Machine Learning and Statistics in Genetics and Genomics

II: Linear regression

Christoph Lippert

Microsoft Research
eScience group

Los Angeles , USA

Microsoft Research

Current topics in computational biology
UCLA
Winter quarter 2014

Linear Regression

- Introduction
- Maximum likelihood estimation
- Genome-wide association studies

Gaussian distribution

- Central limit theorem
- Linear transformations
- Marginal distributions
- Common Gaussian calculations

Outline

Outline

Regression

Noise model and likelihood

- Given a dataset $\mathcal{D} = \{\boldsymbol{x}_n, y_n\}_{n=1}^N$, where $\boldsymbol{x}_n = \{x_{n,1}, \ldots, x_{n,D}\}$ is D dimensional, fit parameters $\boldsymbol{\theta}$ of a regressor f with added Gaussian noise:

$$y_n = f(\boldsymbol{x}_n; \boldsymbol{\theta}) + \epsilon_n \quad \text{where} \quad p(\epsilon \,|\, \sigma^2) = \mathcal{N}\left(\epsilon \mid 0, \sigma^2\right).$$

- Equivalent likelihood formulation:

$$p(\boldsymbol{y} \,|\, \boldsymbol{X}) = \prod_{n=1}^{N} \mathcal{N}\left(y_n \mid f(\boldsymbol{x}_n; \boldsymbol{\theta}), \sigma^2\right)$$

Regression

Choosing a regressor

- Choose f to be linear:

$$p(\boldsymbol{y} \mid \boldsymbol{X}) = \prod_{n=1}^{N} \mathcal{N}\left(y_n \mid \boldsymbol{x}_n \cdot \boldsymbol{\beta} + c, \sigma^2 \right)$$

- Consider bias free case, $c = 0$, otherwise include an additional column of ones in each $\boldsymbol{x}_n$.

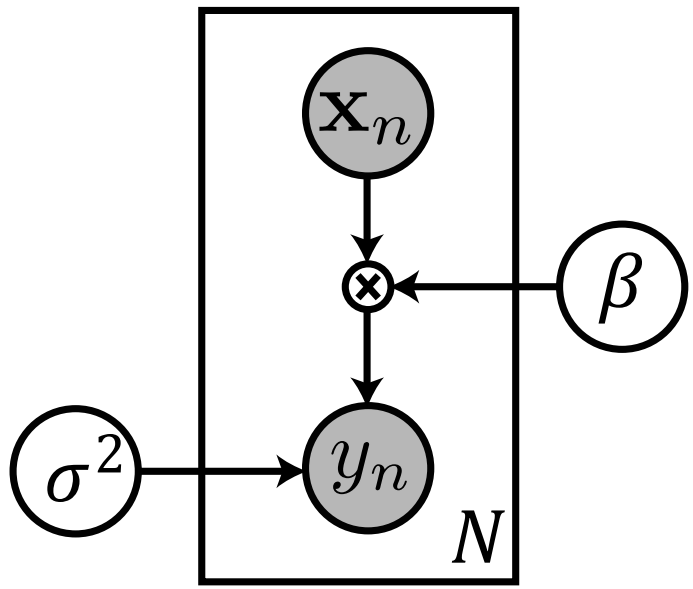
Regression

Choosing a regressor

- Choose f to be linear:

$$p(\boldsymbol{y} \mid \boldsymbol{X}) = \prod_{n=1}^{N} \mathcal{N}\left(y_n \mid \boldsymbol{x}_n \cdot \boldsymbol{\beta} + c, \sigma^2\right)$$

- Consider bias free case, $c = 0$, otherwise include an additional column of ones in each $\boldsymbol{x}_n$.



Equivalent graphical model

Linear Regression

Maximum likelihood

- Taking the logarithm, we obtain

$$\ln p(\boldsymbol{y} \,|\, \boldsymbol{\theta}\sigma^2) = \sum_{n=1}^{N} \ln \mathcal{N}\left(y_n \mid \boldsymbol{x}_n \cdot \boldsymbol{\beta}, \sigma^2 \right)$$

$$= -\frac{N}{2} \ln 2\pi\sigma^2 - \frac{1}{2\sigma^2} \underbrace{\sum_{n=1}^{N} (y_n - \boldsymbol{x}_n \cdot \boldsymbol{\beta})^2}_{\text{Sum of squares}}$$

- The likelihood is maximized when the squared error is minimized.
- Least squares and maximum likelihood are equivalent.

Linear Regression

Maximum likelihood

- Taking the logarithm, we obtain

$$\ln p(\boldsymbol{y} \,|\, \boldsymbol{\theta}\sigma^2) = \sum_{n=1}^{N} \ln \mathcal{N}\left(y_n \mid \boldsymbol{x}_n \cdot \boldsymbol{\beta}, \sigma^2 \right)$$

$$= -\frac{N}{2} \ln 2\pi\sigma^2 - \frac{1}{2\sigma^2} \underbrace{\sum_{n=1}^{N} (y_n - \boldsymbol{x}_n \cdot \boldsymbol{\beta})^2}_{\text{Sum of squares}}$$

- The likelihood is maximized when the squared error is minimized.
- Least squares and maximum likelihood are equivalent.

Linear Regression

Maximum likelihood

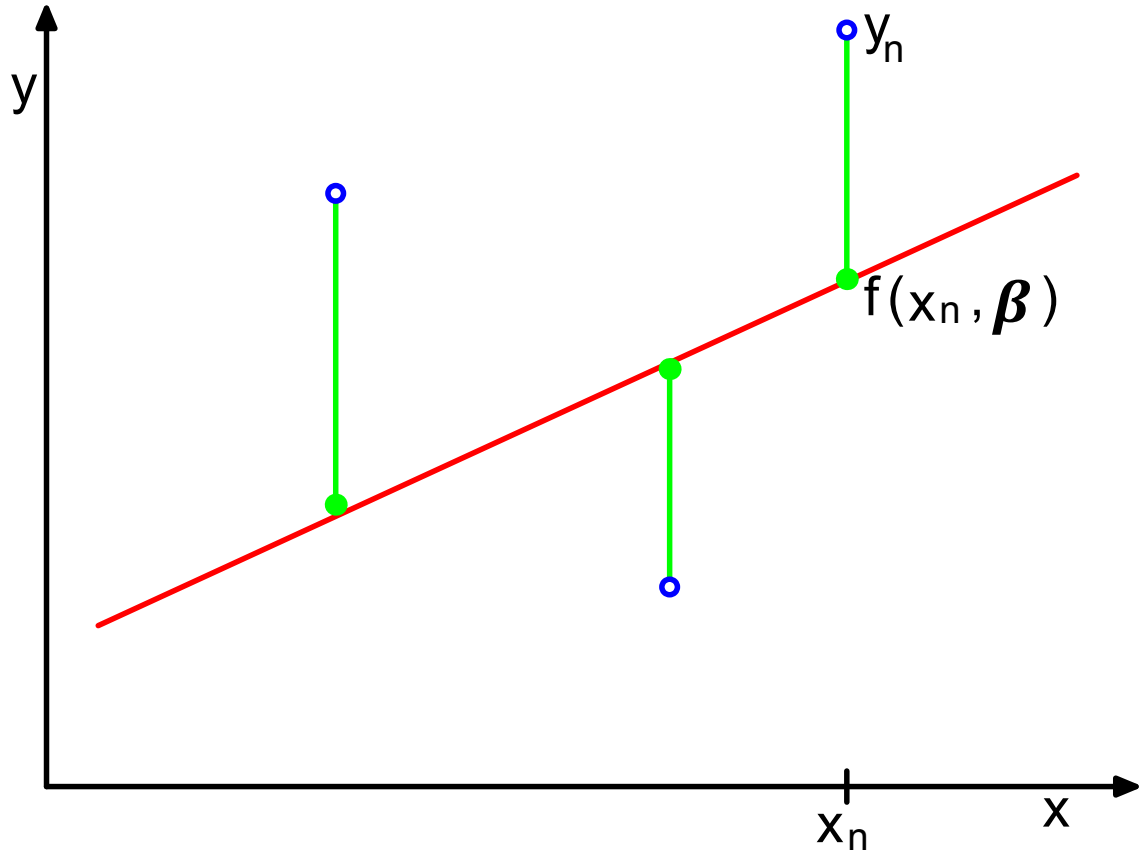
- Taking the logarithm, we obtain

$$\ln p(\boldsymbol{y} \,|\, \boldsymbol{\theta}\sigma^2) = \sum_{n=1}^{N} \ln \mathcal{N}\left(y_n \mid \boldsymbol{x}_n \cdot \boldsymbol{\beta}, \sigma^2 \right)$$

$$= -\frac{N}{2} \ln 2\pi\sigma^2 - \frac{1}{2\sigma^2} \underbrace{\sum_{n=1}^{N} (y_n - \boldsymbol{x}_n \cdot \boldsymbol{\beta})^2}_{\text{Sum of squares}}$$

- The likelihood is maximized when the squared error is minimized.
- Least squares and maximum likelihood are equivalent.

Linear Regression and Least Squares



(C.M. Bishop, Pattern Recognition and Machine Learning)

$$\text{argmin}_{\boldsymbol{\beta}} \frac{1}{2} \sum_{n=1}^{N} (y_n - \boldsymbol{x}_n \cdot \boldsymbol{\beta})^2$$

Linear Regression and Least Squares

- Derivative w.r.t a single weight entry β_i

$$\frac{d}{\mathrm{d}\beta_d} \ln p(\boldsymbol{y} \,|\, \boldsymbol{\beta}, \sigma^2) = \frac{d}{\mathrm{d}\beta_d} \left[-\frac{1}{2\sigma^2} \sum_{n=1}^{N} (y_n - \boldsymbol{x}_n \cdot \boldsymbol{\beta})^2 \right]$$

- Set gradient w.r.t. $\boldsymbol{\beta}$ to zero [vector holding the derivatives $\forall \beta_d$]

$$\nabla_{\boldsymbol{\beta}} \ln p(\boldsymbol{y} \,|\, \boldsymbol{\beta}, \sigma^2) =$$

- Here, the matrix $\boldsymbol{X}$ is defined as $\boldsymbol{X} = \begin{bmatrix} x_{11} & \dots & x1D \\ \dots & \dots & \dots \\ x_{N1} & \dots & x_{ND} \end{bmatrix}$

Linear Regression and Least Squares

- Derivative w.r.t a single weight entry β_i

$$\frac{d}{\mathrm{d}\beta_d} \ln p(\boldsymbol{y} \,|\, \boldsymbol{\beta}, \sigma^2) = \frac{d}{\mathrm{d}\beta_d} \left[-\frac{1}{2\sigma^2} \sum_{n=1}^{N} (y_n - \boldsymbol{x}_n \cdot \boldsymbol{\beta})^2 \right]$$

$$= \left[-\frac{1}{\sigma^2} \sum_{n=1}^{N} \boldsymbol{x}_{nd} (y_n - \boldsymbol{x}_n \cdot \boldsymbol{\beta}) \right]$$

- Set gradient w.r.t. $\boldsymbol{\beta}$ to zero [vector holding the derivatives $\forall \beta_d$]

$$\nabla_{\boldsymbol{\beta}} \ln p(\boldsymbol{y} \,|\, \boldsymbol{\beta}, \sigma^2) =$$

- Here, the matrix $\boldsymbol{X}$ is defined as $\boldsymbol{X} = \begin{bmatrix} x_{11} & \dots & x1D \\ \dots & \dots & \dots \\ x_{N1} & \dots & x_{ND} \end{bmatrix}$

Linear Regression and Least Squares

- Derivative w.r.t a single weight entry β_i

$$\frac{d}{\mathrm{d}\beta_d} \ln p(\boldsymbol{y} \,|\, \boldsymbol{\beta}, \sigma^2) = \frac{d}{\mathrm{d}\beta_d} \left[-\frac{1}{2\sigma^2} \sum_{n=1}^{N} (y_n - \boldsymbol{x}_n \cdot \boldsymbol{\beta})^2 \right]$$

$$= \left[-\frac{1}{\sigma^2} \sum_{n=1}^{N} \boldsymbol{x}_{nd} (y_n - \boldsymbol{x}_n \cdot \boldsymbol{\beta}) \right]$$

- Set gradient w.r.t. $\boldsymbol{\beta}$ to zero [vector holding the derivatives $\forall \beta_d$]

$$\nabla_{\boldsymbol{\beta}} \ln p(\boldsymbol{y} \,|\, \boldsymbol{\beta}, \sigma^2) =$$

- Here, the matrix $\boldsymbol{X}$ is defined as $\boldsymbol{X} = \begin{bmatrix} x_{11} & \dots & x1D \\ \dots & \dots & \dots \\ x_{N1} & \dots & x_{ND} \end{bmatrix}$

Linear Regression and Least Squares

- Derivative w.r.t a single weight entry β_i

$$\frac{d}{\mathrm{d}\beta_d} \ln p(\boldsymbol{y} \,|\, \boldsymbol{\beta}, \sigma^2) = \frac{d}{\mathrm{d}\beta_d} \left[-\frac{1}{2\sigma^2} \sum_{n=1}^{N} (y_n - \boldsymbol{x}_n \cdot \boldsymbol{\beta})^2 \right]$$

$$= \left[-\frac{1}{\sigma^2} \sum_{n=1}^{N} \boldsymbol{x}_{nd} (y_n - \boldsymbol{x}_n \cdot \boldsymbol{\beta}) \right]$$

- Set gradient w.r.t. $\boldsymbol{\beta}$ to zero [vector holding the derivatives $\forall \beta_d$]

$$\nabla_{\boldsymbol{\beta}} \ln p(\boldsymbol{y} \,|\, \boldsymbol{\beta}, \sigma^2) = \frac{1}{\sigma^2} \sum_{n=1}^{N} \boldsymbol{x}_n^\top (y_n - \boldsymbol{x}_n \cdot \boldsymbol{\beta}) = \mathbf{0} \qquad \text{(where } \mathbf{0} \text{ is a vector of 0s)}$$

- Here, the matrix $\boldsymbol{X}$ is defined as $\boldsymbol{X} = \begin{bmatrix} x_{11} & \dots & x1D \\ \dots & \dots & \dots \\ x_{N1} & \dots & x_{ND} \end{bmatrix}$

Linear Regression and Least Squares

- Derivative w.r.t a single weight entry β_i

$$\frac{d}{\mathrm{d}\beta_d} \ln p(\boldsymbol{y} \,|\, \boldsymbol{\beta}, \sigma^2) = \frac{d}{\mathrm{d}\beta_d} \left[-\frac{1}{2\sigma^2} \sum_{n=1}^{N} (y_n - \boldsymbol{x}_n \cdot \boldsymbol{\beta})^2 \right]$$

$$= \left[-\frac{1}{\sigma^2} \sum_{n=1}^{N} \boldsymbol{x}_{nd} (y_n - \boldsymbol{x}_n \cdot \boldsymbol{\beta}) \right]$$

- Set gradient w.r.t. $\boldsymbol{\beta}$ to zero [vector holding the derivatives $\forall \beta_d$]

$$\nabla_{\boldsymbol{\beta}} \ln p(\boldsymbol{y} \,|\, \boldsymbol{\beta}, \sigma^2) = \frac{1}{\sigma^2} \sum_{n=1}^{N} \boldsymbol{x}_n^\top (y_n - \boldsymbol{x}_n \cdot \boldsymbol{\beta}) = \mathbf{0} \qquad \text{(where } \mathbf{0} \text{ is a vector of 0s)}$$

$$\frac{1}{\sigma^2} \boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{X} \cdot \boldsymbol{\beta}) = \mathbf{0}$$

$$\Longrightarrow \boldsymbol{\beta}_{\mathrm{ML}} =$$

- Here, the matrix $\boldsymbol{X}$ is defined as $\boldsymbol{X} = \begin{bmatrix} x_{11} & \dots & x1D \\ \dots & \dots & \dots \\ x_{N1} & \dots & x_{ND} \end{bmatrix}$

Linear Regression and Least Squares

- Derivative w.r.t a single weight entry β_i

$$\frac{d}{\mathrm{d}\beta_d} \ln p(\boldsymbol{y} \,|\, \boldsymbol{\beta}, \sigma^2) = \frac{d}{\mathrm{d}\beta_d} \left[-\frac{1}{2\sigma^2} \sum_{n=1}^{N} (y_n - \boldsymbol{x}_n \cdot \boldsymbol{\beta})^2 \right]$$

$$= \left[-\frac{1}{\sigma^2} \sum_{n=1}^{N} \boldsymbol{x}_{nd} (y_n - \boldsymbol{x}_n \cdot \boldsymbol{\beta}) \right]$$

- Set gradient w.r.t. $\boldsymbol{\beta}$ to zero [vector holding the derivatives $\forall \beta_d$]

$$\nabla_{\boldsymbol{\beta}} \ln p(\boldsymbol{y} \,|\, \boldsymbol{\beta}, \sigma^2) = \frac{1}{\sigma^2} \sum_{n=1}^{N} \boldsymbol{x}_n^\top (y_n - \boldsymbol{x}_n \cdot \boldsymbol{\beta}) = \mathbf{0} \quad \text{(where } \mathbf{0} \text{ is a vector of 0s)}$$

$$\frac{1}{\sigma^2} \boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{X} \cdot \boldsymbol{\beta}) = \mathbf{0}$$

$$\Longrightarrow \boldsymbol{\beta}_{\mathrm{ML}} = \underbrace{(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top}_{\text{Pseudo inverse of } \boldsymbol{X}} \boldsymbol{y}$$

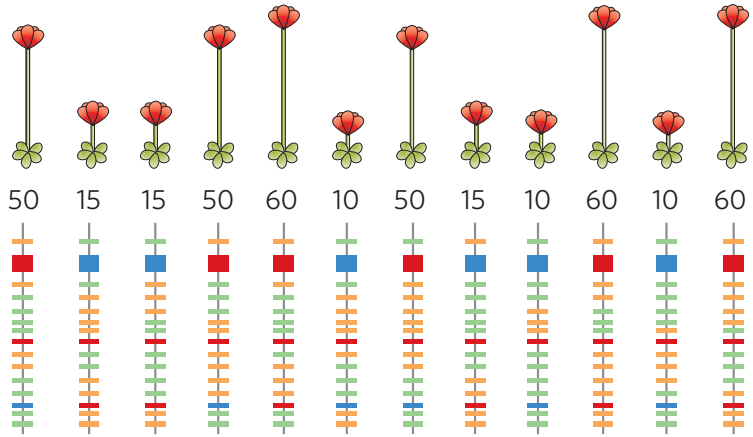
- Here, the matrix $\boldsymbol{X}$ is defined as $\boldsymbol{X} = \begin{bmatrix} x_{11} & \dots & x1D \\ \dots & \dots & \dots \\ x_{N1} & \dots & x_{ND} \end{bmatrix}$

Linear regression

Application: Genome-wide association studies

Given:

- Genetics for multiple individuals
  - e.g.: Single nucleotide polymorphisms (SNPs), microsatelite markers, ...
- *Quantitative* phenotype for the same individuals
  - e.g.: height, gene-expression, ...



