Learning goals of the week:

- How to use the covariance matrix: “Error Matrix”
- Statistical Inference
- Fitting: Least Squares method

Week 4

Error matrix

Error matrix

[Lyons]

(often the covariance matrix is called error matrix)

Take two uncorrelated measurements: e.g. a 2D pdf built form two uncorrelated gaussian: (centred at zero without loss of generality)

$$\begin{aligned} P(x) &= \frac{1}{\sqrt{2\pi}}\frac{1}{\sigma_x}e^{-\frac{1}{2}\frac{x^2}{\sigma_x^2}} \\ P(y) &= \frac{1}{\sqrt{2\pi}}\frac{1}{\sigma_y}e^{-\frac{1}{2}\frac{y^2}{\sigma_y^2}} \end{aligned}$$

$$P(x,y) = \frac{1}{2\pi}\frac{1}{\sigma_x\sigma_y}e^{-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}\right)}$$

The ±1σ error bar in one dimension becomes a 1σ ellipse in 2D:

$$\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} = 1$$

$$(x,y)\begin{pmatrix} \frac{1}{\sigma_x^2} & 0 \\ 0 & \frac{1}{\sigma_y^2} \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = 1$$

Inverse of the error matrix

Error matrix = $\begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix}$

in general $\langle (x_i - \bar{x_i})(x_j - \bar{x_j}) \rangle$

And from here the **correlation coefficient**: $\rho_{x_1x_2} = \frac{cov(x_1,x_2)}{\sqrt{V(x_1)V(x_2)}}$

Error matrix

The matrix notation allows to easily treat the case of correlated variables.
Build a correlation between the two variables by simply rotate the axes (by 30^{o})

$$\begin{aligned} x' &= x\cos\theta - y\sin\theta \\ y' &= x\sin\theta + y\cos\theta \end{aligned}$$

Let's use a numerical example to get the idea:

$\sigma_x = 1/4$ and $\sigma_y = 1/2$

Suppose the ellipse for the uncorrelated case is:

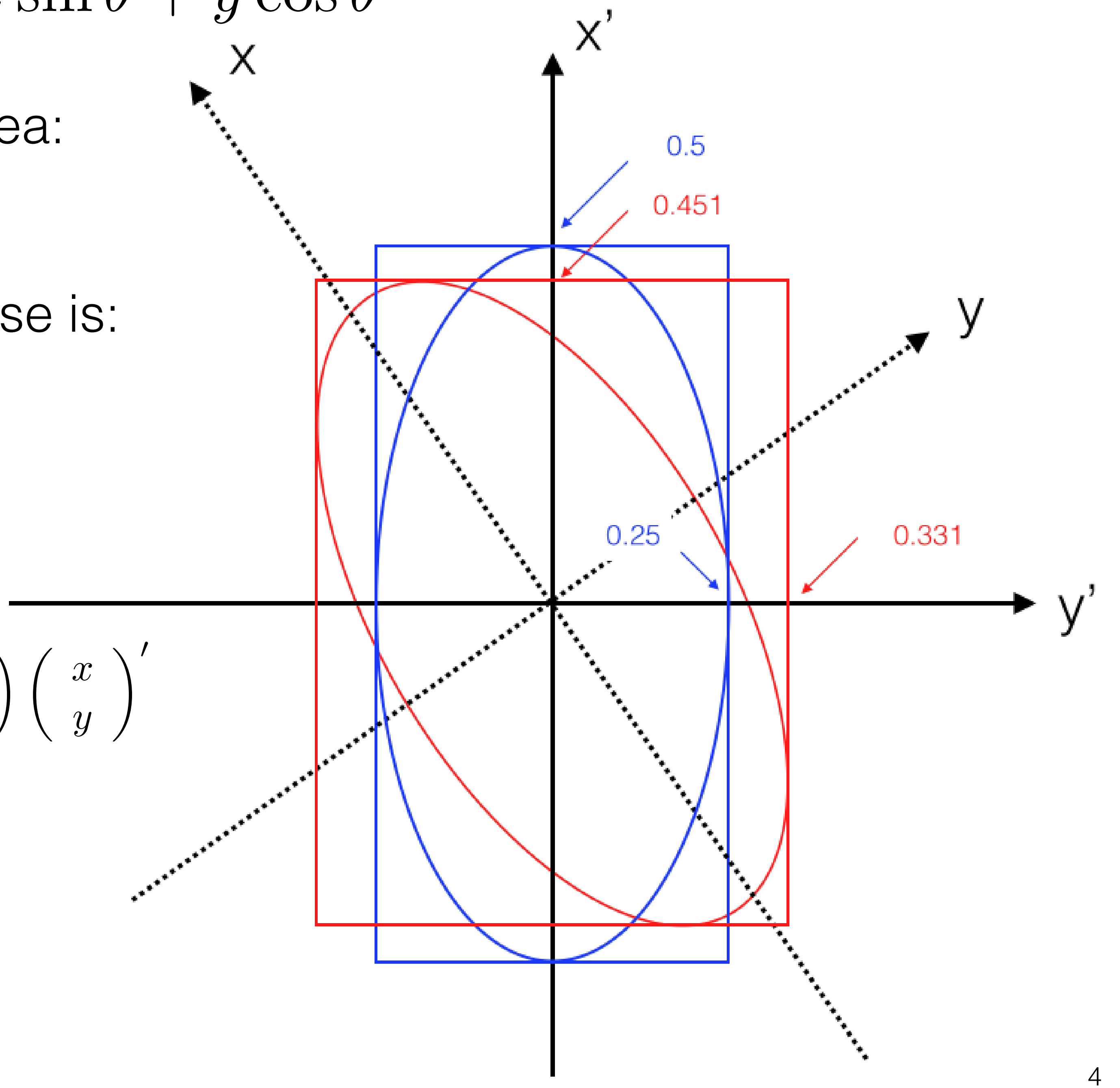
$$16x^2 + 4y^2 = 1$$

After rotation we have:

$$(x\ y)' \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} 1/\sigma_x^2 & 0 \\ 0 & 1/\sigma_y^2 \end{pmatrix} \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}'$$

$$(x\ y)' \begin{pmatrix} 13 & 3\sqrt{3} \\ 3\sqrt{3} & 7 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}'$$

and Error Matrix: $\frac{1}{64}\begin{pmatrix} 7 & -3\sqrt{3} \\ -3\sqrt{3} & 13 \end{pmatrix}$

