

Imputing Missing Data with the Gaussian Copula

Madeleine Udell

Operations Research and Information Engineering
Cornell University

Based on joint work with Yuxuan Zhao (Cornell)

SIAM MDS, June 2020

Data table

age	gender	state	income	education	...
29	F	CT	\$53,000	college	...
57	?	NY	\$19,000	high school	...
?	M	CA	\$102,000	masters	...
41	F	NV	\$23,000	?	...
⋮	⋮	⋮	⋮		

Data table

age	gender	state	income	education	...
29	F	CT	\$53,000	college	...
57	?	NY	\$19,000	high school	...
?	M	CA	\$102,000	masters	...
41	F	NV	\$23,000	?	...
⋮	⋮	⋮	⋮		

goals:

- ▶ impute missing entries?
- ▶ identify related features?

Data table

n examples (patients, respondents, assets)

p features (tests, questions, performance indicators)

$$\begin{bmatrix} X \end{bmatrix} = \begin{bmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{np} \end{bmatrix}$$

- ▶ x^i , i th row of X , is feature vector for i th example
- ▶ x_j , j th column of X , gives values for j th feature across all examples

Classic approach: low rank matrix completion

- ▶ efficient algorithms
- ▶ theoretical guarantees
- ▶ works well when
 - ▶ true rank r is small compared to $\min(n,p)$
 - ▶ or when true matrix is approximately low rank
- ▶ works poorly for ordinal or mixed data
 - ▶ requires additional parameters (e.g., a loss function) to map from low rank parameter to ordinal observations
 - ▶ fit with cross-validation \implies slow!

skinny data matrices are not approximately low rank

skinny data matrices

- ▶ (-) are unlikely to be (very) low rank
- ▶ (+) have lots of observations per column, so can estimate more complex model per column

Outline

Gaussian copula model

Approximate EM

Performance

Gaussian copula model

We say $x \sim \text{GC}(\Sigma, f) \in \mathbf{R}^p$ follows the *Gaussian copula model* with parameters Σ and $f : \mathbf{R}^p \rightarrow \mathbf{R}^p$ if

- ▶ *copula*: $z \sim \mathcal{N}(0, \Sigma)$
- ▶ *marginals*: $x = f(z)$ for $f = (f_1, \dots, f_n)$ entrywise monotonic, i.e.,

$$x_j = f_j(z_j), \quad j = 1, \dots, p$$

Gaussian copula model

We say $x \sim \text{GC}(\Sigma, f) \in \mathbf{R}^p$ follows the *Gaussian copula model* with parameters Σ and $f : \mathbf{R}^p \rightarrow \mathbf{R}^p$ if

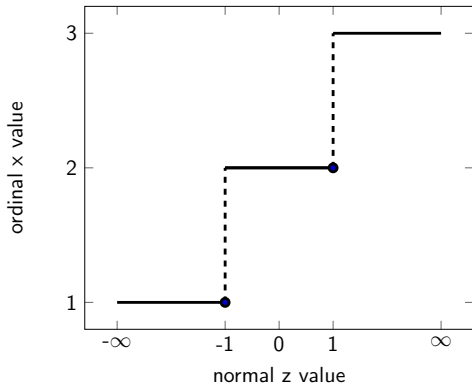
- ▶ *copula*: $z \sim \mathcal{N}(0, \Sigma)$
- ▶ *marginals*: $x = f(z)$ for $f = (f_1, \dots, f_n)$ entrywise monotonic, i.e.,

$$x_j = f_j(z_j), \quad j = 1, \dots, p$$

why use a copula?