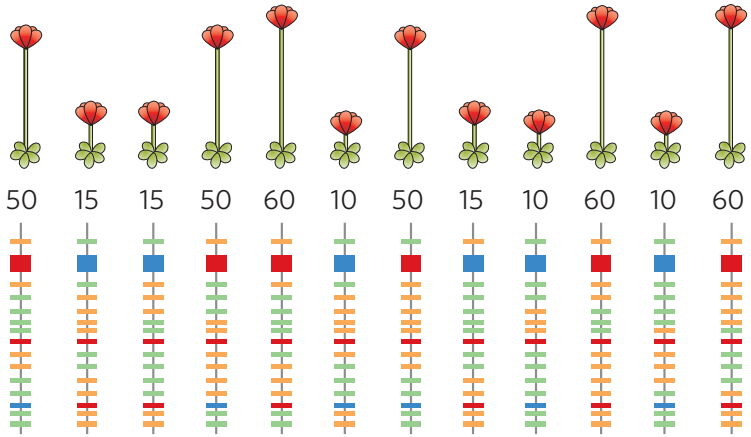
- Try to find genetic markers, that explain the variance in the phenotype.
- Use linear regression!

Linear regression

Application: Genome-wide association studies

Given:

- Genetics for multiple individuals
  - e.g.: Single nucleotide polymorphisms (SNPs), microsatelite markers, ...
- *Quantitative* phenotype for the same individuals
  - e.g.: height, gene-expression, ...
- Try to find genetic markers, that explain the variance in the phenotype.
- Use linear regression!

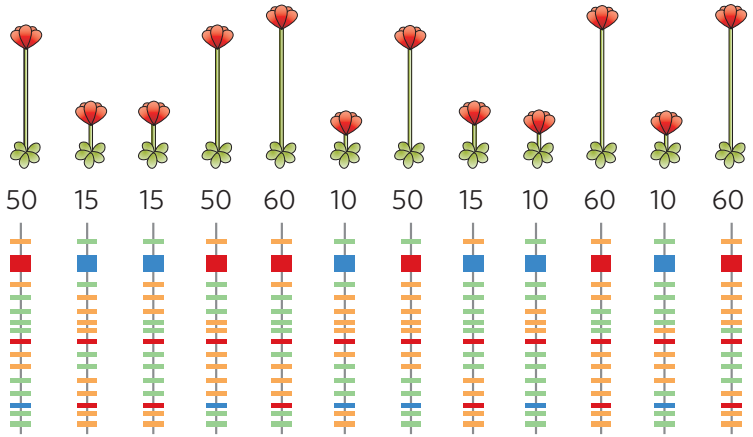


Linear regression

Application: Genome-wide association studies

Given:

- Genetics for multiple individuals
  - e.g.: Single nucleotide polymorphisms (SNPs), microsatelite markers, ...
- *Quantitative* phenotype for the same individuals
  - e.g.: height, gene-expression, ...



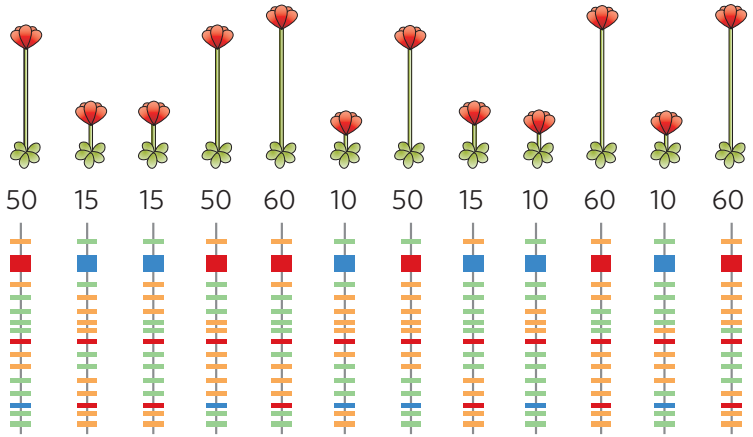
- Try to find genetic markers, that explain the variance in the phenotype.
- Use linear regression!

Linear regression

Application: Genome-wide association studies

Given:

- Genetics for multiple individuals
  - e.g.: Single nucleotide polymorphisms (SNPs), microsatelite markers, ...
- *Quantitative* phenotype for the same individuals
  - e.g.: height, gene-expression, ...



Goal:

- Try to find genetic markers, that explain the variance in the phenotype.
- Use linear regression!

Linear regression

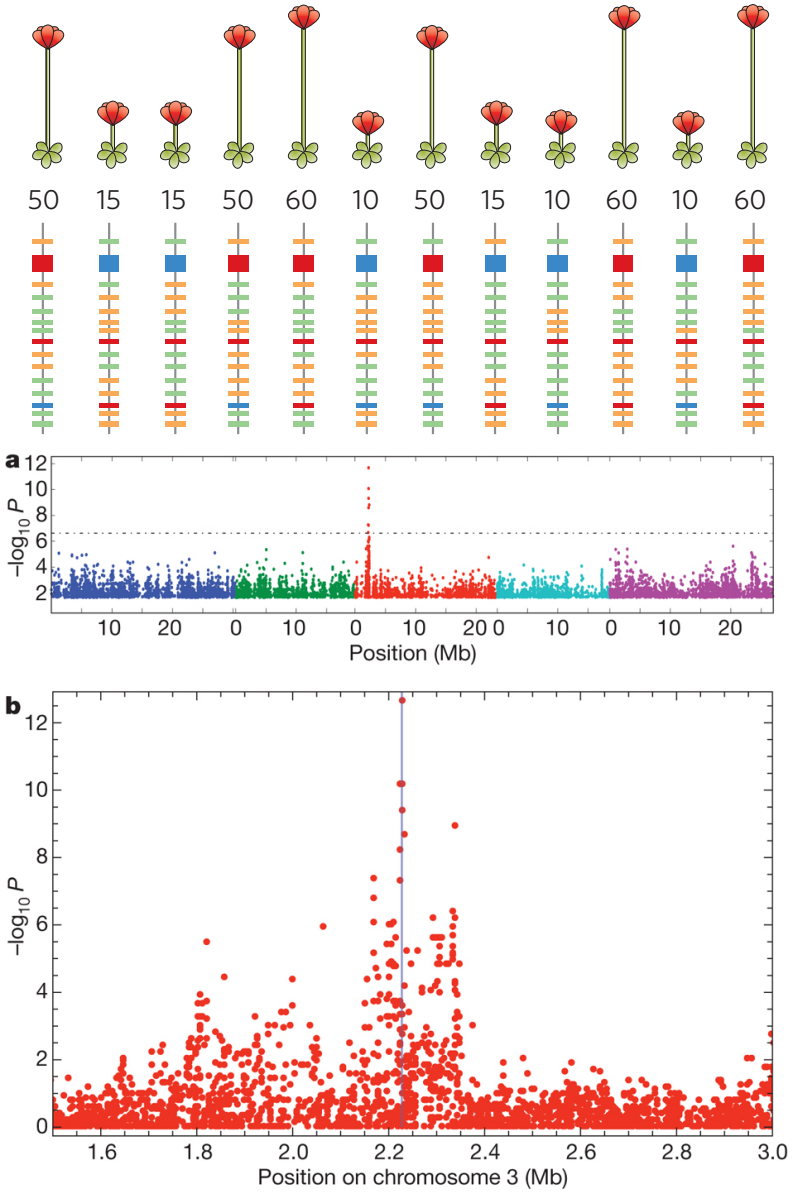
Application: Genome-wide association studies

Given:

- Genetics for multiple individuals
  - e.g.: Single nucleotide polymorphisms (SNPs), microsatelite markers, ...
- *Quantitative* phenotype for the same individuals
  - e.g.: height, gene-expression, ...

Goal:

- Try to find genetic markers, that explain the variance in the phenotype.
- Use linear regression!



Linear regression

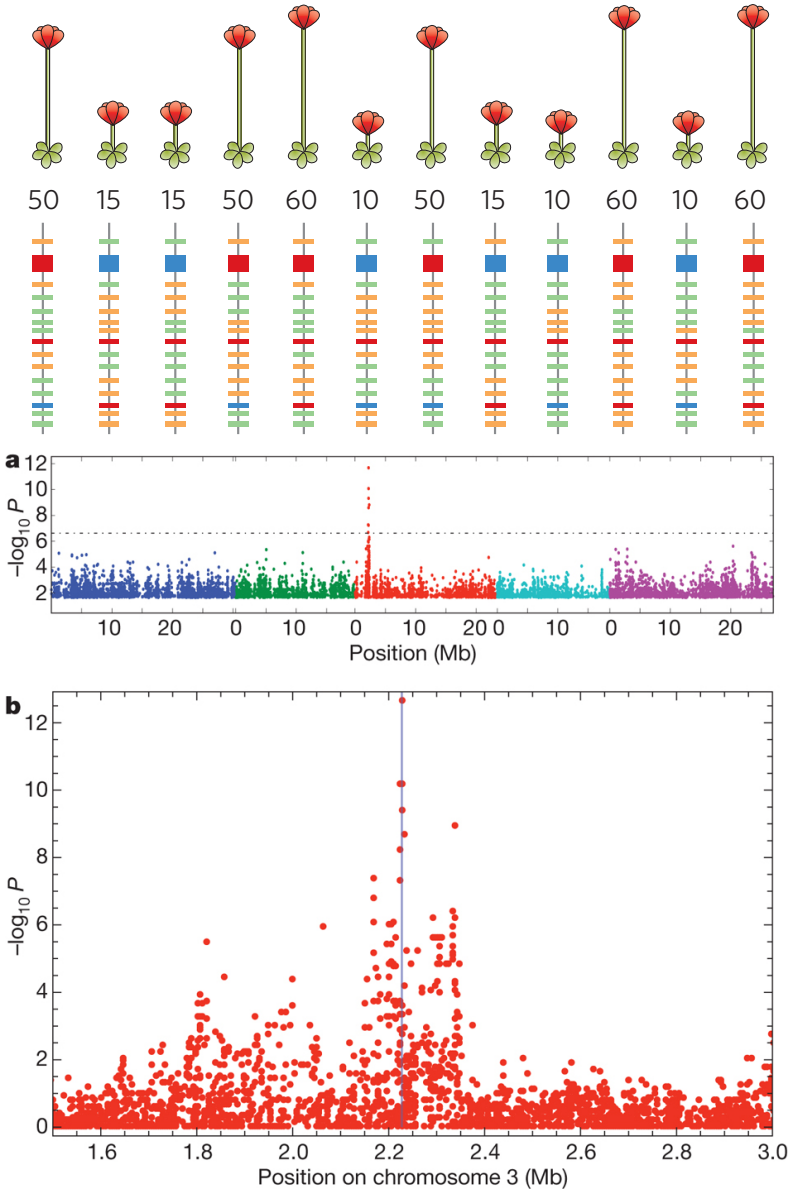
Application: Genome-wide association studies

Given:

- Genetics for multiple individuals
  - e.g.: Single nucleotide polymorphisms (SNPs), microsatelite markers, ...
- *Quantitative* phenotype for the same individuals
  - e.g.: height, gene-expression, ...

Goal:

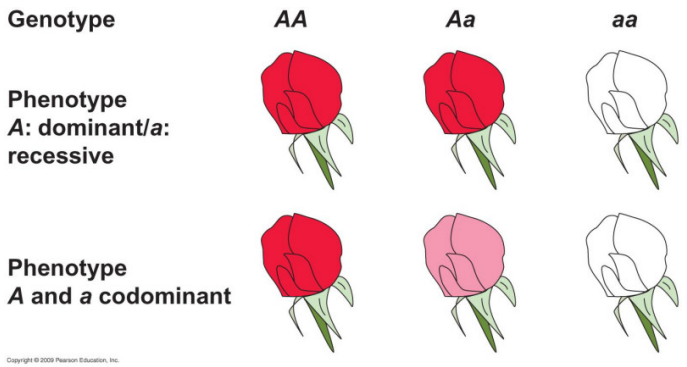
- Try to find genetic markers, that explain the variance in the phenotype.
- Use linear regression!



Genetics 101

Some definitions

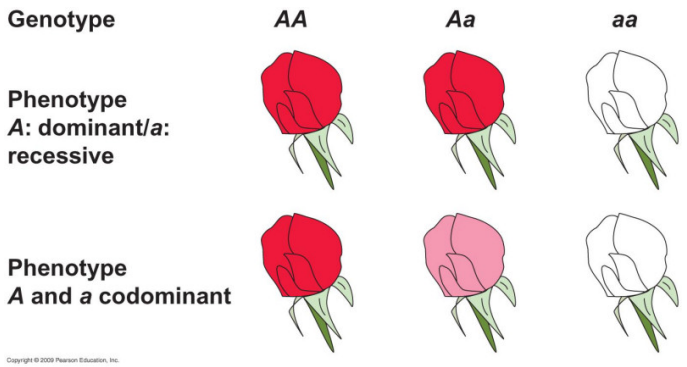
- **Genotype** x denotes the genetic state of an individual.
  - Denoted by $x_{n:}$ for individual n.
- **Phenotype** denotes the state of a trait of an individual.
  - Denoted by y_n for individual n.
- A **Locus** is a position or limited region in the genome.
  - Denoted by x_s for locus (or SNP) s.
- An **allele** is the genetic state of a locus.



Genetics 101

Some definitions

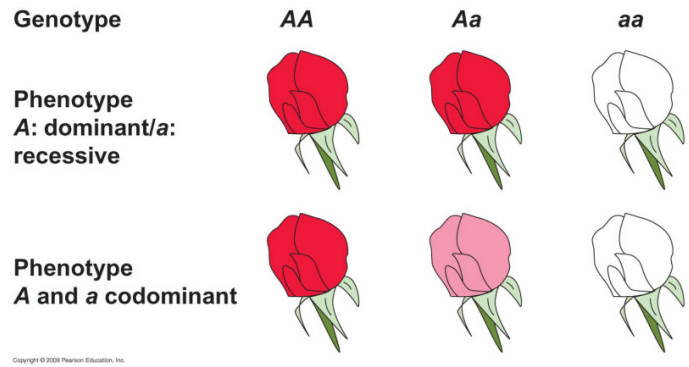
- **Genotype** x denotes the genetic state of an individual.
  - Denoted by $x_{n:}$ for individual n.
- **Phenotype** denotes the state of a trait of an individual.
  - Denoted by y_n for individual n.
- A **Locus** is a position or limited region in the genome.
  - Denoted by x_s for locus (or SNP) s.
- An **allele** is the genetic state of a locus.

