Naive imputation implicitly regularizes high-dimensional linear models

Ayme Alexis
LPSM, Sorbonne Université, France
alexis.ayme@sorbonne-universite.fr

Two different approaches exist to handle missing values for prediction: either imputation, prior to fitting any predictive algorithms, or dedicated methods able to natively incorporate missing values. While imputation is widely (and easily) used, it is unfortunately biased when low-capacity predictors (such as linear models) are applied afterward. However, in practice, naive imputation exhibits good predictive performance. In this paper, we study the impact of imputation in a high-dimensional linear model with MCAR missing data. We prove that zero imputation performs an implicit regularization closely related to the ridge method, often used in high-dimensional problems. Leveraging on this connection, we establish that the imputation bias is controlled by a ridge bias, which vanishes in high dimension. As a predictor, we argue in favor of the averaged SGD strategy, applied to zero-imputed data. We establish an upper bound on its generalization error, highlighting that imputation is benign in the $d \gg \sqrt{n}$ regime. Experiments illustrate our findings.

Joint work with Claire Boyer (LPSM, Sorbonne Université), Aymeric Dieuleveut (CMAP, Ecole Polytechnique), Erwan Scornet (CMAP, Ecole Polytechnique).