- ▶ models *nonlinear rulers*
- ▶ separates interactions from marginals
- ▶ Sklar's theorem: any multivariate joint distribution can be written in terms of univariate marginal distribution functions and a copula which describes the dependence structure

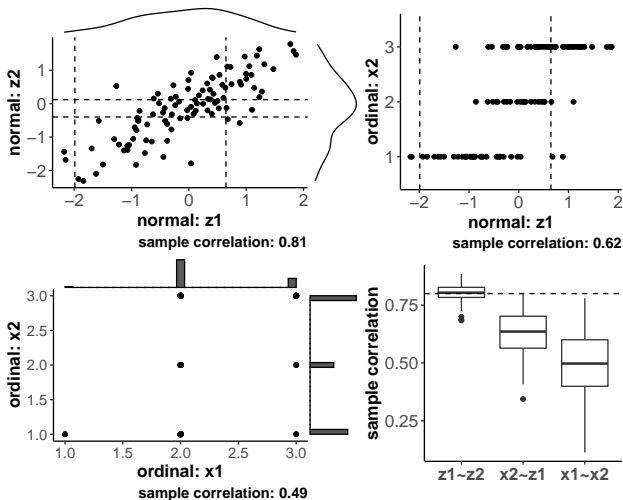
Example: cutoff function f generates ordinal data



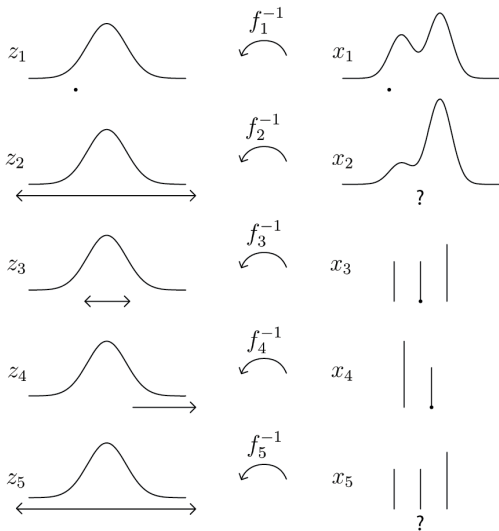
- ▶ when x has a continuous distribution, f is invertible
- ▶ when x has a discrete distribution, f is not invertible
- ▶ we write f^{-1} for the set-valued mapping

$$z \in f^{-1}(x) \iff x = f(z)$$

Marginals distort correlations



Missing data with the Gaussian copula



Related work

- ▶ Hoff et al. 2007: Gaussian copula for ordinal data, fit with MCMC
- ▶ Han and Liu 2014: Gaussian copula for real valued data
- ▶ Ganti et al. 2015: low rank model with nonlinear marginals, not probabilistic
- ▶ Anderson-Bergman et al. 2018: motivated by copula model; method fits low rank mean but assumes diagonal covariance

Related work

- ▶ Hoff et al. 2007: Gaussian copula for ordinal data, fit with MCMC
- ▶ Han and Liu 2014: Gaussian copula for real valued data
- ▶ Ganti et al. 2015: low rank model with nonlinear marginals, not probabilistic
- ▶ Anderson-Bergman et al. 2018: motivated by copula model; method fits low rank mean but assumes diagonal covariance

our method:

- ▶ handles real, boolean, ordinal, and missing data
- ▶ fully probabilistic (\implies confidence intervals for estimates)
- ▶ no hyperparameters (rank, etc)
- ▶ fit with fast approximate EM

Outline

Gaussian copula model

Approximate EM

Performance

Estimating the model

Gaussian copula model

- ▶ $z \sim \mathcal{N}(0, \Sigma)$
- ▶ $x = f(z)$, $f = (f_1, \dots, f_n)$ entrywise monotonic

two parameters to estimate

- ▶ *marginals*: monotone functions f_1, \dots, f_p
- ▶ *copula*: correlation matrix $\Sigma \in \mathbf{R}^{p \times p}$

Given parameters, imputation is easy

- ▶ invertible marginals $f = (f_1, \dots, f_p)$ (or estimates)
- ▶ copula Σ
- ▶ observed entries x_{obs} of new row $x \in \mathbf{R}^p$, $\text{obs} \subset \{1, \dots, p\}$
- ▶ missing entries $\text{mis} = \{1, \dots, p\} \setminus \text{obs}$

impute missing entries using normality of z_{mis} :