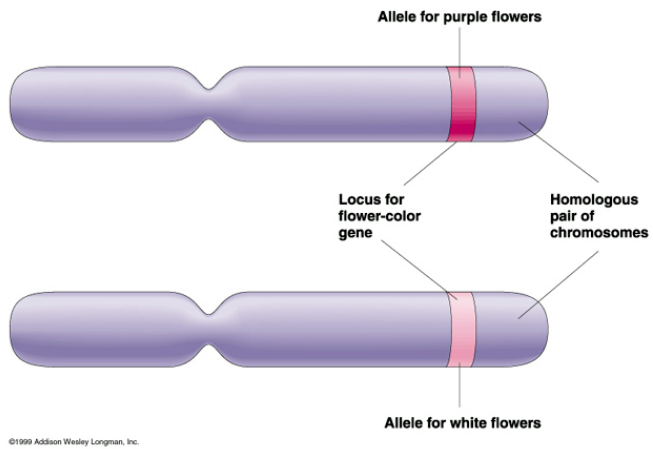


Genetics 101

Some definitions

- **Genotype** x denotes the genetic state of an individual.
  - Denoted by $x_{n:}$ for individual n.
- **Phenotype** denotes the state of a trait of an individual.
  - Denoted by y_n for individual n.
- A **Locus** is a position or limited region in the genome.
  - Denoted by x_s for locus (or SNP) s.
- An **allele** is the genetic state of a locus.

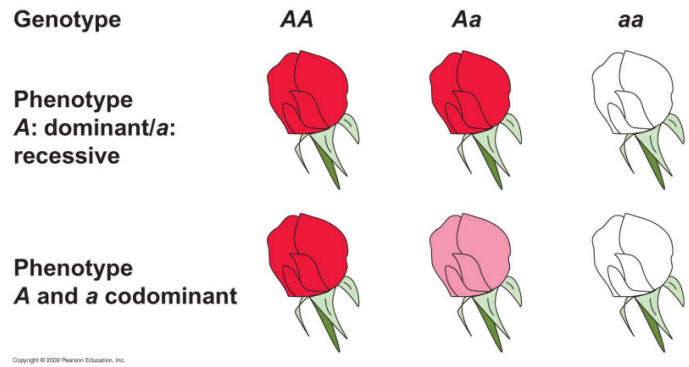


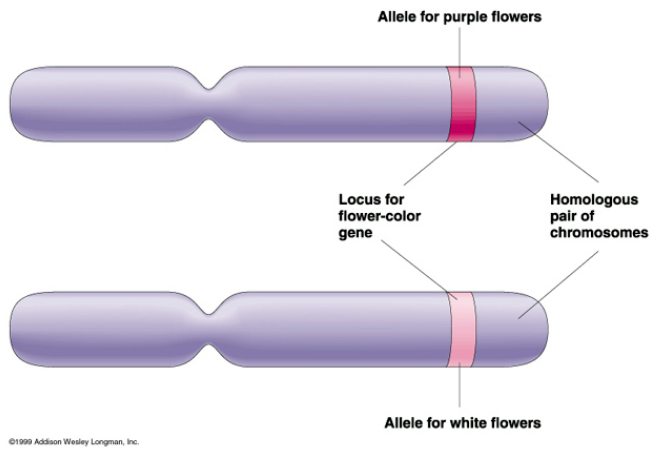


Genetics 101

Some definitions

- **Genotype** x denotes the genetic state of an individual.
  - Denoted by $x_{n:}$ for individual n.
- **Phenotype** denotes the state of a trait of an individual.
  - Denoted by y_n for individual n.
- A **Locus** is a position or limited region in the genome.
  - Denoted by x_s for locus (or SNP) s.
- An **allele** is the genetic state of a locus.

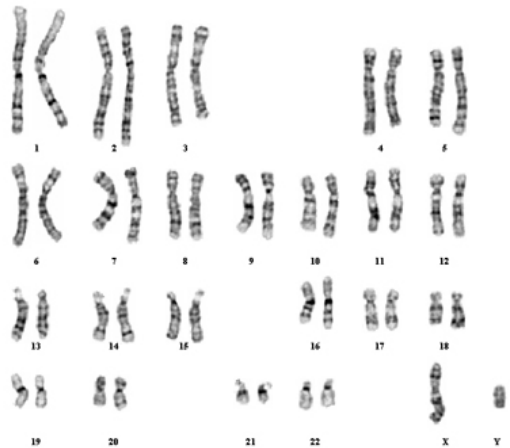




Genetics 101

More definitions

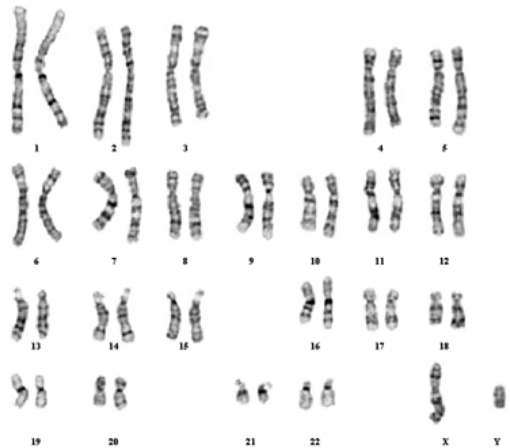
- An organism/cell is **haploid** if it only has one chromosome set or identical chromosome sets.
  - e.g. *A. thaliana*, sperm cells or inbred lab strains
- An organism/cell is **diploid** if it has two separately inherited homologous chromosomes.
  - e.g. *human*
- An organism/cell is **polyploid** if it has more than two homologous chromosomes.
  - e.g. *sugar cane* is hexaploid.



Genetics 101

More definitions

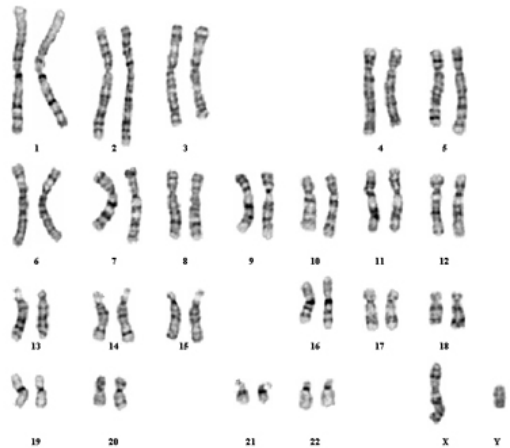
- An organism/cell is **haploid** if it only has one chromosome set or identical chromosome sets.
  - e.g. *A. thaliana*, sperm cells or inbred lab strains
- An organism/cell is **diploid** if it has two separately inherited homologous chromosomes.
  - e.g. *human*
- An organism/cell is **polyploid** if it has more than two homologous chromosomes.
  - e.g. *sugar cane* is hexaploid.



Genetics 101

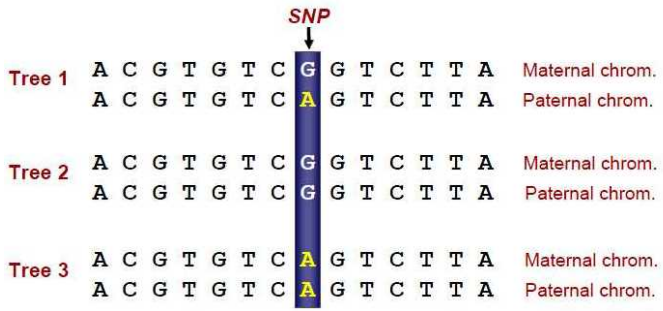
More definitions

- An organism/cell is **haploid** if it only has one chromosome set or identical chromosome sets.
  - e.g. *A. thaliana*, sperm cells or inbred lab strains
- An organism/cell is **diploid** if it has two separately inherited homologous chromosomes.
  - e.g. *human*
- An organism/cell is **polyploid** if it has more than two homologous chromosomes.
  - e.g. *sugar cane* is hexaploid.



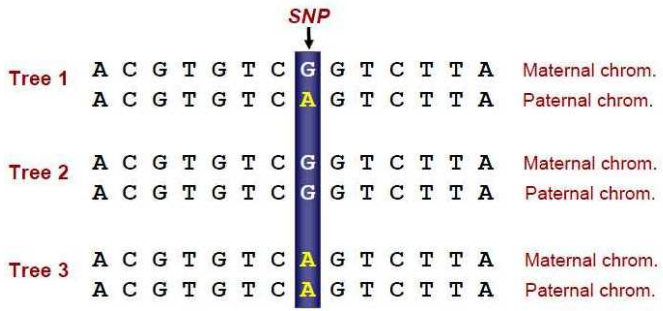
Even more definitions

- **Haplotype** denotes an individual's state of a single set of chromosomes (paternal or maternal).
- A locus s is **heterozygous** if it differs between paternal and maternal haplotypes.
  - heterozygous allele usually encoded as $x_s = 1$
- A locus is **homozygous** if it matches between paternal and maternal haplotypes.
  - homozygous *major* allele usually encoded as $x_s = 0$
  - homozygous *minor* allele usually encoded as $x_s = 2$
- x_{ns} counts the number of minor alleles (mutations) at SNP s.



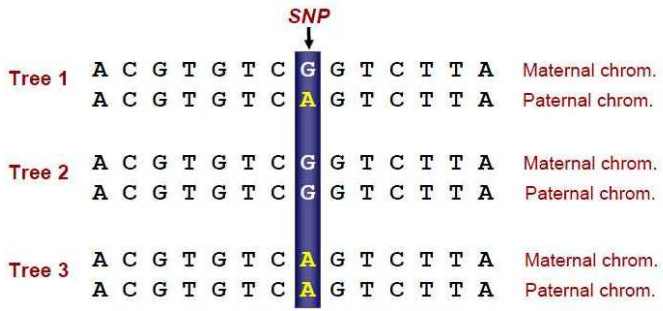
Even more definitions

- **Haplotype** denotes an individual's state of a single set of chromosomes (paternal or maternal).
- A locus s is **heterozygous** if it differs between paternal and maternal haplotypes.
  - heterozygous allele usually encoded as $x_s = 1$
- A locus is **homozygous** if it matches between paternal and maternal haplotypes.
  - homozygous *major* allele usually encoded as $x_s = 0$
  - homozygous *minor* allele usually encoded as $x_s = 2$
- x_{ns} counts the number of minor alleles (mutations) at SNP s.



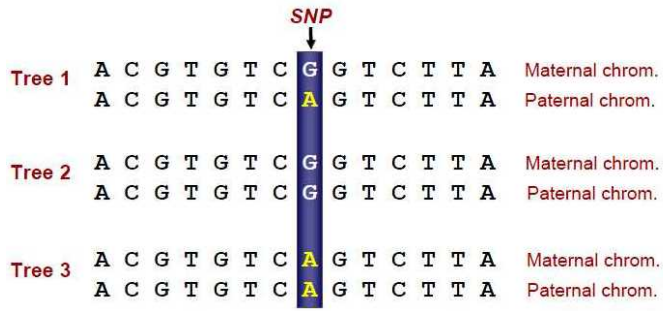
Even more definitions

- **Haplotype** denotes an individual's state of a single set of chromosomes (paternal or maternal).
- A locus s is **heterozygous** if it differs between paternal and maternal haplotypes.
  - heterozygous allele usually encoded as $x_s = 1$
- A locus is **homozygous** if it matches between paternal and maternal haplotypes.
  - homozygous *major* allele usually encoded as $x_s = 0$
  - homozygous *minor* allele usually encoded as $x_s = 2$
- x_{ns} counts the number of minor alleles (mutations) at SNP s.



Even more definitions

- **Haplotype** denotes an individual's state of a single set of chromosomes (paternal or maternal).
- A locus s is **heterozygous** if it differs between paternal and maternal haplotypes.
  - heterozygous allele usually encoded as $x_s = 1$
- A locus is **homozygous** if it matches between paternal and maternal haplotypes.
  - homozygous *major* allele usually encoded as $x_s = 0$
  - homozygous *minor* allele usually encoded as $x_s = 2$
- x_{ns} counts the number of minor alleles (mutations) at SNP s.

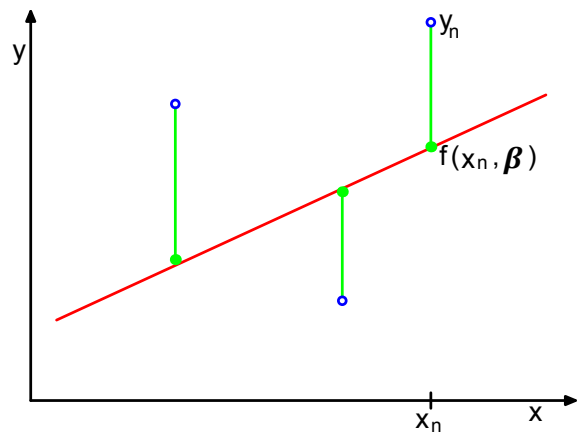


Linear regression

Example: Genome-wide association study

- Model the phenotype of individual n as a linear function of the SNP x_{ns}
- $x_{ns} \in 0, 1, 2$ counts the number of mutations that individual n has at the position s.

$$y_n = \underbrace{x_{ns}\beta}_{\text{genetic effect}} + \underbrace{\epsilon_n}_{\text{noise}}$$



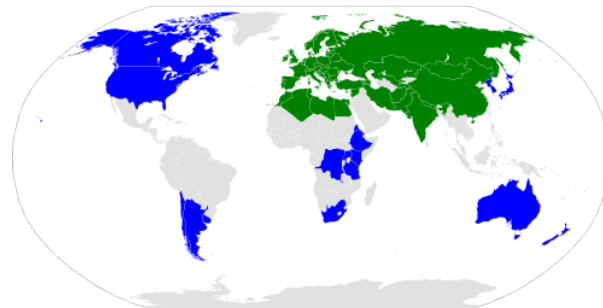
(C.M. Bishop, Pattern Recognition and Machine Learning)

Genome-wide association study

Flowering time in *A. thaliana*

Arabidopsis thaliana

- Grows under rough conditions
- Distributed over the whole globe (it "travels" far)
- Genetics model organism ("lab rat" among plants)
- Mostly self-fertilizing (mostly homozygous)
- Model for flowering

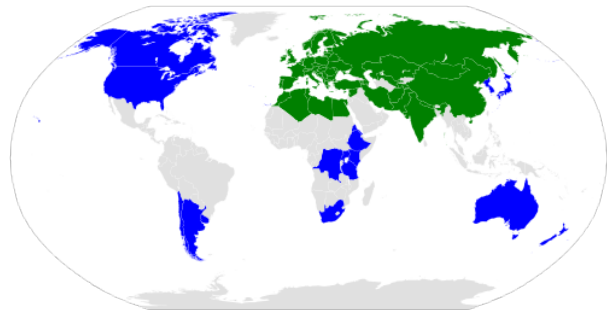


Genome-wide association study

Flowering time in *A. thaliana*

Arabidopsis thaliana

- Grows under rough conditions
- Distributed over the whole globe (it "travels" far)
- Genetics model organism ("lab rat" among plants)
- Mostly self-fertilizing (mostly homozygous)
- Model for flowering

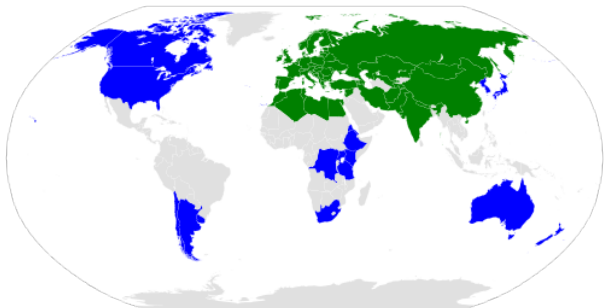


Genome-wide association study

Flowering time in *A. thaliana*

Arabidopsis thaliana

- Grows under rough conditions
- Distributed over the whole globe (it "travels" far)
- Genetics model organism ("lab rat" among plants)
- Mostly self-fertilizing (mostly homozygous)
- Model for flowering

