Natural Language Processing
Poisson mixtures

Mathematical Statistics

Margot POUPONNEAU
Yangjiawei XUE
Ilyass EL KANSOULI
Lucas MONTEIRO

Introduction

Dataset : https://www.kaggle.com/rtatman/blog-authorship-corpus

We consider 18 documents, each one has between 20,000 and 40,000 words. In each document, we count the number of occurrences of a given word.

Samples are of the form : $\{6, 0, 3, .., 0, 1\}$ meaning that a given word appears 6 times in the 1^{st} document, 0 time in the 2^{nd} document, [...], 1 time in the last document.

We will study the distribution these different given words in the corpus of documents and we will compare their empirical distributions with 4 kinds of Poisson mixtures : Poisson, Negative binomial, 2-Poisson and K-Mixture.

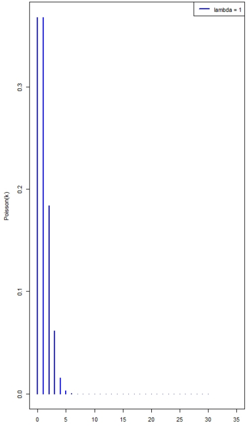
Poisson distribution (Q1)

$\lambda > 0$ and $Supp(P) = \mathbb{N}$

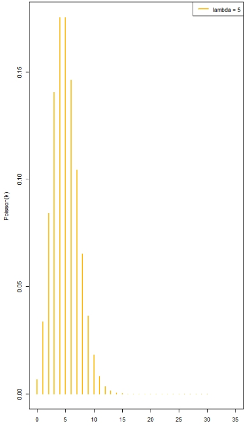
$$\mathbb{P}_P(k) = exp(-\lambda)\frac{\lambda^k}{k!}$$

$\mathbb{E}_P(k) = \lambda$ $\qquad$ $Var_P(k) = \lambda$

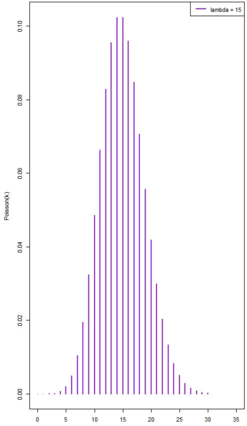
$\lambda = 1$



$\lambda = 5$



$\lambda = 15$

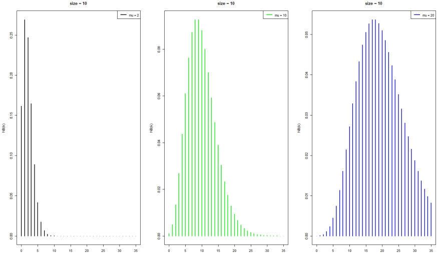


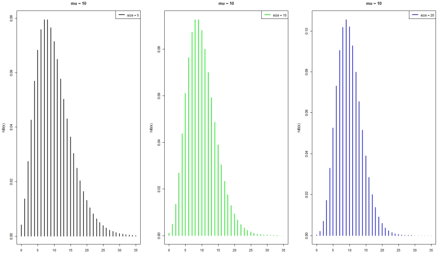
Negative Binomial distribution (Q1)

$P \in [0,1], Q = 1 + P$ and $Supp(NB) = \mathbb{N}$

$\mathbb{P}_{NB}(k) = \binom{N+K-1}{K} P^K Q^{-N-K}$

$\mathbb{E}_{NB}(k) = NP$ $\qquad$ $Var_{NB}(k) = NPQ$





We change the mean

$\mu = 2$

$\mu = 10$

$\mu = 20$

We change the size

N $= 5$

N $= 10$

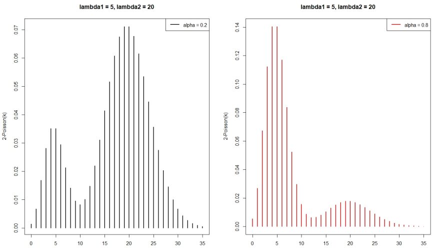
N $= 20$

2-Poisson distribution (Q1)

