Journal of Data Science 19 (1), 75–95
January 2021
DOI: 10.6339/21-JDS1008
Statistical Data Science

# Spline Pattern-Mixture Models for Missing Data

Ye Yang[1,*] and Roderick J.A. Little[2]

[1]*Center for Biologics Evaluation and Research, FDA, Silver Spring, MD, USA*
[2]*Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA*

**Abstract**

We consider a continuous outcome subject to nonresponse and a fully observed covariate. We propose a spline proxy pattern-mixture model (S-PPMA), an extension of the proxy pattern-mixture model (PPMA) (Andridge and Little, 2011), to estimate the mean of the outcome under varying assumptions about nonresponse. S-PPMA improves the robustness of PPMA, which assumes bivariate normality between the outcome and the covariate, by modeling the relationship via a spline. Simulations indicate that S-PPMA outperforms PPMA when the data deviate from normality and are missing not at random, with minor losses of efficiency when the data are normal.

**Keywords** *missing data; missing not at random; nonignorable nonresponse; nonresponse bias*

## 1 Introduction

Missing data are a common problem in many data sets. In this article we consider data where our goal is to estimate the mean of a variable $Y$ with $n_0$ observed values ($\{Y_i\}, i = 1, \ldots, n_0$) and $n_1$ missing values ($\{Y_i\}, i = n_0 + 1, \ldots, n_0 + n_1$), when there is a set of $p$ auxiliary variables $Z_1, \ldots, Z_p$ that are fully observed ($\{Z_{i1}, \ldots, Z_{ip}\}, i = 1, \ldots, n, n = n_0 + n_1$). Define the response indicator $R$ taking values 1 if $Y$ is observed and 0 if $Y$ is missing. It is common to use methods that assume $Y$ is missing at random (MAR) in the sense that $R$ is independent of $Y$ given the observed covariates $Z_1, \ldots, Z_p$ (Rubin, 1976). Such methods include weighting class adjustments and imputation. Our methods build on a robust MAR imputation method called penalized spline of propensity prediction (PSPP) (Little and An, 2004; Zhang and Little, 2009; Yang and Little, 2015). This method (a) estimates the propensity that $R = 1$ given $Z_1, \ldots, Z_p$ based on a logistic regression of $R$ on $Z_1, \ldots, Z_p$, using all the data, and (b) imputes $Y$ based on the regression of $Y$ on a penalized spline of the estimated propensity, with other covariates being included parametrically if they improve the predictions.

MAR-based methods are generally biased in cases where the missingness is missing not at random (MNAR), meaning that missingness of $Y$ depends not only on covariates $Z_1, \ldots, Z_p$ but also on the value of $Y$ itself. Schouten (2007) proposes a selection strategy for weighting variables that relaxes the MAR assumption. The method uses a generalized regression estimator to estimate the mean with auxiliary variables selected to minimize the maximal absolute bias under MNAR. The selection strategy, however, is based on parameters estimated under the MAR assumption and thus may be invalid if the missingness mechanism deviates markedly from MAR. Pfeffermann and Sikov (2011) propose a method for estimating the mean under MNAR by specifying models for the outcome and propensity, which is allowed to depend on both the outcome and auxiliary variables. The method assumes known population totals for

*Corresponding author. Email: yeya@umich.edu.

Received November, 2020; Accepted January, 2021

some or all of the auxiliary variables in the two models and estimates the model parameters in a way that takes into account the known population totals.

The bivariate normal pattern-mixture model (BNPM) (Little, 1993, 1994) assumes a bivariate normal distribution for a single observed covariate $X$ and an outcome $Y$ within strata defined by respondents and nonrespondents, with a different mean and covariance matrix in each stratum. Parameters of BNPM are identified by assumptions about the missingness mechanism. For instance, under MAR, where missingness is assumed to depend on $X$ but not $Y$, the parameters of the regression of $Y$ on $X$ are the same for respondents and nonrespondents; as a result, the maximum likelihood (ML) estimate for the mean of $Y$ is the regression estimate, $\hat{\mu}_Y = \bar{Y}^{(1)} + \frac{s_{XY}}{s_{XX}}(\bar{X} - \bar{X}^{(1)})$, where $\bar{X}$ is the sample mean of $X$, $\bar{X}^{(1)}$ is the respondent mean of $X$, $\bar{Y}^{(1)}$ is the respondent mean of $Y$, $s_{XY}$ is the respondent covariance of $X$ and $Y$, and $s_{XX}$ is the respondent variance of $X$. When missingness is MNAR and is assumed to depend on $Y$ but not $X$, the parameters of the regression of $X$ on $Y$ are the same for respondents and nonrespondents; Little (1994) shows that the resulting ML estimate of the mean of $Y$ is $\hat{\mu}_Y = \bar{Y}^{(1)} + \frac{s_{YY}}{s_{XY}}(\bar{X} - \bar{X}^{(1)})$, where $s_{YY}$ is the respondent variance of $Y$. The approach is easily extended to allow missingness of $Y$ to depend on $Y^* = X + \lambda Y$ for some known $\lambda$, a parameter that can then be varied in a sensitivity analysis. ML, Bayesian and multiple imputation (MI) approaches to inference for this BNPM model are described in Little (1994).

An advantage of the BNPM model is that it does not need to specify an explicit functional form for the missingness mechanism, the mechanism entering in the form of restrictions on the model parameters. The modification of MAR regression estimation to MNAR models is straightforward, as seen in the estimate of the mean of $Y$ above. However, validity of the estimates depends on bivariate normality of $X$ and $Y$, which is a strong assumption. For example, if $X$ is normal and $Y$ given $X$ is normal with conditional mean a quadratic function of $X$, then the regression of $X$ on $Y$ is no longer linear, and ML estimates under the BNPM model are biased. In this article we study the impact of such forms of misspecification on inferences for the mean of $Y$.

We also propose a modification of the BNPM model, spline-BPNM (S-BPNM), which replaces a parametric linear regression by a penalized spline, extending the PSPP method (which assumes MAR) to MNAR situations; in the case where missingness depends on $Y$, we model the regression of $X$ on $Y$ using a flexible penalized spline, rather than assuming a linear relationship. The resulting estimate of the mean of $Y$ is shown in simulations to be more robust than BNPM to the distributional relationship between $X$ and $Y$. The approach can also be generalized to the case where missingness depends on $Y^* = X + \lambda Y$ for some known value of $\lambda$.

We also consider cases with more than one covariate. In that context, proxy pattern-mixture model analysis (Andridge and Little, 2011) extends the BNPM model to data with an outcome $Y$ and a set of $p$ observed covariates $Z_1, \ldots, Z_p$. The PPMA method replaces the set of covariates by a proxy $X$, the single best predictor of $Y$ given the covariates, estimated by regressing $Y$ on $Z_1, \ldots, Z_p$ for the respondents. The method then fits the pattern-mixture model in Little (1994) to $Y$ and $X$. Bayesian forms of PPMA take into account the estimation of the coefficients of $Z$ in the proxy variable $X$. This analysis relies on the bivariate normality assumption between the proxy $X$ and $Y$, which is violated when some or all of the covariates $Z_1, \ldots, Z_p$ used to estimate $X$ are not normally distributed. We propose a more flexible version of PPMA, which we call spline-PPMA (S-PPMA), which relaxes the bivariate normality assumption between the proxy and $Y$ by replacing the linear regression of $X$ on $Y^*$ implied by the bivariate normality with a penalized spline, allowing for a non-linear relationship between the variables.

We conduct simulations to examine the performance of the new S-PPMA model, and in particular to address the following questions:

1. How do inferences under S-BNPM and S-PPMA models compare with the original BNPM and PPMA methods in terms of bias, root mean squared error (RMSE) and coverage, for data sets generated under a variety of distributional assumptions?
2. How sensitive are S-BPNM and S-PPMA models to alternative assumptions about the missingness mechanism?

In the next section, we present the S-BNPM and S-PPMA models in detail. We then assess their performance in simulation studies under a variety of distributional assumptions for the auxiliary variables and missingness mechanisms.

## 2 Pattern-Mixture Model Analysis

We consider first bivariate data on $X$ and $Y$, with $X$ observed for the entire sample and $Y$ subject to missing data, and let $R = 1$ if $Y$ is observed and $R = 0$ if $Y$ is missing. Little (1994) assumes the BNPM model:

$$(Y, X|\phi^{(r)}, R = r) \sim \mathrm{N}_2\left(\begin{bmatrix}\mu_Y^{(r)}\\ \mu_X^{(r)}\end{bmatrix}, \begin{bmatrix}\sigma_{YY}^{(r)} & \sigma_{XY}^{(r)}\\ \sigma_{XY}^{(r)} & \sigma_{XX}^{(r)}\end{bmatrix}\right), R \sim \text{Bernoulli}(\pi), \tag{1}$$

where $\mathrm{N}_2(\mu, \Sigma)$ denotes the bivariate normal distribution with mean $\mu$ and covariance matrix $\Sigma$. Since we have no data on $Y$ for the nonrespondents ($R = 0$), we cannot estimate all of the parameters in (1) for $R = 0$ without further assumptions. If assume that the missingness of $Y$ depends only on $X$, we can factor the joint distribution of $X$, $Y$, and $R$ into:

$$p(X, Y, R|\phi, \pi) = p(Y|X, R, \phi)p(X|R, \phi)p(R|\pi),$$

where $p()$ is the probability density function. Under the bivariate normality assumption and the property that the distribution of $Y$ given $X$ is independent of $R$, the parameters of the regression of $Y$ on $X$ are the same for $R = 1$ and $R = 0$, leading to a just-identified model. Little (1994) derives the ML estimates; in particular the ML estimate for $\hat{\mu}_Y$, the mean of $Y$ averaging over $R$, is:

$$\hat{\mu}_Y = \bar{Y}^{(1)} + \frac{s_{XY}}{s_{XX}}(\bar{X} - \bar{X}^{(1)}). \tag{2}$$

Suppose now that the missingness of $Y$ depends on $Y$ but not $X$. This implies that the parameters of the regression of $X$ on $Y$ are the same for $R = 1$ and $R = 0$, again leading to a just-identified model. The resulting ML for $\hat{\mu}_Y$ averaging over $R$ is:

$$\hat{\mu}_Y = \bar{Y}^{(1)} + \frac{s_{YY}}{s_{XY}}(\bar{X} - \bar{X}^{(1)}). \tag{3}$$

More generally, suppose that the missingness of $Y$ depends on the value of $Y^* = X + \lambda Y$ for a given $\lambda$. Little (1994) shows that the ML estimate for $\hat{\mu}_Y$ averaging over $R$ is then:

$$\hat{\mu}_Y = \bar{Y}^{(1)} + \frac{\lambda s_{YY} + s_{XY}}{\lambda s_{XY} + s_{XX}}(\bar{X} - \bar{X}^{(1)}). \tag{4}$$

It is easy to see that (4) reduces to (2) when the data is MAR ($\lambda = 0$), and to (3) when missingness depends only on $Y$ ($\lambda = \infty$). In practice, the data often provide no information about

the value of $\lambda$. Little (1994) suggests a sensitivity analysis to capture the uncertainty about $\lambda$ by estimating $\hat{\mu}_Y$ over a range of $\lambda$. Large differences in $\hat{\mu}_Y$ over $\lambda$ suggest that inferences on $\hat{\mu}_Y$ are sensitive to assumptions about the missingness mechanism. Alternatively, we can specify a prior distribution that reflects the uncertainty about the choice of $\lambda$.

