Advances in Clustering of Incomplete Data with Skewed and Heavy-tailed Clusters

SKEW, 2026- Andriette Bekker

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

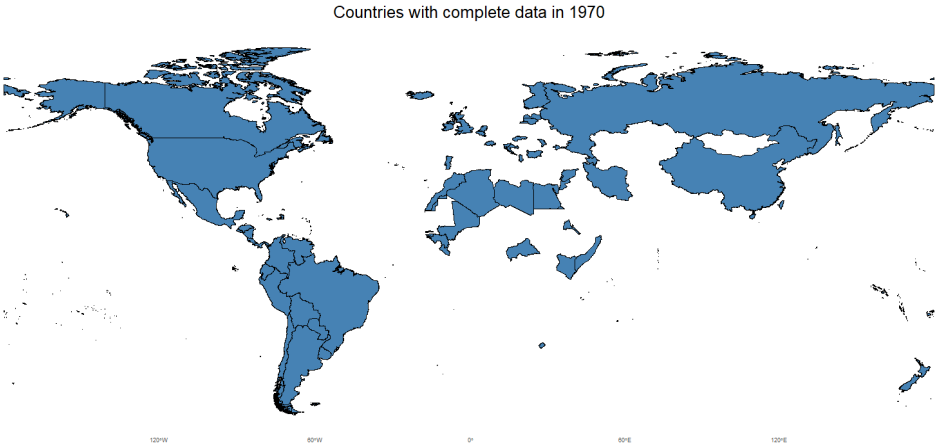
Meet the team!

Session: Snapshot

- Setting the scene
- Proposal
- Simulation studies
- Data applications (CO_2 Emissions dataset & Sleep and Health study dataset)

The impact of missing data

Consider EDGARv8.1 database, CO_2 emissions profiles per country:

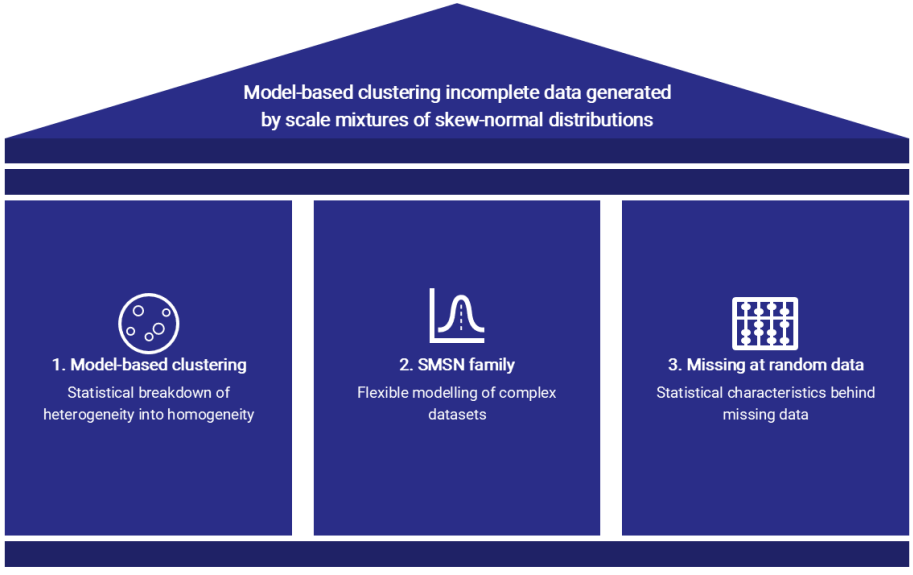


The impact of missing data: Problems

Data is relied upon for 'scientific' and more objective decision-making, but:

- Inconsistent data coverage. Many countries (especially in the global south) lack complete, current emissions inventories.
- Decision-making risk. Actions based only on data-rich countries bias global climate policy.
- Disproportionate ripple. Countries in Africa and other low-income regions are typically most affected by decisions based on incomplete datasets.

Our proposal



Overview: Model-Based Clustering

Basis:

- Continuous random vector $\boldsymbol{X}$.
- G number of components.
- Some probability density function (pdf) f with a collection of parameters $\boldsymbol{\theta}_g$.
- A mixing probability $\pi_g \in (0, 1]$ so that $\sum_{g=1}^{G} \pi_g = 1$.

The Finite Mixture Model (FMM):

$$p(\boldsymbol{x}) = \sum_{g=1}^{G} \pi_g f(\boldsymbol{x}; \boldsymbol{\theta}_g).$$

We focus on how to carefully choose f. The popular choice for f is the pdf of a normal distribution.

Flexibility in clustering

Drawbacks of the normal distribution:

- Assumes the data's distribution is symmetric.
- The marginal distributions have thin tails, i.e. sensitive to outliers.

How to develop a mixture model that circumvents these? Possible solution we explore:
Skewed distributions. In particular, the skew-normal distribution.

