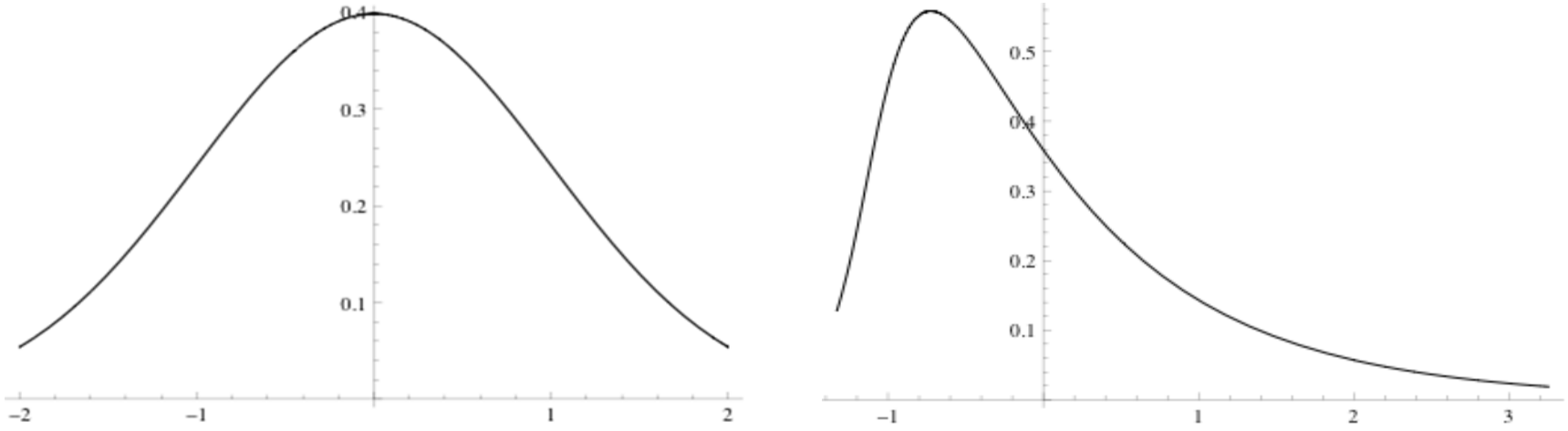
The Multivariate Percentile Power Method Transformation



Dr. Jennifer Koran
Mathematics Colloquium
Southern Illinois University Carbondale
November 10, 2016

Power Method (PM) Transformation

Headrick (2010):

$$p(Z) = \sum_{i=1}^{m} c_i Z^{i-1}$$

$$f_Z(z) = \varphi(z) = (2\pi)^{-\frac{1}{2}} \exp\{-z^2/2\}$$

$$F_Z(z) = \Phi(z) = \int_{-\infty}^{z} \varphi(u)du\,, -\infty < z < +\infty$$

The conventional moment based Fleishman third-order power method

Headrick (2010), based on Fleishman (1978):

$$\alpha_1 = 0 = c_1 + c_3$$

$$\alpha_2 = 1 = c_2^2 + 2c_3^2 + 6c_2c_4 + 15c_4^2$$

$$\alpha_3 = 8c_3^3 + 6c_2^2c_3 + 72c_2c_3c_4 + 270c_3c_4^2$$

$$\alpha_4 = 3c_2^4 + 60c_2^2c_3^2 + 60c_3^4 + 60c_2^3c_4 + 936c_2c_3^2c_4 + 630c_2^2c_4^2 + 4500c_3^2c_4^2 + 3780c_2c_4^3 + 10395c_4^4 - 3.$$

The percentile based power method uses four moment-like parameters

Karian and Dudewicz (2011, pp. 172-173):

Median: $\gamma_1 = \theta_{0.50}$

Inter-decile range: $\gamma_2 = \theta_{0.90} - \theta_{0.10}$

Left-right tail-weight ratio : $\gamma_3 = \frac{\theta_{0.50} - \theta_{0.10}}{\theta_{0.90} - \theta_{0.50}}$ *"percentile skew"*

Tail-weight factor: $\gamma_4 = \frac{\theta_{0.75} - \theta_{0.25}}{\gamma_2}$ *"percentile kurtosis"*

Restrictions:

$-\infty < \gamma_1 < +\infty, \;\; \gamma_2 \geq 0, \;\; \gamma_3 \geq 0, \;\; 0 \leq \gamma_4 \leq 1$

A symmetric distribution implies that $\gamma_3 = 1$.

