{-1} \sqrt{I^{[a]}} \left( \sqrt{I^{[a]}} \delta^{[a]} + \sqrt{I^{[a]}}^{-1} g^{[a]} \right).
\]

Therefore, the parameter estimator can be expressed as

\[
\delta^{[a+1]} = \left( I^{[a]} + S_\lambda \right)^{-1} \sqrt{I^{[a]}} z^{[a]},
\]

where \( z^{[a]} = \mu^{[a]} + \epsilon^{[a]} \) with \( \mu^{[a]} = \sqrt{I^{[a]}} \delta^{[a]} \) and \( \epsilon^{[a]} = \sqrt{I^{[a]}}^{-1} g^{[a]} \). The square root of \( I \) and its inverse are obtained by eigen-value decomposition. Note that, to within an additive constant, pseudodata vector \( z \) 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 \( \delta \). From likelihood theory, \( \epsilon \sim \mathcal{N}(0, I) \) and \( z \sim \mathcal{N}(\mu_z, I) \), where \( I \) is an identity matrix, \( \mu_z = \sqrt{I} \delta^0 \) and \( \delta^0 \) is the true parameter vector. The predicted value vector for \( z \) is \( \hat{\mu}_z = \sqrt{I} \hat{\delta} A_\lambda z \), where \( A_\lambda = \sqrt{I} (I + S_\lambda)^{-1} I \). Since our goal is to estimate \( \lambda \) so that the smooth terms’ complexity which is not supported by the data is suppressed, the smoothing parameter vector is estimated so that \( \hat{\mu}_z \) is as close as possible to \( \mu_z \). Therefore, we use

\[
\mathbb{E} \left( \| \mu_z - \hat{\mu}_z \|^2 \right) = \mathbb{E} \left( \| (z - \epsilon) - A_\lambda z \|^2 \right) = \mathbb{E} \left( \| z - A_\lambda z - \epsilon \|^2 \right)
\]

\[
= \mathbb{E} \left( \| z - A_\lambda z \|^2 \right) + \mathbb{E} \left( -\epsilon^T \epsilon - 2 \epsilon^T \hat{\mu}_z + 2 \epsilon^T A_\lambda \mu_z + 2 \epsilon^T A_\lambda \epsilon \right)
\]

\[
= \mathbb{E} \left( \| z - A_\lambda z \|^2 \right) - \hat{n} + 2 \text{tr}(A_\lambda),
\]

where \( \hat{n} = 3n \) and \( \text{tr}(A_\lambda) \) 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 \( \epsilon \) and that a scalar is its own trace. In practice, \( \lambda \) is estimated by minimizing an
estimate of (1), i.e.

\[ V(\lambda) = \|\mu - \hat{\mu}\|^2 = \|z - A\lambda z\|^2 - \hat{n} + 2\text{tr}(A\lambda). \]  

(2)

Given \( \delta^{[a+1]} \), the problem becomes

\[ \lambda^{[a+1]} = \arg \min_{\lambda} V(\lambda) = \|z^{[a+1]} - A_{\lambda^{[a]}} z^{[a+1]}\|^2 - \hat{n} + 2\text{tr}(A_{\lambda^{[a]}}), \]