Beyond symmetry: The skew-normal distribution

The pdf of a multivariate skew-normal distribution $SN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \lambda_0)$:

$$f_{SN}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}) = \frac{1}{\Phi_1\left(\frac{\lambda_0}{\sqrt{1+\boldsymbol{\lambda}^\top \boldsymbol{\lambda}}}\right)} \phi_p(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \Phi_1\left(\boldsymbol{\lambda}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu}) + \lambda_0\right)^{1}$$

- Φ_1 is the Cumulative Distribution Function of a univariate standard normal distribution.
- The extra parameter $\boldsymbol{\lambda}$ controls asymmetry.
- λ_0 also affects skewness and is typically set to 0.

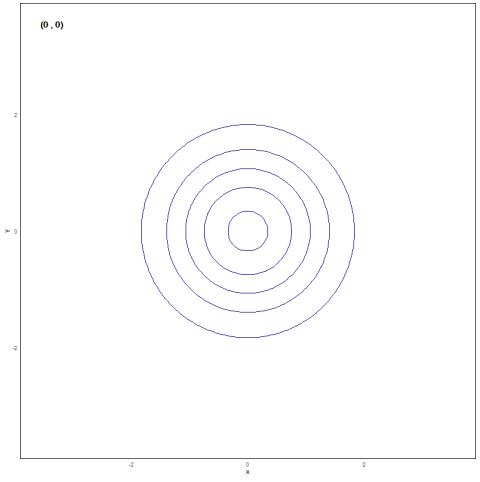
The pdf of a multivariate skew-normal distribution $SN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$:

$$f_{SN}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}) = 2\phi_p(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \Phi_1\left(\boldsymbol{\lambda}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu})\right)^{2}.$$

[1] Arnold B. C., Beaver R. J. (2002)

[2] Lachos, V., et al. (2010)

Visualisation



$SN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}) \to N_p(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ as $\boldsymbol{\lambda} \to \mathbf{0}$.

Scale mixtures of the skew normal (SMSN)

Skewness? Covered! ✓.

How to address heavier tailed data?

The scale mixture distribution is the consequence of a random component that is superimposed on the scale matrix of the reference distribution [3].
Discrete scaling variable:

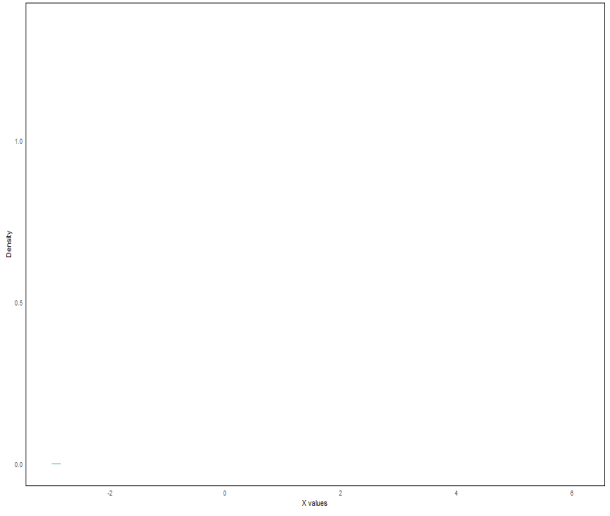
$$f_{SMSN}(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{u=0}^{\infty} f_{SN}(x;\boldsymbol{\mu}, k(u)\boldsymbol{\Sigma}, \boldsymbol{\lambda}) h(u;\boldsymbol{\psi})$$

Continuous scaling variable:

$$f_{SMSN}(\boldsymbol{x};\boldsymbol{\theta}) = \int_{0}^{\infty} f_{SN}(x;\boldsymbol{\mu}, k(u)\boldsymbol{\Sigma}, \boldsymbol{\lambda}) h(u;\boldsymbol{\psi}) du.$$

[3] Branco, M. D., & Dey, D. K. (2001)

Visualisation



Visualisation

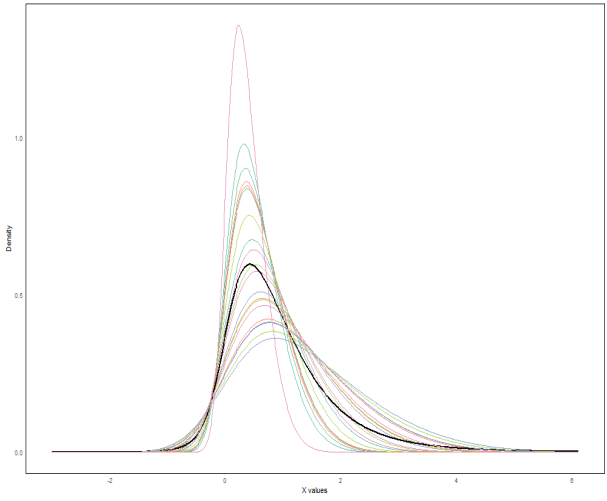


Figure: $f_{SMSN}(\boldsymbol{x};\boldsymbol{\theta}) = \mathbb{E}_U[f_{SN}(x;\boldsymbol{\mu}, k(u)\boldsymbol{\Sigma}, \boldsymbol{\lambda})]$

