Missing observations in a binary covariate

Code and details for ‘Bayesian models for missing and misclassified variables using integrated nested Laplace approximations’

Emma Skarstein, Leonardo Bastos, Håvard Rue and Stefanie Muff

```
# Re-run simulation study or load pre-generated results?
run_study <- TRUE
```

```
library(ggplot2)
library(INLA)
library(inlamisclass)
library(plyr)
```

In this example, we impute missing values in a binary variable using INLA within importance sampling. We assume no misclassification.

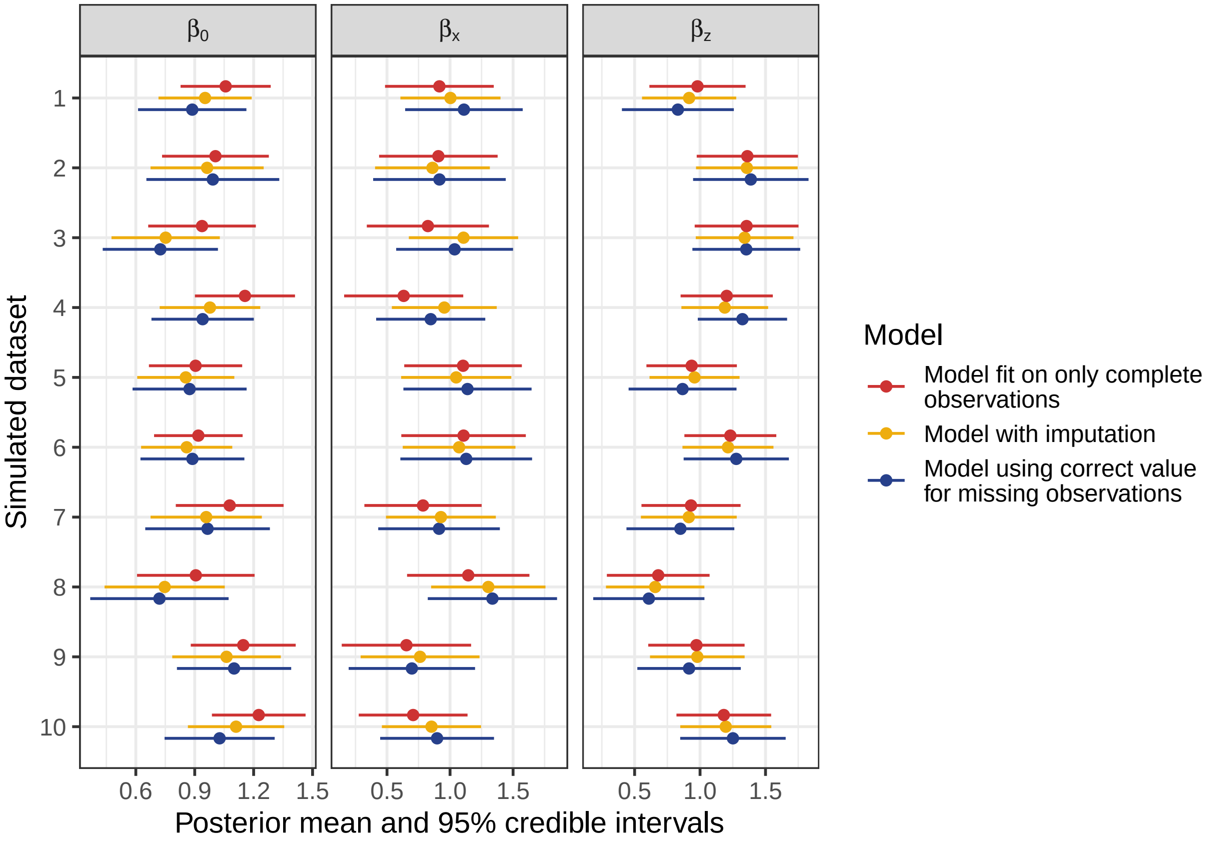
```
n <- 100 # No. of observations
n_runs <- 10 # No. of simulated data sets
niter <- 100000 # No. of iterations for importance sampling

# Suffix giving number of iterations and sample size when saving data and models
name_append <- paste0("n", n, "_", "niter", niter)
```

```
generate_missing <- function(n, n_miss){
  data <- inlamisclass:::generate_data(n, p = 2, betas = c(1, 1, 1),
                                       alphas = c(-0.5, 0.25))
  data$w <- data$x
  data$w[1:n_miss] <- NA
  return(data)
}
```

```
set.seed(1)

all_runs <- list()

for(i in 1:n_runs){
  # Generate data
  data_missing <- generate_missing(n = n, n_miss = 20)

  # Check correct model
  correct_coef <- inla(y ~ x + z, data = data_missing[complete.cases(data_missing),])$summary.fixed
  correct_coef

  # Attenuated version
  naive_coef <- inla(y ~ w + z, data = data_missing)$summary.fixed
  naive_coef

  # Adjusted version
  inla_mod <- inla_is_misclass(formula_moi = y ~ w + z,
                               formula_imp = w ~ z,
                               alpha = c(-0.5, 0.25),
                               data = data_missing,
                               niter = niter,
                               missing_only = TRUE)

  # Extracting relevant stuff
  naive_summ <- data.frame(naive_coef[, c(1,2,3,5)])
  naive_summ$variable <- c("beta.0", "beta.x", "beta.z")

  correct_summ <- data.frame(correct_coef[, c(1,2,3,5)])
  correct_summ$variable <- c("beta.0", "beta.x", "beta.z")

  inla_summ <- make_results_df(inla_mod)$moi
  inla_summ$variable <- c("beta.0", "beta.x", "beta.z")
  colnames(inla_summ) <- c("variable", "mean", "X0.025quant", "X0.975quant")

  all_mods <- dplyr::bind_rows(naive = naive_summ,
                               inla_is = inla_summ,
                               correct = correct_summ,
                               .id = "model")

  all_mods$iteration <- as.factor(i)
  all_runs <- rbind(all_runs, all_mods)
}

saveRDS(all_runs, file = paste0("results/missing_", name_append, ".rds"))
```

```
all_runs <- readRDS(paste0("results/missing_", name_append, ".rds"))
all_runs$Model <- factor(all_runs$model, levels = c("naive", "inla_is", "correct"))
all_runs$Model <- plyr::revalue(all_runs$Model,
                               c("naive" = "Model fit on only complete \nobservations",
                                 "inla_is" = "Model with imputation",
                                 "correct" = "Model using correct value \nfor missing observations"))
all_runs$labels <- paste0(gsub("\\.", "[", all_runs$variable), "]")
colors <- c("brown3", "darkgoldenrod2", "royalblue4")
```

```
ggplot(all_runs, aes(x = mean, y = iteration, color = Model)) +
  geom_point(position = position_dodge2(width = 0.5, reverse = TRUE)) +
  geom_linerange(aes(xmin = X0.025quant, xmax = X0.975quant),
                 position = position_dodge2(width = 0.5, reverse = TRUE)) +
  scale_y_discrete(limits = rev) +
  scale_color_manual(values = colors) +
  facet_grid(cols = vars(labels), scales = "free", labeller = label_parsed) +
  xlab("Posterior mean and 95% credible intervals") +
  ylab("Simulated dataset") +
  theme_bw()
```



```r
ggsave(paste0("figures/missing_simulated_", name_append, ".pdf"), height = 5, width = 8)
```