

Bayesian multimodeling: Distributions, expectation, likelihood

MIPT

2023

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 $w(\omega) : \Omega \rightarrow \mathbb{R}$ 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 , $\sum_i p_i = 1$.

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

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

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

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

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

Maximum likelihood

$$X \sim f(x, w),$$

$$L(X, w) = \prod_{x \in X} f(x, w),$$

$$\hat{w} \equiv \arg \max_w L(X, w).$$

We can also use a logarithmic form:

$$\log L(X, w) = \sum_{x \in X} \log f(x, w),$$

$$\hat{w} \equiv \arg \max_w \log L(X, w).$$

Maximum likelihood variations

Score function:

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

Maximum likelihood estimator 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:

$$\hat{w}_n \xrightarrow{P} w$$

- Asymptotic normality: $n \rightarrow \infty$

$$\hat{w} \sim \mathcal{N}(w, I^{-1}(w))$$

- Efficiency: MLE's dispersion is minimal over all the unbiased estimations
- Invariance: $g(\hat{w})$ is MLE over $g(w)$

Likelihood maximization

Likelihood maximization is a KL divergence minimization:

$$\max_w L(X, w) \iff \min KL(p^*(X) | p(X|w)).$$

Proof sketch

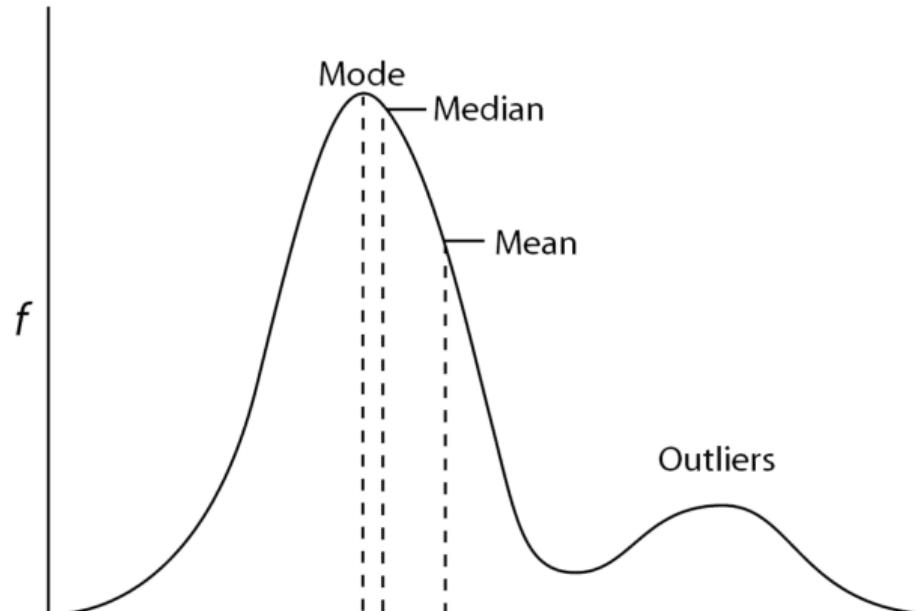
$$\begin{aligned} KL(p^*(X) | p(X|w)) &= E_{x \sim p^*(X)} \log \left(\frac{p^*(X)}{p(X|w)} \right) = \\ &= \text{Const} - E_{x \sim p^*(X)} \log p(X|w) \approx^{3\text{E4}} \\ &\approx \text{Const} - L(X, w). \end{aligned}$$

Central tendency

Empirical mean is an arithmetic mean for the given data.

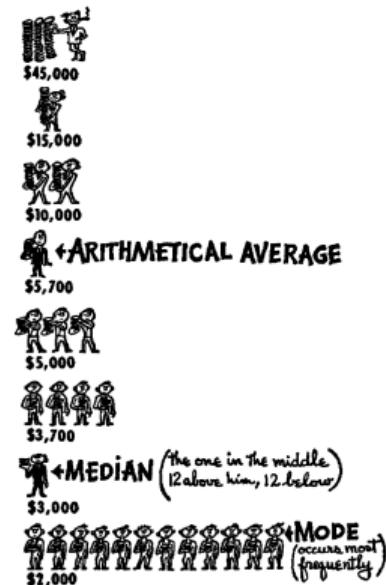
Empirical median is a central element in the variational series.

