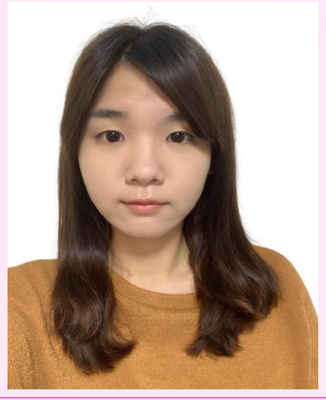
Mixtures of unrestricted skew normal factor analyzers with missing information

國立中興大學

具遺失訊息之混合非限制型的偏斜常態因子分析器



Student: Hong-Ying Chen Advisor: Tsung-I Lin

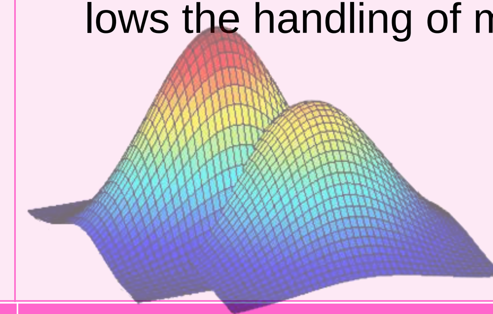
Institute of Statistics, National Chung Hsing University, Taichung, Taiwan

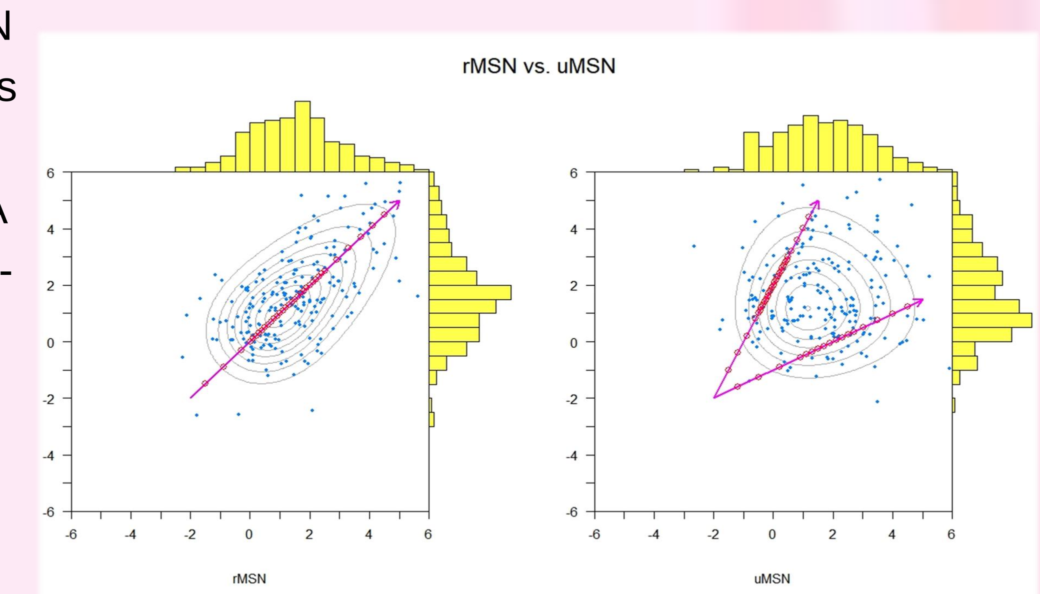
Abstract

Mixtures of factor analyzers (**MFA**) based on the restricted skew normal distribution (rMSN) has been shown to be a flexible tool for modeling **asymmetrical high-dimensional data with heterogeneity**. However, the rMSN distribution is oft-criticized a lack of sufficient ability to accommodating skewness arising from more than one feature space. This thesis presents an alternative extension of MFA by assuming the unrestricted skew normal (uMSN) distribution for the component factors. In particular, the proposed mixtures of uMSN factor analyzers (**MuSNFA**) can simultaneously accommodate multiple directions of skewness and the occurrence of missing values or nonresponses. Under the missing at random mechanism, we develop a computationally feasible **expectation conditional maximization (ECM) algorithm** for computing the maximum likelihood estimates of model parameters. Practical aspects related to model-based clustering, prediction of factor scores and missing values are also discussed. The utility of the proposed methodology is illustrated with the analysis of simulated data and an application to the Pima Indian women diabetes data containing genuine **missing values**.