$\lambda_1 > 0, \lambda_2 > 0$ and $Supp(2P) = \mathbb{N}$

$$\mathbb{P}_{2P}(k) = \alpha exp(-\lambda_1)\frac{\lambda_1^k}{k!} + (1-\alpha)exp(-\lambda_2)\frac{\lambda_2^k}{k!}$$

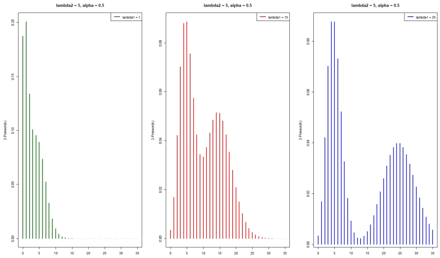
$\mathbb{E}_{2P}(k) = \alpha\lambda_1 + (1-\alpha)\lambda_2$ $\quad\quad$ $Var_{2P}(k) = \alpha^2\lambda_1 + (1-\alpha)^2\lambda_2$



We change α

$\alpha = 0.2$

$\alpha = 0.8$



We change one parameter λ_1

$\lambda_1 = 1$

$\lambda_1 = 15$

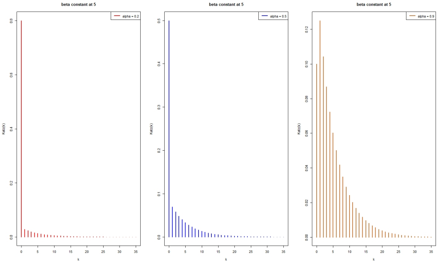
$\lambda_1 = 25$

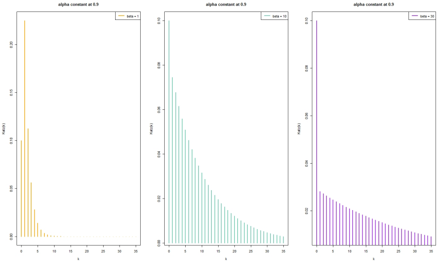
K-Mixture distribution (Q1)

$0 < \alpha < 1, \beta > 0$ and $Supp(K) = \mathbb{N}$

$$\mathrm{P}_K(k) = (1-\alpha)\delta_{k,0} + \left(\frac{\alpha}{\beta+1}\right)\left(\frac{\beta}{\beta+1}\right)^k$$

$$\mathbb{E}_K(k) = \alpha\beta \qquad \sigma_K^2 = \alpha\beta\left[(2-\alpha)\beta+1\right]$$





We change α

$\alpha = 0.2$

$\alpha = 0.5$

$\alpha = 0.9$

We change β

$\beta = 1$

$\beta = 10$

$\beta = 30$

Method of Moments (Q2)

Poisson, NB, K-Mixture

Poisson : $\bar{t} = \hat{\lambda}$

Negative Binomial :

$$\bar{t} = \hat{N}\hat{P} \implies \hat{N} = \frac{\bar{t}}{\hat{P}}$$

$$\sigma_E^2 = \hat{N}\hat{P}(1+\hat{P})$$

$$\implies \sigma_E^2 = \bar{t}(1+\hat{P})$$

$$\implies \hat{P} = \frac{\sigma_E^2}{\bar{t}} - 1$$

K-Mixture :

$$\bar{t} = \hat{\alpha}\hat{\beta} \implies \hat{\alpha} = \frac{\bar{t}}{\hat{\beta}}$$