Substitute the standard normal distribution percentiles (z_u) into parameter equations

$$\gamma_1 = p(z_{0.50}) = c_1$$

$$\gamma_2 = p(z_{0.90}) - p(z_{0.10}) = 2c_2 z_{0.90} + 2c_4 z_{0.90}^3$$

$$\gamma_3 = \frac{p(z_{0.50}) - p(z_{0.10})}{p(z_{0.90}) - p(z_{0.50})} = 1 - \frac{2c_3 z_{0.90}}{c_2 + c_3 z_{0.90} + 2c_4 z_{0.90}^2}$$

$$\gamma_4 = \frac{p(z_{0.75}) - p(z_{0.25})}{\gamma_2} = \frac{2c_2 z_{0.75} + 2c_4 z_{0.75}^3}{2c_2 z_{0.90} + 2c_4 z_{0.90}^3}$$

where $z_{0.50} = 0, z_{0.90} = 1.281\cdots,\ z_{0.75} = 0.6744\cdots$ from the standard normal distribution.

closed-form expressions for the Percentile PM coefficients

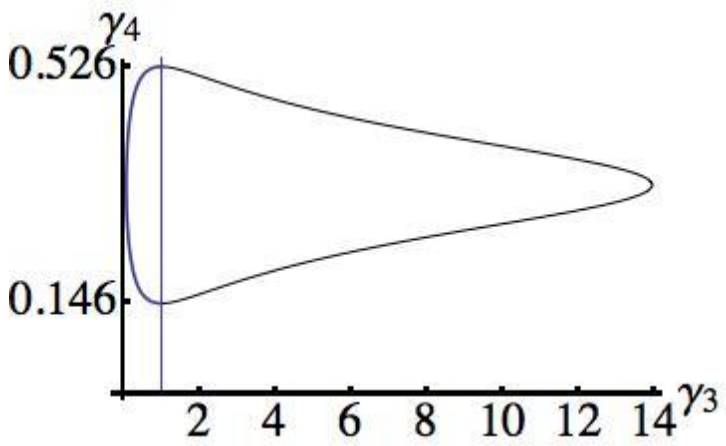
$$c_1 = \gamma_1$$

$$c_2 = \frac{\gamma_2(\gamma_4 z_{0.90}^3 - z_{0.75}^3)}{2z_{0.90}^3 z_{0.75} - 2z_{0.90} z_{0.75}^3}$$

$$c_3 = \frac{\gamma_2(1-\gamma_3)}{2(1+\gamma_3)z_{0.90}^2}$$

$$c_4 = -\frac{\gamma_2(\gamma_4 z_{0.90} - z_{0.75})}{2z_{0.90}^3 z_{0.75} - 2z_{0.90} z_{0.75}^3}$$

Boundary conditions for Percentile PM pdfs



Univariate Percentile PM Transformation process

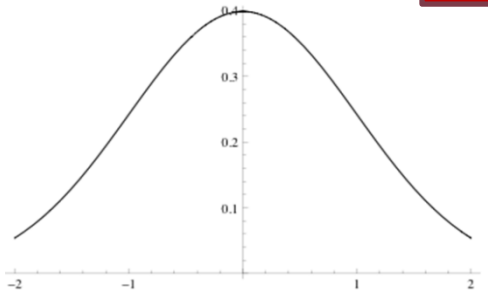
$$c_1 = \gamma_1$$

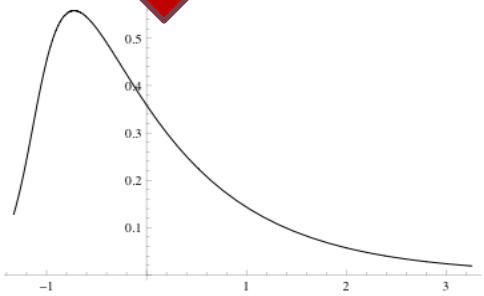
$$c_2 = \frac{\gamma_2(\gamma_4 z_{0.90}^3 - z_{0.75}^3)}{2z_{0.90}^3 z_{0.75} - 2z_{0.90} z_{0.75}^3}$$

$$c_3 = \frac{\gamma_2(1-\gamma_3)}{2(1+\gamma_3)z_{0.90}^2}$$

$$c_4 = -\frac{\gamma_2(\gamma_4 z_{0.90} - z_{0.75})}{2z_{0.90}^3 z_{0.75} - 2z_{0.90} z_{0.75}^3}$$

$$p(Z) = \sum_{i=1}^{m} c_i Z^{i-1}$$





Simulating correlated data

$$p_J(Z_J) = J$$

$$p_K(Z_K) = K$$

?