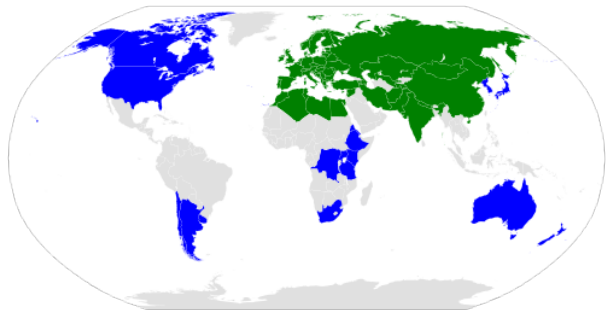


Genome-wide association study

Flowering time in *A. thaliana*

Arabidopsis thaliana

- Grows under rough conditions
- Distributed over the whole globe (it "travels" far)
- Genetics model organism ("lab rat" among plants)
- Mostly self-fertilizing (mostly homozygous)
- Model for flowering

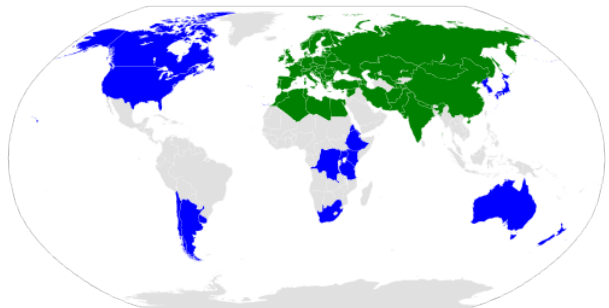


Genome-wide association study

Flowering time in *A. thaliana*

Arabidopsis thaliana

- Grows under rough conditions
- Distributed over the whole globe (it "travels" far)
- Genetics model organism ("lab rat" among plants)
- Mostly self-fertilizing (mostly homozygous)
- Model for flowering

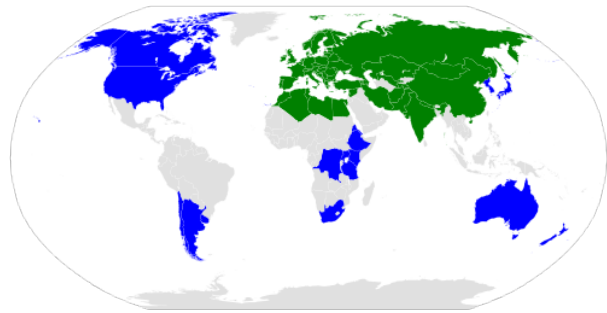


Genome-wide association study

Flowering time in *A. thaliana*

Arabidopsis thaliana

- Grows under rough conditions
- Distributed over the whole globe (it "travels" far)
- Genetics model organism ("lab rat" among plants)
- Mostly self-fertilizing (mostly homozygous)
- Model for flowering

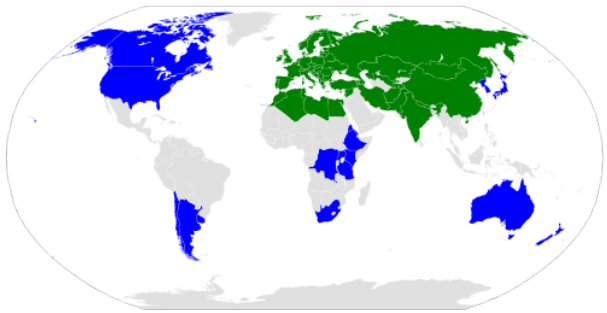


Genome-wide association study

Flowering time in *A. thaliana*

GWAS data [Atwell et al. Nat. Gen. 2010]

- y: Phenotype:
  "Flowering time at 10° Celsius (days)"
- X: Genotype:
  214,553 SNPs, all homozygous
- 194 samples
- show demo.

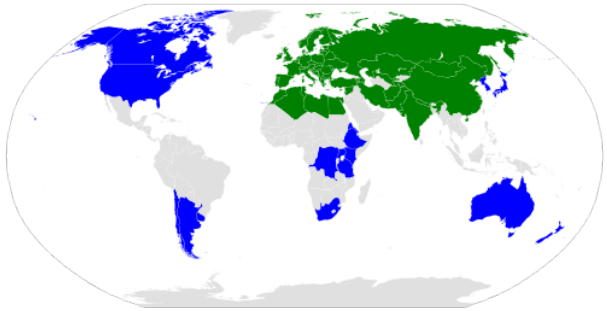


Genome-wide association study

Flowering time in *A. thaliana*

GWAS data [Atwell et al. Nat. Gen. 2010]

- y: Phenotype:
  "Flowering time at 10° Celsius (days)"
- X: Genotype:
  214,553 SNPs, all homozygous
- 194 samples
- show demo.

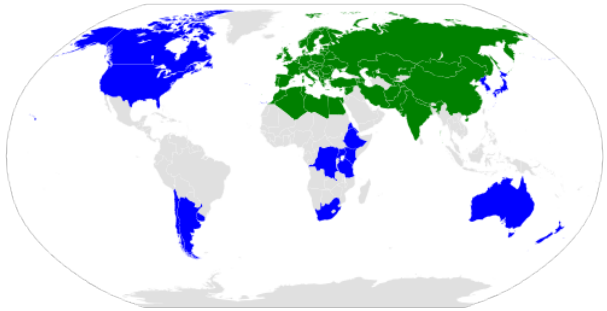


Genome-wide association study

Flowering time in *A. thaliana*

GWAS data [Atwell et al. Nat. Gen. 2010]

- y: Phenotype:
  "Flowering time at 10° Celsius (days)"
- X: Genotype:
  214,553 SNPs, all homozygous
- 194 samples
- show demo.

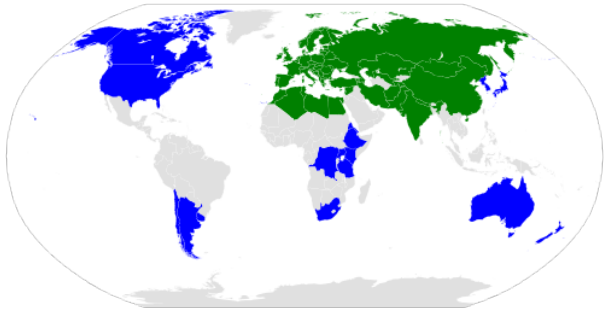


Genome-wide association study

Flowering time in *A. thaliana*

GWAS data [Atwell et al. Nat. Gen. 2010]

- y: Phenotype:
  "Flowering time at 10° Celsius (days)"
- X: Genotype:
  214,553 SNPs, all homozygous
- 194 samples
- show demo.



Outline

Gaussian distribution

Univariate vs. multivariate

- The joint distribution of N samples x_n from independent univariate normal distributions with means μ_n and variances σ_n^2 $\forall n \in [1, \ldots, N]$ is given as

$$\prod_{n=1}^{N} \mathcal{N}\left(x_n \mid \mu_n, \sigma_n^2\right)$$

- Special case: all x_n are also *identically* distributed with mean μ and variance σ^2

$$p(x_1, \ldots, x_N) = \prod_{n=1}^{N} \mathcal{N}\left(x_n \mid \mu, \sigma^2\right) = \mathcal{N}\left(\boldsymbol{x} \mid \mu \cdot \mathbf{1}, \sigma^2 \cdot \boldsymbol{I}\right)$$

- From now on we'll mostly argue about multivariate normal distributions.

Gaussian distribution

Univariate vs. multivariate

- The joint distribution of N samples x_n from independent univariate normal distributions with means μ_n and variances σ_n^2 $\forall n \in [1, \ldots, N]$ is given as

$$\prod_{n=1}^{N} \mathcal{N}\left(x_n \mid \mu_n, \sigma_n^2\right) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left[-\frac{1}{2\sigma_n^2}(x_n - \mu_n)^2\right]$$

- Special case: all x_n are also *identically* distributed with mean μ and variance σ^2

$$p(x_1, \ldots, x_N) = \prod_{n=1}^{N} \mathcal{N}\left(x_n \mid \mu, \sigma^2\right) = \mathcal{N}\left(\boldsymbol{x} \mid \mu \cdot \mathbf{1}, \sigma^2 \cdot \boldsymbol{I}\right)$$

- From now on we'll mostly argue about multivariate normal distributions.

Gaussian distribution

Univariate vs. multivariate

- The joint distribution of N samples x_n from independent univariate normal distributions with means μ_n and variances σ_n^2 $\forall n \in [1, \ldots, N]$ is given as

$$\prod_{n=1}^{N} \mathcal{N}\left(x_n \mid \mu_n, \sigma_n^2\right) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left[-\frac{1}{2\sigma_n^2}(x_n - \mu_n)^2\right]$$

$$= (2\pi)^{-\frac{N}{2}} \left(\prod_{n=1}^{N} \sigma_n^2\right)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\sum_{n=1}^{N} \sigma_n^{-2}(x_n - \mu_n)^2\right]$$

- Special case: all x_n are also *identically* distributed with mean μ and variance σ^2

$$p(x_1, \ldots, x_N) = \prod_{n=1}^{N} \mathcal{N}\left(x_n \mid \mu, \sigma^2\right) = \mathcal{N}\left(\boldsymbol{x} \mid \mu \cdot \mathbf{1}, \sigma^2 \cdot \boldsymbol{I}\right)$$

- From now on we'll mostly argue about multivariate normal distributions.

Gaussian distribution

Univariate vs. multivariate

- The joint distribution of N samples x_n from independent univariate normal distributions with means μ_n and variances σ_n^2 $\forall n \in [1, \ldots, N]$ is given as

$$\begin{aligned}
\prod_{n=1}^{N} \mathcal{N}\left(x_n \mid \mu_n, \sigma_n^2\right) &= \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left[-\frac{1}{2\sigma_n^2}(x_n - \mu_n)^2\right] \\
&= (2\pi)^{-\frac{N}{2}} \left(\prod_{n=1}^{N} \sigma_n^2\right)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\sum_{n=1}^{N} \sigma_n^{-2}(x_n - \mu_n)^2\right] \\
&= (2\pi)^{-\frac{N}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right),
\end{aligned}$$

where we introduced $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \sigma_N^2 \end{bmatrix}$, $\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}$ and $\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_N \end{bmatrix}$.

- Special case: all x_n are also *identically* distributed with mean μ and variance σ^2

$$p(x_1, \ldots, x_N) = \prod_{n=1}^{N} \mathcal{N}\left(x_n \mid \mu, \sigma^2\right) = \mathcal{N}\left(\boldsymbol{x} \mid \mu \cdot \mathbf{1}, \sigma^2 \cdot \boldsymbol{I}\right)$$

- From now on we'll mostly argue about multivariate normal distributions.

Gaussian distribution

Univariate vs. multivariate

- The joint distribution of N samples x_n from independent univariate normal distributions with means μ_n and variances σ_n^2 $\forall n \in [1, \ldots, N]$ is given as

$$\begin{aligned}
\prod_{n=1}^{N} \mathcal{N}\left(x_n \mid \mu_n, \sigma_n^2\right) &= \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left[-\frac{1}{2\sigma_n^2}(x_n - \mu_n)^2\right] \\
&= (2\pi)^{-\frac{N}{2}} \left(\prod_{n=1}^{N} \sigma_n^2\right)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\sum_{n=1}^{N} \sigma_n^{-2}(x_n - \mu_n)^2\right] \\
&= (2\pi)^{-\frac{N}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right),
\end{aligned}$$

where we introduced $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \sigma_N^2 \end{bmatrix}$, $\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}$ and $\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_N \end{bmatrix}$.

- Special case: all x_n are also *identically* distributed with mean μ and variance σ^2

$$p(x_1, \ldots, x_N) = \prod_{n=1}^{N} \mathcal{N}\left(x_n \mid \mu, \sigma^2\right) = \mathcal{N}\left(\boldsymbol{x} \mid \mu \cdot \mathbf{1}, \sigma^2 \cdot \boldsymbol{I}\right)$$

- From now on we'll mostly argue about multivariate normal distributions.

Gaussian distribution

Univariate vs. multivariate

- The joint distribution of N samples x_n from independent univariate normal distributions with means μ_n and variances σ_n^2 $\forall n \in [1, \ldots, N]$ is given as

$$\begin{aligned}\prod_{n=1}^{N} \mathcal{N}\left(x_n \mid \mu_n, \sigma_n^2\right) &= \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left[-\frac{1}{2\sigma_n^2}(x_n - \mu_n)^2\right] \\ &= (2\pi)^{-\frac{N}{2}} \left(\prod_{n=1}^{N} \sigma_n^2\right)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\sum_{n=1}^{N} \sigma_n^{-2}(x_n - \mu_n)^2\right] \\ &= (2\pi)^{-\frac{N}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right),\end{aligned}$$

where we introduced $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \sigma_N^2 \end{bmatrix}$, $\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}$ and $\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_N \end{bmatrix}$.

- Special case: all x_n are also *identically* distributed with mean μ and variance σ^2

$$p(x_1, \ldots, x_N) = \prod_{n=1}^{N} \mathcal{N}\left(x_n \mid \mu, \sigma^2\right) = \mathcal{N}\left(\boldsymbol{x} \mid \mu \cdot \mathbf{1}, \sigma^2 \cdot \boldsymbol{I}\right)$$

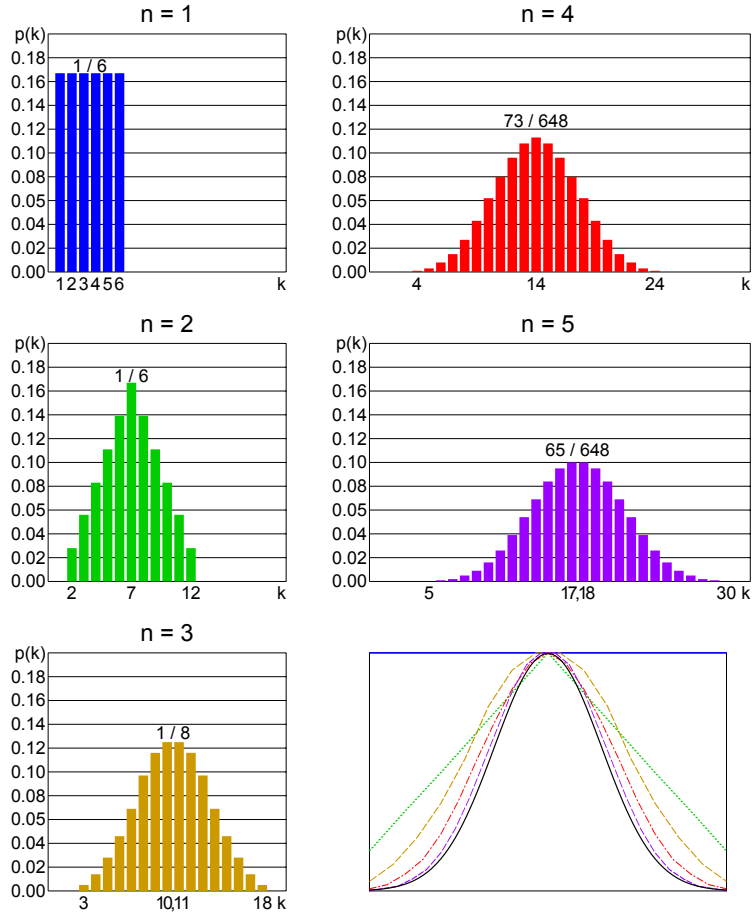
- From now on we'll mostly argue about multivariate normal distributions.

Distributions of sums and means

- Throw a dice N times.
  - Distribution of the sum?
- Independent samples from a uniform distribution on the interval (0,1).
  - Distribution of the mean?

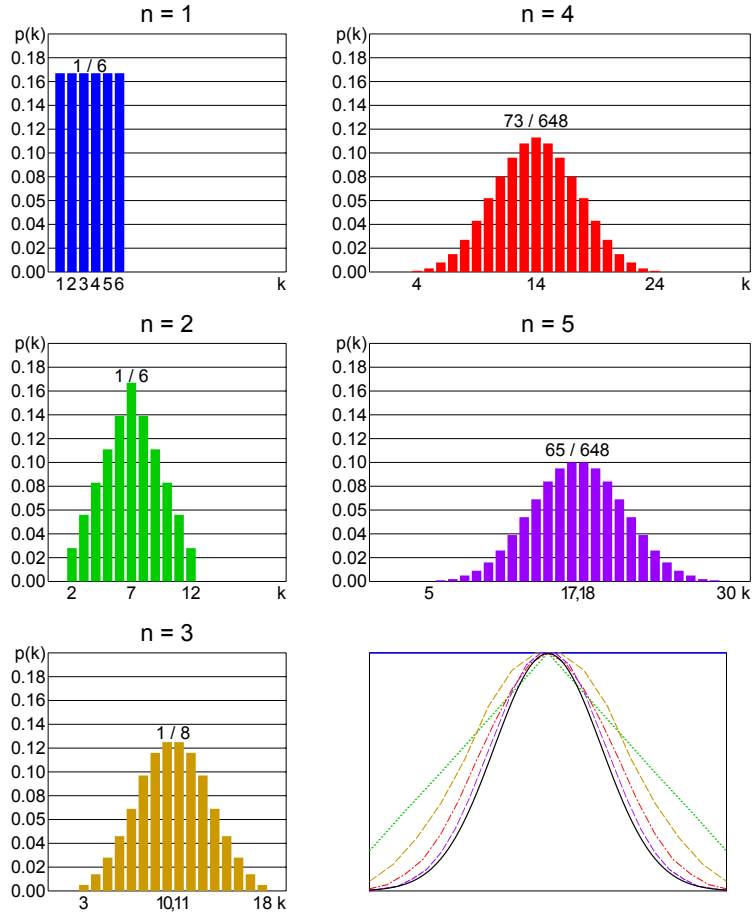
Distributions of sums and means

- Throw a dice N times.
  - Distribution of the sum?
- Independent samples from a uniform distribution on the interval (0,1).
  - Distribution of the mean?



