Normal scores transform for dependence analysis of continuous variables.

Data $(y_{i1}, \ldots, y_{id})$, $i = 1, \ldots, n$.

(a) Rank transform to normal scores

Rank the jth variable vector $(y_{1j}, \ldots, y_{nj})$ in increasing order to get ranks $R_{1j}, \ldots, R_{nj}$ (permutation of $1, \ldots, n$). $\hat{z}_{ij} = \Phi^{-1}((R_{ij} - 0.5)/n)$ for $i = 1, \ldots, n$, consist of normal scores transform for jth variable.

(b) Parametric transform to normal scores

jth variable univariate model $F_j(\cdot; \eta_j)$ is fitted to $(y_{1j}, \ldots, y_{nj})$.

$\hat{u}_{ij} = F_j(y_{ij}; \hat{\eta}_j)$; $\hat{z}_{ij} = \Phi^{-1}(\hat{u}_{ij})$ for normal scores transform.

Bivariate plots: look for deviations from ellptical shape clouds.

For $j \neq k$, plot $(\hat{z}_{ij}, \hat{z}_{ik})$, $i = 1, \ldots, n$.

Discrete version of normal scores; fit Gaussian copula

For multivariate discrete data, fit discretized multivariate Gaussian with well chosen univariate marginal models.

Compared observed and expected counts for each bivariate margin, and look for possible structural deviations from the Gaussian dependence model.

Longitudinal doctor visits

Application to repeated measures data

Data set: Riphahn, Wambach. Million 2003, J Applied Econometrics; analysis: Greene 2008, Economics Letters, Nikoloulopoulos, Joe, Chaganty 2011 with NB regression models.

A count response variable (#doctor visits in last quarter) is measured at $d = 5$ consecutive years 1984–1988 for each of $n = 295$ subjects.

Ignoring the repeated counts over subjects, 33% of the counts were 0, maximum is 60, median is 2. Ignoring the covariates, the sample means, variances and dispersion indices for the 5 time points suggest overdispersion of counts relative to Poisson.

NB: ϑ convolution parameter, p probability parameter, $\xi = p^{-1} - 1 \geq 0$ so that mean is $\mu = \vartheta\xi = \vartheta(1-p)/p$, variance is $\sigma^2 = \mu(1+\xi) = \vartheta(1-p)/p^2$

$$f_{NB}(y; \vartheta, \xi) = \frac{\Gamma(\vartheta + y)}{\Gamma(\vartheta)\, y!} \frac{\xi^y}{(1+\xi)^{\vartheta+y}}, \quad y = 0, 1, 2, \ldots, \ \vartheta > 0, \ \xi > 0,$$

$\vartheta \to \infty$, $\xi \to 0$ with $\vartheta\xi$ fixed to get Poisson.

GP: ϑ convolution parameter, $0 \leq \varrho < 1$; mean $\mu = \vartheta/(1-\varrho)$, variance is $\sigma^2 = \vartheta/(1-\varrho)^3$.

$$f_{GP}(y; \vartheta, \varrho) = \frac{\vartheta(\vartheta + \varrho y)^{y-1}}{y!} e^{-\vartheta - \varrho y}, \quad y = 0, 1, 2, \ldots, \ \vartheta > 0, \ 0 \leq \varrho < 1.$$

---

count regression where $\log \mu = \boldsymbol{\beta}^T \boldsymbol{x}$; $\boldsymbol{x}$=covariate

NB1: convolution parameter depends on $\boldsymbol{x}$, dispersion index doesn't.

GP1: convolution parameter depends on $\boldsymbol{x}$, dispersion index doesn't.

NB2: dispersion index depends on $\boldsymbol{x}$, convolution parameter doesn't; similar for GP2.

Step 1: Univariate margins NB1, NB2, GP1 with several covariates. Ignoring serial dependence, compare univariate regression models composite likelihood (CL) estimation and CL Akaike/Bayesian information criteria.

1. Looking at the diagnostics of comparing univariate model-based expected frequencies versus observed frequencies, the GP1 and NB1 models are much better fits than NB2.

2. Based on the maximum composite log-likelihood, GP1 fits a little better than NB1.
3. The GP2 model had problems with the constraint of $1 - \varrho(\boldsymbol{x}) = \vartheta/\mu(\boldsymbol{x}) \in (0, 1]$ for some subsets of covariates so we do not include it in further comparisons.

parameter	NB1	NB2	GP1	GP1
intercept	1.73 (0.16)	2.11 (0.17)	1.69 (0.16)	1.66 (0.16)
ifemale	0.34 (0.09)	0.27 (0.12)	0.36 (0.09)	0.40 (0.09)
agec	0.18 (0.05)	0.17 (0.07)	0.19 (0.05)	0.19 (0.05)
agec^2	0.11 (0.05)	0.07 (0.06)	0.12 (0.05)	0.11 (0.05)
healthsat	-0.15 (0.02)	-0.21 (0.02)	-0.15 (0.02)	-0.15 (0.02)
handicap	0.67 (0.15)	0.78 (0.20)	0.69 (0.14)	0.70 (0.14)
iuniversity	-0.52 (0.26)	-0.82 (0.39)	-0.53 (0.27)	
ξ: overdispersion-1	4.80 (0.50)		5.78 (0.70)	5.84 (0.71)
ϑ convolution		0.77 (0.06)		
neg. comp. loglik.	3252	3267	3240	3245
tr($\boldsymbol{HV}$)	18.2	21.3	17.7	15.3
CLAIC	6541	6577	6515	6521
CLBIC	6608	6656	6580	6577

Table 1: comparison of univariate count regression models; estimates (standard errors) based on one-wise composite likelihood, minimum of L=negative composite log-likelihood, penalty tr($\boldsymbol{HV}$), CLAIC and CLBIC; asymptotic covariance $\boldsymbol{V}$ of $(\boldsymbol{\beta}, \xi, \theta)$ vector was estimated with the delete-one jackknife, $\boldsymbol{H}$=Hessian of L. The best univariate regression model with all criteria is GP1; the number of covariates is 5 for CLBIC and 6 for CLAIC.

