Basic Problem of Statistical Inference

Assume that we have a set of observations

$$S = \{x_1, x_2, \ldots, x_N\}, \quad x_j \in \mathbb{R}^n.$$

The problem is to infer on the underlying probability distribution that gives rise to the data S.

- Statistical modeling
- Statistical analysis.

Parametric or non-parametric?

- *Parametric problem*: The underlying probability density has a *specified form* and depends on *a number of parameters*. The problem is to infer on those parameters.
- *Non-parametric problem:* No analytic expression for the probability density is available. Description consists of defining the dependency/non-dependency of the data. Numerical exploration.

Typical situation for parametric model: The distribution is the probability density of a random variable $X : \Omega \to \mathbb{R}^n$.

- Parametric problem suitable for inverse problems
- Model for a learning process

LAW OF LARGE NUMBERS

General result ("Statistical law of nature"):

Assume that $X_1, X_2, \ldots$ are independent and identically distributed random variables with finite mean μ and variance σ^2. Then,

$$\lim_{n\to\infty} \frac{1}{n}\big(X_1 + X_2 + \cdots + X_n\big) = \mu$$

almost certainly.

Almost certainly means that with probability one,

$$\lim_{n\to\infty} \frac{1}{n}\big(x_1 + x_2 + \cdots + x_n\big) = \mu,$$

x_j being a realization of X_j.

Example

Sample

$$S = \{x_1, x_2, \ldots, x_N\}, \quad x_j \in \mathbb{R}^2.$$

Parametric model: x_j realizations of

$$X \sim \mathcal{N}(x_0, \Gamma),$$

with unknown mean $x_0 \in \mathbb{R}^2$ and covariance matrix $\Gamma \in \mathbb{R}^{2\times 2}$.

Probability density of X:

$$\pi(x \mid x_0, \Gamma) = \frac{1}{2\pi \mathrm{det}(\Gamma)^{1/2}} \exp\left(-\frac{1}{2}(x - x_0)^{\mathrm{T}}\Gamma^{-1}(x - x_0)\right).$$

Problem: Estimate the parameters x_0 and Γ.

The Law of Large Number suggests that we calculate

$$x_0 = \mathrm{E}\{X\} \approx \frac{1}{n}\sum_{j=1}^{n} x_j = \widehat{x}_0. \tag{1}$$

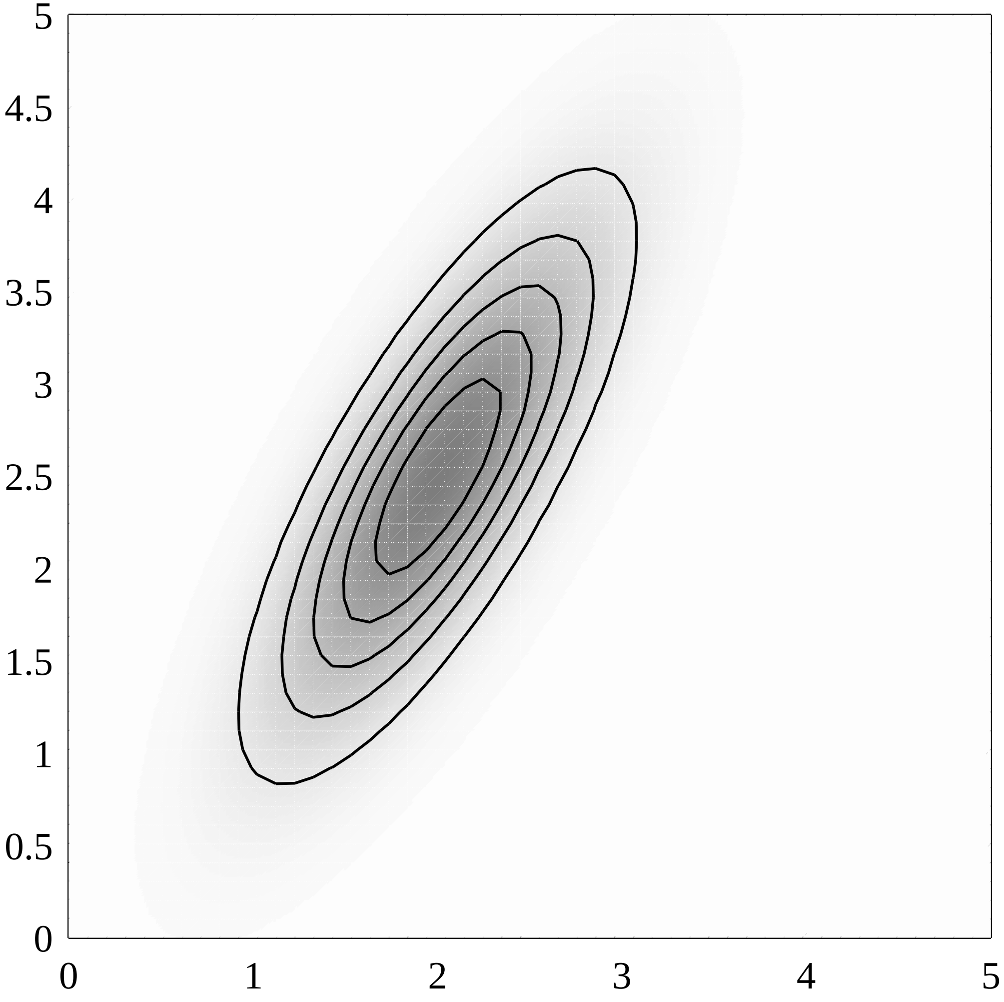
Covariance matrix: observe that if $X_1, X_2, \ldots$ are i.i.d, so are $f(X_1), f(X_2), \ldots$ for any function $f : \mathbb{R}^2 \mapsto \mathbb{R}^k$.