$$Corr(Z_J, Z_K) \neq Corr(\mathrm{J}, K)$$

Intermediate correlation

Specified correlation

Multivariate Conventional PM

Vale and Maurelli (1983)

$$\rho_{jk} = E\left[p(Z_j)p(Z_k)\right]$$
$$= r_{jk}\left(c_{j2}c_{k2} + 3c_{j4}c_{k2} + 3c_{j2}c_{k4} + 9c_{j4}c_{k4} + 2c_{j1}c_{k1}r_{jk} + 6c_{j4}c_{k4}r_{jk}^2\right)$$

Specified Correlation Matrix
P

	1	2	3	4
1	1			
2	0.80	1		
3	0.70	0.60	1	
4	0.65	0.50	0.45	1

Intermediate Correlation Matrix
R

	1	2	3	4
1	1			
2	0.897	1		
3	0.831	0.666	1	
4	0.750	0.580	0.489	1

Multivariate Percentile PM with Spearman correlation

$$\xi_{jk} = \frac{6}{\pi}\left\{\left(\frac{n-2}{n-1}\right)\sin^{-1}\left(\frac{r_{jk}}{2}\right) + \left(\frac{1}{n-1}\right)\sin^{-1}\left(r_{jk}\right)\right\}$$

Specified Correlation Matrix Ξ				
	1	2	3	4
1	1			
2	0.80	1		
3	0.70	0.60	1	
4	0.65	0.50	0.45	1

Intermediate Correlation Matrix, $n = 25$ R				
	1	2	3	4
1	1			
2	0.835	1		
3	0.739	0.639	1	
4	0.689	0.536	0.484	1

Multivariate Percentile PM Transformation process

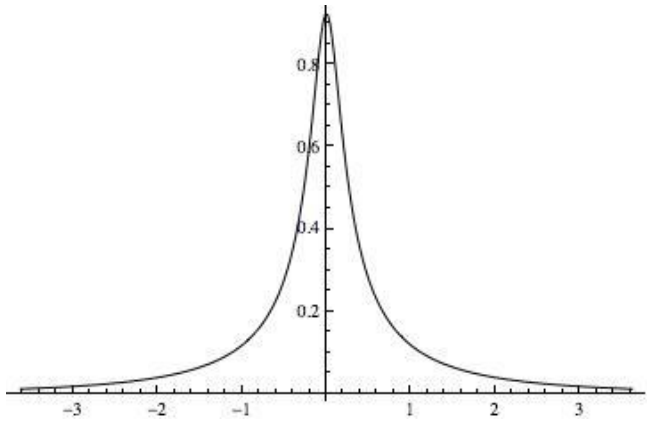
1. Specify percentiles and obtain polynomial coefficients to transform each variable
2. Specify Spearman correlations for each pair of variables
3. Solve for intermediate Pearson correlations
4. Simulate random normal variates with the intermediate Pearson correlations
5. Substitute the random normal variates into the polynomial equations using the coefficients from Step 1

The Simulation and Monte Carlo Study

- Fortran algorithm
- generate 25,000 independent sample estimates for the specified parameters
  - conventional skew (α_3) and kurtosis (α_4) and
  - left-right tail-weight ratio (γ_3) and tail-weight factor (γ_4)
- $n = 25$ and $n = 750$
- Bias-corrected accelerated bootstrapped median estimates, using 10,000 resamples [Spotfire S+]