Given parameters, imputation is easy

- ▶ invertible marginals $f = (f_1, \dots, f_p)$ (or estimates)
- ▶ copula Σ
- ▶ observed entries x_{obs} of new row $x \in \mathbf{R}^p$, $\text{obs} \subset \{1, \dots, p\}$
- ▶ missing entries $\text{mis} = \{1, \dots, p\} \setminus \text{obs}$

impute missing entries using normality of z_{mis} :

- ▶ map observations to latent $z_{\text{obs}} = \{f_j^{-1}(x_j) : j \in \text{obs}\}$

Given parameters, imputation is easy

- ▶ invertible marginals $f = (f_1, \dots, f_p)$ (or estimates)
- ▶ copula Σ
- ▶ observed entries x_{obs} of new row $x \in \mathbf{R}^p$, $\text{obs} \subset \{1, \dots, p\}$
- ▶ missing entries $\text{mis} = \{1, \dots, p\} \setminus \text{obs}$

impute missing entries using normality of z_{mis} :

- ▶ map observations to latent $z_{\text{obs}} = \{f_j^{-1}(x_j) : j \in \text{obs}\}$
- ▶ latent missing z_{mis} are normal given z_{obs} :

$$z_{\text{mis}} | z_{\text{obs}} \sim \mathcal{N}(\Sigma_{\text{mis,obs}} \Sigma_{\text{obs,obs}}^{-1} z_{\text{obs}}, \Sigma_{\text{mis,mis}} - \Sigma_{\text{mis,obs}} \Sigma_{\text{obs,obs}}^{-1} \Sigma_{\text{obs,mis}})$$

Given parameters, imputation is easy

- ▶ invertible marginals $f = (f_1, \dots, f_p)$ (or estimates)
- ▶ copula Σ
- ▶ observed entries x_{obs} of new row $x \in \mathbf{R}^p$, $\text{obs} \subset \{1, \dots, p\}$
- ▶ missing entries $\text{mis} = \{1, \dots, p\} \setminus \text{obs}$

impute missing entries using normality of z_{mis} :

- ▶ map observations to latent $z_{\text{obs}} = \{f_j^{-1}(x_j) : j \in \text{obs}\}$
- ▶ latent missing z_{mis} are normal given z_{obs} :

$$z_{\text{mis}} | z_{\text{obs}} \sim \mathcal{N}(\Sigma_{\text{mis,obs}} \Sigma_{\text{obs,obs}}^{-1} z_{\text{obs}}, \Sigma_{\text{mis,mis}} - \Sigma_{\text{mis,obs}} \Sigma_{\text{obs,obs}}^{-1} \Sigma_{\text{obs,mis}})$$

- ▶ predict with mean $\hat{z}_{\text{mis}} = \mathbb{E}[z_{\text{mis}} | z_{\text{obs}}] = \Sigma_{\text{mis,obs}} \Sigma_{\text{obs,obs}}^{-1} z_{\text{obs}}$

Given parameters, imputation is easy

- ▶ invertible marginals $f = (f_1, \dots, f_p)$ (or estimates)
- ▶ copula Σ
- ▶ observed entries x_{obs} of new row $x \in \mathbf{R}^p$, $\text{obs} \subset \{1, \dots, p\}$
- ▶ missing entries $\text{mis} = \{1, \dots, p\} \setminus \text{obs}$

impute missing entries using normality of z_{mis} :

- ▶ map observations to latent $z_{\text{obs}} = \{f_j^{-1}(x_j) : j \in \text{obs}\}$
- ▶ latent missing z_{mis} are normal given z_{obs} :

$$z_{\text{mis}} | z_{\text{obs}} \sim \mathcal{N}(\Sigma_{\text{mis,obs}} \Sigma_{\text{obs,obs}}^{-1} z_{\text{obs}}, \Sigma_{\text{mis,mis}} - \Sigma_{\text{mis,obs}} \Sigma_{\text{obs,obs}}^{-1} \Sigma_{\text{obs,mis}})$$

