Defining the General Transformation to Normality: A Proposal to Correlate General Nonnormal Distributions

Charles B. Moss

March 24, 2015

Modeling Correlated Non-Normal Variables

- Moss and Shonkwiler - hyperbolic sine transformation to model nonnormality in corn yields

$$z_{it} = \frac{\ln\left(\theta z_i + \sqrt{\theta^2 z_i^2 + 1}\right)}{\theta}. \tag{1}$$

- Ramirez, Moss and Boggess - same transformation to model correlation among potentially nonnormal random variables
- Both use a generalization of the inverse hyperbolic sine transformation introduced by Burbidge, Magee, and Robb.

Heuristic Example

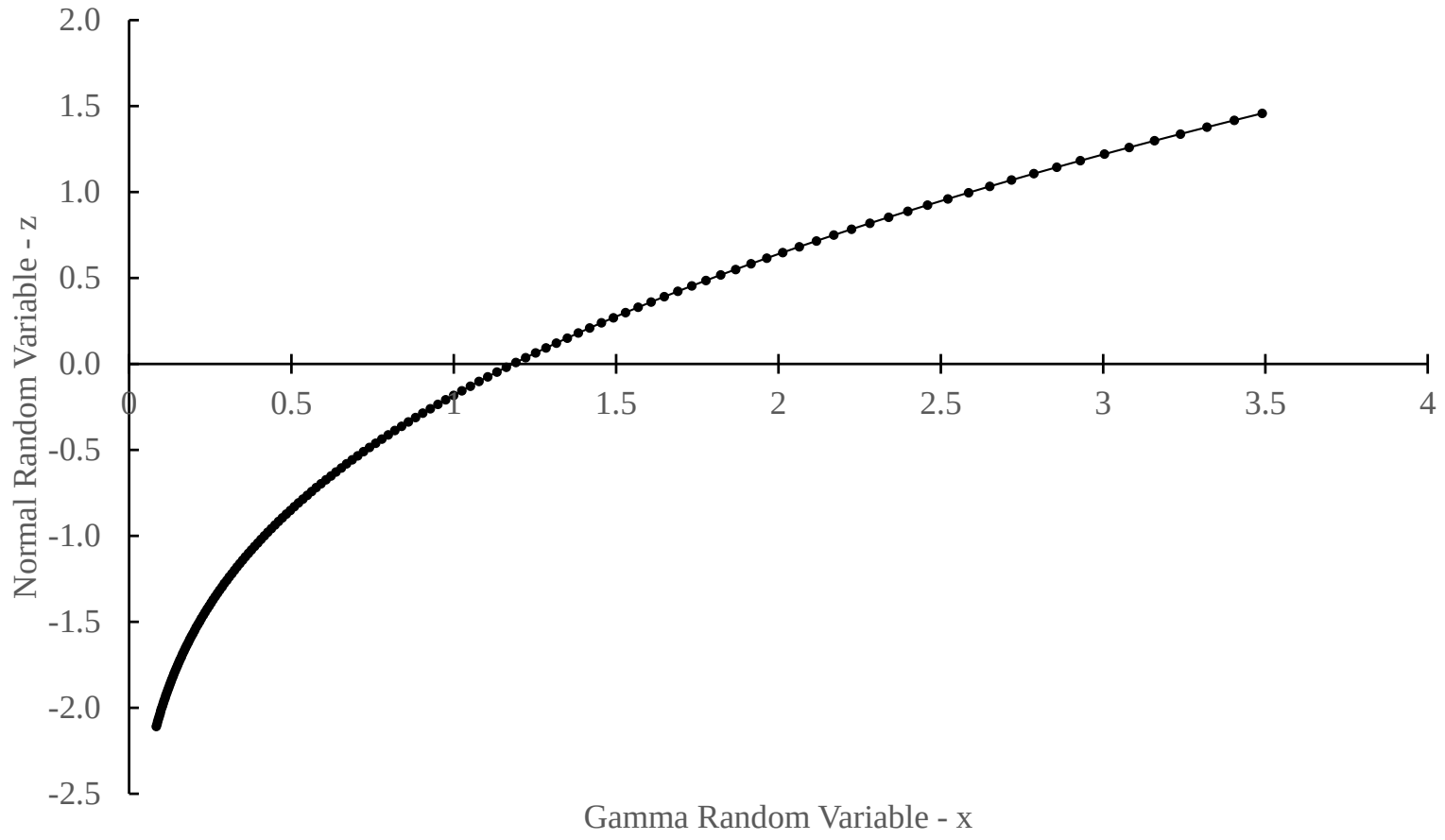
- As an example, assume $x \sim \Gamma[\alpha, \beta]$ there exists a $z \sim \mathrm{N}[0, 1]$ that yields the same probability.
- In this case, assume that $x \sim \Gamma[1.5, 1.0]$
- For any x drawn from this distribution I can define