Distribution 1

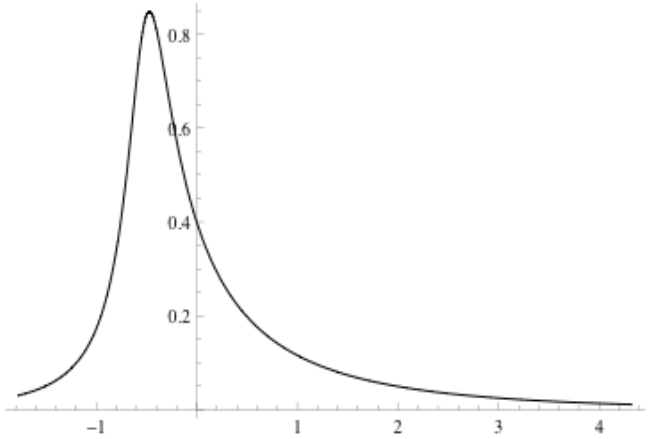
Figure 1. The power method (PM) pdf of Distribution 1.



Conventional PM	Percentile PM	Percentiles
Skew: $\alpha_3 = 0$	Left-right tail-weight ratio: $\gamma_3 = 1.0000$	$\theta(x)_{0.10} = -0.7560$
Kurtosis: $\alpha_4 = 25$	Tail-weight factor : $\gamma_4 = 0.3105$	$\theta(x)_{0.25} = -0.2347$
		$\theta(x)_{0.50} = 0$
$c_1 = 0$	$c_1 = 0$	$\theta(x)_{0.75} = 0.2347$
$c_2 = 0.2553$	$c_2 = 0.4327$	$\theta(x)_{0.90} = 0.7560$
$c_3 = 0$	$c_3 = 0$	
$c_4 = 0.2038$	$c_4 = 0.3454$	

Distribution 2

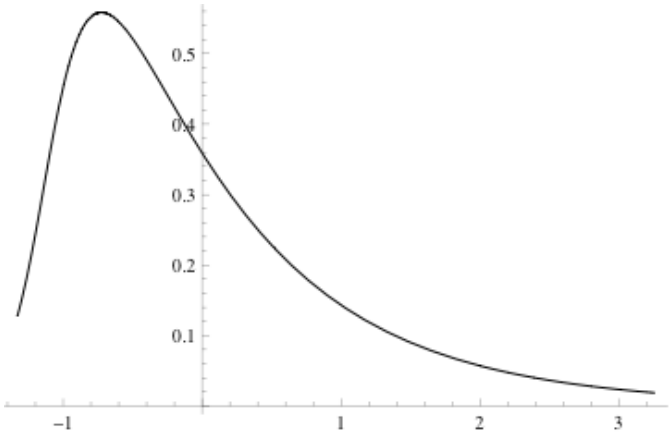
Figure 2. The power method (PM) pdf of Distribution 2.



Conventional PM	Percentile PM	Percentiles
Skew: $\alpha_3 = 3$	Left-right tail-weight ratio: $\gamma_3 = 0.3130$	$\theta(x)_{0.10} = -0.6851$
Kurtosis: $\alpha_4 = 21$	Tail-weight factor : $\gamma_4 = 0.3335$	$\theta(x)_{0.25} = -4652$
		$\theta(x)_{0.50} = -0.2523$
$c_1 = -0.2523$	$c_1 = -0.3203$	$\theta(x)_{0.75} = 0.1901$
$c_2 = 0.4186$	$c_2 = 0.5315$	$\theta(x)_{0.90} = 1.0092$
$c_3 = 0.2523$	$c_3 = 0.3203$	
$c_4 = 0.1476$	$c_4 = 0.1874$	

Distribution 3

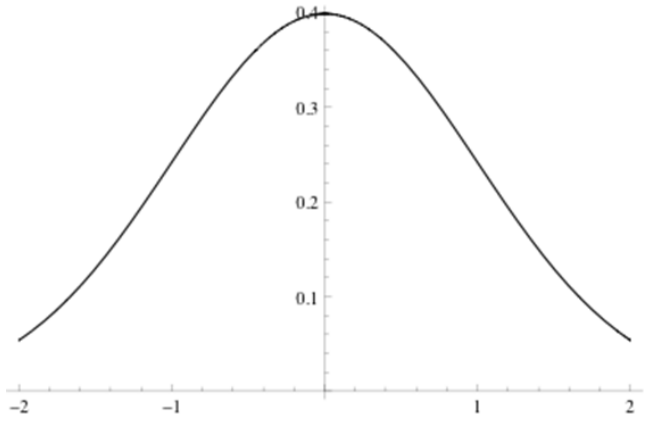
Figure 3. The power method (PM) pdf of Distribution 3.



