Goodness-of-fit tests in many dimensions

André van Hameren

Institut für Physik, Johannes-Gutenberg-Universität,
Mainz, Germany

andrevh@thep.physik.uni-mainz.de

Zeuthen, May 2005

Introduction

Why Goodness-of-fit tests?

- In scientific research, experiment and theory often come as close together as a sample of data and a PDF following which these data are predicted to be distributed.
- Sometimes, the sample is used to fix a number of free parameters in the theory which are translated into parameters for the PDF.
- In the end, however, one has to convince oneself that the PDF actually makes sense.
- One approach is to determine how “probable” it is to find the particular sample assuming that the PDF is correct.
- The translation of this question into the same for the value of a *statistic* evaluated at the sample is a goodness-of-fit (GOF) tests.

The structure of goodness-of-fit tests

A GOF statistic is a function T_N of the data sample $\{\omega_i\}_{i=1}^{N}$ constructed such that, under the hypothesis that the data are i.i.d. distributed in a space Ω following the theoretical PDF P, there is a number t_∞ such that

$$\lim_{N\to\infty} T_N(\{\omega_i\}_{i=1}^{N}) = t_\infty \ .$$

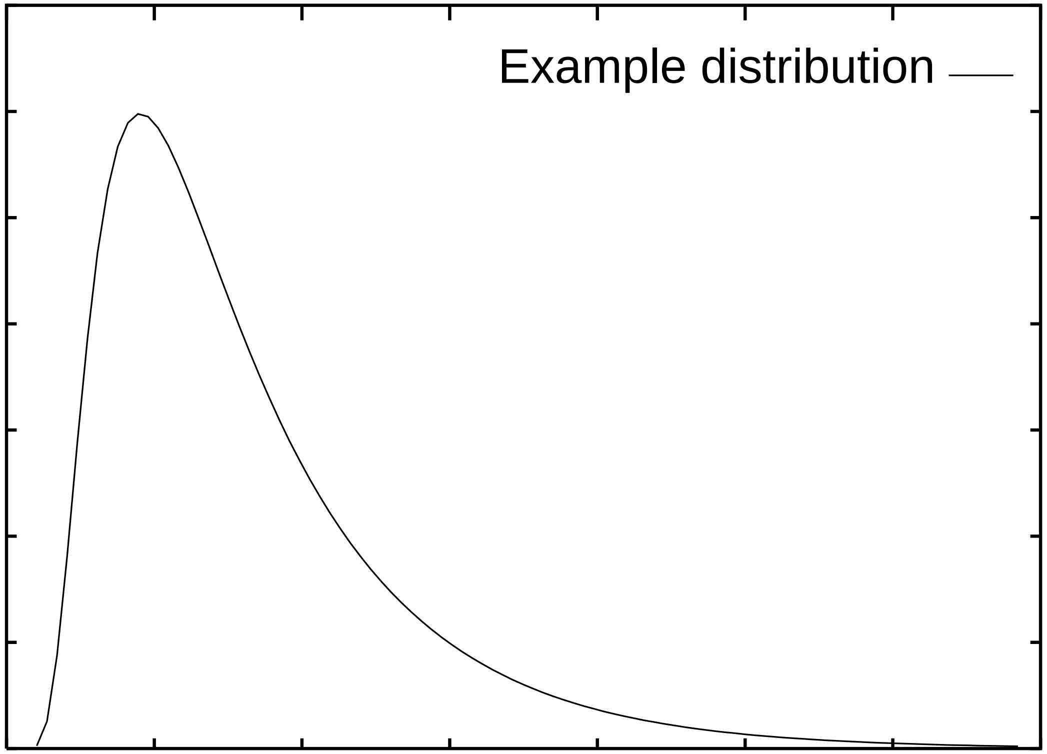
In order to test realistic samples, we need the probability distribution of this function under the hypothesis that the data are indeed distributed following P:

$$\mathcal{P}_N(t) = \int \delta(\,t - T_N(\{\omega_i\}_{i=1}^{N})\,) \prod_{i=1}^{N} P(\omega_i) d\omega_i \ .$$

If a data sample is distributed following P, then one expects $T_N(\{\omega_i\}_{i=1}^{N})$ to be a number under the bulk of $\mathcal{P}_N$.

