random distributions

statistics and data analysis (chapter 3)

Björn Malte Schäfer

Graduate School for Fundamental Physics
Fakultät für Physik und Astronomie, Universität Heidelberg

April 26, 2016

outline: lecture 2 - random distributions

repetition

- distributions of random variables
- properties of distributions
- law of large numbers

numerical exercise

simulate a Gaussian random walk in 2 (d) Cartesian dimensions and show that the expected distance after n steps scales like $\sqrt{n}$. what's the dependence on d?

random distributions

- outcome of a random experiment: random variable x assumes values from the set $\Omega = \{x_1, \ldots, x_n\}$ of discrete results x_i
- set of probabilities $\{p1, \ldots, p_n\}$ with which these results may occur is the probability distribution characterising the random experiment
- If the set of results is continuous, $\Omega = [a, b]$, the result may fall within the interval $[x, x + \mathrm{d}x]$. Then, a function $p(x)$ such that the probability to find the result in this interval is $p(x)\mathrm{d}x$ is also called probability distribution (more exactly: probability density).
- visualisation: histrogram (discrete) or continuous function (probability density)
- **expectation value**

$$\langle x \rangle = \sum_i x_i p_i \quad \text{and} \quad \langle x \rangle = \int \mathrm{d}x\, x p(x)$$

characteristic function $\phi(t)$

- characterisation of a distribution $p(x)$: moments

$$\langle x^n \rangle = \int \mathrm{d}x\, x^n p(x) \tag{1}$$

- characteristic function $\phi(t)$: **Fourier-transform** of $p(x)\mathrm{d}x$:

$$\phi(t) = \int \mathrm{d}x p(x) \exp(-\mathrm{i}tx) \leftrightarrow p(x) = \int \frac{\mathrm{d}t}{2\pi} \phi(t) \exp(+\mathrm{i}tx) \tag{2}$$

- relation to moments: Taylor-expand the exponential:

$$\phi(t) = \int \mathrm{d}x p(x) \sum_n \frac{(-\mathrm{i}tx)^n}{n!} = \sum_n \langle x^n \rangle \frac{(-\mathrm{i}t)^n}{n!}, \quad \langle x^n \rangle = \int \mathrm{d}x x^n p(x) \tag{3}$$

- symmetric distributions have real characteristic functions
- characteristic functions can never be purely imaginary

measurement of moments: estimates

- series for $\phi(t)$ needs to converge absolutely (i.e. for every t-value) $\rightarrow$ all moments need to be finite, and sufficiently small
- then, $p(x)\mathrm{d}x$ can be inferred from the known moments
- but: in a real-world application, only a finite number of random numbers x_i are available, and a small number of **estimates** $\langle x_i^n \rangle$ can be sensibly determined,

$$\langle x_i^n \rangle \equiv \frac{1}{N}\sum_{i=1}^{N} x_i^n \tag{4}$$

  converging to $\langle x^n \rangle$ according to the law of large numbers
- in a physical experiment, $p(x)\mathrm{d}x$ can never be determined, and one has to make an **assumption** about it!

distinguish

always between the **estimates** $\langle x_i^n \rangle$ and the moments $\langle x^n \rangle$

moment generating function $m(t)$

- moment generating function $m(t)$: **Laplace-transform** of $p(x)$

$$m(t) = \int \mathrm{d}x p(x) \exp(tx) = \langle \exp(tx) \rangle = \sum_n \langle x^n \rangle \frac{t^n}{n!} \tag{5}$$

- normally, an integration is necessary for each moment. with $m(t)$ the problem of computing $\langle x^n \rangle$ is reduced to an n-fold differentiation

$$m(t=0) = \int \mathrm{d}x p(x) = \langle x^0 \rangle \tag{6}$$

$$\frac{\mathrm{d}}{\mathrm{d}t} m(t=0) = \int \mathrm{d}x p(x) x \exp(tx)|_{t=0} = \langle x \rangle \tag{7}$$

$$\frac{\mathrm{d}^n}{\mathrm{d}t^n} m(t=0) = \int \mathrm{d}x p(x) x^n \exp(tx)|_{t=0} = \langle x^n \rangle \tag{8}$$

