How to CUSUM under normal distribution

Pierre Ludmann

June 18, 2014

1 The general formula in off-line view

Set

$$L_k = \ln \left[\frac{\left(\sup_{\theta_0} \prod_{i=1}^{k-1} p_{\theta_0}(y_i)\right) \left(\sup_{\theta_1} \prod_{i=k}^{N} p_{\theta_1}(y_i)\right)}{\sup_{\tilde{\theta}} p_{\tilde{\theta}}(y_i)} \right]$$

where the y_i are the ith observation of N. And

$$L = \max_{1 \leq k \leq N} L_k$$

If

$$L \geq h$$

with h a certain threshold then there is a rupture in

$$t_0 = \arg \max_{1 \leq k \leq N} L_k$$

A interpretation of this whole formula is quite simple : trying to get the best distribution assuming there is a change, if the maximal difference is significant the hypothesis test is passed at the time of maximum.

2 Solve the sup

To define a gaussian distribution, one just requires a mean μ and a standard-deviation σ. Then θ is one either or both.

Whatever the pick, one needs to get three sup : it represents a complex problem to compute. That's why we take advantage from the normal distribution. Let divide a L_k, thanks to the fact that $\ln \sup f = \sup \ln f$, in

$$L_k = \frac{\sup_\theta F_0^k(\theta) \sup_\theta F_1^k(\theta)}{\sup_\theta \tilde{F}^k(\theta)}$$

Actually F_0^k, F_1^k and $\tilde{F}^k$ are roughly the same and similar to

$$F(\theta) = \sum_{i=1}^{n} \ln p_\theta(y_i) = \sum_{i=1}^{n} \left[-\ln(\sigma\sqrt{2\pi}) - \frac{1}{2}\left(\frac{y_i - \mu}{\sigma}\right)^2 \right]$$

$$F(\theta) = -n\ln(\sigma\sqrt{2\pi}) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i - \mu}{\sigma}\right)^2$$

Notice F is coercive whatever θ so a mere differential give the sup :

If $\theta = \mu$ then $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i$

If $\theta = \sigma$ then $\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \mu)^2}$ where μ is a set constant

If $\theta = (\mu, \sigma)$ then $\begin{cases} \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i \\ \hat{\sigma} = \sqrt{\frac{1}{n}\left[\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2\right]} \end{cases}$

3 Simplify L_k expression

Be ready for stories without words *i.e.* formula lines.

3.1 Mean change

$$L_k = \sum_{i=1}^{k-1}\left[-\ln(\sigma\sqrt{2\pi}) - \frac{1}{2}\left(\frac{y_i - \mu_0}{\sigma}\right)^2\right] + \sum_{i=k}^{N}\left[-\ln(\sigma\sqrt{2\pi}) - \frac{1}{2}\left(\frac{y_i - \mu_1}{\sigma}\right)^2\right]$$
$$- \sum_{i=1}^{N}\left[-\ln(\sigma\sqrt{2\pi}) - \frac{1}{2}\left(\frac{y_i - \tilde{\mu}}{\sigma}\right)^2\right]$$

$$L_k = -\frac{1}{2\sigma^2}\left[\sum_{i=1}^{k-1}(y_i - \mu_0)^2 + \sum_{i=k}^{N}(y_i - \mu_1)^2 - \sum_{i=1}^{N}(y_i - \tilde{\mu})^2\right]$$

$$L_k = \frac{1}{2\sigma^2}\left[(k-1)\mu_0^2 + (N-k+1)\mu_1^2 - N\tilde{\mu}^2\right]$$

It is already well-simplified but the abstract factor σ remains so we will compute him a value cause we need to take the hypothesis test.

3.2 Standard-deviation change

$$L_k = \sum_{i=1}^{k-1} \left[-\ln(\sigma_0 \sqrt{2\pi}) - \frac{1}{2} \left(\frac{y_i - \mu}{\sigma_0} \right)^2 \right] + \sum_{i=k}^{N} \left[-\ln(\sigma_1 \sqrt{2\pi}) - \frac{1}{2} \left(\frac{y_i - \mu}{\sigma_1} \right)^2 \right] - \sum_{i=1}^{N} \left[-\ln(\tilde{\sigma} \sqrt{2\pi}) - \frac{1}{2} \left(\frac{y_i - \mu}{\tilde{\sigma}} \right)^2 \right]$$

