Bayesian multimodeling: Distributions, expectation, likelihood

MIPT

2022

Random variable

Given:

- A set of elementary events Ω
- A sigma-algebra $\mathfrak{F}$ over Ω
- A system $\mathfrak{B}(\mathbb{R})$ of Borel sets over $\mathbb{R}$

A real-valued function $w(\omega) : \Omega \to \mathbb{R}$ is called random variable, whenever for each $B \in \mathfrak{B}(\mathbb{R})$:

$$\{\omega : w(\omega) \in B\} \in \mathfrak{F}.$$

Random variable characteristics

A discrete random variable w takes a countable set of values $A = \{a_1, a_2, \dots\}$ with probabilities $p_1, p_2, \dots, \sum_i p_1 = 1$.

$f(a_i) = p(w = a_i) = p_i$ is a **probability function**.

A continious random variable is set using **cumulative distribution function (CDF)**:

$$F_w(t) = p(w \leq t)$$

or **probability density functon (PDF)**:

$$f(w) : \int_a^b f(w)\, dw = p(a \leq w \leq b).$$

Maximum likelihood

$$\boldsymbol{X} \sim f(\boldsymbol{x}, w),$$
$$L(\boldsymbol{X}, w) = \prod_{x \in \boldsymbol{X}} f(x, w),$$
$$\hat{w} \equiv \arg\max_{w} L(\boldsymbol{X}, w).$$

Logarithm is simple:

$$\log L(\boldsymbol{X}, w) = \sum_{x \in \boldsymbol{X}} \log f(x, w),$$
$$\hat{w} \equiv \arg\max_{w} \log L(\boldsymbol{X}, w).$$

Maximum likelihood variations

Score function:

$$S(w) \equiv \frac{\partial}{\partial w} \log L(w)$$

Maximum likelihood estimation is a solution of the score equation:

$$S(w) = 0$$

Fisher information:

$$I(w) \equiv -\frac{\partial^2}{\partial w^2} \log L(w)$$

MLE dispersion:

$$\mathbb{D}\hat{\theta} \approx I^{-1}(\hat{w})$$

MLE properties

- Consistency:
$$\hat{w}_n \to^p w$$
- Asymptotic normality: $n \to \infty$
$$\hat{w} \sim \mathcal{N}\left(w, I^{-1}(w)\right)$$
- Efficiency: MLE's dispersion is minimal over all the unbiased estimations
- Invariance: $g\left(\hat{w}\right)$ is MLE for $g\left(w\right)$

Likelihood maximization

Likelihood maximization is a KL divergence mimimization:

$$\max_w L(\boldsymbol{X}, w) \iff \min KL\left(p^*\left(\boldsymbol{X}\right) | p\left(\boldsymbol{X}|w\right)\right).$$

Proof sketch

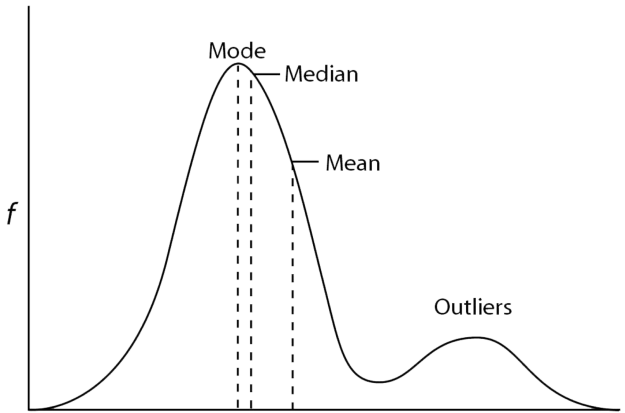
$$KL\left(p^*\left(\boldsymbol{X}\right) | p\left(\boldsymbol{X}|w\right)\right) = \mathrm{E}_{x\sim p^*(\boldsymbol{X})} \log\left(\frac{p^*\left(\boldsymbol{X}\right)}{p\left(\boldsymbol{X}|w\right))}\right) =$$

$$= \mathrm{Const} - \mathrm{E}_{x\sim p^*(\boldsymbol{X})} \log p\left(\boldsymbol{X}|w\right) \approx^{\text{Law of large numbers}}$$

$$\approx \mathrm{Const} - L(\boldsymbol{X}, w).$$

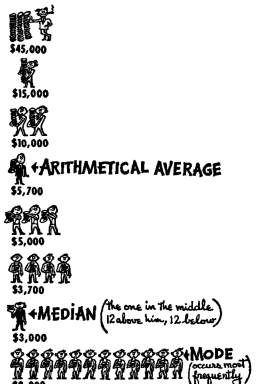
Central tendency

Empirical mean is an arithmetic mean for the given data.
Empirical median is a central element in the variational series.
Empirical mode is the most popular element in the data.