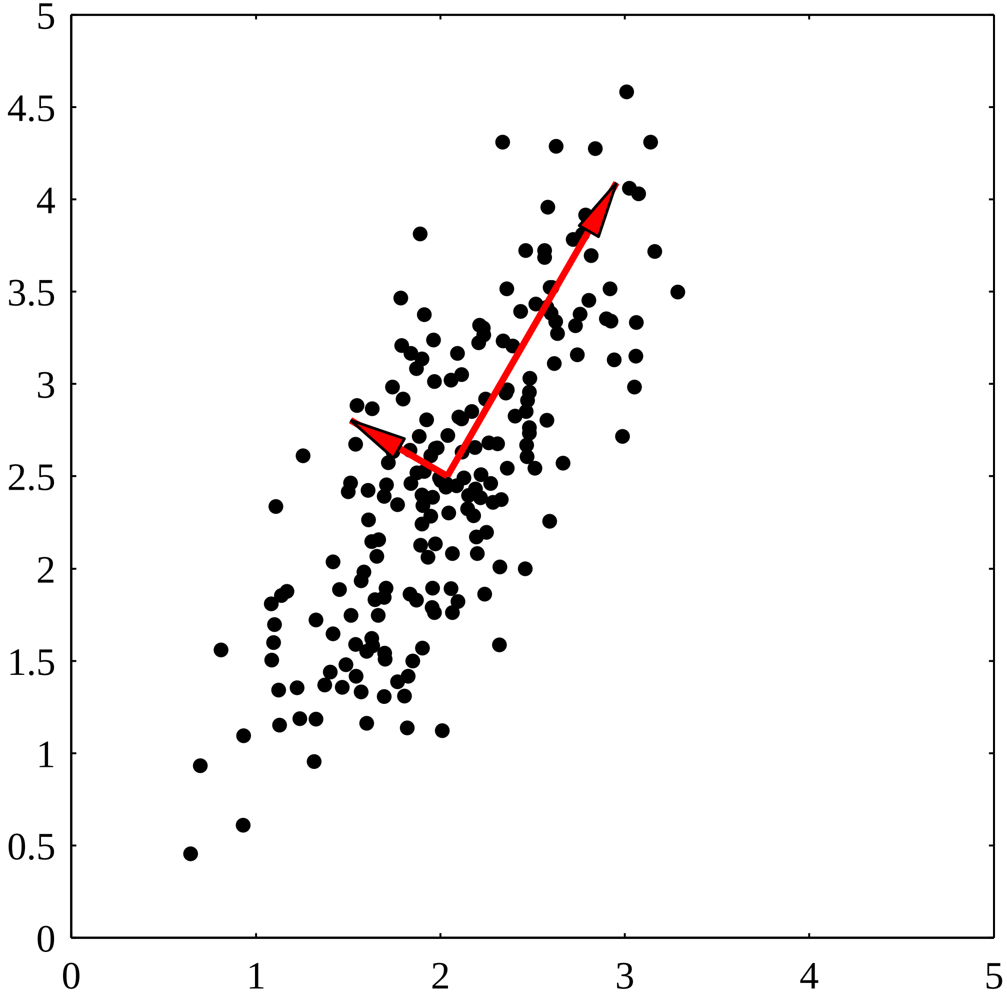
Try

$$\begin{aligned}
\Gamma &= \mathrm{cov}(X) = \mathrm{E}\{(X - x_0)(X - x_0)^{\mathrm{T}}\} \\
&\approx \mathrm{E}\{(X - \widehat{x}_0)(X - \widehat{x}_0)^{\mathrm{T}}\} \\
&\approx \frac{1}{n}\sum_{j=1}^{n}(x_j - \widehat{x}_0)(x_j - \widehat{x}_0)^{\mathrm{T}} = \widehat{\Gamma}.
\end{aligned} \tag{2}$$

Formulas (1) and (2) are known as *empirical mean and covariance*, respectively.

Case 1: Gaussian sample





Sample size $N = 200$.

Eigenvectors of the covariance matrix:

$$\widetilde{\Gamma} = UDU^{\mathrm{T}}, \tag{3}$$

where $U \in \mathbb{R}^{2\times 2}$ is an orthogonal matrix and $D \in \mathbb{R}^{2\times 2}$ is a diagonal,

$$U^{\mathrm{T}} = U^{-1}.$$

$$U = \left[\begin{array}{cc} v_1 & v_2 \end{array}\right], \quad D = \left[\begin{array}{cc} \lambda_1 & \\ & \lambda_2 \end{array}\right],$$