$$F^* = \int_0^x f(x|\alpha, \beta)\, dx \tag{2}$$

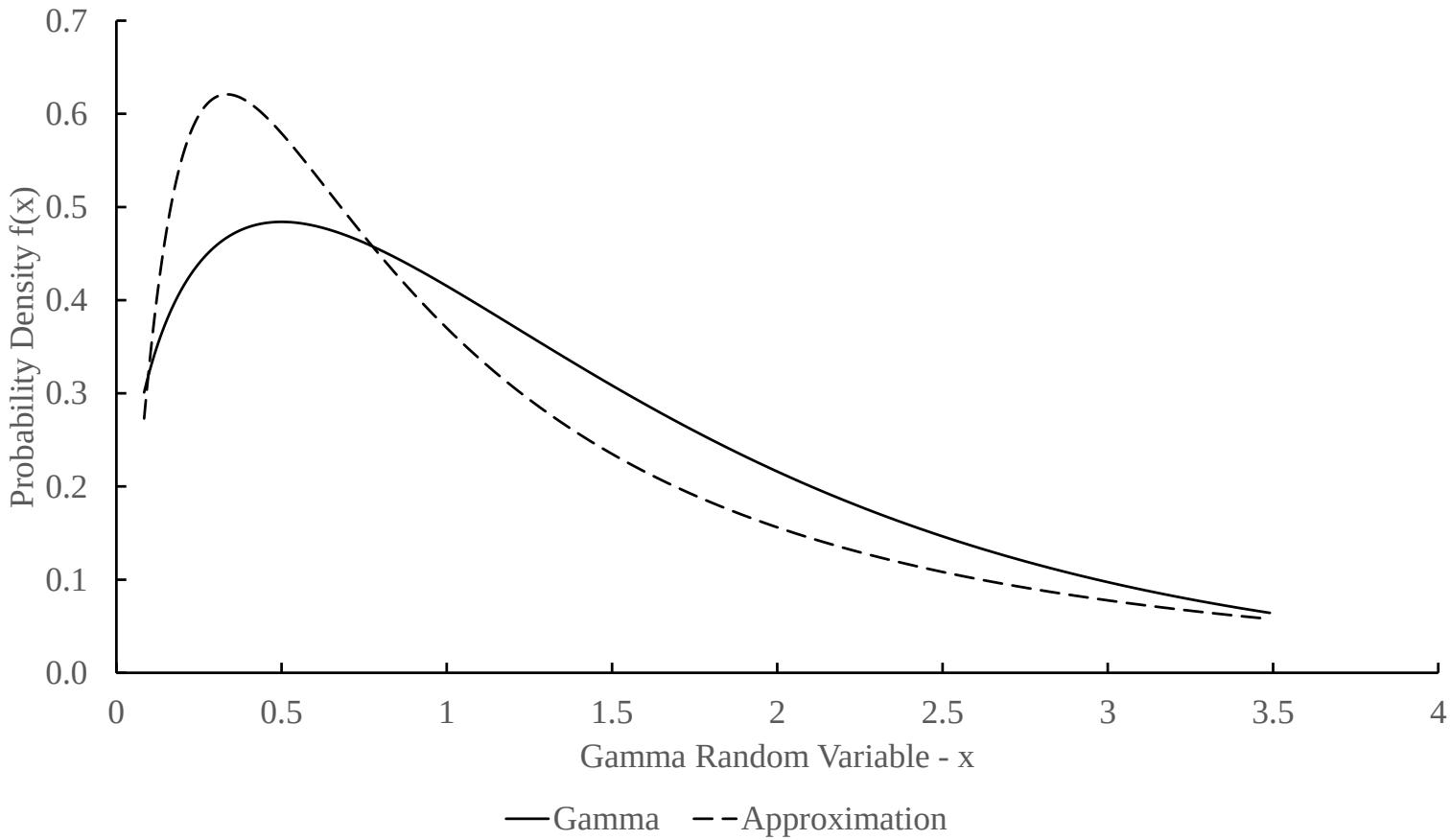
- Using this value it is possible to derive

$$z = G^{-1}(F^*) \ni: G(z) = \int_{-\infty}^{z} g\left(z|\mu, \sigma^2\right) dz \tag{3}$$

