Basic Statistical Analysis of Lattice QCD Data

Gregorio Herdoiza

DESY Zeuthen

Lattice Practices 2008

Outline

Statistical analysis of Monte Carlo data (Markov chain)

- Basic statistics
- Correlation, autocorrelation
- Resampling methods
- Confidence intervals

Lattice QCD data correlations

- Correlations in data generated in in lattice QCD Monte Carlo (MC) simulation:
  - Correlation in MC time $\rightsquigarrow$ autocorrelation
  - Correlated measurements : different observables coming from same ensemble
  - Correlation in Euclidean time-space $\rightsquigarrow$ Green functions

- Neglecting the correlations implies:
  - Underestimating the error
  - Error propagation when combining measurement?
  - Invalidate fitting: goodness-of-fit, error on parameters

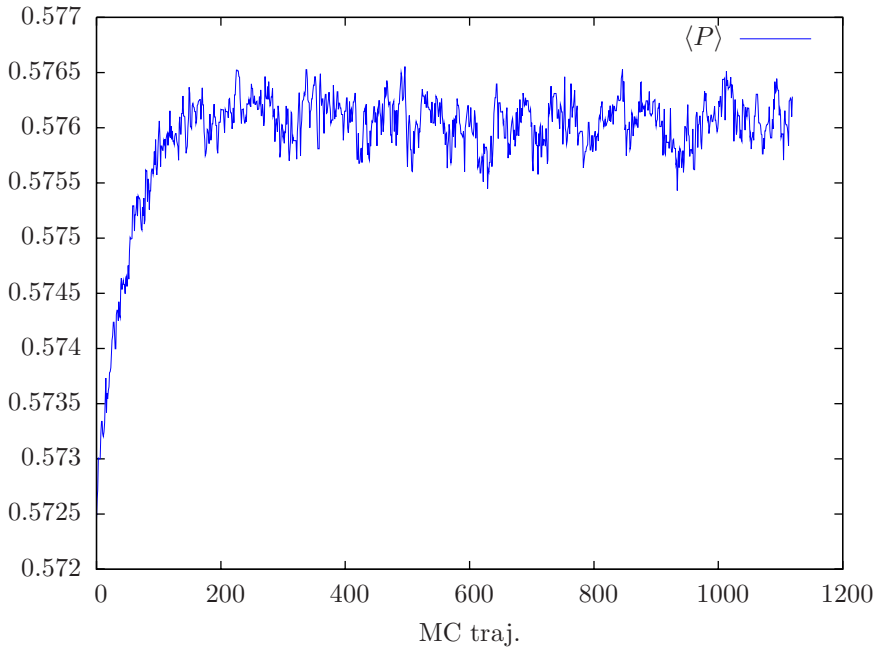
Lattice QCD data equilibrium

- MC simulation contains two phases
  - Equilibration or thermalization
  - Production: equilibrium expectation values are obtained from this statistics

- When is the system thermalized?
  - measure exponential autocorrelation time τ_{exp} : longest autocorrelation time in the system
  - pick N_{therm} such that $\tau_{\text{exp}} \ll N_{\text{therm}}$
  - visual inspection of MC time evolution

Lattice QCD data plaquette

Monte Carlo time evolution of the plaquette (short range quantity)
from a lattice QCD simulation with dynamical quarks



Identify the equilibrium phase and perform statistical analysis ...

Some definitions

Before discussing about autocorrelations, some basic definitions :

- consider a primary observable A *e.g.* average plaquette
  - Assume that the Markov chain has been equilibrated
  - $a_1, a_2, \ldots, a_N$ is a (MC) time series of measurements of A
  - True expectation value of A : $a = \langle a_i \rangle$
  - $\langle \ldots \rangle$ denotes the average over an infinite set of uncorrelated simulations (*i.e.* independent random numbers and initial states).
- The *sample mean* (for a particular simulation) is instead

$$\bar{a} = \frac{1}{N} \sum_{i=1}^{N} a_i$$

  - for the set of simulations: $\langle \bar{a} \rangle = a$
- The unbiased *sample variance* (for a particular simulation) is

$$\sigma^2 = \frac{1}{N-1} \sum_{i=1}^{N} (a_i - \bar{a})^2$$

  - it expresses how much a_i is liable to vary from its mean value
- bias: difference between the expectation value of an estimator and its true value