## 2.1 Spline Pattern-Mixture Model

The BNPM model estimates rely heavily on the bivariate normality assumption between X and Y. For example, $(X, Y)$ is not bivariate normal if (a) the conditional distribution of $Y|X$ is normal with $E(Y|X) = 10 + X$ and the marginal distribution of $X$ is gamma, or (b) $X$ is normal but the regression of $Y$ on $X$ is quadratic in $X$; in such cases the estimates from the BNPM model are potentially biased even under the correct value of $\lambda$. We propose a penalized spline regression (S-BNPM) model for $X$ and $Y$ that relaxes the bivariate normality assumption.

Suppose that missingness depends on the value of $Y^* = X + \lambda Y$ for some known $\lambda > 0$. The conditional distribution of $X|Y^*$ is then the same for respondents and nonrespondents. The S-BNPM method creates multiple imputations of the missing values of $Y^*$ (and hence $Y = (Y^* - X)/\lambda$) so that the regression of $X$ on $Y^*$ for respondents (where $Y^*$ is observed) and nonrespondents (where $Y^*$ is imputed) follows the same spline regression model:

$$X|Y^* = \beta_0 + \beta_1 Y^* + \sum_{k=1}^{K} \gamma_k (Y^* - \kappa_k)_+ + \epsilon, \tag{5}$$

$$\epsilon \sim \mathrm{N}(0, \sigma^2),$$

$$\gamma_k \sim \mathrm{N}(0, \tau^2),$$

where $a_+ = a$ if $a > 0$ and $a_+ = 0$ otherwise, and $\kappa_1 < \ldots < \kappa_K$ are $K$ knots. The model may be fitted to the respondent data using a linear mixed model, treating the splines as random effects and adding a penalization term, $\alpha \sum_{k=1}^{K} \gamma_k^2$, to the log-likelihood to penalize the roughness of the regression function, where $\hat{\alpha}^2 = \hat{\sigma}^2/\hat{\tau}^2$ and the model parameters are estimated via restricted maximum likelihood (REML). Here, we adopt a Bayesian approach by assigning a uniform prior for $\beta$ and inverse gamma ($\upsilon = 10^{-5}, \omega = 10^{-5}$) priors for $\sigma^2$ and $\tau^2$, which have a mean of $\omega/(\upsilon - 1)$ when $\upsilon > 1$ and a variance of $\omega^2/[(\upsilon - 1)^2(\upsilon - 2)]$ when $\upsilon > 2$, and obtain draws from their posterior distributions using a Gibbs sampler (see Supplementary Material for details of the algorithm). We give $\gamma_1, \ldots, \gamma_K$ normal $\mathrm{N}(0, \tau^2)$ priors that result in an equivalent penalization to minimize over-fitting. While estimates are generally insensitive to the choice of the inverse gamma hyperparameters, small values (e.g. $10^{-2}$ or less) should be chosen to yield relatively noninformative but finite priors. In practice, the number of knots, and potentially a polynomial basis for the splines, may be chosen based on the sample size and the observed degree of nonlinearity. We recommend having at least 10 observations for each spline. For simplicity, we consider splines with a linear basis.

We then adopt a hot deck procedure (Andridge and Little, 2010) to impute the missing values of $Y^*$, where the missing value of $Y^*$ is imputed with the observed value of a matched donor with $X$ and $Y^*$ observed. The method involves the following steps:

1. Draw $B$ values of $Y^*$ for each nonrespondent from the distribution of $Y^*|X, R = 0$, estimated under the BNPM model. This results in a pool of $n_1 B$ values of $Y^*$ ($\{Y_p^*\}$, $p = 1, \ldots, n_1 B$). In the simulations in Section 3 a value of $B = 100$ is sufficient.

2. Given each $Y_p^*$ in the pool, draw a value $X_p$ from the posterior predictive distribution of $X|Y^*$ in (5), with parameters estimated from respondents. This results in a set of pairs of $(\{X_p, Y_p^*\}, p = 1, \ldots, n_1 B)$ that form our donor pool.
3. For each nonrespondent $j$, choose a pair $(X_k, Y_k^*)$ from the donor pool $(\{X_p, Y_p^*\}, p = 1, \ldots, n_1 B)$ with the closest value $X_k$ to $X_j$, and impute $Y_j^* = Y_k^*$ (hence $Y_j = (Y_k^* - X_j)/\lambda$) from that pair.
4. Repeat steps 2–3 above for 2000 iterations, deleting the first 1000 as burn-in and using every other 10 iterations to create $D = 100$ multiply-imputed data sets with values of $Y$ imputed. Using multiple imputation combining rules (Little and Rubin, 2020) we obtain $\hat{\mu}_Y$ and its variance:

$$\hat{\mu}_Y = \hat{\mu}_D = \frac{1}{D}\sum_{d=1}^{D} \hat{\mu}_d, \tag{6}$$

$$\text{Var}(\hat{\mu}_Y) = \frac{1}{D}\sum_{d=1}^{D} W_d + \frac{D+1}{D(D-1)}\sum_{d=1}^{D}(\hat{\mu}_d - \bar{\mu}_D)^2, \tag{7}$$

where $\hat{\mu}_d$ and $W_d$ are the estimated marginal mean and variance in the $d^{th}$ imputed data set, respectively. For the MAR assumption of $\lambda = 0$, we apply a Bayesian form of the PSPP method (Zhang and Little, 2009; Yang and Little, 2015). Specifically, we regress $Y$ on a spline of $X$ using the complete cases and impute $Y$ by drawing directly from its predictive posterior distribution in (5) given the observed $X^{(1)}$ for each iteration of the Gibbs algorithm.

The underlying rationale of the procedures is as follows. Since the unobserved $Y^*$ is a covariate in our spline model (5), we cannot impute $Y^*$ by drawing directly from a model. Thus we first create a donor pool of values $(\{Y_{ib}^*\}, b = 1, \ldots, B, i = n_0 + 1, \ldots, n_0 + n_1)$ as draws from the BNPM model. For each donor in the pool, we create a corresponding value of $X$ as a prediction from the spline model (5). We then match each incomplete case to a member of the donor pool with a similar value of $X$, and impute for that case the corresponding value of $Y^*$ from the donor. When the data are normal, the "hot-deck" matching step has little effect on the final imputations of $Y^*$. However, when data deviate from normality, the pairs $(X, Y^*)$ resulting from the hot-deck respect the spline model (5) and hence should improve on the imputations from the BNPM model, which incorrectly assume a linear relationship between $X$ and $Y^*$. In practice, we create multiple initial draws of $Y^*$ for each nonrespondent, as a large value of $B$ allows flexibility in the nonlinearity adjustment by S-BNPM and ensures a close match with the donors for every observed $X$. In the following examples we find a value of $B = 100$ to be sufficient to ensure a near-identical match in $X$.

As in the original BNPM model, the S-BNPM model utilizes the fact that, conditional on the variables contributing to missingness, the regression model parameters are the same for both respondents and nonrespondents. However, the penalized spline improves robustness of the pattern-mixture model by allowing us to model nonlinearity in the relationship between $X$ and $Y$. As suggested in Little (1994), inferences for $\hat{\mu}_Y$ should be displayed for a range of potential values of $\lambda$ to account for uncertainty about the true value of $\lambda$ and to assess sensitivity of inferences to the choice of $\lambda$.

## 2.2 Extensions of Proxy Pattern-Mixture Model Analysis

There may be multiple observed covariates $Z_1, \ldots, Z_p$ that are predictive of $\hat{\mu}_Y$. Andridge and Little (2011) proposed an extension of the pattern-mixture model analysis by taking $X$ as a proxy

obtained by regressing $Y$ on the set of $Z_1, \ldots, Z_p$ and replacing the set of covariates by $X$, the estimated best predictor of $Y$ given $Z_1, \ldots, Z_p$. Proxy pattern-mixture model analysis (PPMA) then estimates $\hat{\mu}_Y$ by applying the pattern-mixture model in Little (1994) to $X$ and $Y$. The advantage of reducing $Z_1, \ldots, Z_p$ to $X$ is simplicity: modelling departures from MAR under one sensitivity parameter $\lambda$ is much simpler than specifying a model with $p$ sensitivity parameters for each of $Z_1, \ldots, Z_p$. Moreover, should missingness depend on some other combination of $Z_1, \ldots, Z_p$ (e.g. $W = c_1 Z_1 + \ldots + c_p Z_p$, where $c_1, \ldots, c_p$ are constants), estimates for the mean of $Y$ are still approximately unbiased since $Y$ is independent of $W$ given $X$.

Andridge and Little (2011) showed that the uncertainty of the estimates of $\hat{\mu}_Y$ depends largely on the degree of correlation between the proxy $X$ and $Y$ as well as the degree of similarity between respondents and nonrespondents with respect to the value of $X$. When $X$ and $Y$ are highly correlated and the values of $X$ are similar for respondents and nonrespondents, information on missing values of $Y$ and evidence on the lack of response bias are both strong, resulting in estimates of $\hat{\mu}_Y$ with high precision. However, if $X$ and $Y$ are weakly correlated and the values of $X$ are much different for respondents and nonrespondents, we have strong evidence for response bias with little information on the missing values of $Y$, resulting in estimates of $\hat{\mu}_Y$ with high uncertainty.