$$L_k = -\frac{1}{2} \left[\frac{1}{\sigma_0^2} \sum_{i=1}^{k-1} (y_i - \mu)^2 + \frac{1}{\sigma_1^2} \sum_{i=k}^{N} (y_i - \mu)^2 - \frac{1}{\tilde{\sigma}^2} \sum_{i=1}^{N} (y_i - \mu)^2 \right] + N \ln(\tilde{\sigma}) - (k-1) \ln(\sigma_0) - (N-k+1) \ln(\sigma_1)$$

$$L_k = N \ln(\tilde{\sigma}) - (k-1) \ln(\sigma_0) - (N-k+1) \ln(\sigma_1)$$

If σ_0 occurs to be zero when $k = 2$ replace $-(k-1)\ln(\sigma_0)$ with $\frac{1}{2} + \frac{1}{2}\ln(2\pi)$. *Idem* with σ_1 for its term when $k = N$. Else, you are facing a piece of straight line.

3.3 Both change

$$L_k = \sum_{i=1}^{k-1} \left[-\ln(\sigma_0 \sqrt{2\pi}) - \frac{1}{2} \left(\frac{y_i - \mu_0}{\sigma_0} \right)^2 \right] + \sum_{i=k}^{N} \left[-\ln(\sigma_1 \sqrt{2\pi}) - \frac{1}{2} \left(\frac{y_i - \mu_1}{\sigma_1} \right)^2 \right] - \sum_{i=1}^{N} \left[-\ln(\tilde{\sigma} \sqrt{2\pi}) - \frac{1}{2} \left(\frac{y_i - \tilde{\mu}}{\tilde{\sigma}} \right)^2 \right]$$

$$L_k = -\frac{1}{2} \left[\frac{1}{\sigma_0^2} \sum_{i=1}^{k-1} (y_i - \mu_0)^2 + \frac{1}{\sigma_1^2} \sum_{i=k}^{N} (y_i - \mu_1)^2 - \frac{1}{\tilde{\sigma}^2} \sum_{i=1}^{N} (y_i - \tilde{\mu})^2 \right] + N \ln(\tilde{\sigma}) - (k-1) \ln(\sigma_0) - (N-k+1) \ln(\sigma_1)$$

$$L_k = N \ln(\tilde{\sigma}) - (k-1) \ln(\sigma_0) - (N-k+1) \ln(\sigma_1)$$

When $k = 2$ replace $-(k-1)\ln(\sigma_0)$ with $\frac{1}{2} + \frac{1}{2}\ln(2\pi)$. *Idem* when $k = N$ for σ_1 term.

Amazingly, there is no such difference between expecting a change in variance or in variance and mean.

4 When stop

One could procede by recursive dichotomy. So the question is : should I stop when don't reach the threshold? That is to say :

$$\text{If } L_{t_0} < h \text{ then } \max_{1\leq k\leq t_0} L'_k < h \text{ and } \max_{t_0\leq k\leq N} L''_k < h$$

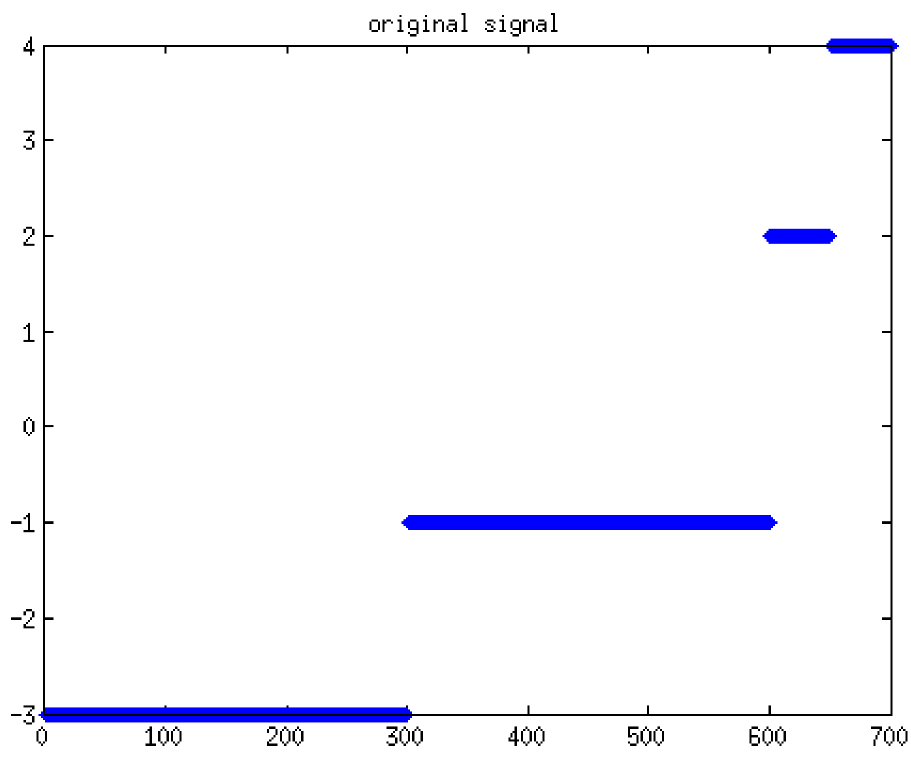
where $t_0 = \arg\max_{1\leq k\leq N} L_k$, L'_k and L''_k is the log-likelihood ratio about each side of t_0.

