FYS4411 MARCH 17, 2022

iid = independent and identically distributed

CLM: final distribution

$$N\left(\mu, \sigma^2/n\right)$$

$$\mu = \int p(x)\, x\, dx$$

$$\left(\sum_{i=1}^{n} p(x_i) x_i \right)$$

sample mean $\bar{\mu} \neq \mu$

$$\bar{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$X = \{x_1, x_2, \ldots x_n\}$$

$$\sigma^2 = \int p(x)(x-\mu)^2 dx$$

$$\neq \bar{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{\mu})^2$$

sample variance

$$p(x_1, x_2 \ldots x_n) = p(x_1)\, p(x_2) \ldots p(x_n)$$

$$\mathrm{cov}(x_i x_j) = \int dx_1\, dx_2 \ldots dx_n$$

$$\times (x_i - \mu)(x_j - \mu)$$
$$\times p(x_1)\, p(x_2) \dots p(x_n)$$

$$\int p(x)\,dx = 1 \quad \wedge \quad \mu = \int p(x)\,dx \cdot x$$

$$\mathrm{cov}(x_i\, x_j) = \int dx_i\, dx_j\, (x_i - \mu)(x_j - \mu) \times p(x_i)\, p(x_j)$$

$$= \int dx_i\, p(x_i)(x_i - \mu) \int dx_j\, (x_j - \mu)\, p(x_j)$$

$$= 0$$

if not iid:

$$\mathrm{cov}(x_i x_j) = \mathbb{E}[x_i x_j] - \mu_i \mu_j$$

sample variance for a data set $X = \{x_1, x_2, \dots x_n\}$

$$\hat{\sigma}^2 \to \sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$$

How do we estimate in a reliable way σ^2. σ^2 defines the standard deviation

$STD = \sqrt{\sigma^2} = \sigma$

Expected value $\boxed{\mu \pm \sigma}$

assume we have m - experiments and each has n - observations.

- each experiments - α - has

$$\mu_\alpha = \frac{1}{n} \sum_{k=1}^{n} x_{\alpha,k}$$

$$\sigma_\alpha^2 = \frac{1}{n} \sum_{k=1}^{n} (x_{\alpha,k} - \mu_\alpha)^2$$

- Repeat m times

$$\mu_m = \frac{1}{m} \sum_{\alpha=1}^{m} \mu_\alpha = \frac{1}{mn} \sum_{\alpha k} x_{\alpha,k}$$

Total variance

$$\sigma_m^2 = \frac{1}{m} \sum_{\alpha=1}^{m} (\mu_\alpha - \mu_m)^2$$

$$\mu_\alpha - \mu_m = \frac{1}{\quad} \sum^{n} (x_{\alpha k} - \mu_m)$$

$$\sigma_m^2 = \frac{1}{m} \sum_{\alpha=1}^{m} \left\{ \mu_\alpha^2 - \mu_\alpha \mu_m - \mu_\alpha \mu_m + \mu_m^2 \right\}$$

$$= \frac{1}{m} \sum_{\alpha=1}^{m} \mu_\alpha^2 - \mu_m^2$$

$$= \frac{1}{m} \sum_{\alpha=1}^{m} \left(\frac{1}{n} \sum_{k=1}^{n} x_{\alpha k} \frac{1}{n} \sum_{l=1}^{n} x_{\alpha l} \right) - \mu_m^2$$

$$= \frac{1}{mn^2} \sum_{\alpha=1}^{m} \sum_{k\,l=1}^{n} (x_{\alpha k} - \mu_m)(x_{\alpha l} - \mu_m)$$

$$= \frac{1}{mn^2} \sum_{\alpha k} (x_{\alpha k} - \mu_m)^2 + \frac{2}{mn^2} \sum_{\alpha=1}^{m} \sum_{k<l}^{n} (x_{\alpha k} - \mu_m) \times (x_{\alpha l} - \mu_m)$$

sample variance of all mn experiments

$$\sigma^2 = \frac{1}{mn} \sum_{\alpha k} (x_{\alpha k} - \mu_m)$$

$$\sigma_m^2 = \frac{\sigma^2}{n} + \underbrace{\mathrm{cov}}_{\geq 0}$$

We want σ_m^2 without having to evaluate cov.