Characteristics of the SMSN family

The stochastic representation

$$\boldsymbol{X} \stackrel{d}{=} \boldsymbol{\mu} + (k(U)\boldsymbol{\Sigma})^{1/2}\left[\frac{\boldsymbol{\lambda}}{\sqrt{1+\boldsymbol{\lambda}^\top\boldsymbol{\lambda}}}T + \left(\boldsymbol{I}_p - \frac{\boldsymbol{\lambda}\boldsymbol{\lambda}^\top}{1+\boldsymbol{\lambda}^\top\boldsymbol{\lambda}}\right)^{1/2}\boldsymbol{V}\right]$$ [4]

with:

- T Truncated standard normal univariate RV on $[0, \infty)$,
- $\boldsymbol{V} \sim N(\boldsymbol{0}, \boldsymbol{I}_p)$,
- T and $\boldsymbol{V}$ are independent.

[4] Azzalini A., Valle, D. (1996)

Characteristics of the SMSN family

Conditionals of the SMSN family:

- $\boldsymbol{X}|u \sim SN_p(\boldsymbol{\mu}, k(u)\boldsymbol{\Sigma}, \boldsymbol{\lambda})$,
- $\boldsymbol{X}|t,u \sim N_p(\boldsymbol{\mu} + \boldsymbol{\Delta}t, k(u)\boldsymbol{\Omega})$,

where $\boldsymbol{\Omega} = \boldsymbol{\Sigma} - \boldsymbol{\Delta}\boldsymbol{\Delta}^\top$, $\quad \boldsymbol{\Delta} = \boldsymbol{\Sigma}^{1/2}\frac{\boldsymbol{\lambda}}{\sqrt{1+\boldsymbol{\lambda}^\top\boldsymbol{\lambda}}}$.

Some examples of the SMSN family: Skew-t

- $k(u) = \frac{1}{u}$
- $u \sim Gam(\frac{\nu}{2}, \frac{\nu}{2})$,

so that

- $\boldsymbol{X} \sim ST_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu)$,

where ν denotes the degrees of freedom.

Contaminated skew-normal

- $k(u) = u + \beta(1-u)$
- $u \sim Bernoulli(\alpha)$,

so that $\beta > 1$

$$f_{CSN}(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{u=0}^{1} f_{SN}(\boldsymbol{x};\boldsymbol{\mu},(u+\beta(1-u))\boldsymbol{\Sigma},\boldsymbol{\lambda})h(u;\alpha)^{5}$$

$$f_{CSN}(\boldsymbol{x};\boldsymbol{\theta}) = \underbrace{\alpha f_{SN}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\lambda})}_{\text{good component}} + \underbrace{(1-\alpha) f_{SN}(\boldsymbol{x};\boldsymbol{\mu},\beta\boldsymbol{\Sigma},\boldsymbol{\lambda})}_{\text{bad component}}$$

Here, $\alpha > 0.5$ is the proportion of typical observations, and β is the degree of contamination.

[5] Tukey J.W. (1960)

Examples of the SMSN family

Distribution	Denoted as	$k(u)$	Distribution of U
Skew-normal	SN_p	u	Degenerate at $u = 1$
Skew-t	ST_p	u^{-1}	$U \sim Gam(\frac{\nu}{2}, \frac{\nu}{2})$
Skew-slash	SS_p	u^{-1}	$U \sim Beta(\alpha, 1)$
Contaminated skew-normal	CSN_p	$u + \beta(1-u)$	$U \sim Bernoulli\,(\alpha)$
Skew-Laplace	SL_p	u	$U \sim Exp\,(1)$
Skew-variance-gamma	SVG_p	u	$U \sim Gam\left(\eta, \frac{\gamma^2}{2}\right)$
Skew-hyperbolic	SH_p	u	$U \sim GIG\,(\eta, \gamma, \psi)$

Missing At Random (MAR)

Partition an observation $\boldsymbol{X}_i$ into:

$$\boldsymbol{X}_i = \begin{bmatrix} \boldsymbol{X}_i^o \\ \boldsymbol{X}_i^m \end{bmatrix},$$

where $\boldsymbol{X}_i^o$ and $\boldsymbol{X}_i^m$ are the observed and missing parts of $\boldsymbol{X}_i$ respectively. Strictly speaking, the missing and observed patterns are potentially different from observation to observation, requiring the notation o_i and m_i, but for notational convenience, we drop the subscript and stick to o and m.

Properties of the SMSN family

Make the partitions:

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_o \\ \boldsymbol{\mu}_m \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{oo} & \boldsymbol{\Sigma}_{om} \\ \boldsymbol{\Sigma}_{mo} & \boldsymbol{\Sigma}_{mm} \end{bmatrix}, \quad \boldsymbol{\Delta} = \begin{bmatrix} \boldsymbol{\Delta}_o \\ \boldsymbol{\Delta}_m \end{bmatrix}, \quad \boldsymbol{\Omega} = \begin{bmatrix} \boldsymbol{\Omega}_{oo} & \boldsymbol{\Omega}_{om} \\ \boldsymbol{\Omega}_{mo} & \boldsymbol{\Omega}_{mm} \end{bmatrix},$$

p_i^o and p_i^m are the dimensions of $\boldsymbol{X}_i^o$ and $\boldsymbol{X}_i^m$, respectively ($p_i^o + p_i^m = p_i$).

