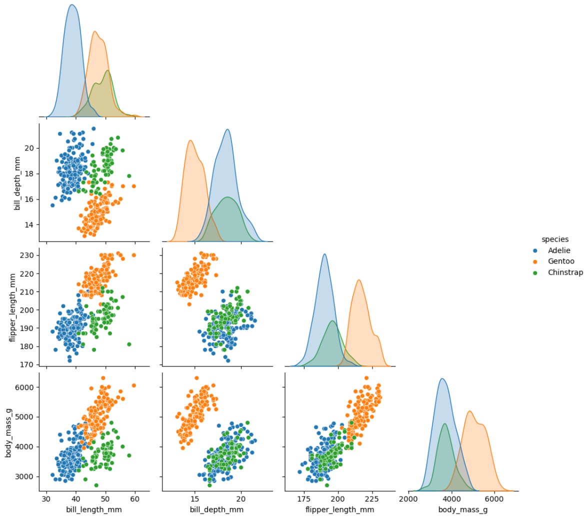
Foundations of Data Science

Lecture 4: Latent variable models

Prof. Gilles Louppe
g.louppe@uliege.be



In Lecture 2, our exploratory data analysis revealed that penguins are clustered by species, with distinctive physical traits.

What if we had not been given the species labels? What underlying factors might explain the observed variations in physical traits?

Probabilistic modeling of data

Data are recorded observations about the world. Mathematically, we can think of data as resulting from a function f that maps real-world entities ω to measurements x,

$$f : \Omega \to \mathcal{X},$$

where

- Ω is the sample space (the set of all possible entities, accounting for all sources of variability),
- $\mathcal{X}$ is the measurement space.

Entities $\omega \in \Omega$ are not observable, only their measurements $x = f(\omega) \in \mathcal{X}$ are.

If the sample space Ω is equipped with a probability function p, then the data $x = f(\omega)$ can be viewed as a random variable with distribution induced by p,

$$x \sim p_r(x) = \int_{\omega \in \Omega} p(\omega)\delta(x - f(\omega))d\omega,$$

where δ is the Dirac delta function.

We call $p_r(x)$ the **data generating process** or the data distribution.

A **parametric probabilistic model** encodes assumptions about how data are generated. It is specified by a parametric family

$$\mathcal{P} = \{p(x \mid \theta) : \theta \in \boldsymbol{\Theta}\},$$

where $p(x \mid \theta)$ is a probability distribution over $\mathcal{X}$, θ are parameters, and $\boldsymbol{\Theta}$ is the parameter space.

$p(x \mid \theta)$ is not the data distribution $p_r(x)$, only a model of it! There is no such thing as a *true* parameter θ.

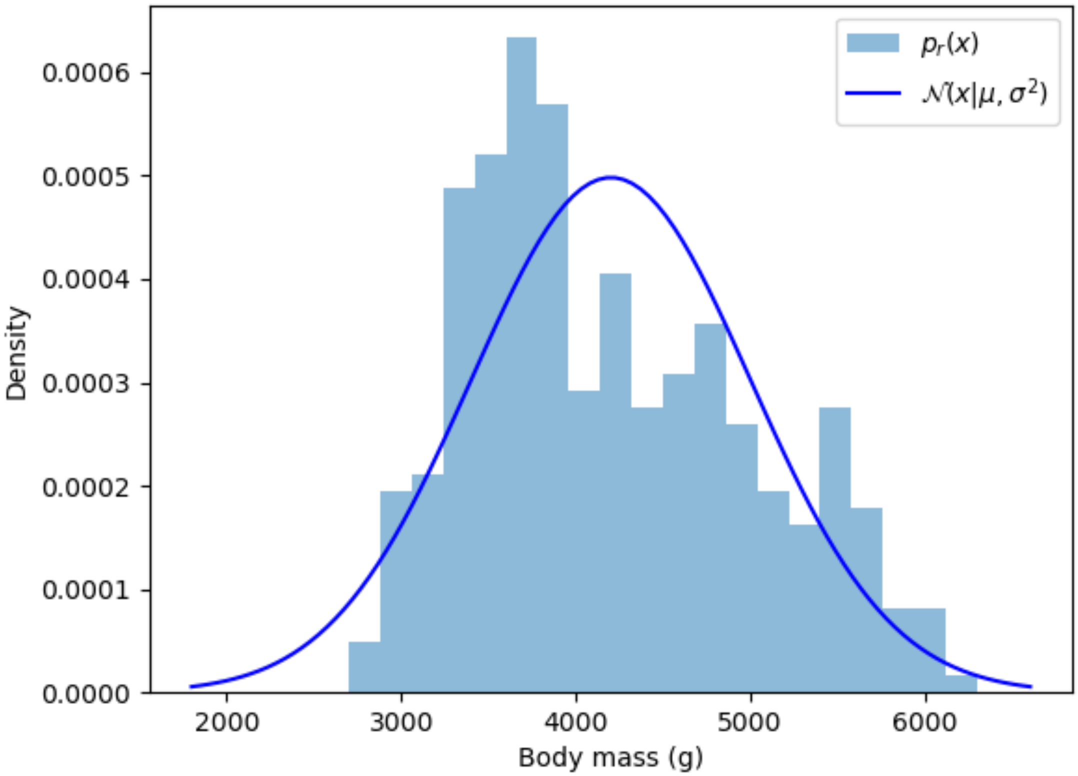
Example: Body masses of penguins could be modeled as

$$p(x \mid \mu, \sigma^2) = \mathcal{N}(x \mid \mu, \sigma^2),$$

where the parameters are $\theta = (\mu, \sigma^2)$, with $\mu \in \mathbb{R}$ and $\sigma^2 > 0$.