Gaussian distribution

- random variable x following a Gaussian distribution
- often the case when x represents an estimator for a parameter with a sufficiently large data sample (central limit theorem)

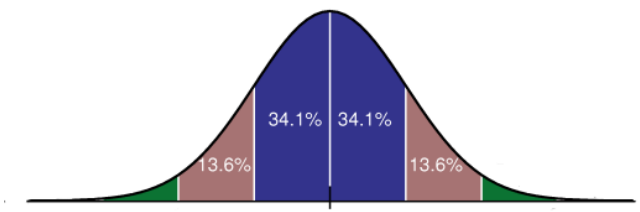
$$f(x,\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$

$$\mu = \int_{-\infty}^{+\infty} x\, f(x,\mu,\sigma)dx \qquad \sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2\, f(x,\mu,\sigma)dx$$

- probability γ that the measured value x will fall within $\pm\delta$ of the true value μ

$$\gamma = 1-\alpha = \frac{1}{\sqrt{2\pi}\sigma} \int_{\mu-\delta}^{\mu+\delta} e^{-(x-\mu)^2/2\sigma^2} dx$$

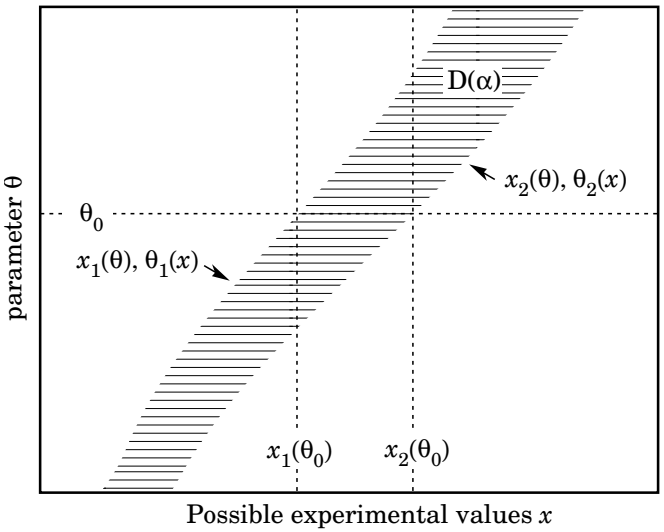
- This is also the probability for the interval $x \pm \delta$ to include μ
- The choice $\delta = \sigma$ gives an interval called the standard error which has $\gamma = 1-\alpha = 68.27\%$



α	δ	α	δ
0.3173	1σ	0.2	1.28σ
0.0455	2σ	0.1	1.64σ
0.0027	3σ	0.05	1.96σ

Confidence interval

- Confidence intervals often used for results where interpretation of uncertainties is non-trivial (i.e. non-Gaussian assumption)
- Definition of a confidence interval with probability γ=CL% (confidence level)
  - goal : locate a region which contains the *true value* of a parameter θ with a probability γ
  - x is a measurement and θ the unknown parameter for which we want to construct a confidence interval
  - for a given probability γ and for every value of θ one can find a set of values $x_1(\theta, \alpha)$ and $x_2(\theta, \alpha)$ such that :



$$P(x_1 < x < x_2; \theta) = \gamma = \int_{x_1}^{x_2} f(x; \theta) dx$$

Autocorrelations definitions

Recall : $a_1, a_2, \ldots, a_N$ is a time series of measurements of A such that $a = \langle \bar{a} \rangle = \langle a_i \rangle$

The *true* autocorrelation function is

$$\Gamma(t) = \Gamma(-t) = \langle (a_i - a)(a_{i+t} - a) \rangle$$

- Correlates deviation of i'th estimate for A with its deviation after $t \geq 0$ updates
- The variance of the measured value $\bar{a}$ of A is

$$\sigma^2 \equiv \left\langle (\bar{a} - a)^2 \right\rangle = \frac{1}{N^2} \sum_{i,j=1}^{N} \Gamma(i-j)$$