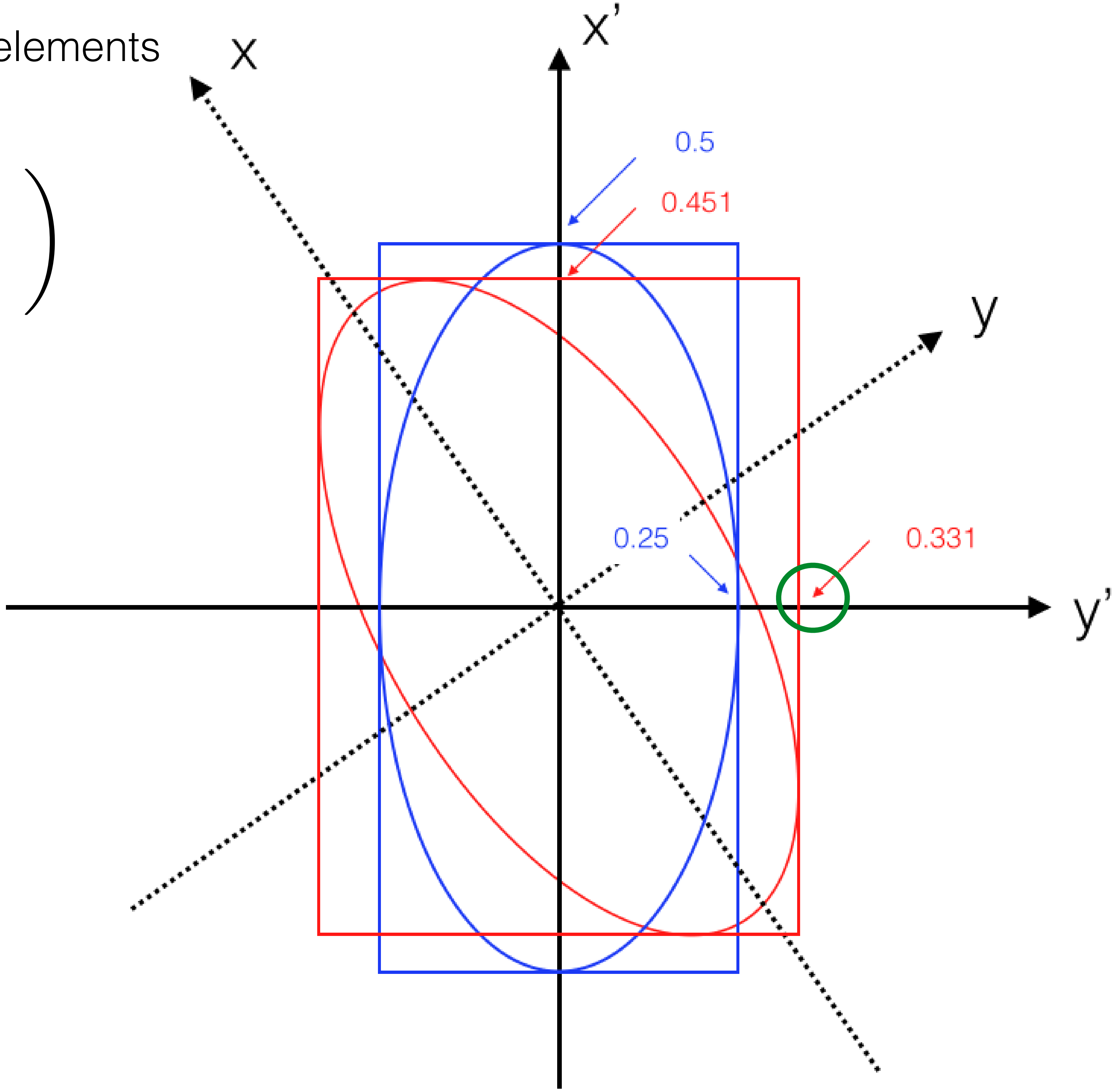


Error matrix

What is the meaning of the elements of the error matrix ?

$$\frac{1}{64}\begin{pmatrix} \boxed{7} & -3\sqrt{3} \\ -3\sqrt{3} & 13 \end{pmatrix}$$

$$\sigma^2_{x'} = 7/64$$

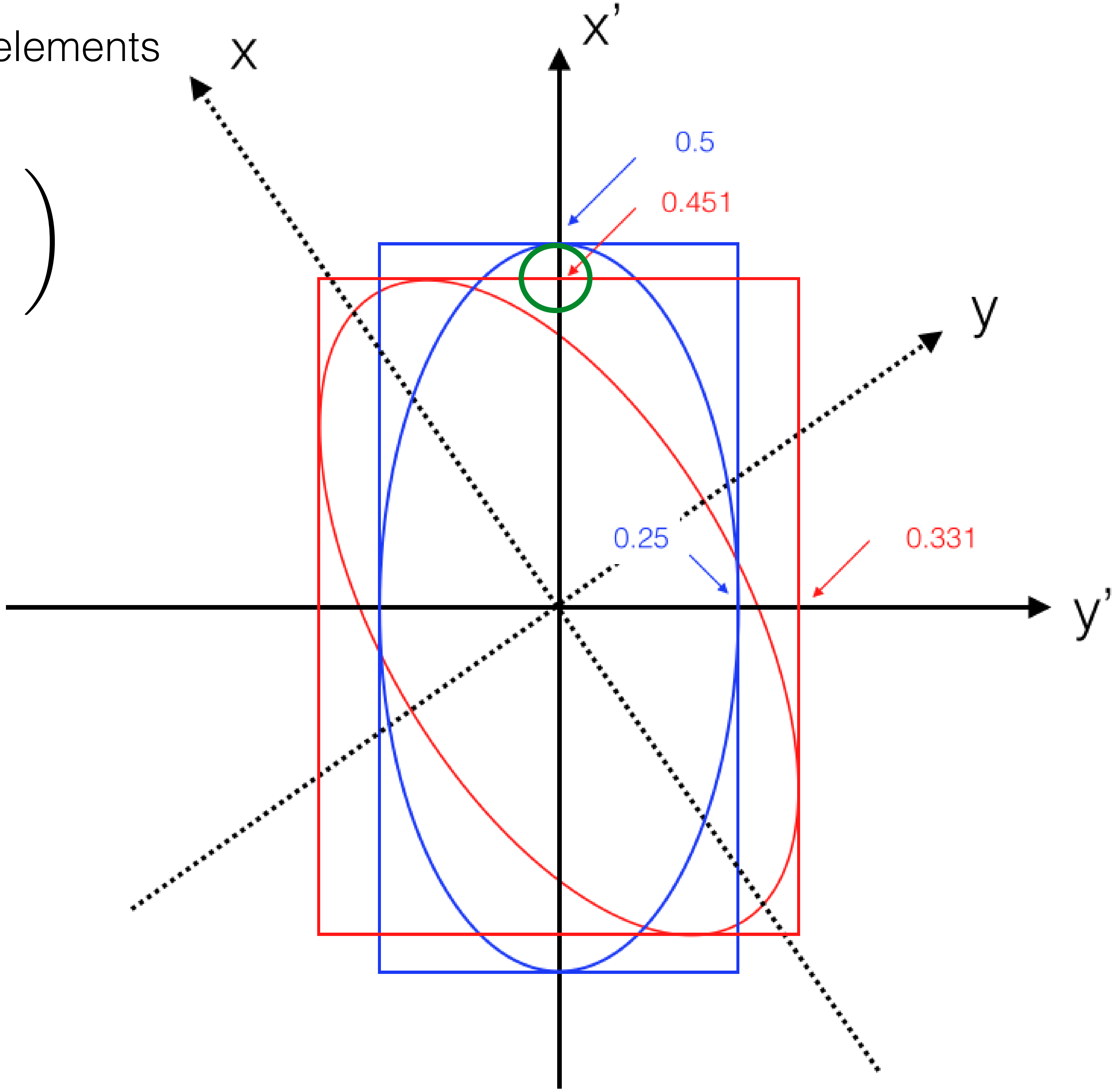


Error matrix

What is the meaning of the elements of the error matrix ?

$$\frac{1}{64}\begin{pmatrix} 7 & -3\sqrt{3} \\ -3\sqrt{3} & \boxed{13} \end{pmatrix}$$