- ▶ predict with mean $\hat{z}_{\text{mis}} = \mathbb{E}[z_{\text{mis}} | z_{\text{obs}}] = \Sigma_{\text{mis,obs}} \Sigma_{\text{obs,obs}}^{-1} z_{\text{obs}}$
- ▶ map back to observed space using marginals $\hat{x}_{\text{mis}} = f(\hat{z}_{\text{mis}})$

Given parameters, imputation is easy

- ▶ invertible marginals $f = (f_1, \dots, f_p)$ (or estimates)
- ▶ copula Σ
- ▶ observed entries x_{obs} of new row $x \in \mathbf{R}^p$, $\text{obs} \subset \{1, \dots, p\}$
- ▶ missing entries $\text{mis} = \{1, \dots, p\} \setminus \text{obs}$

impute missing entries using normality of z_{mis} :

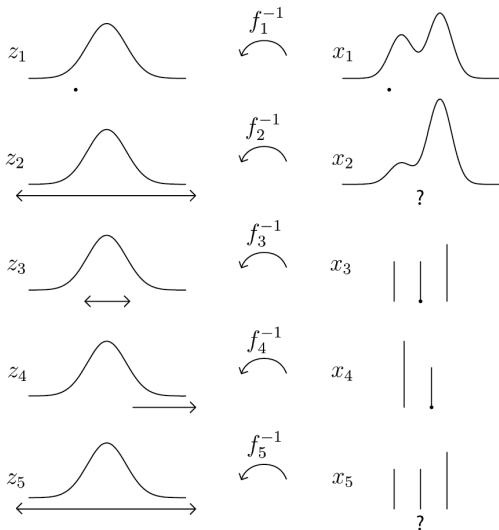
- ▶ map observations to latent $z_{\text{obs}} = \{f_j^{-1}(x_j) : j \in \text{obs}\}$
- ▶ latent missing z_{mis} are normal given z_{obs} :

$$z_{\text{mis}} | z_{\text{obs}} \sim \mathcal{N}(\Sigma_{\text{mis,obs}} \Sigma_{\text{obs,obs}}^{-1} z_{\text{obs}}, \Sigma_{\text{mis,mis}} - \Sigma_{\text{mis,obs}} \Sigma_{\text{obs,obs}}^{-1} \Sigma_{\text{obs,mis}})$$

- ▶ predict with mean $\hat{z}_{\text{mis}} = \mathbb{E}[z_{\text{mis}} | z_{\text{obs}}] = \Sigma_{\text{mis,obs}} \Sigma_{\text{obs,obs}}^{-1} z_{\text{obs}}$
- ▶ map back to observed space using marginals $\hat{x}_{\text{mis}} = f(\hat{z}_{\text{mis}})$

can also compute confidence intervals!

Missing or mixed data: latent vector z not known



latent z is constrained to product of intervals

Given parameters, imputation is easy

- ▶ marginals $f = (f_1, \dots, f_p)$ (or estimates) (**not invertible**)
- ▶ copula Σ
- ▶ observed entries x_{obs} of new row $x \in \mathbf{R}^p$
- ▶ missing entries $\text{mis} = \{1, \dots, p\} \setminus \text{obs}$

impute missing entries using normality of z_{mis} :

- ▶ latent missing z_{mis} are normal given z_{obs} :

$$z_{\text{mis}} | z_{\text{obs}} \sim \mathcal{N}(\Sigma_{\text{mis,obs}} \Sigma_{\text{obs,obs}}^{-1} z_{\text{obs}}, \Sigma_{\text{mis,mis}} - \Sigma_{\text{mis,obs}} \Sigma_{\text{obs,obs}}^{-1} \Sigma_{\text{obs,mis}})$$

- ▶ predict with mean:

$$\begin{aligned} \hat{z}_{\text{mis}} &= \mathbb{E}[z_{\text{mis}} | x_{\text{obs}}] = \mathbb{E}[\Sigma_{\text{mis,obs}} \Sigma_{\text{obs,obs}}^{-1} z_{\text{obs}} | x_{\text{obs}}] \\ &= \Sigma_{\text{mis,obs}} \Sigma_{\text{obs,obs}}^{-1} \mathbb{E}[z_{\text{obs}} | x_{\text{obs}}] \end{aligned}$$