Motivation

- Unlike the rMSN distribution, the uMSN distribution is able to capture skewness in more than one direction.
- Aim: propose a skew extension of MFA based on the uMSN distribution and allows the handling of missing values.





Methodology

MuSNFA model

$$\boldsymbol{Y}_j = \mu_i + \boldsymbol{B}_i\boldsymbol{U}_{ij} + \varepsilon_{ij}, \quad \text{with probability } \pi_i \ (i = 1, \ldots, g)$$

$$\boldsymbol{U}_{ij} \sim uSN_q\left(-c\boldsymbol{\Delta}_i^{-1/2}\boldsymbol{\Lambda}_i\boldsymbol{1}_q, \boldsymbol{\Delta}_i^{-1}, \boldsymbol{\Delta}_i^{-1/2}\boldsymbol{\Lambda}_i\right), \ \varepsilon_{ij} \sim N_p(\boldsymbol{0}, \boldsymbol{D}_i), \ \boldsymbol{U}_{ij} \perp \varepsilon_{ij}$$

where $\boldsymbol{\Lambda}_i$ be a q*q **skewness parameter** matrix, and $\boldsymbol{\Delta}_i = \boldsymbol{I}_q + (1 - c^2)\boldsymbol{\Lambda}_i\boldsymbol{\Lambda}_i^T$ with $c=\sqrt{2/\pi}$.

- Introduce two permutation matrices $\mathbf{O}_j(\mathrm{p}_j^o \times \mathrm{p})$ and $\boldsymbol{M}_j((\mathrm{p}-\mathrm{p}_j^o) \times \mathrm{p})$ such that $\boldsymbol{Y}_j = \boldsymbol{O}_j^T\boldsymbol{Y}_j^o + \boldsymbol{M}_j^T\boldsymbol{Y}_j^m$. The hierarchical representation of MuSNFA is

$$\boldsymbol{Y}_j{}^o|(Z_{ij}=1) \sim CFUSN_{p_j^o,q}(\mu_{ij}^o - c\boldsymbol{\alpha}_{ij}^o\boldsymbol{1}_q, \boldsymbol{\Sigma}_{ij}^{oo}, \boldsymbol{\alpha}_{ij}^o)$$

$$\boldsymbol{Z}_j \sim \mathcal{M}(1\,; \pi_1, \ldots, \pi_g),$$

where $\boldsymbol{Y}_j^o = \boldsymbol{O}_j\boldsymbol{Y}_j, \boldsymbol{Y}_j^m = \boldsymbol{M}_j\boldsymbol{Y}_j$, $\mu_{ij}^o = \boldsymbol{O}_j\mu_i$, $\boldsymbol{\Sigma}_{ij}^{oo} = \boldsymbol{O}_j\,\Sigma_i\boldsymbol{O}_j^T$, $\boldsymbol{\alpha}_{ij}^o = \boldsymbol{O}_j\alpha_i$, $\Sigma_i = \boldsymbol{B}\Delta_i^{-1}\boldsymbol{B}^T + \boldsymbol{D}_i$,

$\boldsymbol{\alpha}_i = \boldsymbol{B}\Delta_i^{-1/2}\boldsymbol{\Lambda}_i$, and $\boldsymbol{Z}_j$ is a set of binary allocation indicators.

- The component pdf of $\boldsymbol{Y}_j{}^o|(Z_{ij}=1)$ is given by

$$\boldsymbol{\varphi}(\mathbf{y}_j^o; \boldsymbol{\theta}_i) = 2^q\,\phi_{p_j^o}\left(\mathbf{y}_j^o; \boldsymbol{\mu}_{ij}^o - c\boldsymbol{\alpha}_{ij}^o\mathbf{1}_q, \boldsymbol{\Omega}_{ij}^{oo}\right) \times \Phi_q\left(\boldsymbol{\alpha}_{ij}^{o\,T}\boldsymbol{\Omega}_{ij}^{oo\,-1}(\mathbf{y}_j^o - \boldsymbol{\mu}_{ij}^o + c\boldsymbol{\alpha}_{ij}^o\mathbf{1}_q); \mathbf{I}_q - \boldsymbol{\alpha}_{ij}^{o\,T}\boldsymbol{\Omega}_{ij}^{oo\,-1}\boldsymbol{\alpha}_{ij}^o\right),$$

where $\boldsymbol{\Omega}_{ij}^{oo} = \boldsymbol{O}_j\,\boldsymbol{\Omega}_i\boldsymbol{O}_j^T$ with $\boldsymbol{\Omega}_i = \boldsymbol{\Sigma}_i + \boldsymbol{\alpha}_i\boldsymbol{\alpha}_i^T$.