Marginal distribution of the SMSN family:

$\boldsymbol{X}_i^o | u_i \sim SN_{p_i^o}(\boldsymbol{\mu}_o, k(u_i)\boldsymbol{\Sigma}_{oo}, \dot{\boldsymbol{\lambda}}_o)$, with $\dot{\boldsymbol{\lambda}}_o = \frac{\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Delta}_o}{\sqrt{1-\boldsymbol{\Delta}_o^\top \boldsymbol{\Sigma}_{oo}^{-1}\boldsymbol{\Delta}_o}}$

Properties of the SMSN family

Conditional distributions of the SMSN family:

$$\boldsymbol{X}_i^m|\boldsymbol{x}_i^o, u_i \sim SN_{p_i^m}(\boldsymbol{\mu}^*, k(u_i)\boldsymbol{\Sigma}^*, \boldsymbol{\lambda}^*, k(u_i)^{1/2}\lambda_0^*),\ \boldsymbol{X}_i^m|\boldsymbol{x}_i^o, t_i, u_i \sim N_{p_i^m}(\boldsymbol{m}^* + t_i\boldsymbol{\psi}^*, k(u_i)\boldsymbol{\Omega}^*)$$

with

- $\boldsymbol{\mu}^* = \boldsymbol{\mu}_m + \boldsymbol{\Sigma}_{mo}\boldsymbol{\Sigma}_{oo}^{-1}(\boldsymbol{x}_i^o - \boldsymbol{\mu}_o)$
- $\boldsymbol{\Sigma}^* = \boldsymbol{\Sigma}_{mm} - \boldsymbol{\Sigma}_{mo}\boldsymbol{\Sigma}_{oo}^{-1}\boldsymbol{\Sigma}_{om}$
- $\lambda_0^* = \frac{\boldsymbol{\Delta}_o^\top \boldsymbol{\Sigma}_{oo}^{-1}(\boldsymbol{x}_i^o - \boldsymbol{\mu}_o)}{\sqrt{1-\boldsymbol{\Delta}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Delta}}}$
- $\boldsymbol{\lambda}^* = \frac{\boldsymbol{\Sigma}^{-1/2}\left[\boldsymbol{\Delta}_m - \boldsymbol{\Sigma}_{mo}\boldsymbol{\Sigma}_{oo}^{-1}\boldsymbol{\Delta}_o\right]}{\sqrt{1-\boldsymbol{\Delta}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Delta}}}$

with

- $\boldsymbol{m}^* = \boldsymbol{\mu}_m + \boldsymbol{\Omega}_{mo}\boldsymbol{\Omega}_{oo}^{-1}(\boldsymbol{x}_i^o - \boldsymbol{\mu}_o)$
- $\boldsymbol{\psi}^* = \boldsymbol{\Delta}_m - \boldsymbol{\Omega}_{mo}\boldsymbol{\Omega}_{oo}^{-1}\boldsymbol{\Delta}_o$
- $\boldsymbol{\Omega}^* = \boldsymbol{\Omega}_{mm} - \boldsymbol{\Omega}_{mo}\boldsymbol{\Omega}_{oo}^{-1}\boldsymbol{\Omega}_{om}$

Properties of the SMSN family

For the special case of the CSN distribution, the conditional distributions can be compartmentalised as:

The good component ($u_i = 1$):

- $\boldsymbol{X}_i^m | \boldsymbol{x}_i^o, u_i \sim SN_{p_i^m}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*, \boldsymbol{\lambda}^*, \lambda_0^*)$
- $\boldsymbol{X}_i^m | \boldsymbol{x}_i^o, u_i \sim N_{p_i^m}(\boldsymbol{m}^* + t_i \boldsymbol{\psi}^*, \boldsymbol{\Omega}^*)$

The bad component ($u_i = 0$):

- $\boldsymbol{X}_i^m | \boldsymbol{x}_i^o, u_i \sim SN_{p_i^m}(\boldsymbol{\mu}^*, \beta_g \boldsymbol{\Sigma}^*, \boldsymbol{\lambda}^*, \beta_g^{1/2} \lambda_0^*)$
- $\boldsymbol{X}_i^m | \boldsymbol{x}_i^o, t_i, u_i \sim N_{p_i^m}(\boldsymbol{m}^* + \beta_g^{1/2} t_i \boldsymbol{\psi}^*, \beta_g \boldsymbol{\Omega}^*)$

The EM algorithm

The Expectation-Maximisation algorithm fits a model to incomplete data. It consists of two steps.

1. The E-step. Computes the expected loglikelihood function using the most recent parameter updates.
2. The M-step. Updates the parameters using the expected values from the E-step.