Empirical mode is the most popular element in the data



Central tendency

(Huff, 1954):



Median

- α -quantile:

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

or, equivalently:

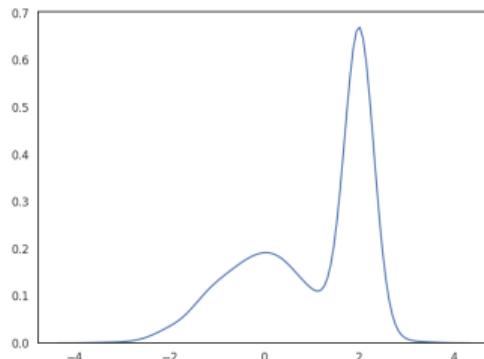
$$w_\alpha = F^{-1}(\alpha) = \inf\{w: F(w) \geq \alpha\}$$

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

Mode

A mode is a maximum of the CDF or PDF:

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



Distribution can have multiple modes:

$$w \sim \alpha_1 \mathcal{N}(\mu_1, \sigma_1^2) + \alpha_2 \mathcal{N}(\mu_2, \sigma_2^2).$$

Expectation

Average value of the random variable w :

$$Ew = \int w dF(w).$$

- can be undetermined;
- linear;
- does not depend on the 0-measure values;
- Law of large numbers:

$$\bar{w}_n \rightarrow_{n \rightarrow \infty} Ew;$$

- Central Limit Theorem:

$$\sqrt{n} \frac{\bar{w}_n - Ew}{\sqrt{Dw}} \rightarrow \mathcal{N}(0, 1).$$

Expectation: convergence

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

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

$$Ef(w) \rightarrow Ef(w^*)?$$

- f is continuous and limited, then $Ef(w) \xrightarrow{d} Ef(w^*)$.
- f is continuous almost everywhere, then $f(w) \xrightarrow{d/p/n.h.} f(w^*)$ (Mann–Wald theorem).
- A.s. convergence: Lebesgue theorem (swap limit and expectation).