General Transformation between Gamma and Normality



Comparison of Actual and Approximated Gamma Distribution



- The next step is to estimate a general monotonic mapping between the variables.
- In this case, a variant of the natural logarithm would seem appropriate

$$z = \gamma_0 + \gamma_1 \ln(x) \tag{4}$$

- Estimating this transformation with ordinary least squares yields $\hat{\gamma}_0 = -0.0590$ (0.0129) and $\hat{\gamma}_1 = -0.9290$ (0.0103).

Empirical Example

	Original Data			Detrended Yields		
Year	Cotton	Soybeans	Potatoes	Cotton	Soybeans	Potatoes
1960	327	26	122	732.444	35.487	257.869
1965	353	26	148	720.903	34.609	271.288
1970	436	28	162	766.362	35.730	272.708
1975	346	24	194	638.821	30.852	292.127
1980	610	22	194	865.280	27.973	279.547
1985	693	26	226	910.739	31.095	298.967
1990	640	19	219	820.197	23.216	279.386
1995	472	26	210	614.656	29.338	257.806
2000	480	19	286	585.115	21.460	321.225
2005	762	32	273	829.574	33.581	295.645
2010	766	30	250	796.033	30.703	260.064
2014	914	43	240	914.000	43.000	240.000

- Using this data, I computed the empirical cumulative density function defined as

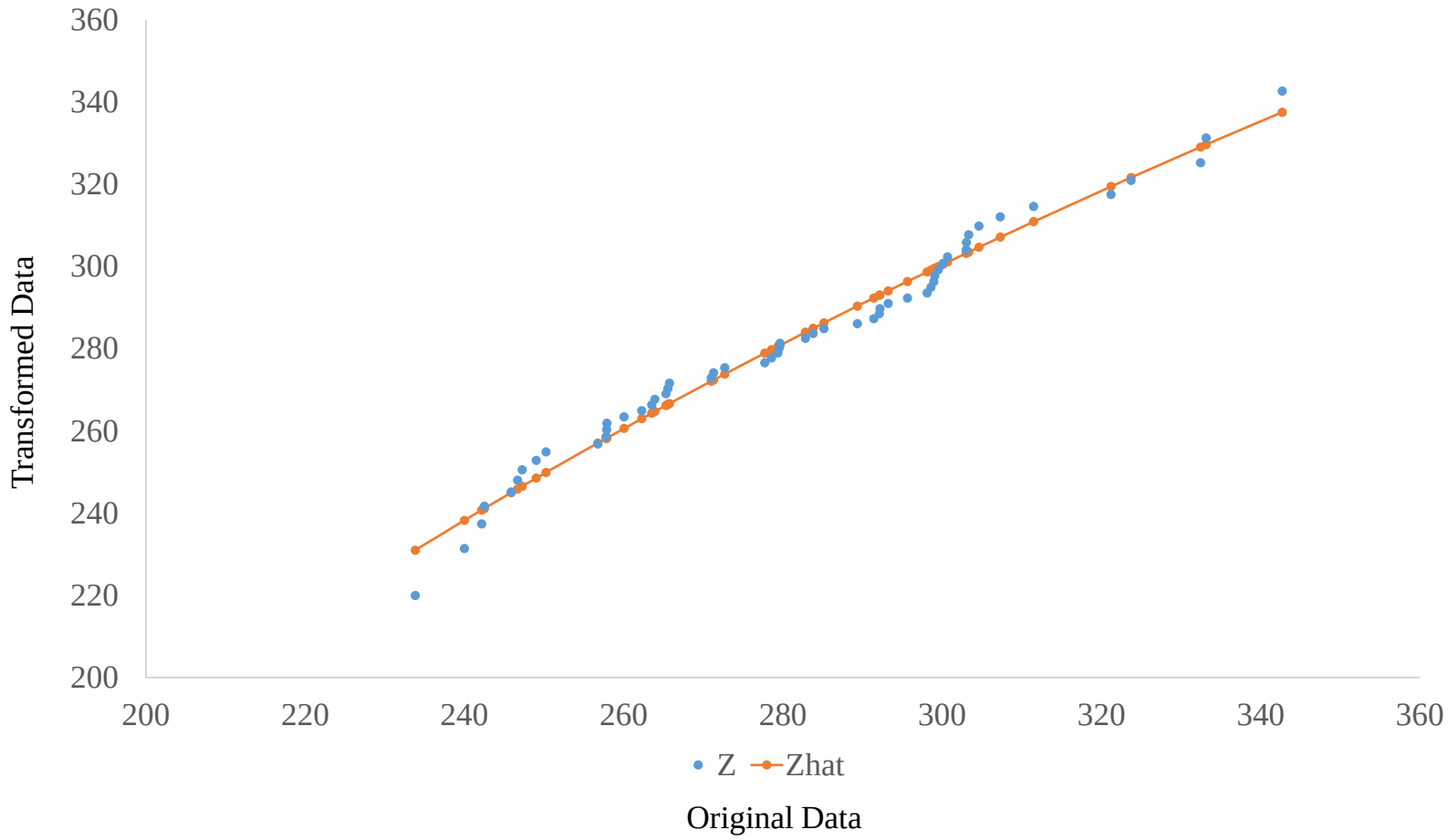
$$\tilde{F}(x_{1i}) = \frac{1}{N} \sum_{x_{1j} \leq x_{1i}} 1. \tag{5}$$