$$\begin{aligned}
\hat{\sigma}^2_{Katz} &= \hat{\alpha}\hat{\beta}\big[(2-\hat{\alpha})\hat{\beta}+1\big] \\
&= \hat{\alpha}\hat{\beta}\big[2\hat{\beta}-\hat{\alpha}\hat{\beta}\big] + \hat{\alpha}\hat{\beta} \\
&= \bar{t}\big[2\hat{\beta}-\bar{t}\big] + \bar{t} \\
&= 2\bar{t}\hat{\beta} - (\bar{t})^2 + \bar{t}
\end{aligned}$$

$$2\bar{t}\hat{\beta} = \hat{\sigma}^2_{Katz} + (\bar{t})^2 - \bar{t}$$

$$\hat{\beta} = \frac{1}{2\bar{t}}\left(\hat{\sigma}^2_{Katz} + (\bar{t})^2 - \bar{t}\right)$$

Method of Moments (Q2)

2-Poisson

Moment generating function :

$$m(t) = \mathbb{E}_{2P}(e^{tx}) = \sum_{x=0}^{+\infty} e^{tx} \left[\alpha \frac{e^{\lambda_1} \lambda_1^x}{x!} + (1-\alpha) \frac{e^{\lambda_2} \lambda_2^x}{x!} \right]$$

$$= \alpha e^{\lambda_1 (e^t - 1)} + (1-\alpha) e^{\lambda_2 (e^t - 1)}$$

We compute the first 3 derivatives of m and evaluate them at $t = 0$

We find the 3 first theorical moments :

$$R_1 = \alpha\lambda_1 + (1-\alpha)\lambda_2$$

$$R_2 = \alpha(\lambda_1^2 + \lambda_1) + (1-\alpha)(\lambda_2^2 + \lambda_2)$$

$$R_3 = \alpha(\lambda_1^3 + 3\lambda_1^2 + \lambda_1) + (1-\alpha)(\lambda_2^3 + 3\lambda_2^2 + \lambda_2)$$

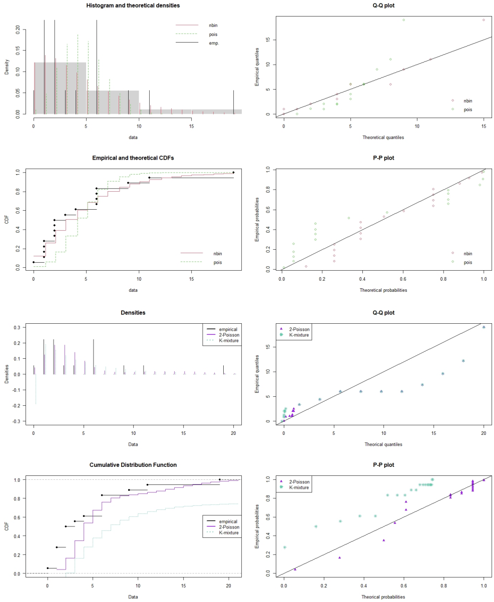
Finally, λ_1 and λ_2 are solutions of $a\lambda^2 + b\lambda + c = 0$

where $a = \hat{R_1}^2 + \hat{R_1} - \hat{R_2}$, $b = \hat{R_1}^2 - \hat{R_1}\hat{R_2} + 2\hat{R_1} - 3\hat{R_2} + \hat{R_3}$,

$c = \hat{R_2}^2 - \hat{R_1}^2 + \hat{R_1}\hat{R_2} - \hat{R_1}\hat{R_3}$ and $\hat{R_1}, \hat{R_2}, \hat{R_3}$ are 3 first empirical moments

AMAZING (Q3)

With QQ-plot, we can see if empirical and theoritical quantiles coincide or not



Poisson : $\hat{\lambda} = 4.56$

Neg-Bin : $\hat{N} = 1.52$, $\hat{\mu} = 4.56$

2-Poisson: $\hat{\lambda}_1 = 13.4$, $\hat{\lambda}_2 = 2.69$, $\hat{\alpha} = 0.17$