- The naive variance (*i.e.* asuming independent measurents) is : $\sigma_0^2 = \Gamma(0)/N$

  (*i.e.* $\tau_{\text{int}} = 1/2$)

- At large N, the *true* variance can be written by :

$$\sigma^2 = 2\tau_{\text{int}}\sigma_0^2 \qquad \text{where} \qquad \tau_{\text{int}} = \frac{1}{2} + \sum_{t=1}^{\infty} \frac{\Gamma(t)}{\Gamma(0)}$$

  τ_{int} is the *integrated* autocorrelation time.

- At large t, the autocorrelation function is

$$\Gamma(t) \propto \exp(-t/\tau_{\text{exp}}) \qquad \text{for } t \to \infty$$

  where τ_{exp} is the *exponential* autocorrelation time.

Autocorrelation time properties

- The integrated autocorrelation time τ_{int}
  - depends on the details of the algorithm, on the observable, on the parameters (quark masses, ...) critical slowing down
  - it is related to the number of update steps needed in order to have indenpendent measurments (*e.g.* in units of trajectory lengths)
  - encodes the efficiency of the algorithm for a determination of a given quantity
  - necessary to quote the "error on the error"

- Difficult measurement
  - "error on the error" ...
  - the calculation can in practice be ambiguous unless the time series is long
  - a systematic error is introduced by replacing the infinite sum by a finite summation window W

- Choice of the summation window W
  - large compared to the decay time τ_{exp} $\rightsquigarrow$ small systematic error
  - not too large to avoid contribution with negligible signal but large noise

Autocorrelation time estimates

- estimator for the true autocorrelation function

$$\bar{\Gamma}(t) = \frac{1}{N-t}\sum_{i=1}^{N-t}(a_i - \bar{a})(a_{i+t} - \bar{a})$$

- variance of the normalized autocorrelation function : $\bar{\rho}(t) = \bar{\Gamma}(t)/\bar{\Gamma}(0)$

$$\left\langle \delta\rho(t)^2 \right\rangle \simeq \frac{1}{N}\sum_{k=1}^{t+\Lambda}\left\{\bar{\rho}(k+t) + \bar{\rho}(k-t) - 2\bar{\rho}(k)\bar{\rho}(t)\right\}^2$$

the choice of Λ is not critical.

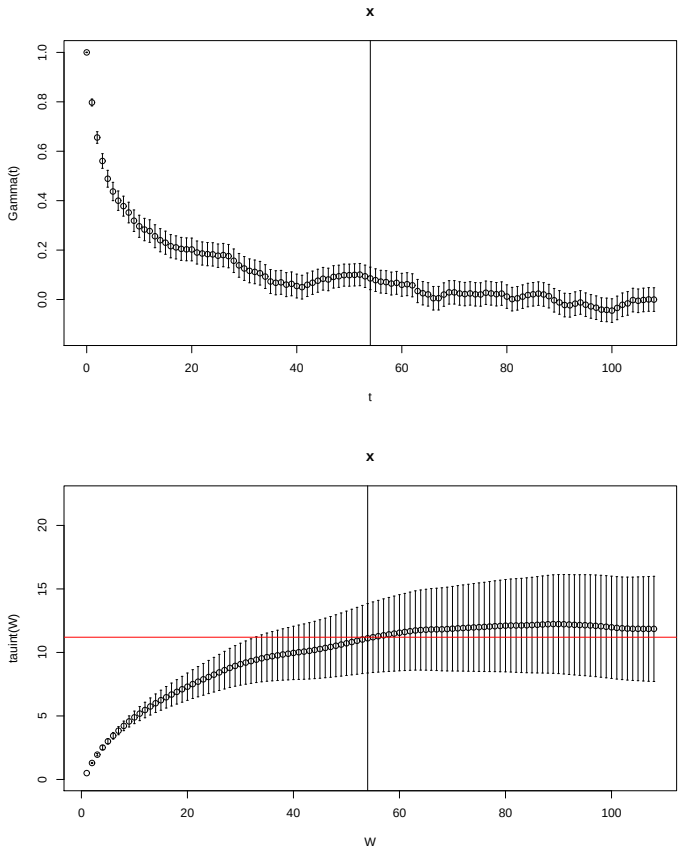
- estimate of the integrated autocorrelation time $\tau_{\rm int}$:

$$\tau_{\rm int} = \frac{1}{2} + \sum_{t=1}^{W}\bar{\rho}(t)$$

different prescriptions exist for the choice of W

- error on $\tau_{\rm int}$:

$$\left\langle \delta\tau_{\rm int}^2 \right\rangle \sim \frac{4W+2}{N}\tau_{\rm int}^2$$

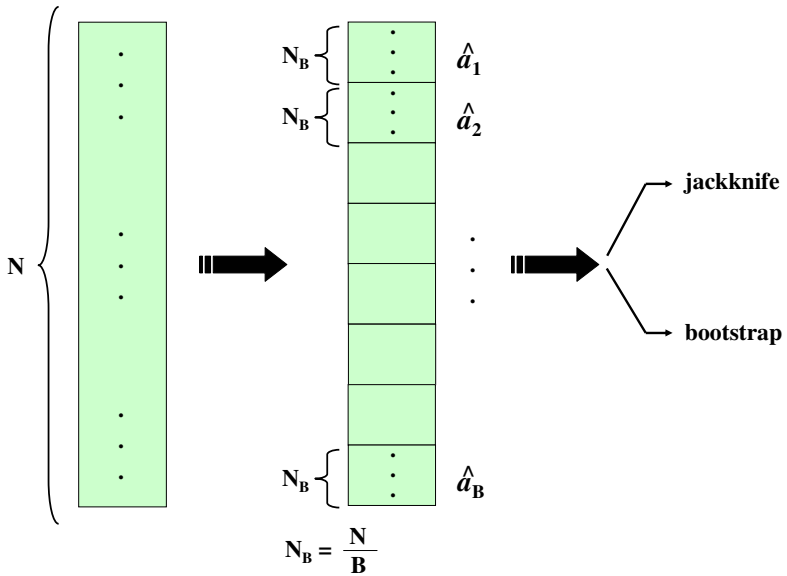


and generalisation to derived quantities

(*i.e.* non-linear functions of primary observables) ...

Binning

- $a_1, a_2, \ldots, a_N$ is a time series of measurements of A
- history of length $N = BN_B$ is divided into B blocks each of them containing N_B succesive measurements

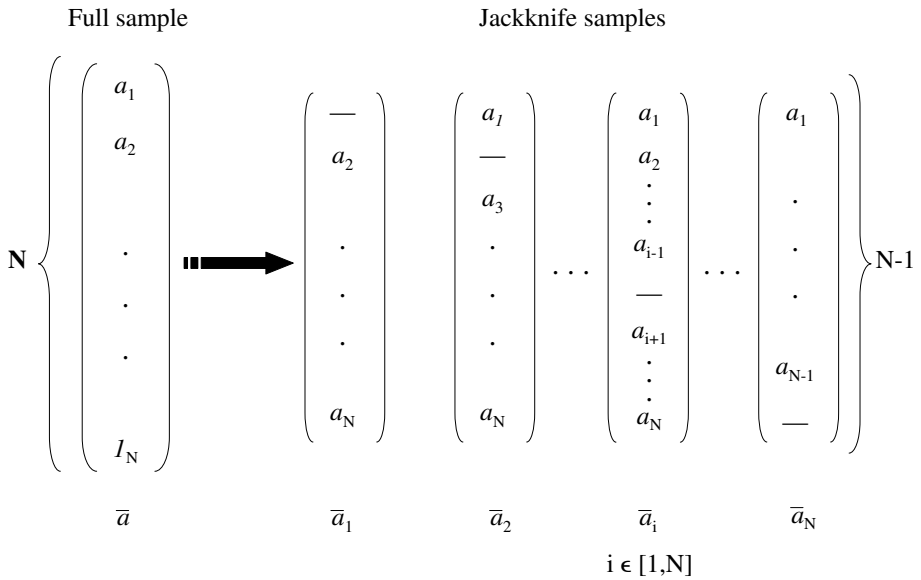


- Purpose of binning:
  - For large N_B, binned data becomes Gaussian $\rightsquigarrow$ Gaussian error analysis
  - For binned data autocorrelations are reduded. Can be neglected for large enough N_B

