UNIVERSIDAD
DE GRANADA

STATISTICAL MODELS WITH VARIATIONAL METHODS

LUIS ANTONIO ORTEGA ANDRÉS

Bachelor's Thesis

Computer Science and Mathematics

Tutor
Serafín Moral Callejón

FACULTY OF SCIENCE
H.T.S. OF COMPUTER ENGINEER AND TELECOMMUNICATIONS

Granada, Tuesday 30th June, 2020

ABSTRACT

Some introduction about how important Variational methods are nowadays and what this project is about.

CONTENTS

Part I

BASIC CONCEPTS

In this part we will introduce the underlying concepts of probability and graph theory that we will need.

1

PROBABILITY

All our theory will be made under the assumption that there is a *referential set* Ω, set of all possible outcomes of an experiment. Any subset of Ω will be called an *event*.

Definition 1. Let $\mathcal{P}(\Omega)$ be the power set of Ω. Then, $\mathcal{F} \subset \mathcal{P}(\Omega)$ is called a *σ-algebra* if it satisfies:

- $\Omega \in \mathcal{F}$.
- $\mathcal{F}$ is closed under complementation.
- $\mathcal{F}$ is closed under countable unions.

From these properties it follows that $\emptyset \in \mathcal{F}$ and that $\mathcal{F}$ is closed under countable intersections.

The tuple $(\Omega, \mathcal{F})$ is called a *measurable space*.

Definition 2. A *probability* P over $(\Omega, \mathcal{F})$ is a mapping $P : \mathcal{F} \to [0, 1]$ which satisfies

- $P(\alpha) \geq 0 \ \ \forall \alpha \in \mathcal{F}$.
- $P(\Omega) = 1$.
- P is countably additive, that is, if $\{\alpha_i\}_{i \in \mathbb{N}} \subset \mathcal{F}$, is a countable collection of pairwise disjoint sets, then

$$P\Big(\bigcup_{i \in \mathbb{N}} \alpha_i\Big) = \sum_{i \in \mathbb{N}} P(\alpha_i).$$

The first condition guarantees non negativity. The second one states that the *trivial event* has the maximal possible probability of 1. The third condition implies that given a set of pairwise disjoint events, the probability of either one of them occurring is equal to the sum of the probabilities of each one.

From these conditions follows

- $P(\emptyset) = 0$.
- $P(\alpha \cup \beta) = P(\alpha) + P(\beta) - P(\alpha \cap \beta)$.

The triple $(\Omega, \mathcal{F}, P)$ is called a *probability space*.

Definition 3. Given two events $\alpha, \beta \in \mathcal{F}$, with $P(\beta) \neq 0$, the conditional probability of α given β is defined as the quotient of the probability of the joint events and the probability of β:

$$P(\alpha \mid \beta) = \frac{P(\alpha \cap \beta)}{P(\beta)}.$$

Theorem 1. ***(Bayes' theorem)***. *Let α, β be two events of an experiment, given that $P(\beta) \neq 0$. Then*

$$P(\alpha \mid \beta) = \frac{P(\beta \mid \alpha)P(\alpha)}{P(\beta)}.$$

Example 1. Consider a study where the relation of a disease d and an habit h is being investigated. Suppose that $P(d) = 10^{-5}$, $P(h) = 0.5$ and $P(h \mid d) = 0.9$. What is the probability that a person with habit h will have the disease d?

$$P(d \mid h) = \frac{P(d \cap h)}{P(h)} = \frac{P(h \mid d)P(d)}{P(h)} = \frac{0.9 \times 10^{-5}}{0.5} = 1.8 \times 10^{-5}.$$

If we set the probability of having habit h to a much lower value as $P(h) = 0.001$, then the above calculation gives approximately $1/100$. Intuitively, a smaller number of people have the habit and most of them have the desease. This means that the relation between having the desease and the habit is stronger now compared with the case where more people had the habit.

Definition 4. We say that two events $\alpha, \beta \in \mathcal{F}$ are *independent* if knowing one of them does not give any extra information about the other. Mathematically,

$$P(\alpha \cap \beta) = P(\alpha)P(\beta), \quad P(\alpha \mid \beta) = P(\alpha).$$

Let $\gamma \in \mathcal{F}$, we say that α and β are *conditionally independent* on γ, $\alpha \perp\!\!\!\perp \beta \mid \gamma$ if and only if

$$P(\alpha \cup \beta \mid \gamma) = P(\alpha \mid \gamma)P(\beta \mid \gamma).$$

Otherwise, they are said to be *conditionally dependent* on γ, $\alpha \not\perp\!\!\!\perp \beta \mid \gamma$.