- ▶ map back to observed space using marginals $\hat{x}_{\text{mis}} = f(\hat{z}_{\text{mis}})$

Estimating the marginals

- ▶ Gaussian matrix Z :
 - ▶ columns $\mathcal{N}(0, 1)$
- ▶ (partially) observed matrix X :
 - ▶ column j has observed empirical distribution with CDF F_j

Estimating the marginals

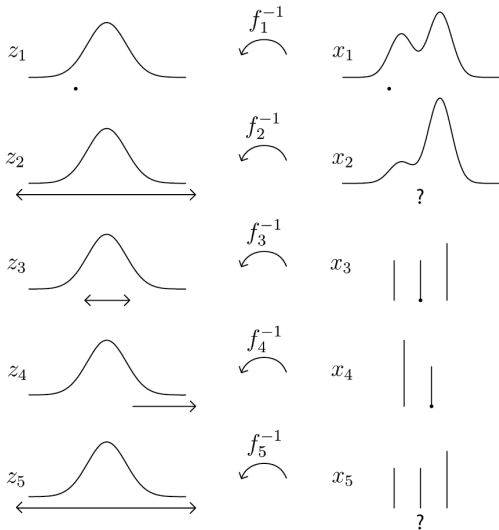
- ▶ Gaussian matrix Z :
 - ▶ columns $\mathcal{N}(0, 1)$
- ▶ (partially) observed matrix X :
 - ▶ column j has observed empirical distribution with CDF F_j

estimate f_j for each column $j = 1, \dots, p$ by matching quantiles:

$$\hat{f}_j^{-1}(x_j^i) = \Phi^{-1} \left(\frac{n_j}{n_j + 1} \hat{F}_j(x_j^i) \right).$$

our convention: for discrete data, \hat{F}_j is set-valued.

Matching quantiles



Warmup: estimate copula for complete cts data

given

- ▶ observed matrix $X \in \mathbf{R}^{n \times p}$
- ▶ invertible marginals $f = (f_1, \dots, f_p)$ (or estimates)

estimate copula Σ using maximum likelihood estimation:

Warmup: estimate copula for complete cts data

given

- ▶ observed matrix $X \in \mathbf{R}^{n \times p}$
- ▶ invertible marginals $f = (f_1, \dots, f_p)$ (or estimates)

estimate copula Σ using maximum likelihood estimation:

- ▶ compute latent $z^i = f^{-1}(x^i)$ for each row x^i

Warmup: estimate copula for complete cts data

given

- ▶ observed matrix $X \in \mathbf{R}^{n \times p}$
- ▶ invertible marginals $f = (f_1, \dots, f_p)$ (or estimates)

estimate copula Σ using maximum likelihood estimation:

- ▶ compute latent $z^i = f^{-1}(x^i)$ for each row x^i
- ▶ log likelihood is

$$\ell(\Sigma; X) = c - \frac{\log(|\Sigma|)}{2} - \frac{1}{2} \text{Tr}(\Sigma^{-1} \sum_{i=1}^n \frac{1}{n} z^i (z^i)^\top),$$

maximized by the empirical covariance $G = \frac{1}{n} \sum_{i=1}^n z^i (z^i)^\top$

Warmup: estimate copula for complete cts data

given

- ▶ observed matrix $X \in \mathbf{R}^{n \times p}$
- ▶ invertible marginals $f = (f_1, \dots, f_p)$ (or estimates)

estimate copula Σ using maximum likelihood estimation:

- ▶ compute latent $z^i = f^{-1}(x^i)$ for each row x^i
- ▶ log likelihood is

$$\ell(\Sigma; X) = c - \frac{\log(|\Sigma|)}{2} - \frac{1}{2} \text{Tr}(\Sigma^{-1} \sum_{i=1}^n \frac{1}{n} z^i (z^i)^\top),$$

maximized by the empirical covariance $G = \frac{1}{n} \sum_{i=1}^n z^i (z^i)^\top$