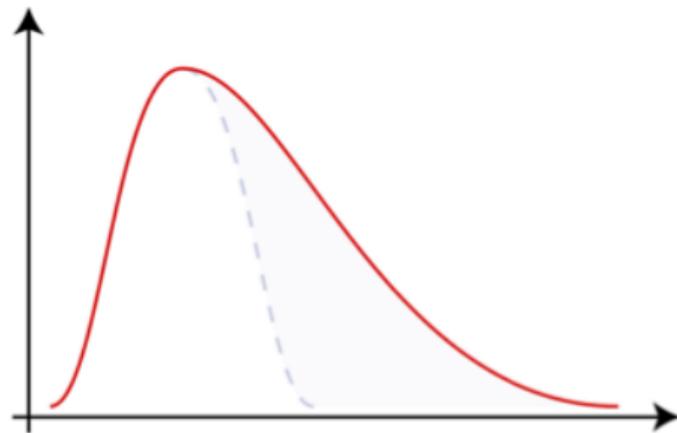
Other moments

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

Other moments

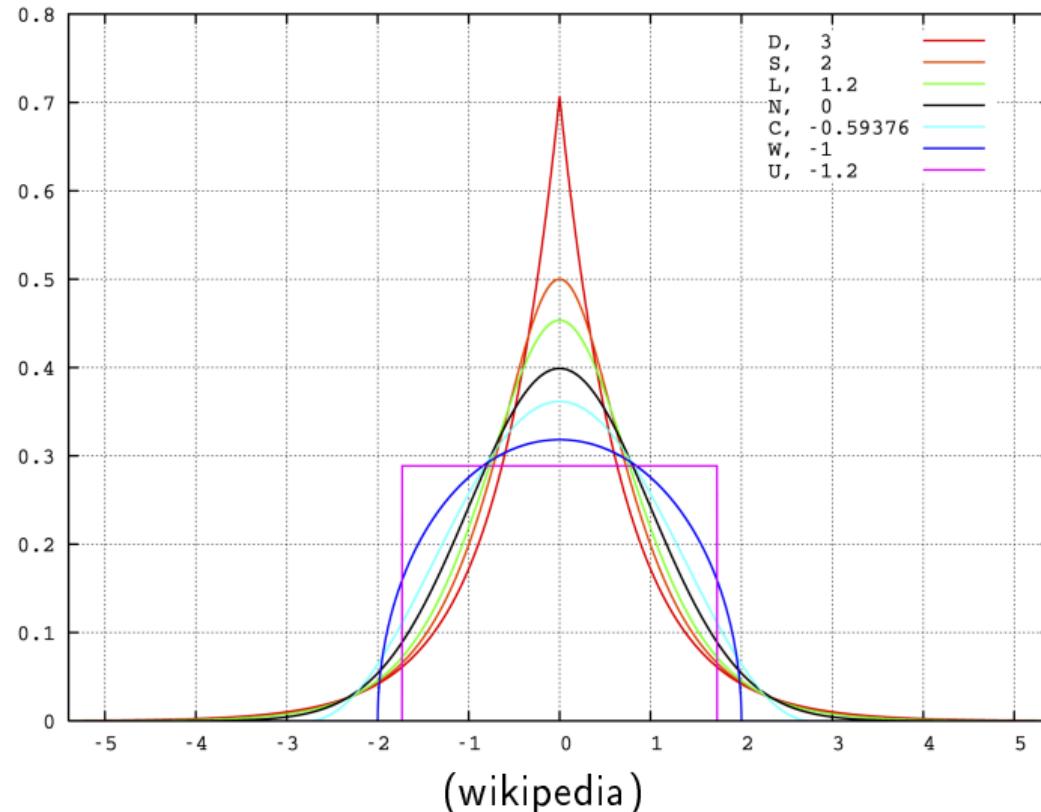


Negative Skew

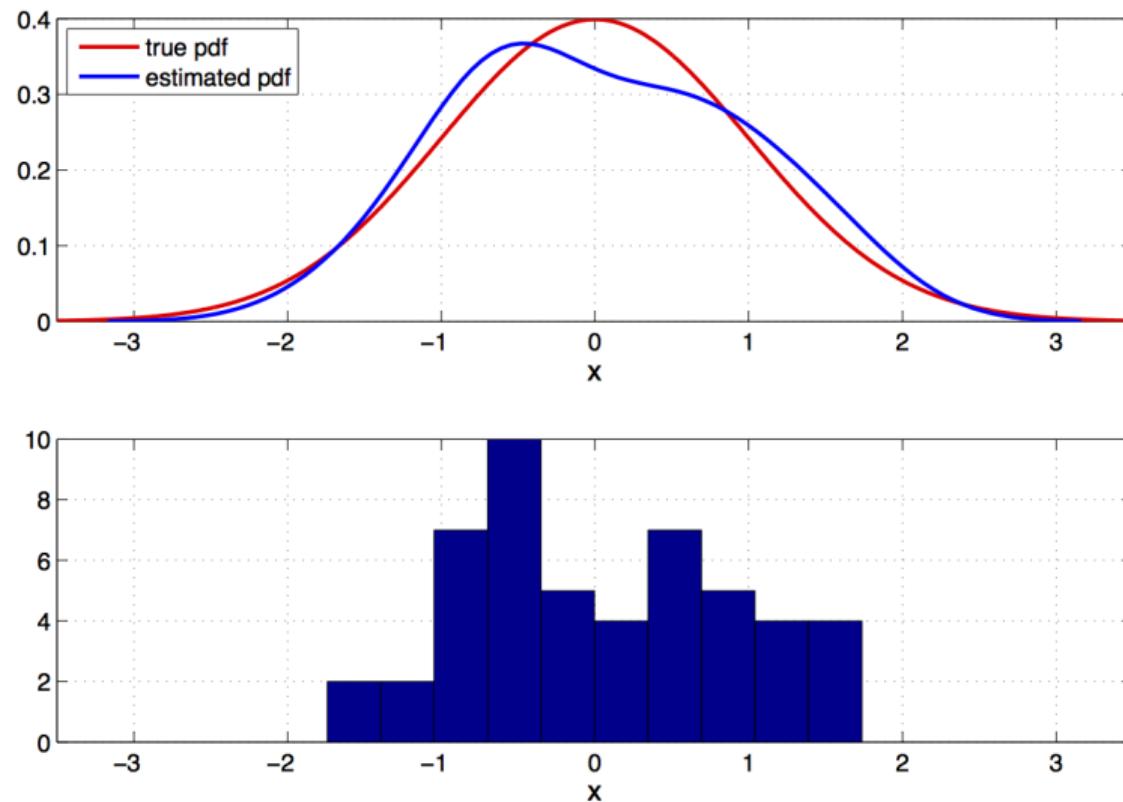


Positive Skew

Other moments



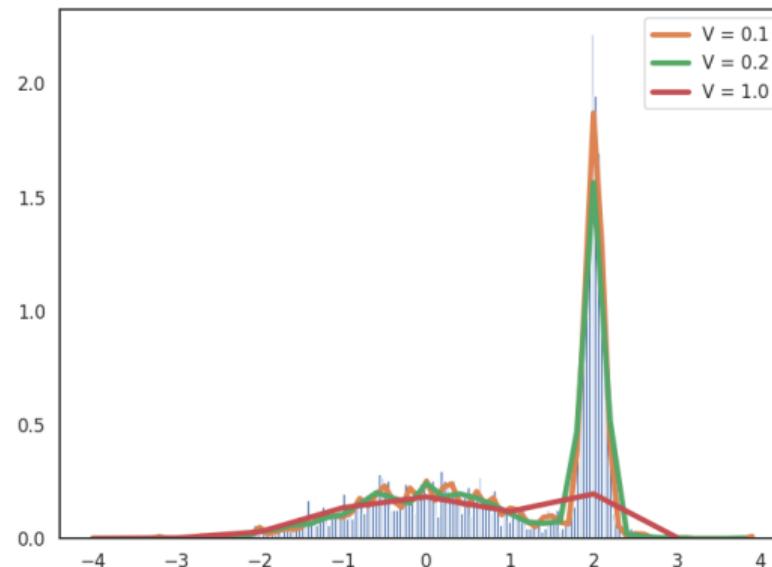
PDF estimation



KDE

Estimation using kernel function K :