$$\sigma^2_{y'} = 13/64$$



Error matrix

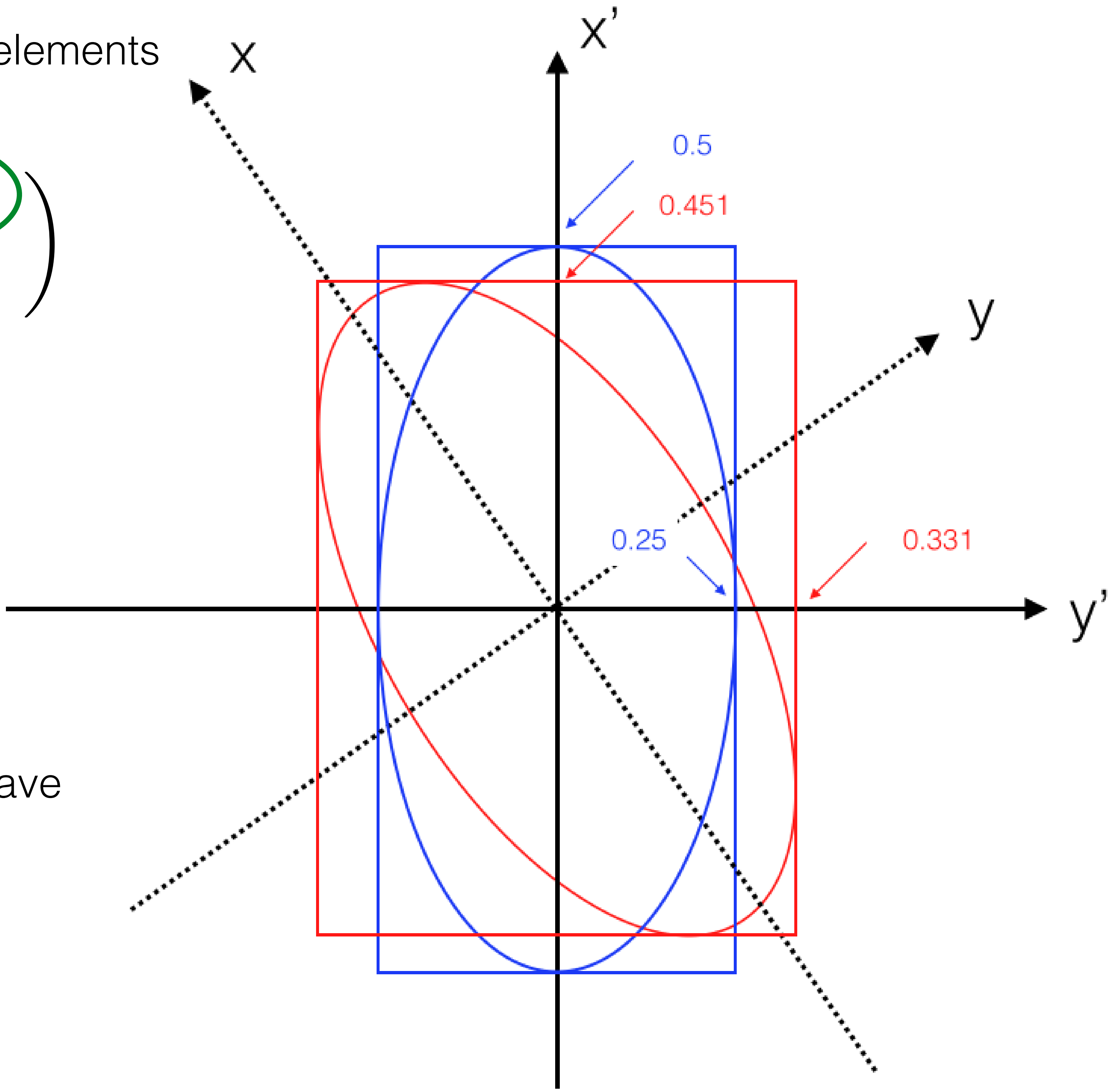
What is the meaning of the elements of the error matrix ?

$$\frac{1}{64}\begin{pmatrix} 7 & -3\sqrt{3} \\ -3\sqrt{3} & 13 \end{pmatrix}$$

$$\rho\sigma_{x'}\sigma_{y'}$$

Knowing σx′ and σy′ from the diagonal elements we have

$$\rho = 0.54$$

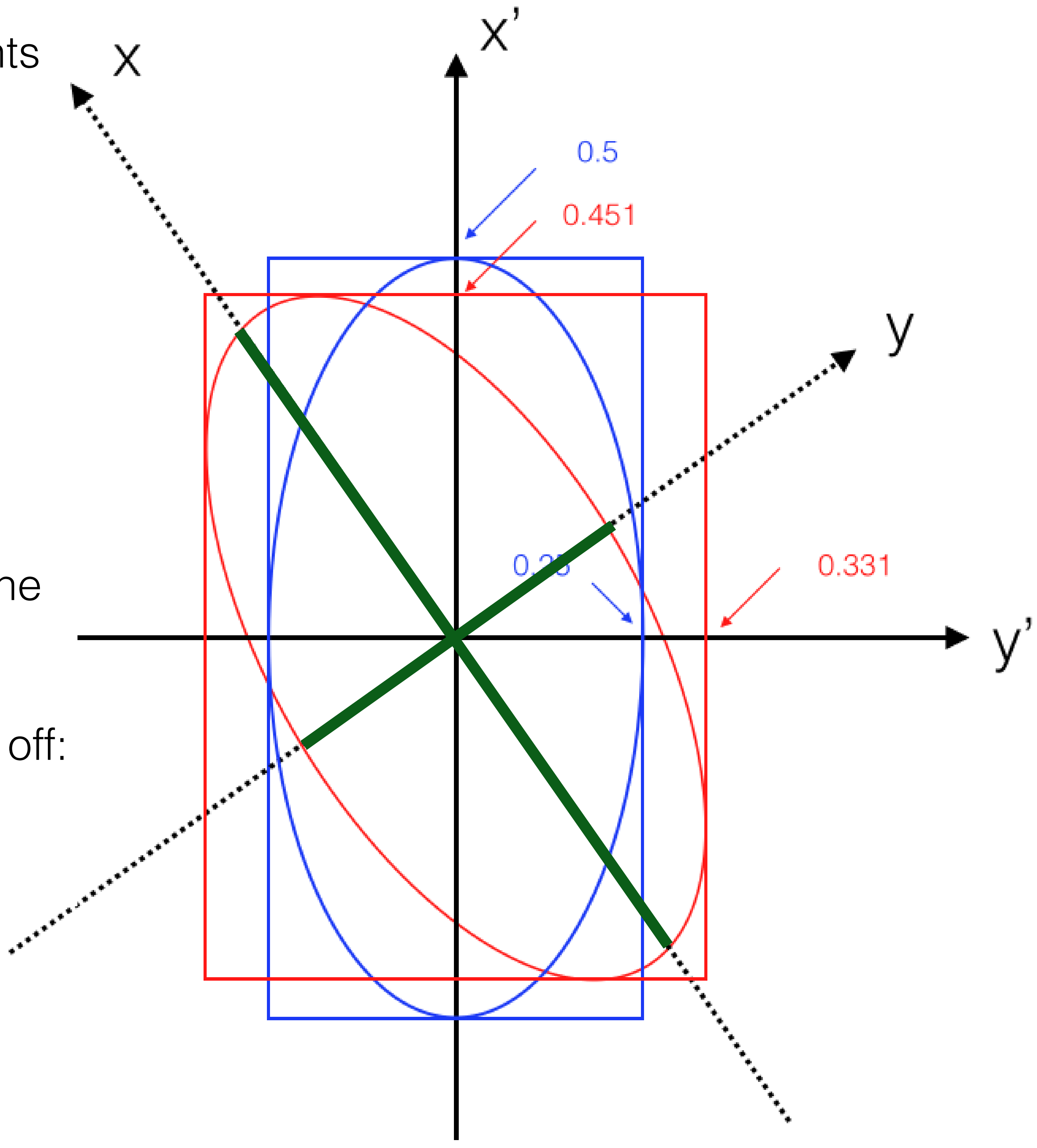


Error matrix

What is the meaning of the elements of the error matrix ?

$$\frac{1}{64}\begin{pmatrix} 7 & -3\sqrt{3} \\ -3\sqrt{3} & 13 \end{pmatrix}$$

The semi-axes of the ellipse are the square roots of the eigenvalues of the error matrix (here we know the diagonalized matrix, i.e. before rotation, and we can just read them off: 0.25, 0.5)