The algorithm iterates between the E and M steps until the algorithm has converged.
Note that the EM algorithm relies on initial values and a criterion to monitor convergence status. There are numerous techniques that can fulfill these needs.

Model based clustering using the SMSN family

Introduce the membership random variable Z_{ig}:

$$Z_{ig} = \begin{cases} 1 & \text{if } \boldsymbol{X}_i \text{ belongs to the } g^{th} \text{ cluster} \\ 0 & \text{if } \boldsymbol{X}_i \text{ does not belong to the } g^{th} \text{ cluster} \end{cases}$$

Cluster memberships are not observed, and are treated as missing values, but help produce a more attractive 'complete' data loglikelihood:

$$\begin{aligned} \mathcal{L} &= \prod_{i=1}^{n} \prod_{g=1}^{G} \left[\pi_g f_{SMSN}(\boldsymbol{x}_i; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \boldsymbol{\lambda}_g, \boldsymbol{\theta}_g) \right]^{z_{ig}} \\ &= \prod_{i=1}^{n} \prod_{g=1}^{G} \left[\pi_g f_N(\boldsymbol{x}_i; \boldsymbol{\mu}_g + \boldsymbol{\Delta}_g t_i, k(u_i)\boldsymbol{\Omega}_g) f_{TN}(t_i; 0, k(u_i)) h(u_i; \boldsymbol{\theta}_g) \right]^{z_{ig}} \end{aligned}$$

Model based clustering using the SMSN family

The complete likelihood $\mathcal{L}$ depends on four sources of missing data values:

1. $\boldsymbol{Z}_g = \{Z_{ig}\}_{i=1}^n$, which depends on $\boldsymbol{X}^o$.
2. $\boldsymbol{U} = \{U_i\}_{i=1}^n$, which depends on $\boldsymbol{X}^o$.
3. $\boldsymbol{T} = \{t_i\}_{i=1}^n$,which depends on $\boldsymbol{X}^o$ and $\boldsymbol{U}$.
4. $\boldsymbol{X}^m = \{\boldsymbol{X}_i^m\}_{i=1}^n$, which depends on $\boldsymbol{X}^o$, $\boldsymbol{U}$, and $\boldsymbol{T}$.

Model based clustering using the SMSN family

The complete likelihood $\mathcal{L}$ depends on four sources of missing data values:

1. $\boldsymbol{Z}_g = \{Z_{ig}\}_{i=1}^n$, which depends on $\boldsymbol{X}^o$.
2. $\boldsymbol{U} = \{U_i\}_{i=1}^n$, which depends on $\boldsymbol{X}^o$.
3. $\boldsymbol{T} = \{t_i\}_{i=1}^n$,which depends on $\boldsymbol{X}^o$ and $\boldsymbol{U}$.
4. $\boldsymbol{X}^m = \{\boldsymbol{X}_i^m\}_{i=1}^n$, which depends on $\boldsymbol{X}^o$, $\boldsymbol{U}$, and $\boldsymbol{T}$.

$*$ For the CSN distribution, U_i is given added interpretation:

$$U_i = \begin{cases} 1 & \text{if } \boldsymbol{X}_i \text{ is a typical observation,} \\ 0 & \text{if } \boldsymbol{X}_i \text{ is atypical/outlier.} \end{cases}$$

Model based clustering using the SMSN family

The complete likelihood $\mathcal{L}$ depends on four sources of missing data values:

1. $\boldsymbol{Z}_g = \{Z_{ig}\}_{i=1}^n$, which depends on $\boldsymbol{X}^o$.
2. $\boldsymbol{U} = \{U_i\}_{i=1}^n$, which depends on $\boldsymbol{X}^o$.
3. $\boldsymbol{T} = \{t_i\}_{i=1}^n$,which depends on $\boldsymbol{X}^o$ and $\boldsymbol{U}$.
4. $\boldsymbol{X}^m = \{\boldsymbol{X}_i^m\}_{i=1}^n$, which depends on $\boldsymbol{X}^o$, $\boldsymbol{U}$, and $\boldsymbol{T}$.

$*$ For the CSN distribution, U_i is given added interpretation:

$$U_i = \begin{cases} 1 & \text{if } \boldsymbol{X}_i \text{ is a typical observation,} \\ 0 & \text{if } \boldsymbol{X}_i \text{ is atypical/outlier.} \end{cases}$$

The E-step computes the expected values of missing values 1.-4. and their products, producing an expected likelihood function $Q = \mathbb{E}[\mathcal{L}|\boldsymbol{X}^o]$.
The M-step then maximises (the natural logarithm of) Q with respect to the model's parameters.