- For ease of estimation, we reparametrize

$$\widetilde{\boldsymbol{B}}_i \triangleq \boldsymbol{B}_i\Delta_i^{-1/2} \text{ and } \widetilde{\boldsymbol{U}}_{ij} \triangleq \Delta_i^{1/2}\boldsymbol{U}_{ij}$$

It follows that $\boldsymbol{Y}_j = \mu_i + \widetilde{\boldsymbol{B}}_i\widetilde{\boldsymbol{U}}_{ij} + \varepsilon_{ij}$, where $\widetilde{\boldsymbol{U}}_{ij} \sim uSN_q(-c\boldsymbol{\Lambda}_i\boldsymbol{1}_q, \boldsymbol{I}_q, \boldsymbol{\Lambda}_i)$.

- Let $\boldsymbol{\Theta} = \{\pi_1, \ldots, \pi_{g-1}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_g\}$ represent all the unknown parameters of the mixture model, where $\boldsymbol{\theta}_i = (\boldsymbol{\mu}_i, \boldsymbol{B}_i, \boldsymbol{D}_i, \boldsymbol{\Lambda}_i)$.

The complete data likelihood function of $\boldsymbol{\Theta}$ given $\boldsymbol{Y}_c = (\boldsymbol{y}^o, \boldsymbol{y}^m, \widetilde{\boldsymbol{U}}, \boldsymbol{\gamma}, \boldsymbol{Z})$ is

$$\mathrm{L}_c(\boldsymbol{\Theta} \mid \boldsymbol{Y}_c) = \prod_{j=1}^{n} f(\boldsymbol{y}_j^o \mid Z_{ij}=1)f(\boldsymbol{Z}_j) = \prod_{j=1}^{n}\prod_{i=1}^{g}\{\pi_i\varphi(\boldsymbol{y}_j^o \mid \boldsymbol{\theta}_i)\}^{z_{ij}}$$

$$= \prod_{j=1}^{n}\prod_{i=1}^{g}\left\{\pi_i\phi_{p_j^o}(\boldsymbol{y}_j^o; \mu_{ij}^o + \widetilde{\boldsymbol{B}}_{ij}^o\widetilde{\mathrm{U}}_{ij}, \boldsymbol{D}_{ij}^{oo})\phi_q(\widetilde{\boldsymbol{U}}_{ij}; -c\boldsymbol{\Lambda}_i\boldsymbol{1}_q + \boldsymbol{\Lambda}_i\boldsymbol{\gamma}_j, \boldsymbol{I}_q)\right\}^{z_{ij}}$$

Proposition:

For notational simplicity, the symbol "$\mid \cdots$" stands for conditioning on $\mathbf{y}_j^o$ and $Z_{ij} = 1$.

Given the hierarchical representation, we have the following conditional expectations:

$$\mathrm{E}\left(\widetilde{\mathbf{U}}_{ij} \mid \cdots\right) = \mathbf{W}_{ij}^{oo}\{\mathbf{v}_{ij}^o + \boldsymbol{\Lambda}_i[\mathrm{E}(\boldsymbol{\gamma}_j - c\mathbf{1}_q \mid \cdots)]\}, \boldsymbol{W}_{ij}^{oo} = \boldsymbol{I}_q + \widetilde{\boldsymbol{B}}_i{}^T\boldsymbol{C}_{ij}^{oo}\widetilde{\boldsymbol{B}}_i{}^{-1}, \boldsymbol{v}_{ij}^o = \widetilde{\boldsymbol{B}}_i{}^T\boldsymbol{C}_{ij}^{oo}(\boldsymbol{y}_j - \boldsymbol{\mu}_i),$$