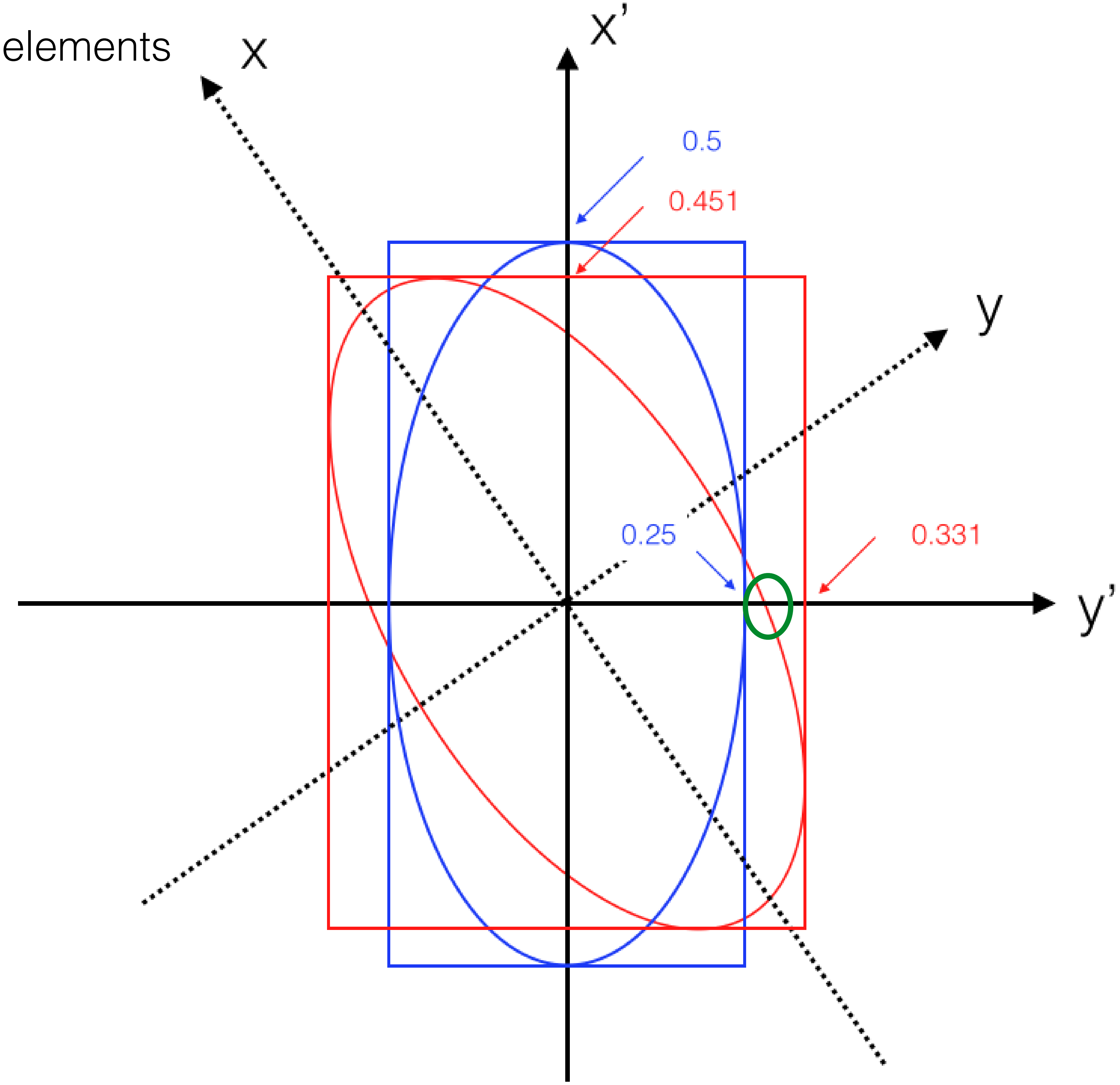


Inverse Error matrix

What is the meaning of the elements of the inverse error matrix ?

$$\begin{pmatrix} \boxed{13} & 3\sqrt{3} \\ 3\sqrt{3} & 7 \end{pmatrix}$$

$$\sqrt{1/13} = 0.277$$

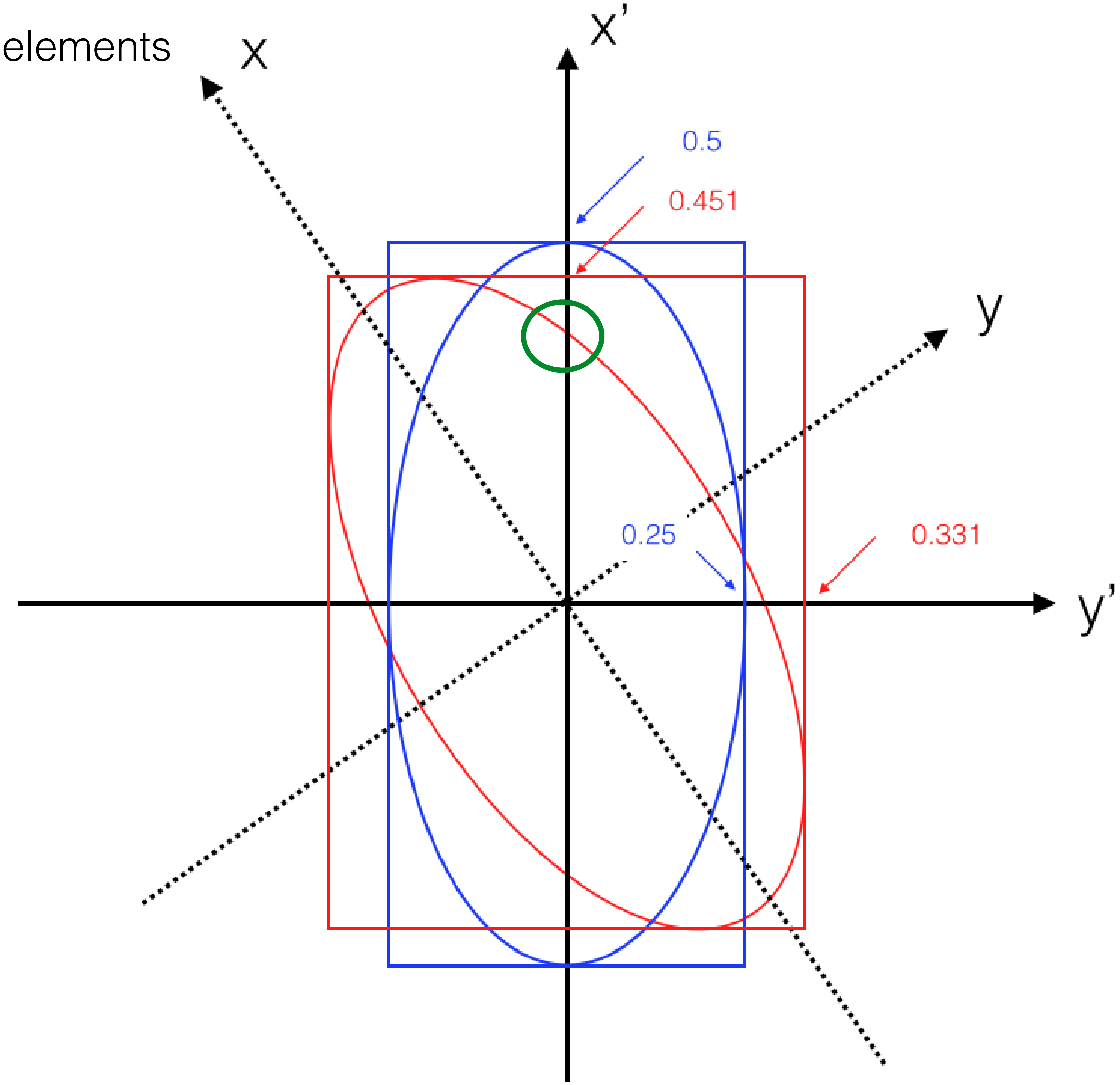


Inverse Error matrix

What is the meaning of the elements of the inverse error matrix ?

$$\begin{pmatrix} 13 & 3\sqrt{3} \\ 3\sqrt{3} & \boxed{7} \end{pmatrix}$$

$$\sqrt{1/7} = 0.378$$



Working with systematic uncertainties

The easiest way to work with systematic uncertainties when you have to work out the uncertainty on a function of several variables is to use the matrix notation:

Suppose the variables x_1 and x_2 are affected by a common systematic uncertainty S, then:

$$V_{i,j}^{tot} = \begin{pmatrix} \sigma_1^2 + S^2 & S^2 \\ S^2 & \sigma_2^2 + S^2 \end{pmatrix}$$

or in case of a systematic uncertainty given as a percentage:

$T = \epsilon x_i$ (e.g. $\epsilon = 0.01$ for a 1%)

$$V_{i,j}^{tot} = \begin{pmatrix} \sigma_1^2 + \epsilon^2 x_1^2 & \epsilon^2 x_1 x_2 \\ \epsilon^2 x_1 x_2 & \sigma^2 + \epsilon^2 x_2^2 \end{pmatrix}$$

Working with systematic uncertainties

Example Consider two variables x and y with two sources of uncertainties: a statistical (s_x, s_y) with *no* correlation and a systematic (c_x, c_y) with *full* correlation

$$\begin{aligned} x &= x_0 \pm s_x \text{ (stat)} \pm c_x \text{ (syst)} && (3.7.40) \\ y &= y_0 \pm s_y \text{ (stat)} \pm c_y \text{ (syst)} && (3.7.41) \end{aligned}$$

Because the uncertainty's are already separated into a correlated and uncorrelated category, they can be summed up in quadrature at the matrix level, yielding:

$$V_{ij}^{tot} = \begin{pmatrix} s_x^2 & 0 \\ 0 & s_y^2 \end{pmatrix} + \begin{pmatrix} c_x^2 & c_{xy} \\ c_{yx} & c_y^2 \end{pmatrix} = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}, \tag{3.7.42}$$

where ρ is the correlation coefficient $\rho = \frac{c_{xy}}{\sigma_x \sigma_y}$ and $\sigma_i^2 = s_i^2 + c_i^2$ is the sum of the squared individual uncertainties for x and y, respectively. □

Working with systematic uncertainties

Example Take three variables x_1, x_2, x_3 with statistical uncertainties $\sigma_1, \sigma_2, \sigma_3$, a common systematic uncertainty S and a second systematic uncertainty T shared by only x_1 and x_2. In this case the covariance matrix reads:

$$V_{i,j}^{tot} = \begin{pmatrix} \sigma_1^2 + S^2 + T^2 & S^2 + T^2 & S^2 \\ S^2 + T^2 & \sigma_2^2 + S^2 + T^2 & S^2 \\ S^2 & S^2 & \sigma_3^2 + S^2 \end{pmatrix} \tag{3.7.43}$$

□

Statistical inference

Statistical Inference

Probability: pdf —> compute probability of any outcome
Statistics: give a set of data points sampling a distribution, infer the characteristics of the parent distribution

Two main classes of problems:
Parameter estimation: from the data assumed to follow a pdf —> estimate its parameters (point estimation)
Hypothesis testing: test if the data collected follow a given distribution

The two classes are strongly related !

Frequentist / Bayesian approach to probability produce two approaches Frequentist / Bayesian to statistical inference. (We will use frequentists likelihood)

Parameters estimation

We call "statistics" any function of the data (i.e. data itself is a random variable)
Estimator = statistics to estimate a parameter (again a random variable)

parameter θ estimator of the parameter $\hat{\theta}$

For any given parameter you can define several estimators:

Example: parameter ETHZ students stature. A few possible estimators are:

- add all h_i and divide by N
- add only the first 15, divide by 15; ignore the rest
- add all h_i and divide by N-1
- just quote it to be 1.82 m
- multiply all h_i and take Nth-root of result
- choose the most popular height (mode)
- take shortest and tallest and divide by 2
- …

How do you choose the one that better suits your needs ?

Estimators properties

Estimators $\hat{\theta}$ are random variables —> they are distributed according to a pdf $g(\hat{\theta}|\theta)$ (which must depend on θ).

The estimator is chosen based on your experience and taking into account:

- bias
- consistency
- efficiency
- robustness

Bias

an estimator is unbiased if its expectation value is equal to its true value

$$< \hat{\theta} >= \theta$$

The bias of an estimator is: $\boxed{b_n =< \hat{\theta} > -\theta}$

n is the size of the sample used for the estimation. Some estimators are unbiased only for large n (i.e. “asymptotically”)

Example: the mean is an asymptotically unbiased estimator $\langle\bar{\mu}\rangle$ —> μ

Often knowing the bias allows to build an unbiased estimator by correcting for it

Estimators properties

Consistency

an estimator is consistent if by adding more data to your experiment you obtain a smaller variance

$$\text{if } \forall\epsilon > 0,\ \lim_{n\to\infty} P(|\hat{\theta} - \theta| \geq \epsilon) = 0$$

Example:

The distribution $N(\bar{x};\mu,\sigma^2/n)$ will tend to a delta for n —> infinity

Efficiency

if it has the smallest possible variance (see later likelihood and Variance bound)

$$\epsilon = \frac{\text{minimal Variance of } \hat{\theta}}{\text{Variance of } \hat{\theta}}$$

Robustness

if it is not sensitive to the details of the parent distribution

Example: estimation of the mean

This estimator of the mean is:

$$\hat{\mu} = \frac{1}{n}\sum_i x_i$$

unbiased

$$\langle\hat{\mu}\rangle = \langle \tfrac{1}{n}\sum_i x_i\rangle = \tfrac{1}{n}\sum_i\langle x_i\rangle = \mu$$

consistent

$$V(\hat{\mu}) = \frac{1}{n}\sigma^2$$

efficiency

depends on the pdf: for a uniform distribution the estimator $\hat{\mu} = 0.5(x_{max} + x_{min})$ has a smaller variance

The robustness depends on the pdf: in general to increase the robustness of the estimator we can cut out the tails of the distribution —> truncated mean.
Price to pay —> in general the estimator will be biased

Examples:
- in skating the final score is a truncated mean (exclude the highest/lowest scores)
- the mean of distributions w/o moments (infinite integral) is taken as the truncated mean

Example: estimation of the variance

Knowing the true mean of the pdf we define: $$s_1^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

unbiased

$$\begin{aligned}
< s_1^2 > &= \frac{1}{n}\langle\sum(x_i - \mu)^2\rangle \\
&= \frac{1}{n}\left(\langle\sum x_i^2\rangle - 2\mu\langle\sum x_i\rangle + n\mu^2\right) \\
&= \frac{1}{n}\left(n\langle x^2\rangle - 2n\mu\langle x\rangle + n\mu^2\right) \qquad (independent\ x_i\ so\ :\langle\sum x_i^2\rangle = n\langle x^2\rangle) \\
&= \langle x^2\rangle - 2\mu^2 + \mu^2 \\
&= \sigma^2 - \mu^2 + \mu^2 \qquad (\sigma^2 = \langle x^2\rangle - \mu^2) \\
&= \sigma^2
\end{aligned}$$

Example: estimation of the variance

For the more common case where we don't know the true mean we defined:

$$s_x^2 = \frac{1}{n}\sum(x_i - \bar{x})^2 = \bar{x^2} - \bar{x}^2$$

from which we have (try to work it out):

$$\langle s_x^2 \rangle = \frac{1}{n}\left(\langle \sum x_i^2 \rangle - \frac{1}{n} \langle \left(\sum x_i \right)^2 \rangle \right) = \frac{1}{n}(n-1)\sigma^2$$

s_x is a biased estimator of the variance. That's because we used the estimated mean instead of the true mean: The spread of the data around the sample mean is smaller than the spread around the true mean.

$$s^2 = \frac{n}{n-1} s_x^2 = \frac{n}{n-1}(\bar{x^2} - \bar{x}^2) = \boxed{\frac{1}{n-1}\sum(x_i - \bar{x})^2}$$

This is where the "n-1" terms come from in the sample variance definition (see week 1)

(the normalization is given by the number of degrees of freedom of the problem not by the number of points)

Fits

What is fitting ?

The basic quantities we used so far to characterize a sample of data are the mean and variance. We have just seen how to build their estimators.

Often you (want to) know more about your data. E.g. you have a model (function/distribution) to describe them.