- ▶ scale $\hat{\Sigma}$ to correlation matrix:
set $D = \mathbf{diag}(G)$ and scale $\hat{\Sigma} \leftarrow D^{-1/2} G D^{-1/2}$

Estimating the copula: missing or discrete

given

- ▶ observed entries x_{obs}^i for each row $x^i \in \mathbf{R}^p$, $i = 1, \dots, n$
- ▶ marginals $f = (f_1, \dots, f_p)$ (or estimates)

idea: estimate copula Σ using maximum likelihood estimation?
log likelihood is

$$\ell(\Sigma; \{x_{\mathcal{O}_i}\}_{i=1}^n) = \frac{1}{n} \sum_{i=1}^n \log \left(\int_{z^i \in f^{-1}(x^i)} \phi(z^i; 0, \Sigma) dz^i \right),$$

where $\phi(z; 0, \Sigma)$ is the pdf of a Gaussian with mean 0 and correlation Σ .

Estimating the copula: missing or discrete

given

- ▶ observed entries x_{obs}^i for each row $x^i \in \mathbf{R}^p$, $i = 1, \dots, n$
- ▶ marginals $f = (f_1, \dots, f_p)$ (or estimates)

idea: estimate copula Σ using maximum likelihood estimation?
log likelihood is

$$\ell(\Sigma; \{x_{\mathcal{O}_i}\}_{i=1}^n) = \frac{1}{n} \sum_{i=1}^n \log \left(\int_{z^i \in f^{-1}(x^i)} \phi(z^i; 0, \Sigma) dz^i \right),$$

where $\phi(z; 0, \Sigma)$ is the pdf of a Gaussian with mean 0 and correlation Σ .

- ▶ $z^i \in f^{-1}(x^i)$ is product of intervals
- ▶ truncated Gaussian integral is hard to compute or optimize

Estimating the copula: missing or discrete

given

- ▶ observed entries x_{obs}^i for each row $x^i \in \mathbf{R}^p$, $i = 1, \dots, n$
- ▶ marginals $f = (f_1, \dots, f_p)$ (or estimates)

idea: estimate copula Σ using maximum likelihood estimation?
log likelihood is

$$\ell(\Sigma; \{x_{\mathcal{O}_i}\}_{i=1}^n) = \frac{1}{n} \sum_{i=1}^n \log \left(\int_{z^i \in f^{-1}(x^i)} \phi(z^i; 0, \Sigma) dz^i \right),$$

where $\phi(z; 0, \Sigma)$ is the pdf of a Gaussian with mean 0 and correlation Σ .

- ▶ $z^i \in f^{-1}(x^i)$ is product of intervals
- ▶ truncated Gaussian integral is hard to compute or optimize
- ▶ instead: use EM

EM algorithm for Gaussian Copula

EM algorithm for Gaussian Copula:

- ▶ **Input:** observed entries X_{obs} , marginals f
 - ▶ **Initialize:** $t = 0$, $\hat{\Sigma} = I$.
 - ▶ For $t = 0, 1, 2, \dots$
 1. E-step: $G = \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[zz^T | x_{\text{obs}}^i, \hat{\Sigma}, f \right]$.
 2. M-step: $D = \mathbf{diag}(G)$; scale $\hat{\Sigma} \leftarrow D^{-1/2} G D^{-1/2}$.
- until convergence.
- ▶ **Output:** $\hat{\Sigma}$.

EM algorithm for Gaussian Copula

EM algorithm for Gaussian Copula:

- ▶ **Input:** observed entries X_{obs} , marginals f
- ▶ **Initialize:** $t = 0$, $\hat{\Sigma} = I$.
- ▶ For $t = 0, 1, 2, \dots$
 1. E-step: $G = \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[zz^T | x_{\text{obs}}^i, \hat{\Sigma}, f \right]$.
 2. M-step: $D = \text{diag}(G)$; scale $\hat{\Sigma} \leftarrow D^{-1/2} G D^{-1/2}$.until convergence.
- ▶ **Output:** $\hat{\Sigma}$.

properties:

- ▶ every step increases likelihood $\mathbb{P}(X_{\text{obs}}; \hat{\Sigma})$
- ▶ consistent estimate for Σ if data MAR
- ▶ E step is challenging; our approximate EM method uses a fast approximation based on 1D integrals.