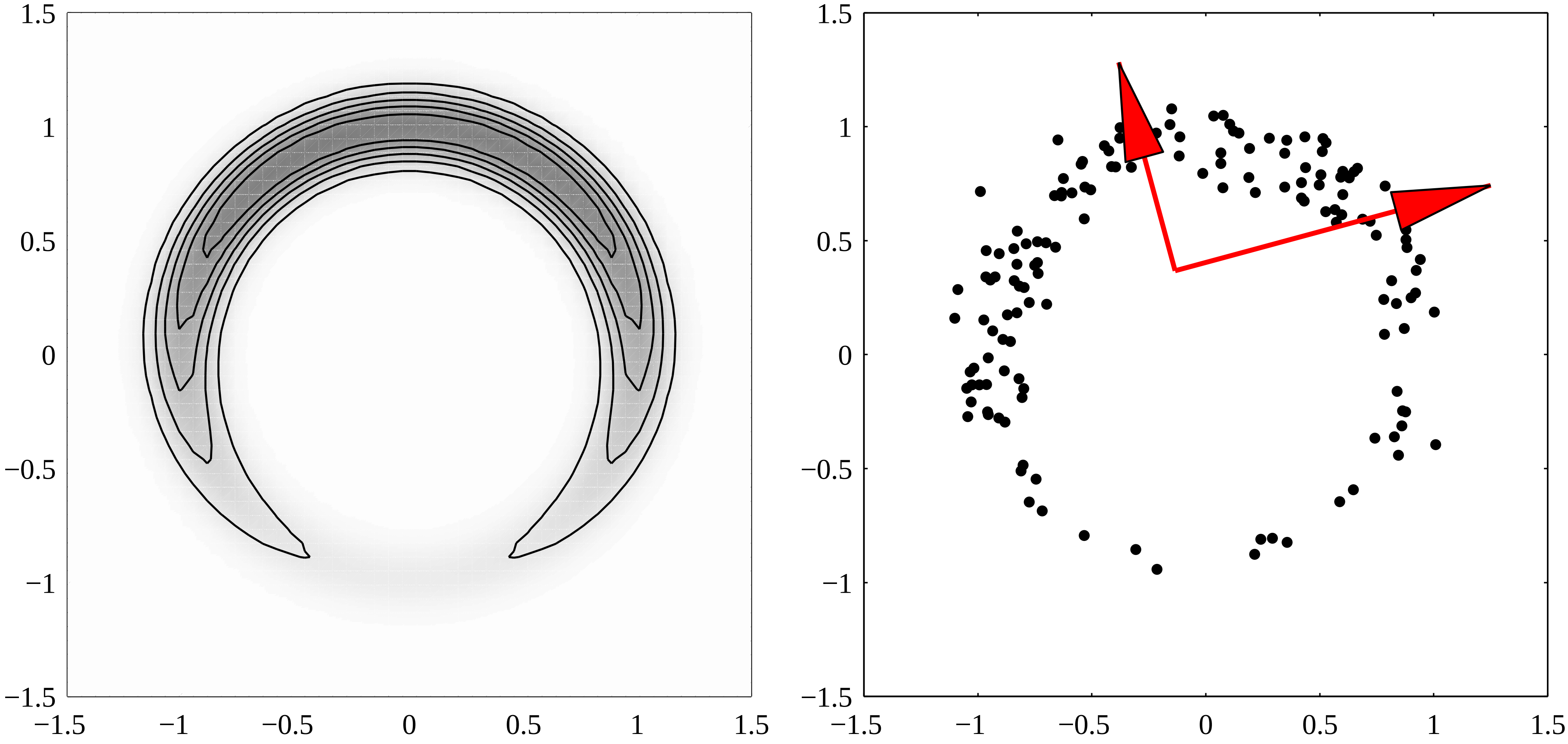
$$\widetilde{\Gamma} v_j = \lambda_j v_, \quad , j = 1, 2.$$

Scaled eigenvectors,

$$v_{j,\text{scaled}} = 2\sqrt{\lambda_j} v_j,$$

where $\sqrt{\lambda_j}$ =standard deviation (STD).

CASE 2: NON-GAUSSIAN SAMPLE



ESTIMATE OF NORMALITY/NON-NORMALITY

Consider the sets

$$B_\alpha = \{x \in \mathbb{R}^2 \mid \pi(x) \geq \alpha\}, \quad \alpha > 0.$$

If π is Gaussian, B_α is an ellipse or $\emptyset$.

Calculate the integral

$$\mathrm{P}\{X \in B_\alpha\} = \int_{B_\alpha} \pi(x)dx. \tag{4}$$

We call B_α the *credibility ellipse* with credibility p, $0 < p < 1$, if

$$\mathrm{P}\{X \in B_\alpha\} = p, \text{ giving } \alpha = \alpha(p). \tag{5}$$

Assume that the Gaussian density π has the center of mass and covariance matrix $\widetilde{x}_0$ and $\widetilde{\Gamma}$ estimated from the sample S of size N.

If S is normally distributed,

$$\#\big\{x_j \in B_{\alpha(p)}\big\} \approx pN. \tag{6}$$

Deviations due to non-normality.

How do we calculate the quantity?

