Statistical and numerical methods of data analysis

Volker Blobel
Universität Hamburg

Sommersemester 2002

"The result of this experiment was inconclusive, so we had to use statistics."

(Overheard at international physics conference)

- *Statistical* methods are needed in order to extract meaningful and significant information from experimental data
- Data analysis requires stable *numerical* methods and efficient *algorithms*

Statistics helps to solve these problems

- *Point estimation:* Find the "best" value for a parameter (and make error propagation).
- *Interval estimation:* Find a range within which the true value should lie, with a given confidence.
- *Hypothesis testing:* Compare two hypotheses. Find which one is better supported by the data
- *Goodness-of-fit Testing:* Find how well one hypothesis is supported by the data
- *Decision making:* Make the best decision, based on data.

(Fred James)

These problem should not be mixed!

Example: neutrino mass

Data (1996/99) on the electron-neutrino mass m_ν (tritium β decay)

Experiment	m_ν^2 (eV^2)		m_ν 95 % C.L. limit
Tokyo	-65	$\pm 85 \pm 65$	< 13.0 eV
Los Alamos	-147	$\pm 68 \pm 41$	< 9.3 eV
Zürich	-24	$\pm 48 \pm 61$	< 11.7 eV
Mainz	-39	$\pm 34 \pm 15$	< 7.2 eV
Livermore (LLNL)	-130	$\pm 20 \pm 15$	< 7.0 eV
Moskau (INR)	-22	± 4.8	< 4.4 eV
Lobashev 1999	-1.9	$\pm 3.4 \pm 2.2$	< 2.5 eV
Weinheimer 1999	-3.7	$\pm 5.3 \pm 2.1$	< 2.8 eV

First error of m_ν^2 is statistical, second systematic. Upper limits for the neutrino mass are calculated using a certain method for 95 % confidence level.

PDG 2000: $m_\nu < 3\text{eV}$ Interpretation of tritium β decay is complicated by anomalies near the endpoint ("unknown effects cause an accumulation of events near the endpoint of the electron spectrum"), and the limits are not without ambiguity.

Example: Number of light neutrinos

Mark II collaboration at SLAC (1989): analysis which assumed Standard Model couplings but which fit for the Z mass and the number of neutrinos N_ν with result:

$$N_\nu = 2.8 \pm 0.6$$

95 % C.L. limit $N_\nu < 3.9$ excludes fourth massless neutrino

- Existence of at least three types on neutrinos was well established
- Ability to exclude $N_\nu = 4$ in this experiment due to downward fluctuation of measured N_ν from 3 to 2.8

PDG 2000: $N_\nu = 2.994 \pm 0.012$ from Standard Model fits to LEP data.

Literature

R. Barlow Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences, Wiley, 1989

V. Blobel and E. Lohrmann Statistische und numerische Methoden der Datenanalyse, Teubner, 1998

S. Brandt Datenanalyse, Spektrum, 1999

G. Cowan Statistical Data Analysis, Oxford, 1998

L. Lyons Statistics for nuclear and particle physicists, Cambridge, 1989

B. Roe Probability and Statistics in Experimental Physics, Springer, 1992

W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, Numerical Recipes in [. . . C, Fortran], The Art of Scientific Computing

R. Sedgewick, Algorithms in . . . , Addison-Wesley

Chapter titles

- Probability
- Example: Application of linear algebra
- Statistics
- Systematic errors
- Monte Carlo methods
- Algorithms
- Parameter estimation and information
- Numerical linear algebra
- Least squares method
- Confidence intervals and limits
- The maximum likelihood method
- Function minimization – optimization
- Case study: track recognition and reconstruction
- Multivariate analysis and classification
- Inverse problems – deconvolution

A. Probability

"The theory of probability and statistics is concerned with situtations in which an observation is made and which contains an element of inherent variability outwith the observer's control. He knows that if he repeated the observation under conditions which were identical in so far as he could control them, the second observation would not necessarily agree exactly with the first." (S.D. Silvey, Statistical inference)

Historically the interest in statistical problems started with investigations of games of chance.

In games of chance there are often many different outcomes, which a-priori are equally likeli (tossing a coin, throwing a dice); this can be used to make a *combinatorical* definition of probability.

Similarly one can give an *empiricial* definition of probability/frequency.

Two definitions of probability

Statistical methods are based on calculations of probability.

In mathematics, probability is an abstract concept which obeys certain rules.