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

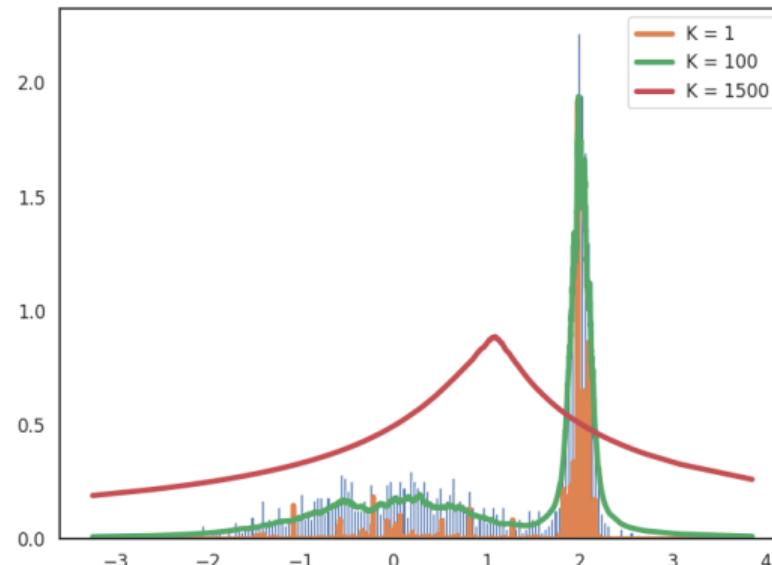


KDE

KNN can be considered as a generalization of KDE:

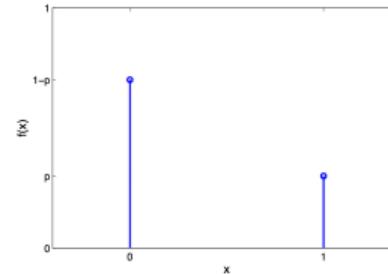
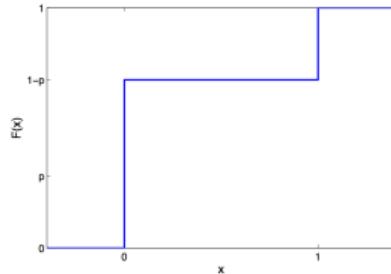
$$p(x) = \frac{1}{N} \sum_{x_i \in X} \frac{1}{h} K\left(\frac{x - x_i}{h_i}\right),$$