This would assume

- body masses cluster around a central value μ,
- variability is symmetric and controlled by σ^2,
- extreme values are rare (body masses are normally distributed).


$p_r(x)$
$\mathcal{N}(x|\mu, \sigma^2)$
Density
0.0000
0.0001
0.0002
0.0003
0.0004
0.0005
0.0006
2000
3000
4000
5000
6000
Body mass (g)

Frequentist inference

In the Frequentist framework, θ is treated as an unknown but fixed quantity to be estimated from observed data x_{obs}. The data are assumed to be generated from the model for some unknown parameter value θ^*,

$$x_{\text{obs}} \sim p(x \mid \theta^*).$$

Fitting the model to data consists in finding a point estimate $\hat{\theta}$ of θ^* (or a confidence region thereof) that best explains the observed data.

Bayesian inference

In the Bayesian framework, θ is treated as a random variable with prior distribution $p(\theta)$ encoding beliefs about plausible parameter values before observing any data.

A **Bayesian model** therefore specifies a joint distribution over data and parameters,

$$p(x, \theta) = p(x \mid \theta)p(\theta),$$

where

- $p(x \mid \theta)$ is the likelihood,
- $p(\theta)$ is the prior over parameters.

Fitting a Bayesian model to observed data x_{obs} consists in computing the posterior distribution of the parameters given the data. Using Bayes' rule,

$$p(\theta \mid x_{\text{obs}}) = \frac{p(x_{\text{obs}} \mid \theta)p(\theta)}{p(x_{\text{obs}})}.$$

Depending on the structure of the model, this computation may be easy, difficult, or even intractable.

Prior predictive checks

Often, the forward model $p(x|\theta)$ is understood, but the prior $p(\theta)$ is more subjective and harder to justify.

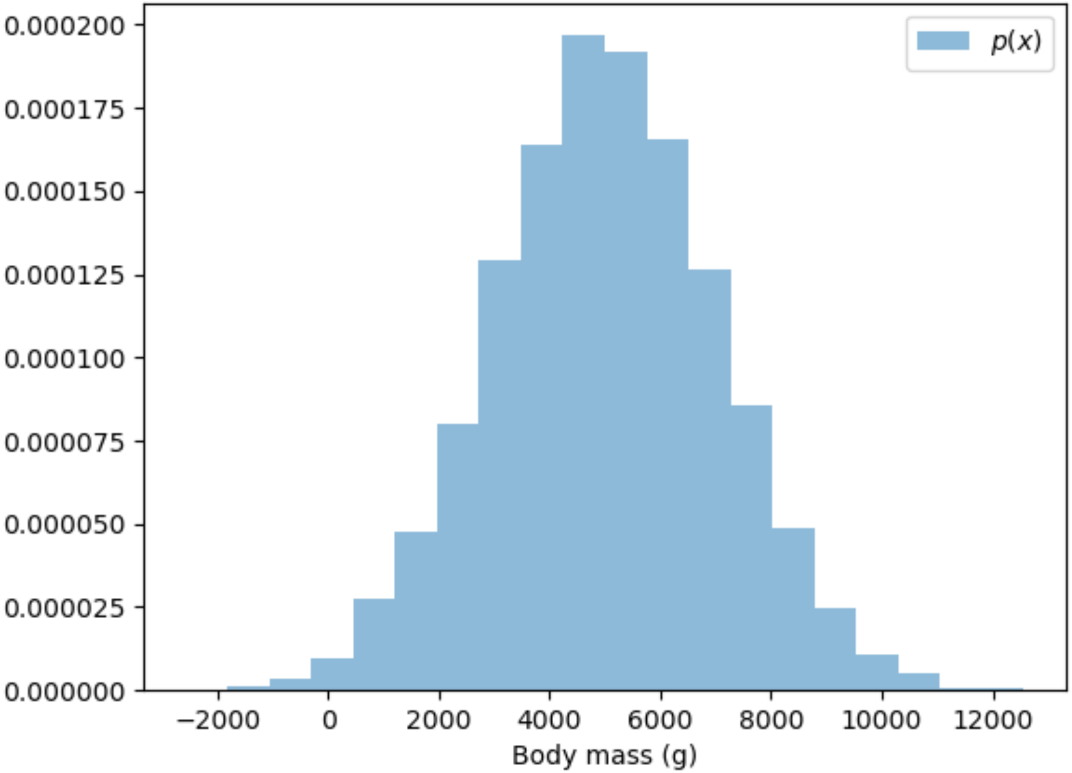
The consequences of prior choices in the context of the generative model can be assessed through **prior predictive checks**, which involve simulating data from the model using only the prior distributions, without conditioning on any observed data.

The prior predictive distribution is given by

$$p(x) = \int p(x|\theta)p(\theta)d\theta.$$

This distribution defines the data that we expect to observe under the model assumptions encoded in the prior. It should be examined to ensure that it aligns with domain knowledge and expectations about the data.

$$p(\mu, \sigma^2) = \mathcal{N}(\mu|5000, 2000^2) \times \text{Uniform}(\sigma^2|0, 100)$$
$$p(x|\mu, \sigma^2) = \mathcal{N}(x|\mu, \sigma^2)$$



Latent variable models

Joint distribution

A **latent variable model** is a probabilistic model that assumes unobserved (latent) variables z that mediate the relationship between observed data x and model parameters θ.

It specifies a joint distribution over observed variables, latent variables, and parameters,

$$p(x, z, \theta) = p(x \mid z, \theta) p(z \mid \theta) p(\theta),$$

where x is the observed data, z are the latent variables, and θ are the parameters.