Eigenvalue decomposition:

$$
\begin{aligned}
(x-\widetilde{x}_0)^{\mathrm{T}}\widetilde{\Gamma}^{-1}(x-\widetilde{x}_0) &= (x-\widetilde{x}_0)^{\mathrm{T}}UD^{-1}U^{\mathrm{T}}(x-\widetilde{x}_0) \\
&= \|D^{-1/2}U^{\mathrm{T}}(x-\widetilde{x}_0)\|^2,
\end{aligned}
$$

since U is orthogonal, i.e., $U^{-1} = U^{\mathrm{T}}$, and we wrote

$$
D^{-1/2} = \begin{bmatrix} 1/\sqrt{\lambda_1} & \\ & 1/\sqrt{\lambda_2} \end{bmatrix}.
$$

We introduce the change of variables,

$$
w = f(x) = W(x-\widetilde{x}_0), \quad W = D^{-1/2}U^{\mathrm{T}}.
$$

Write the the integral in terms of the new variable w,

$$\begin{aligned}\int_{B_\alpha} \pi(x)dx &= \frac{1}{2\pi\big(\det(\widetilde{\Gamma})\big)^{1/2}} \int_{B_\alpha} \exp\left(-\frac{1}{2}(x-\widetilde{x}_0)^{\mathrm{T}}\widetilde{\Gamma}^{-1}(x-\widetilde{x}_0)\right)dx \\ &= \frac{1}{2\pi\big(\det(\widetilde{\Gamma})\big)^{1/2}} \int_{B_\alpha} \exp\left(-\frac{1}{2}\|W(x-\widetilde{x}_0\|^2\right)dx \\ &= \frac{1}{2\pi}\int_{f(B_\alpha)} \exp\left(-\frac{1}{2}\|w\|^2\right)dw,\end{aligned}$$

where we used the fact that

$$dw = \det(W)dx = \frac{1}{\sqrt{\lambda_1\lambda_2}}dx = \frac{1}{\det(\widetilde{\Gamma})^{1/2}}dx.$$

Note:

$$\det(\widetilde{\Gamma}) = \det(UDU^{\mathrm{T}}) = \det(U^{\mathrm{T}}UD) = \det(D) = \lambda_1\lambda_2.$$

The equiprobability curves for the density for w are circles centered around the origin, i.e.,

$$f(B_\alpha) = D_\delta = \left\{ w \in \mathbb{R}^2 \mid \|w\| < \delta \right\}$$

for some $\delta > 0$.

Solve δ: Integrate in radial coordinates (r, θ),

$$\begin{aligned} \frac{1}{2\pi} \int_{D_\delta} \exp\left(-\frac{1}{2}\|w\|^2\right) dw &= \int_0^\delta \exp\left(-\frac{1}{2}r^2\right) r dr \\ &= 1 - \exp\left(-\frac{1}{2}\delta^2\right) = p, \end{aligned}$$

implying that

$$\delta = \delta(p) = \sqrt{2 \log\left(\frac{1}{1-p}\right)}.$$

To see if the sample points x_j is within the confidence ellipse with confidence p, it is enough to check if the condition

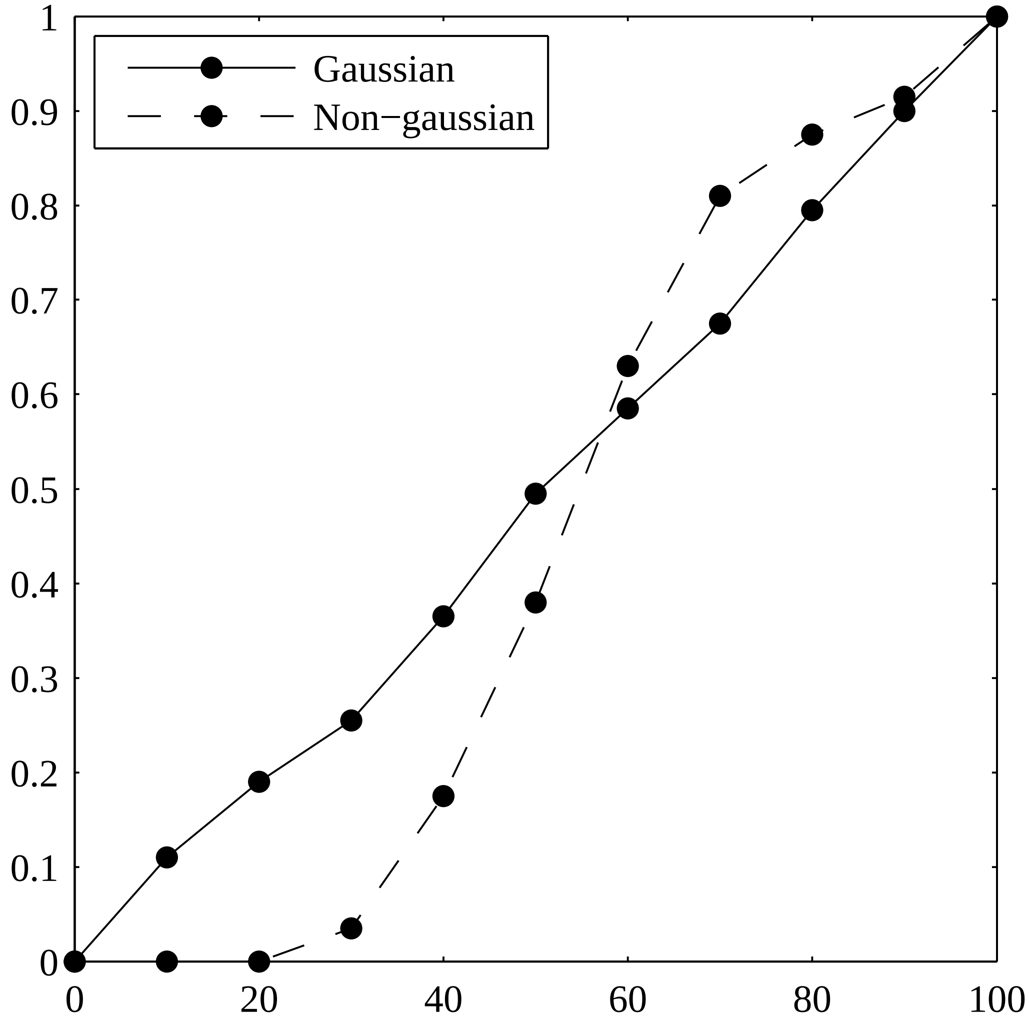
$$\|w_j\| < \delta(p), \quad w_j = W(x_j - \widetilde{x}_0), \quad 1 \leq j \leq N$$

is valid.