$$\mathrm{E}(\widetilde{\mathbf{U}}_{ij}\boldsymbol{\gamma}_j{}^T \mid \cdots) = \mathbf{W}_{ij}^{oo}\{\mathbf{v}_{ij}^o\mathrm{E}(\boldsymbol{\gamma}_j{}^T \mid \cdots) + \boldsymbol{\Lambda}_i[\mathrm{E}(\boldsymbol{\gamma}_j\boldsymbol{\gamma}_j{}^T \mid \cdots) - c\mathbf{1}_q\mathrm{E}(\boldsymbol{\gamma}_j{}^T \mid \cdots)]\} \text{ and}$$

$$\mathrm{E}\left(\widetilde{\mathbf{U}}_{ij}\widetilde{\mathbf{U}}_{ij}{}^T \mid \cdots\right) = \left\{\boldsymbol{I}_q + \mathrm{E}\left(\widetilde{\mathbf{U}}_{ij} \mid \cdots\right)\mathbf{v}_{ij}^{oT} + [\mathrm{E}(\widetilde{\mathbf{U}}_{ij}\boldsymbol{\gamma}_j{}^T \mid \cdots) - c\mathrm{E}\left(\widetilde{\mathbf{U}}_{ij} \mid \cdots\right)\mathbf{1}_q^T]\boldsymbol{\Lambda}_i\right\}\mathbf{W}_{ij}^{oo\,T},$$

where $\mathrm{E}(\boldsymbol{\gamma}_j{}^T \mid \cdots)$ and $\mathrm{E}(\boldsymbol{\gamma}_j\boldsymbol{\gamma}_j{}^T \mid \cdots)$ are the first-two moments of the TMVN distribution.

ECM algorithm

Given the initial parameter values $\hat{\boldsymbol{\theta}}^{(0)}$

E-step :
Calculate the Q function
$\mathrm{Q}(\boldsymbol{\Theta}|\hat{\boldsymbol{\Theta}}^{(k)}) = E\{\ell_c(\boldsymbol{\Theta} \mid \boldsymbol{Y}_c) | \boldsymbol{y}^o, \hat{\boldsymbol{\Theta}}^{(k)}\}$

CM-steps :
Given current parameter estimates, conditionally maximize $\mathrm{Q}(\boldsymbol{\Theta}|\hat{\boldsymbol{\Theta}}^{(k)})$ to obtain updated parameter values.

$\hat{\pi}_i{}^{(k)}, \hat{\mu}_i{}^{(k)}, \hat{\boldsymbol{B}}_i{}^{(k)}, \hat{\boldsymbol{D}}_i{}^{(k)}, \hat{\Lambda}_i{}^{(k)}$
Convergence ?

