Online Mixed Value Imputation with Gaussian Copula

Eric Landgrebe, Yuxuan Zhao, Madeleine Udell

Cornell University

July 17, 2020

Problem Statement and Model

- Problem Statement
  - Want to impute missing entries
  - Matrix $X \in \mathbb{R}^{n \times d}$ with missing values
  - iid rows, with columns representing features
  - Continuous, ordinal and binary entries
  - Online setting: must impute as rows come in (also useful for minibatch fitting)
- Gaussian Copula
  - Latent vector $z \sim N(0, \Sigma)$
  - Scaled by elementwise monotonic function to produce observations $x = f(z)$
  - f is uniquely determined by the observed marginals, and for ordinals it is a threshold function
  - Given f, Σ we can impute latent variables using the conditional mean and scale them through f

Algorithm

- Want to estimate Σ to maximize observed log likelihood ℓ_{obs}
  - Hard to do exactly, because the observed log likelihood takes the form of an integral of a Gaussian density over all of $\mathbb{R}$ for missing dimensions, and an integral over interval for observed ordinal dimensions.
  - In the offline setting, do this iteratively using an approximate Expectation Maximization (EM) algorithm [1] updating

$$\Sigma_{t+1} = \underset{\Sigma}{\operatorname{argmax}}\, \mathbb{E}_{\Sigma_t}[\ell_{obs}(X;\Sigma)] = \sum_{i=1}^{n} \frac{1}{n} \mathbb{E}[z^i (z^i)^\top | X^i, \Sigma_t]$$

    - The maximizer is an "empirical covariance" over latent vectors weighted by liklihood under the previous estimate
  - In the online setting with batch indices S_t and $\gamma_t \in (0, 1]$ we use an online EM Algorithm [2] to update Σ as

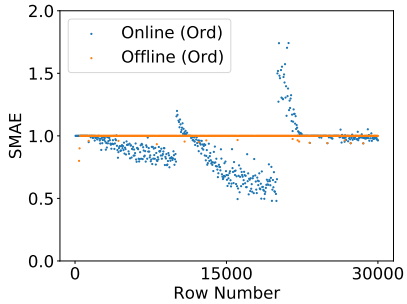
$$\Sigma_{t+1} = \gamma_t \left(\underset{\Sigma}{\operatorname{argmax}}\, \mathbb{E}_{\Sigma_t}[\ell_{obs}(X_{S_t};\Sigma)] \right) + (1-\gamma_t)\Sigma_t$$

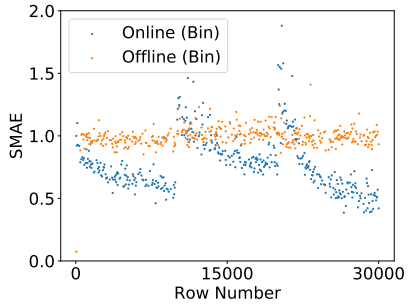
Results

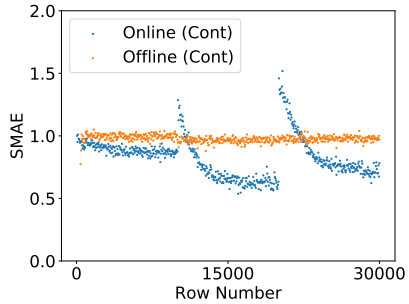
- Minibatch method is faster with comparable error on real data ($> 75\%$ missing) [3]

Method	Runtime (s)	MAE	RMSE
Standard EM unthreaded	2411.071	0.582	0.882
Standard EM threaded	1033.083	0.582	0.882
Minibatch EM unthreaded	446.805	0.585	0.887
Minibatch EM unthreaded (longer timeout)	893.929	0.583	0.884
Online EM threaded	249.551	0.598	0.898

- Adapts to changing data distribution online (SMAE of 1 is median imputation, lower is better)







References

Yuxuan Zhao and Madeleine Udell.

Missing value imputation for mixed data through gaussian copula, 2019.

Olivier Cappé and Eric Moulines.

On-line expectation–maximization algorithm for latent data models.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 71(3):593–613, 2009.

F Maxwell Harper and Joseph A Konstan.

The movielens datasets: History and context.
Acm transactions on interactive intelligent systems (tiis), 5(4):1–19, 2015.