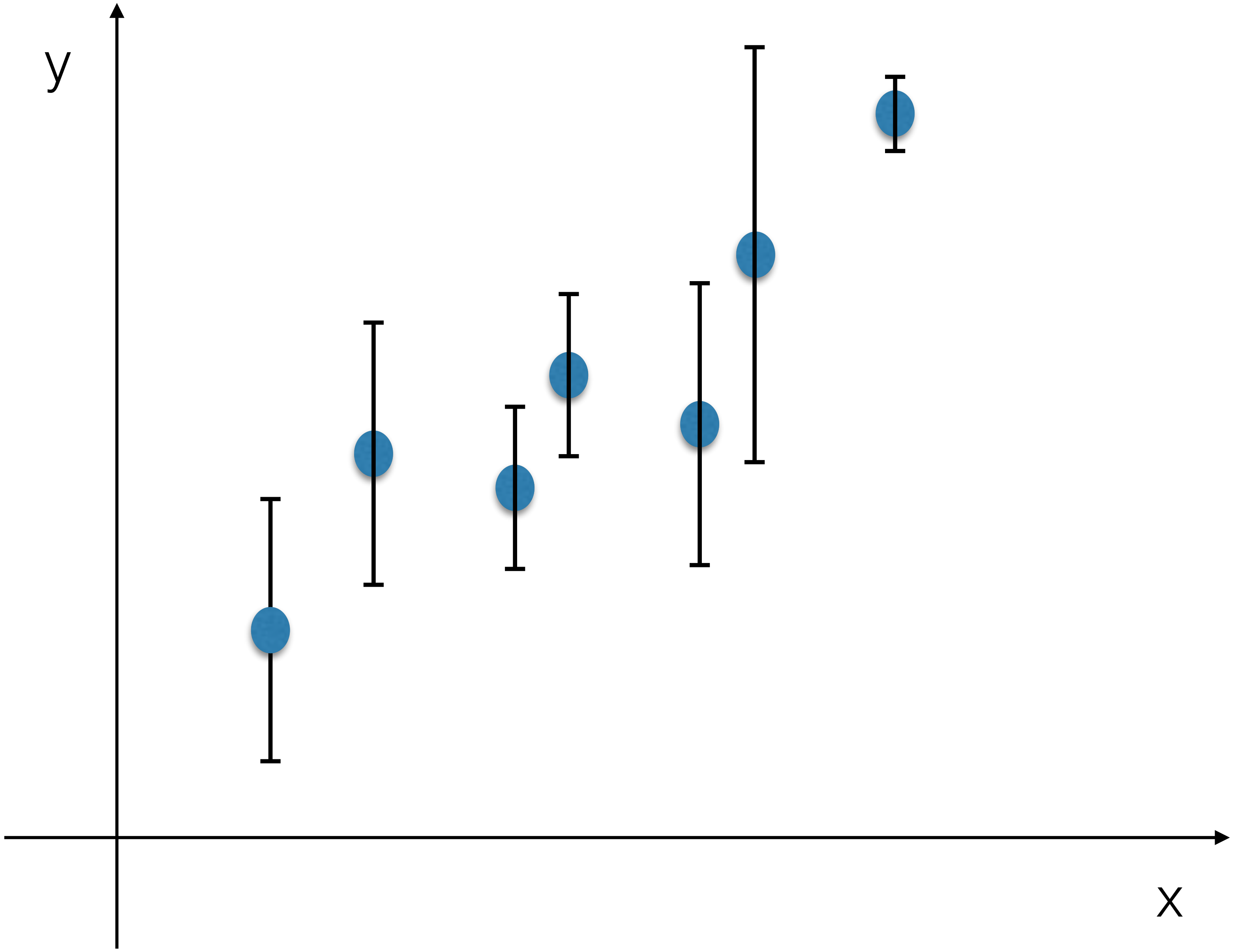
How can you generically build an estimator for the (parameters of) the model ?

We will describe two general methods for parameters estimation:

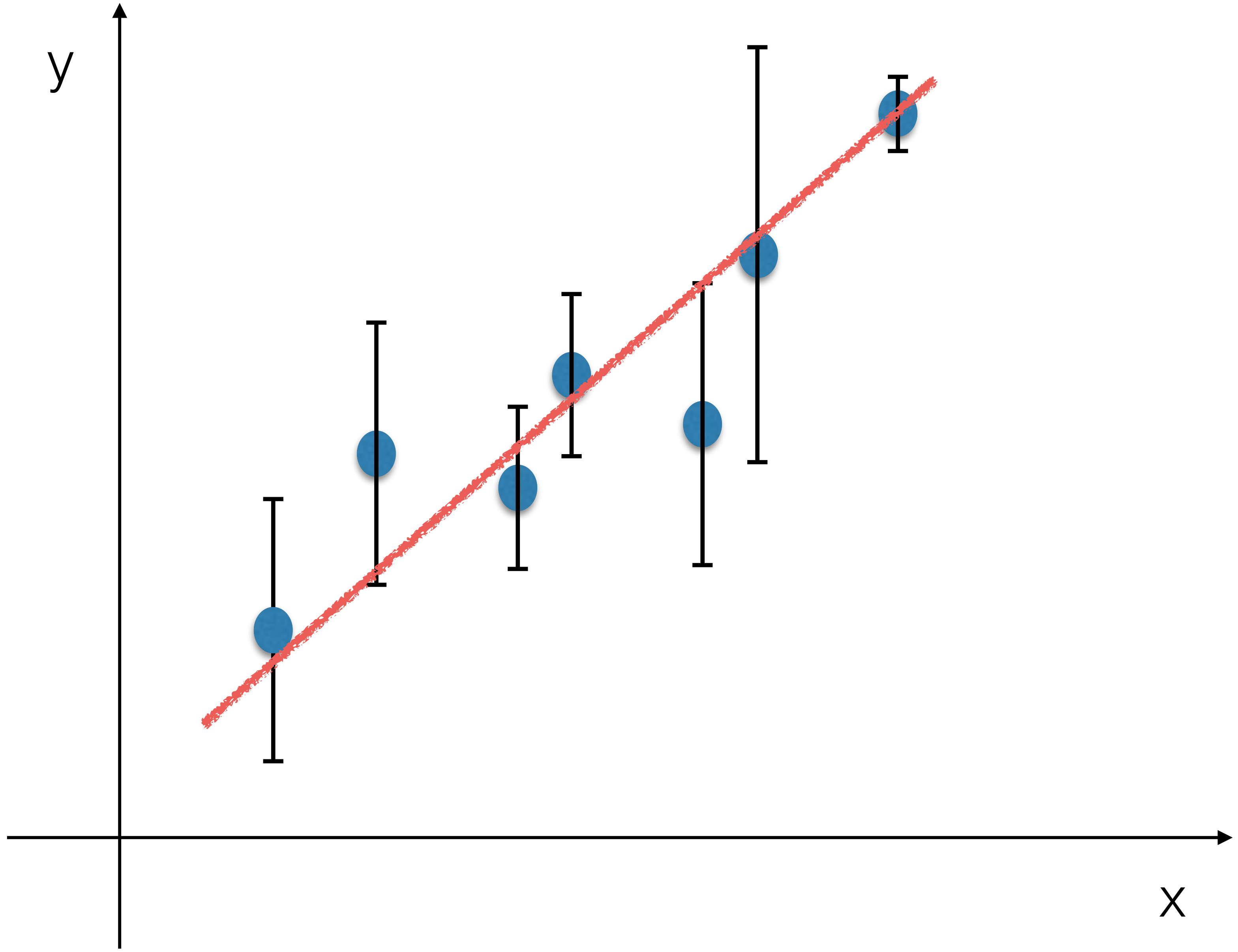
- least squares
- maximum likelihood

What is fitting ?