More generally, for a dataset of N observations $\{x_1, \ldots, x_N\}$, a latent variable model specifies a joint distribution

$$p(x_{1:N}, z_{1:N}, \theta) = \left(\prod_{i=1}^{N} p(x_i \mid z_i, \theta) p(z_i \mid \theta) \right) p(\theta),$$

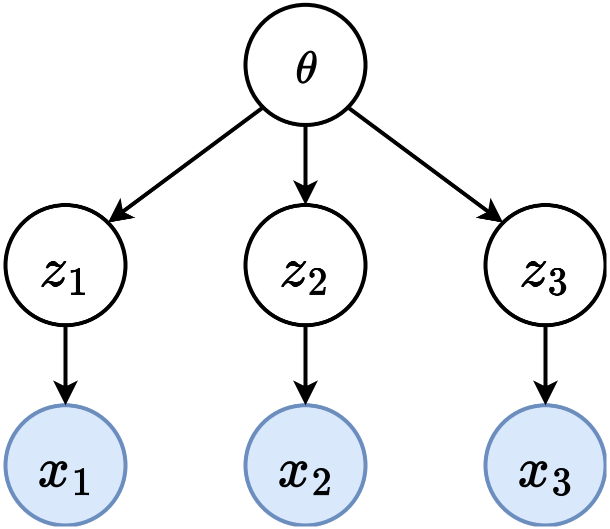
where z_i are the latent variables associated with observation x_i.

The factorization assumes that each observation x_i is mediated by its own latent variable z_i, and that observations are conditionally independent given their latent variables and the parameters. All are governed by shared parameters θ.

Graphical model representation

Latent variable models can be represented using graphical models, where nodes represent variables (observed, latent, or parameters) and edges represent (possible) dependencies between them.

The graphical model illustrates the structure of the factorization of the joint distribution and the flow of the generative process.



$$p(x_{1:3}, z_{1:3}, \theta) = \left(\prod_{n=1}^{3} p(x_n \mid z_n, \theta) p(z_n \mid \theta) \right) p(\theta)$$

Shaded nodes represent observed variables, unshaded nodes represent latent variables or parameters.

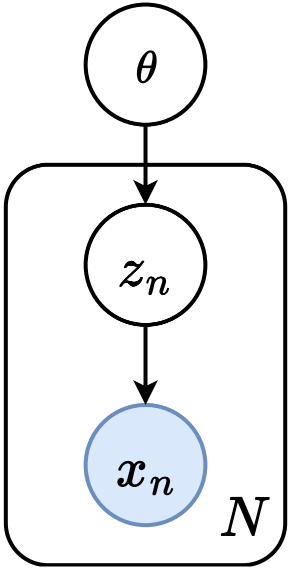


Plate notation can be used to compactly represent repeated structures in the graphical model.

Hyperparameters

Conditional distributions in a latent variable model may depend on additional parameters called **hyperparameters**, denoted $\alpha, \beta, \ldots$. These are assumed to be fixed nonrandom quantities[1].

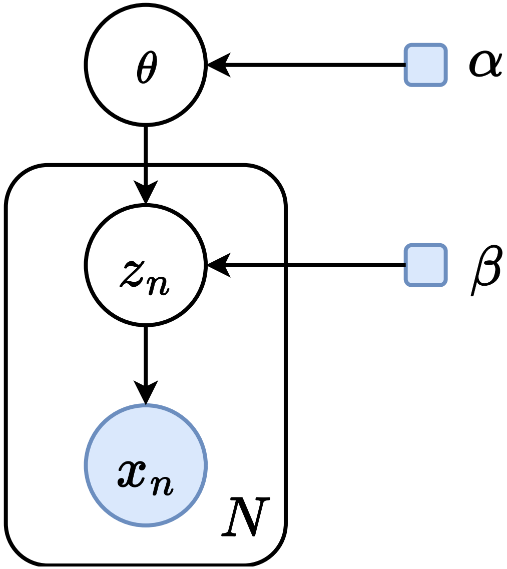
For instance, the prior distribution of parameters may depend on hyperparameters,

$$p(\theta \mid \alpha),$$

or the prior distribution of latent variables may depend on hyperparameters,

$$p(z \mid \theta, \beta).$$

[1] Estimating hyperparameters from data is possible and will be discussed later in the course.



Small squares denote fixed hyperparameters.

Inference

Fitting a latent variable model to observed data x_{obs} consists in computing the posterior distribution of latent variables and parameters given the data. Using Bayes' rule,

$$p(z, \theta \mid x_{\text{obs}}) = \frac{p(x_{\text{obs}} \mid z, \theta) p(z \mid \theta) p(\theta)}{p(x_{\text{obs}})}.$$

The posterior distribution is used to examine the particular hidden structure that is manifested in the observed data. It can also be used to make predictions about new, unseen data, through the posterior predictive distribution,

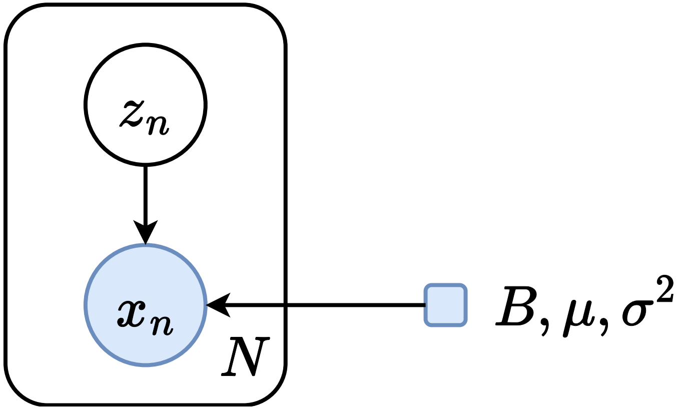
$$p(x_{\text{new}} \mid x_{\text{obs}}) = \iint p(x_{\text{new}} \mid z, \theta) p(z, \theta \mid x_{\text{obs}}) dz d\theta.$$

Example 1: (Probabilistic) PCA

In probabilistic PCA, each observation $\boldsymbol{x}_i \in \mathbb{R}^d$ is assumed to be generated from a lower-dimensional latent variable $z_i \in \mathbb{R}^m$ through a linear transformation plus Gaussian noise.