There are two operational definitions of probability:

- Frequentists probability is defined as the limiting frequency of favorable outcomes in a large number of *identical* experiments.
  Probability can be measured. Used in most scientific work because it is objective. It is the probability of Quantum Mechanics. Can only be applied to repeatable phenomena.

- Bayesian probability is defined as the degree of belief in a favourable outcome of a *single* experiment.
  More general. Can be applied to unrepeatable phenomena.

Frequentists definition of probability

Combinatorical definition. If an event can happen in n different ways, which are all equal probable, and if k of the different ways have the probability $\mathcal{A}$, then the probability for event $\mathcal{A}$ is given by

$$P(\mathcal{A}) = \frac{k}{n}.$$

Empirical definition of probability. An observation is repeated independently under identical conditions n times; if property $\mathcal{A}$ is observed in k of the n cases, the ratio k/n is the empirical probability for $\mathcal{A}$. The probability $P(\mathcal{A})$ can be defined as the ratio in the limit for n going to infinity:

$$P(\mathcal{A}) = \lim_{n\to\infty} \frac{k}{n}.$$

Both definitions can be critisized: the first one can be considered as a vicious circle, and the second one is based on an impossible experiment.

Formal definition of probability – Kolmogorov 1931

Elementary events e_i are considered, which *exclude* each other:

$$\begin{aligned} e_i &= \text{elementary event} \\ \Omega &= \text{set of all elementary events} \\ P(e_i) &= \text{probability of event } e_i \end{aligned}$$

The probability for elementary events has to follow the following rules:

1. $P(e_i) \geq 0$ for all i
2. $P(e_i \vee e_j) = P(e_i) + P(e_j) \quad (\vee = \text{or})$
3. $\sum_\Omega P(e_i) = 1$

Non-elementary events $\mathcal{A}$ and $\mathcal{B}$ are formed from elementary events.

Combination of probability

Given two types of events, $\mathcal{A}$ and $\mathcal{B}$. $P(\mathcal{A})$ is the probability for the occurence of event $\mathcal{A}$ and $P(\mathcal{B})$ is the probability for the occurence of event $\mathcal{B}$.

$P(\mathcal{A} \text{ or } \mathcal{B}) = P(\mathcal{A}) + P(\mathcal{B}) - P(\mathcal{A} \text{ and } \mathcal{B})$	$\mathcal{A}$ or $\mathcal{B}$
$P(\mathcal{A} \text{ and } \mathcal{B}) = 0$	if $\mathcal{A}$ and $\mathcal{B}$ exclude each other
$P(\mathcal{A} \text{ or } \mathcal{B}) = P(\mathcal{A}) + P(\mathcal{B})$	if $\mathcal{A}$ and $\mathcal{B}$ exclude each other
$P(\mathcal{A} \text{ or } \overline{\mathcal{A}}) = P(\mathcal{A}) + P(\overline{\mathcal{A}}) = 1$	$(\overline{\mathcal{A}} = \text{not } \mathcal{A})$
$P(\mathcal{A} \text{ and } \mathcal{B}) = P(\mathcal{A}) \cdot P(\mathcal{B}\|\mathcal{A}) = P(\mathcal{B}) \cdot P(\mathcal{A}\|\mathcal{B})$	$\mathcal{A}$ and $\mathcal{B}$
$P(\mathcal{A} \text{ and } \mathcal{B}) = P(\mathcal{A}) \cdot P(\mathcal{B})$	if $\mathcal{A}$ and $\mathcal{B}$ independent

$P(\mathcal{B}|\mathcal{A})$ is the probability for event $\mathcal{B}$, if event $\mathcal{A}$ has happened, and is called *conditional* probability.

Bayes' theorem

It relates the two conditional probabilities $P(\mathcal{A}|\mathcal{B})$ and $P(\mathcal{B}|\mathcal{A})$:

$$\boxed{P(\mathcal{A}|\mathcal{B}) = P(\mathcal{B}|\mathcal{A}) \cdot \frac{P(\mathcal{A})}{P(\mathcal{B})}}$$

which follows directly from

$$P(\mathcal{A} \text{ and } \mathcal{B}) = P(\mathcal{A}) \cdot P(\mathcal{B}|\mathcal{A}) = P(\mathcal{B}) \cdot P(\mathcal{A}|\mathcal{B})$$

Generalized to n classes of events with properties $\mathcal{A}_i$, $(i = 1, 2, \ldots n)$ and $P(\mathcal{B}) = \sum_i P(\mathcal{B}|\mathcal{A}_i) \cdot P(\mathcal{A}_i)$

$$P(\mathcal{A}_i|\mathcal{B}) = \frac{P(\mathcal{B}|\mathcal{A}_i) P(\mathcal{A}_i)}{\sum_{i=1}^{n} P(\mathcal{B}|\mathcal{A}_i) \cdot P(\mathcal{A}_i)}$$

Special interpretation of Bayes' theorem $\rightarrow$ Bayesian probability.