Simulation experiments

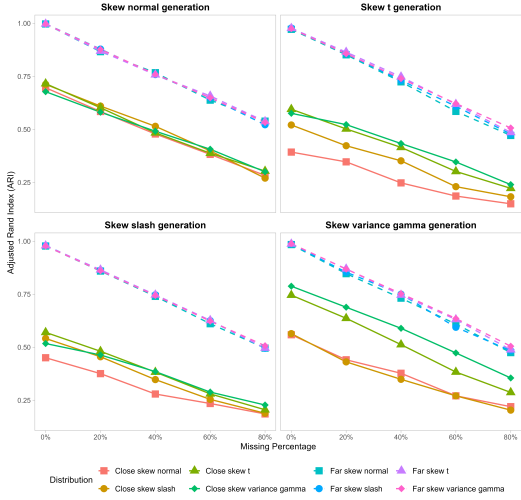
Purpose- See how observations are clustered when the appropriate model is used compared to ill-fitting models.
Design - Clustering performance is assessed while controlling for confounding effects, such as:

- Two levels of cluster overlap - Far and close.
- Two sample sizes - small and large.
- Extent of missingness - Ranging from none to 80%.

Measure - The Adjusted Rand Index (ARI) is used.

Simulation experiments: SMSN

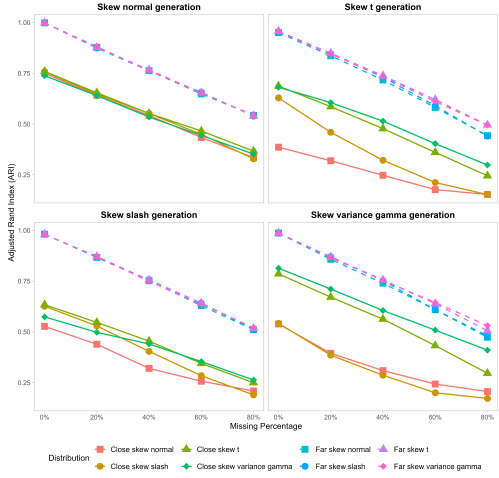
Figure: Average ARI values across 200 replications for datasets of size $n = 200$, randomly generated from a two-component mixture.



- When data comes from the skew normal distribution, the cluster performances are comparable.
- Thereafter, there is a hierarchy in performance, with the best performer being the skew variance gamma.

Simulation experiments: SMSN

Figure: Average ARI values across 200 replications for datasets of size $n = 500$, randomly generated from a two-component mixture.



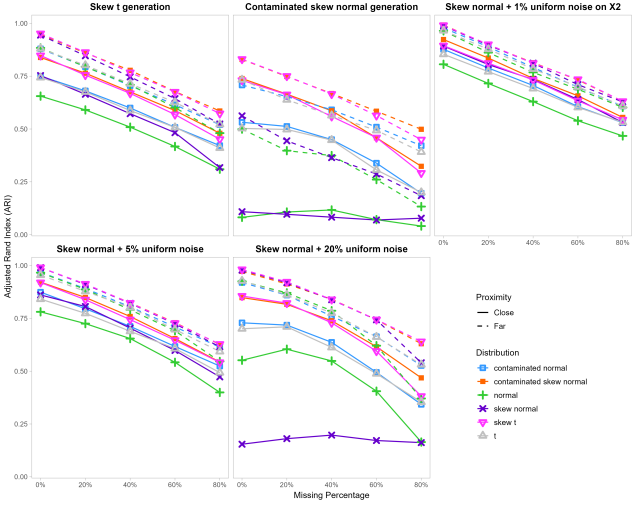
Simulation experiments: CSN

Since the FMCMSN model has the capabilities to handle atypical points, we consider data generation from a FMSN contaminated by noise:

- 1% of the observations randomly replaced by $(0, x_{i2}^*)$, where $x_{i2}^* \sim U(10, 15)$.
- 5% of points randomly replaced by noise observed from a $U(0, 10)$ distribution.
- 20% of points randomly replaced by noise observed from a $U(0, 10)$ distribution.

Simulation experiments: CSN

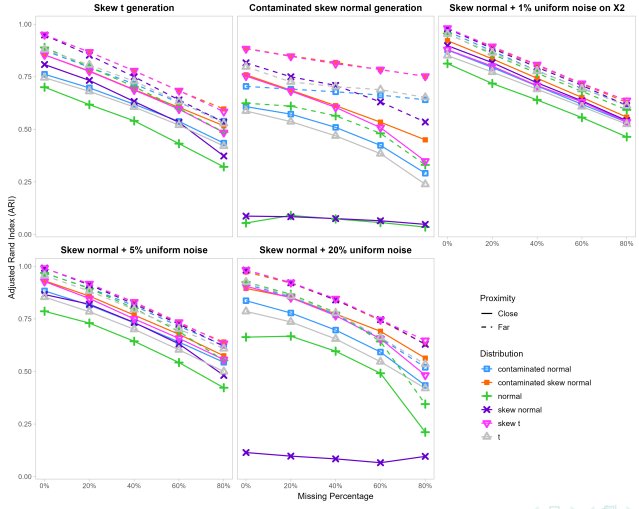
Figure: Average ARI values across 100 replications for datasets of size $n = 300$, randomly generated from a two-component mixture.