Conventional PM	Percentile PM	Percentiles
Skew: $\alpha_3 = 2$	Left-right tail-weight ratio: $\gamma_3 = 0.2841$	$\theta(x)_{0.10} = -0.9207$
Kurtosis: $\alpha_4 = 7$	Tail-weight factor : $\gamma_4 = 0.1894$	$\theta(x)_{0.25} = -0.6717$
		$\theta(x)_{0.50} = -0.2600$
$c_1 = -0.2600$	$c_1 = -0.2908$	$\theta(x)_{0.75} = 0.3882$
$c_2 = 0.7616$	$c_2 = 0.8516$	$\theta(x)_{0.90} = 1.2547$
$c_3 = 0.2600$	$c_3 = 0.2908$	
$c_4 = 0.0531$	$c_4 = 0.0593$	

Distribution 4

Figure 4. The power method (PM) pdf of Distribution 4.



Conventional PM	Percentile PM	Percentiles
Skew: $\alpha_3 = 0$	Left-right tail-weight ratio: $\gamma_3 = 0.0000$	$\theta(x)_{0.10} = -1.2816$
Kurtosis: $\alpha_4 = 0$	Tail-weight factor : $\gamma_4 = 0.1226$	$\theta(x)_{0.25} = -0.6745$
		$\theta(x)_{0.50} = 0$
$c_1 = 0$	$c_1 = 0$	$\theta(x)_{0.75} = 0.6745$
$c_2 = 1$	$c_2 = 1$	$\theta(x)_{0.90} = 1.2816$
$c_3 = 0$	$c_3 = 0$	
$c_4 = 0$	$c_4 = 0$	

Marginal Results n = 25

Skew (α_3) and Kurtosis (α_4) results for the Conventional PM.

Dist	Parameter	Estimate	95% Bootstrap C.I.	Standard Error	Relative Bias %
1	$\alpha_3 = 0$	-0.0223	-0.0497,0.0045	0.013660	--
	$\alpha_4 = 25$	4.4560	4.4011,4.5261	0.030200	-82.18
2	$\alpha_3 = 3$	1.5750	1.5579,1.5911	0.008122	-47.50
	$\alpha_4 = 21$	3.6960	3.6452,3.7525	0.027010	-82.40
3	$\alpha_3 = 2$	1.2780	1.2677,1.2893	0.005561	-36.10
	$\alpha_4 = 7$	1.5230	1.4849,1.5662	0.020430	-78.24
4	$\alpha_3 = 0$	0.0034	-0.0038,0.0103	0.003626	--
	$\alpha_4 = 0$	-0.1786	-0.1906,-0.1678	0.005579	--

Left-right tail-weight ratio (γ_3) and tail-weight factor (γ_4)results for Percentiles PM

Dist	Parameter	Estimate	95% Bootstrap C.I.	Stand. Error	Relative Bias %
1	$\gamma_3 = 1.0000$	1.0050	0.9942, 1.0154	0.005348	--
	$\gamma_4 = 0.3105$	0.3208	0.3191, 0.3227	0.000947	--
2	$\gamma_3 = 0.3430$	0.3466	0.3438, 0.3497	0.001485	1.04
	$\gamma_4 = 0.3868$	0.3972	0.3954, 0.3993	0.000983	2.70
3	$\gamma_3 = 0.4361$	0.4472	0.4444, 0.4501	0.001464	2.53
	$\gamma_4 = 0.4872$	0.4960	0.4943, 0.4980	0.001003	1.80
4	$\gamma_3 = 1.0000$	0.9978	0.9912, 1.0045	0.003380	--
	$\gamma_4 = 0.5263$	0.5294	0.5279, 0.5310	0.000801	--

Correlation Results n = 25

Correlation results for the Conventional PM, $n = 25$

Parameter	Estimate	95% Bootstrap C.I.	Standard Error	RSE	Relative Bias %
$\rho_{12}^{*} = 0.80$	0.8275	0.8258 , 0.8290	0.002612	0.0032	3.43
$\rho_{13}^{*} = 0.70$	0.7358	0.7340 , 0.7376	0.001944	0.0026	5.12
$\rho_{14}^{*} = 0.65$	0.6959	0.6943 , 0.6976	0.001575	0.0023	7.07
$\rho_{23}^{*} = 0.60$	0.6209	0.6185 , 0.6236	0.002075	0.0033	3.48
$\rho_{24}^{*} = 0.50$	0.5376	0.5354 , 0.5400	0.001595	0.0030	7.52
$\rho_{34}^{*} = 0.45$	0.4677	0.4650 , 0.4700	0.001638	0.0035	3.93