Outline

Gaussian copula model

Approximate EM

Performance

Copula-EM imputes well and quickly

Synthetic data experiments:

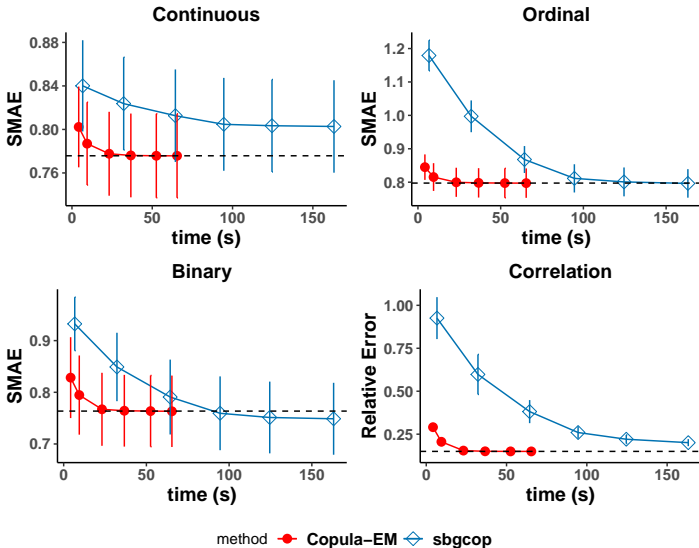
- ▶ *Complex*: $p = 15$, $n = 2000$. Correlation Σ is random. First five columns have exponential distributions, next five are binary, and last five are ordinal with five levels.

Define scaled mean absolute error

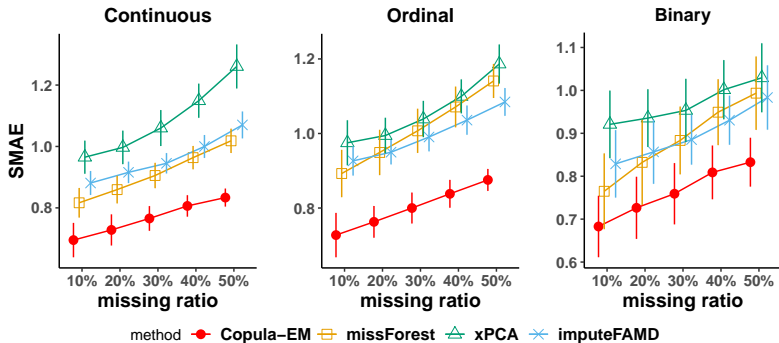
$$\text{SMAE}(\hat{x}, x, \text{obs}) := \frac{\|\hat{x} - x\|_1}{\|\text{median}(x) - x\|_1}$$

(Median imputation has $\text{SMAE} = 1$.)

Copula-EM imputes faster than MCMC



Gaussian copula has best accuracy



Gaussian copula for survey data

- ▶ General Social Survey data
- ▶ 1 continuous, 17 ordinal variables with 2–48 levels
- ▶ 25% of data missing
- ▶ mask additional 10% of data

Table: Imputation Error (SMAE) on Five GSS Variables

Variable	Copula-EM	sbgcop	missForest	xPCA	imputeFAMD
CLASS	0.736(0.11)	1.098(0.15)	0.782(0.09)	0.795(0.08)	0.797(0.10)
LIFE	0.758(0.13)	0.994(0.19)	0.828(0.17)	0.783(0.11)	0.821(0.11)
HEALTH	0.886(0.08)	1.257(0.17)	1.143(0.18)	0.908(0.10)	0.947(0.04)
HAPPY	0.894(0.08)	1.332(0.13)	1.079(0.15)	1.003(0.15)	1.001(0.10)
INCOME	0.897(0.05)	0.952(0.03)	0.944(0.18)	1.090(0.15)	0.996(0.01)

