Probabilistic Missing Value Imputation for Mixed Categorical and Ordered Data

Yuxuan Zhao, Alex Townsend, Madeleine Udell

Goal: impute missing entries in mixed data

age	gender	state	income	education	⋯
29	F	CT	$53,000	college	⋯
57	?	CA	?	high school	⋯
?	M	?	$102,000	masters	⋯
41	F	NV	$23,000	?	⋯
⋮	⋮	⋮	⋮		

Goal: impute missing entries in mixed data

age	gender	state	income	education	⋯
29	F	CT	\$53,000	college	⋯
57	?	CA	?	high school	⋯
?	M	?	\$102,000	masters	⋯
41	F	NV	\$23,000	?	⋯
⋮	⋮	⋮	⋮		

Crude continuous approximation:

- Ordinal: integer encoding such as 1-5 rating
- Categorical: one-hot encoding (vector with 0 or 1)

Imputed values cannot guarantee to satisfy the integer/one-hot restrictions

Our imputation approach

Estimate the conditional distribution of missing entries given observation

Our imputation approach

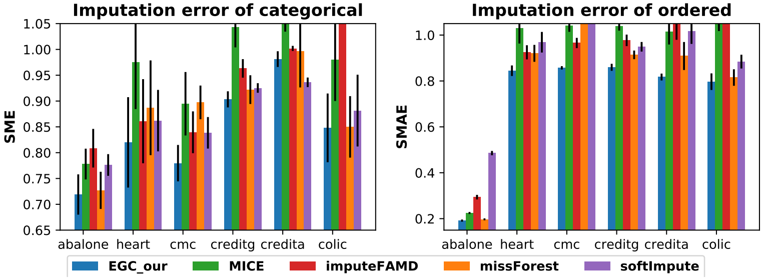
Estimate the conditional distribution of missing entries given observation

Our proposition: the extended Gaussian copula

A distribution model for continuous, ordinal and categorical mixed data

- A transformation of latent Gaussian vector
  - Categorical is modeled via the argmax transformation
- No assumption on the marginal distribution

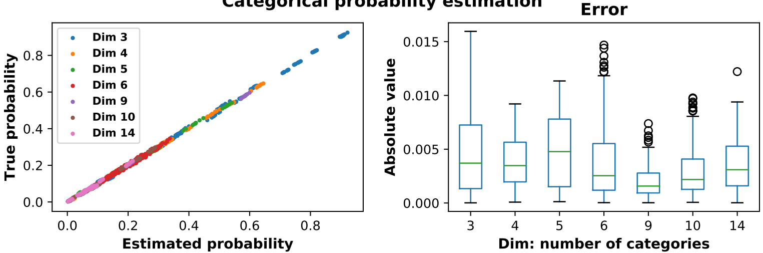
Results



Imputation error of categorical

Imputation error of ordered

Categorical probability estimation



Error