- one can get the moments as well from the characteristic function by differentiation and setting $t = 0$, only one has to correct for the powers of i

cumulants

- what about Taylor-expanding the **logarithm** of the characteristic function?

$$\ln\phi(t) = \sum_n \frac{(\mathrm{i}t)^n}{n!}\kappa_n$$

- the coefficients κ_n are called **cumulants**
- with the cumulants, one can write every probability density $p(x)$ in the form $\exp(\sum_n \kappa_n t^n)$
- but are a bit more practical when dealing with deviations from Gaussianity:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- a Gaussian distribution has only two cumulants: mean and variance

$$\phi(t) = \exp(\mathrm{i}\mu t - \sigma^2 t^2/2)$$

- $\ln\phi(t)$ is a polynomial of order two for a Gaussian: $\kappa_1 = \mu$, $\kappa_2 = \sigma^2$

converting between moments and cumulants

- moments $\langle x^n\rangle$ and cumulants κ_n convey the same information, but converting between them is not straightforward!
- the relation is nonlinear, but at least one only needs to know all cumulants up to order n for the moments up to order n
- remember $\ln\phi(t) = K(t)$ with the cumulant generating function (K)

$$\phi(t) = \langle\exp(\mathrm{i}tx)\rangle = \sum_n \langle x^n\rangle \frac{(\mathrm{i}t)^n}{n!} \quad \text{and} \quad K(t) = \ln\langle\exp(\mathrm{i}tx)\rangle) = \sum_n \kappa_n \frac{(it)^n}{n!} \tag{9}$$

- obviously, $\ln\phi(t) = K(t)$ and $K(t) = \exp(\phi(t))$
- first cumulant:

$$\kappa_1 = \frac{\mathrm{d}}{\mathrm{d}t}K(t) = \frac{\mathrm{d}}{\mathrm{d}t}\exp(\phi(t)) = K(t)\frac{\mathrm{d}}{\mathrm{d}t}\phi(t) = \langle x\rangle \tag{10}$$

  if evaluated at $t = 0$. continue by induction!
- did you notice that partition sums in statistical mechanics are cumulant generating functions?

cumulative distribution and percentiles

- cumulative distribution function of a probability distribution $p(x)$ is defined by

$$P_j = \sum_i^j p_i \quad \text{and} \quad P(x) = \int_a^x \mathrm{d}x\, p(x)$$

  i.e. it gives the probability for the random variable to be $\leq x_j$ or $\leq x$

- percentiles q_ϵ are defined to contain a certain fraction of the possible results of a random experiment

$$P_j = \epsilon \quad \text{and} \quad P(x) = \epsilon$$

  if $\epsilon = 0.25$, percentiles are called quartiles

- **median** is the $\epsilon = 0.5$ percentile

$$P(x) = 0.5 \rightarrow P(x = q_{0.5}) = \int_a^{q_0.5} \mathrm{d}x\, p(x)$$

characterisation of random distributions

- probability density
- moments or cumulants
- cumulative distribution
- percentiles

question

what are the above defined quantities for a dice, for a Gaussian distribution, for a Planck distribution, for a Maxwell distribution?

Bernoulli-distribution

- single random experiment with two possible results, x_1 and x_2, which occur with the probabilities $p_1 = p$ and $p_2 = 1 - p$
- Bernoulli: probability of getting k **favourable** results in n trials

$$B(n, p, k) = \binom{n}{k} p^k (1 - p^{n-k}) \quad \text{with} \quad \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

question (typical Bernoulli question)

the fraction of female students in physics in Heidelberg is 0.2. how probable is it that the statistics course is attended by 18 female students when the total attendance is 40?