Plot

$$p \mapsto \frac{1}{N} \#\big\{x_j \in B_{\alpha(p)}\big\}$$

Example


Gaussian
Non−gaussian
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
20
40
60
80
100

Matlab code

```
N = length(S(1,:));           % Size of the sample
xmean = (1/N)*(sum(S')');     % Mean of the sample
CS = S - xmean*ones(1,N);     % Centered sample
Gamma = 1/N*CS*CS';           % Covariance matrix

% Whitening of the sample

[V,D] = eig(Gamma);           % Eigenvalue decomposition
W = diag([1/sqrt(D(1,1));1/sqrt(D(2,2))])*V';
WS = W*CS;                    % Whitened sample
normWS2 = sum(WS.^2);
```

```
% Calculating percentual amount of scatter points that are
% included in the confidence ellipses


 rinside = zeros(11,1);
 rinside(11) = N;
 for j = 1:9
     delta2 = 2*log(1/(1-j/10));
     rinside(j+1) = sum(normWS2<delta2);
 end
 rinside = (1/N)*rinside;

 plot([0:10:100],rinside,'k.-','MarkerSize',12)
```

Which one of the following formulae?

$$\widehat{\Gamma} = \frac{1}{N}\sum_{j=1}^{N}(x_j - \widehat{x}_0)(x_j - \widehat{x}_0)^{\mathrm{T}},$$

or

$$\widetilde{\Gamma} = \frac{1}{N}\sum_{j=1}^{N} x_j x_j^{\mathrm{T}} - \widetilde{x}_0\widetilde{x}_0^{\mathrm{T}}.$$

The former, please.

EXAMPLE

Calibration of a measurement instrument:

- Measure a dummy load whose output known
- Subtract from actual measurement
- Analyze the noise

Discrete sampling; Output is a vector of length n.

Noise vector $x \in \mathbb{R}^n$ is a realization of

$$X : \Omega \to \mathbb{R}^n.$$

Estimate mean and variance

$$x_0 = \frac{1}{n}\sum_{j=1}^{n} x_j \text{ (offset)}, \quad \sigma^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j - x_0)^2.$$

Improving Signal-to-Noise Ratio (SNR):

- Repeat the measurement
- Average
- Hope that the target is stationary

Averaged noise:

$$x = \frac{1}{N}\sum_{k=1}^{N} x^{(k)} \in \mathbb{R}^n.$$

How large must N be to reduce the noise enough?

Averaged noise x is a realization of a random variable

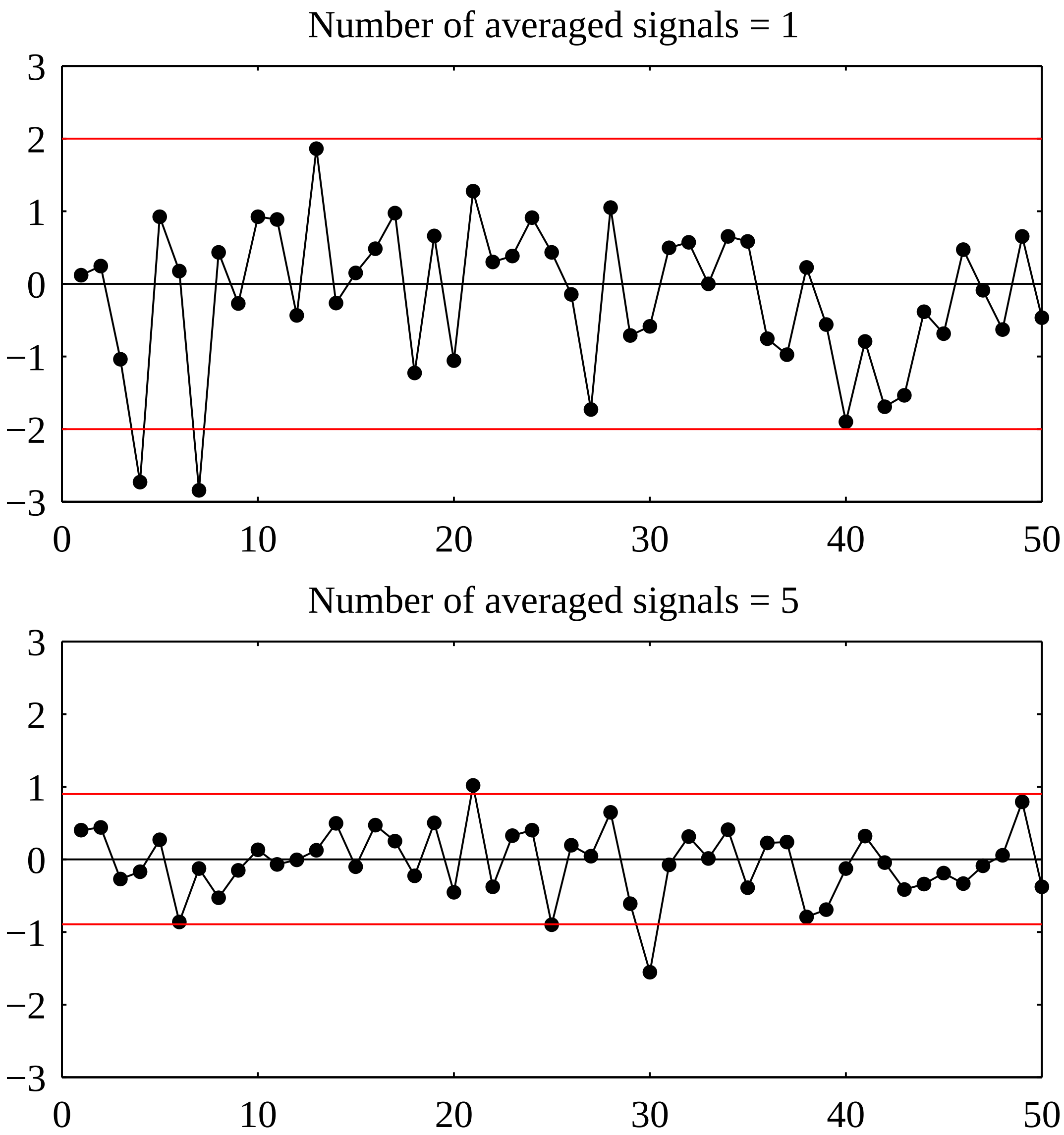
$$X = \frac{1}{N}\sum_{k=1}^{N} X^{(k)} \in \mathbb{R}^n.$$