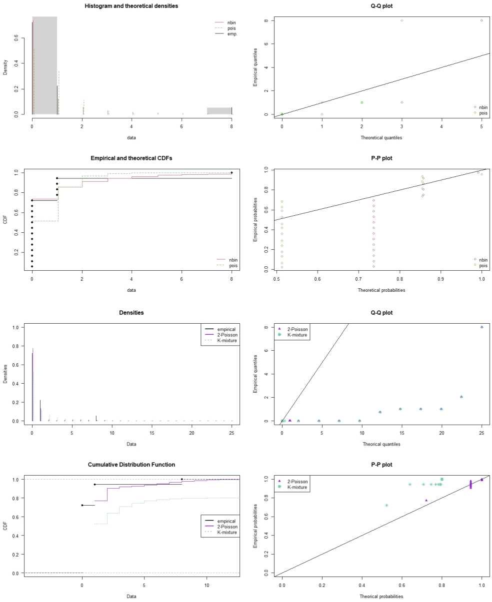
K-Mixture : $\hat{\alpha} = 1.19$, $\hat{\beta} = 3.82$

$\hat{\alpha} \notin]0, 1[$

This is a big problem

$\Rightarrow \mathbb{P}_{K(\hat{\alpha},\hat{\beta})}(x = 0) < 0$!!!

INVESTMENT (Q3)



Poisson : $\hat{\lambda} = 0.67$

Neg-Bin : $\hat{N} = 0.22$, $\hat{\mu} = 0.67$

2-Poisson: $\hat{\lambda}_1 = 0.17$, $\hat{\lambda}_2 = 6$, $\hat{\alpha} = 0.08$

K-Mixture : $\hat{\alpha} = 0.48$, $\hat{\beta} = 1.4$

$\hat{\alpha} \in]0, 1[$

This time it's good

$\implies \mathbb{P}_{K(\hat{\alpha},\hat{\beta})}(x = 0) > 0$

10 words and RMS (Q4)

5 common words : amazing, bad, interesting, love, good

5 specialized words : investment, war, construction, software, animal

$$RMS = \sqrt{\sum_{w \in words} (est_w - obs_w)^2}$$

	Poisson	Neg-Binomial	2-Poisson	K-Mixture
Mean	0	69.12	62.45	48.65
Variance	583.2	639.73	560.79	2987.93
IDF	12.12	43.20	5.96	7.97
Burstiness	5.15	68.03	58.06	48.67
Adaptation	0.22	4.88	0.98	0.26
Entropy	13	66.05	4.17	NaN

Table: Root Mean Square (RMS) for the 10 words considered

We get "NaN" for Entropy of the K-Mixture because the computation of entropy requires $log_2(\mathbb{P}_{K(\hat{\alpha},\hat{\beta})}(x=0))$ but since $\mathbb{P}_{K(\hat{\alpha},\hat{\beta})}(x=0) < 0$ for several words, this computation is impossible

Goodness of fit χ^2 (Q5)

$$\chi^2 = \sum_{w \in words} \frac{(obs_w - est_w)^2}{est_w}$$

	Poisson	Neg-Binomial	2-Poisson	K-Mixture
Amazing	24.92	19.03	23.29	23.23 *
Investment	10.83	7.27	7.34	6.99
Bad	22.29	11.86	9.42	7.02 *
Animal	35.79	33.59	32.64	30.47

Table: χ^2 statistic

* indicates that we remove $x = 0$ (because of negative probability)

Conclusion

When we consider the χ^2 statistic, Negative binomial and 2-Poisson distributions seem to give better results than Poisson distribution.

K-Mixture is special in our examples, for several words, it gives inconsistent results but when results are consistent, K-Mixture seems to be better than 2-Poisson and Negative binomial.

We propose the following inequality to find if K-Mixture will be consistent or not :

$$0 < \alpha < 1 \implies 0 < \frac{\bar{t}}{\beta} < 1 \implies \bar{t} < B_E - 1$$

When this inequality holds, K-Mixture is supposed to give $\alpha \in]0, 1[$