Bernoulli-distribution

- normalisation: use generalised binomial formula

$$\sum_k B(n,p,k) = \sum_k \binom{n}{k} p^k (1-p)^{n-k} = (p + (1-p))^n = 1$$

- $B(n,p,k)$ is symmetric **only** for $q = 0.5$:

$$B(n, 0.5, k) = \binom{n}{k} \frac{1}{2^n}$$

Bernoulli-distribution: mean and variance

- consider a mathematical trick:

$$g_k(p,q) = \binom{n}{k} p^k q^{n-k}$$

  with independent p, q
- then, one has a derivative relation

$$k g_k(p,q) = p\frac{\partial}{\partial p} g_k(p,q) = \frac{\partial}{\partial \ln p} g_k(p,q)$$

- substitute into the binomial formula

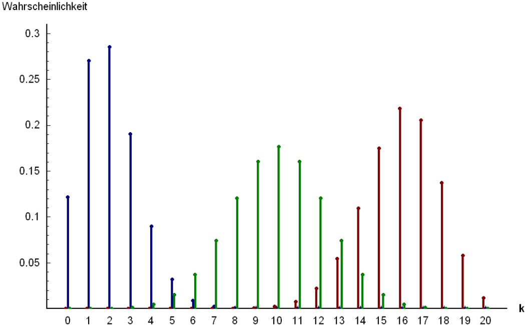
$$\langle k\rangle = \sum_k k g_k(p,q) = \frac{\partial}{\partial \ln p}\sum_k g_k(p,q) = \frac{\partial}{\partial \ln p}(p+q)^n = np(p+q)^n = np$$

- for the second moment, use

$$k(k-1)g_k(p,q) = p^2\frac{\partial^2}{\partial p^2} g_k(p,q)$$

- this implies for the variance $\sigma^2 = \langle k^2\rangle - \langle k\rangle^2 = np(1-p)$

Bernoulli-distribution: visualisation



source: wikipedia

- Bernoulli-distribution for $p = 0.1, 0.5, 0.8$ and $n = 20$

Bernoulli-distribution: random walk

- example for the binomial distribution is the basic random walk
- take n equally sized steps along a coordinate axis
- probability p for stepping right, and $1 - p$ for stepping left
- if k steps are taken to the right, the distance after n steps is $x = k - (n - k) = 2k - n$
- consequently, the average distance is

$$\langle x \rangle = 2\langle k \rangle - n = 2np - n = n(2p - 1)$$

- variance of the random walk:

$$\sigma^2 = \langle x^2 \rangle - \langle x \rangle^2 = 4np(1 - p)$$

fluid mechanics

what's the relation between the $\sqrt{n}$-law just derived with diffusive processes? children know that (but they don't know that they do)!

Poisson-distribution

- sequence of **rare** events, characterised by 3 properties
  1. within a time interval $\mathrm{d}t$, there occurs either no or one event
  2. the probability for an event to occur within $\mathrm{d}t$ is $g\mathrm{d}t$, with constant g
  3. independent of the preceding events

  what is the probability $p_n(t)$ for the n events to have occurred within the time t?

- generally, let $\lambda = gt$ characterise the probability distribution of n rare and independent events, then the Poisson distribution

$$p_\lambda(n) = \frac{\lambda^n}{n!}\exp(-\lambda)$$

  describes this random experiment

question

show that the Poisson distribution is normalised

Poisson-distribution

- expectation value

$$\langle n\rangle = \sum_{n=0} \frac{n\lambda^n}{n!}\exp(-\lambda) = \exp(-\lambda)\sum_{n=1}\frac{\lambda\lambda^{n-1}}{(n-1)!} = \lambda\exp(-\lambda)\sum_{n=0}\frac{\lambda^n}{n!} = \lambda$$