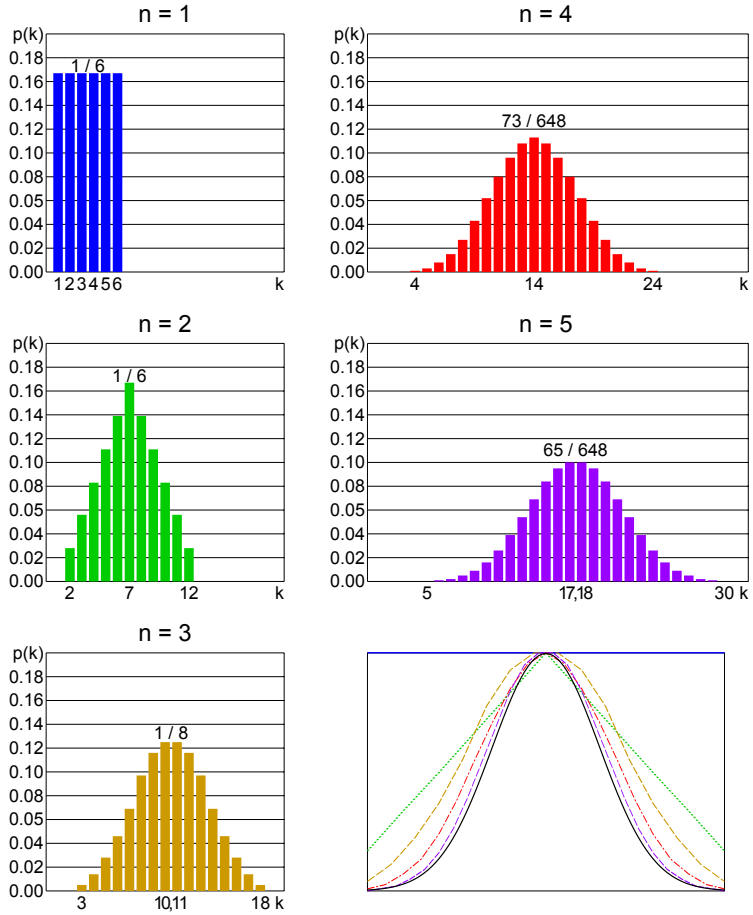
Distributions of sums and means

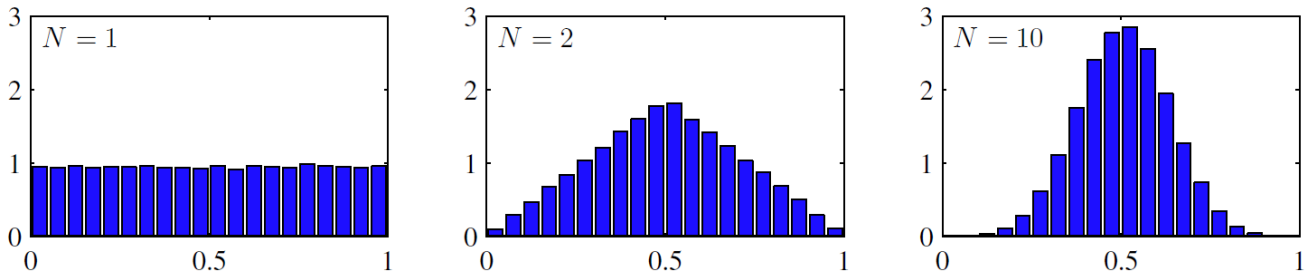
- Throw a dice N times.
  - Distribution of the sum?
- Independent samples from a uniform distribution on the interval (0,1).
  - Distribution of the mean?



Distributions of sums and means

- Throw a dice N times.
  - Distribution of the sum?
- Independent samples from a uniform distribution on the interval (0,1).
  - Distribution of the mean?





(C.M. Bishop, Pattern Recognition and Machine Learning)

Central limit theorem

- $\boldsymbol{x_n} \sim P(\boldsymbol{x})$, unknown distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.
- Given N independent, identically distributed (i.i.d.) samples from $P(x)$: $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$, let $\bar{\boldsymbol{x}}^{(N)}$ be the sample mean

$$\sqrt{N}\left(\bar{\boldsymbol{x}}^{(N)} - \boldsymbol{\mu}\right) = \sqrt{N}\underbrace{\left(\frac{1}{N}\left(\sum_{n=1}^{N} \boldsymbol{x}_n\right) - \boldsymbol{\mu}\right)}_{\to 0 \text{ (a.s.)}}$$

$$\Rightarrow P\left(\sqrt{N}\left(\bar{\boldsymbol{x}}^{(N)} - \boldsymbol{\mu}\right)\right) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{\Sigma})$$

- Only mean and covariance retained when averaging i.i.d. variables.
- Distribution becomes Gaussian
- Gaussian is a "limit distribution"
  Implication: Once something is Gaussian it usually stays Gaussian under many operations
  - Normal distribution is closed under linear transformations

Central limit theorem

- $\boldsymbol{x_n} \sim P(\boldsymbol{x})$, unknown distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.
- Given N independent, identically distributed (i.i.d.) samples from $P(x)$: $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$, let $\bar{\boldsymbol{x}}^{(N)}$ be the sample mean

$$\sqrt{N}\left(\bar{\boldsymbol{x}}^{(N)} - \boldsymbol{\mu}\right) = \sqrt{N} \underbrace{\left(\frac{1}{N}\left(\sum_{n=1}^{N} \boldsymbol{x}_n\right) - \boldsymbol{\mu}\right)}_{\to 0 (a.s.)}$$

$$\Rightarrow P\left(\sqrt{N}\left(\bar{\boldsymbol{x}}^{(N)} - \boldsymbol{\mu}\right)\right) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$$

- Only mean and covariance retained when averaging i.i.d. variables.
- Distribution becomes Gaussian
- Gaussian is a "limit distribution"
  Implication: Once something is Gaussian it usually stays Gaussian under many operations
  - Normal distribution is closed under [illegible]

Central limit theorem

- $\boldsymbol{x_n} \sim P(\boldsymbol{x})$, unknown distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.
- Given N independent, identically distributed (i.i.d.) samples from $P(x)$: $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$, let $\bar{\boldsymbol{x}}^{(N)}$ be the sample mean

$$\sqrt{N}\left(\bar{\boldsymbol{x}}^{(N)} - \boldsymbol{\mu}\right) = \sqrt{N} \underbrace{\left(\frac{1}{N}\left(\sum_{n=1}^{N} \boldsymbol{x}_n\right) - \boldsymbol{\mu}\right)}_{\to 0 (a.s.)}$$

$$\Rightarrow P\left(\sqrt{N}\left(\bar{\boldsymbol{x}}^{(N)} - \boldsymbol{\mu}\right)\right) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$$

- Only mean and covariance retained when averaging i.i.d. variables.
- Distribution becomes Gaussian
- Gaussian is a "limit distribution"
  Implication: Once something is Gaussian it usually stays Gaussian under many operations
  - Normal distribution is closed under [illegible]

Central limit theorem

- $\boldsymbol{x_n} \sim P(\boldsymbol{x})$, unknown distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.
- Given N independent, identically distributed (i.i.d.) samples from $P(x)$: $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$, let $\bar{\boldsymbol{x}}^{(N)}$ be the sample mean

$$\sqrt{N}\left(\bar{\boldsymbol{x}}^{(N)} - \boldsymbol{\mu}\right) = \sqrt{N} \underbrace{\left(\frac{1}{N}\left(\sum_{n=1}^{N} \boldsymbol{x}_n\right) - \boldsymbol{\mu}\right)}_{\to 0 (a.s.)}$$

$$\Rightarrow P\left(\sqrt{N}\left(\bar{\boldsymbol{x}}^{(N)} - \boldsymbol{\mu}\right)\right) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$$

- Only mean and covariance retained when averaging i.i.d. variables.
- Distribution becomes Gaussian
- Gaussian is a "limit distribution"
  Implication: Once something is Gaussian it usually stays Gaussian under many operations
  - Normal distribution is closed under [illegible]

Central limit theorem

- $\boldsymbol{x_n} \sim P(\boldsymbol{x})$, unknown distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.
- Given N independent, identically distributed (i.i.d.) samples from $P(x)$: $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$, let $\bar{\boldsymbol{x}}^{(N)}$ be the sample mean

$$\sqrt{N}\left(\bar{\boldsymbol{x}}^{(N)} - \boldsymbol{\mu}\right) = \sqrt{N} \underbrace{\left(\frac{1}{N}\left(\sum_{n=1}^{N} \boldsymbol{x}_n\right) - \boldsymbol{\mu}\right)}_{\to 0 (a.s.)}$$

$$\Rightarrow P\left(\sqrt{N}\left(\bar{\boldsymbol{x}}^{(N)} - \boldsymbol{\mu}\right)\right) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$$

- Only mean and covariance retained when averaging i.i.d. variables.
- Distribution becomes Gaussian
- Gaussian is a "limit distribution"
  Implication: Once something is Gaussian it usually stays Gaussian under many operations
  - Normal distribution is *closed* under linear transformation, marginalization, conditioning.

Central limit theorem

- $\boldsymbol{x_n} \sim P(\boldsymbol{x})$, unknown distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.
- Given N independent, identically distributed (i.i.d.) samples from $P(x)$: $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$, let $\bar{\boldsymbol{x}}^{(N)}$ be the sample mean

$$\sqrt{N}\left(\bar{\boldsymbol{x}}^{(N)} - \boldsymbol{\mu}\right) = \sqrt{N} \underbrace{\left(\frac{1}{N}\left(\sum_{n=1}^{N} \boldsymbol{x}_n\right) - \boldsymbol{\mu}\right)}_{\to 0 (a.s.)}$$

$$\Rightarrow P\left(\sqrt{N}\left(\bar{\boldsymbol{x}}^{(N)} - \boldsymbol{\mu}\right)\right) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$$

- Only mean and covariance retained when averaging i.i.d. variables.
- Distribution becomes Gaussian
- Gaussian is a "limit distribution"
  Implication: Once something is Gaussian it usually stays Gaussian under many operations
  - Normal distribution is *closed* under linear transformation marginalization, conditioning.

Central limit theorem

- $\boldsymbol{x_n} \sim P(\boldsymbol{x})$, unknown distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.
- Given N independent, identically distributed (i.i.d.) samples from $P(x)$: $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$, let $\bar{\boldsymbol{x}}^{(N)}$ be the sample mean

$$\sqrt{N}\left(\bar{\boldsymbol{x}}^{(N)} - \boldsymbol{\mu}\right) = \sqrt{N} \underbrace{\left(\frac{1}{N}\left(\sum_{n=1}^{N} \boldsymbol{x}_n\right) - \boldsymbol{\mu}\right)}_{\to 0 (a.s.)}$$

$$\Rightarrow P\left(\sqrt{N}\left(\bar{\boldsymbol{x}}^{(N)} - \boldsymbol{\mu}\right)\right) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$$

- Only mean and covariance retained when averaging i.i.d. variables.
- Distribution becomes Gaussian
- Gaussian is a "limit distribution"
  Implication: Once something is Gaussian it usually stays Gaussian under many operations
  - Normal distribution is *closed* under linear transformation marginalization, conditioning.

Gaussian distribution

Linear transformation

For any random variable $\boldsymbol{x}$ with mean $\mathbb{E}[\boldsymbol{x}]$ and covariance $\mathbb{C}[\boldsymbol{x}]$

- $\mathbb{E}[\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}] = \boldsymbol{A}\mathbb{E}[\boldsymbol{x}]+\boldsymbol{b}$; (linearity of expectation)

$$\mathbb{E}[\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}] = \sum_{x\in\mathcal{X}}(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b})p(x) = \boldsymbol{A}\underbrace{\left(\sum_{x\in\mathcal{X}}\boldsymbol{x}p(x)\right)}_{\mathbb{E}[\boldsymbol{x}]}+\boldsymbol{b}$$

- $\mathbb{C}[\boldsymbol{A}\boldsymbol{x}] = \boldsymbol{A}\mathbb{C}[\boldsymbol{x}]\boldsymbol{A}^\top$

$$\begin{aligned}\mathbb{C}[\boldsymbol{A}\boldsymbol{x}] &= \mathbb{E}\left[\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{A}^\top\right]-\mathbb{E}[\boldsymbol{A}\boldsymbol{x}]\mathbb{E}[\boldsymbol{A}\boldsymbol{x}]^\top\\ &= \boldsymbol{A}\left(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]-\mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{x}]^\top\right)\boldsymbol{A}^\top\\ &= \boldsymbol{A}\mathbb{C}[\boldsymbol{x}]\boldsymbol{A}^\top\end{aligned}$$

- Gaussian distribution only has mean and covariance

Gaussian distribution

Linear transformation

For any random variable $\boldsymbol{x}$ with mean $\mathbb{E}[\boldsymbol{x}]$ and covariance $\mathbb{C}[\boldsymbol{x}]$

- $\mathbb{E}[\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}] = \boldsymbol{A}\mathbb{E}[\boldsymbol{x}]+\boldsymbol{b}$; (linearity of expectation)

$$\mathbb{E}[\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}] = \sum_{x\in\mathcal{X}}(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b})p(x) = \boldsymbol{A}\underbrace{\left(\sum_{x\in\mathcal{X}}\boldsymbol{x}p(x)\right)}_{\mathbb{E}[\boldsymbol{x}]}+\boldsymbol{b}$$

- $\mathbb{C}[\boldsymbol{A}\boldsymbol{x}] = \boldsymbol{A}\mathbb{C}[\boldsymbol{x}]\boldsymbol{A}^\top$

$$\begin{aligned}\mathbb{C}[\boldsymbol{A}\boldsymbol{x}] &= \mathbb{E}\left[\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{A}^\top\right] - \mathbb{E}[\boldsymbol{A}\boldsymbol{x}]\,\mathbb{E}[\boldsymbol{A}\boldsymbol{x}]^\top \\ &= \boldsymbol{A}\left(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] - \mathbb{E}[\boldsymbol{x}]\,\mathbb{E}[\boldsymbol{x}]^\top\right)\boldsymbol{A}^\top \\ &= \boldsymbol{A}\mathbb{C}[\boldsymbol{x}]\boldsymbol{A}^\top\end{aligned}$$

- Gaussian distribution only has mean and covariance

Gaussian distribution

Linear transformation

For any random variable $\boldsymbol{x}$ with mean $\mathbb{E}[\boldsymbol{x}]$ and covariance $\mathbb{C}[\boldsymbol{x}]$

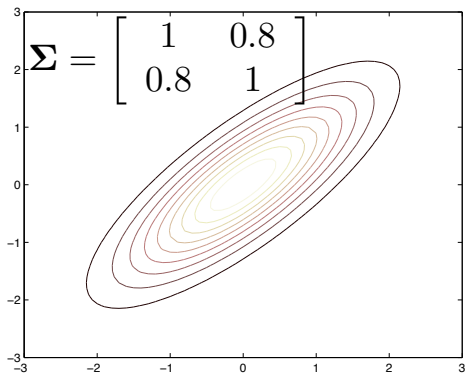
- $\mathbb{E}[\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}] = \boldsymbol{A}\mathbb{E}[\boldsymbol{x}]+\boldsymbol{b}$; (linearity of expectation)

$$\mathbb{E}[\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}] = \sum_{x\in\mathcal{X}}(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b})p(x) = \boldsymbol{A}\underbrace{\left(\sum_{x\in\mathcal{X}}\boldsymbol{x}p(x)\right)}_{\mathbb{E}[\boldsymbol{x}]}+\boldsymbol{b}$$