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

$$x_n \mid z_n \sim \mathcal{N}\left(Bz_n + \mu, \sigma^2\right)$$



The joint distribution $p(z, x|B, \mu, \sigma^2)$ factorizes as $p(z)p(x|z, B, \mu, \sigma^2)$, where

- $p(z) = \mathcal{N}(z|0, I)$ assumes latent variables are standard Gaussian,
- $p(x|z, B, \mu, \sigma^2) = \mathcal{N}(x|Bz + \mu, \sigma^2 I)$ assumes a linear Gaussian observation model, with $B \in \mathbb{R}^{d\times m}$ the loading matrix, $\mu \in \mathbb{R}^d$ the mean vector, and σ^2 the noise variance.

Therefore, using Gaussian identities, the joint distribution is Gaussian and can be written as

$$p(z, x|B, \mu, \sigma^2) = \mathcal{N}\left(\begin{bmatrix} z \\ x \end{bmatrix} \middle| \begin{bmatrix} 0 \\ \mu \end{bmatrix}, \begin{bmatrix} I & B^T \\ B & BB^T + \sigma^2 I \end{bmatrix}\right).$$

The posterior distribution $p(z|x, B, \mu, \sigma^2)$ is also Gaussian,

$$p(z|x, B, \mu, \sigma^2) = \mathcal{N}(z|m, C),$$

where

- $m = B^T(BB^T + \sigma^2 I)^{-1}(x - \mu)$ is the posterior mean,
- $C = I - B^T(BB^T + \sigma^2 I)^{-1}B$ is the posterior covariance.

When $\sigma^2 \to 0$,

- $m = B^T(BB^T + \sigma^2 I)^{-1}(x - \mu) \to B^T(BB^T)^{-1}(x - \mu)$. If the columns of B are orthonormal, then $B^T(BB^T)^{-1} = B^T$, so $m \to B^T(x - \mu)$, which corresponds to the PCA projection of x onto the subspace spanned by the columns of B.
- $C = I - B^T(BB^T + \sigma^2 I)^{-1}B \to I - B^T(BB^T)^{-1}B$. If the columns of B are orthonormal, then $B^T B = I$, so $C \to 0$, indicating that the posterior distribution collapses to a point mass at the PCA projection.

Probabilistic PCA recovers classical PCA in the limit of vanishing noise!

The hyperparameters B, μ, σ^2 can be estimated from data $x_{1:N}$ using maximum (marginal) likelihood estimation,

$$(\hat{B}, \hat{\mu}, \hat{\sigma}^2) = \arg \max_{B,\mu,\sigma^2} \prod_{i=1}^{N} p(x_i|B, \mu, \sigma^2),$$

where $p(x|B, \mu, \sigma^2) = \int p(x|z, B, \mu, \sigma^2)p(z)dz$ is the marginal likelihood.

Since the joint distribution is Gaussian, the marginal likelihood is also Gaussian,