Now we are going to introduce the concept of *random variable* and some properties as we have done with events.

Definition 5. A function $f : \Omega_1 \to \Omega_2$ between two measurable spaces $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ is said to be *measurable* if $f^{-1}(\alpha) \in \mathcal{F}_1$ for every $\alpha \in \mathcal{F}_2$.

Definition 6. A *random variable* is a measurable function $X : \Omega \to E$ from a probability space $(\Omega, \mathcal{F}, P)$ to a measurable space $(E, \mathcal{F}')$ verifying $X(\omega) \in \mathcal{F}' \ \forall \omega \in \Omega$.

The probability of X taking a value on a measurable set $S \in E$ is written as

$$P_X(S) = P(X \in S) = P(\{a \in \Omega \mid X(a) \in S\}).$$

We could make a question like "How likely is that the value of X equals a?". This is the same as asking for the probability of the set $\{\omega \in \Omega \mid X(w) = a\}$.

We will set the following notation that is going to be used, that is: random variables will be denoted with an upper case letter like X and a set of variables with a bold symbol like $\boldsymbol{X}$. The meaning of $P(state)$ will be clear without a reference to the variable. Otherwise $P(X = state)$ will be used. Using a lower case letter like $P(x)$ will denote the probability of the corresponding upper case variable X taking a specific value.

Definition 7. The *cumulative distribution function* of a real-valued random variable X is defined as

$$F_X(x) = P(X \leq x),$$

where the right-hand side represents the probability of the random variable taking value below or equal to x.

Definition 8. When the image of a random variable X is countable, the random variable is called a *discrete random variable* and its *probability mass function* p gives the probability of it being equal to some value:

$$p(x) = P(X = x).$$

In case the image is uncountable and real, then X is called a *continuous random variable* and if there exists is a non-negative Lebesgue-integrable f such that

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)du,$$

then it is called its *probability density function*.

A *mixed random variable* is a random variable who is neither discrete nor continuous, it can be realized as the sum of a discrete and continuous random variables. An example of a random variable of mixed type would be based on an experiment where a coin is flipped and a random positive number is chose only if the result of the coin toss is heads, -1 otherwise.

From now on, $P(x)$ will denote $f_X(x)$ when X is a continuous random variable. We define the *probability distribution* P_X of a random variable X over the probability space $(\Omega, \mathcal{F}, P)$ as the pushforward measure of it, that is, $P_X = PX^{-1}$.

Integrate notation will be used in both continuous and discrete cases, where the last one can be interpreted as integration with respect to the *counting measure* defined as

$$\#(dx) = \sum_{n \in \mathcal{I}} \delta(x - n)dx,$$

where $\mathcal{I}$ is the set of values X can take, and δ is the Dirac distribution. Given this measure, integration corresponds to summation as

$$\int_x P(x)\#(dx) = \sum_{n \in \mathcal{I}} \int_x P(x)\delta(x - n)dx = \sum_{n \in I} P(n).$$

Where we used that $\int f(x)\delta(x - x_0) = f(x_0)$. Given this, from now on, we will use the integration notation for both discrete and continuous variables given that the integrals will be respect to the counting measure when needed.

Definition 9. As we did for events, we can define the *conditional probability* over random variables, let X, Y be two random variables, then

$$P(x \mid y) = \frac{P(x,y)}{P(y)}.$$

It is required that $P(y) \neq 0$ for the conditional probability to be defined.

We can also enunciate the *Bayes' theorem* as

$$P(x,y) = \frac{P(y \mid x)P(x)}{P(y)}$$

Clearly an arbitrary number of variables can be considered in both cases.

Definition 10. The *marginal distribution* of a subset of random variables is the probability distribution of the variables contained in that subset.

Let X, Y be two random variables, it follows that

$$P(x) = \int_y P(x,y).$$

Definition 11. Let $\boldsymbol{X} = \{X_1, X_2, \ldots, X_n\}$ be a set of random variables, the *joint probability distribution* for $\boldsymbol{X}$ is function that gives the probability of each random variable X_i falling in a particular range or discrete set of values for that variable. It is called a *multi-variate distribution*.

When using only two random variables, then is called a *bi-variate distribution*.

This distribution can be expressed either in terms of a joint cumulative distribution function

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = P(X_1 \leq x_1, \ldots, X_n \leq x_n)^1,$$

or using a probability density or mass function.

Definition 12. We say that two random variables X and Y are *independent* if knowing one of them doesn't give any extra information about the other. Mathematically,

$$P(x,y) = P(x)P(y).$$

From this it follows that if X and Y are independent, then $P(x \mid y) = P(x)$.