- $\mathbb{C}[\boldsymbol{A}\boldsymbol{x}] = \boldsymbol{A}\mathbb{C}[\boldsymbol{x}]\boldsymbol{A}^\top$

$$\begin{aligned}\mathbb{C}[\boldsymbol{A}\boldsymbol{x}] &= \mathbb{E}\left[\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{A}^\top\right] - \mathbb{E}[\boldsymbol{A}\boldsymbol{x}]\,\mathbb{E}[\boldsymbol{A}\boldsymbol{x}]^\top \\ &= \boldsymbol{A}\left(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] - \mathbb{E}[\boldsymbol{x}]\,\mathbb{E}[\boldsymbol{x}]^\top\right)\boldsymbol{A}^\top \\ &= \boldsymbol{A}\mathbb{C}[\boldsymbol{x}]\boldsymbol{A}^\top\end{aligned}$$

- Gaussian distribution only has mean and covariance



Gaussian distribution

Linear transformation

For any random variable $\boldsymbol{x}$ with mean $\mathbb{E}[\boldsymbol{x}]$ and covariance $\mathbb{C}[\boldsymbol{x}]$

- $\mathbb{E}[\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}] = \boldsymbol{A}\mathbb{E}[\boldsymbol{x}]+\boldsymbol{b}$; (linearity of expectation)

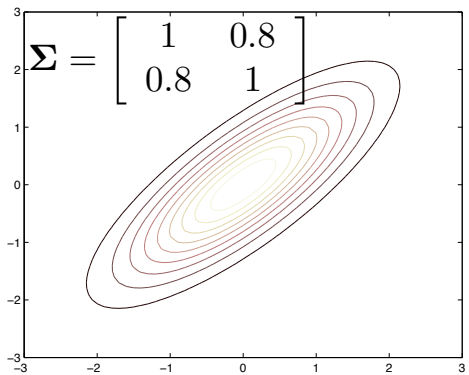
$$\mathbb{E}[\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}] = \sum_{x\in\mathcal{X}}(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b})p(x) = \boldsymbol{A}\underbrace{\left(\sum_{x\in\mathcal{X}}\boldsymbol{x}p(x)\right)}_{\mathbb{E}[\boldsymbol{x}]}+\boldsymbol{b}$$

- $\mathbb{C}[\boldsymbol{A}\boldsymbol{x}] = \boldsymbol{A}\mathbb{C}[\boldsymbol{x}]\boldsymbol{A}^\top$

$$\begin{aligned}\mathbb{C}[\boldsymbol{A}\boldsymbol{x}] &= \mathbb{E}\left[\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{A}^\top\right]-\mathbb{E}[\boldsymbol{A}\boldsymbol{x}]\mathbb{E}[\boldsymbol{A}\boldsymbol{x}]^\top\\ &= \boldsymbol{A}\left(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]-\mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{x}]^\top\right)\boldsymbol{A}^\top\\ &= \boldsymbol{A}\mathbb{C}[\boldsymbol{x}]\boldsymbol{A}^\top\end{aligned}$$

- Gaussian distribution only has mean and covariance

$$\begin{aligned}&\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})\\ \Rightarrow &\boldsymbol{y}=\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}\sim\mathcal{N}\left(\boldsymbol{A}\boldsymbol{\mu}+\boldsymbol{b},\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top\right)\end{aligned}$$



Marginal distribution of a Gaussian

Independent case

Given a Gaussian distributed random variable $\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, with mean and covariance

$$\mathbb{E}[\boldsymbol{x}] = \begin{bmatrix} \mathbb{E}[x_1] \\ \mathbb{E}[x_2] \end{bmatrix}, \qquad \mathbb{C}[\boldsymbol{x}] = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

- What is $p(x_1)$?

$$p(x_1) = \int_{x_2} \mathcal{N}\left(\begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{bmatrix} \mid \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \right) \mathrm{d}\, x_2 =$$

$$= \int_{x_2} \mathcal{N}\left(x_1 \mid \mu_1, \sigma_1^2\right) \mathcal{N}\left(x_1 \mid \mu_1, \sigma_1^2\right) \mathrm{d}\, x_2 =$$

$$= \mathcal{N}\left(x_1 \mid \mu_1, \sigma_1^2\right) \underbrace{\int_{x_2} \mathcal{N}\left(x_1 \mid \mu_1, \sigma_1^2\right) \mathrm{d}\, x_2}_{1} =$$

$$= \mathcal{N}\left(x_1 \mid \mu_1, \sigma_1^2\right)$$

- observation: The marginal distribution is Gaussian (Closure property!)

Marginal distribution of a Gaussian

Independent case

Given a Gaussian distributed random variable $\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, with mean and covariance

$$\mathbb{E}[\boldsymbol{x}] = \begin{bmatrix} \mathbb{E}[x_1] \\ \mathbb{E}[x_2] \end{bmatrix}, \qquad \mathbb{C}[\boldsymbol{x}] = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

- What is $p(x_1)$?

$$\begin{aligned} p(x_1) &= \int_{x_2} \mathcal{N}\left(\begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{bmatrix} \mid \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \right) \mathrm{d}\, x_2 = \\ &= \int_{x_2} \mathcal{N}\left(x_1 \mid \mu_1 , \sigma_1^2 \right) \mathcal{N}\left(x_1 \mid \mu_1 , \sigma_1^2 \right) \mathrm{d}\, x_2 = \\ &= \mathcal{N}\left(x_1 \mid \mu_1 , \sigma_1^2 \right) \underbrace{\int_{x_2} \mathcal{N}\left(x_1 \mid \mu_1 , \sigma_1^2 \right) \mathrm{d}\, x_2}_{1} = \\ &= \mathcal{N}\left(x_1 \mid \mu_1 , \sigma_1^2 \right) \end{aligned}$$

- observation: The marginal distribution is Gaussian (Closure property!)

Marginal distribution of a Gaussian

Independent case

Given a Gaussian distributed random variable $\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, with mean and covariance

$$\mathbb{E}[\boldsymbol{x}] = \begin{bmatrix} \mathbb{E}[x_1] \\ \mathbb{E}[x_2] \end{bmatrix}, \qquad \mathbb{C}[\boldsymbol{x}] = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

- What is $p(x_1)$?

$$\begin{aligned} p(x_1) &= \int_{x_2} \mathcal{N}\left(\begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{bmatrix} \mid \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \right) \mathrm{d}\, x_2 = \\ &= \int_{x_2} \mathcal{N}\left(x_1 \mid \mu_1 , \sigma_1^2 \right) \mathcal{N}\left(x_1 \mid \mu_1 , \sigma_1^2 \right) \mathrm{d}\, x_2 = \\ &= \mathcal{N}\left(x_1 \mid \mu_1 , \sigma_1^2 \right) \underbrace{\int_{x_2} \mathcal{N}\left(x_1 \mid \mu_1 , \sigma_1^2 \right) \mathrm{d}\, x_2}_{1} = \\ &= \mathcal{N}\left(x_1 \mid \mu_1 , \sigma_1^2 \right) \end{aligned}$$

- observation: The marginal distribution is Gaussian (Closure property!)

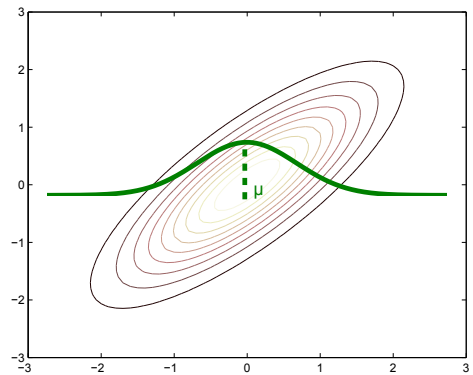
Marginal distribution of a Gaussian

Multivariate case

Given a $(N_1 + N_2)$-dimensional Gaussian distributed random variable $\boldsymbol{x} = \left[\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top\right]^\top$, with mean and covariance

$$\mathbb{E}[\boldsymbol{x}] = \begin{bmatrix} \mathbb{E}[\boldsymbol{x}_1] \\ \mathbb{E}[\boldsymbol{x}_2] \end{bmatrix}, \qquad \mathbb{C}[\boldsymbol{x}] = \begin{bmatrix} \mathbb{C}[\boldsymbol{x}_1] & \mathbb{C}[\boldsymbol{x}_1, \boldsymbol{x}_2] \\ \mathbb{C}[\boldsymbol{x}_1, \boldsymbol{x}_2]^\top & \mathbb{C}[\boldsymbol{x}_2] \end{bmatrix}$$

- What is $p(\boldsymbol{x}_1)$? ($\boldsymbol{x}_2$ N_1-dimensional)
- Can we find a linear transformation $\boldsymbol{A}$, such that $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{x}_1$?



- Attention: **Dimensionality** of the distribution is now N_1, instead of N!
  If we used formula for N-dimensional normal distribution, then the distribution is **not normalized**!

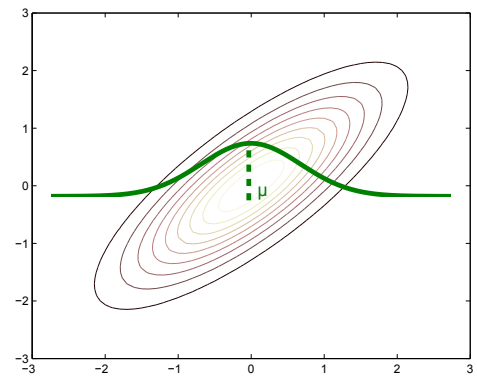
Marginal distribution of a Gaussian

Multivariate case

Given a $(N_1 + N_2)$-dimensional Gaussian distributed random variable $\boldsymbol{x} = \left[\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top\right]^\top$, with mean and covariance

$$\mathbb{E}[\boldsymbol{x}] = \begin{bmatrix} \mathbb{E}[\boldsymbol{x}_1] \\ \mathbb{E}[\boldsymbol{x}_2] \end{bmatrix}, \qquad \mathbb{C}[\boldsymbol{x}] = \begin{bmatrix} \mathbb{C}[\boldsymbol{x}_1] & \mathbb{C}[\boldsymbol{x}_1, \boldsymbol{x}_2] \\ \mathbb{C}[\boldsymbol{x}_1, \boldsymbol{x}_2]^\top & \mathbb{C}[\boldsymbol{x}_2] \end{bmatrix}$$

- What is $p(\boldsymbol{x}_1)$? ($\boldsymbol{x}_2$ N_1-dimensional)
- Can we find a linear transformation $\boldsymbol{A}$, such that $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{x}_1$?



- Attention: **Dimensionality** of the distribution is now N_1, instead of N!
  If we used formula for N-dimensional normal distribution, then the distribution is **not normalized**!

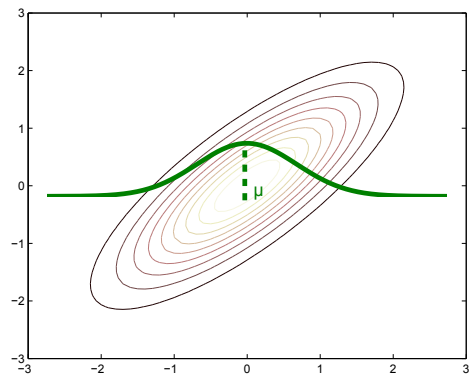
Marginal distribution of a Gaussian

Multivariate case

Given a $(N_1 + N_2)$-dimensional Gaussian distributed random variable $\boldsymbol{x} = \left[\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top\right]^\top$, with mean and covariance

$$\mathbb{E}[\boldsymbol{x}] = \begin{bmatrix} \mathbb{E}[\boldsymbol{x}_1] \\ \mathbb{E}[\boldsymbol{x}_2] \end{bmatrix}, \qquad \mathbb{C}[\boldsymbol{x}] = \begin{bmatrix} \mathbb{C}[\boldsymbol{x}_1] & \mathbb{C}[\boldsymbol{x}_1, \boldsymbol{x}_2] \\ \mathbb{C}[\boldsymbol{x}_1, \boldsymbol{x}_2]^\top & \mathbb{C}[\boldsymbol{x}_2] \end{bmatrix}$$

- What is $p(\boldsymbol{x}_1)$? ($\boldsymbol{x}_2$ N_1-dimensional)
- Can we find a linear transformation $\boldsymbol{A}$, such that $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{x}_1$?



- Attention: **Dimensionality** of the distribution is now N_1, instead of N!
  If we used formula for N-dimensional normal distribution, then the distribution is **not normalized**!

Marginal distribution of a Gaussian

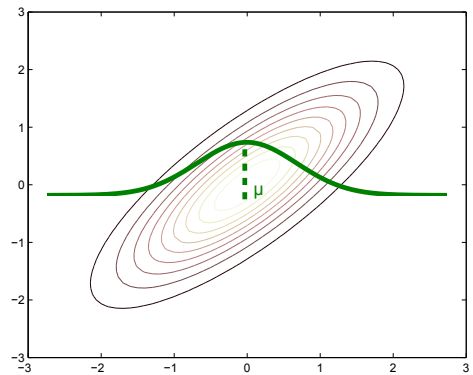
Multivariate case

Given a $(N_1 + N_2)$-dimensional Gaussian distributed random variable $\boldsymbol{x} = \left[\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top\right]^\top$, with mean and covariance

$$\mathbb{E}[\boldsymbol{x}] = \begin{bmatrix} \mathbb{E}[\boldsymbol{x}_1] \\ \mathbb{E}[\boldsymbol{x}_2] \end{bmatrix}, \qquad \mathbb{C}[\boldsymbol{x}] = \begin{bmatrix} \mathbb{C}[\boldsymbol{x}_1] & \mathbb{C}[\boldsymbol{x}_1, \boldsymbol{x}_2] \\ \mathbb{C}[\boldsymbol{x}_1, \boldsymbol{x}_2]^\top & \mathbb{C}[\boldsymbol{x}_2] \end{bmatrix}$$

- What is $p(\boldsymbol{x}_1)$? ($\boldsymbol{x}_2$ N_1-dimensional)
- Can we find a linear transformation $\boldsymbol{A}$, such that $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{x}_1$?

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{I}_{N_1} & \boldsymbol{0} \end{bmatrix}$$



- Attention: **Dimensionality** of the distribution is now N_1, instead of N!
  If we used formula for N-dimensional normal distribution, then the distribution is **not normalized**!

Marginal distribution of a Gaussian

Multivariate case

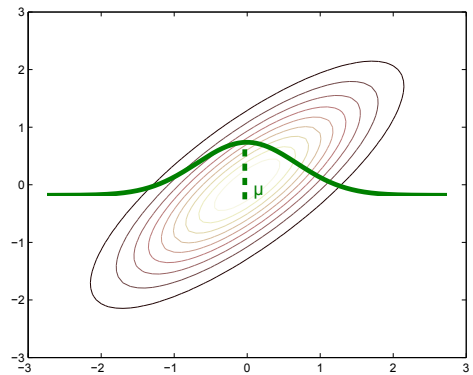
Given a $(N_1 + N_2)$-dimensional Gaussian distributed random variable $\boldsymbol{x} = \left[\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top\right]^\top$, with mean and covariance

$$\mathbb{E}[\boldsymbol{x}] = \begin{bmatrix} \mathbb{E}[\boldsymbol{x}_1] \\ \mathbb{E}[\boldsymbol{x}_2] \end{bmatrix}, \qquad \mathbb{C}[\boldsymbol{x}] = \begin{bmatrix} \mathbb{C}[\boldsymbol{x}_1] & \mathbb{C}[\boldsymbol{x}_1, \boldsymbol{x}_2] \\ \mathbb{C}[\boldsymbol{x}_1, \boldsymbol{x}_2]^\top & \mathbb{C}[\boldsymbol{x}_2] \end{bmatrix}$$

- What is $p(\boldsymbol{x}_1)$? ($\boldsymbol{x}_2$ N_1-dimensional)
- Can we find a linear transformation $\boldsymbol{A}$, such that $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{x}_1$?

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{I}_{N_1} & \boldsymbol{0} \end{bmatrix}$$

$$\begin{aligned} p(\boldsymbol{A}x) =& \mathcal{N}\left(\boldsymbol{A}\boldsymbol{x} \mid \boldsymbol{A}\mathbb{E}[\boldsymbol{x}], \boldsymbol{A}\mathbb{C}[\boldsymbol{x}]\boldsymbol{A}^\top\right) \\ =& \mathcal{N}\left(\boldsymbol{x}_1 \mid \mathbb{E}[\boldsymbol{x}_1], \mathbb{C}[\boldsymbol{x}_1]\right) \end{aligned}$$



- Attention: **Dimensionality** of the distribution is now N_1, instead of N!
  If we used formula for N-dimensional normal distribution, then the distribution is **not normalized**!