$$p(x|B, \mu, \sigma^2) = \mathcal{N}(x|\mu, BB^T + \sigma^2 I).$$

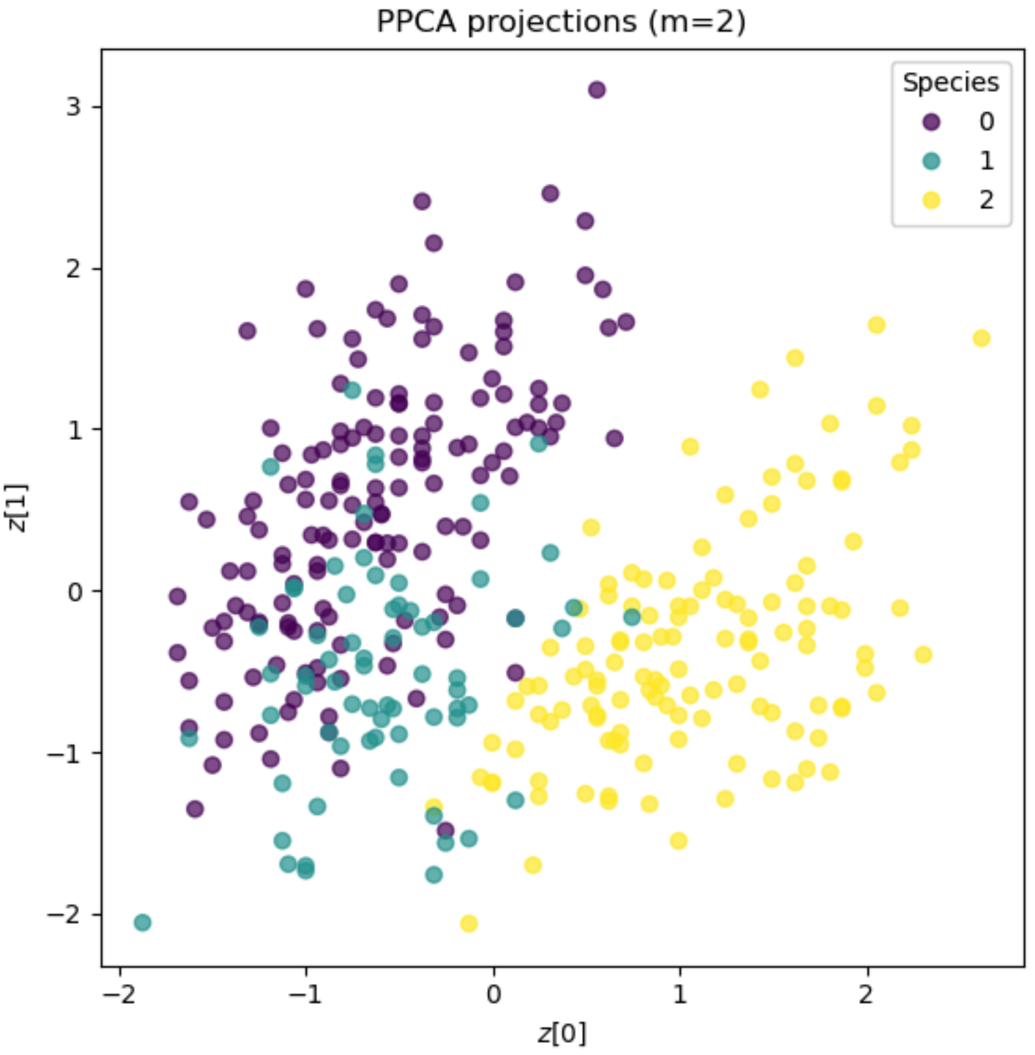
Therefore, writing $BB^T + \sigma^2 I = \Sigma$, maximum likelihood estimation reduces to

$$\begin{aligned}(\hat{B}, \hat{\mu}, \hat{\sigma}^2) &= \arg \max_{B,\mu,\sigma^2} \prod_{i=1}^{N} \mathcal{N}(x_i|\mu, \Sigma) \\ &= \arg \min_{B,\mu,\sigma^2} \sum_{i=1}^{N} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) + N \log |\Sigma| \\ &= \arg \min_{B,\mu,\sigma^2} \operatorname{tr}(\Sigma^{-1} S) + N \log |\Sigma|,\end{aligned}$$

where $S = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)(x_i - \mu)^T$ is the sample covariance matrix.

The solution can be derived in closed form, yielding

- $\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i$ (the sample mean),
- $\hat{B} = U_m(\Lambda_m - \hat{\sigma}^2 I)^{1/2} R$, where U_m are the top m eigenvectors of S, Λ_m are the corresponding eigenvalues, and R is an arbitrary rotation matrix,
- $\hat{\sigma}^2 = \frac{1}{d-m} \sum_{j=m+1}^{d} \lambda_j$, where λ_j are the eigenvalues of S.


PPCA projections (m=2)
Species
0
1
2
z[1]
3
2
1
0
−1
−2
z[0]
−2
−1
0
1
2

Deriving PCA from a latent variable model provides a **probabilistic interpretation of PCA projections as the most likely latent variables that could have generated the observed data**.

It also enables direct extensions such as

- Independent Component Analysis (ICA), which assumes non-Gaussian latent variables,
- Factor Analysis, which assumes a more general noise covariance structure,
- Bayesian PCA, which places a prior distribution over the hyperparameters.

Example 2: Mixture models

Mixture models assume that data are generated from a mixture of several underlying distributions, each corresponding to a different cluster or component.

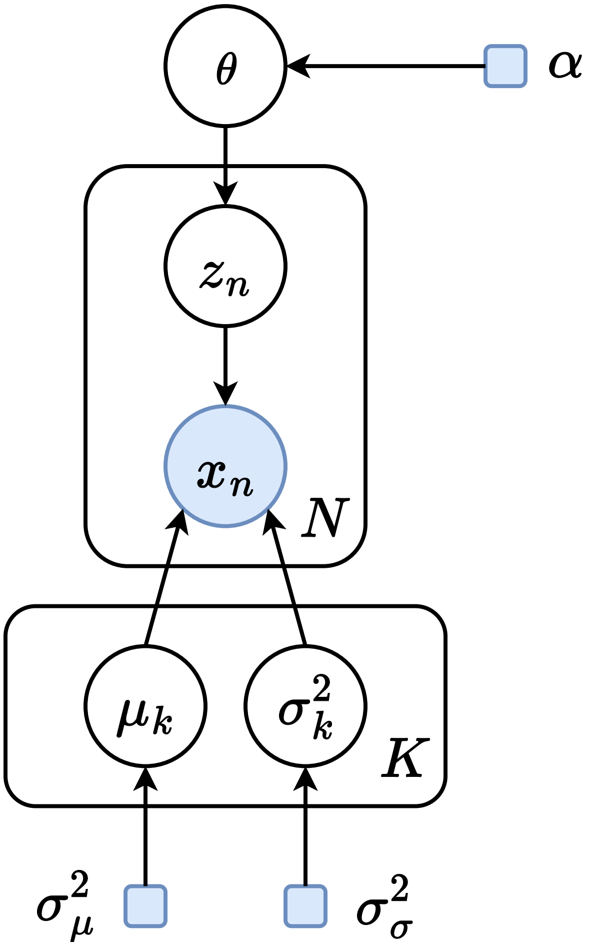
$$\theta \sim \mathrm{Dirichlet}(\alpha)$$

$$z_n \sim \mathrm{Categorical}(\theta)$$

$$x_n \sim \mathcal{N}\left(x_{z_n}, \sigma^2_{z_n}\right)$$

$$\mu_k \sim \mathcal{N}\left(0, \sigma^2_\mu\right)$$

$$\sigma^2_k \sim \mathrm{Lognormal}\left(0, \sigma^2_\sigma\right)$$



For a Gaussian mixture model with K components, each observation $x_i \in \mathbb{R}^d$ is assumed to be generated by first selecting a component $z_i \in \{1, \ldots, K\}$ according to a categorical distribution, then sampling x_i from a Gaussian distribution associated with that component.