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



Bernoulli distribution

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



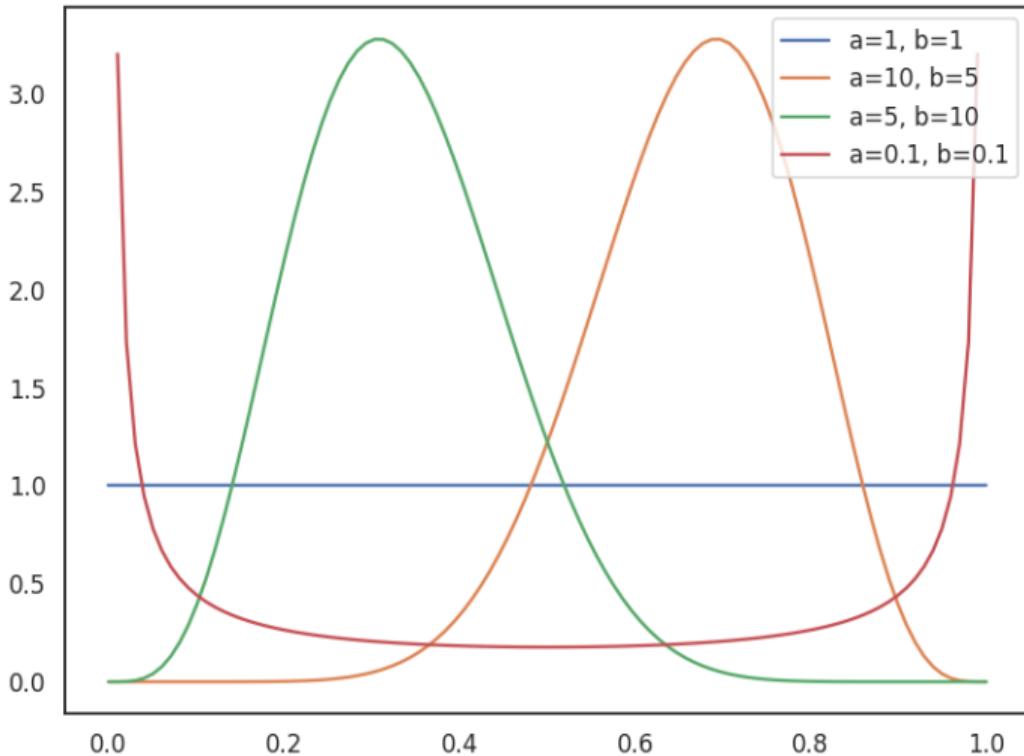
$$F(w) = \begin{cases} 0, & w < 0, \\ 1 - p, & 0 \leq w < 1, \\ 1, & w \geq 1. \end{cases}$$