- ARI now compares cluster performance across the FMCMSN and its competitors.
- The results: The FMCMSN model performs the best overall.

Simulation experiments: CSN

Figure: Average ARI values across 100 replications for datasets of size $n = 800$, randomly generated from a two-component mixture.



Simulation experiments: Outlier detection

Outlier detection performance of the FMCMSN model is compared with:

- Contaminated Normal (CN)
- t

Measures:

- Accuracy - Number of points correctly classified.
- False positive rate - FPR.
- True positive rate - TPR.

Simulation experiments: Outlier detection

Figure: Average FPR values for $n = 300$.

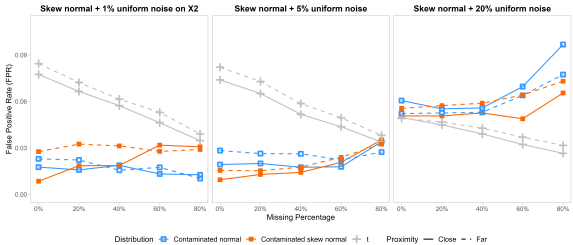
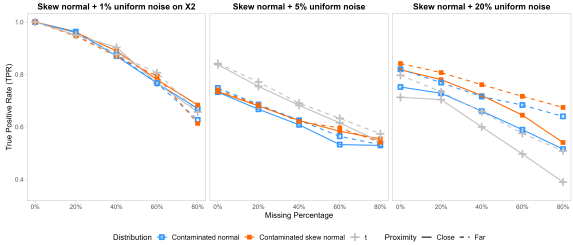


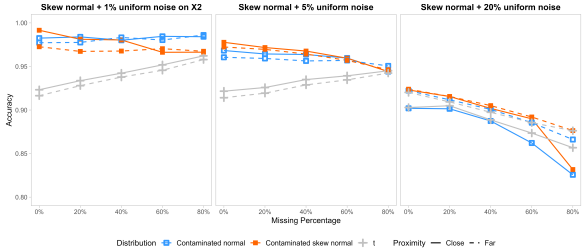
Figure: Average TPR values for $n = 300$.



- The TPRs are comparable.
- The FPRs: as missingness increases, the FMCMSN does not outperform its symmetric counterpart for 1% of outliers.
- For 20% of outliers, the models are initially comparable when there are no missing values in the FPRs, but the FMCMSN performs best in terms of the TPRs.

Simulation experiments: Outlier detection

Figure: Proportion of accurately classified points for $n = 300$ across 100 random replications.



- For the large sample case (n=800), the trends across the three measures are consistent with the small sample case (n=300).
- For outlier detection, the contaminated models rely on imputed values to make a decision, in contrast to the threshold approach of the Student t. At 80%, it is more likely to impute a missing outlier as a typical point.

Simulation experiments: Outlier detection

Outlier detection performance under high dimensionality. Eliminate confounders - one cluster, p=10.

Illustrative example: global climate emissions

Emissions Database for Global Atmospheric Research (EDGAR).
CO_2 emissions for their most recent year (2022).
Sectors: (X_1) Main activity electricity and heat production, (X_2) Manufacturing industries and construction, (X_3) Road transportation no resuspension, (X_4) Residential and other sectors, (X_5) Oil and natural gas, (X_6) Lime production, and the (X_7) Metal industry.

Illustrative example: global climate emissions

Table: Proportion of missing values by sector in the EDGARv8.1 dataset.

Variable	Name	Missing (%)
X_1	Main activity electricity and heat production	0.48
X_2	Manufacturing industries and construction	0.48
X_3	Road transportation no resuspension	0.00
X_4	Residential and other sectors	0.48
X_5	Oil and natural gas	60.1
X_6	Lime production	55.3
X_7	Metal industry	66.3

Illustrative example: model selection

Mixture component	G	log-likelihood	BIC
Skew-normal	1	−2140.416	−4505.009
Skew-t	1	−2101.591	−4432.696
Skew-slash	1	−2101.246	−4432.006
Skew-variance-gamma	1	−2108.206	−4445.925
Skew-normal	2	−1998.665	−4451.020
Skew-slash	2	−1982.555	−4429.476
Skew-t	2	−1979.074	−4422.514
Skew-variance-gamma	2	−1959.650	**-4383.66**
Skew-normal	3	−1908.094	−4499.392
Skew-slash	3	−1889.803	−4478.824
Skew-t	3	−1888.842	−4476.902
Skew-variance-gamma	3	−1885.642	−4470.502

Illustrative example: Cleveland Children's Sleep and Health Study

Provided by the National Sleep Research Resource (NSRR) repository
Dataset consists of 255 variables and 517 observations with a mix of categorical and continuous variables.
Dataset's dictionary highlights the commonly used variables. These are used in the application.