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 \( V(\lambda) \) and their first and second derivatives with respect to \( \log(\lambda) \) (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 \(-2\ell(\hat{\delta})\). Therefore, dropping irrelevant constants yields \( V(\lambda) \propto -2\ell(\hat{\delta}) + 2\text{tr}(A\lambda) \). 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 \( z \) and \( A \) 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 \( F, C_0, C_{180}, C_{270}, J_0, J_{180}, J_{270}, G_0, G_{180}, G_{270} \) or \( G_{270} \) 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:

```r
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 \( x_2 \) and factor variable \( x_4 \). \( \text{Model} = "BSS" \) 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`. 

5
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 \( \theta \) 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
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 \(\eta_{1i}\) and \(\eta_{2i}\) which were determined by age and place of residence, contained in \(x_i\), an unobserved confounder \(u_i\) and interviewer identity:

\[
\eta_{1i} = \beta_{10} + x_i^T \beta_{11} + \gamma_1 u_i + \beta_{\text{interviewerID}_{1i}},
\]

\[
\eta_{2i} = \beta_{20} + x_i^T \beta_{21} + \gamma_2 u_i + \delta \beta_{\text{interviewerID}_{2i}}.
\]

Individuals were matched to one of 30 interviewers, whose persuasiveness (\(\beta_{\text{interviewerID}_{1i}}\) and \(\beta_{\text{interviewerID}_{2i}}\)) 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 \(\beta_{11}\) and \(\beta_{21}\) 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 (\(\beta_{10}, \gamma_1, \beta_{20}, \gamma_2\)) 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 \(\eta_{1i}\) and \(\eta_{2i}\) using the cumulative distribution functions of the Gaussian, logistic and Weibull. Finally, binary outcomes were
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 ($\delta$ 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.

<table>
<thead>
<tr>
<th>Age Group</th>
<th>Consent equation $\beta_{11}$</th>
<th>HIV status equation $\beta_{21}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age 15-19</td>
<td>(Omitted Category)</td>
<td>(Omitted Category)</td>
</tr>
<tr>
<td>Age 20-24</td>
<td>-0.039</td>
<td>0.229</td>
</tr>
<tr>
<td>Age 25-29</td>
<td>-0.036</td>
<td>0.703</td>
</tr>
<tr>
<td>Age 30-34</td>
<td>0.017</td>
<td>1.036</td>
</tr>
<tr>
<td>Age 35-39</td>
<td>0.081</td>
<td>1.147</td>
</tr>
<tr>
<td>Age 40-44</td>
<td>0.134</td>
<td>1.203</td>
</tr>
<tr>
<td>Age 45-49</td>
<td>0.053</td>
<td>1.063</td>
</tr>
<tr>
<td>Age 50-54</td>
<td>0.028</td>
<td>0.834</td>
</tr>
<tr>
<td>Age 55-59</td>
<td>0.166</td>
<td>0.661</td>
</tr>
<tr>
<td>Rural</td>
<td>0.123</td>
<td>-0.396</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Letter/Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y_{1i}$</td>
<td>Binary selection random variable for $i^{th}$ individual</td>
</tr>
<tr>
<td>$Y_{2i}$</td>
<td>Binary outcome random variable for $i^{th}$ individual</td>
</tr>
<tr>
<td>$y_{1i}$</td>
<td>Observed value for $Y_{1i}$</td>
</tr>
<tr>
<td>$y_{2i}$</td>
<td>Observed value for $Y_{2i}$</td>
</tr>
<tr>
<td>$n$</td>
<td>Sample size</td>
</tr>
<tr>
<td>$z_i$</td>
<td>Generic vector of covariates for $i^{th}$ individual</td>
</tr>
<tr>
<td>$\mathbb{P}$</td>
<td>Probability function</td>
</tr>
<tr>
<td>$C$</td>
<td>Two-place copula function</td>
</tr>
<tr>
<td>$\eta_i$</td>
<td>Generic linear predictor for $i^{th}$ individual containing parameters and covariates</td>
</tr>
<tr>
<td>$\Phi$</td>
<td>Cumulative distribution function of standard univariate Gaussian distribution</td>
</tr>
<tr>
<td>$\theta_i$</td>
<td>Copula parameter measuring the dependence between $Y_{1i}$ and $Y_{2i}$</td>
</tr>
<tr>
<td>$m$</td>
<td>One-to-one transformation mapping the linear predictor to the copula parameter</td>
</tr>
<tr>
<td>$\ell$</td>
<td>Log-likelihood function</td>
</tr>
<tr>
<td>$\mathbb{R}$</td>
<td>Set of real numbers</td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>Overall intercept of generic linear predictor</td>
</tr>
<tr>
<td>$z_{ki}$</td>
<td>$k^{th}$ sub-vector of $z_i$</td>
</tr>
<tr>
<td>$\eta_i(z_{ki})$</td>
<td>Smooth function of $z_{ki}$</td>
</tr>
<tr>
<td>$K$</td>
<td>Generic number of smooth functions</td>
</tr>
<tr>
<td>$J_k$</td>
<td>Generic number of basis functions</td>
</tr>
<tr>
<td>$Z_k[i,j_k], b_{kj_k}(z_{ki})$</td>
<td>$j_k^{th}$ basis function of $z_{ki}$</td>