Central tendency

(Huff, 1954):



Median

- α-quantile, $\alpha \in (0, 1)$:

$$w_\alpha: \quad p(w \leq w_\alpha) \geq \alpha, \quad p(w \geq w_\alpha) \geq 1 - \alpha$$

  equivalent:

$$w_\alpha = F^{-1}(\alpha) = \inf\{w: F(w) \geq \alpha\}$$

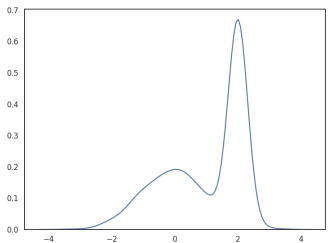
- **A median is a** 0.5-quantile 0.5, the central value of the distribution:

$$\text{median}(w): \quad p(w \leq \text{median}(w)) \geq 0.5, \quad p(w \geq \text{median}(w)) \geq 0.5$$

Mode

A mode is a maximum of the CDF or PDF:

$$\text{mode}(w) = \arg\max_{w} f(w)$$



Distribution can have multiple modes:

$$w \sim \alpha_1 \mathcal{N}(\mu_1, \sigma_1^2) + \alpha_1 \mathcal{N}(\mu_2, \sigma_2^2).$$

Expectation

Average value of the random variable w:

$$\mathrm{E}w = \int w \, dF(w).$$

- can be undetermined;
- linear;
- does not depend on the 0-measure values;
- Law of large numbers:
$$\bar{w}_n \to_{n\to\infty} \mathrm{E}w;$$
- Central Limit Theorem:
$$\sqrt{n}\frac{\bar{w}_n - \mathrm{E}w}{\sqrt{\mathrm{D}w}} \to \mathcal{N}(0,1).$$

Expectation: convergence

Often we need to investigate the convergence of the random variable, which is constructed as a composition of functions:

$$w \to^{\text{d/p/a.s.}} w^*,$$

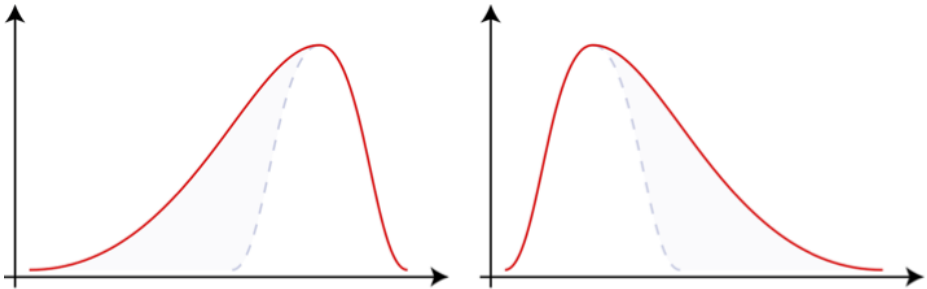
$$\mathrm{E}f(w) \to \mathrm{E}f(w^*)?$$

- f is continuous and limited, then $\mathrm{E}f(w) \to^{d} \mathrm{E}f(w^*)$.
- f is continuous almost everywhere, then $f(w) \to^{\text{d/p/a.s.}} f(w^*)$ (Mann–Wald theorem).
- A.s. convergence: Lebesgue theorem (swap limit and expectation).