The structure of goodness-of-fit tests

$\mathcal{P}_N(t)$



$$t = T_N(\{\omega_i\}_{i=1}^N)$$

Difficulty in many dimensions

- It is inconvenient when $\mathcal{P}_N$ depends on P:
  - It can be hard to calculate $\mathcal{P}_N$;
  - We would prefer to have universal tests.
- The straightforward solution is to map the data-points $\omega \mapsto \varphi(\omega)$ such that the Jacobian of the mapping satisfies

$$\left|\det \frac{\partial \varphi^k(\omega)}{\partial \omega^l}\right| = P(\omega) \ .$$

  If the original data are distributed following P, the mapped data are distributed following the uniform distribution
  $\Longrightarrow$ we only need test statistics for the uniform distribution.
- Easy only for one-dimensional cases: $\varphi(\omega) = \int_{-\infty}^{\omega} P(\omega')\, d\omega'$.
- Examples are the Kolmogorov-Smirnov and the Cramér-von Mises test.

Construction of GOF statistics

If $\{\omega_i\}_{i=1}^{N}$ is distributed following P over $\Omega = [0,1)^{\Delta}$, and fP quadratically integrable over Ω, then

$$\langle f \rangle_N \stackrel{N\to\infty}{\longrightarrow} \langle fP \rangle \ .$$

So choose a class of functions $\mathcal{F}$ and a measure μ on $\mathcal{F}$ and define

$$T_N = \left(\int_{\mathcal{F}} |\langle f \rangle_N - \langle fP \rangle|^r \, \mu(df) \right)^{1/r} \ .$$

- We restrict ourselves to the quadratic case $r = 2$, with μ Gaussian.
- Take the square, and add a factor N so that the distribution does not become trivial for $N \to \infty$.
- Introduce a help-function Q to stay flexible for now:

$$T_N = N \int_{\mathcal{F}} |\langle fQ \rangle_N - \langle fQP \rangle|^2 \, \mu(df) \ .$$

Construction of GOF statistics

In terms of the two-point function c of μ, the statistic is

$$
\begin{aligned}
T_N &= \frac{1}{N}\sum_{i,j=1}^{N} c(\omega_i,\omega_j)Q(\omega_i)Q(\omega_j) \\
&- 2\int_\Omega \left(\sum_{i=1}^{N} c(\omega_i,\eta)Q(\omega_i)Q(\eta)\right)P(\eta)d\eta \\
&+ N\int_\Omega\int_\Omega c(\eta,\zeta)Q(\eta)Q(\zeta)\,P(\eta)P(\zeta)d\eta d\zeta\ .
\end{aligned}
$$

So the computational complexity grows quadratically with the sample size N. The expectation value

$$
E(T_N) = \int_\Omega c(\eta,\eta)Q(\eta)^2\,P(\eta)d\eta - \int_\Omega\int_\Omega c(\eta,\zeta)Q(\eta)Q(\zeta)\,P(\eta)P(\zeta)d\eta d\zeta
$$

is independent of N. The variance is $(1-2/N)\times$ [independent of N].

Spectral decomposition

Gaussian measures are completely determined by their two-point function, which we shall use as starting point.

We define a two-point function through its spectral decomposition: choose a set of functions $(u_n)_{n=1}^M$ and weights $(\sigma_n^2)_{n=1}^M$ and define

$$c(\eta, \zeta) = \sum_{n=1}^{M} \sigma_n^2 u_n(\eta) u_n(\zeta) \ .$$

A member f from $\mathcal{F}$ is then given in terms of some string $(x_n)_{n=1}^M$ by