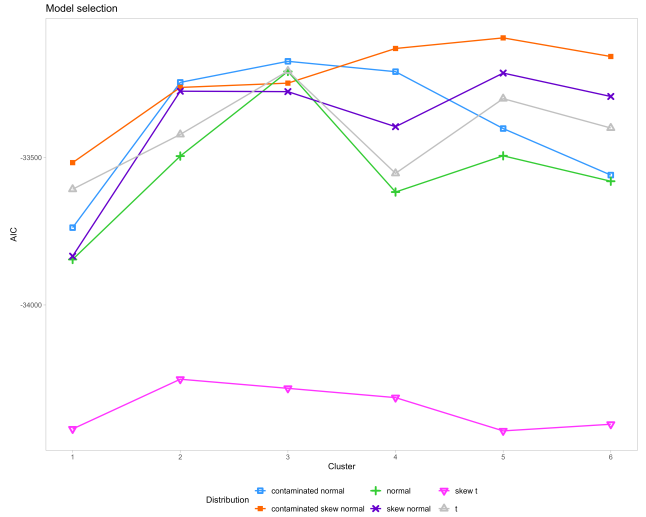
Illustrative example: Cleveland Children's Sleep and Health Study

Table: Proportion of missing values per variable from the CCSHS dataset.

Variable	Name	Missing (%)
bpsys	Systolic blood pressure	0.00
bydias	Diastolic blood pressure	0.00
bmi	Body Mass Index (BMI)	0.00
mslp	Average daily total sleep duration	13.15
cslp	Coefficient of variation of daily total sleep duration	15.67
mseff	Average daily sleep efficiency	13.15
mrigrams	Mean total grams per day	0.00
pbmi_mom	Body Mass Index (BMI) of subject's mother	18.18
pbmi_dad	Body Mass Index (BMI) of subject's father	67.12

Illustrative example: Cleveland Children's Sleep and Health Study

Figure: AIC values (vertical axis) vs the number of clusters chosen (horizontal axis).



Conclusions

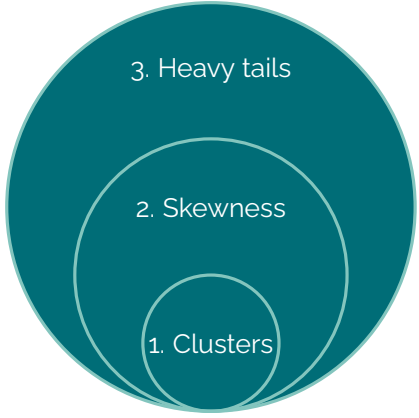
1 Heterogeneity from homogeneous subgroups.

Conclusions

2 Skew-normal encompasses normality.

1 Heterogeneity from homogeneous subgroups.

Conclusions

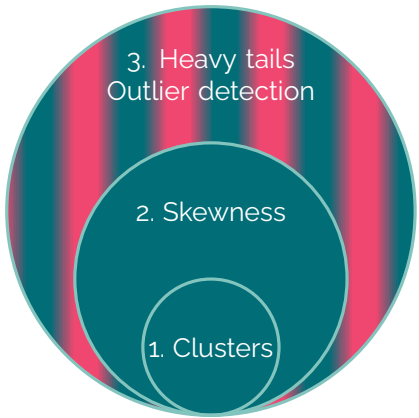


3 SMSN family, each with their own advantages and niches. The best fit can be chosen from.

2 Skew-normal encompasses normality.

1 Heterogeneity from homogeneous subgroups.

Conclusions



3 SMSN family, each with their own advantages and niches. The best fit can be chosen from. The CSN model adds interpretation to heavy tails.

2 Skew-normal encompasses normality.

1 Heterogeneity from homogeneous subgroups.

Conclusions

4 MAR mechanism preserving homogeneity, skewness, heavy-tailedness in imputations as the model is fitted to the data.

3 SMSN family, each with their own advantages and niches. The best fit can be chosen from. The CSN model adds interpretation to heavy tails.

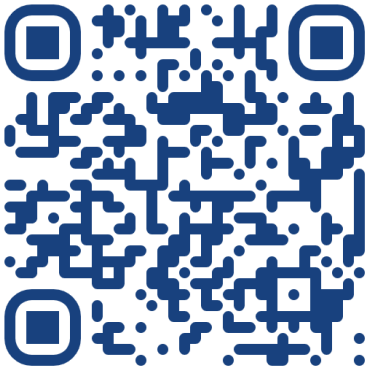
2 Skew-normal encompasses normality.

1 Heterogeneity from homogeneous subgroups.

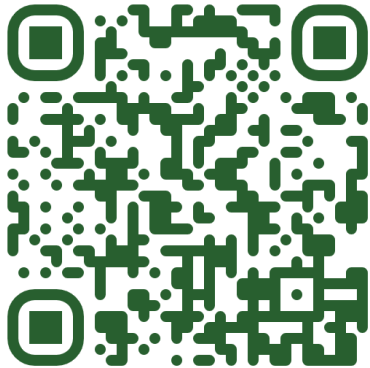
Itching to scrutinise?

Have a look at our papers!

Finite mixtures of SMSN distributions



Finite mixtures of CSN distributions



Acknowledgements

Recognition is given to the following institutions for their support towards this research:

Thank you!