Reverend Thomas Bayes (1702-1761), "An essay towards solving a problem in the doctrine of chances, Philos. Trans. R. Society **53** (1763) 370, reprinted in Biometrika **45** (1958) 293

Application of Bayes theorem

A test for a rare, very dangerous illnes K is made.
The result is positive! What does it mean?

The illness K is rare: a-priori $P(K) = 0.001 \qquad P(\bar{K}) = 0.999$

The test result is not safe, their are known error probabilities:

in case of illness	$P(+\|K) = 0.98$	$P(-\|K) = 0.02$
no illness	$P(+\|\bar{K}) = 0.03$	$P(-\|\bar{K}) = 0.97$

Application of Bayes theorem to calculate $P(K|+)$:

$$\begin{aligned} P(K|+) &= \frac{P(+|K)\cdot P(K)}{P(+|K)\cdot P(K) + P(+|\bar{K})\cdot P(\bar{K})} = \frac{0.98\cdot 0.001}{0.98\cdot 0.001 + 0.03\cdot 0.999} \\ &= 0.033 \end{aligned}$$

Thus the probability to have illness K after a positive test is 3.3 %.

Subjective probability – the Bayesian view

$A, B, \ldots$ are hypotheses, i.e. statements, that are either true of false
$P(A) =$ degree of belief that A is true.

This definition of "probability" is consistent with the Kolmogorov axioms.
Subjective probability allows a more natural treatment of phenomena that are not repeatable.

Examples for hypotheses:

- the particle is a K meson
- nature is supersymmetric
- the Higgs mass is 115 GeV
- it will rain tomorrow

Sometimes one has to take decisions on the basis of degree of belief (take an umbrella or not).

Relative frequency – the Frequentists view

A, B, ... are outcomes of a repeatable experiment
$P(A)$ is limit of frequency

$$P(A) = \lim_{n\to\infty} \frac{\text{outcome is } A}{n}$$

This frequency interpretation is often most natural for data analysis in physics.

Examples for frequencies:

- decay probability of $J/\Psi \to \mu^+\mu^-$ $(\Gamma_i/\Gamma = (5.88 \pm 0.10)\%)$
- detector acceptance
- upper limit of electron neutrino mass is 3 eV (95 % C.L.)

The probability of hypotheses A, B, ... is either 1 (true) of 0 (false).

Point estimation – Frequentist

An **Estimator** is a function of the data which will be used to *estimate* (physicists: *measure*) the value x of the parameter.

Task: Find that function which give estimates of x "closest to the true value" (highest precision, zero or small bias) for x. The probability $P\,(\text{data}|\text{true value of } x)$ is known, and the estimate is a function of the data.

Under very general conditions, the best estimator will be the one which maximizes the **Likelihood** $\mathcal{L}\,(x)$.

Point estimation – Bayesian

For point estimation, rewrite Bayes' Theorem:

$$P(\text{hyp}|\text{data}) = \frac{P(\text{data}|\text{hyp})\, P(\text{hyp})}{P(\text{data})}$$

and if the hypothesis concerns the value x of a parameter:

$$P(x|\text{data}) = \frac{P(\text{data}|x)\, P(x)}{P(\text{data})}$$

which is a probability density function (pdf) in the unknown x. This pdf must be normalized: $\int_x P(x|\text{data})\, \mathrm{d}x = 1$.

Assigning names to the different factors:

$$\text{Posterior probability}\,(x) \propto \mathcal{L}(x) \times \text{Prior probability}\,(x)$$

with point estimate taken at the maximum value of the Posterior probability. Objective choice of Prior probability?

Example: Frequentists and Bayesian probability

Mathematical urn with N_R red balls and N_W white balls (urn with replacement).