Definition 13. Let X, Y and Z be three random variables, then X and Y are *conditionally independent* given Z if and only if

$$P(x,y \mid z) = P(x \mid z)P(y \mid z),$$

in that case we will denote $X \perp\!\!\!\perp Y \mid Z$. If X and Y are not conditionally independent, they are *conditionally dependent* $X \not\perp\!\!\!\perp Y \mid Z$

1 Where $\boldsymbol{x} = (x_1, \ldots, x_n)$

Both independence definitions can be made over sets of variables $\boldsymbol{X}, \boldsymbol{Y}$ and $\boldsymbol{Z}$ in a straight forward way.

Definition 14. We say that a set of n random variables $\{X_1, \ldots, X_n\}$ defined to assume values in $I \subset \mathbb{R}$ are *independent and identically distributed (i.i.d)* if and only if they are independent, i.e,

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = F_{X_1}(x_1) \ldots F_{X_n}(x_n) \ \forall x_1, \ldots, x_n \in I,$$

and are identically distributed

$$F_{X_1}(x_1) = F_{X_k}(x_k) \ \forall k \in \{2, \ldots, n\} \text{ and } \forall x \in I.$$

Definition 15. A *multi-variate random variable* or *random vector* is a column vector $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ whose components are random variables that can be defined over different probability spaces.

Note that we use the same symbol $\boldsymbol{X}$ for random vectors and sets of variables, but the meaning will be clear within the context.

2

DISTRIBUTIONS

In this section we will summarize some concepts concerning probability distributions among with some of the most used ones.

From now on, let X be a random variable and P its probability distribution.

Definition 16. The *mode* X_* of the probability distribution P is the state of X where the distribution takes it's highest value

$$X_* = \arg\max_x P(x).$$

A distribution could have more than one mode, in this case we say it is *multi-modal*.

Definition 17. The notation $\mathbb{E}\left[X\right]$ is used to denote the *average* or *expectation* of the values a real-valued variable takes respect to its distribution. It is worth mentioning that it might not exists. If X is non-negative, it is defined as

$$\mathbb{E}\left[X\right] = \int_\Omega X(\omega) dP(\omega)^1.$$

For a general variable X it is defined as $\mathbb{E}\left[X\right] = \mathbb{E}\left[X^+\right] - \mathbb{E}\left[X^-\right]$. Where

$$X^+(\omega) = \max(X(\omega), 0), \quad X^-(\omega) = \min(X(\omega), 0).$$

Suppose now that X is a real-valued random variable, in case it is also continuous, the expectation is

$$\mathbb{E}\left[X\right] = \int_{-\infty}^{+\infty} x f(x) dx.$$

And if it is discrete, let x_i be the values X can take, the expectation takes the form

$$\mathbb{E}\left[X\right] = \sum_{i=1}^{+\infty} x_i p(x_i) dx.$$

Let $g : \mathbb{R} \to \mathbb{R}$ be a measurable function, then $g \circ X$ is another random variable and we can talk about $\mathbb{E}\left[g(X)\right]$, so in case X is continuous, we have that

$$\mathbb{E}\left[g(X)\right] = \int_{-\infty}^{+\infty} g(x) f(x) dx.$$

1 P is a measure over Ω

In the above definition, the expectation is calculated over the probability distribution P of the random variable, in future sections this distribution is unknown and a guessed distribution Q will be used. In this cases when the distribution is not clear from the context the notation $\mathbb{E}_Q\left[X\right]$ will be used.

Definition 18. We define the k^{th} moment of a distribution as the average of X^k over the distribution

$$\mu_k = \mathbb{E}\left[X^k\right]$$

For $k = 1$ it is typically denoted as μ. Note μ_k can also denote the k^{th} element in the mean vector of a multi-variate variable.

Definition 19. The *variance* of a distribution is defined as

$$Var(X) = \sigma^2 = \mathbb{E}\left[(X - \mathbb{E}\left[X\right])^2\right] = \mathbb{E}\left[X\right]^2 - \mathbb{E}\left[X^2\right]$$

The square root of the variance σ is called the *standard deviation*.

When using a multi-variate distribution $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ we can talk about the *covariance matrix* $\boldsymbol{\Sigma}$ whose elements are

$$\begin{aligned}\boldsymbol{\Sigma}_{ij} &= \mathbb{E}\left[\left(X_i - \mathbb{E}\left[X_i\right]\right)\left(X_j - \mathbb{E}\left[X_j\right]\right)\right] = \mathbb{E}\left[(X_i - \mu_i)(X_j - \mu_j)\right] \\ &= \mathbb{E}\left[X_i X_j\right] - \mathbb{E}\left[X_i\right]\mathbb{E}\left[X_j\right]\end{aligned}$$

The following result will be helpful later on.