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

- example: coin flipping

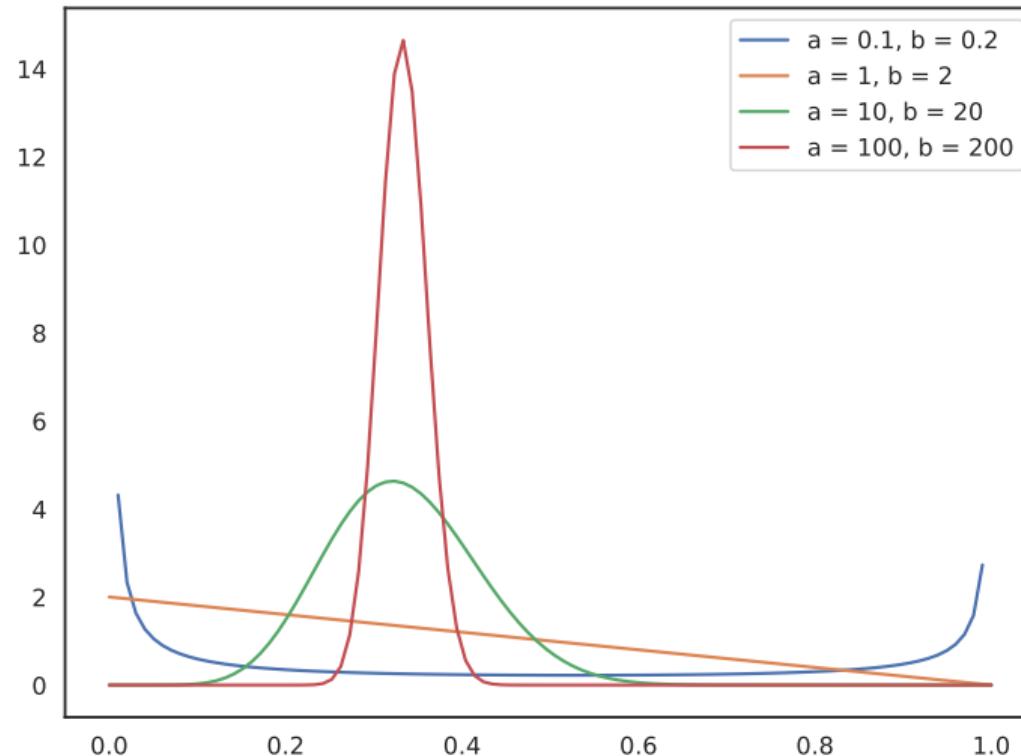
Beta-distribution

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



Beta-distribution

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



Beta-distribution

distribution

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

Multinomial distribution

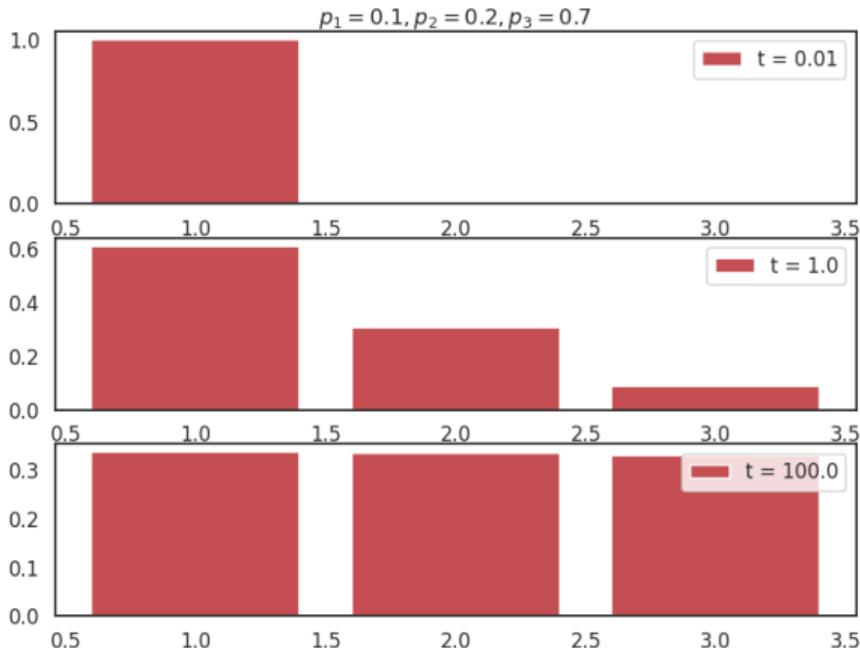
- Generalization of Bernoulli distribution for a larger number of events:

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

- Can be parametrized using softmax:

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

Multinomial distribution



Dirichlet distribution

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

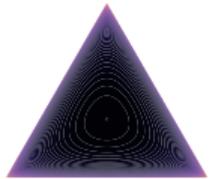
- Generalization of Beta-distribution for a larger number of events
- Support: $K - 1$ -simplex:

$$w_i \geq 0, \sum w_i = 1.$$

- 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:

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

Dirichlet distribution



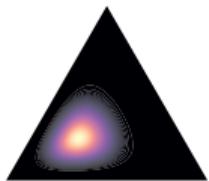
$\bar{\alpha} = [1, 1, 1]$, $t = 0.9$



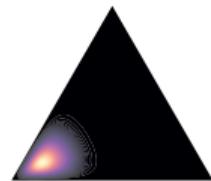
$t = 1.0$



$t = 10.0$



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



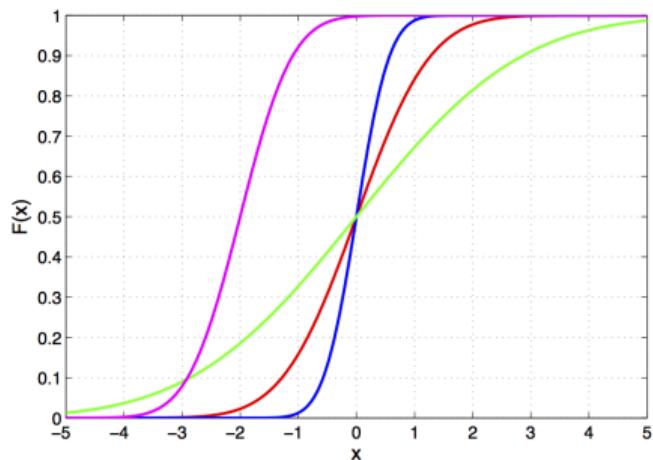
$[0.75, 0.125, 0.125]$



$[0.9, 0.05, 0.05]$

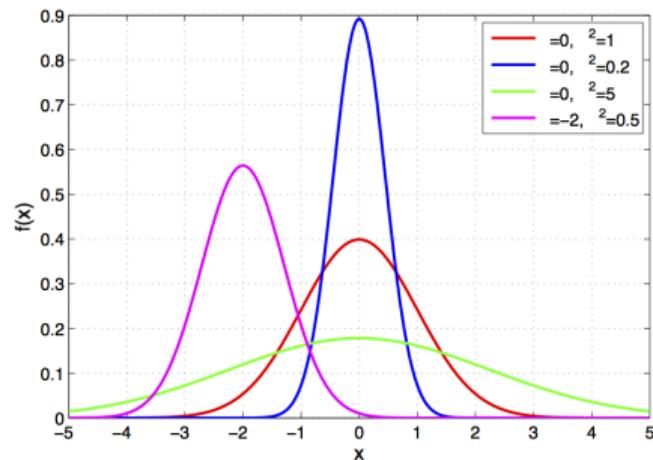
Normal distribution

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



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

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



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

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

Normal distribution

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

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

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

Student distribution

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

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

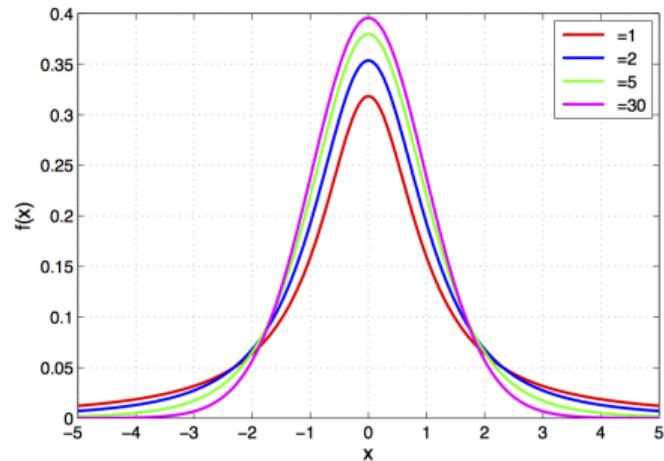
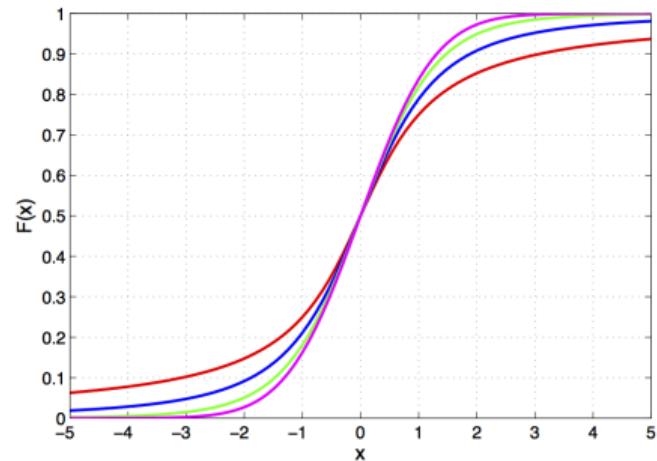
- if $w \sim St(\nu)$, then