### 2.3 Spline Proxy Pattern-Mixture Model

As in the bivariate case, validity of the proxy pattern-mixture model proposed by Andridge and Little (2011) when data are MNAR relies on the assumption of bivariate normality between the proxy $X$ and $Y$, which is violated when some or all of the $Z_1, \ldots, Z_p$ used to obtain $X$ are not normally distributed. Suppose, for example, $Z$ is a fully observed standard normal variable and $Y$ given $Z$ is normal with mean $Z + Z^2$. Let $X$ be a proxy from the regression of $Y$ on $Z$ and $Z^2$. When the data is MAR, $X$ is an unbiased predictor of $Y$, hence estimates from the pattern-mixture model under $\lambda = 0$ are unbiased. However, when missingness depends on $Y$, the resulting proxy $X$ is no longer an unbiased predictor of $Y$ since the regression coefficients in the regression of $Y$ on $Z$ and $Z^2$ based on the respondents are biased for the nonrespondents. Since $X$ is some function of $Z$ and $Z^2$ which is not normally distributed, the assumption of bivariate normality, hence linearity, with $Y$ fails, resulting in biased estimates for all values of $\lambda$.

We propose a modification of the proxy pattern-mixture model that relaxes the assumption of bivariate normality between $X$ and $Y$. Suppose, as before, $X$ is the predicted value of $Y$ based on regression of $Y$ on $Z_1, \ldots, Z_p$ for the complete cases, and that missingness depends on the value of $Y^*$. The conditional distribution of $X$ given $Y^*$ is independent of $R$ and the regression coefficients of $X$ on $Y^*$ are the same for both respondents and nonrespondents. The model proposed in Andridge and Little (2011) assumes linearity between $X$ and $Y$, and hence $Y^*$, which as discussed may not be appropriate when $X$ and $Y$ are not bivariate normal. Thus, we propose a spline proxy pattern-mixture model analysis (S-PPMA) to describe the relationship between $X$ and $Y$. Under S-PPMA, we first estimate the proxy based on a complete-case regression of $Y$ on $Z_1, \ldots, Z_p$ as in Andridge and Little (2011), and set $X$ as the predicted value of $Y$ from this regression. Then, we apply a penalized spline model to $X$ and $Y$ and estimate $\hat{\mu}_Y$ as discussed in Section 2.1. As in the bivariate model, we believe S-PPMA will further enhance the robustness of PPMA by relaxing the bivariate normality assumption.

In the next section, we describe simulation studies to assess the performance of S-PPMA under various distributions of $Z_1, \ldots, Z_p$, $Y$, and missingness mechanisms. For comparison we include estimates from the proxy pattern-mixture model proposed in Andridge and Little (2011).

## 3 Simulation Studies

We assess the performance of S-PPMA for inferences about the mean of $Y$ with respect to average bias, root mean square error, 95% confidence interval width, and rate of confidence interval non-coverage over 1000 replications and six scenarios. For each replication, we construct 95% confidence intervals and estimate the non-coverage rate as the proportion of the 1000 confidence intervals that do not cover the true value, where 95% CI $= (\hat{\mu}_Y - t_{n-1,0.975}\sqrt{\text{Var}(\hat{\mu}_Y)}, \hat{\mu}_Y + t_{n-1,0.975}\sqrt{\text{Var}(\hat{\mu}_Y)})$, $t_{n-1,0.975}$ is the $97.5^{th}$ percentile of the t-distribution with $n-1$ degrees of freedom, and $\text{Var}(\hat{\mu}_Y)$ is the estimated variance of the mean in (7). Confidence interval widths (CIW) are computed as CIW $= 2t_{n-1,0.975}\sqrt{\text{Var}(\hat{\mu}_Y)}$, or for credibility intervals, the difference between the $97.5^{th}$ and $2.5^{th}$ percentiles of the posterior distribution of $\hat{\mu}_Y$. For all simulations, we set sample sizes of $n = 100$ and $n = 400$ over 1000 replications and apply the penalized spline models using $K = 2$ and $K = 5$ knots, respectively.

For the first scenario, we assume bivariate normal data of $X$ and $Y$ and compare estimates of the mean of $Y$ under the BNPM and S-BNPM models. For scenarios 2–5, we assume a set of fully observed covariates $Z_1, \ldots, Z_p$. Here, we first obtain the proxy $X$ from a correctly specified regression of $Y$ on $Z_1, \ldots, Z_p$ using the respondent sample. Then, we estimate the mean of $Y$ using three methods:

1. We apply the S-PPMA model to $X$ and $Y$ using a penalized spline in (5). (S-PPMA)
2. We assume bivariate normality between $X$ and $Y$ and estimate $\hat{\mu}_Y$ via maximum likelihood in (4) as originally proposed in Andridge and Little (2011). Variance is estimated using 200 bootstrap samples. (PPMA-ML)
3. We assume bivariate normality between $X$ and $Y$ and draw $\hat{\mu}_Y$ from its posterior distribution as described in Little (1994). The 95% credibility intervals and coverage are based on draws from the posterior distribution. (PPMA-BAYES)

Let $\lambda_T$ be the true, unobservable value of $\lambda$ generating missing data, and let $\lambda_A$ be the assumed value of $\lambda$ in our models. For each scenario, we simulate nonresponse using $\lambda_T = 0$, 1 and $\infty$. To assess sensitivity of inferences to $\lambda_A$, we produce estimates under $\lambda_A = 0$, 1 and $\infty$ for each value of $\lambda_T$, one of which corresponds to the true underlying value of $\lambda_T$. While inferences under additional values of $\lambda_A$ may be explored, we chose these three values to capture a range of potential missingness mechanisms. In the following section, only results for which $\lambda_A = \lambda_T$ are shown (for rest, see Supplementary Material).

### 3.1 Scenario 1: Bivariate Normal Data

We assume a fully observed covariate $X$ and a $Y$ that is bivariate normal with $X$ and subject to missingness. The data is generated under the following pattern-mixture model with sample sizes of $n = 100$ and $n = 400$:

$$\begin{aligned} R &\sim \text{Bernoulli}(0.5), \\ X, Y | R = 1 &\sim \text{N}_2\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0.5\\0.5 & 1\end{bmatrix}\right), \\ X | R = 0 &\sim \text{N}(1, 1). \end{aligned}$$

In this and all subsequent scenarios, nonresponse rates are approximately 50%. For simplicity we only display results at $n = 400$, as results for $n = 100$ are generally similar (see Supplementary Material). Figure 1 displays the performances of each estimator in terms of av-

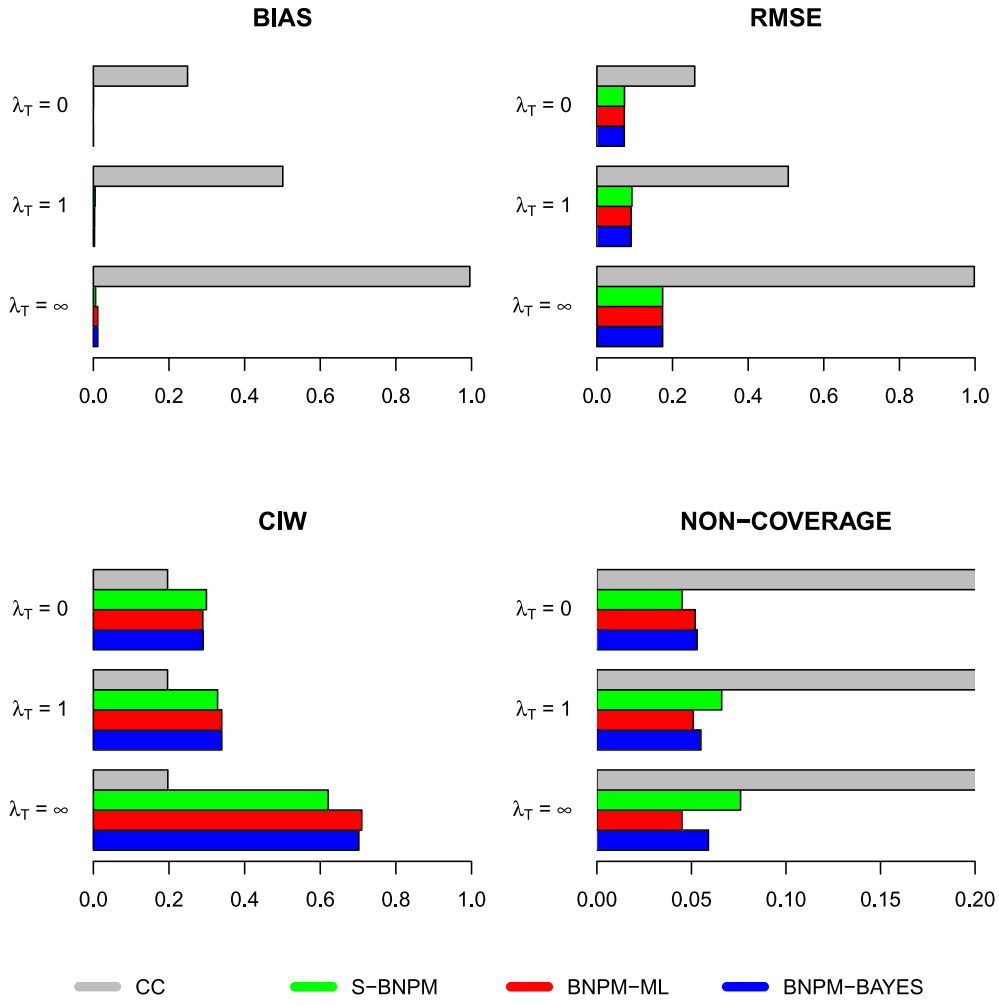


Figure 1: Results for scenario 1 where $\lambda_A = \lambda_T$.

erage bias, root mean squared error (RMSE), 95% CIW, and its corresponding non-coverage rate out of 1000 replications when $\lambda_A = \lambda_T$. In the figure, the true missingness of $Y$ depends on $X + \lambda_T Y$ for $\lambda_T = 0$, 1, and $\infty$. Results show little differences between the methods in bias, RMSE, and CIW regardless of $\lambda_A$ in all values of $\lambda_T$ (results for $\lambda_A \neq \lambda_T$ in Supplementary Material). As expected, when $\lambda_A = \lambda_T$, all estimates are approximately unbiased and non-coverages are near the nominal 5%, as BNPM is the correct model for the data. Moreover, CIW increases as $\lambda_T$ increases, reflecting a rise in uncertainty as a result of nonresponse due to $Y$. We notice that the CIW for S-BNPM at $\lambda_A = \infty$ is narrower than that for BNPM under both ML and Bayes for all values of $\lambda_T$. This may be due to a small correlation of 0.5 between $X$ and $Y$, which may lead to a large value of $\frac{s_{YY}}{s_{XY}}$ in (3) and consequently an extreme $\hat{\mu}_Y$. In S-BNPM, the process of generating multiple initial draws of the missing $Y$ and matching on the donor pool based on predictions from the spline model helps to alleviate this problem as draws of $X$ from extreme values of $Y$ are less likely to be matched to observed values of $X$, leading to less extreme imputations in this particular scenario.