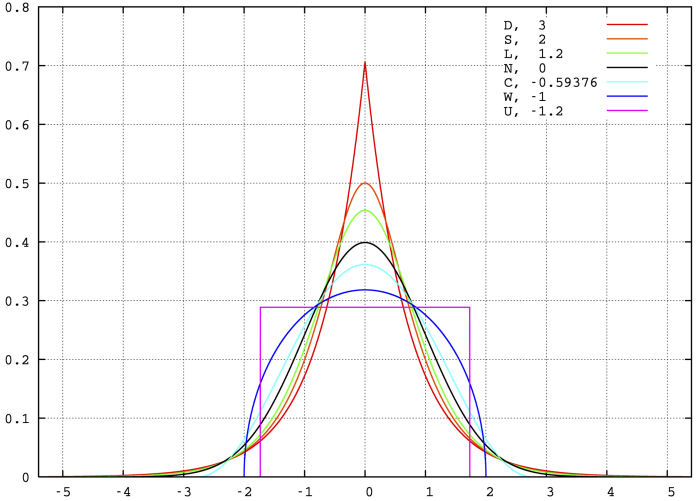
Other moments

- Second moment $\mathrm{D}w = \mathrm{E}(w - \mathrm{E}w)^2$: dispersion.
- Third moment $\frac{\mathrm{E}(w-\mathrm{E}w)^3}{\mathrm{D}w^{3/2}}$: asymmetry coefficient.
- Forth moment $\frac{\mathrm{E}(w-\mathrm{E}w)^4}{\mathrm{D}w^2} - 3$: Excess coefficient, pointedness of the PDF.

Other moments

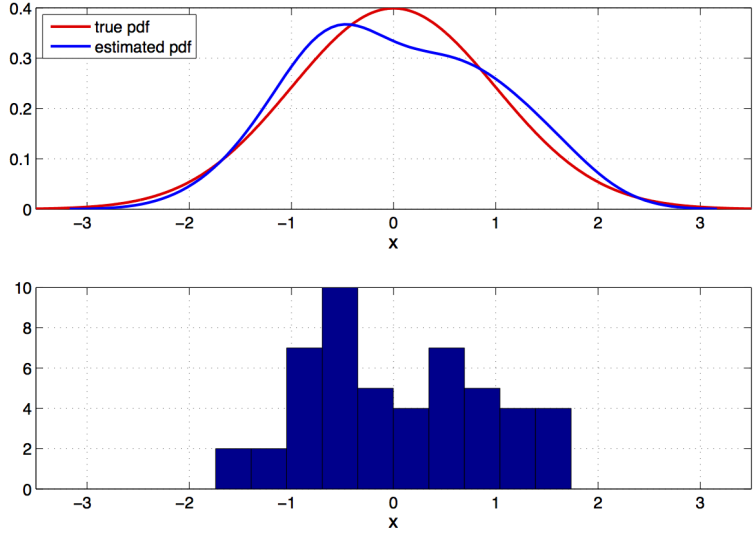


Other moments



(wikipedia)

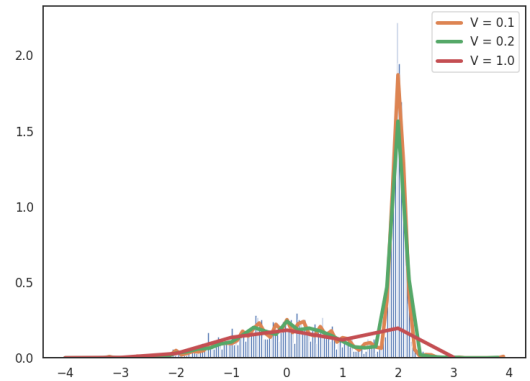
PDF estimation



KDE

Estimation using kernel function k:

$$p(x) = \frac{1}{N} \sum_{x_i \in \boldsymbol{X}} \frac{1}{h} K \left(\frac{x - x_i}{h} \right)$$

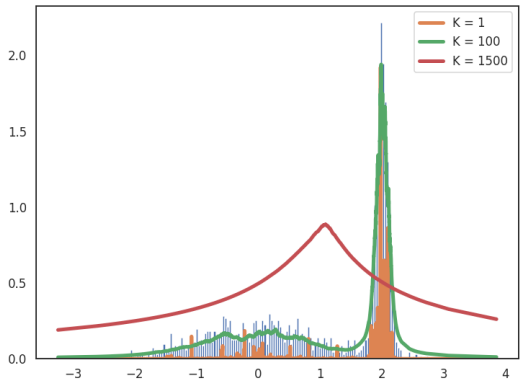


KDE

KNN can be considered as a generalization of KDE:

