BvM for Missing Data

Stefan Franssen — *Oxford/ Sorbonne*
Joint work with Judith Rousseau

Coarsening data

For this presentation, we restrict to linear regression with missing data.

Linear regression with coarsened data

Standard linear regression set up:

$$Y|\, X \sim N\left(\beta^T X, \sigma\epsilon\right).$$

Data and coarsening mechanism:

$$X \sim G$$
$$R|\, X, Y \sim \mathbb{P}_R(\cdot|\, X, Y).$$

We only observe R and some measurable transformation $T(X, Y, R)$.

Missing and coarsening at random

The literature generally makes a distinction into 3 different cases:

- Missing/Coarsening completely at random: the probability of coarsening only depends on the coarsening level r,

$$\mathbb{P}_R(R = r|\, X, Y) = p_r.$$

- Missing/Coarsening at random: the probability of coarsening only depends on the observed data r, t,

$$\mathbb{P}_R(R = r|\, X, Y) = f(R, T(X, Y, R)).$$

- Missing/Coarsening not at random: the data is not coarsened at random.

In our project we will assume the data is coarsened at random. There is an extensive literature on these problems already. We refer to [1, 2] for reviews.

Existing methods for handling coarsened Data

The missing data problem has been extensively studied, and many solutions have been proposed. We will give a list of popular methods, but note that this is far from exhaustive.

- List-wise deletion;
- Imputation;
- Inverse probability weighting;
- Expectation Maximization algorithm;
- General likelihood methods;
- Bayesian solutions.

Limitations

Each of the listed methods suffer from at least one of the following:

- Biased estimation;
- Inefficient estimation;
- Complicated methods;
- Sensitivity to model misspecification;
- Lack of theoretical guarantees.

Frequentist Bayes as a solution

Bayesian models are simple to specify and have good practical performance. By putting a nonparametric prior on the distribution of X, we can resolve a large part of the issue of model misspecification. Then a semiparametric Bernstein-von Mises theorem would provide theoretical guarantees of efficiency.

The semiparametric Bernstein-von Mises theorem

The Bernstein-von Mises theorem states that, assuming that the following two conditions hold:

- The *LAN Expansion*,
- The *Change of Measure condition*,

then the posterior distribution of β will be approximately normal, centred on an efficient estimator and with the efficient variance. For details and examples, see [3–5].

Models we aim to study

- Random histograms;
- Gaussian processes;
- (Log) Spline/wavelet bases.

Progress so far

LAN Expansion

Denote r_e the value of R such that $T(X, Y, r_e) = (X, Y)$, a complete case. Assume that the probability $\mathbb{P}_R(R = r_e|\, X = x, Y = y) \geq \delta > 0$ (x, y)-almost surely. Then the LAN-expansion holds.

Change-of-Measure condition

We have a few sketches of proof strategies that can prove this condition, under some assumptions. Assuming, among other things, that the prior on the distribution of X is undersmoothing, then the change of measure condition holds.

Future extensions

Nonparametric estimation

Assuming the probability of a complete case is bounded from below by a positive constant, we would naively expect that we do not lose in the rate of estimation. However, this is not automatic. Arguably, missing data can even lead Bayesian methods to become inconsistent. We aim to give theoretical guarantees that we do not lose in rate.

Robustness

We aim to study the effect of misspecification of the parametric part of the model. Similar with misspecified parametric models, we conjecture that posterior will be asymptotically Gaussian, centred on an efficient estimator, but with possibly a wrong asymptotic variance.

Causal inference

There is a close relation between missing data and causal inference via the potential outcomes framework. In the potential outcomes framework, you will never see a complete case, so we need to substitute a consistency argument for the complete case argument.

References

1. Tsiatis, A. A. *Semiparametric Theory and Missing Data* xvi+383. ISBN: 0-387-32448-8 978-0-387-32448-7 (Springer, New York, 2006).
2. Little, R. J. & Rubin, D. B. *Statistical Analysis with Missing Data* (John Wiley & Sons, 2019).
3. Castillo, I. & Rousseau, J. A Bernstein–von Mises Theorem for Smooth Functionals in Semiparametric Models. *The Annals of Statistics* **43,** 2353–2383. ISSN: 0090-5364 (2015).
4. Franssen, S., Nguyen, J. & van der Vaart, A. *The Bernstein-von Mises Theorem for Semiparametric Mixtures* Nov. 2024. arXiv: 2412.00219 [math].
5. Rivoirard, V. & Rousseau, J. *Bernstein Von Mises Theorem for Linear Functionals of the Density* Comment: 36 pages. Aug. 2009. arXiv: 0908.4167 [math]. (2025).