Resampling : Bootstrap

$$X = \{x_1, x_2 \ldots x_n\}$$

m-bootstrap

1) compute μ, σ^2
2) reshuffle data randomly by selecting n points with replacement

length $= n$

$$X^1 = \{x_3, x_5, x_5, \ldots x_{n-50}\}$$

compute $\sigma^{2\,1}$ μ^1

3) repeat m-times

4) Find/compute

$$\sigma_m^2 = \frac{1}{m} \sum_{i=1}^{m} \sigma_i^2$$

Time consuming when n/m are large

$\Rightarrow$ Blocking method.

$n > 10^5 \sim 10^6$

sample mean

$$\mu_\alpha = \mu = \frac{1}{n} \sum_{k=1}^{n} x_{\alpha,k}$$

$$\mu_m = \frac{1}{mn} \sum_{\alpha=1}^{m} \sum_{k=1}^{n} x_{\alpha,k}$$

Total variance

$$\sigma_m^2 = \frac{\sigma^2}{n} + \frac{2}{mn^2} \sum_{\alpha=1}^{m} \sum_{k<l}^{n} (x_{\alpha k} - \mu_m)(x_{\alpha l} - \mu_m)$$

$$\sigma^2 = \frac{1}{mn} \sum_{\alpha} \sum_{k} (x_{\alpha k} - \mu_m)^2$$

introduce a shorthand notation

$$f_d = \frac{1}{nm} \sum_{\alpha=1}^{m} \sum_{k=1}^{n-d} (x_{\alpha k} - \mu_m)$$

$$\times (x_{\alpha\, k+d} - \mu_m)$$

$$d = |k - l|$$

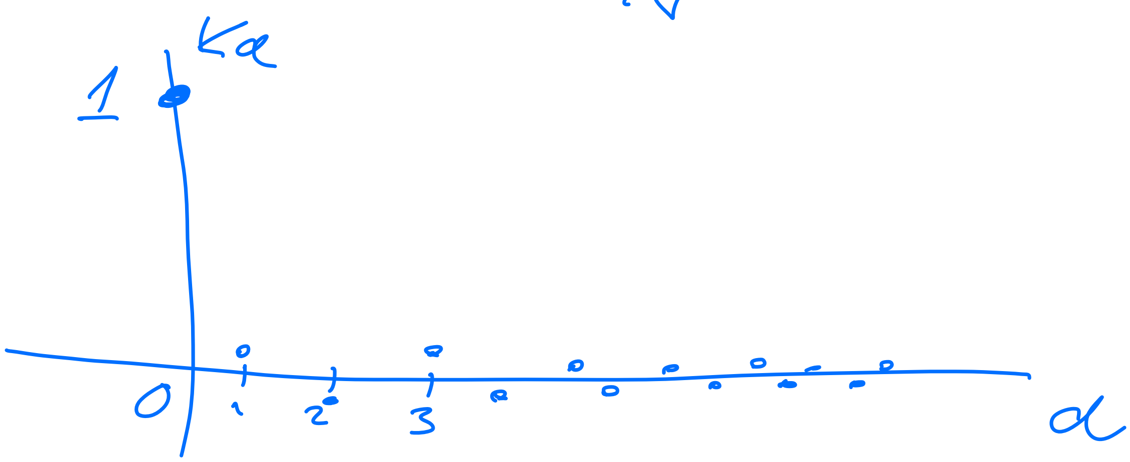
autocorrelation function

$$K_d = \frac{f_d}{\sigma^2}$$

$$\underset{d=0}{f_0} = \frac{1}{m}\frac{1}{n}\sum_{\alpha=1}^{m}\sum_{k=1}^{n}(x_{\alpha k} - \mu_m)^2$$

$$= \sigma^2$$

$$K_0 = \sigma^2/\sigma^2 = \underline{1}$$



$$\sigma_m^2 = \frac{\sigma^2}{n} + \frac{2}{n}\sum_{d=1}^{n-1} f_d$$

$$= \frac{\sigma^2}{n}\left[1 + 2\sum_{d=1}^{n-1} K_d\right]$$

Blocking method:
Flyvbjerg - Petersen
J. Chem Phys 91, 461 (1989)

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\sigma^2(\mu) = \mathbb{E}[x^2] - \mu^2$$