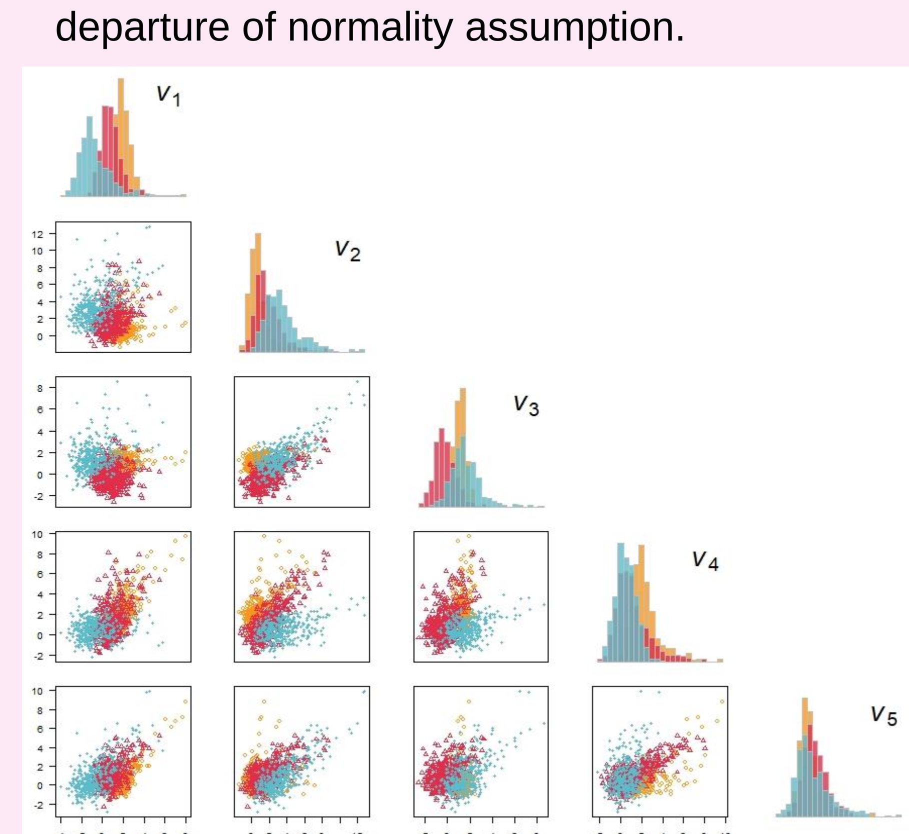
No: Iter = iter+1

Yes: Output $\hat{\boldsymbol{\theta}}^{(k)}$

Simulation & Real data

Simulation

- A simulation study is undertaken to examine the performance of the proposed MuSNFA model in comparison with the MFA and MrSNFA approaches under varying proportions of **synthetic missing values.**
- 100 Monte Carlo samples are created from the **3-component MFA model with p=5 and q=2 factors**, while the latent factors are independently generated from the chi-square distribution with one degree of freedom χ_1^2 for yielding strong effects of skewness and kurtosis.
- MuSNFA model outperforms the other two competing approaches in almost all trials.
- The MFA model has the worst performance due to a lack of sufficient robustness against serious departure of normality assumption.



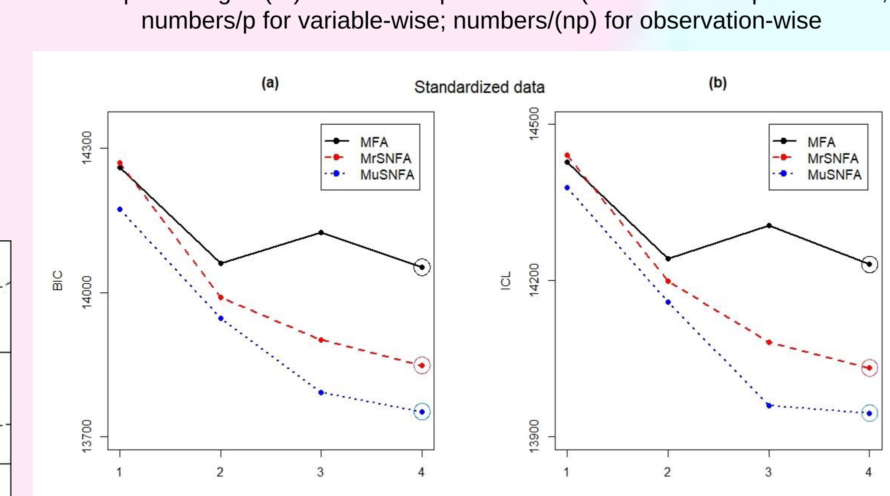
Missing rate	Criterion		$n=300$ MFA	$n=300$ MrSNFA	$n=300$ MuSNFA	$n=600$ MFA	$n=600$ MrSNFA	$n=600$ MuSNFA	$n=1200$ MFA	$n=1200$ MrSNFA	$n=1200$ MuSNFA
$r=10\%$	BIC	Mean	4790.81	4756.70	4735.59	8958.33	8828.85	8782.34	17664.59	17352.88	17243.15
		Freq	4	14	82	0	4	96	0	6	94
	ICL	Mean	4902.15	4864.91	4834.12	9246.50	9112.59	9022.93	18265.53	17944.62	17717.26
		Freq	6	11	83	0	6	94	0	2	98
	CCR	Mean	0.834	0.842	0.870	0.834	0.858	0.886	0.832	0.873	0.911
		Freq	25	19	56	12	25	63	8	5	87
	ARI	Mean	0.623	0.634	0.689	0.617	0.646	0.710	0.617	0.675	0.759
		Freq	23	17	60	12	24	64	8	3	89
	MSE	Mean	1.184	1.139	1.112	1.183	1.151	1.130	1.142	1.097	1.066
		Freq	14	35	51	6	28	66	2	16	82
$r=30\%$	BIC	Mean	3757.54	3702.84	3693.55	7281.53	7134.67	7113.46	14331.74	13997.08	13938.48
		Freq	0	16	84	0	12	88	0	9	91
	ICL	Mean	3972.24	3882.21	3871.68	7747.61	7552.67	7513.12	15312.47	14892.26	14750.88
		Freq	0	24	76	0	20	80	0	7	93
	CCR	Mean	0.781	0.745	0.766	0.803	0.759	0.803	0.813	0.782	0.826
		Freq	49	16	35	34	5	61	28	12	60
	ARI	Mean	0.475	0.439	0.473	0.508	0.451	0.523	0.526	0.486	0.563
		Freq	47	16	37	32	5	63	25	14	61
	MSE	Mean	1.529	1.485	1.457	1.432	1.393	1.379	1.419	1.353	1.335
		Freq	16	29	55	15	26	59	1	28	71