- variance

$$\langle n^2\rangle = \sum_{n=0} n^2 p_\lambda(n) = \ldots = \lambda^2 - \lambda$$

such that $\sigma^2 = \langle n^2\rangle - \langle n\rangle^2 = \lambda$

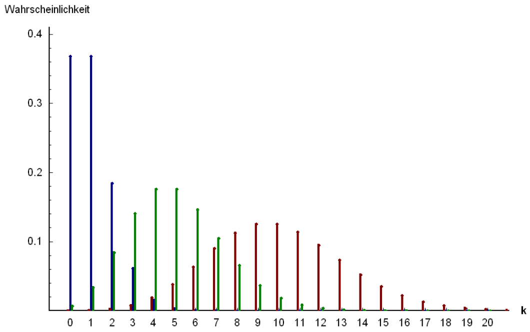
Poisson distribution

has **equal** mean and variance!

question

show that $\langle n^2\rangle = \lambda^2 - \lambda$ for a Poisson-distribution

Poisson-distribution: visualisation



source: wikipedia

- Poisson distribution for $\lambda = 1, 5, 10$

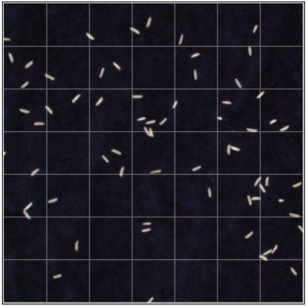
Poisson errors in counting experiments

- imagine drawing n numbers x from an arbitrary distribution $p(x)$. you can expect a total of $k = np_i$ of the draws to fall inside a histrogram bin with the integrated probability

$$p_i = \int_{x_i}^{x_{i+1}} \mathrm{d}x\, p(x) \tag{11}$$

- Bernoulli-statistics: if there are k draws of values inside the interval at the probability p_i, $n - k$ draws need to fall outside the interval with the probability $1 - p_i$ and the order does not matter
- consequently, the number of random numbers in an interval has mean np and variance $np(1 - p)$
- fine bins, small probabilities: approximate with **Poisson-statistics**: mean and variance are both $\lambda = np_i$, the standard deviation $\sqrt{\lambda}$

Poisson-distribution: example



source wikipedia

- sprinkle $n = 66$ rice grains on a 7×7-grid, $\lambda = 66/49 \simeq 1.35$

k	0	1	2	3	4	5
counts	15	15	11	5	1	2
$49 \times p_\lambda(k)$	12.7	17.2	11.6	5.2	1.7	0.5

Gauss-distribution

- consider now a symmetric binomial distribution $B(n, 1/2, k)$ in the limit of very many repetitions n of the random experiment
- in comparison to the width n of the interval from which k can be drawn, the relative standard deviation $\sigma/n = 1/(2\sqrt{n})$ becomes very small and thus the probability distribution must form a sharp peak around its expectation value at $k = n/2$
- expand the logarithm of $B(n, 1/2, k)$ around its peak at $k = n/2$ into a Taylor series and keep the terms up to second order:

$$\ln B(n, 1/2, k) = \ln B(n, 1/2, n/2) + \frac{1}{2}\frac{\partial^2}{\partial k^2}B(n, 1/2, k = n/2)\left(k - \frac{n}{2}\right)^2$$

Gauss-distribution

- first derivative

$$\frac{\partial}{\partial k}\ln B(n,1/2,k) = -\frac{\partial \ln k!}{k} + \frac{\partial \ln(n-k)!}{k} + \ln p - \ln(1-p)$$

- logarithm: $\partial \ln n!/n = \ln n$

- first derivative

$$\frac{\partial}{\partial k}\ln B(n,1/2,k) = -\ln k + \ln(n-k) + \ln p - \ln(1-p)$$

- second derivative, evaluated at $k = n/2$

$$\frac{\partial^2}{\partial k^2}\ln B(n,1/2,k) = -\frac{4}{n} = -\frac{1}{\sigma^2}$$

- functional form of the Gaussian probability density

$$B(n,1/2,k) \propto \exp\left(-\frac{(k-n/2)^2}{2\sigma^2}\right)$$

Gauss-distribution

- limit $n \to \infty$: $k \to x \in \mathbb{R}$

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-x_0)^2}{2\sigma^2}\right)$$