- Combine with resampling methods : jackknife and bootstrap

Resampling methods jackknife

- jackknife : remove one data point at a time from the sample and look at the variation of the resulting average



- jackknife samples are highly correlated $\rightsquigarrow$ the resulting variance is too small and must be corrected by multiplying by $(N-1)$
- if data is correlated: eliminate blocks of data to form each jackknife sample
  $\rightsquigarrow$ estimate of $\tau_{\rm int}$
- naive binning (instead of jackknife): averages evaluated over only N_B events
  $\rightsquigarrow$ stability of fits?

Resampling methods jackknife

- jackknife bins

$$a^{(k)} = \frac{1}{N - N_B} \left(\sum_{i=1}^{N} a_i - \sum_{j=1}^{N_B} a_{(k-1)N_B+j} \right) \qquad k = 1, \ldots, B$$

- the jackknife error is given by

$$\bar{\sigma}^2_{F\,\text{jack}} = \frac{B-1}{B} \sum_{k=1}^{B} (f(a^{(k)}) - \bar{F})^2$$

where $\bar{F} = f(\bar{a})$ is an estimator of the desired quantity $F = f(A)$.

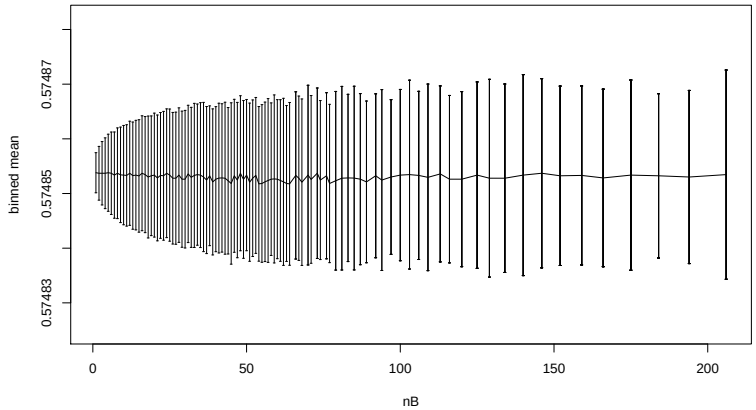
- unbiased estimator :

$$\bar{F} \longrightarrow B\,\bar{F} + (1 - B)\,\frac{1}{B} \sum_{k=1}^{B} f(a^{(k)})$$

- the error of an arbitrary quantity $f(\bar{a})$ can be computed
  - $\rightsquigarrow$ determine errors of fit parameters
- correlation between data can be take into account
  - $\rightsquigarrow$ simple with respect to error propagation calculations

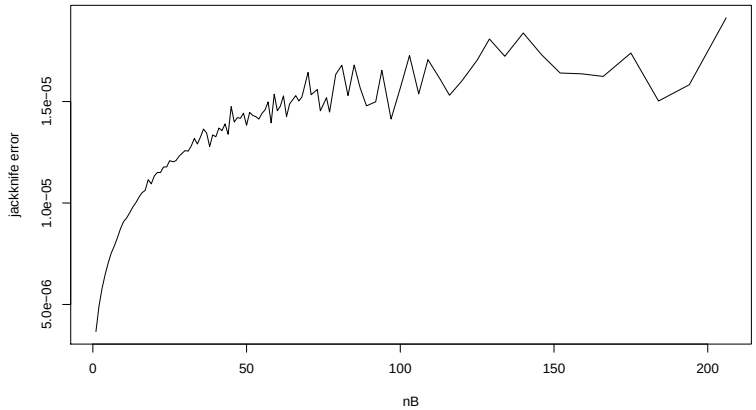
Resampling methods autocorrelations

- Dependence of jackknife error on the bin size N_B
  - $N_B = 1 \quad \rightsquigarrow$ naive error
  - For $N_B > \tau_{\rm exp}$ the autocorrelations are essentially reduced to those between nearest neighbor bins