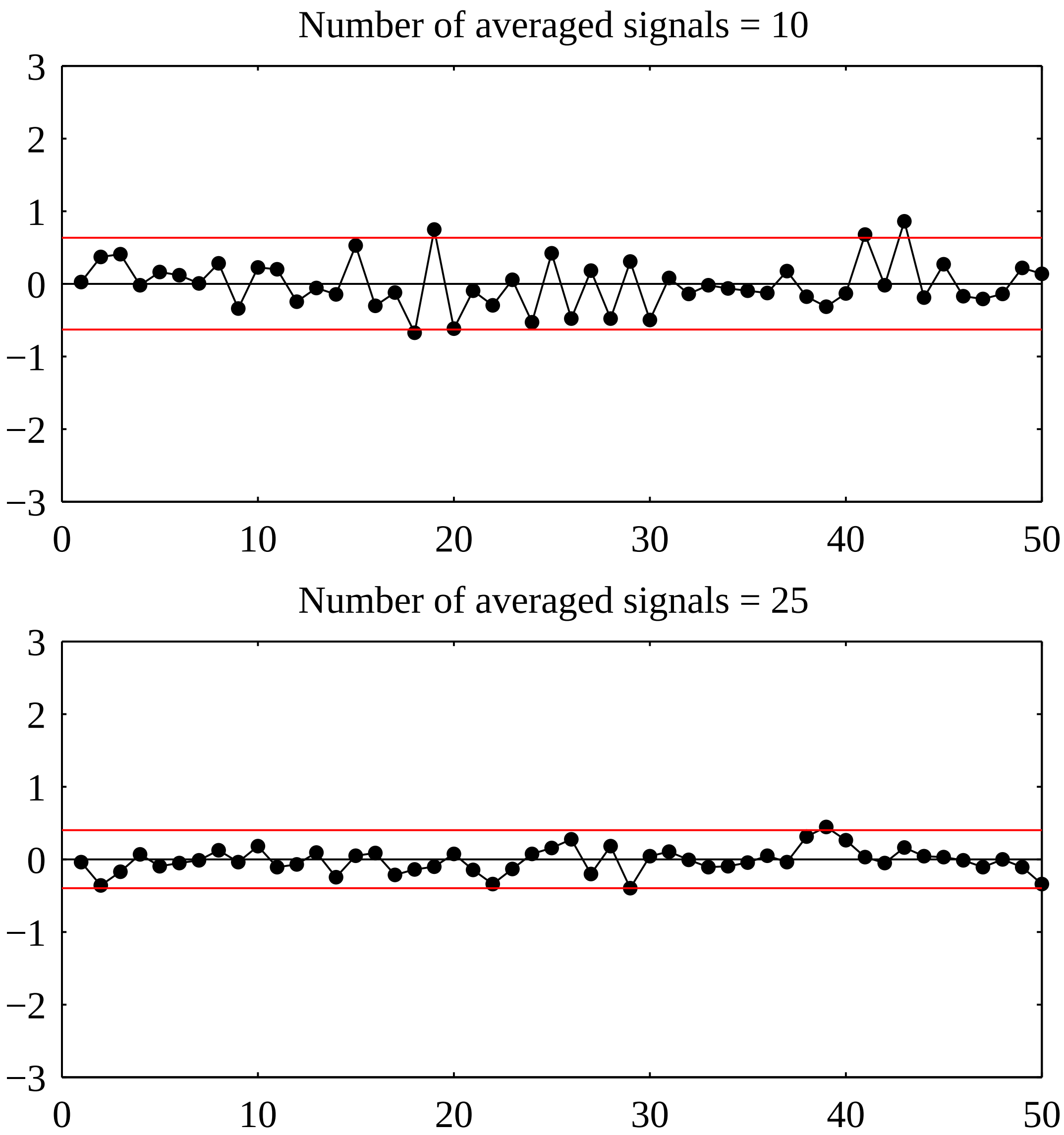
If $X^{(1)}, X^{(2)}, \ldots$ i.i.d., X is asymptotically Gaussian by Central Limit Theorem, and its variance is

$$\mathrm{var}(X) = \frac{\sigma^2}{N}.$$

Repeat until the variance is below a given threshold,

$$\frac{\sigma^2}{N} < \tau^2.$$


Number of averaged signals = 1
3
2
1
0
−1
−2
−3
0
10
20
30
40
50
Number of averaged signals = 5
3
2
1
0
−1
−2
−3
0
10
20
30
40
50