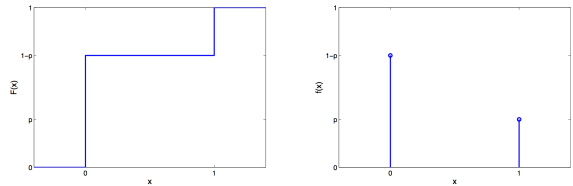
$$p(x) = \frac{1}{N} \sum_{x_i \in \boldsymbol{X}} \frac{1}{h} K\left(\frac{x - x_i}{h_i}\right),$$

$$h_i = \text{dist}\left(x_i, \text{neighbour}\left(x_i, K\right)\right)$$



Bernoulli distribution

$w \in \{0,1\} \sim Ber\,(p)\,, \quad p \in (0,1)$



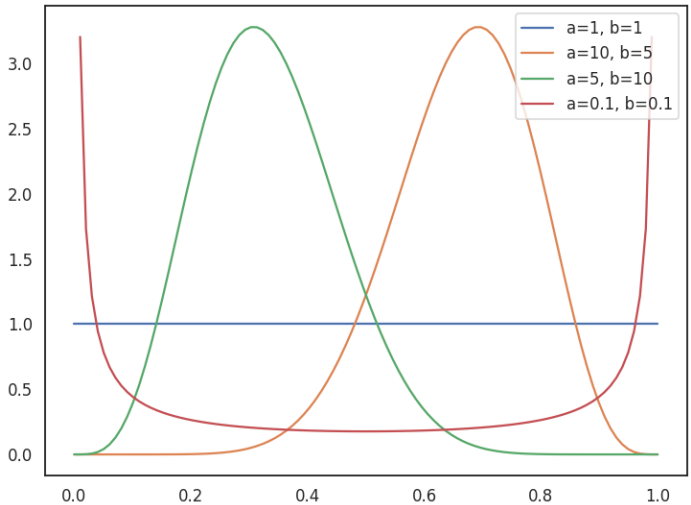
$$F\,(w) = \begin{cases} 0, & w < 0, \\ 1-p, & 0 \le w < 1, \\ 1, & w \ge 1. \end{cases}$$

$$f\,(w) = \begin{cases} 1-p, & w = 0, \\ p, & w = 1. \end{cases}$$

- example: coin flipping

Beta-distribution

$$p(w) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} w^{a-1}(1-w)^{b-1}$$



Beta-distribution

- corresponds to the *prior* beliefs about Bernoulli distribution
- interpretation: "effective number of events $w = 1, w = 0$"
- With $n \to \infty$ converges to δ-distribtuion with PDF concentration at MLE for Bernoulli distribution.

Multinomial distribution

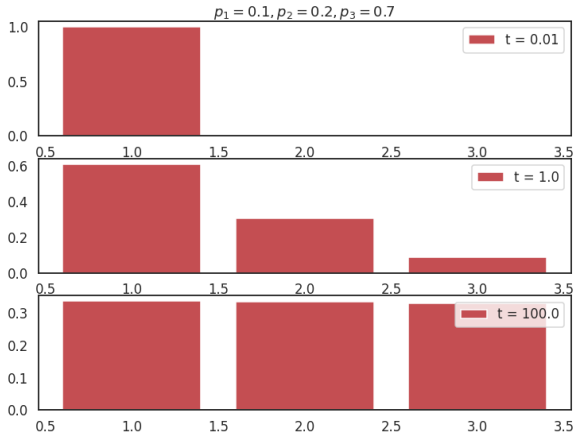
- Generalization of Bernoulli distribution for a larger number of events:

$$p(w = w_i | p_1, \ldots, p_n) = p_i.$$

- Can be parametrized using softmax:

$$\mathbf{softmax}(\log \boldsymbol{p}, t) = \frac{\mathbf{exp}(-\mathbf{log}\boldsymbol{p} \cdot t^{-1})}{\sum_{i=1}^{n} \exp(-logp_i \cdot t^{-1})}$$

Multinomial distribution



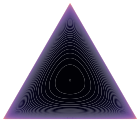
Dirichlet distribution

$$p(w_1, \ldots, w_K, \boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{K} w_i^{\alpha_i - 1}$$

- Generalization of Beta-distribution for a larger number of events
- Support: $K-1$-simplex:
$$w_i \geq 0, \sum w_i = 1.$$
- Interpretation: a probability of each of K mutually exclusive events equals to w_i given that each event was observed $\alpha_i - 1$ times
- Can be parametrized using the following expression:

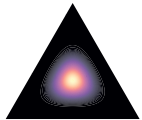
$$\boldsymbol{\alpha} = \bar{\boldsymbol{\alpha}} \cdot t, ||\boldsymbol{\alpha}|| = 1.$$