$m = 1$

$$\sigma_m^2 = \sigma_1^2 = \frac{\sigma^2}{n} + \frac{2}{n^2}\sum_{k<l}(x_k-\mu)$$
$$\times (x_l - \mu)$$

$$= \sigma^2(\mu)$$

$$\gamma_{ij} = \mathbb{E}[x_i x_j] - (\mathbb{E}[\mu])^2$$

$$= \gamma_t \qquad t = |i-j|$$

$$\sigma^2(\mu) = \frac{1}{n^2}\sum_{ij}\gamma_{ij}$$

$$= \frac{1}{n}\left[\gamma_0 + 2\sum_{t=1}^{n-1}\left(1-\frac{t}{n}\right)\gamma_t\right]$$

$\ldots$ $t=1$

σ^2

$\gamma_0' = \frac{1}{2}\gamma_0 + \frac{1}{2}\gamma_1$

Algorithm:

Transform data

$$X = \{x_1 \, x_2 \, \ldots \, x_n\}$$

into half as large a data set

$$\{x_1', x_2', \ldots x_{n'}'\} \quad n' = \frac{1}{2}n$$

$$x_i' = \frac{1}{2}\left[x_{2i-1} + x_{2i}\right]$$

$$\boxed{\mu' = \mu} \quad \wedge \quad \sigma^2(\mu') = \sigma^2(\mu)$$

$$X = \{x_1 \; x_2 \; x_3 \; x_4 \; x_5 \; x_6\}$$

$$= \{1, 2, 3, 4, 5, 6\}$$

$$\mu = \frac{1}{6}\sum_{i=1}^{6} x_i = 7/2 = \mu' = \frac{1}{3}\sum_{i=1}^{3} x_i'$$

$$\sigma^2(\mu) = \frac{1}{6}\sum_{i=1}^{6}(x_i)^2 - \mu^2 = 8/3$$

$$\sigma^{2\prime}(\mu') = \frac{1}{3}\sum_{i=1}^{3}\left(\frac{x_{2i-1}+x_{2i}}{2}\right)^2 - \mu'^2$$

$$= \frac{1}{3}\left(\frac{(x_1+x_2)^2}{4} + \frac{(x_3+x_4)^2}{4} + \frac{(x_5+x_6)^2}{4}\right) - \mu'^2 = 8/3$$