### 3.2 Scenario 2: Bivariate Non-normal Data

Suppose $X$ is a fully observed, gamma-distributed covariate and $Y$ is normal conditional on $X$ and is subject to missingness. We generate the data under a selection model with sample sizes of $n = 100$ and $n = 400$:

$$X \sim \text{Gamma}(1, 0.25),$$
$$Y|X \sim \text{N}(10 + X, 1).$$

We generate missing values of $Y$ under the following models to reflect both MAR and MNAR scenarios, assuming an unobserved latent variable $U$:

$$U|X, Y \sim \text{N}(-1.5 + 0.5X, 1), \qquad \text{(A. } \lambda_T = 0\text{)}$$
$$U|X, Y \sim \text{N}(-2.5 + 0.15(X + Y), 1), \qquad \text{(B. } \lambda_T = 1\text{)}$$
$$U|X, Y \sim \text{N}(-3.5 + 0.25Y, 1), \qquad \text{(C. } \lambda_T = \infty\text{)}$$

where $Y$ is missing if $U > 0$ and observed otherwise.

In this scenario we include estimates from the true model, which models $Y^*$ on $U$ and $X$ for $\lambda_T > 0$, since $Y^*$ and $U$ are bivariate normal conditional on $X$. Since $U$ is unobserved, we estimate $U$ by introducing a latent variable $U^*$, and produce posterior draws of the missing $Y^*$ iteratively by the following steps:

1. Initialize values of $Y^{*(0)}$ and $U^*$ by setting $Y^{*(0)}$ as predictions from the regression of $Y^*$ on $X$ under the complete cases, and draw $U^*$ from a normal distribution with variance 1 and mean $Z_{\hat{\pi}} - \bar{Y} + Y$, where $\hat{\pi}$ is the nonresponse rate, $Z_\alpha$ is the $\alpha^{th}$ percentile of the standard normal distribution, and $\bar{Y}$ is the estimated mean combining the observed $Y^{*(1)}$ and the initialized $Y^{*(0)}$. For respondents, positive values of $U^*$ are discarded and redrawn until all values are negative. Likewise for nonrespondents, we discard and redraw negative values of $U^*$.
2. At the $i$th iteration, obtain posterior predictive draws of $Y^{*(0)}_{(i)}|U^{*(0)}_{(i-1)}, X^{(0)}$ under a linear regression model with parameters estimated from $Y^*|U^*_{(i-1)}, X$ under the entire imputed sample, using values of $Y^{*(0)}_{(i-1)}$ and $U^{*(r)}_{(i-1)}$ drawn from the previous iteration.
3. Obtain posterior predictive draws of $U^*_{(i)}|Y^{*(1)}, Y^{*(0)}_{(i)}, X$ under a linear regression model for the entire sample, where $Y^{*(0)}_{(i)}$ are predictive draws for the missing $Y^{*(0)}$ at the current iteration. We again discard and redraw all positive values of $U^*_{(i)}$ for respondents and negative values of $U^*_{(i)}$ for nonrespondents.
4. Repeat steps 2 and 3 over 1000 iterations, discarding the first 100 as burn-in. We then apply (6) and (7) over the 900 sets of drawn values of $Y^{*(0)}$ to estimate the mean and variance.

For $\lambda_T = 0$, we impute the missing $Y$ based on posterior predictive draws from the regression of $Y$ on $X$ on the complete cases.

Figure 2 displays results under $\lambda_A = \lambda_T$ for $\lambda_T = 0$, 1, and $\infty$. As in Scenario 1, we only display results at $n = 400$, as results for $n = 100$ are generally similar (see Supplementary Material). When $\lambda_T = 0$, all methods are unbiased, with S-BNPM having slightly higher RMSE and more conservative 95% confidence intervals. Since data is MAR and $Y|X$ is normal with a mean that is linear on $X$, the BNPM model is correctly specified and thus it is not surprising that its estimates are unbiased and have better precision than S-BNPM. However, when $\lambda_T = 1$, linearity assumptions for $X|Y^*$ are violated, and consequently we see bias and under-coverage by BNPM. Here, S-BNPM shows reductions in bias and to a lesser extent RMSE, and achieves

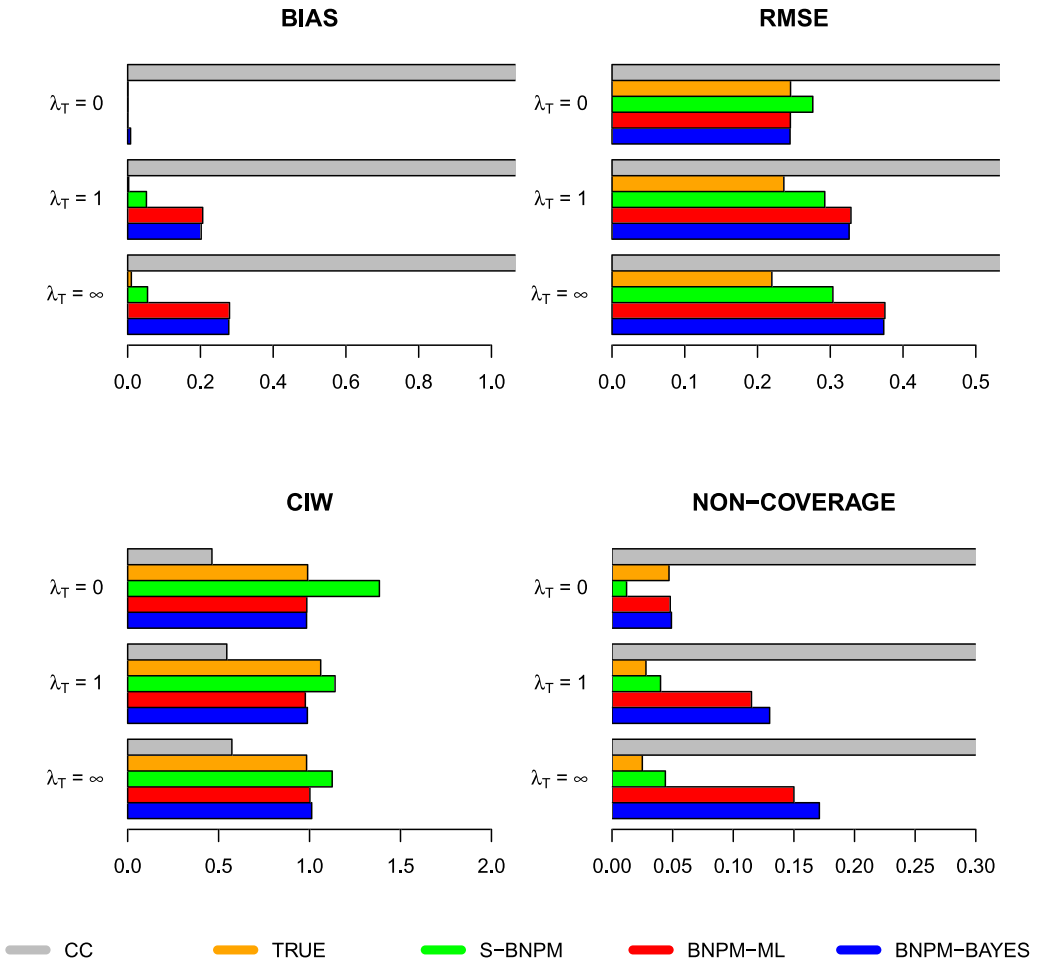


Figure 2: Results for scenario 2 where $\lambda_A = \lambda_T$.

near nominal 5% non-coverage with a minor penalty in RMSE and precision compared to the true model. The more the data deviate from MAR, the higher the gains in bias and RMSE from S-BNPM, as evident in the results under $\lambda_T = \infty$. S-BNPM shows a noticeable improvement in RMSE over BNPM and still yields close to nominal non-coverage. Robustness to normality, however, comes at the price of precision, as S-BNPM tends to yield wider intervals than both BNPM and the true model.

### 3.3 Scenario 3: Set of Normal $Z$'s

In this scenario, we assume a set of covariates that are normally distributed. Let $Z_1$, $Z_2$, $Z_3$ be fully observed covariates with distributions:

$$\begin{aligned} Z_1 &\sim \mathrm{N}(0, 1), \\ Z_2 &\sim \mathrm{N}(0, 1), \\ Z_3 &\sim \mathrm{N}(0, 1), \\ Y|Z_1, Z_2, Z_3 &\sim \mathrm{N}(15 + Z_1 + 2Z_2 + Z_3, 1). \end{aligned}$$

Let $Y$ be missing under the following logistic models:

$$\text{Logit}[\Pr(R=0)] = 0.5(Z_1 + 2Z_2 + Z_3), \quad \text{(A. } \lambda_T = 0\text{)}$$
$$\text{Logit}[\Pr(R=0)] = -3.5 + 0.25(0.98Z_1 + 1.95Z_2 + 0.98Z_3 + Y), \quad \text{(B. } \lambda_T = 1\text{)}$$
$$\text{Logit}[\Pr(R=0)] = -7.5 + 0.5Y, \quad \text{(C. } \lambda_T = \infty\text{)}$$
$$\text{Logit}[\Pr(R=0)] = 2Z_2, \quad \text{(D)}$$
$$\text{Logit}[\Pr(R=0)] = -7.5 + 0.5(2Z_2 + Y). \quad \text{(E)}$$

For each missingness mechanism, we obtain the proxy $X$ by regressing $Y$ on $Z_1$, $Z_2$, and $Z_3$ apply the estimators to $X$ and $Y$. Figure 3 shows results for $\lambda_A = \lambda_T$, with $\lambda_T = 0$, 1, and $\infty$ under $n = 400$ (see Supplementary Material for rest of results). In addition there are two nonresponse mechanisms, D and E, that do not correspond to any $\lambda_T$. When $\lambda_T = 0$, $Y$ is MAR, $\lambda_A = 0$ is the correct assumption about nonresponse and as a result all estimators are approximately unbiased and yield similar RMSE, confidence interval widths, and near-nominal non-coverage of 5%. For values of $\lambda_A = 1$ and $\infty$ when $\lambda_T = 0$, all three methods exhibit bias, with negligible differences in RMSE, CIW, and non-coverage (not shown). Similarly when $\lambda_T = 1$ and $\infty$, values of $\lambda_A$ such that $\lambda_A = \lambda_T$ result in negligible bias and near nominal non-coverage for all estimators. For values of $\lambda_A$ such that $\lambda_A \neq \lambda_T$, all methods are biased with higher than nominal non-coverage, as expected given that the assumptions about nonresponse are wrong. Results for mechanism D (not shown) are generally similar to those of A, where $\lambda_T = 0$. Here, all methods have negligible bias and nominal non-coverage at $\lambda_A = 0$ and yield similar RMSE and CIW at all values of $\lambda_A$. In mechanism E, all methods have minor bias at $\lambda_A = 1$ and cover the true mean at a rate close to 95%, with minor differences in RMSE and CIW regardless of $\lambda_A$. In this scenario, nonresponse mechanisms D and E do not deviate much from mechanisms A and B, which explains the similarity of results.