Given $N_R \geq 0$ $N_W \geq 0$ $N_T = N_R + N_W > 0$

Required: estimate for $P_R = N_R/N_T$

	Frequentist	Bayesian
no ball	P unknown (no data)	$P = 1/2$
100 balls, 40 red	$P = 0.40 \pm 0.05$ (68 % C.L.)	$P = 0.405$
1 ball, red	$P \neq 0$	P $= 0.6$

B. Example: Application of linear algebra

Solution of system of linear equations

$$\boldsymbol{y} \text{ measured} \qquad \boxed{\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}} \qquad \boldsymbol{x} \text{ unknown}$$

Discussion of a special case with migration matrix

$$\boldsymbol{A} = \begin{pmatrix} 1-\varepsilon & \varepsilon & 0 & 0 & 0 \\ \varepsilon & 1-2\varepsilon & \varepsilon & 0 & 0 \\ 0 & \varepsilon & 1-2\varepsilon & \varepsilon & 0 \\ 0 & 0 & \varepsilon & 1-2\varepsilon & \varepsilon \\ 0 & 0 & 0 & \varepsilon & 1-\varepsilon \end{pmatrix}$$

Numerical values in simulation: dimension $n = m = 20$ and $\varepsilon = 0.22$.

Solution in this special case is rather simple:

$$\widehat{\boldsymbol{x}} = \boldsymbol{A}^{-1}\boldsymbol{y} \qquad \boldsymbol{V}(\widehat{\boldsymbol{x}}) = \boldsymbol{A}^{-1}\boldsymbol{V}_y \left(\boldsymbol{A}^{-1}\right)^T$$

but result often wildly oscillating.

Orthogonal decomposition

Matrix $\boldsymbol{A}$ is decomposed into a diagonal matrix $\boldsymbol{D}$ and a rotation matrix $\boldsymbol{U}$:

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T \qquad \boldsymbol{D} = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{pmatrix} \qquad \boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{1}$$

$$\text{Transform:} \quad \boldsymbol{U}^T \cdot \ | \qquad \begin{aligned} \boldsymbol{y} &= \boldsymbol{A}\boldsymbol{x} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T\boldsymbol{x}\,. \\ \underbrace{\boldsymbol{U}^T\boldsymbol{y}}_{\boldsymbol{c}} &= \boldsymbol{D}\underbrace{\boldsymbol{U}^T\boldsymbol{x}}_{\boldsymbol{b}} \end{aligned}$$

Vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are transformed to vectors $\boldsymbol{b}$ and $\boldsymbol{c}$:

$$\begin{aligned} \boldsymbol{x} \to \boldsymbol{b} &= \boldsymbol{U}^T\boldsymbol{x} \qquad \text{or} \qquad b_j = \boldsymbol{u}_j^T\boldsymbol{x} \qquad j = 1\ldots n \\ \boldsymbol{y} \to \boldsymbol{c} &= \boldsymbol{U}^T\boldsymbol{y} \qquad \text{or} \qquad c_j = \boldsymbol{u}_j^T\boldsymbol{y} \qquad j = 1\ldots n \end{aligned}$$

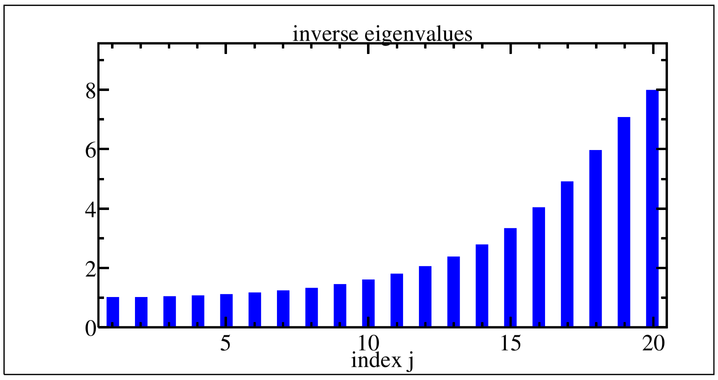
$$\text{back transformation} \quad \boldsymbol{x} = \boldsymbol{U}\boldsymbol{b} \qquad \boldsymbol{y} = \boldsymbol{U}\boldsymbol{c}$$

Eigenvalues

Transitions $\boldsymbol{x} \leftrightarrow \boldsymbol{y}$ are now done by multiplication/divison of the coefficients b_j and c_j by the eigenvalues:

Folding	$\boldsymbol{x} \to \boldsymbol{y}$	$b_j \to b_j \cdot \lambda_j = c_j$	$j = 1 \ldots n$
Unfolding	$\boldsymbol{y} \to \boldsymbol{x}$	$c_j \to \dfrac{c_j}{\lambda_j} = b_j$	$j = 1 \ldots n\,.$

. . . but look at the inverse of the eigenvalue for $n = m = 20$ and $\varepsilon = 0.22$:

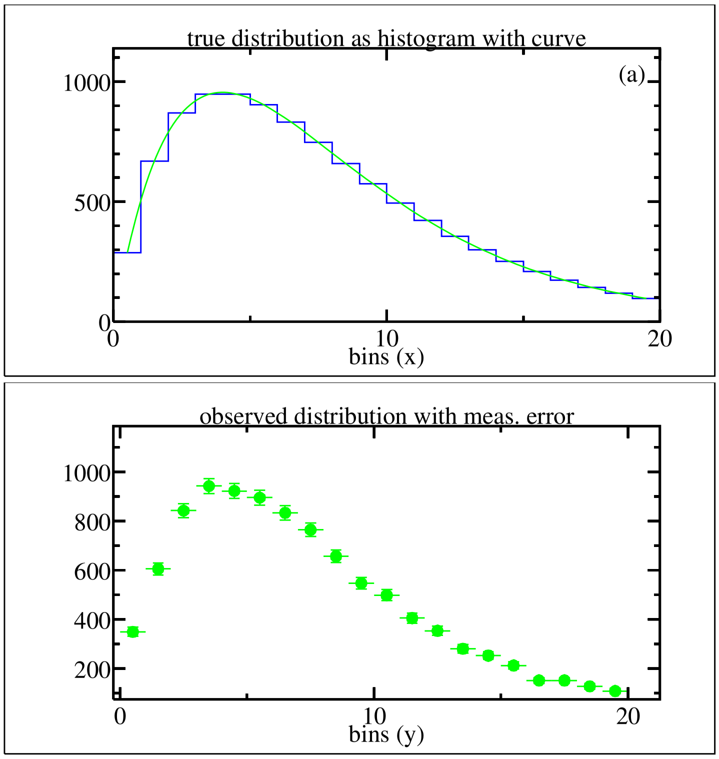


Simulation of a measurement

Starting from an ideal distribution (shape of $x \exp(-ax)$) a measured distribution is simulated:

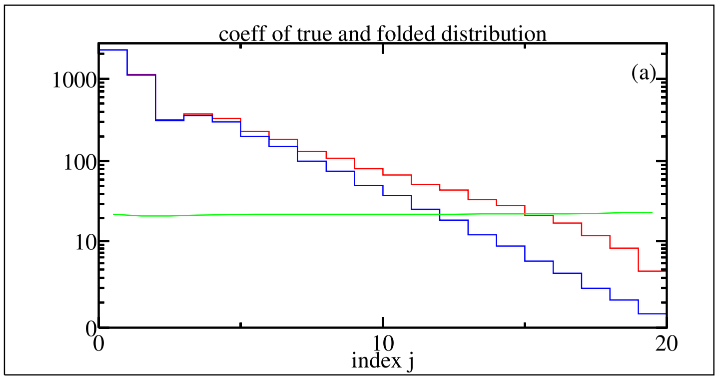
top original true distribution

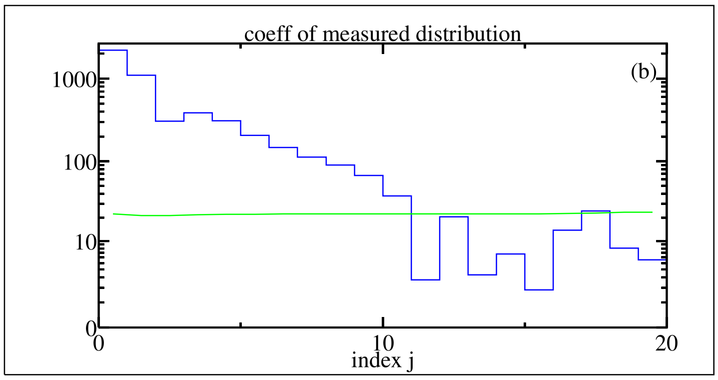
bottom simulated measured distribution



Coefficients

Transformation from distribution to coefficients:





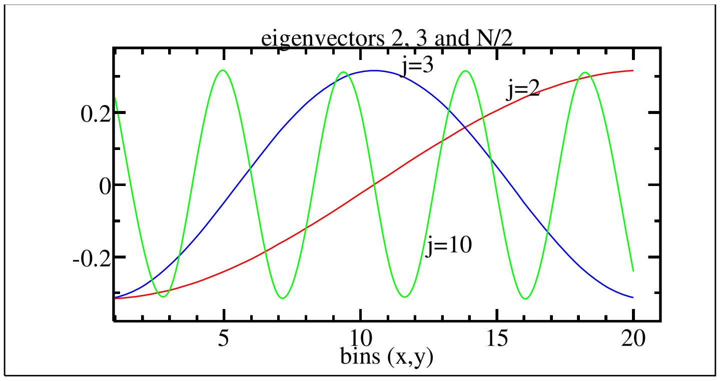
top coefficients b_j and c_j of the true and the folded distribution (no measurement errors)

bottom coefficients c_j of the measured distribution (with measurement errors)

Green constant line indicates level of statistical errors.

Eigenvectors

The eigenvectors can be visualized by a function interpolating the bins.

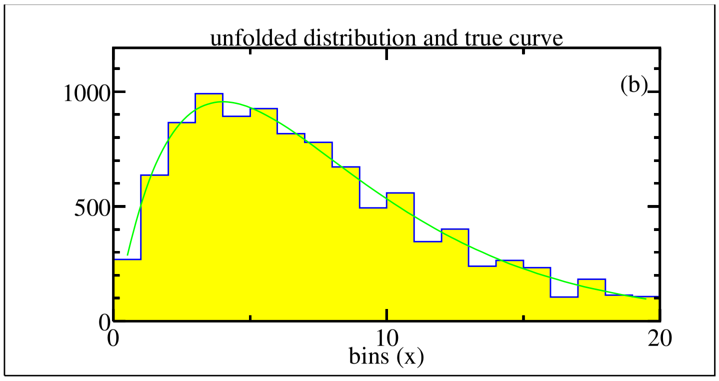


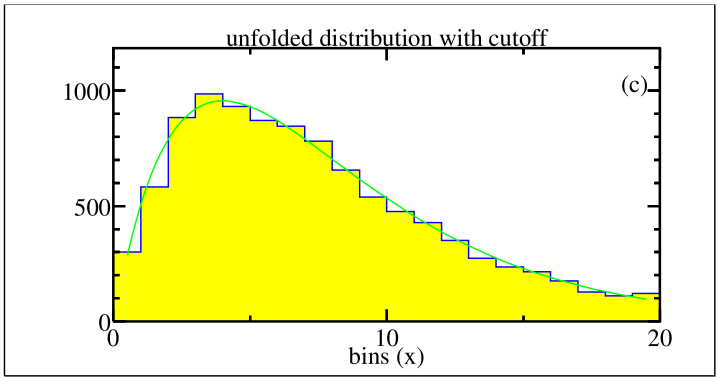
Components of $\boldsymbol{u}_1$ with eigenvalue 1 all have the same value $\rightarrow$ coefficient c_1 is proportional to the sum of the distribution.

There are $j-1$ zeros in eigenvector $\boldsymbol{u}_j$ $\rightarrow$ more and more oscillating with increasing index j.

Solution of linear system of equations

...without and with sharp cut in the number of contributions:





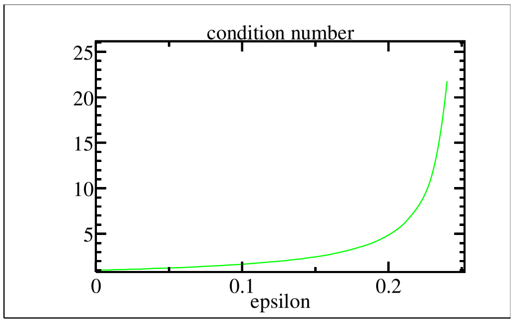
top (unfolding) with all 20 coefficients

bottom (unfolding) with a sharp cut-off after 10 coefficients

Condition number and migration parameter

The plot below shows the condition number as a function of the migration parameter ε:

$$\text{condition number} \quad \kappa = \frac{\lambda_{\text{max}}}{\lambda_{\text{min}}}$$



Sharp increase above $\varepsilon = 0.2$.