Correlation results for the Percentiles PM, $n = 25$

Parameter	Estimate	95% Bootstrap C.I.	Standard Error	RSE	Relative Bias %
$\xi_{12} = 0.80$	0.8141	0.8123 , 0.8146	0.002007	0.0025	1.76
$\xi_{13} = 0.70$	0.7138	0.7119 , 0.7162	0.002005	0.0028	1.97
$\xi_{14} = 0.65$	0.6658	0.6646 , 0.6685	0.001954	0.0029	2.43
$\xi_{23} = 0.60$	0.6142	0.6115 , 0.6154	0.001719	0.0028	2.37
$\xi_{24} = 0.50$	0.5154	0.5131 , 0.5177	0.001809	0.0035	3.09
$\xi_{34} = 0.45$	0.4646	0.4631 , 0.4685	0.001534	0.0033	3.23

Marginal Results n = 750

Skew (α_3) and Kurtosis (α_4) results for the Conventional PM.

Dist	Parameter	Estimate	95% Bootstrap C.I.	Standard Error	Relative Bias %
1	$\alpha_3 = 0$	2.562	2.5383, 2.5823	0.01117	-26.5
	$\alpha_4 = 25$	22.15	21.6873, 22.6698	0.24850	-81.5
2	$\alpha_3 = 3$	2.180	2.1668, 2.1944	0.00697	-12.3
	$\alpha_4 = 21$	13.36	13.0936, 13.6467	0.14100	-49.7
3	$\alpha_3 = 2$	-0.0051	-0.0265, 0.0163	0.01100	-----
	$\alpha_4 = 7$	18.57	18.2203, 18.9412	0.18330	-53.4
4	$\alpha_3 = 0$	1.54	1.5246, 1.5539	0.00743	-15.8
	$\alpha_4 = 0$	12.91	12.6537, 13.1903	0.13610	-45.0

Left-right tail-weight ratio (γ_3) and tail-weight factor (γ_4)results for Percentiles PM.

Dist	Parameter	Estimate	95% Bootstrap C.I.	Stand. Error	Relative Bias %
1	$\gamma_3 = 1.0000$	1.0000	0.9978, 1.0020	0.001062	--
	$\gamma_4 = 0.3105$	0.3108	0.3105, 0.3112	0.000171	0.11
2	$\gamma_3 = 0.3430$	0.3432	0.3426, 0.3438	0.000308	--
	$\gamma_4 = 0.3868$	0.3873	0.3869, 0.3877	0.000203	0.14
3	$\gamma_3 = 0.4361$	0.4359	0.4353, 0.4364	0.000287	--
	$\gamma_4 = 0.4872$	0.4874	0.4870, 0.4877	0.000189	--
4	$\gamma_3 = 1.0000$	1.0000	0.9991, 1.0014	0.000539	--
	$\gamma_4 = 0.5263$	0.5264	0.5261, 0.5267	0.000159	--

Correlation Results n = 750

Correlation results for the Conventional PM, $n = 750$

Parameter	Estimate	95% Bootstrap C.I.	Standard Error	RSE	Relative Bias %
$\rho_{12}^{*} = 0.80$	0.8012	0.8009 , 0.8018	0.000622	0.0008	0.15
$\rho_{13}^{*} = 0.70$	0.7037	0.7033 , 0.7042	0.000496	0.0007	0.53
$\rho_{14}^{*} = 0.65$	0.6546	0.6542 , 0.6549	0.000330	0.0005	0.71
$\rho_{23}^{*} = 0.60$	0.6007	0.6001 , 0.6012	0.000464	0.0008	0.11
$\rho_{24}^{*} = 0.50$	0.5022	0.5018 , 0.5026	0.000266	0.0005	0.45
$\rho_{34}^{*} = 0.45$	0.4506	0.4501 , 0.4510	0.000271	0.0006	0.12