- Estimate of $\tau_{\rm int}$ by using :

$$\sigma^2 = 2\tau_{\rm int}\sigma_0^2$$

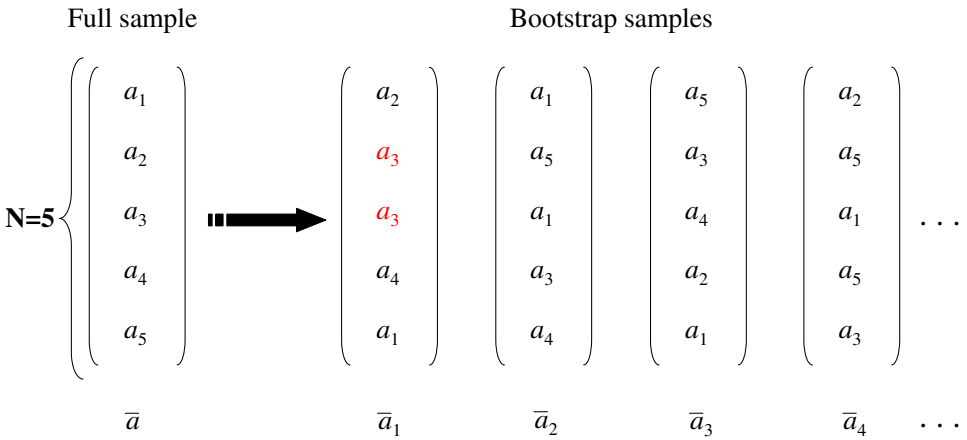


- bin size N_B and the size of the summation window $2W$ of Γ-method play a very similar rôle

Resampling methods bootstrap

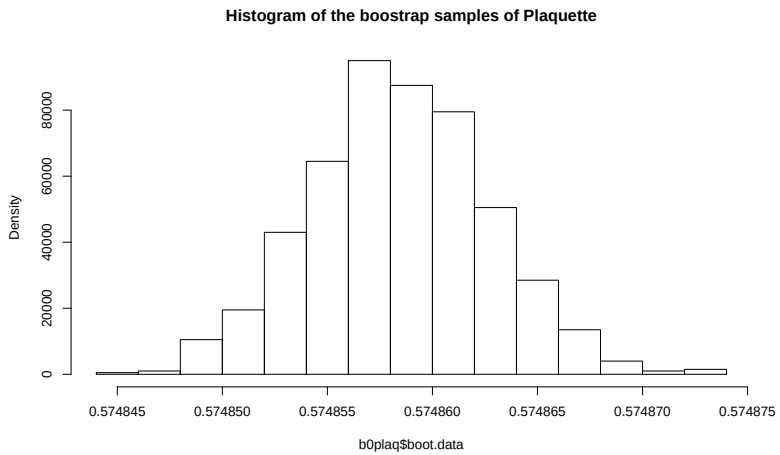
Bootstrap

- jackknife assumes a Gaussian distribution of the sample mean
- bootstrap gives the possibility tosample this distribution
- the chain of measurements is resampled N_b times
- each bootstrap sample contains measurements randomly selected (with replacement) from the full sample