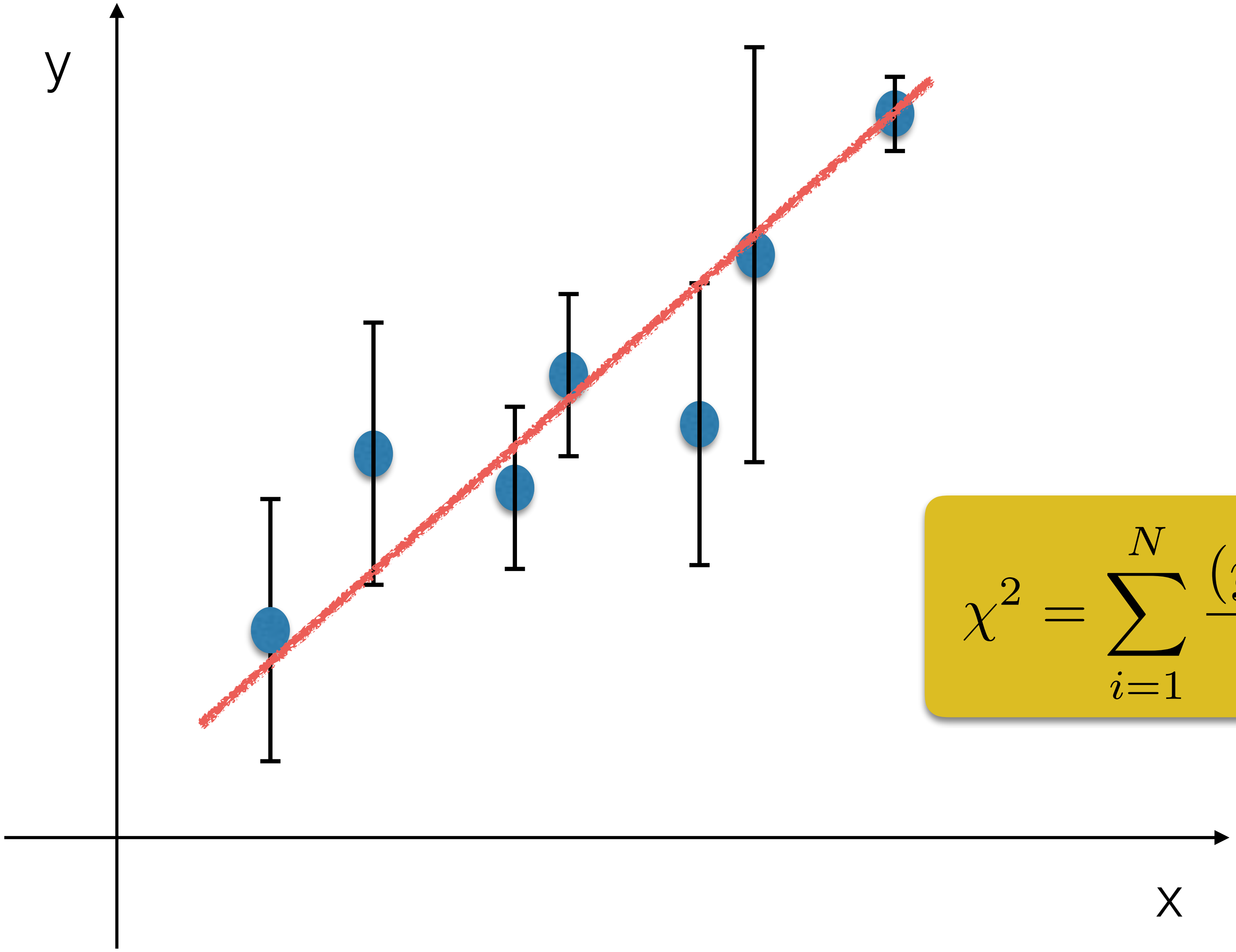
Fit with a straight line



What is fitting ?



What is fitting ?



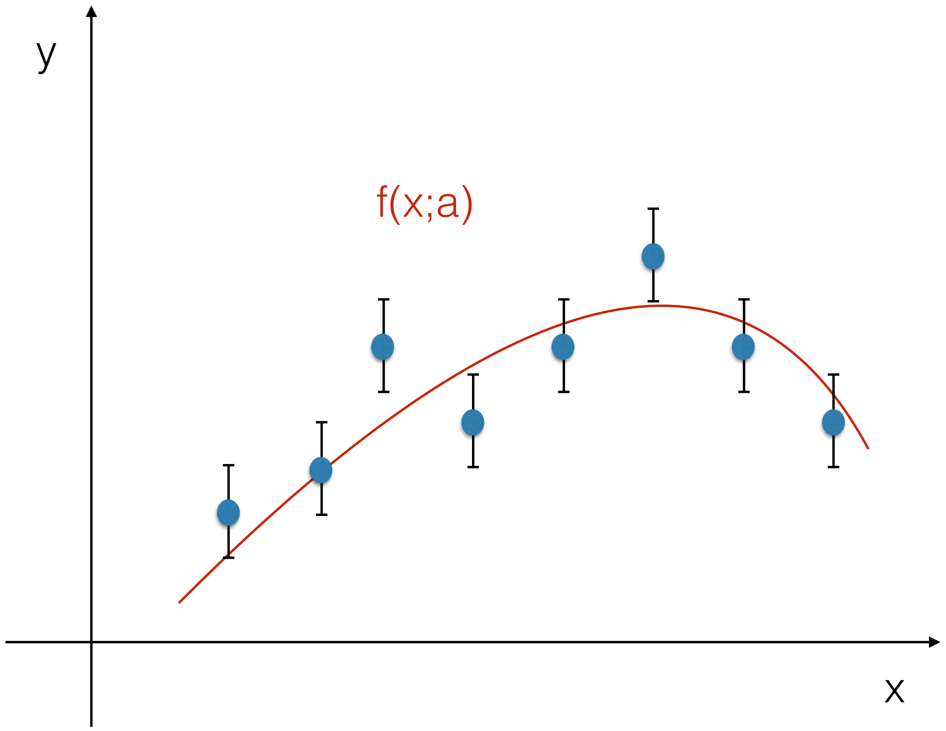
$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - f(x;a))^2}{\sigma_i^2}$$

Principle of Maximum Likelihood

Suppose you have a number of measurements (x_i, y_i) where i=1,N.

Assume you have no uncertainty on the values of x_i while you assume that the uncertainty on y are gaussian distributed: σ_i.

You convinced yourself that the data are described by a model y=f(x; a) that depends only on one parameter “a”.



How do you extract the best estimate of a, given the data you collected ?

Suppose you have the “true” value of a then your measurements would be distributed as $y_i = f(x_i;a)$. You could compute the probability to obtain one of those measurements:

$$P_a(y_i) \propto \frac{1}{\sigma_i} e^{\frac{-(y_i - f(x_i;a))^2}{2\sigma_i^2}}$$

(this is the gaussian assumption)

Principle of Maximum Likelihood

The probability to obtain the complete set of measurements is:

$$P_a(y_1, \ldots, y_N) = Prob_a(y_1) \cdots P_a(y_N) \quad \propto \cdot e^{\chi^2/2}$$

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - f(x;a))^2}{\sigma_i^2}$$

So far is just computing probability. Now invert the logic: you don’t know the true value of a and you want to extract that value from data: “best estimate from data” = $\hat{a}$
This value is the one that maximizes the probability to observe the measurements we have. “PRINCIPLE OF MAXIMUM LIKELIHOOD”

$$\frac{\partial \chi^2}{\partial a} = 0$$ which correspond to minimize the sum of the squares $$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - f(x;a))^2}{\sigma_y^2}$$

“LEAST SQUARES METHOD”