Dirichlet distribution

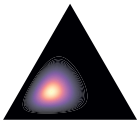


$\bar{\boldsymbol{\alpha}} = [1, 1, 1], t = 0.9$

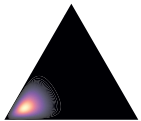
$t = 1.0$



$t = 10.0$



$\bar{\boldsymbol{\alpha}} = [0.5, 0.25, 0.25], t = 30.0$

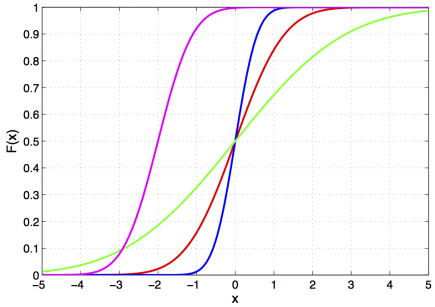


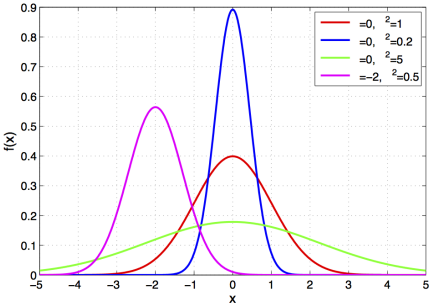
$[0.75, 0.125, 0.125]$

$[0.9, 0.05, 0.05]$

Normal distribution

$w \in \mathbb{R} \sim \mathcal{N}\left(\mu, \sigma^2\right),\ \sigma^2 > 0$





$$F\left(w\right) = \Phi\left(\frac{w-\mu}{\sigma}\right)$$

$$f\left(w\right) = \frac{1}{\sigma}\phi\left(\frac{w-\mu}{\sigma}\right)$$

$$\Phi\left(w\right) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{w} e^{-\frac{t^2}{2}}\,dt$$

$$\phi\left(w\right) = \frac{1}{\sqrt{2\pi}}e^{-\frac{w^2}{2}}$$

Normal distribution

- limit of the sum of the weakly dependent random variables.
- $\mathrm{E}w = \mathrm{median}(w) = \mathrm{mode}(w) = \mu$, $\mathrm{D}w = \sigma^2$, the higher moments equal to zero
- if $w_1, \ldots, w_n$ are independent, $w_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right)$, then $\forall a_1, \ldots, a_n$

$$\sum_{i=1}^{n} a_i w_i \sim \mathcal{N}\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right)$$

- Normal distribution has the maximal differential entropy among all the continuous distribtuions with limited dispersion.
- example: estimation error

Student distribution

- $Ew = 0$ for $\nu > 1$, $\text{median}(w) = \text{mode}(w) = 0$ always
- Let $Z \sim N(0,1)$ and $V \sim \chi^2_\nu$ are independent, then

$$T = \frac{Z}{\sqrt{V/\nu}} \sim St(\nu)$$

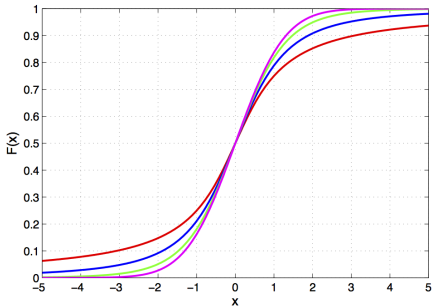
- if $w \sim St(\nu)$, then

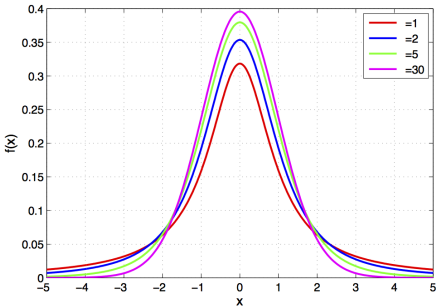
$$Y = \lim_{\nu \to \infty} w \sim \mathcal{N}(0,1)$$