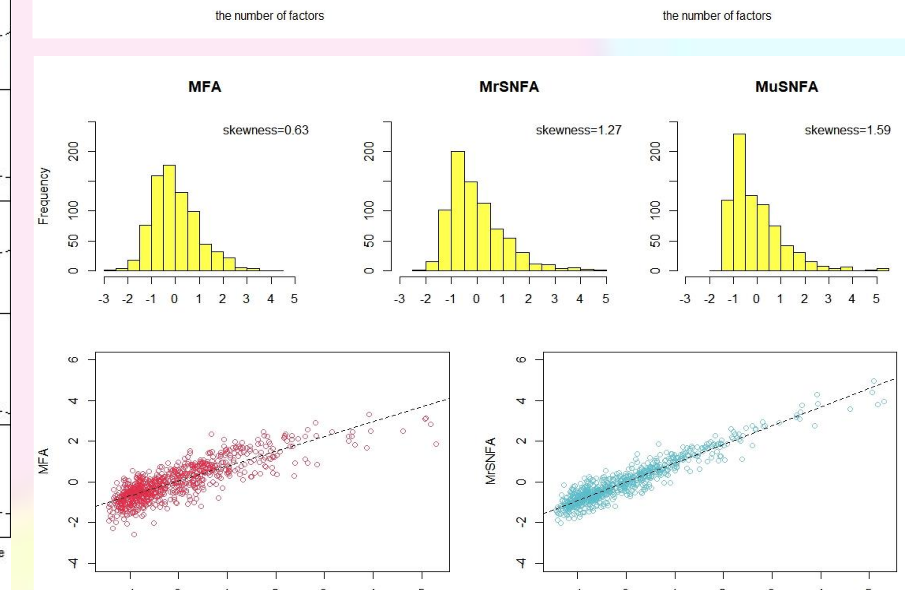
Pima Indians Diabetes Data Set (PIMA data)