- sometimes, the Gaussian distribution is called **central distribution** or **normal** distribution

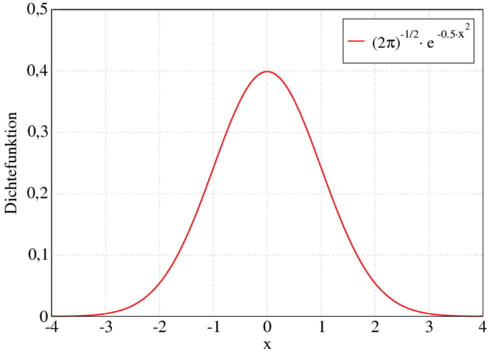
question

show that $\int \mathrm{d}x\, p(x)x^2 = \sigma^2$ for a normalised Gaussian

question

show that the $\mathrm{d}^2p/\mathrm{d}x^2 = 0$ is solved by $x = \pm\sigma$

Gauss-distribution: visualisation



source: wikipedia

- Gauss-distribution for $\mu = 0$ and $\sigma = 1$

relations between Bernoulli-, Poisson- and Gauss-distributions

- in the limit of large n and small p while keeping $np \equiv \lambda$ constant, the Binomial distribution approximates the Poisson distribution:

$$\lim_{n\to\infty} \binom{n}{k} p^k (1-p)^{n-k} = \lim_{n\to\infty} \binom{n}{k} \frac{\lambda^k}{n^k} \left(1-\frac{\lambda}{n}\right)^n = \frac{\lambda^n}{n!}\exp(-\lambda)$$

- if λ is large, the Poisson distribution turns into the Gaussian distribution with mean (and variance) λ
- Stirling's formula

$$\ln n! \simeq n\ln n - n + \frac{1}{2}\ln(2\pi n) \rightarrow \ln p_\lambda(k) \simeq k\ln\lambda - k\ln k + k - \frac{1}{2}\ln(2\pi k) - \lambda$$

- since the mean and the standard deviation of k are λ and $\sqrt{\lambda}$, respectively, $p_\lambda(k)$ forms a sharp peak around $k = \lambda$ for large λ:

$$\ln p_\lambda(k) \simeq -\frac{1}{2}\ln(2\pi k)$$

relations between Bernoulli-, Poisson- and Gauss-distributions

- furthermore at $k = \lambda$,

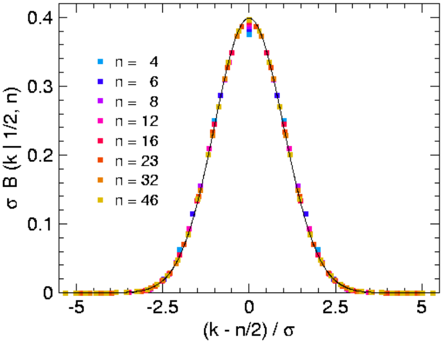
$$\frac{\partial}{\partial k}\ln p_\lambda(k) = -\frac{1}{2\lambda} \to 0 \quad \text{and} \quad \frac{\partial^2}{\partial k^2}\ln p_\lambda(k) = -\frac{1}{k}$$

- carry out a Taylor expansion to second order at $k = \lambda$

$$\ln p_\lambda(k) \simeq -\frac{1}{2}\ln(2\pi k) - \frac{1}{2k}(k-\lambda)^2$$

which is exactly the Taylor expansion of a Gaussian with mean and variance λ

convergence of the Bernoulli-distribution towards the Gauss-distribution



source: wikipedia

- Gauss-distribution is a very good approximation in the limit of large n

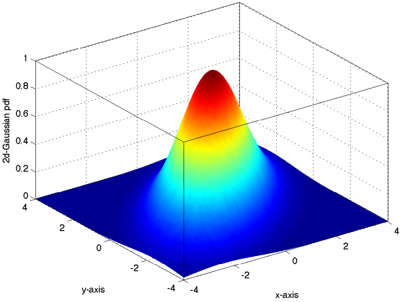
multivariate Gaussian probability density