- can be met during empirical mean estimation

Student distribution

$X \in \mathbb{R} \sim St(\nu),\ \nu > 0$





$$F(x) = \frac{1}{2} + x\Gamma\left(\frac{\nu+1}{2}\right)$$

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

Periodical distribution

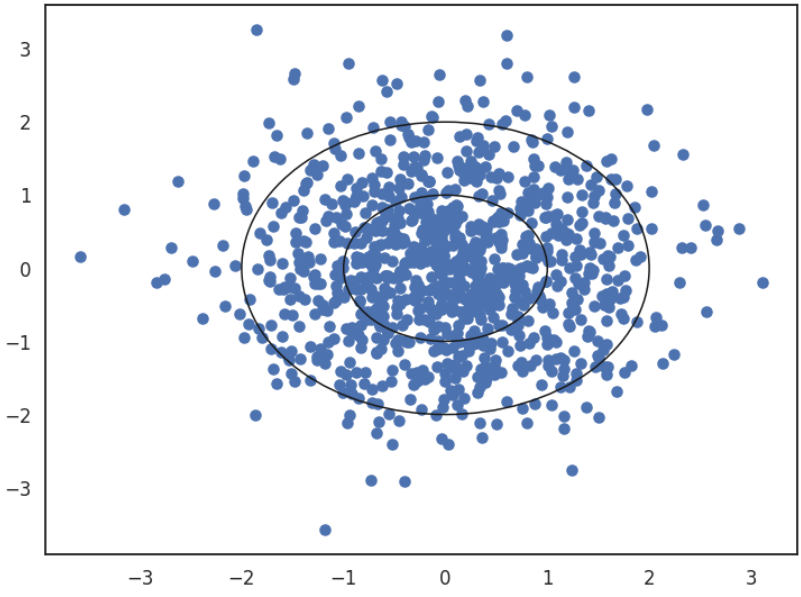
Periodical distribution

Naive approach: use a non-periodical distribution (let's say, Normal?)

Periodical distribution

- Naive approach: use a non-periodical distribution (let's say, Normal?)
  - What if our dataset contains two objects on the circle: (1 degree and 359 degree)?
  - $Ew = 180$, $D = 179$

Periodical distribution



Periodical distribution

- von Mises distribution:

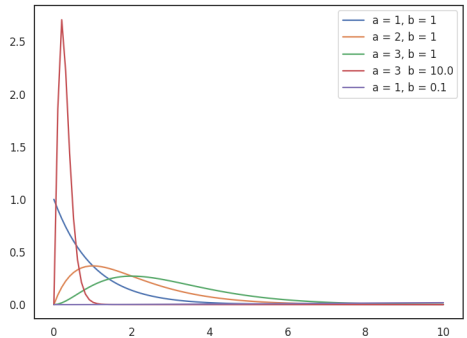
$$p = \frac{\exp(k\cos(x-\mu))}{2\pi I_0(k)}, \quad I_0 \text{is a Bessel function.}$$

- Distribution idea:
  - $[w_1, w_2] \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$
  - Parameterize using polar coordinate system, limit the radius:

$$w_1 = r\cos\phi, w_2 = r\sin\phi, r = \text{Const.}$$

Gamma distribution

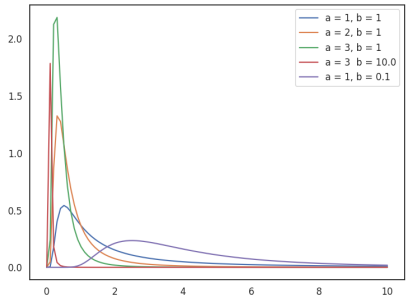
$$p(w, \alpha, \beta) = \frac{\beta^{\alpha} w^{\alpha-1} e^{-\beta}}{\Gamma(\alpha)}$$



- Can be used for prior distribution modeling, for the values inverse to standard deviation.

Inverse Gamma-distribution

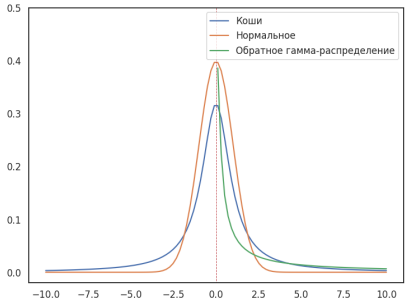
$$p(w, \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} w^{-\alpha-1} \exp(-\beta \cdot w^{-1}).$$



- $w \sim \Gamma(\alpha, \beta) \rightarrow w^{-1} \sim \text{Inv} - \Gamma(\alpha, \beta^{-1})$
- Can be used for prior distribution modeling, for the standard deviation for example.

Cauchy distribution

$$p(w, w_0, \gamma) = \frac{\gamma}{\pi\left((w - w_0)^2 + \gamma^2\right)}.$$



- Alternative to Normal distribution and Inverse Gamma-distribution
- Heavy tails
- The moments are undetermined

References

- Ширяев А. Н. Вероятность-1. – МЦНМО, 2007. – С. 552-552.
- Bishop C. M. Pattern recognition //Machine learning. – 2006. – Т. 128. – №. 9.
- KL minimization: https://wiseodd.github.io/techblog/2017/01/26/kl-mle/
- `Numpy`
- `Scipy`
- `JAX!`