Correlation results for the Percentiles PM, $n = 750$

Parameter	Estimate	95% Bootstrap C.I.	Standard Error	RSE	Relative Bias %
$\xi_{12} = 0.80$	0.8001	0.8001 , 0.8005	0.000338	0.0004	0.02
$\xi_{13} = 0.70$	0.7004	0.7000 , 0.7007	0.000322	0.0005	0.05
$\xi_{14} = 0.65$	0.6502	0.6499 , 0.6506	0.000303	0.0005	--
$\xi_{23} = 0.60$	0.6002	0.5999 , 0.6006	0.000302	0.0005	--
$\xi_{24} = 0.50$	0.5000	0.4995 , 0.5005	0.000328	0.0007	--
$\xi_{34} = 0.45$	0.4502	0.4497 , 0.4506	0.000293	0.0007	--

Multivariate Percentile PM with Pearson correlation

$$\rho_{jk} = \frac{E\left[p(Z_j)p(Z_k)\right] - m_j m_k}{\sqrt{v_j v_k}}$$

Where

$$E\left[p(Z_j)p(Z_k)\right]$$
$$= \left(c_{j1}(c_{k1} + c_{k3}) + c_{k3}(c_{k1} + c_{k3})\right)$$
$$+ r_{jk}\left(c_{j2}c_{k2} + 3c_{j4}c_{k2} + 3c_{j2}c_{k4} + 9c_{j4}c_{k4}\right)$$
$$+ r_{jk}^2\left(2c_{j3}c_{k3}\right) + r_{jk}^3\left(6c_{j4}c_{k4}\right)$$

$m_j = c_{j1} + c_{j3}$ and $v_j = c_{j2}^2 + 2c_{j3}^3 + 6c_{j2}c_{j4} + 15c_{j4}^2$

Application

- Secondary analysis of data collected in settings for which privacy rules restrict individually identifiable information by law
  - Education: the Family Educational Rights and Privacy Act (FERPA)
  - Healthcare: the Health Insurance Portability and Accountability Act (HIPAA)
- The researcher may be restricted to the use of descriptive distributional statistics commonly released to the public, such as
  - Means
  - Standard deviations
  - Percentiles
  - Correlations

SAS/IML macro %simPPM

- Available for download with Koran and Headrick (2016; open-access)

http://digitalcommons.wayne.edu/jmasm/vol15/iss1/42

- Options
  - Univariate
  - Multivariate: Spearman or Pearson correlations
- Include the following lines to access the macro

```
filename simppm "directory of file simPPM";
%include simppm(simPPM) / nosource2;
```

Preparing to Call %simPPM

Need the following:

1) the number of variables,
2) the file path and name of percentiles file,
3) the file path and name of specified correlation file,
4) an indication of whether the specified correlations are Pearson or Spearman (1 for Pearson, 2 for Spearman),
5) the desired sample size,
6) a random number seed (optional), and
7) the file path and name for output file.

Empirical example 1: Idaho Standards Achievement Test

- mathematics scale scores for 25 third grade students.
- 2011 scale score to percentile rank conversion tables for the ISAT are publicly available
- Macro call:

```
%simPPM(1, C:\SAS\ex1percentiles.txt,
, , 25, 54321, C:\SAS\ex1simdata.txt)
```

  - ex1percentiles.txt file saved in the folder C:\SAS\

external Percentiles File - five rows

1. 10th percentile
2. 25th percentile
3. 50th percentile
4. 75th percentile
5. 90th percentile

- columns = number of variables to be simulated
- ASCII (text) file
- space delimited
- there cannot be any missing values

ex1percentiles.txt

188.5

197.0

205.8

216.3

228.1

Output file

Percentile Power Method coefficients

C

205.8

13.869234

1.5221864

0.9625014

Empirical example 2: General Social Survey

Percentiles and Pearson correlations from $n = 527$ respondents in the 2012 General Social Survey

1. respondent's age (AGE)
2. "Approximately how much money or the cash equivalent of property have you contributed in each of the fields listed in the past 12 months? b. Education" (TOTEDUC)