The joint distribution $p(\theta, z_{1:N}, x_{1:N}, \mu_{1:K}, \sigma^2_{1:K} | \alpha, \sigma^2_\mu, \sigma^2_\sigma)$ factorizes as

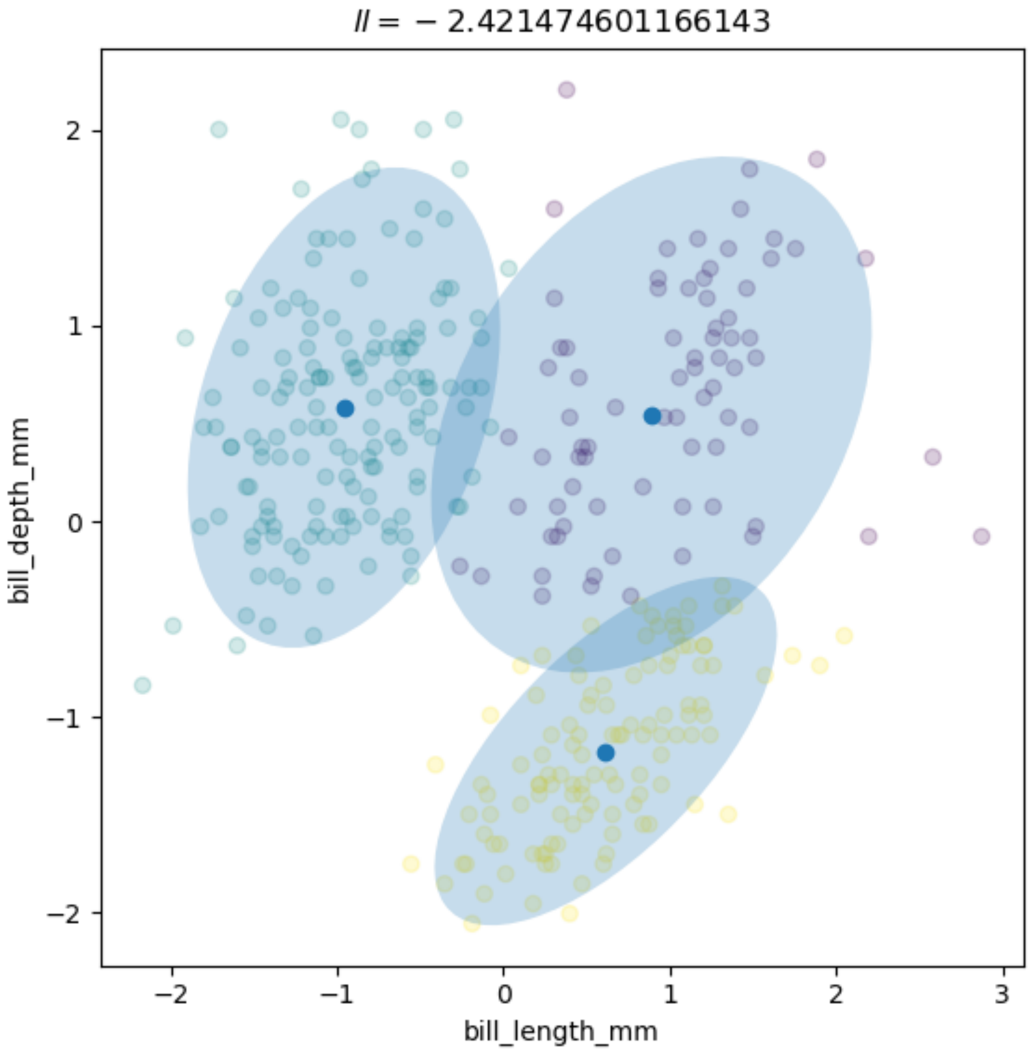
$$p(\theta|\alpha) \prod_{k=1}^{K} p(\mu_k|\sigma^2_\mu) p(\sigma^2_k|\sigma^2_\sigma) \prod_{i=1}^{N} p(z_i|\theta) p(x_i|z_i, \mu_{z_i}, \sigma^2_{z_i}),$$

where

- $p(\theta|\alpha) = \mathbf{Dirichlet}(\alpha)$ is the prior over mixture weights,
- $p(\mu_k|\sigma^2_\mu) = \mathcal{N}(0, \sigma^2_\mu I)$ is the prior over component means,
- $p(\sigma^2_k|\sigma^2_\sigma) = \mathbf{Lognormal}(0, \sigma^2_\sigma)$ is the prior over component variances,
- $p(z_i|\theta) = \mathbf{Categorical}(\theta)$ is the categorical distribution over components,
- $p(x_i|z_i, \mu_{z_i}, \sigma^2_{z_i}) = \mathcal{N}(\mu_{z_i}, \sigma^2_{z_i} I)$ is the Gaussian observation model.