$$f = \sum_{n=1}^{M} x_n u_n \ ,$$

and the measure μ can be written as

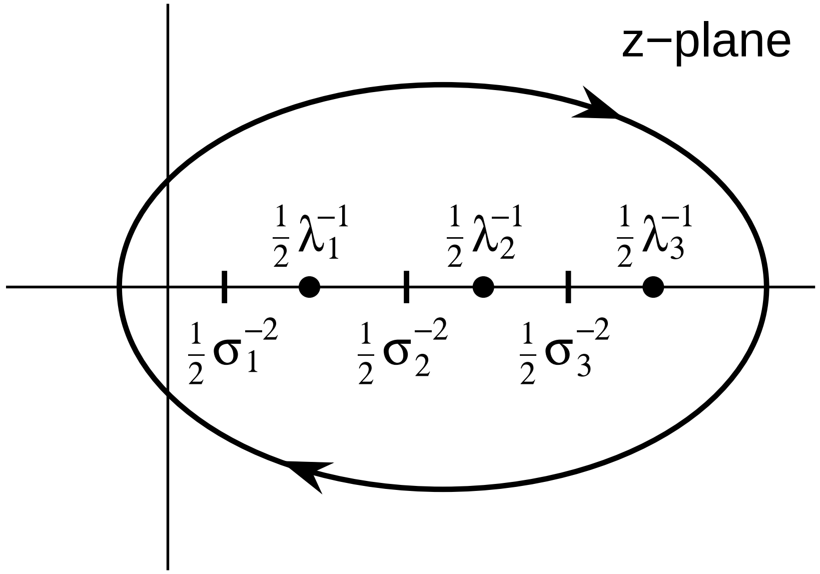
$$\mu(df) = \prod_{n=1}^{M} \frac{\exp(-\frac{1}{2} x_n^2 / \sigma_n^2) dx_i}{\sqrt{2\pi\sigma_n^2}} \ .$$

Calculating the asymptotic distribution

$$T_N = \frac{1}{N}\sum_{i,j=1}^{N} \frac{c(\omega_i,\omega_j)}{\sqrt{P(\omega_i)P(\omega_j)}} - 2\int_\Omega \left(\sum_{i=1}^{N} \frac{c(\omega_i,\eta)}{\sqrt{P(\omega_i)}} \right) \sqrt{P(\eta)}\, d\eta$$
$$+ N \int_\Omega \int_\Omega c(\eta,\zeta) \sqrt{P(\eta)P(\zeta)}\, d\eta\, d\zeta$$

$$c(\eta,\zeta) = \sum_{n=1}^{M} \sigma_n^2 u_n(\eta) u_n(\zeta)$$

$$\mathcal{P}_N(t) = \int_{-\varepsilon-i\infty}^{-\varepsilon+i\infty} \frac{dz}{2\pi i}\, e^{-tz}\, G_N(z)$$



$$G_{N\to\infty}(z) = \left(\prod_{n=1}^{M} (1 - 2z\lambda_n) \right)^{-1/2}$$

Gaussian limits

- For discrepancies (*i.e.* for $P = 1$) it has been shown [vanHameren *et al* 1997] that $\mathcal{P}_{N\to\infty}$ becomes Gaussian when $\Delta \to \infty$, if and only if

$$\frac{\lambda_1}{\sum_{n=1}^{M} \lambda_n} \xrightarrow{\Delta\to\infty} 0 \; .$$

- One can conclude that $\mathcal{P}_{N\to\infty}$ becomes Gaussian when $\Delta \to \infty$ for any P if and only if

$$\frac{\sigma_1^2}{\sum_{n=1}^{M} \sigma_n^2} \xrightarrow{\Delta\to\infty} 0 \; .$$

- <u>Conjecture:</u> P affects $\mathcal{P}_{N\to\infty}$ almost only through the expectation value and the variance for any $\Delta > 1$.

Example

GOF statistic based on *diaphony*

$$u_{\vec{n}}(\eta) = \prod_{k=1}^{\Delta} u_{n_k}(\eta^k) \quad , \quad n_k = 0,1,2,\dots \quad , \quad k = 1,2,\dots,\Delta \ ,$$

$$u_0(x) = 1 \quad , \quad u_{2n-1}(x) = \sqrt{2}\sin(2\pi n x) \quad , \quad u_{2n}(x) = \sqrt{2}\cos(2\pi n x) \ ,$$

