The Pearson Type 4 Algebra

By Olivier Croissant

1 Introduction

The pearson type 4 family of density is characterized by densities parametrized by 4 parameters, following the differential equation:

$$\frac{1}{p}\frac{dp}{dx} = -\frac{a+x}{c_0 + c_1 x + c_2 x^2}$$

such that the equation $c_0 + c_1 x + c_2 x^2 = 0$ has no real solution, that means that the parameters should verify:

$$c_1^2 - 4c_0 c_2 < 0$$

it seems to involves 5 parameters, but the additional condition : $\int_{-\infty}^{\infty} p(x)dx = 1$ imposes a constraints on those 5 parameters that reduces the true parameterization to the one of a 4 parameters family.

2 The explicit density.

Lets rewrite the second order polynomial as;

$$c_0 + c_1 x + c_2 x^2 = C_0 + c_2 (x + C_1)^2$$

the variable transformations involved are then:

$$\begin{cases} C_0 = c_0 - \dfrac{c_1^2}{4 c_2} \\ C_1 = \dfrac{c_1}{2 c_2} \end{cases}$$

And the new differential equation

$$\frac{1}{p}\frac{dp}{dx} = -\frac{a + x}{C_0 + c_2 (x + C_1)^2}$$

That we can rewrite as:

$$\frac{d}{dx}(Log(x)) = \left(-\frac{1}{2\sqrt{c_2}}\right)\frac{2\sqrt{c_2}(x + C_1)}{C_0 + c_2 (x + C_1)^2} + \left(\frac{-a + C_1}{\sqrt{c_2 C_0}}\right)\frac{\sqrt{c_2 / C_0}}{1 + \dfrac{c_2}{C_0}(x + C_1)^2}$$

where we recognize

$$\frac{d}{dx}(Log(x)) = \left(-\frac{1}{2}\right)\frac{U'}{U} + \left(\frac{-a + C_1}{\sqrt{c_2 C_0}}\right)\frac{V'}{1 + V^2}$$

and because we know that

$$\int \frac{U'}{U} = Log(U) \qquad and \qquad \int \frac{V'}{1 + V^2} = ArcTan(V)$$

then we integrate the preceding equation into

$$Log(x) = \left(-\frac{1}{2\sqrt{c_2}}\right)Log(C_0 + c_2(x + C_1)^2) + \left(\frac{-a + C_1}{\sqrt{c_2 C_0}}\right)Tan^{-1}\left(\frac{x + C_1}{\sqrt{C_0 / c_2}}\right) + Cst$$

So the final solution is

$$P(x) = K[C_0 + c_2(x + C_1)^2]^{-(2c_2)^{-1}} Exp\left[-\left(\frac{a - C_1}{\sqrt{c_2 C_0}}\right)Tan^{-1}\left[\frac{x + C_1}{\sqrt{C_0 / c_2}}\right]\right]$$

3 The link with the moments and the cumulants

If we rewrite the preceding equation as

$$\frac{dp}{dx}(c_0 + c_1 x + c_2 x^2) + p(a + x) = 0$$

then by multiplying this equation by x^r and integrating between $-\infty$ and ∞ we get:

$$c_0 \int p' dx + c_1 \int x p' dx + c_2 \int x^2 p' dx + a \int p dx + \int x p dx = 0$$

$$c_0 \int x^r p' dx + c_1 \int x^{r+1} p' dx + c_2 \int x^{r+2} p' dx + a \int x^r p dx + \int x^{r+1} p dx = 0$$

we use the following formulas:

$$\begin{cases} \int_{-\infty}^{\infty} p dx = 1 \\ \int_{-\infty}^{\infty} p' dx = 0 \end{cases} \qquad r \geq 1 \begin{cases} \int_{-\infty}^{\infty} x^r p dx = \mu_r \\ \int_{-\infty}^{\infty} x^r p' dx = -r \int_{-\infty}^{\infty} x^{r-1} p dx = -r\mu_{r-1} \end{cases}$$

and we get

$$\begin{gathered} -c_1 + c_2 \mu_1 + a + \mu_1 = 0 \\ -c_0 r \mu_{r-1} - c_1 (r+1) \mu_r - c_2 (r+2) \mu_{r+1} + a \mu_r + \mu_{r+1} = 0 \end{gathered}$$

we need a sufficient number of equations to determine the unknown parameters s c_0, c_1, c_2. So we need 4 equations;