Computing the posterior distribution $p(\theta, z_{1:N}, \mu_{1:K}, \sigma^2_{1:K} | x_{1:N}, \alpha, \sigma^2_\mu, \sigma^2_\sigma)$ amounts to solving a clustering problem, where each component corresponds to a cluster and the latent variables z_i indicate cluster membership of each observation.

The posterior is typically intractable, requiring approximate inference methods such as Expectation-Maximization (EM) or Variational Inference (VI).


$ll = -2.421474601166143$
bill_depth_mm
bill_length_mm
2
1
0
−1
−2
−2
−1
0
1
2
3

Example 3: Mixed membership models

Nested sets of latent variables can also be used to model more complex generative structures.

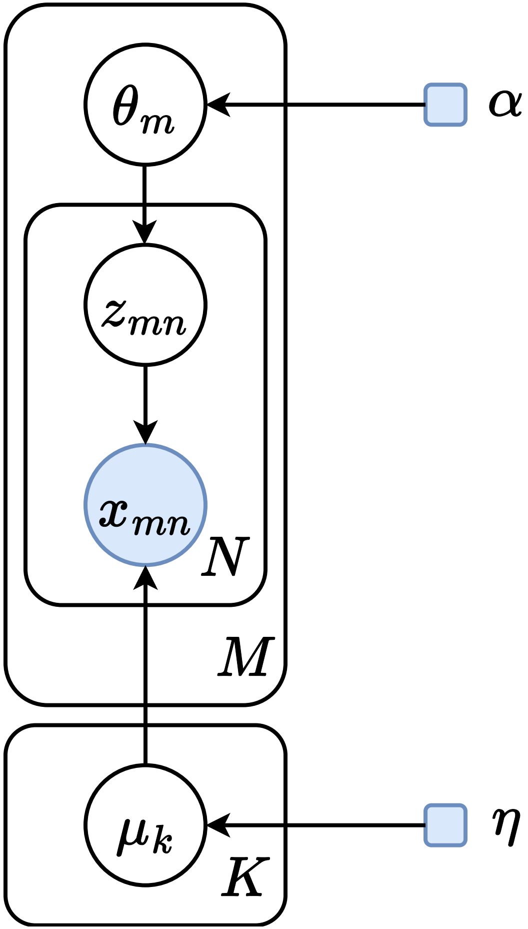
For instance, in mixed membership models of text documents (**latent dirichlet allocation**), each document is assumed to be generated from a mixture of topics, where each topic is characterized by a distribution over words.

$$\theta_m \sim \mathrm{Dirichlet}(\alpha)$$

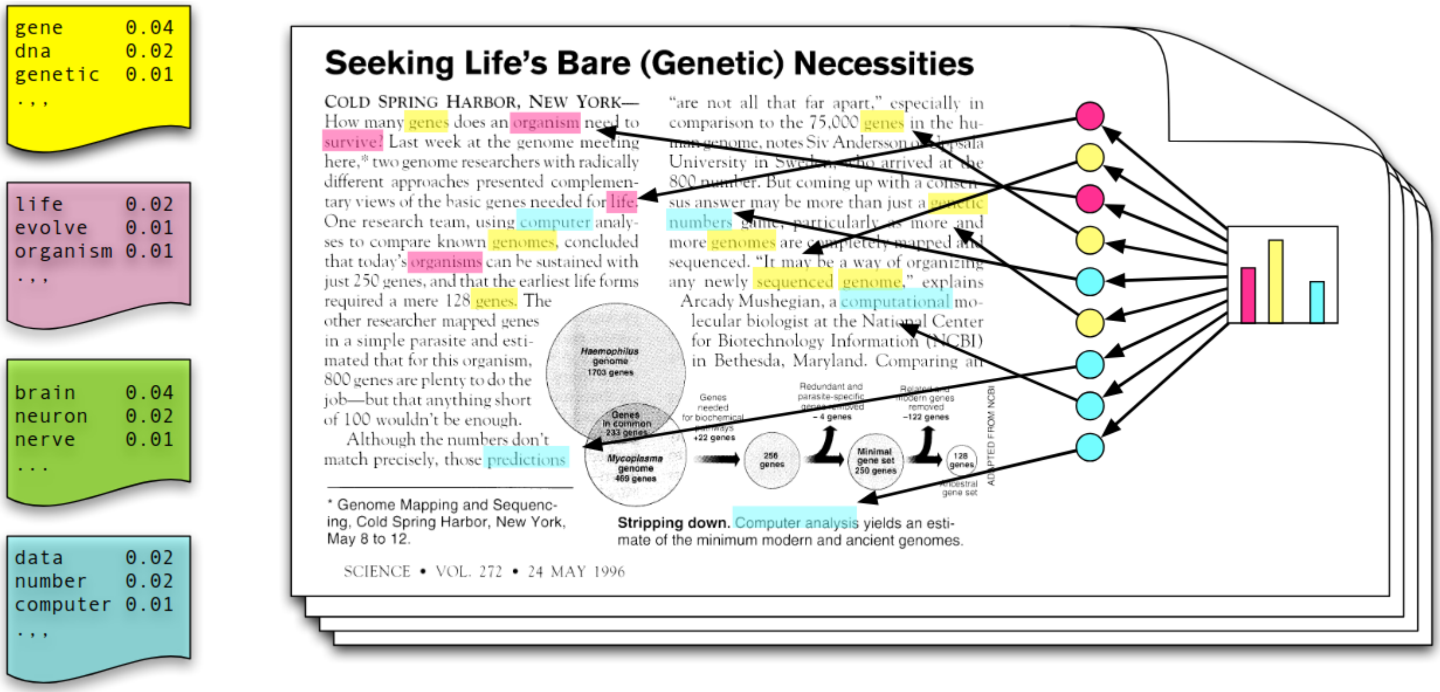
$$z_{mn} \sim \mathrm{Categorical}(\theta_m)$$

$$x_{mn} \sim p(. \mid \mu_{z_{mn}})$$

$$\mu_k \sim p(. \mid \eta)$$



Posterior inference in mixed membership models can be used to discover the underlying topics in a corpus of documents and to infer the topic proportions for each document.