Number of averaged signals = 10
3
2
1
0
−1
−2
−3
0
10
20
30
40
50
Number of averaged signals = 25
3
2
1
0
−1
−2
−3
0
10
20
30
40
50

Maximum Likelihood Estimator: frequentist's approach

Parametric problem,

$$X \sim \pi_\theta(x) = \pi(x \mid \theta), \quad \theta \in \mathbb{R}^k.$$

Independent realizations: Assume that the observations x_j are obtained independently.

More precisely: $X_1, X_2, \ldots, X_N$ i.i.d, x_j is a realization of X_j.

Independency:

$$\pi(x_1, x_2, \ldots, x_N \mid \theta) = \pi(x_1 \mid \theta)\pi(x_2 \mid \theta) \cdots \pi(x_N \mid \theta),$$

or, briefly,

$$\pi(S \mid \theta) = \prod_{j=1}^{N} \pi(x_j \mid \theta),$$

Maximum likelihood (ML) estimator of θ = parameter value that maximizes the probability of the outcome:

$$\theta_{\rm ML} = \arg\max \prod_{j=1}^{N} \pi(x_j \mid \theta).$$

Define

$$L(S \mid \theta) = -\log(\pi(S \mid \theta)).$$

Minimizer of $L(S \mid \theta)$ = maximizer of $\pi(S \mid \theta)$.

Example

Gaussian model

$$\pi(x \mid x_0, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{1}{2\sigma^2}(x - x_0)^2\right), \quad \theta = \begin{bmatrix} x_0 \\ \sigma^2 \end{bmatrix} = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}.$$

Likelihood function is

$$\begin{aligned} \prod_{j=1}^{N} \pi(x_j \mid \theta) &= \left(\frac{1}{2\pi\theta_2}\right)^{N/2} \exp\left(-\frac{1}{2\theta_2}\sum_{j=1}^{N}(x_j - \theta_1)^2\right) \\ &= \exp\left(-\frac{1}{2\theta_2}\sum_{j=1}^{N}(x_j - \theta_1)^2 - \frac{N}{2}\log\left(2\pi\theta_2\right)\right) \\ &= \exp\left(-L(S \mid \theta)\right). \end{aligned}$$

We have

$$\nabla_\theta L(S \mid \theta) = \begin{bmatrix} \dfrac{\partial L}{\partial \theta_1} \\ \\ \dfrac{\partial L}{\partial \theta_2} \end{bmatrix} = \begin{bmatrix} -\dfrac{1}{\theta_2^2}\displaystyle\sum_{j=1}^N x_j + \dfrac{N}{\theta_2^2}\theta_1 \\ -\dfrac{1}{2\theta_2^2}\displaystyle\sum_{j=1}^N (x_j - \theta_1)^2 + \dfrac{N}{2\theta_2} \end{bmatrix}.$$

Setting $\nabla_\theta L(S \mid \theta) = 0$ gives

$$x_0 = \theta_{\mathrm{ML},1} = \frac{1}{N}\sum_{j=1}^N x_j,$$

$$\sigma^2 = \theta_{\mathrm{ML},2} = \frac{1}{N}\sum_{j=1}^N (x_j - \theta_{\mathrm{ML},1})^2.$$

Example

Parametric model

$$\pi(n \mid \theta) = \frac{\theta^n}{n!} e^{-\theta},$$

sample $S = \{n_1, \cdots, n_N\}$, $n_k \in \mathbb{N}$, obtained by independent sampling.

The likelihood density is

$$\pi(S \mid \theta) = \prod_{k=1}^{N} \pi(n_k) = e^{-N\theta} \prod_{k=1}^{N} \frac{\theta^{n_k}}{n_k!},$$

and its negative logarithm is

$$L(S \mid \theta) = -\log \pi(S \mid \theta) = \sum_{k=1}^{N} \big(\theta - n_k \log \theta + \log n_k!\big).$$

Derivative with respect to θ to zero:

$$\frac{\partial}{\partial\theta}L(S \mid \theta) = \sum_{k=1}^{N}\left(1 - \frac{n_k}{\theta}\right) = 0, \tag{7}$$

leading to

$$\theta_{\mathrm{ML}} = \frac{1}{N}\sum_{k=1}^{N} n_k.$$

Warning:

$$\mathrm{var}(N) \approx \frac{1}{N}\sum_{k=1}^{N}\left(n_k - \frac{1}{N}\sum_{j=1}^{N} n_j\right)^2,$$

which is different from the estimate of θ_{ML} obtained above.

Assume that θ is known *a priori* to be relatively large.