$$
\begin{aligned}
-c_1 + c_2\mu_1 + a + \mu_1 &= 0 \\
-c_0 - c_1\mu_1 - c_2 3\mu_2 + a\mu_1 + \mu_2 &= 0 \\
-c_0 2\mu_1 - c_1 3\mu_2 - c_2 4\mu_3 + a\mu_2 + \mu_3 &= 0 \\
-c_0 3\mu_2 - c_1 4\mu_3 - c_2 5\mu_4 + a\mu_3 + \mu_4 &= 0
\end{aligned}
$$

we can write the solution as:

$$
\begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ a \end{bmatrix} = \begin{bmatrix} 0 & 1 & -\mu_1 & -1 \\ 1 & \mu_1 & 3\mu_2 & -\mu_1 \\ 2\mu_1 & 3\mu_2 & 4\mu_3 & -\mu_2 \\ 3\mu_2 & 4\mu_3 & 5\mu_4 & -\mu_3 \end{bmatrix}^{-1} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{bmatrix}
$$

a simple case is when the x axis is such that $\mu_1 = 0$:In this case:

$$
\begin{cases} -c_1 + a = 0 \\ -c_0 - c_2 3\mu_2 + \mu_2 = 0 \\ -c_1 3\mu_2 - c_2 4\mu_3 + a\mu_2 + \mu_3 = 0 \\ -c_0 3\mu_2 - c_1 4\mu_3 - c_2 5\mu_4 + a\mu_3 + \mu_4 = 0 \end{cases} \Leftrightarrow \begin{cases} a = c_1 \\ -c_0 - c_2 3\mu_2 + \mu_2 = 0 \\ -c_1 2\mu_2 - c_2 4\mu_3 + \mu_3 = 0 \\ -c_0 3\mu_2 - c_1 3\mu_3 - c_2 5\mu_4 + \mu_4 = 0 \end{cases}
$$

The solution of the preceding equations could be expressed as:

$$c_0 = \mu_2 - \frac{3\mu_2(3\mu_3^2 + 2\mu_2(3\mu_2^2 - \mu_4))}{12\mu_3^2 - 2\mu_2(-9\mu_2^2 + 5\mu_4)}$$

$$c_1 = \frac{\mu_3}{\mu_2}\left\{\frac{1}{2} - \frac{2(3\mu_3^2 + 2\mu_2(3\mu_2^2 - \mu_4))}{12\mu_3^2 - 2\mu_2(-9\mu_2^2 + 5\mu_4)}\right\}$$

$$c_2 = \frac{3\mu_3^2 + 2\mu_2(3\mu_2^2 - \mu_4)}{12\mu_3^2 - 2\mu_2(-9\mu_2^2 + 5\mu_4)}$$

But the initial data are usually the cumulants that are linked to the moments by:

$$Cu_2 = \mu_2$$

$$Cu_3 = \mu_3$$

$$Cu_4 = \mu_4 - 3\mu_2^2$$

and the solution in function of the cumulants is:

$$c_0 = Cu_2 - \frac{3Cu_2(3Cu_3^2 - 2Cu_2Cu_4)}{12Cu_3^2 - 2Cu_2(5Cu_4 + 6Cu_2^2)}$$

$$c_1 = \frac{Cu_3}{Cu_2}\left\{\frac{1}{2} - \frac{2(3Cu_3^2 - 2Cu_2Cu_4)}{12Cu_3^2 - 2Cu_2(5Cu_4 + 6Cu_2^2)}\right\}$$

$$c_2 = \frac{3Cu_3^2 - 2Cu_2Cu_4}{12Cu_3^2 - 2Cu_2(5Cu_4 + 6Cu_2^2)}$$

That could be simplified into:

$$c_0 = Cu_2(1-3c_2)$$

$$c_1 = \frac{Cu_3}{Cu_2}\left\{\frac{1}{2}-2c_2\right\}$$

$$c_2 = \frac{3Cu_3^2-2Cu_2Cu_4}{12Cu_3^2-2Cu_2(5Cu_4+6Cu_2^2)}$$

Let's define variables that will characterize the skew S and the kurtosis K and rename the second cumulants by σ^2 :

$$\sigma^2 = Cu_2 \qquad S = \frac{Cu_3}{(Cu_2)^{3/2}} \qquad and \qquad K = \frac{Cu_4}{(Cu_2)^2}$$