Resampling methods bootstrap

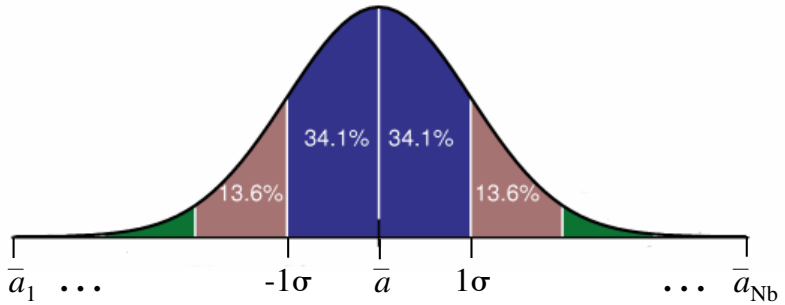
- bootstrap error
  - the desired quantity is computed on each sample
  - the set of N_b estimators sorted numerically
  - confidence interval can be quoted



Order $\overline{a}_i$ $\quad \overline{a}_1 \leq \overline{a}_2 \leq \ldots \leq \overline{a}_i \leq \ldots \leq \overline{a}_{N_b}$

$i \in [1, N_b]$

N_b= number of bootstrap samples



Fitting correlated data

- Consider a Green function $x(t)$ with $t = 1 \ldots D$ the Euclidean time
  - data set is $x^{(n)}(t)$ where $n = 1 \ldots N$ label succesive measurements
  - we assume $x^{(n)}(t)$ are statistically independent versus n (*i.e.* $\tau_{\rm int} = 1/2$)
  - but . . . strongly correlated in t (data from the same configuration)

- Goal : fit $x(t)$ to a function $F(t)$ which depends on P parameters a_p

  Determine :
  - the best values of the parameters a_p
  - the errors of a_P
  - the confidence level that the fit represents the data sample

Fitting correlated data

- Correlated χ^2 fit : finding best fit parameters corresponds to minimizing

$$\chi^2(a) = \sum_{t,t'} (F(t,a) - \bar{x}(t))\, C^{-1}(t,t')\, (F(t',a) - \bar{x}(t'))$$

  with respect to a_p for $p = 1 \ldots P$

- $\bar{x}(t)$ is the sample mean
- $C(t,t')$ is the covariance matrix

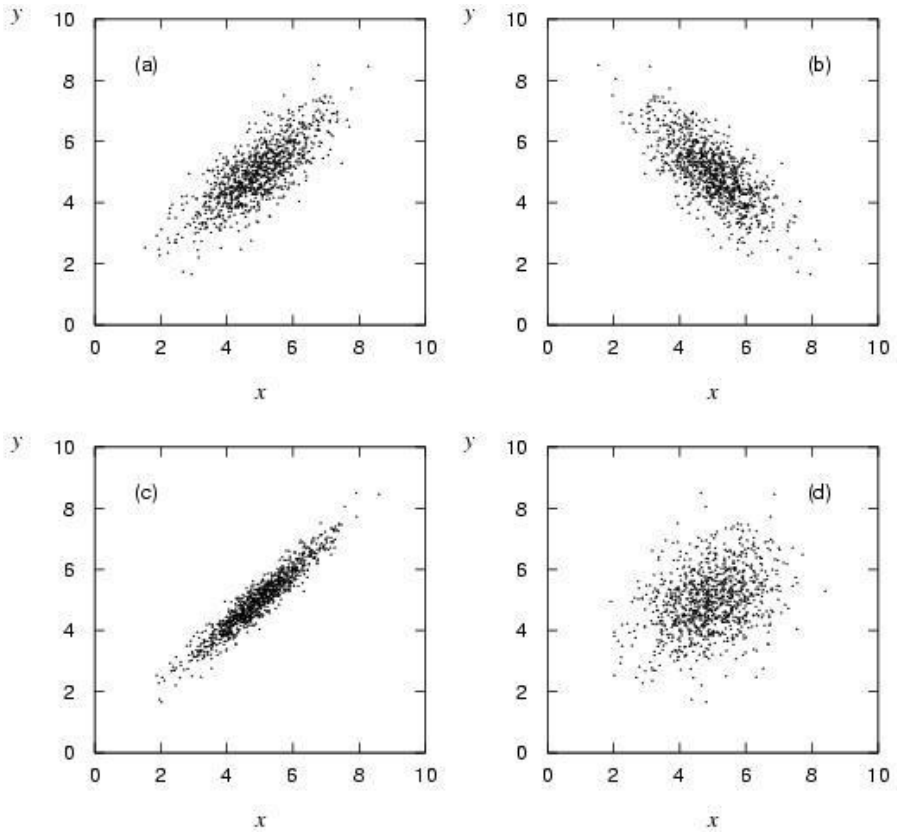
$$C(t,t') = \frac{1}{N-1} \sum_{n=1}^{N} (x^{(n)}(t) - \bar{x}(t))(x^{(n)}(t') - \bar{x}(t'))$$