Gaussian copula yields insights

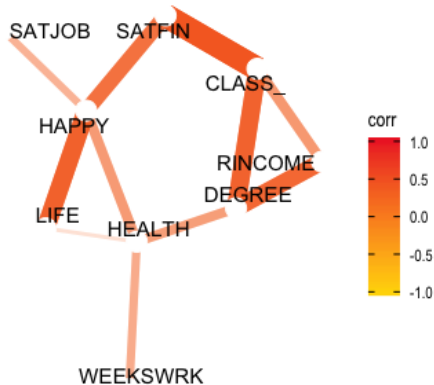


Figure: Large correlations ($|\cdot| > 0.3$) between a few interesting variables from GSS data.

Gaussian copula model outperforms on MovieLens

MovieLens data:

- ▶ 207 movies with ≥ 1000 ratings
- ▶ all users who rate at least one of those 207 movies
- ▶ 75% missing rate
- ▶ mask additional 10% of data

Table: Imputation Error on 207 Movies

Algorithm	MAE	RMSE
Column Median	0.702(0.004)	1.001(0.004)
Copula-EM	0.579(0.004)	0.880(0.005)
GLRM (BvSLoss)	0.595(0.004)	0.892(0.004)
softImpute	0.602(0.004)	0.883(0.004)
xPCA	0.613(0.004)	0.897(0.004)
imputeFAMD	0.646(0.005)	0.991(0.005)
missForest	0.669(0.004)	1.015(0.006)

Gaussian copula model outperforms on MovieLens

MovieLens data:

- ▶ 207 movies with ≥ 1000 ratings
- ▶ all users who rate at least one of those 207 movies
- ▶ 75% missing rate
- ▶ mask additional 10% of data

Table: Imputation Error on 207 Movies

Algorithm	MAE	RMSE
Column Median	0.702(0.004)	1.001(0.004)
Copula-EM	0.579(0.004)	0.880(0.005)
GLRM (BvSLoss)	0.595(0.004)	0.892(0.004)
softImpute	0.602(0.004)	0.883(0.004)
xPCA	0.613(0.004)	0.897(0.004)
imputeFAMD	0.646(0.005)	0.991(0.005)
missForest	0.669(0.004)	1.015(0.006)

selected rank using 5CV is 99 for softImpute, 6 for xPCA, and 8 for GLRM.

Conclusion

Gaussian copula model for data imputation:

- ▶ no hyperparameters
- ▶ outperforms fancier nonparametric methods
- ▶ yields confidence intervals for imputations
- ▶ fast and easy to estimate for moderate p or (with additional low rank assumption) large p

references:

- ▶ *Missing Value Imputation for Mixed Data Through Gaussian Copula*. Y. Zhao and M. Udell, KDD 2020.
- ▶ *Matrix Completion with Quantified Uncertainty through Low Rank Gaussian Copula*. Y. Zhao and M. Udell, <https://arxiv.org/abs/2006.10829> 2020.

Outline

*

References

- Anderson-Bergman, C., Kolda, T. G., and Kincher-Winoto, K. (2018). XPCA: Extending PCA for a combination of discrete and continuous variables. *arXiv preprint arXiv:1808.07510*.
- Ganti, R. S., Balzano, L., and Willett, R. (2015). Matrix completion under monotonic single index models. In *Advances in Neural Information Processing Systems*, pages 1873–1881.
- Han, F. and Liu, H. (2014). High dimensional semiparametric scale-invariant principal component analysis. *IEEE transactions on pattern analysis and machine intelligence*, 36(10):2016–2032.
- Hoff, P. D. et al. (2007). Extending the rank likelihood for semiparametric copula estimation. *The Annals of Applied Statistics*, 1(1):265–283.
- Tsukahara, H. (2005). Semiparametric estimation in copula models. *Canadian Journal of Statistics*, 33(3):357–375.