- assign two values x, y to the outcome of a random experiment $\rightarrow$ multivariate distribution $p(x, y)\mathrm{d}x\mathrm{d}y$
- special importance: **multivariate Gaussian**

$$p(\vec{x})\mathrm{d}\vec{x} = \frac{1}{(2\pi)^{n/2}\sqrt{\det(Q)}}\exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^t Q^{-1}(\vec{x}-\vec{\mu})\right) \tag{12}$$

  with mean μ and covariance matrix Q
- $\phi(\vec{t})$ and $m(\vec{t})$ generalise to the multivariate case
- covariance matrix $Q \equiv \langle x_i x_j \rangle$ is symmetric and positive definite (because of the Cauchy-Schwarz inequality) $\rightarrow$ diagonalisable
- in the diagonal frame, $\Delta = O^t Q O \equiv \mathrm{dia}(\sigma_1^2, \ldots, \sigma_N^2)$

covariance



2d Gaussian probability density

- covariance Q determines the size, axis ratio and orientation

properties of the covariance

- covariance Q is positive definite for two reasons
  1. positive determinant $\det(Q)$ gives proper positive real normalisation
  2. $\vec{x}^t Q^{-1} \vec{x}$ is a positive definite quantity $\rightarrow$ finite normalisation
- positiveness is a consequence of the Cauchy-Schwarz inequality
- correlation coefficient

$$r_{\mu\nu} = \frac{Q_{\mu\nu}}{\sqrt{Q_{\mu\mu} Q_{\nu\nu}}}$$

  always in the range $-1 \leq r_{\mu\nu} \leq +1$
- covariance can be estimated in the same way as the variance

question

give two reasons why the covariance matrix of the multivariate Gaussian is positive definite.

sampling from a multivariate Gaussian

- is it possible to sample **sets of random numbers** from a multivariate Gaussian?
- covariance matrix is $C_{ij} = \langle y_i y_j \rangle$, in vectors $C = \langle \vec{y}\vec{y}^t \rangle$
- apply a similarity transformation $\vec{y} \to A\vec{x}$ with $\langle \vec{x}\vec{x}^t \rangle = \mathrm{id}$

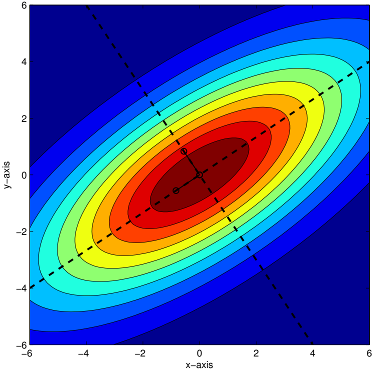
$$C = \langle \vec{y}\vec{y}^t \rangle = \langle A\vec{x}A^t\vec{x}^t \rangle = AA^t \underbrace{\langle \vec{x}\vec{x}^t \rangle}_{=\mathrm{id}} \tag{13}$$

- **Cholesky-decomposition** A of the matrix $C = AA^t$
- sample for an uncorrelated multivariate Gaussian a vector $\vec{x}$ (which has unit covariance), linear transformation $\vec{y} = A\vec{x}$ **correlates** the entries in $\vec{x}$

question

carry out the Cholesky transform of $C = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$ and sample from it in a small python script.