for n from 1 to ∞. The corresponding weights are given by

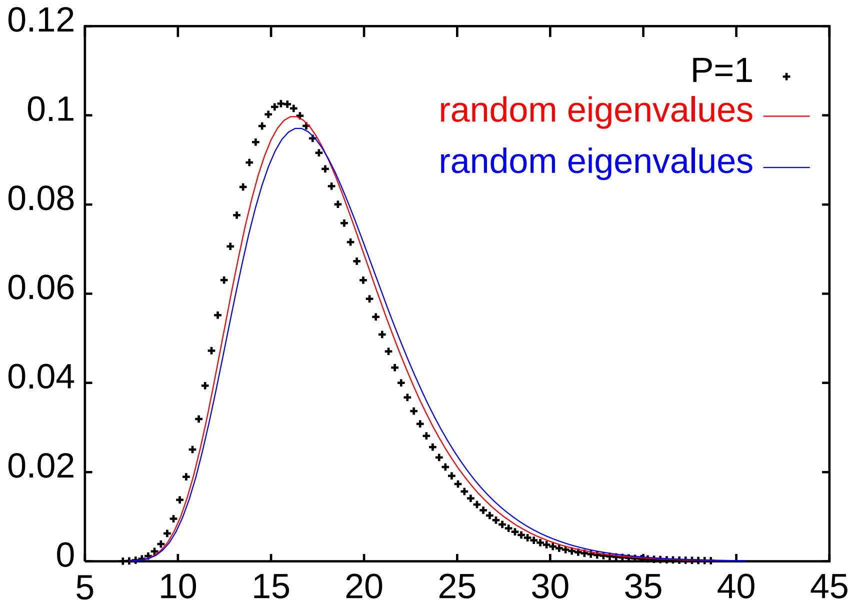
$$\sigma_{\vec{n}} = \prod_{k=1}^{\Delta} \sigma_{n_k} \qquad \text{with} \qquad \sigma_0 = 1 \quad , \quad \sigma_{2n-1} = \sigma_{2n} = \frac{1}{n} \ .$$

The two-point function is equal to

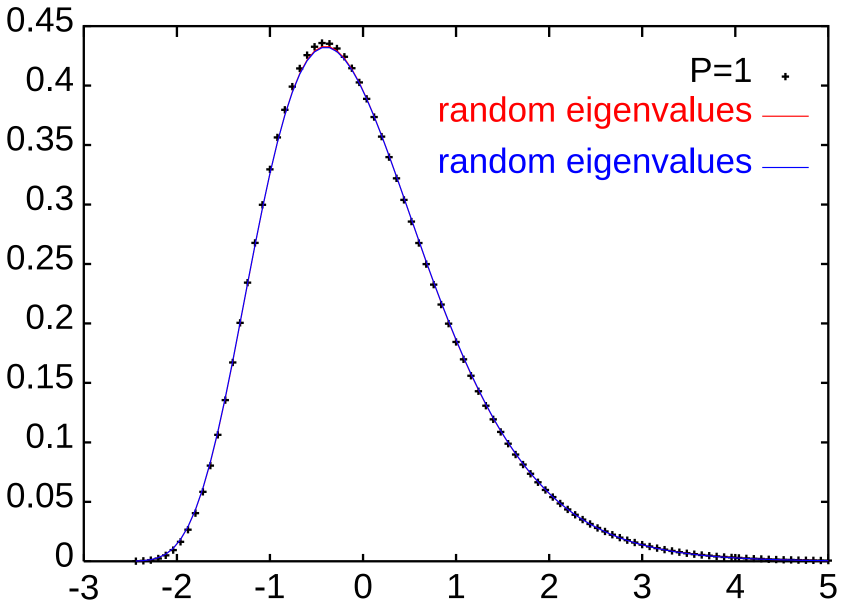
$$\begin{aligned} c(\eta,\zeta) &= \sum_{\vec{n}} \sigma_{\vec{n}}^2 u_{\vec{n}}(\eta) u_{\vec{n}}(\zeta) = \prod_{k=1}^{\Delta} c_1(\eta^k - \zeta^k) \\ c_1(x) &= 1 + \frac{\pi^2}{3} - 2\pi^2 x(1-x) \ . \end{aligned}$$

Example

$$\Delta = 2$$



$\mathcal{P}_{N \to \infty}(t)$



$\sqrt{V}\,\mathcal{P}_{N \to \infty}(\sqrt{V}\,t + E)$

$$E(T_N) = \int_\Omega c(\eta, \eta)\, d\eta - \int_\Omega \int_\Omega c(\eta, \zeta) \sqrt{P(\eta) P(\zeta)}\, d\eta\, d\zeta$$

$$V(T_N) = \int_\Omega \int_\Omega c(\eta, \zeta)^2\, d\eta\, d\zeta + \cdots$$

Conclusion

- We have seen how to construct practical GOF statistics to test the hypothesis that a sample of data-points is distributed in a many-dimensional space following a given PDF;
- The asymptotic test-distribution in the limit of an infinite sample size becomes Gaussian in the limit of infinite dimension.
- We have seen an explicit example of such a statistic, for which the asymptotic test-distribution depends on the PDF almost only through the expectation value and the variance of the statistic for any dimension larger than one