This scenario assumes that all auxiliary variables are normally distributed, resulting in a proxy $X$ that is normal and linear with $Y$ regardless of the nonresponse mechanism. As such, the methods in Andridge and Little (2011) produce valid estimates under the correct value of $\lambda_A$. We again notice that S-PPMA tends to yield slightly more conservative confidence intervals than PPMA, which suggests there is some penalty in precision from fitting a more robust model when normality assumptions are met.

### 3.4 Scenario 4: Varying Distributions of $Z$

Let $Z_1$, $Z_2$, $Z_3$ be fully observed covariates with the following distributions:

$$Z_1 \sim \text{N}(0, 1),$$
$$Z_2 \sim \text{Gamma}(1, 1),$$
$$Z_3 \sim \text{Bernoulli}(0.5),$$
$$Y|Z_1, Z_2, Z_3 \sim \text{N}(10 + Z_1 + 4Z_2 + Z_3, 1).$$

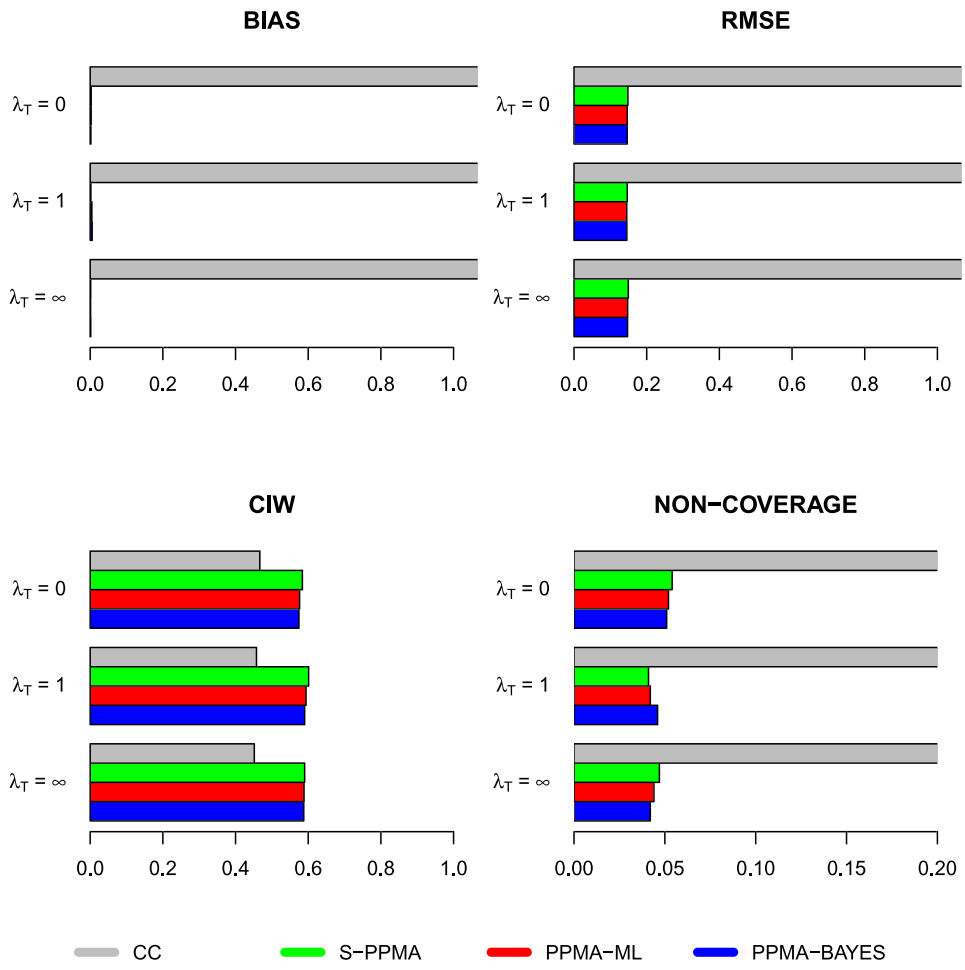


Figure 3: Results for scenario 3 where $\lambda_A = \lambda_T$.

Let $Y$ be missing under the following logistic models:

$$\text{Logit}[\Pr(R=0)] = -2 + 0.5(Z_1 + 4Z_2 + Z_3), \quad \text{(A. } \lambda_T = 0\text{)}$$
$$\text{Logit}[\Pr(R=0)] = -4.5 + 0.25(0.98Z_1 + 3.9Z_2 + 0.98Z_3 + Y), \quad \text{(B. } \lambda_T = 1\text{)}$$
$$\text{Logit}[\Pr(R=0)] = -7 + 0.5Y, \quad \text{(C. } \lambda_T = \infty\text{)}$$
$$\text{Logit}[\Pr(R=0)] = -1 + Z_2, \quad \text{(D)}$$
$$\text{Logit}[\Pr(R=0)] = -4 + 0.25(2Z_2 + Y). \quad \text{(E)}$$

We obtain the proxy by regressing $Y$ on $Z_1$, $Z_2$, and $Z_3$ using respondent data and apply the estimators under $\lambda_A = 0$, 1 and $\infty$. Results for which $\lambda_A = \lambda_T$ under $n = 400$ are shown in Figure 4 (see Supplementary Material for rest of results). Mechanisms D and E do not correspond to any value of $\lambda_T$. In this scenario we vary the distributions of the auxiliary variables and the conditional mean of $Y$ given $Z_1$, $Z_2$, and $Z_3$ is dominated by a gamma distributed $Z_2$. For $\lambda_T = 0$ where $Y$ is MAR, all three methods yield approximately unbiased means with close to nominal non-coverage when the correct value of $\lambda_A = 0$ is used. Under the incorrect values of $\lambda_A = 1$ and $\infty$, however, the S-PPMA has lower bias, lower RMSE, and lower non-coverage rate than the linear models albeit with more conservative confidence intervals (not shown).

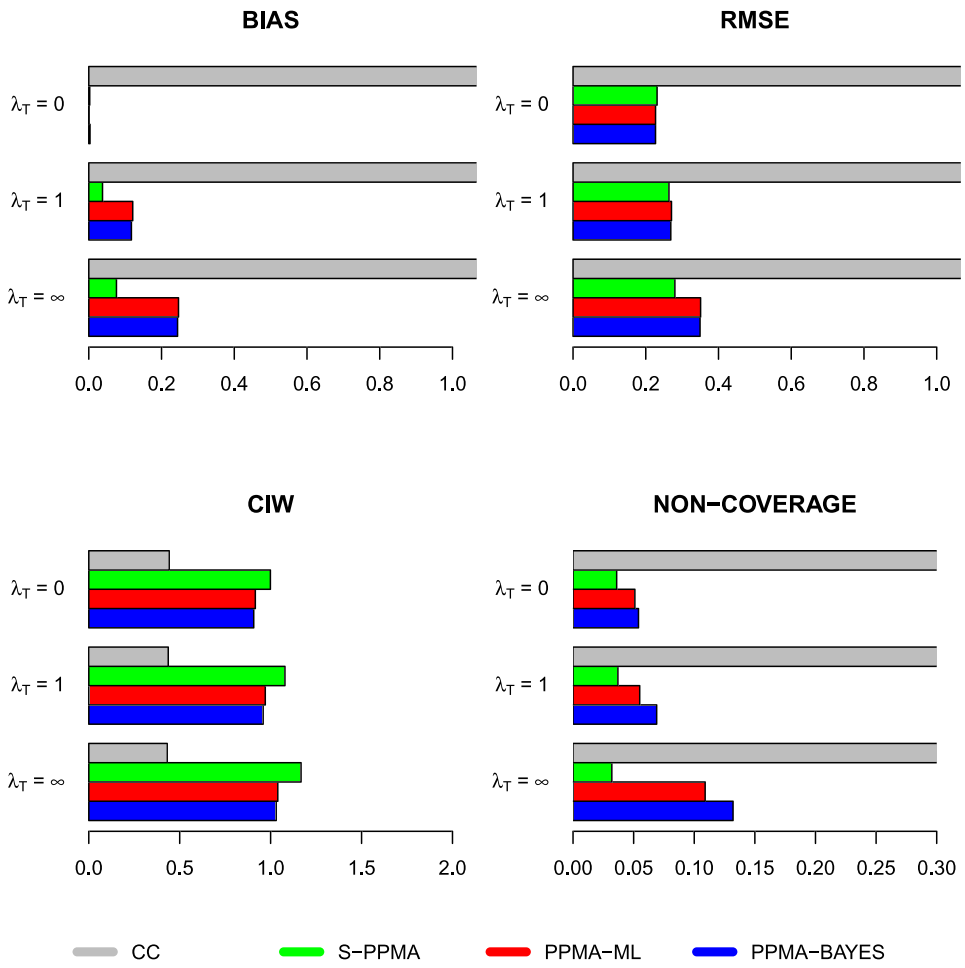


Figure 4: Results for scenario 4 where $\lambda_A = \lambda_T$.