Then the equations are:

$$c_0 = \sigma^2(1-3c_2)$$

$$c_1 = \frac{S\sigma}{2}\{1-4c_2\}$$

$$c_2 = \frac{2K-3S^2}{12+10K-12S^2}$$

4 Link with the normal case

The differential equation in the normal case could be written as

$$\frac{1}{p}\frac{dp}{dx} = -\left(\frac{a+x}{C_0}\right)$$

with the clear identifications:

$$\begin{cases} a = -\mu_1 & c_1 = C_1 = 0 \\ c_0 = C_0 = \sigma & c_2 = 0 \end{cases}$$

the density of the normal distribution, solution of the preceding differential equation is;

$$p_N(x) = \frac{1}{\sqrt{2\pi C_0}} e^{-\frac{(x+a)^2}{2C_0}}$$

If we look at the way the preceding density tends toward the pearson distribution we have to make a taylor development of $Log[p(x)/p_N(x)]$ We will of course do it in the case where $E[x] = 0$ We have the following formulation for C_0, C_1:

The quantity to analyse:

$$Log\left[\frac{p(x)}{p_N(x)}\right] = \frac{x^2}{2C_0} - \frac{1}{2c_2}Log[C_0 + c_2(x+C_1)^2] - \left(\frac{a-C_1}{\sqrt{c_2 C_0}}\right)Tan^{-1}\left[\frac{x+C_1}{\sqrt{C_0/c_2}}\right] + Cst$$

we are not interested in the value of the constant Cst, but only in the dependency of the preceding expression on x. Lets do the following change of variables

$$c_2 = \frac{2K - 3S^2}{12 + 10K - 12S^2}$$

$$C_0 = \sigma^2\left[(1-3c_2) - \frac{\left(\frac{S}{2}\{1-4c_2\}\right)^2}{4c_2}\right] \qquad C_1 = \sigma\frac{\frac{S}{2}\{1-4c_2\}}{2c_2}$$

: $\eta = c_2/C_0 \qquad x = y - C_1$

and we know that $a = 2c_2C_1 = 2\eta C_0C_1$

Then we can put the ratio like

$$Log\left[\frac{p(x)}{p_N(x)e^{Cst}}\right] = \frac{1}{2C_0}\left(y^2 - \frac{Log(1+\eta y^2)}{\eta}\right) + \frac{C_1}{C_0}(1-2\eta C_0)\left(\frac{Tan^{-1}(\sqrt{\eta}y)}{\sqrt{\eta}} - y\right) - 2\eta C_1 y$$

where we can appreciate the convergence of the ratio, knowing that for z small:

$Log(1+z) \approx z \qquad and \qquad Tan^{-1}(z) \approx z$

So now, we can go backward and write an expression of the pearson type 4 density like a perturbation of the normal density:

Where G is a constant that should be determined such that

$$\int_{\infty}^{\infty} p(x)dx = 1$$

Of course, G=1 in the normal case.

In order to normalize the computation it is better to introduce

$$p(x) = \frac{G}{\sqrt{2\pi C_0}}$$

$$e^{-\frac{x^2}{2C_0} + \frac{\left((x+C_1)^2 - \frac{Log(1+\eta(x+C_1)^2)}{\eta}\right)}{2C_0} + \frac{C_1}{C_0}(1-2c_2)\left(\frac{Tan^{-1}(\sqrt{\eta}(x+C_1))}{\sqrt{\eta}} - (x+C_1)\right) - \frac{2c_2C_1}{C_0}x}$$

the abnormal abcisse z defined by: $z = y/(\sqrt{C_0})$, then

$$p(x) =$$

$$\frac{G}{\sqrt{2\pi C_0}} e^{-\frac{x^2}{2C_0} + \frac{\left(z^2 - \frac{Log(1+c_2z^2)}{c_2}\right)}{2} + \frac{C_1}{\sqrt{C_0}}(1-2c_2)\left(\frac{Tan^{-1}(\sqrt{c_2}z)}{\sqrt{c_2}} - z\right) - 2\frac{c_2}{C_0}C_1x}$$

It is now interesting to reintroduce the linear term in y into the square of the normal density to induce a shift of the center of the normal density due to the abnormality:

$$p(x) =$$

$$\frac{G}{\sqrt{2\pi C_0}} e^{-\frac{(x+2c_2C_1)^2}{2C_0} + \frac{\left(z^2 - \frac{Log(1+c_2z^2)}{c_2}\right)}{2} + \frac{C_1}{\sqrt{C_0}}(1-2c_2)\left(\frac{Tan^{-1}(\sqrt{c_2}z)}{\sqrt{c_2}} - z\right)}$$

The nature of the preceding convergence is complex.It is not uniform with S and K.We have to assume that $\eta \to 0$ while C_1 stay bounded which is the case when $S \to 0$ and $K \to 0$ with S/K remaining constant.This is doable in the domain (S,K) accessible to the parameters because the constant term of the inequation described

in the next paragraph of this document is negative. (= -576).So we have found a perturbation shema compatible with an acceptable induced topology on the initial parameters S,K