- Properties of the covariance matrix
  - real symmetric positive-definite $D \times D$ matrix
  - rank is $N-1 \quad \rightsquigarrow \quad C$ will have $D-(N-1)$ zero eigenvalues if $N \leq D$
  - C develops small eigenvalues when $N \gtrsim D \quad \rightsquigarrow \quad$ impact on χ^2

- For sufficiently large N : expected value of χ^2 is the number of d.o.f. $D - P$
- The correlation is matrix is

$$\rho(t,t') = \frac{C(t,t')}{\sqrt{C(t,t)C(t',t')}} \in [0,1]$$

Correlation matrix



Side remark : correlation between systematic errors

$\rightsquigarrow$ add errors in quadrature only if they are uncorrelated

Correlated fits

- if $x(t)$ are uncorrelated

$$\chi^2(a) = \sum_t \frac{(F(t,a) - \overline{x}(t))^2}{\sigma(t)^2}$$

- if $F(t,a)$ is a linear function of a
  - minimum of χ^2 from linear algebra
  - this value quantifies the consistency between the measured values and the fitted form
  - covariance matrix of the obtained fit parameters a_P is given by :

$$\widetilde{C}^{-1}(p,p') = \frac{1}{2}\frac{\partial^2\chi^2}{\partial a_p \partial a_{p'}}$$

- If $F(t,a)$ is non-linear :
  - minimization through an iteration procedure
  - the minimun $\chi^2(a)$ is biased and is not guaranteed to follow a χ^2 distribution
    $\rightsquigarrow$ only when N is large

Correlated fits

- For small sample size $N \gtrsim D$: spurious small eigenvalues of the correlation matrix which increase χ^2
- Illustration : [C.Michael, hep-lat/9412087]
  - take N samples from a Gaussian distribution $\rightsquigarrow$ $x^{(n)}(t)$ with $n = 1 \ldots N$
  - the sample average is $\bar{x}(t)$ and true value $x(t) = 0$
  - compute χ^2 from the fit $\bar{x}(t) = F(t) = 0$
  - values of χ^2/D (averages from 10000 samples)

N	$D=1$	$D=3$	$D=5$	$D=7$	$D=10$	$D=15$
10	1.29	1.83	3.03	9.16	∞	∞
20	1.12	1.25	1.44	1.74	2.38	6.45
30	1.07	1.15	1.27	1.38	1.59	2.22
40	1.05	1.13	1.17	1.25	1.39	1.70
50	1.04	1.08	1.13	1.21	1.30	1.49
100	1.02	1.03	1.06	1.09	1.12	1.19

- Careful use of correlated χ^2 with N data samples of D data unless $N > \max(D^2, 10(D+1))$
  - if N is small : one can try to model the correlations
  - it is reasonable to use an uncorrelated χ^2 fit and estimate the errors on the parameters by bootstrap
  - it may be difficult to estimate the goodness of fit

Conclusions

- Autocorrelations in Markov chains
  - Autocorrelation function
  - Binning
  - Resampling : jackknife, bootstrap
- Correlations among observables
- References
  - B. Berg "Markov Chain Monte Carlo Simulations and their Statistical Analysis"
  - Numerical Recipes
  - M. C. K. Yang and David H. Robinson, Understanding and Learning Statistics by Computer, (World Scientific, Singapore, 1986)
  - Ulli Wolff, "Monte Carlo errors with less errors", hep-lat/0306017
- Tutorial
  - go to `http://www-zeuthen.desy.de/~herdoiza/tutorial`