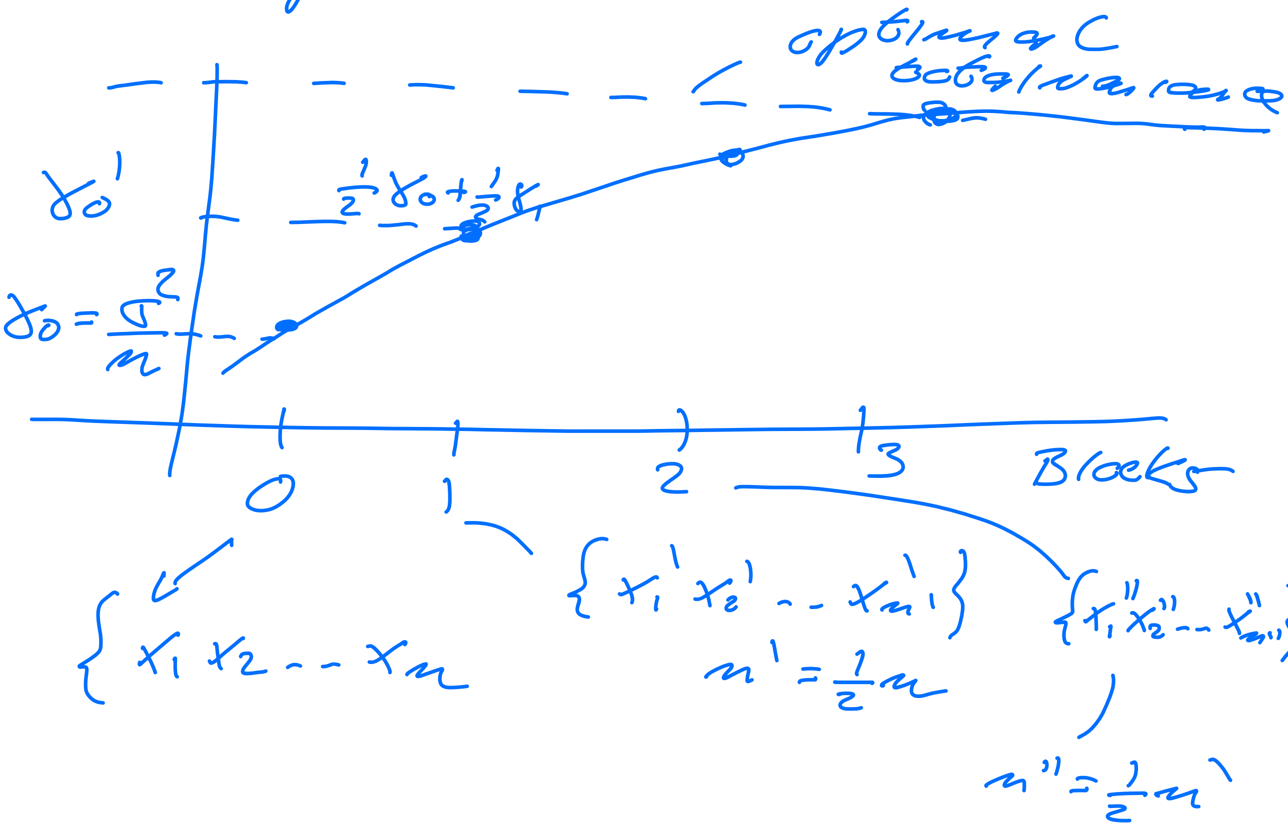
x_i' : $\gamma_{ij}' = \gamma_t'$

x_i : $\gamma_{ij} = \gamma_t$

$$\gamma_t' = \begin{cases} \frac{1}{2}\gamma_0 \overset{\sigma^2}{} + \frac{1}{2}\gamma_1 & t=0 \\ \frac{1}{4}\gamma_{2t-1} + \frac{1}{2}\gamma_{2t} + \frac{1}{4}\gamma_{2t+1} & t>0 \end{cases}$$

$$\gamma_0' = \frac{1}{2}\gamma_0 + \frac{1}{2}\gamma_1$$

continue splitting the data in halves, till γ_t' reaches a constant value after t operations



applied to n-data sets when $n \sim 10^5$ or larger, convenient to have $n=2^k$, $k \sim 20$ or larger.

—— One body densities ——

$$\rho(\vec{r}_1) = \int d\vec{r}_2\, d\vec{r}_3 \ldots d\vec{r}_n$$

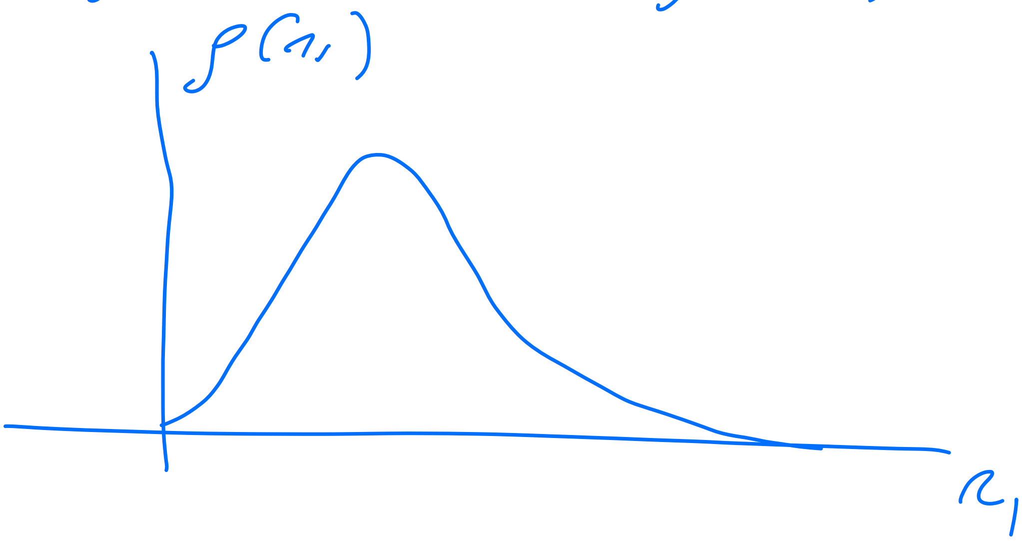
$\times \; |\Psi_T(\vec{r}_1, \vec{r}_2, \dots \vec{r}_N; \vec{\alpha}_{opt})|$

non-interacting case

$\vec{r}_1 \longrightarrow r_1 = \sqrt{x_1^2 + y_1^2}$

make a table of x_1 and y_1 values and compute

$$\rho(x_1, y_1) = \rho(r_1)$$



interaction added

- repulsive
- attractive