For $\lambda_T = 1$ and $\infty$, the PPMA estimates exhibit small bias even when $\lambda_A = \lambda_T$, most likely as a result of lack of linearity between $X$ and $Y$ due to MNAR and some of the auxiliary variables being non-normal. The S-PPMA estimates at the correct $\lambda_A$ show low bias and non-coverages close to 5%, which may be explained by the spline's ability to model nonlinearity between $X$ and $Y$. It is worth noting, however, that despite the bias PPMA still achieves good coverage at $\lambda_A = \lambda_T = 1$. In terms of RMSE, S-PPMA has no noticeable gains over PPMA under $\lambda_A = \lambda_T = 1$, and larger gains when $\lambda_A = \lambda_T = \infty$. This suggests that as dependence of nonresponse on $Y$ increases, the degree of nonlinearity adjustment by the penalized spline increases. Robustness to $\lambda_T$ comes at the expense of precision, as the penalized spline yields wider intervals under all values of $\lambda_A$ for any $\lambda_T$. However, it is important to note that values of $Y$ tend to be much lower for respondents than nonrespondents as a result of the nonresponse mechanism, which leads to sparse data and extrapolation at higher values of $Y$. Thus, wider interval widths by the spline may be a reflection of uncertainty in imputing the missing values by extrapolating a nonlinear model. For mechanism D (not shown), there are no significant differences in RMSE and CIW regardless of $\lambda_A$, with negligible bias at $\lambda_A = 0$ and close to nominal coverage at both $\lambda_A = 0$ and 1 for all methods. In mechanism E, both S-BNPM and BNPM yield similar estimates with nominal non-coverage at $\lambda_A = 1$.

### 3.5 Scenario 5: Quadratic Term in Mean of $Y$

Let $Z_1$ and $Z_2$ be fully observed covariates with the following distributions:

$$Z_1 \sim \mathrm{N}(0, 1),$$
$$Z_2 \sim \mathrm{N}(0, 1),$$
$$Y|Z_1, Z_2 \sim \mathrm{N}(10 + Z_1 + Z_2 + 2Z_2^2, 1).$$

Let $Y$ be missing under the following mechanisms:

$$\mathrm{Logit}[\Pr(R = 0)] = -1 + 0.5(Z_1 + Z_2 + 2Z_2^2), \quad \text{(A. } \lambda_T = 0\text{)}$$
$$\mathrm{Logit}[\Pr(R = 0)] = -3 + 0.25(0.97Z_1 + 0.97Z_2 + 1.95Z_2^2 + Y), \quad \text{(B. } \lambda_T = 1\text{)}$$
$$\mathrm{Logit}[\Pr(R = 0)] = -6 + 0.5Y, \quad \text{(C. } \lambda_T = \infty\text{)}$$
$$\mathrm{Logit}[\Pr(R = 0)] = 4Z_2, \quad \text{(D)}$$
$$\mathrm{Logit}[\Pr(R = 0)] = -5.5 + 0.5(4Z_2 + Y). \quad \text{(E)}$$

We estimate the proxy $X$ by regressing $Y$ on $Z_1$, $Z_2$, and $Z_2^2$ using the complete cases and apply the estimators under the different values of $\lambda_A$. Here we introduce a quadratic term in the conditional mean of Y. For $\lambda_T = 0$, when data is MAR, the estimated proxies are unbiased estimates of $Y$ since they are based on a correctly specified regression model. As a result all methods are unbiased with close to nominal 5% non-coverage when we assume the correct value of $\lambda_A = 0$, with the spline having slightly wider interval widths (Figure 5). For other values of $\lambda_A$, the S-PPMA shows smaller bias, lower RMSE, and much higher coverage rate than their linear counterparts, and still achieves near nominal non-coverage under the incorrect assumption of $\lambda_A = 1$ (not shown).

For $\lambda_A = \lambda_T = 1$, where missingness depends equally on both $Y$ and the auxiliary variables, estimates under $\lambda_A = 0$ (see Supplementary Material) are similarly biased and intervals undercover the true value for all methods, which is not surprising since the assumption of $\lambda_T$ is incorrect. However, S-PPMA has minor bias under $\lambda_A = 1$, which is the correct assumption in this case shown in Figure 5, and near nominal non-coverage rates under both assumptions of $\lambda_A = 1$ and $\lambda_A = \infty$, where the PPMA estimates are biased and undercover the true value. With respect to RMSE, S-PPMA shows increasing gains over PPMA as $\lambda_A$ increases.

When $\lambda_T = \infty$, where missingness depends only on $Y$, the penalized spline is again approximately unbiased with nominal non-coverage under the correct assumption of $\lambda_A = \infty$, while the linear models are heavily biased. This is due to nonlinearity between $X$ and $Y$ caused by the quadratic $Z_2^2$ term in the mean of $Y$, violating the bivariate normality assumption required in PPMA. Although the spline yields more conservative intervals, possibly from extrapolating nonlinearity, its ability to model nonlinearity results in estimates that are unbiased and have good coverage rates. This is especially important when the missingness mechanism is MNAR, where the proxy $X$ is no longer unbiased and has a nonlinear relationship with $Y$. It is interesting to note, however, that in this and the previous scenario, PPMA shows slightly lower RMSE at the wrong assumption of $\lambda_A = 1$ when the true value is $\lambda_T = \infty$ (see Supplementary Material).

In mechanism D (not shown), the methods show low bias and similar RMSE at all values of $\lambda_A$, with the ML estimate of BNPM having significantly wider intervals than S-BNPM and the Bayesian estimate of BNPM, resulting in better coverage. In mechanism E, all methods are generally biased and fail to achieve nominal non-coverage regardless of $\lambda_A$, with small differences in RMSE. Again the ML estimate of BNPM tends to yield much wider intervals that result in better coverage.

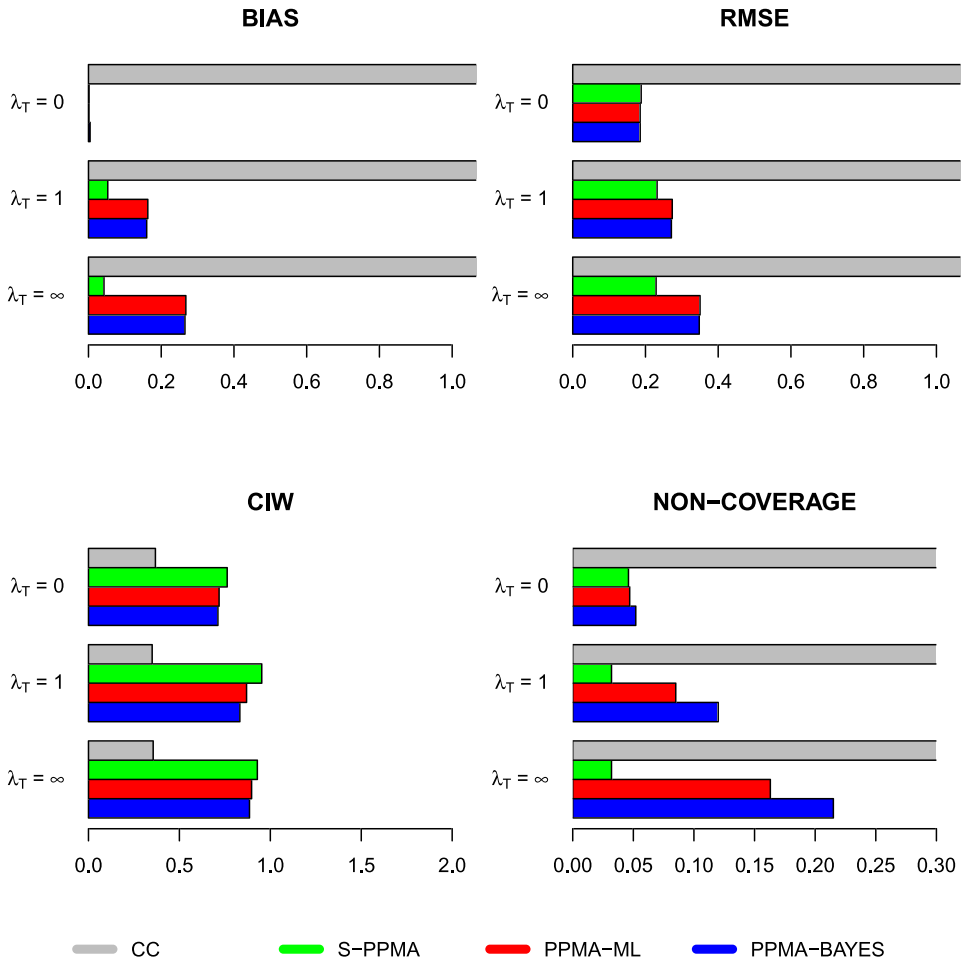


Figure 5: Results for scenario 5 where $\lambda_A = \lambda_T$.

### 3.6 Scenario 6: Interaction Term in Mean of $Y$

Let $Z_1$ and $Z_2$ be fully observed covariates with the following distributions:

$$Z_1 \sim \mathrm{N}(0, 1),$$
$$Z_2 \sim \mathrm{N}(0, 1),$$
$$Y|Z_1, Z_2 \sim \mathrm{N}(20 + Z_1 + Z_2 + 2Z_1Z_2, 1).$$

Let $Y$ be missing under the following mechanisms:

$$\mathrm{Logit}[\Pr(R = 0)] = Z_1 + Z_2 + 2Z_1Z_2, \quad \text{(A. } \lambda_T = 0\text{)}$$
$$\mathrm{Logit}[\Pr(R = 0)] = -5 + 0.25(0.98Z_1 + 0.98Z_2 + 1.96Z_1Z_2 + Y), \quad \text{(B. } \lambda_T = 1\text{)}$$
$$\mathrm{Logit}[\Pr(R = 0)] = -10 + 0.5Y, \quad \text{(C. } \lambda_T = \infty\text{)}$$
$$\mathrm{Logit}[\Pr(R = 0)] = 5Z_2, \quad \text{(D)}$$
$$\mathrm{Logit}[\Pr(R = 0)] = -10 + 0.5(5Z_2 + Y). \quad \text{(E)}$$

In this last scenario, we let the conditional mean of $Y$ be a function of two normally distributed variables and their interaction. We then model $Y$ using a correctly specified regression