For the multivariate model, proceed with the GP1 regression model that includes `iuniversity`.

Step 2: Some preliminary analysis of dependence based on polychoric correlations with the fitted univariate count regression

The matrix of latent (polychoric) correlations from fitting a multivariate Gaussian copula:

$$\widehat{\boldsymbol{R}} = \begin{pmatrix} 1.00 & 0.54 & 0.46 & 0.34 & 0.38 \\ 0.54 & 1.00 & 0.35 & 0.36 & 0.34 \\ 0.46 & 0.35 & 1.00 & 0.34 & 0.28 \\ 0.34 & 0.36 & 0.34 & 1.00 & 0.35 \\ 0.38 & 0.34 & 0.28 & 0.35 & 1.00 \end{pmatrix}$$

1. This matrix has some small deviations from an exchangeable correlation matrix.
2. Comparison of bivariate marginal expected versus observed counts suggest possibly more dependence in the joint lower tail than Gaussian.

margin		count							
		0	1	2	3	4	5	6	≥ 7
1	$\boldsymbol{E}_1$	105	50	32	22	16	13	10	48
	$\boldsymbol{O}_1$	105	36	40	25	27	8	8	46
2	$\boldsymbol{E}_2$	101	51	33	23	17	13	10	48
	$\boldsymbol{O}_2$	105	36	39	24	19	12	15	45
(1,2)	$\boldsymbol{E}_{12}$								
	0	62.6	18.8	9.0	5.0	3.0	1.9	1.3	3.2
	1	17.1	10.7	6.6	4.3	2.9	2.0	1.5	4.5
	2	8.0	6.4	4.4	3.1	2.3	1.7	1.3	4.5
	3	4.4	4.1	3.1	2.3	1.8	1.3	1.0	4.1
	4	2.7	2.8	2.2	1.7	1.4	1.1	0.9	3.7
	5	1.7	1.9	1.6	1.3	1.1	0.9	0.7	3.3
	6	1.1	1.4	1.2	1.0	0.9	0.7	0.6	2.9
	≥ 7	3.0	4.5	4.5	4.2	3.8	3.4	3.0	21.9
	$\boldsymbol{O}_{12}$								
	0	75	12	8	4	2	1	0	3
	1	8	9	6	2	4	9	4	3
	2	10	4	6	4	5	3	2	6
	3	6	4	4	4	2	1	1	3
	4	2	3	4	5	3	3	2	5
	5	0	0	3	1	0	0	1	3
	6	2	1	3	2	0	0	0	0
	≥ 7	2	3	5	2	3	4	5	22

Bivariate Gaussian model-based expected counts compared with observed counts for margins 1, 2 and (1,2); $\boldsymbol{E}$ is the symbol for expected counts and $\boldsymbol{O}$ is the symbol for observed counts. The univariate model is GP1 with 6 covariates.

Step 3: compare copula models

copula	estimated parameter(s)	-loglik.	AIC	BIC
independence		3240	6480	6480
exchangeable Gaussian	0.37	3129	6260	6264
exchangeable Frank	2.19	3133	6267	6271
exchangeable Gumbel	1.27	3138	6279	6283
exchangeable r.Gumbel	1.30	3150	6301	6305
D-vine with Gaussian	0.54,0.35,0.35,0.35	3118	6255	6292
	0.35,0.28,0.19			
	0.13,0.21,0.18			
D-vine with Frank	3.54,2.20,2.14,2.25	3119	6258	6295
	2.35,1.71,1.29			
	0.76,1.00,1.03			
D-vine with r.Gumbel	1.61,1.33,1.32,1.31	3117	6254	6291
	1.28,1.23,1.15			
	1.08,1.15,1.14			

Doctor visits: copula models and second stage negative log-likelihoods with univariate parameters held fixed. r.Gumbel is an abbreviation for the reflected/survival Gumbel copula. The negative log-likelihood from the independence copula is the same as the one-wise negative composite log-likelihood in Table 1; it is included in this table as a baseline for comparisons. The dependence parameters for D-vines are $\theta_{12}, \theta_{23}, \theta_{34}, \theta_{45}$, $\theta_{13;2}, \theta_{24;3}, \theta_{35;4}$,

$\theta_{14;23}, \theta_{25;34}, \theta_{15;234}$. The best dependence model (among the above) based on AIC is D-vine with r.Gumbel and the best model based on BIC is exchangeable Gaussian.

Multivariate discrete data, not longitudinal.

Get polychoric correlation matrix.

Check if 1-factor structure or Markov tree dependence is a good approximation. If not, consider 2-factor, 2-truncated vine, etc.

After getting good dependence structure, use copula model with asymmetric copulas if this is suggested from bivariate marginal comparisons of observed versus expected based on discretized Gaussian.