Simulate 1000 responses to these two survey items

Arrange the 10th, 25th, 50th, 75th, and 90th percentiles for the AGE and TOTEDUC

Percentiles File (ex2percentiles.txt)**:**

27.176 15.500
33.667 25.350
41.444 90.339
48.901 180.737
59.500 600.529

Macro call:

```
%simPPM(2, C:\SAS\ex2percentiles.txt,
C:\SAS\ex2correlations.txt, 1, 1000, 7654321,
C:\SAS\ex2simdata.txt)
```

external Correlation File

Correlations File (ex2correlations.txt)**:**

```
1   .10
.10 1
```

- ASCII (text) file
- space delimited
- as many rows and columns as there are variables to be simulated
- the variables appear in the same order as in the percentiles file
- the correlations arranged in a full symmetric matrix with ones on the diagonal.
- there cannot be any missing values

Output

Percentile Power Method coefficients

C

41.444 90.339

10.78791 71.871855

1.1532084 132.53707

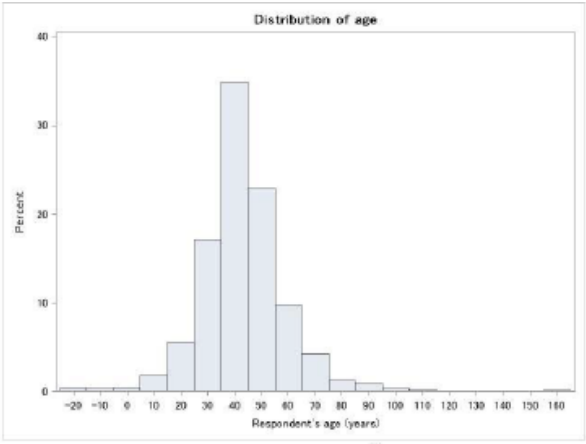
1.1102007 95.214843

Intermediate Pearson correlations

PI

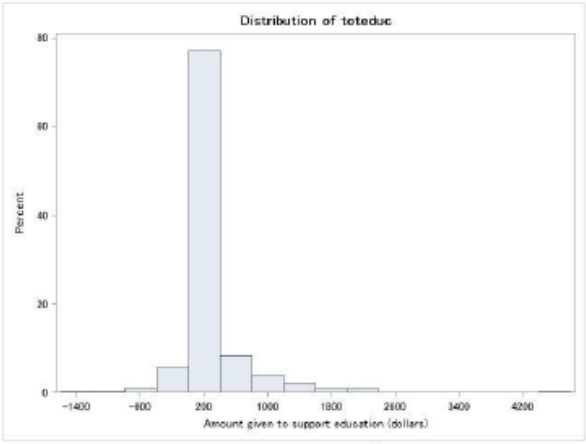
1 0.1322937

0.1322937 1



A

Percentiles	
10th	26.700
25th	34.787
50th	41.984
75th	50.998
90th	60.617



B

Percentiles	
10th	34.588
25th	74.442
50th	95.034
75th	237.418
90th	637.174

The Multivariate Percentile Power Method transformation:

- matches distributions for which conventional skew and kurtosis are unavailable but percentiles are available
- is superior to the multivariate conventional power method in matching nonnormal distributions, especially for small sample size
- has a unique, closed form solution for the polynomial coefficients

open-access download

Journal of Modern Applied Statistical Methods:

http://digitalcommons.wayne.edu/jmasm/vol15/iss1/42

presenter contact information

Dr. Jennifer Koran

jkoran@siu.edu

References

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika, 43,* 521-532. doi: 10.1007/BF02293811

Headrick, T. C. (2010). *Statistical simulation: power method polynomials and other transformations*. Chapman & Hall/CRC.

Karian, Z. A., & Dudewicz, E. J. (2011). *Handbook of fitting statistical distributions with R*. Boca Raton FL: CRC Press.

Koran, J., & Headrick, T.C. (2016). A percentile-based power method in SAS: Simulating multivariate non-normal continuous distributions. *Journal of Modern Applied Statistical Methods, 15*(1). Available from http://digitalcommons.wayne.edu/jmasm/vol15/iss1/42

Koran, J., Headrick, T.C., & Kuo,T.-C. (2015). Simulating univariate and multivariate nonnormal distributions through the method of percentiles. *Multivariate Behavioral Research*, 50, 216-232. doi: 10.1080/00273171.2014.963194

Vale, C. D., & Maurelli, V. A. (1983). Simulating multivariate nonnormal distributions. *Psychometrika*, *48*, 465-471. doi: 10.1007/BF02293687