Marginal distribution of a Gaussian

Multivariate case

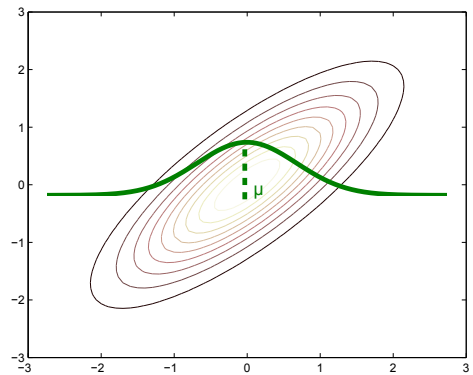
Given a $(N_1 + N_2)$-dimensional Gaussian distributed random variable $\boldsymbol{x} = \left[\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top\right]^\top$, with mean and covariance

$$\mathbb{E}[\boldsymbol{x}] = \begin{bmatrix} \mathbb{E}[\boldsymbol{x}_1] \\ \mathbb{E}[\boldsymbol{x}_2] \end{bmatrix}, \qquad \mathbb{C}[\boldsymbol{x}] = \begin{bmatrix} \mathbb{C}[\boldsymbol{x}_1] & \mathbb{C}[\boldsymbol{x}_1, \boldsymbol{x}_2] \\ \mathbb{C}[\boldsymbol{x}_1, \boldsymbol{x}_2]^\top & \mathbb{C}[\boldsymbol{x}_2] \end{bmatrix}$$

- What is $p(\boldsymbol{x}_1)$? ($\boldsymbol{x}_2$ N_1-dimensional)
- Can we find a linear transformation $\boldsymbol{A}$, such that $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{x}_1$?

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{I}_{N_1} & \boldsymbol{0} \end{bmatrix}$$

$$\begin{aligned} p(\boldsymbol{A}x) =& \mathcal{N}\left(\boldsymbol{A}\boldsymbol{x} \mid \boldsymbol{A}\mathbb{E}[\boldsymbol{x}], \boldsymbol{A}\mathbb{C}[\boldsymbol{x}]\boldsymbol{A}^\top\right) \\ =& \mathcal{N}\left(\boldsymbol{x}_1 \mid \mathbb{E}[\boldsymbol{x}_1], \mathbb{C}[\boldsymbol{x}_1]\right) \end{aligned}$$



- Attention: **Dimensionality** of the distribution is now N_1, instead of N!
  If we used formula for N-dimensional normal distribution, then the distribution is **not normalized**!

Sum of Gaussian distributed variables

Given a N-dimensional Gaussian distributed random variable $\boldsymbol{x}$, with mean $\mathbb{E}[\boldsymbol{x}]$ and covariance $\mathbb{C}[\boldsymbol{x}]$

- What is $p\left(\sum_{n=1}^{N} x_n\right)$?
- Can we find a linear transformation $\boldsymbol{A}$, such that $\boldsymbol{A}\boldsymbol{x} = \sum_{n=1}^{N} x_n$?
- Attention: **Dimensionality** of the distribution is now 1, instead of N!
  If we used formula for N-dimensional normal distribution, then the distribution is **not normalized**!

Sum of Gaussian distributed variables

Given a N-dimensional Gaussian distributed random variable $\boldsymbol{x}$, with mean $\mathbb{E}[\boldsymbol{x}]$ and covariance $\mathbb{C}[\boldsymbol{x}]$

- What is $p\left(\sum_{n=1}^{N} x_n\right)$?
- Can we find a linear transformation $\boldsymbol{A}$, such that $\boldsymbol{A}\boldsymbol{x} = \sum_{n=1}^{N} x_n$?
- Attention: **Dimensionality** of the distribution is now 1, instead of N!
  If we used formula for N-dimensional normal distribution, then the distribution is **not normalized**!

Sum of Gaussian distributed variables

Given a N-dimensional Gaussian distributed random variable $\boldsymbol{x}$, with mean $\mathbb{E}[\boldsymbol{x}]$ and covariance $\mathbb{C}[\boldsymbol{x}]$

- What is $p\left(\sum_{n=1}^{N} x_n\right)$?
- Can we find a linear transformation $\boldsymbol{A}$, such that $\boldsymbol{A}\boldsymbol{x} = \sum_{n=1}^{N} x_n$?

- Attention: **Dimensionality** of the distribution is now 1, instead of N!
  If we used formula for N-dimensional normal distribution, then the distribution is **not normalized**!

Sum of Gaussian distributed variables

Given a N-dimensional Gaussian distributed random variable $\boldsymbol{x}$, with mean $\mathbb{E}[\boldsymbol{x}]$ and covariance $\mathbb{C}[\boldsymbol{x}]$

- What is $p\left(\sum_{n=1}^{N} x_n\right)$?
- Can we find a linear transformation $\boldsymbol{A}$, such that $\boldsymbol{A}\boldsymbol{x} = \sum_{n=1}^{N} x_n$?

$$\boldsymbol{A} = \mathbf{1}_N = [1, \ldots, 1]$$

- Attention: **Dimensionality** of the distribution is now 1, instead of N!
  If we used formula for N-dimensional normal distribution, then the distribution is **not normalized**!

Sum of Gaussian distributed variables

Given a N-dimensional Gaussian distributed random variable $\boldsymbol{x}$, with mean $\mathbb{E}[\boldsymbol{x}]$ and covariance $\mathbb{C}[\boldsymbol{x}]$

- What is $p\left(\sum_{n=1}^{N} x_n\right)$?
- Can we find a linear transformation $\boldsymbol{A}$, such that $\boldsymbol{A}\boldsymbol{x} = \sum_{n=1}^{N} x_n$?

$$\boldsymbol{A} = \mathbf{1}_N = [1, \ldots, 1]$$

$$p(\boldsymbol{A}x) = \mathcal{N}\left(\mathbf{1}\boldsymbol{x} \mid \mathbf{1}_N\mathbb{E}[\boldsymbol{x}]\,,\, \mathbf{1}_N\mathbb{C}[\boldsymbol{x}]\mathbf{1}_N^\top\right)$$

$$= \mathcal{N}\left(\sum_{n=1}^{N} x_n \mid \sum_{n=1}^{N}\mathbb{E}[x_n]\,,\, \sum_{n=1}^{N}\sum_{m=1}^{N}\mathbb{C}[x_n\,,\,x_m]\right)$$

- Attention: **Dimensionality** of the distribution is now 1, instead of N!
  If we used formula for N-dimensional normal distribution, then the distribution is **not normalized**!

Sum of Gaussian distributed variables

Given a N-dimensional Gaussian distributed random variable $\boldsymbol{x}$, with mean $\mathbb{E}[\boldsymbol{x}]$ and covariance $\mathbb{C}[\boldsymbol{x}]$

- What is $p\left(\sum_{n=1}^{N} x_n\right)$?
- Can we find a linear transformation $\boldsymbol{A}$, such that $\boldsymbol{A}\boldsymbol{x} = \sum_{n=1}^{N} x_n$?

$$\boldsymbol{A} = \mathbf{1}_N = [1, \ldots, 1]$$

$$p(\boldsymbol{A}x) = \mathcal{N}\left(\mathbf{1}\boldsymbol{x} \mid \mathbf{1}_N\mathbb{E}[\boldsymbol{x}]\,,\, \mathbf{1}_N\mathbb{C}[\boldsymbol{x}]\mathbf{1}_N^\top\right)$$

$$= \mathcal{N}\left(\sum_{n=1}^{N} x_n \mid \sum_{n=1}^{N}\mathbb{E}[x_n]\,,\, \sum_{n=1}^{N}\sum_{m=1}^{N}\mathbb{C}[x_n\,,\,x_m]\right)$$

- Attention: **Dimensionality** of the distribution is now 1, instead of N!
  If we used formula for N-dimensional normal distribution, then the distribution is **not normalized**!

Whitening in a Gaussian distribution

Univariate case

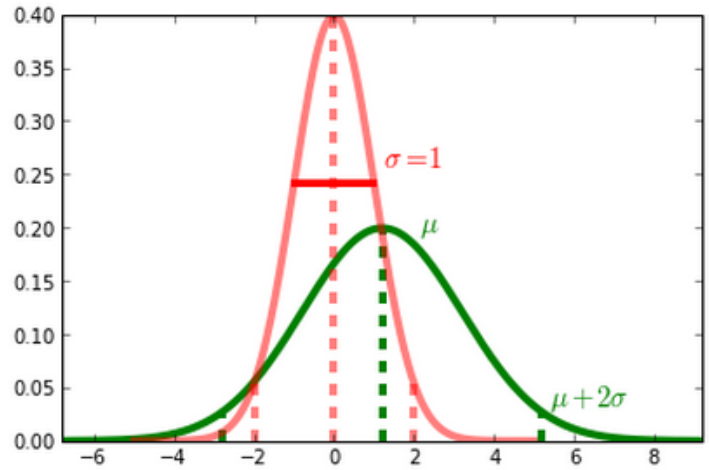
- Given a univariate Gaussian distribution with arbitrary mean μ and variance σ^2

$$x \sim \mathcal{N}\left(\mu\,,\,\sigma^2\right)$$

- Can I find a linear transformation $y = ax + b$, such that the resulting distribution has unit variance and zero mean?

$$y \sim \mathcal{N}\left(0\,,\,1\right) \Leftrightarrow$$

- Useful: Many statistical software packages only consider implement statistical tests standard normal distribution $\mathcal{N}(0\,,\,1)$.



Whitening in a Gaussian distribution

Univariate case

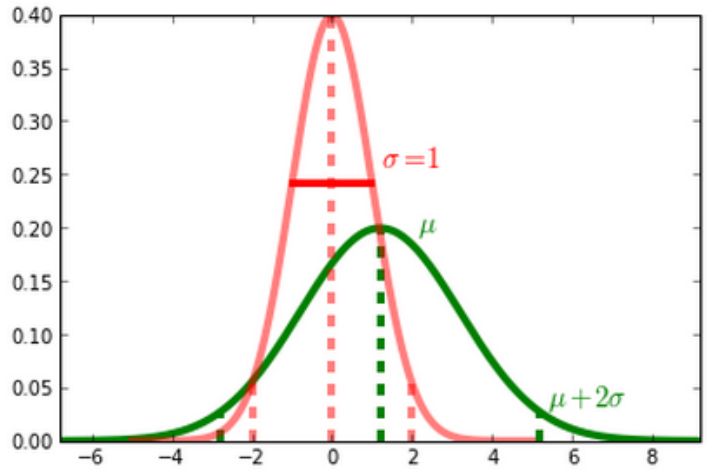
- Given a univariate Gaussian distribution with arbitrary mean μ and variance σ^2

$$x \sim \mathcal{N}\left(\mu\,,\,\sigma^2\right)$$

- Can I find a linear transformation $y = ax + b$, such that the resulting distribution has unit variance and zero mean?

$$y \sim \mathcal{N}\left(0\,,\,1\right) \;\Leftrightarrow$$

- Useful: Many statistical software packages only consider implement statistical tests standard normal distribution $\mathcal{N}(0\,,\,1)$.



Whitening in a Gaussian distribution

Univariate case

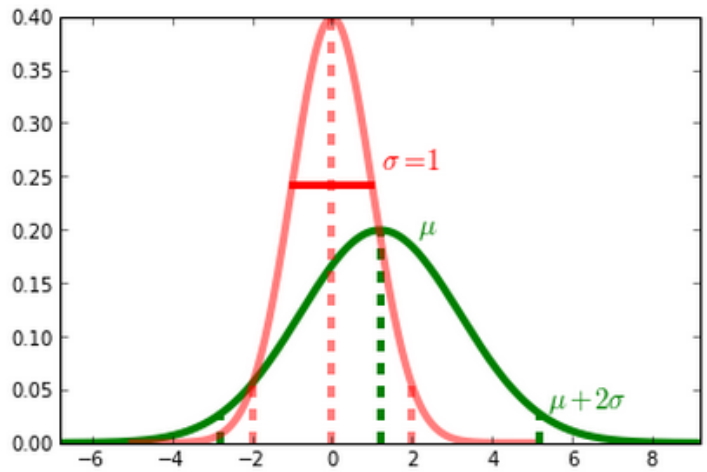
- Given a univariate Gaussian distribution with arbitrary mean μ and variance σ^2

$$x \sim \mathcal{N}\left(\mu\,,\,\sigma^2\right)$$

- Can I find a linear transformation $y = ax + b$, such that the resulting distribution has unit variance and zero mean?

$$y \sim \mathcal{N}\left(0\,,\,1\right) \;\Leftrightarrow\; y = \frac{x-\mu}{\sigma} \;\Leftrightarrow$$

- Useful: Many statistical software packages only consider implement statistical tests standard normal distribution $\mathcal{N}(0\,,\,1)$.



Whitening in a Gaussian distribution

Univariate case

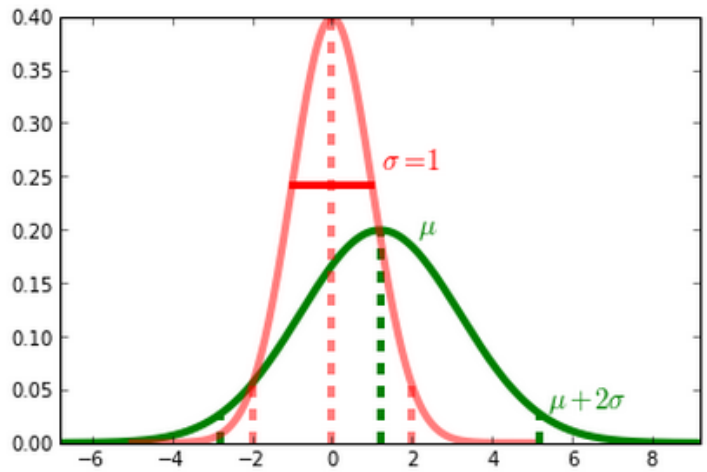
- Given a univariate Gaussian distribution with arbitrary mean μ and variance σ^2

$$x \sim \mathcal{N}\left(\mu\,,\,\sigma^2\right)$$

- Can I find a linear transformation $y = ax + b$, such that the resulting distribution has unit variance and zero mean?

$$y \sim \mathcal{N}\left(0\,,\,1\right) \;\Leftrightarrow\; y = \frac{x-\mu}{\sigma} \;\Leftrightarrow\; a = \frac{1}{\sigma}, \quad b = \frac{\mu}{\sigma}.$$

- Useful: Many statistical software packages only consider implement statistical tests standard normal distribution $\mathcal{N}(0\,,\,1)$.



Whitening in a Gaussian distribution

Univariate case

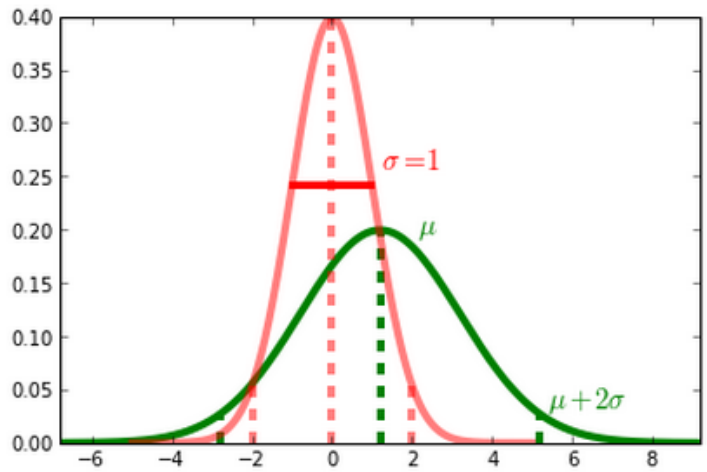
- Given a univariate Gaussian distribution with arbitrary mean μ and variance σ^2

$$x \sim \mathcal{N}\left(\mu\,,\,\sigma^2\right)$$

- Can I find a linear transformation $y = ax + b$, such that the resulting distribution has unit variance and zero mean?

$$y \sim \mathcal{N}\left(0\,,\,1\right) \;\Leftrightarrow\; y = \frac{x-\mu}{\sigma} \;\Leftrightarrow\; a = \frac{1}{\sigma}, \quad b = \frac{\mu}{\sigma}.$$

- Useful: Many statistical software packages only consider implement statistical tests standard normal distribution $\mathcal{N}(0\,,\,1)$.