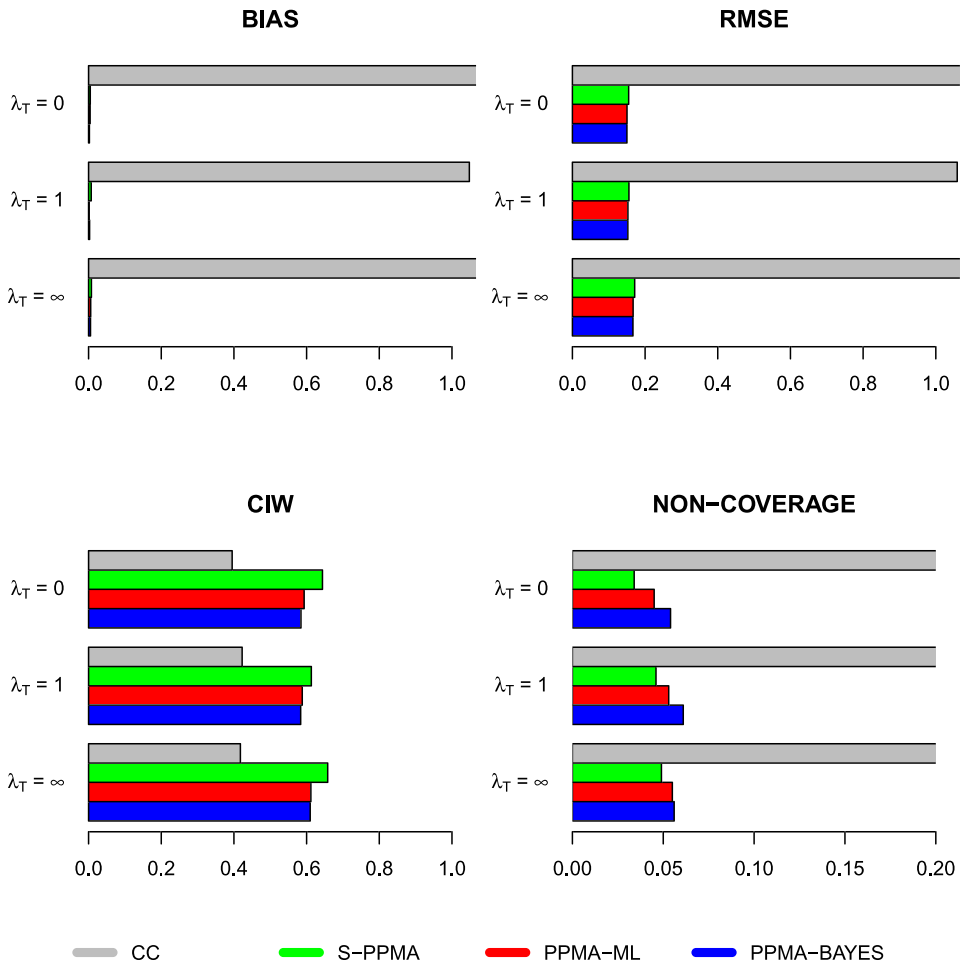


Figure 6: Results for scenario 6 where $\lambda_A = \lambda_T$.

on $Z_1$, $Z_2$, and $Z_1Z_2$ for the respondents, and obtain the predicted values of $Y$ as our proxy $X$. As in all scenarios, Figure 6 shows that when missingness is at random, all methods are unbiased, yield similar RMSE, and achieve nominal non-coverage under $\lambda_A = 0$ since the proxy $X$ itself is unbiased for $Y$. However, under the incorrect values of $\lambda_A = 1$ and $\lambda_A = \infty$, the S-PPMA shows significantly larger bias, RMSE, and CIW than PPMA (see Supplementary Material).

When $\lambda_T = 1$, all methods have negligible bias under the correct value of $\lambda_A = 1$ as shown in Figure 6, and achieve close to 5% non-coverage. There are generally minor differences in RMSE between the methods regardless of the assumption in $\lambda_A$, though S-PPMA tends to be slightly more conservative in terms of interval widths. For $\lambda_T = \infty$, all methods yield low bias with similar RMSE at $\lambda_A = \infty$ and nominal non-coverage. All methods have similar bias, RMSE, CIW, and coverage at all other values of $\lambda_A$ (not shown). Although the mean of $Y$ in this scenario depends on the interaction of $Z_1$ and $Z_2$, which is not normally distributed, the model assuming linearity between $X$ and $Y$ still yields good estimates of the mean under MNAR when $\lambda_A = \lambda_T$. This may be because the distribution of $Z_1Z_2$ does not result in a drastic departure from normality in the proxy $X$, so the bivariate normality assumption between $X$ and $Y$ still approximately holds.

In the results for mechanism D (not shown), which does not correspond to any value of $\lambda_T$, estimates at $\lambda_A = 0$ are generally unbiased with minor differences in RMSE, and achieve close to nominal non-coverage with the exception of the Bayesian BNPM. In mechanism E, all methods show some bias at all values of $\lambda_A$ with S-BNPM yielding higher RMSE than BNPM at $\lambda_A = \infty$.

## 4 Example: Child Asthma Study

We apply S-PPMA and PPMA to an asthma study conducted by the University of Michigan Schools of Public Health and Medicine. The study consists of children with asthma from Detroit elementary and middle schools, whose aim is to evaluate the effectiveness of an educational intervention in reducing asthma symptoms. The main outcome of interest is the average number of nights the child experiences asthma symptoms per month, collected at baseline and one-year follow-up. Our goal is to estimate the mean change in nights of symptoms per month from baseline to follow-up, which is subject to dropout. However, since it is well documented that asthma severity naturally declines as the child ages, we restrict our attention to only those in the control group with symptoms at baseline.

Out of 133 children ages 6 to 14 with asthma symptoms at baseline in the control group, 41 (31%) dropped out before follow-up information was obtained. Since dropout may be attributed to asthma severity, we apply the S-PPMA and PPMA models to estimate the mean change in nights of symptoms per month. Only age and measurement at baseline are significantly associated with the outcome, with baseline age also being significantly associated with response. We first obtain our proxy by regressing change in nights per month on its baseline value and age using the respondent sample. We then apply the S-PPMA and PPMA models to estimate the mean change in nights of symptoms per month.

Figure 7 shows the distributions of baseline age and nights per month in our data. Both variables show deviations from normality, particularly nights of symptoms per month. Figure 8 displays scatter plots for the relationship between $X$ and $Y$ along with the average regression lines for PPMA and S-PPMA. For the regression of $Y$ on $X$ under the assumption of $\lambda = 0$, both PPMA and S-PPMA yield near identical regression lines. However, differences can be seen for the regression of $X$ on $Y$ under the assumption of $\lambda = \infty$, where S-PPMA seems to provide a minor improvement in fit. As such, we expect some differences between estimates from S-PPMA and PPMA, particularly at $\lambda = \infty$. Figure 9 shows estimates of the mean change under each method. Each line represents the mean and its 95% confidence interval for S-PPMA (PS) and PPMA, which is estimated using both maximum likelihood bootstrap (ML) and posterior draws (PD). To assess sensitivity to our assumption about $\lambda$, we display estimates under $\lambda =$ 0, 0.5, 1, 4, and $\infty$. Results show that the mean change in symptoms per month generally decreases as we place more weight on our outcome to response, which suggests that children with higher decrease in symptoms may be less likely to participate in the follow-up survey. As expected from Figure 8, estimates for PPMA and S-PPMA at $\lambda = 0$ are similar, with differences between the methods being most pronounced at $\lambda = \infty$. There are minor differences between the PPMA estimates, with the posterior draws generally producing more conservative intervals than maximum likelihood. As in the simulations, interval lengths tend to widen slightly as $\lambda$ increases due to increasing uncertainty when missingness depends on the outcome. S-PPMA is to a small degree less sensitive to assumptions about $\lambda$ than PPMA, as estimates of mean change are within 0.1 nights of each other for values of $\lambda > 0$, whereas estimates from PPMA

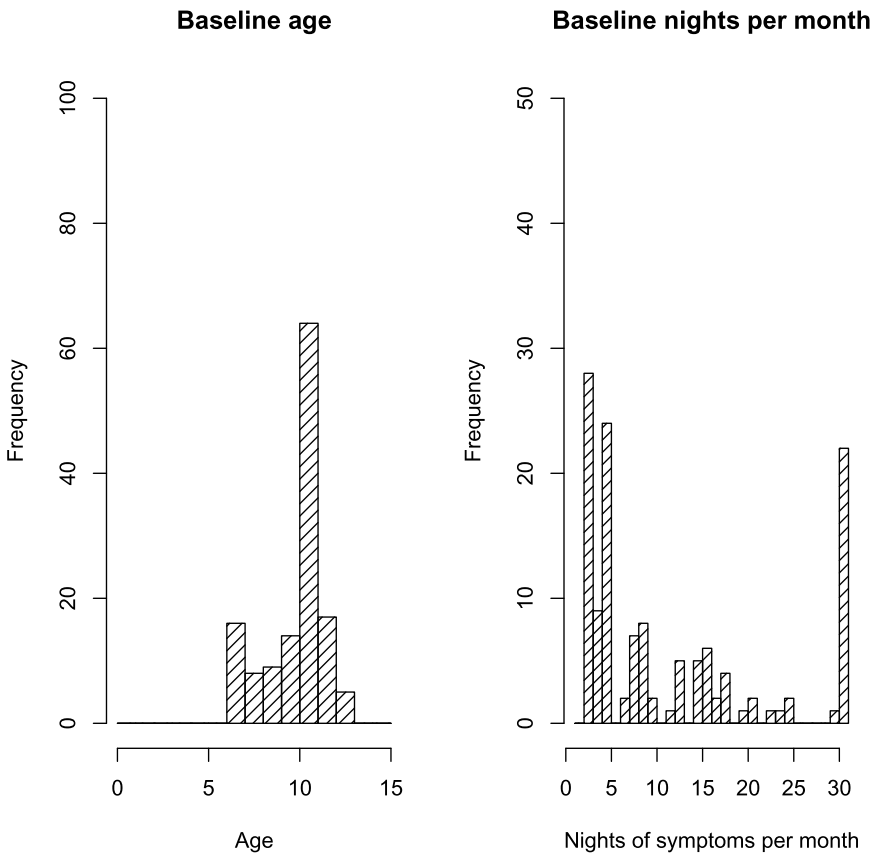


Figure 7: Distributions of baseline covariates.

are generally within 0.4 nights as $\lambda$ varies from 1/2 to $\infty$. In terms of precision, S-PPMA tends to be more conservative than ML but has slightly narrower interval widths than PD.

In practice, one might choose some intermediate value of $\lambda$ (e.g. $\lambda = 1$) since it represents a more conservative assumption about the missingness mechanism. However, lack of sensitivity to $\lambda$ allows for more robustness of estimates to the assumptions about missingness, which is important since any belief regarding $\lambda$ cannot be tested.