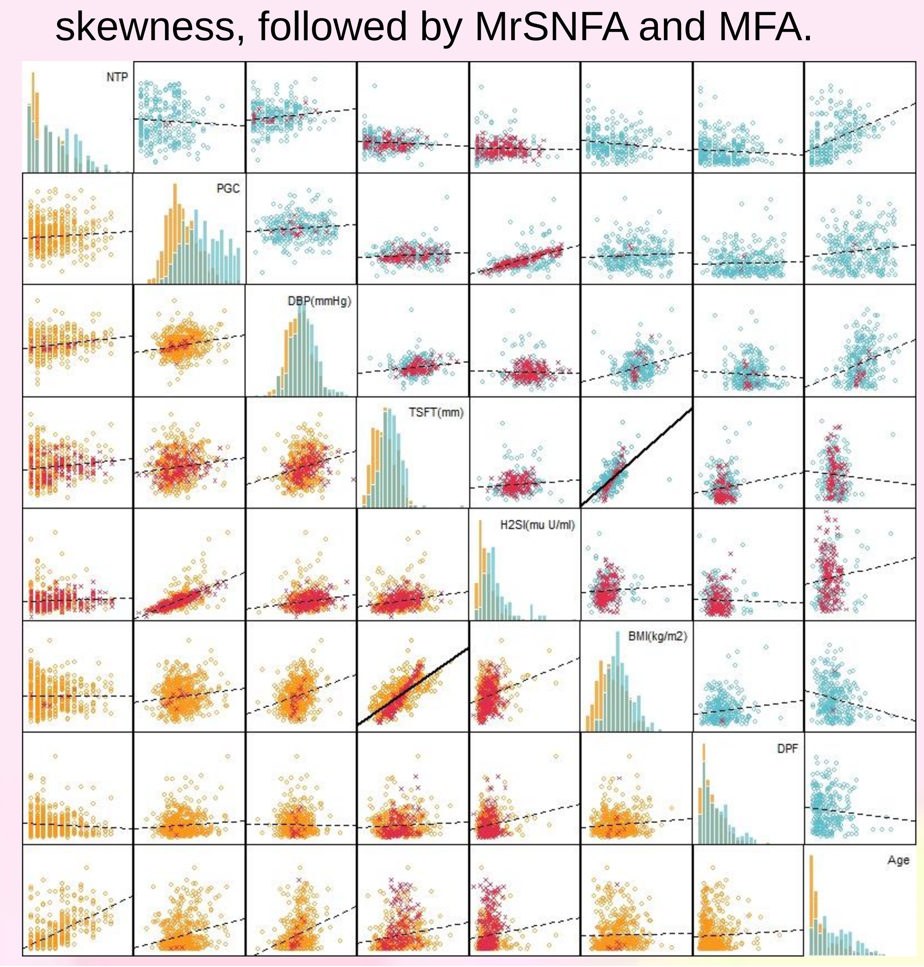
- There are (p=8) attributes measured on 500 non-diabetes and 278 diabetes female patients.
- We consider the fitting of 2-component MFA, MrSNFA and MuSNFA models to this dataset with q ranging from 1-4.
- The factor scores predicted by the three approaches are all positively skewed, but the scattering patterns somewhat different.
- The MuSNFA offer the strongest degree of skewness, followed by MrSNFA and MFA.

	Number of missing items 0	1	2	3	4	5	Patients	NA-patient	NA-var	NA-obs
Non-diabetes	262 (52.40)	95 (19.00)	124 (24.80)	13 (2.60)	6 (1.20)	0 (0.00)	500 (65.10)	238 (30.99)	5 (62.5)	406 (10.15)
Diabets	130 (48.50)	47 (17.54)	75 (27.99)	15 (5.60)	1 (0.37)	0 (0.00)	268 (34.90)	138 (17.97)	5 (62.5)	246 (11.47)
Total	392 (51.04)	142 (18.49)	199 (25.91)	28 (3.65)	7 (0.91)	0 (0.00)	768 (100)	376 (48.96)	8 (100)	652 (10.61)

The percentages (%) are listed in parentheses (numbers/n for patient-wise; numbers/p for variable-wise; numbers/(np) for observation-wise







Conclusions and future work

- The proposed MuSNFA model is a new model-based clustering tool for handling asymmetric high-dimensionality data with possibly missing values.
- Experimental results demonstrate that the proposed MuSNFA model may outperform MFA and MrSNFA on providing a better fit, improved classification performance, and more accurate prediction for missing values.
- Future research can expand the capabilities of the current method to analyze the data with multimodality, multiple skewness, missing values and censored responses simultaneously , see Lin et al. (2018) and Lin and Wang (2020) in the context of multivariate linear mixed models.

References

- Ghahramani and Hinton(1997) The EM algorithm for factor analyzers. Technical Report No. CRG-TR-96-1. The University of Toronto, Toronto
- Lin TI, Lachos VH, Wang WL (2018) Multivariate longitudinal data analysis with censored and intermittent missing responses. Stat Med 37:2822-2835
- Lin TI, McLachlan GJ and Lee SX (2016) Extending mixtures of factor models using the restricted multivariate skew-normal distribution. J Multivariate Anal 143:398-413
- Lin TI, Wang WL (2020) Multivariate-t linear mixed models with censored responses, intermittent missing values and heavy tails. Stat Meth Med Res 29:1288-1304
- Meng XL, Rubin DB (1993) Maximum likelihood estimation via the ECM algorithm: a general framework. Biometrika 80:267-278