</tr>
<tr>
<td>$\beta_{kj_k}$</td>
<td>Regression coefficient associated with $b_{kj_k}(z_{ki})$</td>
</tr>
<tr>
<td>$\eta$</td>
<td>Vector containing the $\eta_i$ values for all individuals</td>
</tr>
<tr>
<td>$Z_k$</td>
<td>$k^{th}$ design matrix containing the $J_k$ basis functions for $k^{th}$ covariate</td>
</tr>
<tr>
<td>$\beta_k$</td>
<td>Coefficient vector associated with $Z_k$</td>
</tr>
<tr>
<td>$1_n$</td>
<td>$n$-dimensional vector made up of ones</td>
</tr>
<tr>
<td>$Z$</td>
<td>Overall design matrix made up of $1_n$ and $Z_k$ for $k = 1, \ldots, K$</td>
</tr>
<tr>
<td>$\beta$</td>
<td>Overall coefficient vector associated with $Z$</td>
</tr>
<tr>
<td>$\lambda_k$</td>
<td>$k^{th}$ smoothing parameter controlling the trade-off between fit and smoothness</td>
</tr>
<tr>
<td>$D_k$</td>
<td>$k^{th}$ smoothing penalty whose structure depends on the type smooth employed</td>
</tr>
<tr>
<td>$D_\lambda$</td>
<td>Overall smoothing penalty for one equation made up of 0 and $\lambda_kD_k$ for $k = 1, \ldots, K$</td>
</tr>
<tr>
<td>$I$</td>
<td>Identity matrix</td>
</tr>
<tr>
<td>$r$</td>
<td>Region $r$</td>
</tr>
<tr>
<td>$q$</td>
<td>Region $q$</td>
</tr>
<tr>
<td>$\wedge$</td>
<td>AND operator</td>
</tr>
<tr>
<td>$N_r$</td>
<td>Total number of neighbors for region $r$</td>
</tr>
<tr>
<td>$R$</td>
<td>Number of regions</td>
</tr>
<tr>
<td>$d_k(z_k)$</td>
<td>Vector with $j_k^{th}$ element given by $\partial^2b_{kj_k}(z_k)/\partial z_k^2$</td>
</tr>
</tbody>
</table>

Table 4: Definition of letters and symbols used in the paper and their corresponding meanings.
<table>
<thead>
<tr>
<th>Letter/Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_{10}, \beta_{20}, \beta_{30}$</td>
<td>Overall levels of the linear predictors $\eta_{1i}, \eta_{2i}, \eta_{3i}$</td>
</tr>
<tr>
<td>$x_i$</td>
<td>Vector of discrete and binary variables associated with the selection and the outcome</td>
</tr>
<tr>
<td>$\beta_{11}, \beta_{21}$</td>
<td>Vectors of parameters for the selection and outcome equations associated with $x_i$</td>
</tr>
<tr>
<td>$s_{vkr}$</td>
<td>Smooth functions of continuous covariates for $v = 1, 2$ and $k = 1, 2, 3$</td>
</tr>
<tr>
<td>$s_{\text{spatial}}$</td>
<td>Spatial regional effects for $v = 1, 2, 3$</td>
</tr>
<tr>
<td>$\beta_{\text{interviewerID}_i}$</td>
<td>Random effects for the set of binary variables defined by interviewer identity</td>
</tr>
<tr>
<td>$\hat{}$</td>
<td>Estimate or estimator of argument</td>
</tr>
<tr>
<td>$\tau$</td>
<td>Transpose</td>
</tr>
<tr>
<td>$\delta$</td>
<td>Overall parameter vector of the model</td>
</tr>
<tr>
<td>$S_\lambda$</td>
<td>Overall penalty matrix of the model’s parameters</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>Overall smoothing parameter vector</td>
</tr>
<tr>
<td>$\ell_p$</td>
<td>Penalized log-likelihood function</td>
</tr>
<tr>
<td>$a$</td>
<td>Iteration index</td>
</tr>
<tr>
<td>$g$</td>
<td>Gradient vector</td>
</tr>
<tr>
<td>$g_p$</td>
<td>Penalized gradient vector</td>
</tr>
<tr>
<td>$\mathcal{H}$</td>
<td>Hessian matrix</td>
</tr>
<tr>
<td>$\mathcal{H}_p$</td>
<td>Penalized Hessian matrix</td>
</tr>
<tr>
<td>$\hat{\ell}_p$</td>
<td>Quadratic approximation of $\ell_p$</td>
</tr>
<tr>
<td>$p$</td>
<td>Step update</td>
</tr>
<tr>
<td>$| \cdot |$</td>
<td>Euclidean norm</td>
</tr>
<tr>
<td>$r_a$</td>
<td>Radius of the trust region</td>
</tr>
<tr>
<td>$E$</td>
<td>Expected value</td>
</tr>
<tr>
<td>$\mathcal{I}$</td>
<td>$-\mathcal{H}$</td>
</tr>
<tr>
<td>$\epsilon$</td>
<td>$\sqrt{\mathcal{I}}^{-1}g$</td>
</tr>
<tr>
<td>$z$</td>
<td>$\sqrt{\mathcal{I}}\delta + \epsilon$</td>
</tr>
<tr>
<td>$\sim$</td>
<td>Distribution</td>
</tr>
<tr>
<td>$\delta^0$</td>
<td>True overall parameter vector</td>
</tr>
<tr>
<td>$\mathcal{N}$</td>
<td>Multivariate Gaussian distribution</td>
</tr>
<tr>
<td>$\mu_z$</td>
<td>$\sqrt{\mathcal{I}}\delta^0$</td>
</tr>
<tr>
<td>$A_\lambda$</td>
<td>Hat matrix $\sqrt{\mathcal{I}} (\mathcal{I} + S_\lambda)^{-1} \sqrt{\mathcal{I}}$</td>
</tr>
<tr>
<td>$\mathcal{V}(\lambda)$</td>
<td>Smoothing criterion</td>
</tr>
<tr>
<td>$\arg\min$</td>
<td>Argument of the minimum</td>
</tr>
<tr>
<td>$\hat{n}$</td>
<td>$3n$</td>
</tr>
<tr>
<td>$\text{tr}$</td>
<td>Trace operator</td>
</tr>
<tr>
<td>$w_i$</td>
<td>Survey weight for $i^{th}$ individual</td>
</tr>
<tr>
<td>$\bar{x}_i$</td>
<td>Mean characteristics of each interviewer’s interviewees</td>
</tr>
</tbody>
</table>

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