In order to improve the computability of the preceding formula we will substract from the exponential a quantity that usually is important (value in 0);

$$H = \frac{C_1}{\sqrt{C_0}} \qquad \Xi = \frac{\left(H^2 - \frac{Log(1 + c_2 H^2)}{c_2}\right)}{2} + H(1 - 2c_2)\left(\frac{Tan^{-1}(\sqrt{c_2}H)}{\sqrt{c_2}} - H\right)$$

IF we introduce the factor θ then the formula simplify to:

$$c_2 = \frac{2K - 3S^2}{12 + 10K - 12S^2} \qquad \theta = \frac{S(1 - 4c_2)}{4c_2}$$

$$C_0 = \sigma^2[(1 - c_2(3 + \theta^2))] \qquad C_1 = \sigma\theta$$

5 Limitations induced by the parametrization

The constraint restraining the possible values for the moments is:

$$c_1^2 - 4c_0c_2 < 0 = (Cu_3)^2\left(\frac{1}{4} - c_2\right)^2 < (Cu_2)^3(1 - 3c_2)$$

If we replace $c_2 = \frac{2K - 3S^2}{12 + 10K - 12S^2}$ we get:

$$K^2(S^2-160)+K(-12S^4+12S^3+660S^2-360S+672)+$$
$$36S^6-72S^5-468S^4+360S^3-1116S^2+432S-576<0$$

which are is the limitation to check before attempting the parametrization

6 Case where the 1-moment is not null

The degenerecence $c1 = a$ is due to the constraint $\mu_1 = 0$

The relationship between the bar parameters and the primed parameters re given by the shift

$$x = x' + E[x] \qquad and \qquad p'(x') = p(x)$$

so

$$a' = a + E[x] \qquad c'_0 = c_0 + c_1 E[x] + c_2 (E[x])^2$$
$$c'_1 = c_1 + 2c_2 E[x] \qquad c'_2 = c_2$$

The inverse relation being:

$$a = a' - E[x] \qquad c_0 = c'_0 - c'_1 E[x] + 3c'_2 (E[x])^2$$
$$c_1 = c'_1 - 2c'_2 E[x] \qquad c_2 = c'_2$$

7 Shape of the density

The density have one mode located at $-c_1$. that means that all the weight of the density will be hold by the neighbors of

$$Mode = -c_1 = -\frac{S\sigma}{2}\{1-4c_2\}$$

Of course the second cumulant give us an indication of how far from this mode goes the important values of this density.

8 Summary

Given the Cumulants for a centered variable ($(\mu_1 = 0)$

$$\sigma^2 = Cu_2 = Variance = E[(X - E[X])^2]$$

$$Cu_3 = Skew = E[(X - E[X])^3]$$

$$Cu_4 = Kurtosis = E[(X - E[X])^4] - 3\sigma^4$$

we define:

$$S = \frac{Cu_3}{(Cu_2)^{3/2}} \qquad K = \frac{Cu_4}{(Cu_2)^2} \qquad c_2 = \frac{2K - 3S^2}{12 + 10K - 12S^2} \qquad \theta = \frac{S}{4c_2}\{1 - 4c_2\}$$

and

$$C_1 = \sqrt{Cu_2}\theta \qquad C_0 = Cu_2[1 - c_2[3 + \theta^2]] \qquad z = \frac{x + C_1}{\sqrt{C_0}}$$

$$H = \frac{C_1}{\sqrt{C_0}} \qquad \Xi = \frac{\left(H^2 - \frac{Log(1 + c_2 H^2)}{c_2}\right)}{2} + H(1 - 2c_2)\left(\frac{Tan^{-1}(\sqrt{c_2}H)}{\sqrt{c_2}} - H\right)$$

Then the density is defined by

c_2 defines the abnormality While θ describes the structure of this abnormality.

C_0 simply describe the “modified variance” of the variable x.

$$p(x) = \frac{G}{\sqrt{2\pi C_0}} e^{-\frac{(x+2c_2C_1)^2}{2C_0} + \frac{\left(z^2 - \frac{Log(1+c_2z^2)}{c_2}\right)}{2} + \frac{C_1}{\sqrt{C_0}}(1-2c_2)\left(\frac{Tan^{-1}(\sqrt{c_2}z)}{\sqrt{c_2}} - z\right) - \Xi}$$

z is the abnormal abcisse that should be close to zero to minimize the abnormality in the computation of density.The abnormal abcisse is normalized and should compared directly to1.Its action will intervene through a multiplication by the square root of c_2.

Ξ is a pre-normalization that improve the numerical behavior of any software implementation