# 5 Discussion

Most nonresponse adjustment methods assume MAR, which can be a strong and untestable assumption. An advantage of the PPMA model is it allows us to make inferences about the mean of an outcome variable without assuming MAR. Moreover, the model does not require us to specify a propensity model, since it assumes that missingness depends only on the value of $X + \lambda Y$. The method simplifies nonresponse adjustment by combining a set of auxiliary variables into a single measure $X$ and models departures from MAR using a single sensitivity parameter $\lambda$. In our proposed extension to the PPMA model, we model the relationship between $X$ and $Y$ through a spline. An advantage of this approach is that it does not require $X$ and $Y$ to be bivariate normal, which is assumed in PPMA, since splines allow us to model nonlinearity between the variables. As a result, we do not require the auxiliary variables to be normally distributed, as

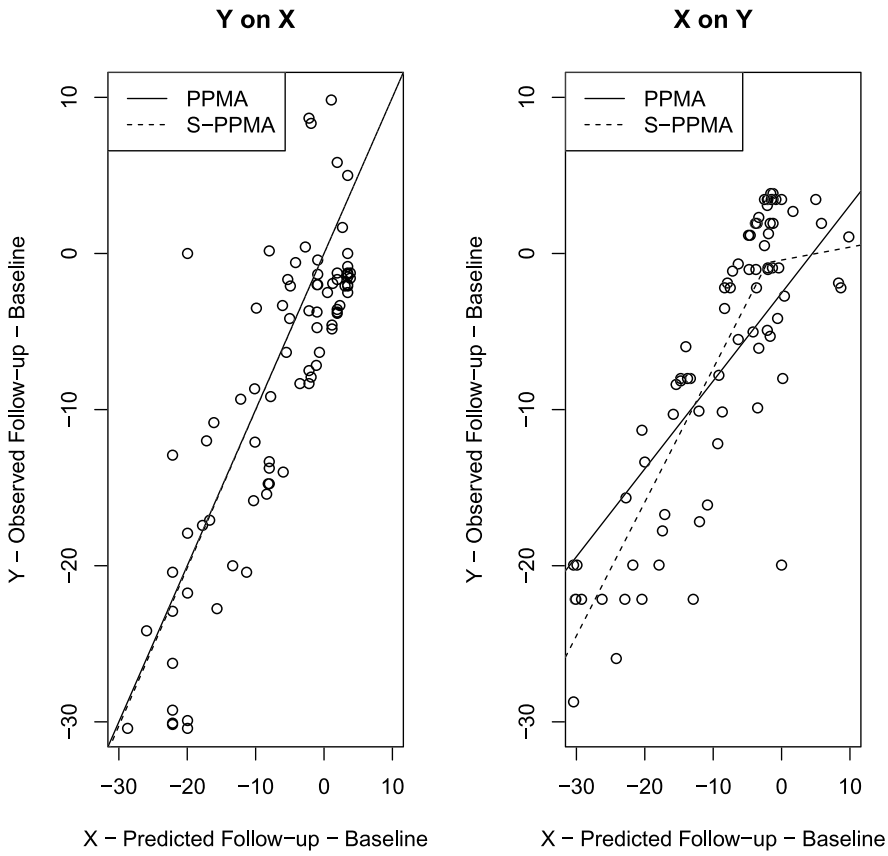


Figure 8: Relationship between $X$ and $Y$.

the model is robust to non-normal distributions of the auxiliary variables. It is important to note, however, that we do not specify a joint distribution between $X$ and $Y$. Thus S-PPMA is approximately viewed as a method.

While S-PPMA utilizes initial values of $Y$ generated from the potentially incorrect PPMA model, the additional steps of spline modelling and hot deck imputation helps to adjust for this nonlinearity. Our simulations show that the proposed S-PPMA model with penalized spline consistently yields approximately unbiased estimates with near nominal non-coverage regardless of the distributions of the auxiliary variables when the correct value of $\lambda$ is used. Compared to the original PPMA proposed in Andridge and Little (2011), S-PPMA has shown to yield estimates that are more robust to covariate distributions, though with a slight penalty in precision when the PPMA model is correct. The gains in bias and RMSE are particularly noticeable the more the auxiliary variables deviate from normality. Results for a smaller sample size of $n = 100$ (see Supplementary Material) show similar trend, where S-PPMA provide some gains in bias and RMSE when covariates are not normal and missingness is not at random, though differences in bias and RMSE tend to be less pronounced than in larger sample sizes. Moreover, the bootstrap variance estimates of PPMA tend to be more conservative than their Bayesian counterpart, leading to better coverages.

It may be tempting to estimate the value of $\lambda$ by specifying a prior distribution. However, any inference about $\lambda$ would be driven entirely by the prior since the data contains no information about $\lambda$. Thus we recommend conducting a sensitivity analysis by applying the S-PPMA

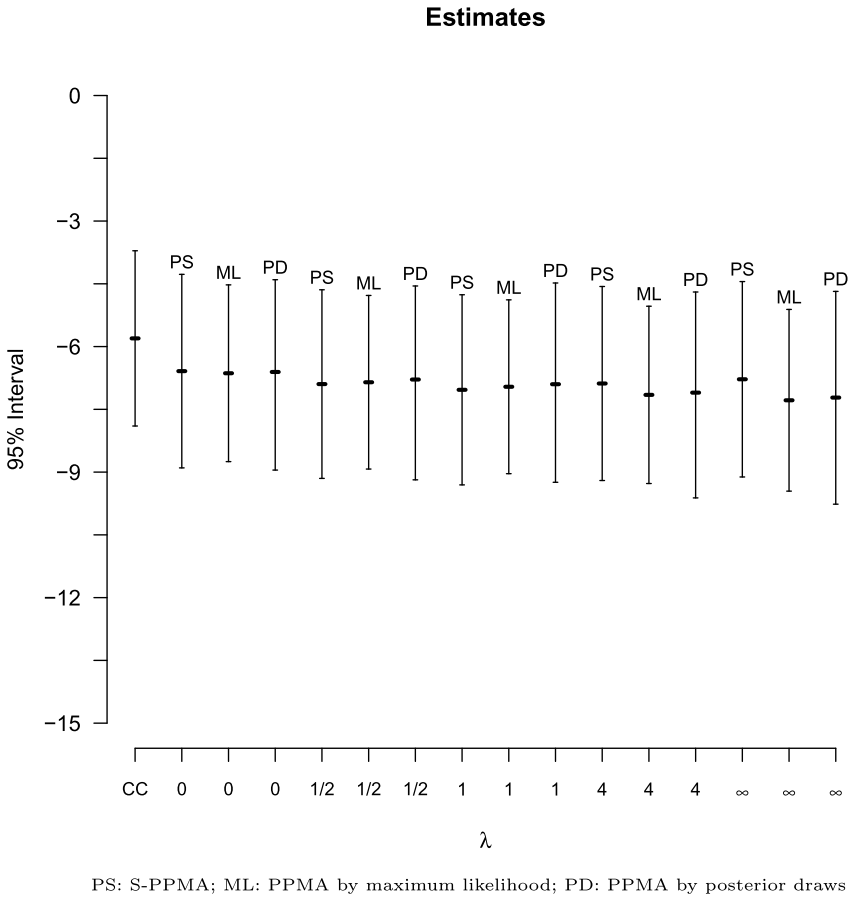


Figure 9: Estimates for mean change in nights of symptoms per month.

model over a range of $\lambda$. The sensitivity analysis reflects our uncertainty about the nonresponse mechanism by displaying estimates of the mean over different values of $\lambda$, ranging from MAR ($\lambda = 0$) to the more extreme MNAR that assumes missingness depends only on the outcome itself ($\lambda = \infty$). Comparing estimates over a range of $\lambda$ helps to provide us an idea of how sensitive our inferences are to the missingness mechanism.

Our examples assume that the variables used to predict the outcome are fully observed, which may not be the case since often both outcome and covariates are missing at the same time, as is the case in unit nonresponse. Extension to the S-PPMA model incorporating additional assumptions about missingness of the covariates may be explored. In our simulations, S-PPMA tends to yield wider confidence intervals than the bivariate normal model particularly for $\lambda > 0$. This may be attributed to the fact that when the data is MNAR, values of the outcome for the nonrespondents may be drastically different than the respondents, leading to extrapolation. Estimation becomes particularly tricky when the relationship between $Y$ and $X$ is nonlinear. Thus, the precision of the penalized spline at high values of $\lambda$ may be a reflection of our uncertainty in extrapolating a nonlinear model.

The S-PPMA and PPMA models assume that missingness depends only on the value of $X + \lambda Y$, where $X$ is a function of the covariates $Z_1, \ldots, Z_p$. In reality, there are infinite ways in which data is missing. For example, missingness of $Y$ may depend only on some subset of

$Z_1, \ldots, Z_p$, which would not be reflected by $X + \lambda Y$ for any $\lambda$. While we may place additional sensitivity parameters on the auxiliary variables, it will reduce simplicity of the model. Finally, we assume that our outcome variable, $Y$, is continuous and limit our inferences to the mean. Extensions to the S-PPMA model are needed to model non-continuous outcome variables.

## Disclaimer

This work was performed while Ye Yang was a doctoral student at the University of Michigan. This article reflects the views of the authors and should not be construed to represent the Food and Drug Administration's views or policies.

## Supplementary Material

Please refer to the Supplementary Material document for:

1. A detailed description of the Gibbs sampling algorithm for the penalized spline prediction.
2. Results from all six simulation scenarios, including estimates from $n = 100$ and $n = 400$ and where $\lambda_A = \lambda_T$ and $\lambda_A \neq \lambda_T$.
3. R code and workspace for the simulations.

## References

Andridge RR, Little RJA (2010). A review of hot deck imputation for survey nonresponse. *International Statistical Review*, 78(1): 40–64.

Andridge RR, Little RJA (2011). Proxy pattern-mixture analysis for survey nonresponse. *Journal of Official Statistics*, 27: 153–180.

Little RJA (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88: 125–134.

Little RJA (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika*, 81: 471–483.

Little RJA, An H (2004). Robust likelihood-based analysis of multivariate data with missing values. *Statistica Sinica*, 14: 949–968.

Little RJA, Rubin DB (2020). *Statistical Analysis with Missing Data*. Wiley, Third Edition.

Pfeffermann D, Sikov A (2011). Imputation and estimation under nonignorable nonresponse in household surveys with missing covariate information. *Journal of Official Statistics*, 27: 181–209.

Rubin DB (1976). Inference and missing data. *Biometrika*, 63: 581–592.

Schouten B (2007). A selection strategy for weighting variables under a not-missing-at-random assumption. *Journal of Official Statistics*, 23: 51–68.

Yang Y, Little RJA (2015). A comparison of doubly robust estimators of the mean with missing data. *Journal of Statistical Computation and Simulation*, 85: 3383–3403.

Zhang G, Little RJA (2009). Extensions of the penalized spline of propensity prediction method of imputation. *Biometrics*, 65(3): 911–918.