- Following Equation 4 I then compute the value of yield that would give the same cumulative density function value. Unfortunately, none of the data yields a marked depiction from normality.
- taking potatoes as an example, I apply the logarithmic form depicted in Equation 4.
- The result is are the estimates $\hat{\gamma}_0 = -1,289.1$ (278.6739) and $\hat{\gamma}_1 = 278.67.4$ (5.1965).

Transformed Distributions

Cumulative Distribution	Cotton		Soybeans		Potatoes	
	X_1	Z_1	X_2	Z_2	X_3	Z_3
0.009	529.099	534.389	21.460	22.284	233.838	219.955
0.027	585.115	580.394	23.216	24.139	240.000	231.368
0.045	609.132	604.668	25.230	25.117	242.192	237.389
0.064	614.656	622.014	25.811	25.816	242.516	241.692
0.082	628.623	635.833	26.392	26.374	245.870	245.121
0.100	638.821	647.496	27.527	26.844	246.676	248.014
0.118	668.508	657.696	27.919	27.255	247.257	250.544
0.136	676.082	666.838	27.973	27.623	249.032	252.812
0.155	676.936	675.179	27.987	27.960	250.256	254.881
0.173	688.887	682.894	28.379	28.271	256.774	256.795
0.191	692.590	690.107	28.405	28.561	257.806	258.584
⋮	⋮	⋮	⋮	⋮	⋮	⋮
0.991	1072.247	1028.946	43.000	42.221	342.709	342.641

Estimated Transformed Potato Yields



Transformed Potato Yields Paired with Cotton Yields

Year	Cotton	Potatoes	Transformed Potatoes
1960	732.444	257.869	258.223
1965	720.903	271.288	272.360
1970	766.362	272.708	273.815
1975	638.821	292.127	292.984
1980	865.280	279.547	280.718
1985	910.739	298.967	299.434
1990	820.197	279.386	280.557
1995	614.656	257.806	258.155
2000	585.115	321.225	319.445
2005	829.574	295.645	296.320
2010	796.033	260.064	260.585
2014	914.000	240.000	238.211

- The parameters of transformation along with the variance covariance matrix for yields can be estimated using maximum likelihood

$$z_{1i} = x_{1i}$$

$$z_{2i} = x_{2i}$$

$$z_{3i} = \gamma_0 + \gamma_1 \ln\left(x_{3i}\right)$$

$$f\left(z, \gamma_0, \gamma_1, \sigma^2\right) \propto |\Omega|^{-N/2} \prod_{i=1}^{N} \exp\left[-\frac{1}{2}\left(z_i - \mu\right)' \Omega^{-1} \left(z_i - \mu\right)\right] \frac{\gamma_1}{x_{3i}} \tag{6}$$

Discussion

- This paper outlines a generalization of the approach used by Moss and Shonkwiler to model correlated non-normal random variables such as yields.
- I consider a general mapping function with the only restriction that the mapping be positively monotonic.
- I use cotton, soybean and potato data for North Florida. Unfortunately, each of these distributions appear to be normal.