Whitening in a Gaussian distribution

Multivariate case

- Given a multivariate Gaussian distribution with arbitrary mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$

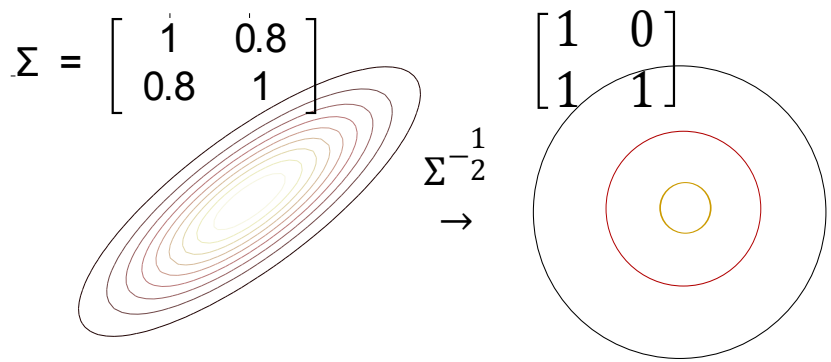
$$\mathcal{N}\left(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$$

- Can I find a linear transformation $f(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\mu}$, such that the resulting distribution has unit covariance $\boldsymbol{I}$ and zero mean?

- Matrix square-root of $\boldsymbol{\Sigma}^{-1}$:

$$\boldsymbol{A} = \boldsymbol{\Sigma}^{-\frac{1}{2}}, \quad \boldsymbol{b} = \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\mu}$$

$$\mathcal{N}\left(\boldsymbol{y} \mid \boldsymbol{0}, \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-\frac{\top}{2}}\right)$$



Whitening in a Gaussian distribution

Multivariate case

- Given a multivariate Gaussian distribution with arbitrary mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$

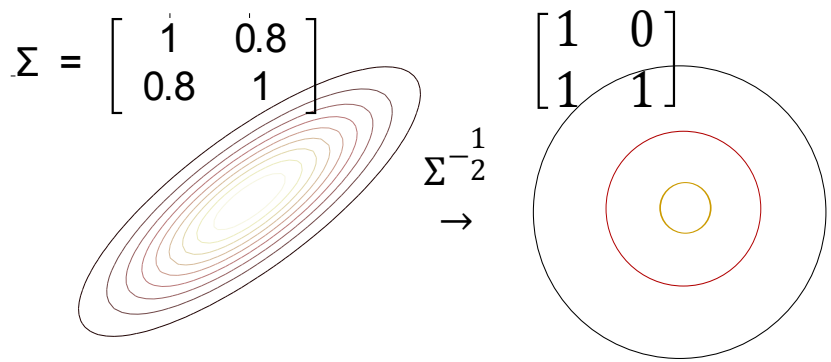
$$\mathcal{N}\left(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$$

- Can I find a linear transformation $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\mu}$, such that the resulting distribution has unit covariance $\boldsymbol{I}$ and zero mean?

- Matrix square-root of $\boldsymbol{\Sigma}^{-1}$:

$$\boldsymbol{A} = \boldsymbol{\Sigma}^{-\frac{1}{2}}, \quad \boldsymbol{b} = \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\mu}$$

$$\mathcal{N}\left(\boldsymbol{y} \mid \boldsymbol{0}, \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-\frac{\top}{2}}\right)$$



Whitening in a Gaussian distribution

Multivariate case

- Given a multivariate Gaussian distribution with arbitrary mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$

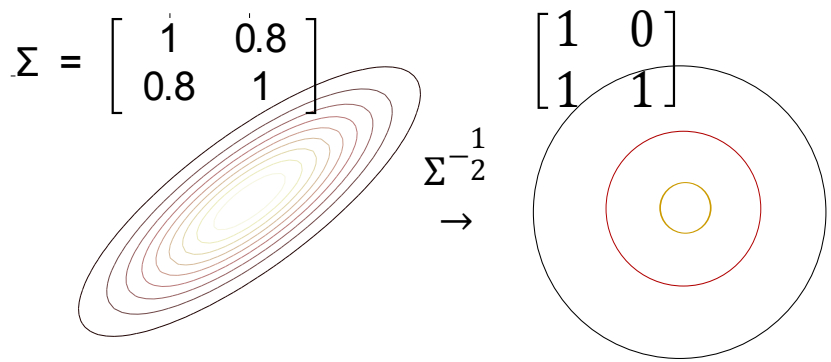
$$\mathcal{N}\left(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$$

- Can I find a linear transformation $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\mu}$, such that the resulting distribution has unit covariance $\boldsymbol{I}$ and zero mean?

- Matrix square-root of $\boldsymbol{\Sigma}^{-1}$:

$$\boldsymbol{A} = \boldsymbol{\Sigma}^{-\frac{1}{2}}, \quad \boldsymbol{b} = \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\mu}$$

$$\mathcal{N}\left(\boldsymbol{y} \mid \boldsymbol{0}, \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-\frac{\top}{2}}\right)$$



Matrix square root

A matrix $\boldsymbol{B}^{1/2}$ is called matrix square-root of a N-by-N symmetric matrix $\boldsymbol{B}$ iff

$$\boldsymbol{B}^{1/2}\boldsymbol{B}^{\top/2} = \boldsymbol{B}.$$

Examples of matrix square roots:

- **Cholesky factor** L(triangular matrix)
  - Cholesky factorization of B : $B = LL^{\top}$,
  - Computational complexity: $O(N^3)$.
  - The Cholesky factorization is only available for positive definite matrices.
- **Symmetric matrix square-root** $U\Lambda^{1/2}U^{\top}$ (unique),
  - U *orthogonal* matrix. The columns are the eigenvectors of B

$$UU^{\top} = U^{\top}U = I$$

  - $\Lambda = \mathrm{diag}[\lambda_1, \ldots, \lambda_N]$ diagonal matrix of eigenvalues λ_n.
  - Computational complexity: $O(N^3)$

$$B = U\Lambda^{1/2}\underbrace{U^{\top}U}_{I}\Lambda^{1/2}U^{\top}$$

- $\Rightarrow \Lambda^{1/2}U^{\top}$ is also a matrix square root! (non-symmetric)

Matrix square root

A matrix $\boldsymbol{B}^{1/2}$ is called matrix square-root of a N-by-N symmetric matrix $\boldsymbol{B}$ iff

$$\boldsymbol{B}^{1/2}\boldsymbol{B}^{\top/2} = \boldsymbol{B}.$$

Examples of matrix square roots:

- **Cholesky factor** $\boldsymbol{L}$(triangular matrix)
  - Cholesky factorization of $\boldsymbol{B}$: $\boldsymbol{B} = \boldsymbol{L}\boldsymbol{L}^{\top}$,
  - Computational complexity: $O(N^3)$.
  - The Cholesky factorization is only available for positive definite matrices.
- **Symmetric matrix square-root** $\boldsymbol{U}\boldsymbol{\Lambda}^{1/2}\boldsymbol{U}^{\top}$ (unique),
  - $\boldsymbol{U}$ *orthogonal* matrix. The columns are the eigenvectors of $\boldsymbol{B}$

$$\boldsymbol{U}\boldsymbol{U}^{\top} = \boldsymbol{U}^{\top}\boldsymbol{U} = \boldsymbol{I}$$

  - $\boldsymbol{\Lambda} = \mathrm{diag}[\lambda_1, \ldots, \lambda_N]$ diagonal matrix of eigenvalues λ_n.
  - Computational complexity: $O(N^3)$

$$\boldsymbol{B} = \boldsymbol{U}\boldsymbol{\Lambda}^{1/2}\underbrace{\boldsymbol{U}^{\top}\boldsymbol{U}}_{\boldsymbol{I}}\boldsymbol{\Lambda}^{1/2}\boldsymbol{U}^{\top}$$

  - $\Rightarrow \boldsymbol{\Lambda}^{1/2}\boldsymbol{U}^{\top}$ is also a matrix square root! (non-symmetric)

Matrix square root

A matrix $\boldsymbol{B}^{1/2}$ is called matrix square-root of a N-by-N symmetric matrix $\boldsymbol{B}$ iff

$$\boldsymbol{B}^{1/2}\boldsymbol{B}^{\top/2} = \boldsymbol{B}.$$

Examples of matrix square roots:

- **Cholesky factor** $\boldsymbol{L}$(triangular matrix)
  - Cholesky factorization of $\boldsymbol{B}$: $\boldsymbol{B} = \boldsymbol{L}\boldsymbol{L}^{\top}$,
  - Computational complexity: $O(N^3)$.
  - The Cholesky factorization is only available for positive definite matrices.
- **Symmetric matrix square-root** $\boldsymbol{U}\boldsymbol{\Lambda}^{1/2}\boldsymbol{U}^{\top}$ (unique),
  - $\boldsymbol{U}$ *orthogonal* matrix. The columns are the eigenvectors of $\boldsymbol{B}$

$$\boldsymbol{U}\boldsymbol{U}^{\top} = \boldsymbol{U}^{\top}\boldsymbol{U} = \boldsymbol{I}$$

  - $\boldsymbol{\Lambda} = \mathrm{diag}[\lambda_1, \ldots, \lambda_N]$ diagonal matrix of eigenvalues λ_n.
  - Computational complexity: $O(N^3)$

$$\boldsymbol{B} = \boldsymbol{U}\boldsymbol{\Lambda}^{1/2}\underbrace{\boldsymbol{U}^{\top}\boldsymbol{U}}_{\boldsymbol{I}}\boldsymbol{\Lambda}^{1/2}\boldsymbol{U}^{\top}$$

- $\Rightarrow \boldsymbol{\Lambda}^{1/2}\boldsymbol{U}^{\top}$ is also a matrix square root! (non-symmetric)

Matrix square root

A matrix $\boldsymbol{B}^{1/2}$ is called matrix square-root of a N-by-N symmetric matrix $\boldsymbol{B}$ iff

$$\boldsymbol{B}^{1/2}\boldsymbol{B}^{\top/2} = \boldsymbol{B}.$$

Examples of matrix square roots:

- **Cholesky factor** $\boldsymbol{L}$(triangular matrix)
  - Cholesky factorization of $\boldsymbol{B}$: $\boldsymbol{B} = \boldsymbol{L}\boldsymbol{L}^{\top}$,
  - Computational complexity: $O(N^3)$.
  - The Cholesky factorization is only available for positive definite matrices.
- **Symmetric matrix square-root** $\boldsymbol{U}\boldsymbol{\Lambda}^{1/2}\boldsymbol{U}^{\top}$ (unique),
  - $\boldsymbol{U}$ *orthogonal* matrix. The columns are the eigenvectors of $\boldsymbol{B}$

$$\boldsymbol{U}\boldsymbol{U}^{\top} = \boldsymbol{U}^{\top}\boldsymbol{U} = \boldsymbol{I}$$

  - $\boldsymbol{\Lambda} = \mathrm{diag}[\lambda_1, \ldots, \lambda_N]$ diagonal matrix of eigenvalues λ_n.
  - Computational complexity: $O(N^3)$

$$\boldsymbol{B} = \boldsymbol{U}\boldsymbol{\Lambda}^{1/2}\underbrace{\boldsymbol{U}^{\top}\boldsymbol{U}}_{\boldsymbol{I}}\boldsymbol{\Lambda}^{1/2}\boldsymbol{U}^{\top}$$

- $\Rightarrow \boldsymbol{\Lambda}^{1/2}\boldsymbol{U}^{\top}$ is also a matrix square root! (non-symmetric)

Matrix square root

Application: Sampling from a multivariate Gaussian distribution

- We often need samples from a multivariate Gaussian distribution (For example: MCMC sampling)
- Sampling from a univariate Gaussian distribution is easy.

$$x \sim \mathcal{N}(0\,,\,1)$$

- How can we use N samples from a univariate Gaussian to imitate a single sample from a multivariate Gaussian with covariance $\boldsymbol{\Sigma}$?

$$\boldsymbol{x} \sim \mathcal{N}\left(\boldsymbol{\mu}\,,\,\boldsymbol{\Sigma}\right)$$

- Use matrix square $\boldsymbol{\Sigma}^{1/2}$ root of the covariance matrix to transform the samples!
- **Algorithm:**
  1. Compute a square root of $\boldsymbol{\Sigma}$ (e.g. from Cholesky decomposition)
  2. Get N samples x_n from a standard normal distribution $\mathcal{N}(0\,,\,1)$
  3. Stack the samples into a vector $\boldsymbol{x} = [x_1, \ldots, x_N]$
  4. Multiply $\boldsymbol{x}$ by $\boldsymbol{\Sigma}^{1/2}$
  5. Add $\boldsymbol{\mu}$.
  6. done.

Matrix square root

Application: Sampling from a multivariate Gaussian distribution

- We often need samples from a multivariate Gaussian distribution (For example: MCMC sampling)
- Sampling from a univariate Gaussian distribution is easy.

$$x \sim \mathcal{N}(0\,,\,1)$$

- How can we use N samples from a univariate Gaussian to imitate a single sample from a multivariate Gaussian with covariance $\boldsymbol{\Sigma}$?

$$\boldsymbol{x} \sim \mathcal{N}\left(\boldsymbol{\mu}\,,\,\boldsymbol{\Sigma}\right)$$

- Use matrix square $\boldsymbol{\Sigma}^{1/2}$ root of the covariance matrix to transform the samples!
- **Algorithm:**
  1. Compute a square root of $\boldsymbol{\Sigma}$ (e.g. from Cholesky decomposition)
  2. Get N samples x_n from a standard normal distribution $\mathcal{N}(0\,,\,1)$
  3. Stack the samples into a vector $\boldsymbol{x} = [x_1, \ldots, x_N]$
  4. Multiply $\boldsymbol{x}$ by $\boldsymbol{\Sigma}^{1/2}$
  5. Add $\boldsymbol{\mu}$.
  6. done.

Matrix square root

Application: Sampling from a multivariate Gaussian distribution

- We often need samples from a multivariate Gaussian distribution (For example: MCMC sampling)
- Sampling from a univariate Gaussian distribution is easy.

$$x \sim \mathcal{N}(0\,,\,1)$$

- How can we use N samples from a univariate Gaussian to imitate a single sample from a multivariate Gaussian with covariance $\boldsymbol{\Sigma}$?

$$\boldsymbol{x} \sim \mathcal{N}\left(\boldsymbol{\mu}\,,\,\boldsymbol{\Sigma}\right)$$

- Use matrix square $\boldsymbol{\Sigma}^{1/2}$ root of the covariance matrix to transform the samples!
- **Algorithm:**
  1. Compute a square root of $\boldsymbol{\Sigma}$ (e.g. from Cholesky decomposition)
  2. Get N samples x_n from a standard normal distribution $\mathcal{N}(0\,,\,1)$
  3. Stack the samples into a vector $\boldsymbol{x} = [x_1, \ldots, x_N]$
  4. Multiply $\boldsymbol{x}$ by $\boldsymbol{\Sigma}^{1/2}$
  5. Add $\boldsymbol{\mu}$.
  6. done.

Matrix square root

Application: Sampling from a multivariate Gaussian distribution

- We often need samples from a multivariate Gaussian distribution (For example: MCMC sampling)
- Sampling from a univariate Gaussian distribution is easy.

$$x \sim \mathcal{N}(0\,,\,1)$$

- How can we use N samples from a univariate Gaussian to imitate a single sample from a multivariate Gaussian with covariance $\boldsymbol{\Sigma}$?

$$\boldsymbol{x} \sim \mathcal{N}\left(\boldsymbol{\mu}\,,\,\boldsymbol{\Sigma}\right)$$

- Use matrix square $\boldsymbol{\Sigma}^{1/2}$ root of the covariance matrix to transform the samples!
- **Algorithm:**
  1. Compute a square root of Σ (e.g. from Cholesky decomposition)
  2. Get N samples x_n from a standard normal distribution $\mathcal{N}(0\,,\,1)$
  3. Stack the samples into a vector $x = [x_1, \ldots, x_N]$
  4. Multiply x by $\boldsymbol{\Sigma}^{1/2}$
  5. Add $\boldsymbol{\mu}$.
  6. done.

Matrix square root

Application: Sampling from a multivariate Gaussian distribution

- We often need samples from a multivariate Gaussian distribution (For example: MCMC sampling)
- Sampling from a univariate Gaussian distribution is easy.

$$x \sim \mathcal{N}(0\,,\,1)$$

- How can we use N samples from a univariate Gaussian to imitate a single sample from a multivariate Gaussian with covariance $\boldsymbol{\Sigma}$?

$$\boldsymbol{x} \sim \mathcal{N}\left(\boldsymbol{\mu}\,,\,\boldsymbol{\Sigma}\right)$$

- Use matrix square $\boldsymbol{\Sigma}^{1/2}$ root of the covariance matrix to transform the samples!
- **Algorithm:**
  1. Compute a square root of $\boldsymbol{\Sigma}$ (e.g. from Cholesky decomposition)
  2. Get N samples x_n from a standard normal distribution $\mathcal{N}(0\,,\,1)$
  3. Stack the samples into a vector $\boldsymbol{x} = [x_1, \ldots, x_N]$
  4. Multiply $\boldsymbol{x}$ by $\boldsymbol{\Sigma}^{1/2}$
  5. Add $\boldsymbol{\mu}$.
  6. done.