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

- 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

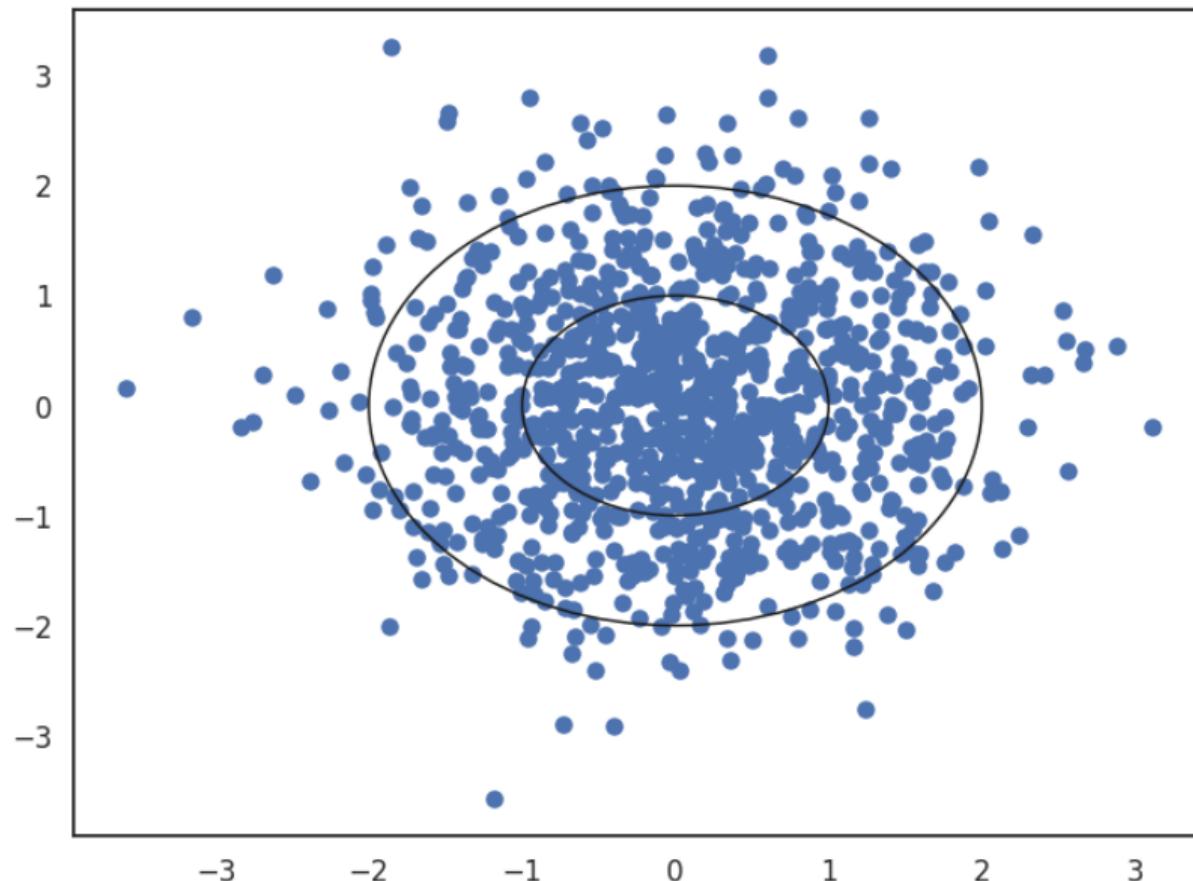
Periodical distribution

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

Periodical distribution

- 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:

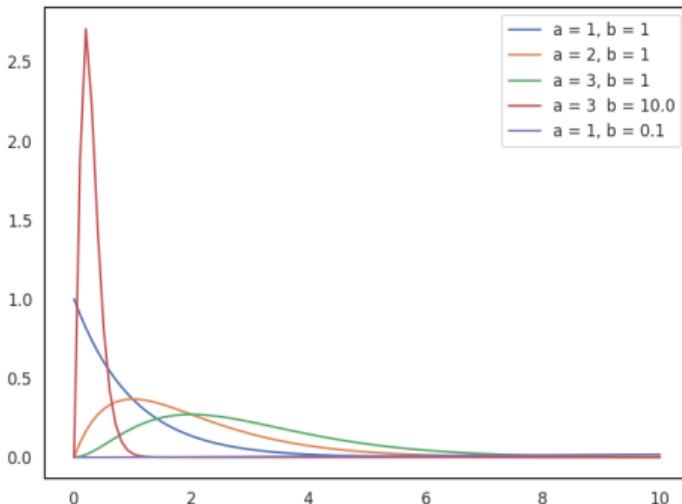
- ▶ $[w_1, w_2] \sim \mathcal{N}(\mu, \sigma^2)$
- ▶ Parameterize using polar coordinate system, limit the radius:

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

- See also Anton Bishuk's 2021 task 2 about Kent distribution.

Gamma distribution

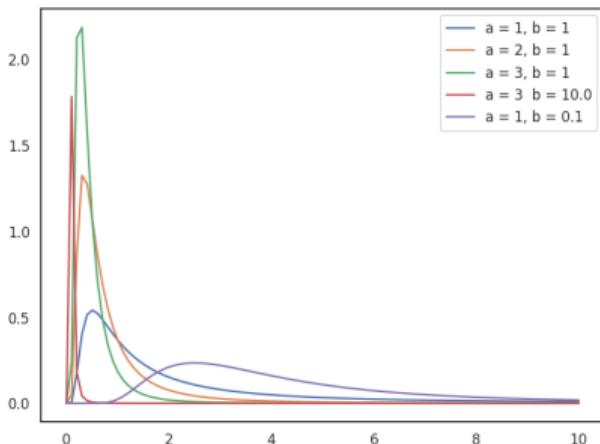
$$p(w, \alpha, \beta) = \frac{\beta^\alpha w^{\alpha-1} e^{-\beta}}{\Gamma(\alpha)}$$



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

Inverse Gamma-distribution

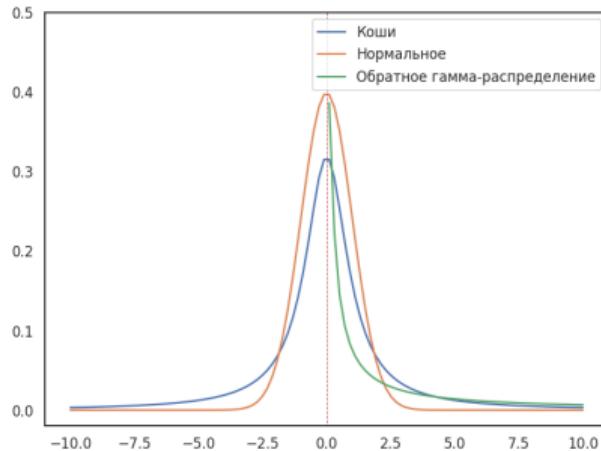
$$p(w, \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} w^{-\alpha-1} \exp(-\beta \cdot w^{-1}).$$



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

Cauchy distribution

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



- 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