Use Gaussian approximation:

$$\begin{aligned}\prod_{j=1}^{N} \pi_{\text{Poisson}}(n_j \mid \theta) &\approx \left(\frac{1}{2\pi\theta}\right)^{N/2} \exp\left(-\frac{1}{2\theta}\sum_{j=1}^{N}(n_j-\theta)^2\right) \\ &= \left(\frac{1}{2\pi}\right)^{N/2} \exp\left(-\frac{1}{2}\left[\frac{1}{\theta}\sum_{j=1}^{N}(n_j-\theta)^2 + N\log\theta\right]\right).\end{aligned}$$

$$L(S \mid \theta) = \frac{1}{\theta}\sum_{j=1}^{N}(n_j-\theta)^2 + N\log\theta.$$

An approximation for $\theta_{\rm ML}$: Minimize

$$L(S \mid \theta) = \frac{1}{\theta}\sum_{j=1}^{N}(n_j - \theta)^2 + N\log\theta.$$

Write

$$\frac{\partial}{\partial\theta}L(S \mid \theta) = -\frac{1}{\theta^2}\sum_{j=1}^{N}(n_j - \theta)^2 - \frac{2}{\theta}\sum_{j=1}^{N}(n_j - \theta) + \frac{N}{\theta} = 0,$$

or

$$-\sum_{j=1}^{N}(n_j - \theta)^2 - 2\sum_{j=1}^{N}\theta(n_j - \theta) + N\theta = N\theta^2 + N\theta - \sum_{j=1}^{N}n_j^2 = 0,$$

giving

$$\theta = \left(\frac{1}{4} + \frac{1}{N}\sum_{j=1}^{N}n_j^2\right)^{1/2} - \frac{1}{2} \qquad \left(\neq \frac{1}{N}\sum_{j=1}^{N}n_j\right).$$

Example

Multivariate Gaussian model,

$$X \sim \mathcal{N}(x_0, \Gamma),$$

where $x_0 \in \mathbb{R}^n$ is unknown, $\Gamma \in \mathbb{R}^{n\times n}$ is symmetric positive definite (SPD) and known.

Model reduction: assume that x_0 depends on *hidden parameters* $z \in \mathbb{R}^k$ through a linear equation,

$$x_0 = Az, \quad A \in \mathbb{R}^{n\times k}, \quad z \in \mathbb{R}^k. \tag{8}$$

Model for an inverse problem: z is the true physical quantity that in the *ideal case* is related to the observable x_0 through the linear model (8).

Noisy observations:

$$X = Az + E, \quad E \sim \mathcal{N}(0, \Gamma).$$

Obviously,

$$\mathrm{E}\{X\} = Az + \mathrm{E}\{E\} = Az = x_0,$$

and

$$\mathrm{cov}(X) = \mathrm{E}\big\{(X - Az)(X - Az)^{\mathrm{T}}\big\} = \mathrm{E}\big\{EE^{\mathrm{T}}\big\} = \Gamma.$$

The probability density of X, given z, is

$$\pi(x \mid z) = \frac{1}{(2\pi)^{n/2}\det(\Gamma)^{1/2}} \exp\left(-\frac{1}{2}(x - Az)^{\mathrm{T}}\Gamma^{-1}(x - Az)\right).$$

Independent observations:

$$S = \{x_1, \ldots, x_N\}, \quad x_j \in \mathbb{R}^n.$$

Likelihood function

$$\prod_{j=1}^{N} \pi(x_j \mid z) \propto \exp\left(-\frac{1}{2} \sum_{j=1}^{N} (x_j - Az)^{\mathrm{T}} \Gamma^{-1} (x_j - Az) \right)$$

is maximized by minimizing

$$\begin{aligned} L(S \mid z) &= \frac{1}{2} \sum_{j=1}^{N} (x_j - Az)^{\mathrm{T}} \Gamma^{-1} (x_j - Az) \\ &= \frac{N}{2} z^{\mathrm{T}} \left[A^{\mathrm{T}} \Gamma^{-1} A \right] z - z^{\mathrm{T}} \left[A^{\mathrm{T}} \Gamma^{-1} \sum_{j=1}^{N} x_j \right] + \frac{1}{2} \sum_{j=1}^{N} x_j^{\mathrm{T}} \Gamma^{-1} x_j. \end{aligned}$$

Zeroing of the gradient gives

$$\nabla_z L(S \mid z) = N\big[A^{\mathrm{T}}\Gamma^{-1}A\big]z - A^{\mathrm{T}}\Gamma^{-1}\sum_{j=1}^{N} x_j = 0,$$

i.e., the maximum likelihood estimator z_{ML} is the solution of the linear system

$$\big[A^{\mathrm{T}}\Gamma^{-1}A\big]z = A^{\mathrm{T}}\Gamma^{-1}\overline{x}, \quad \overline{x} = \frac{1}{N}\sum_{j=1}^{N} x_j.$$

The solution may not exist; All depends on the properties of the model reduction matrix $A \in \mathbb{R}^{n\times k}$.

Particular case: *one* observation, $S = \{x\}$,

$$L(z \mid x) = (x - Az)^{\mathrm{T}}\Gamma^{-1}(x - Az).$$

Eigenvalue decomposition of the covariance matrix,

$$\Gamma = UDU^{\mathrm{T}},$$

or,

$$\Gamma^{-1} = W^{\mathrm{T}}W, \quad W = D^{-1/2}U^{\mathrm{T}},$$

we have

$$L(z \mid x) = \|W(Az - x)\|^2.$$

Hence, the problem reduces to a *weighted least squares problem*