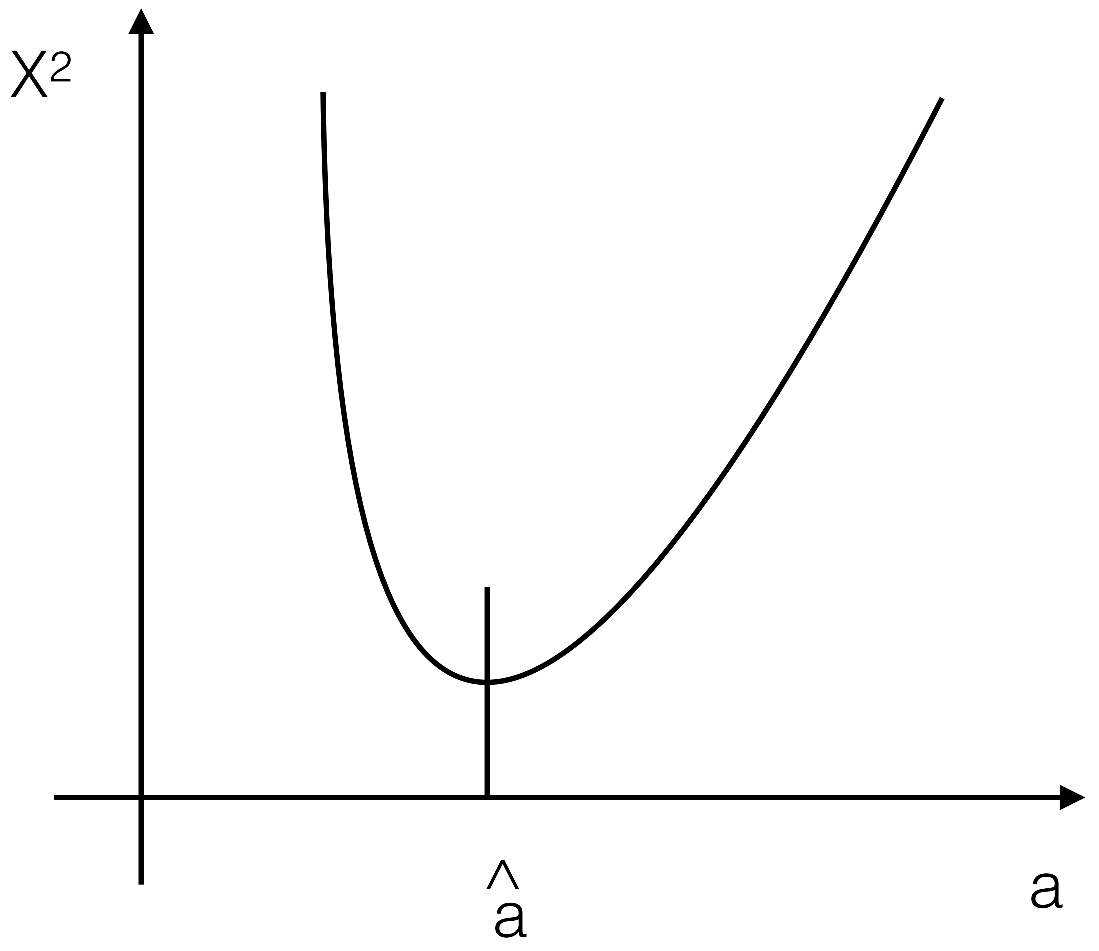
(the method is trivially generalized to any number of parameters by taking the corresponding partial derivatives)

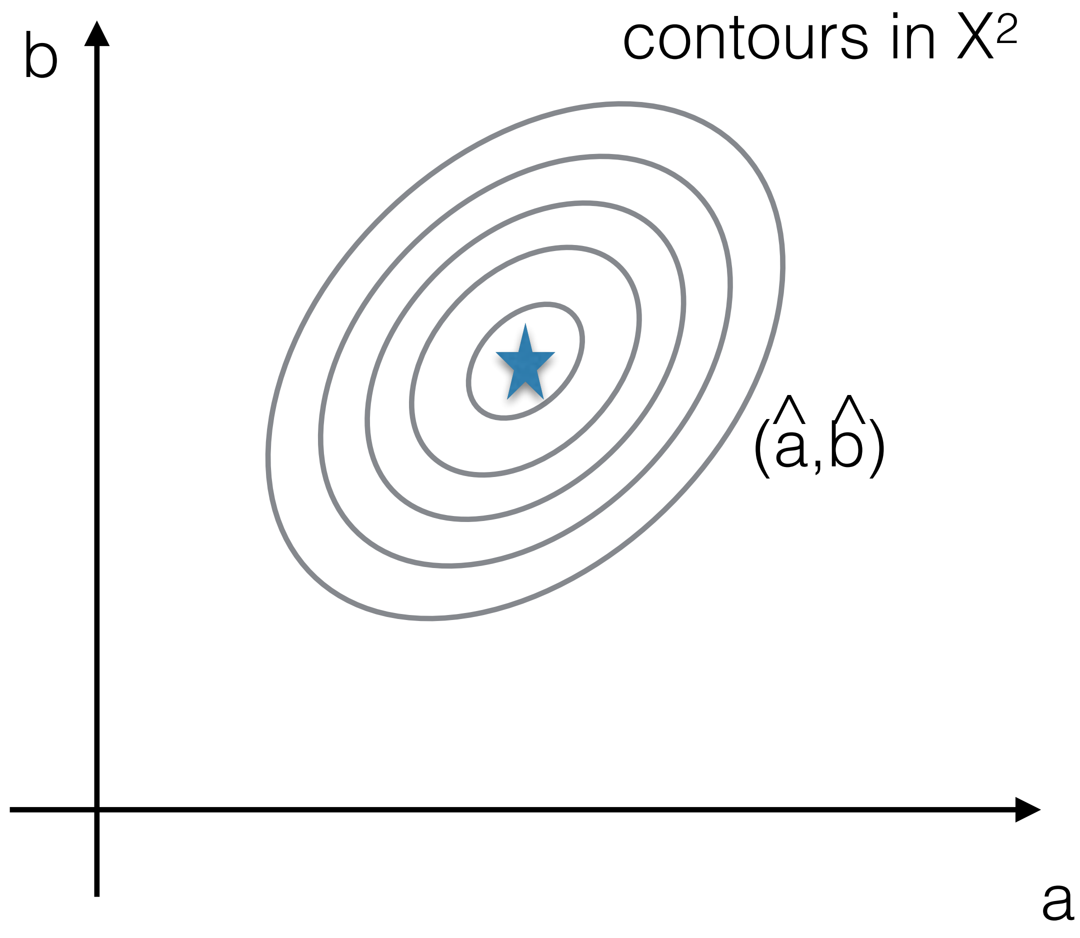
Least squares method

Interpret the chi-square as a “distance between the data and the model”, the best estimate of the parameter “a” is obtained by minimizing that distance:

$$\frac{d\chi^2}{da} = \sum_{i=1}^{N} \frac{d}{da} f(x_i;a) \cdot \frac{y_i - f(x_i;a)}{\sigma_i^2} = 0$$

(for p-parameters the minimization is to be performed in p-dimensions)





Example: f = mx

Assume a linear model for the data (x_i, y_i) and that the measurements are independent. The chi-square to minimize is:

$$\chi^2 = \sum_i \left(\frac{y_i - mx_i}{\sigma_i} \right)^2$$

Assume for simplicity that $\sigma_i = \sigma \ \forall i$

$$\begin{aligned} \frac{\partial \chi^2}{\partial m} &= -\frac{2}{\sigma^2} \sum_i (x_i y_i - m x_i^2) \\ & \sum_i (x_i y_i - m x_i^2) = 0 \\ \sum_i x_i y_i &= m \sum_i x_i^2 \\ \hat{m} &= \sum_i \frac{x_i y_i}{N\overline{x^2}} = \frac{\overline{xy}}{\overline{x^2}} \qquad \left(\sum x_i^2 = N\overline{x^2} \right) \end{aligned}$$

and by error propagation $V(\hat{m}) = \sum_i \left(\frac{x_i}{N\overline{x^2}} \right)^2 \sigma^2 = \frac{\sigma^2}{N\overline{x^2}}$

Example: f = mx + b

When the intercept is not zero:

$$\begin{aligned}\hat{m} &= \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} \\ \hat{b} &= \bar{y} - \hat{m}\bar{x}\end{aligned}$$

and the elements of the covariance matrix are:

$$\begin{aligned}\sigma_m^2 &= V(\hat{m}) = \frac{\sigma^2}{N(\overline{x^2} - \bar{x}^2)} \\ \sigma_b^2 &= V(\hat{b}) = \frac{\sigma^2\overline{x^2}}{N(\overline{x^2} - \bar{x}^2)} \\ cov(\hat{m}, \hat{b}) &= -\frac{\sigma^2\bar{x}}{N(\overline{x^2} - \bar{x}^2)}\end{aligned}$$