gene 0.04
dna 0.02
genetic 0.01
. . .
life 0.02
evolve 0.01
organism 0.01
. . .
brain 0.04
neuron 0.02
nerve 0.01
. . .
data 0.02
number 0.02
computer 0.01
. . .
Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed – 4 genes
Minimal gene set 250 genes
Related and modern genes removed –122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

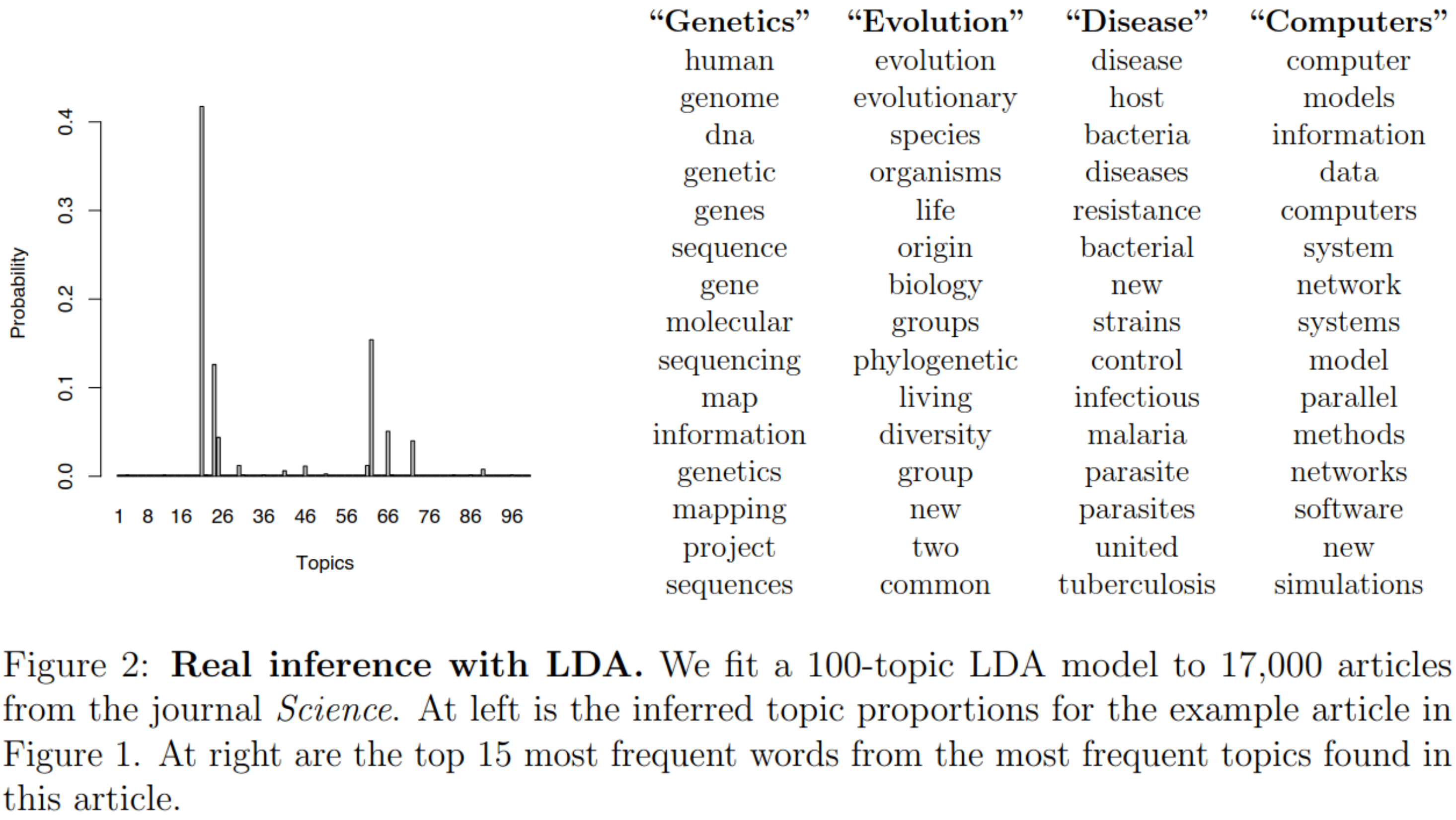
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

Figure 1: **The intuitions behind latent Dirichlet allocation.** We assume that some number of "topics," which are distributions over words, exist for the whole collection (far left). Each document is assumed to be generated as follows. First choose a distribution over the topics (the histogram at right); then, for each word, choose a topic assignment (the colored coins) and choose the word from the corresponding topic. The topics and topic assignments in this figure are illustrative—they are not fit from real data. See Figure 2 for topics fit from data.

—
Credits: Blei, 2011.



“Genetics”	**“Evolution”**	**“Disease”**	**“Computers”**
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Figure 2: **Real inference with LDA.** We fit a 100-topic LDA model to 17,000 articles from the journal *Science*. At left is the inferred topic proportions for the example article in Figure 1. At right are the top 15 most frequent words from the most frequent topics found in this article.

—

Credits: Blei, 2011.

The end.