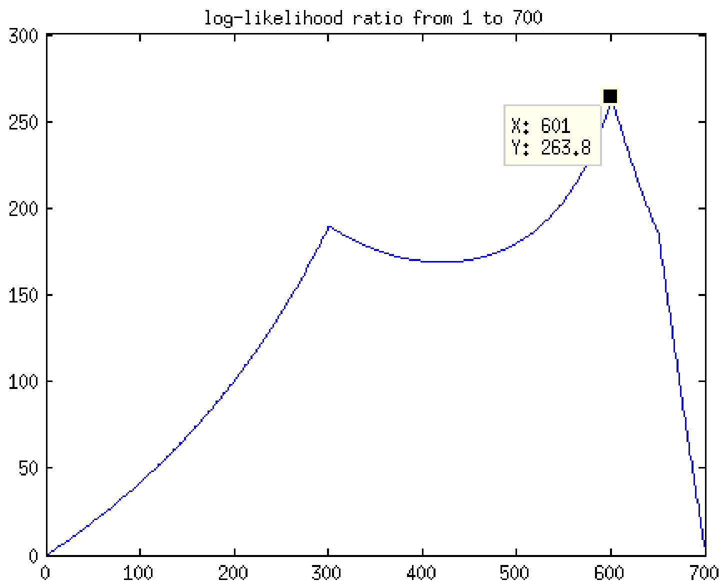
It will be true for any threshold <u>iff</u>

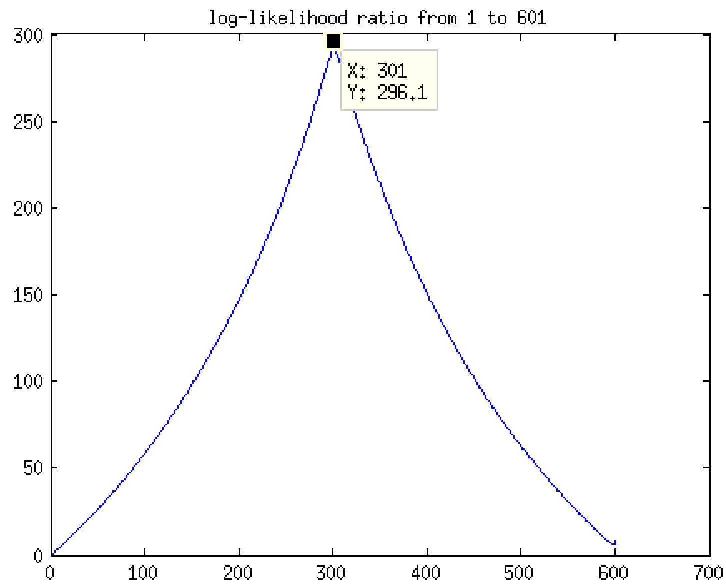
$$L_{t_0} \geq \max_k \left\{L'_k, L''_k\right\}$$

Unfortunately it is merely false according the following signals :

4.1 Mean change







4.2 Standard-deviation change

Take $\mathcal{N}(0, 1.9053 \cdot 10^{-6})$ for the first thousand samples, then $\mathcal{N}(0, 365.41)$ for the following hundred and finally $\mathcal{N}(0, 1.5599 \cdot 10^{-6})$ for the last thousand

4.3 Both change

Idem