covariance



cut through a 2d Gaussian probability density

- covariance is positive definite: can be diagonalised
- diagonal frame: individual random numbers are uncorrelated

conditions on a multivariate Gaussian

- if you want to determine the width of a multivariate Gaussian: 3 answers
  1. width along the coordinate axes
  2. effective width projected onto the coordinate axes
  3. width in the principal axis frame
- marginalisation: projection of the Gaussian onto a coordinate axis
- conditionalisation: cut through a Gaussian along a coordinate axis

conditionalisation: simple case

- impose a condition $y = 0$ on a bivariate Gaussian $p(x, y)\mathrm{d}x\mathrm{d}y$:

$$p_c(x) = p(x|y = 0) \tag{14}$$

- subsitute into the Gaussian:

$$p_c(x) = \frac{1}{\sqrt{(2\pi)^2 \det(Q)}} \exp\left(-\frac{1}{2}\begin{pmatrix} x \\ 0 \end{pmatrix}^t \begin{pmatrix} \langle x^2\rangle & \langle xy\rangle \\ \langle xy\rangle & \langle y^2\rangle \end{pmatrix}^{-1} \begin{pmatrix} x \\ 0 \end{pmatrix}\right) \tag{15}$$

- which yields:

$$p_c(x) = \frac{1}{\sqrt{(2\pi)^2 \det(Q)}} \exp\left(-\frac{1}{2}\frac{x^2}{\langle x^2\rangle(1-r^2)}\right) \tag{16}$$

- smaller variance: $\sigma^2 = (1 - r^2)\langle x^2\rangle \leq \langle x^2\rangle$

question

verify the last equation! why does the sign of r not matter? what about the normalisation?

conditionalisation: general case

- the same idea works as well if the condition is not equal to the mean
- split up mean μ and covariance matrix Q

$$\vec{\mu} = \begin{pmatrix} \vec{\mu}_1 \\ \vec{\mu}_2 \end{pmatrix} \quad \text{and} \quad Q = \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix} \tag{17}$$

  where the index 1 refers to the free part of the pdf, and the index 2 to the constraint variables

- then one obtains the **Schur-complement**

$$\bar{\mu} = \vec{\mu}_1 + Q_{12}Q_{22}^{-1}(\vec{a} - \vec{\mu}_2) \quad \text{and} \quad \bar{Q} = Q_{11} - Q_{12}Q_{22}^{-1}Q_{21} \tag{18}$$

  i.e. a Gaussian $p(\vec{x}_1|\vec{x}_2 = \vec{a})$ with mean $\bar{\mu}$ and covariance $\bar{Q}$

- depending on the condition, the variance and mean changes!

marginalisation

- integrate over all y-values for a given x

$$p_m(x) = \int \mathrm{d}y\, p(x,y) \tag{19}$$

- the integration can be carried out by completing the square

$$p_m(x) = \frac{1}{\sqrt{(2\pi)^2 \det(Q)}} \exp(-\chi^2/2) \tag{20}$$

with $\chi^2 = (\langle x^2\rangle y^2 - 2\langle xy\rangle xy + \langle y^2\rangle x^2)/\det(Q)$

- results in

$$p_m(x) = \sqrt{\frac{\det(Q)}{2\pi\langle x^2\rangle}} \exp\left(-\frac{1}{2}\frac{\langle y^2\rangle}{\langle x^2\rangle}(1-r^2)x^2\right) \tag{21}$$

- new variance: $\sigma^2 = \langle x^2\rangle/\langle y^2\rangle/(1-r^2)$

principal value decomposition

- assume you've got n samples $\{x_i\}$ from a multivariate Gaussian
- estimate the mean

$$\bar{x}_i = \frac{1}{n}\sum_m (x_i)_m \tag{22}$$

- and the covariance

$$Q_{ij} = \frac{1}{n}\sum_m (x_i - \bar{x}_i)_m (x_j - \bar{x}_j)_m \tag{23}$$

  by summing over the samples indexed by m
- one can now write down a multivariate Gaussian with these two estimates
- the eigenvalues are the variances in the principal axis system and the eigenvectors determine the transformation into that frame

summary

- distributions can be quantified using
  - moments
  - histograms, and cumulative histograms
  - cumulants
  - percentiles
- 3 most relevant distributions
  1. Bernoulli-distribution (binomial distribution)
  2. Poisson-distribution
  3. Gauss-distribution (normal, central distribution)
- multivariate Gaussians: covariance matrix
- conditions: simple in the case of univariate distributions, multivariate distributions require Schur-complement