Proposition 1. *Let $\mathcal{X} = \{X_1, \ldots, X_n\}$ be a set of random variables, $\mathcal{X}_0 \subset \mathcal{X}$ and $P(\mathcal{X}), P(\mathcal{X}_0)$ their probability distributions. It follows that the expectation of a function g over $\mathcal{X}_0$, verifies*

$$\mathbb{E}_{P(\mathcal{X})}\left[g(\mathcal{X}_0)\right] = \mathbb{E}_{P(\mathcal{X}_0)}\left[g(\mathcal{X}_0)\right].$$

That is, we only need to know the marginal distribution of the subset in order to carry out the average.

Proof. Let $\mathcal{I} = (i_1, \ldots, i_k)$ be the indexes corresponding to $\mathcal{X}_0$, then

$$\begin{aligned}\mathbb{E}_{P(\mathcal{X})}\left[g(\mathcal{X}_0)\right] &= \int_{x_1} \cdots \int_{x_n} g(x_{i_1}, \ldots, x_{i_k}) f(x_1, \ldots, x_n) \\ &= \int_{x_{i_1}} \cdots \int_{x_{i_k}} g(x_{i_1}, \ldots, x_{i_k}) \int \cdots \int f(x_1, \ldots, x_n) d_{x_1}, \ldots, d_{x_n} \\ &= \int_{x_{i_1}} \cdots \int_{x_{i_k}} g(x_{i_1}, \ldots, x_{i_k}) f(x_{i_1}, \ldots, x_{i_k}) = \mathbb{E}_{P(\mathcal{X}_0)}\left[g(\mathcal{X}_0)\right].\end{aligned}$$

Where in the second-last equality we used marginalization. □

We are going to discuss now some examples of probability distributions that are going to be used from now on.

2.1 DISCRETE DISTRIBUTIONS

Bernoulli Distribution

The Bernoulli distribution describes a discrete binary variable X that takes the value 1 with probability p and the value 0 with probability $1 - p$.

$$P(x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases}.$$

Categorical Distribution

A generalization of the Bernoulli Distribution when the variable can take more than two states is the *Categorical Distribution*. Let $Dom(X) = \{1, \ldots, N\}$, then X follows a categorical distribution of parameters $\theta = (\theta_1, \ldots, \theta_N)$ if and only if

$$P(x \mid \theta) = \prod_{i=1}^{N} \theta_i^{\mathbb{I}[x=i]} \text{ and } \sum_{i=1}^{N} \theta_i = 1.$$

Binomial Distribution

The binomial distribution describes the number of successes in a sequence of independent Bernoulli Trials. A discrete binary random variable X follows a *binomial distribution* of parameters $n \in \mathbb{N}$ and $p \in [0, 1]$, denoted as $X \sim B(n, p)$ if and only if

$$P(x = k) = \binom{n}{k} p^k (1 - p)^{n-k}.$$

Multinomial Distribution

The multinomial distribution is a generalization of the binomial distribution which describes the result of a sequence of independent trials in a categorical distribution. A discrete random variable X follows a *multinomial distribution* of parameters $n \in \mathbb{N}$, $\boldsymbol{p} = (p_1, \ldots, p_K)$ such that $\sum p_k = 1$ if and only if

$$P(\boldsymbol{x} = x_1, \ldots, x_K) = \begin{cases} \frac{n!}{x_1! \ldots x_K!} \prod_{k=1}^{K} p_k^{x_k} & \text{if } \sum_k x_k = n \\ 0 & \text{otherwise} \end{cases}$$

2.2 CONTINUOUS DISTRIBUTIONS

Uni-Variate Normal Distribution

The *normal* or *Gaussian distribution* is a type of continuous probability distribution for real-valued random variables.

Definition 20. We say the real valued random variable X follows a *normal distribution* of parameters $\mu, \sigma \in \mathbb{R}$, denoted as $X \sim N(\mu, \sigma)$ if and only if, its probability density function exists and is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

The parameter μ is the mean or expectation of the distribution and σ is its standard deviation.

The simplest case of a normal distribution is known as *standard normal distribution*, denoted as Z. It is a special case where $\mu = 0$ and $\sigma = 1$, then its density function is

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}.$$

One of the properties of the normal distribution is that if $X \sim N(\mu, \sigma)$, $a, b \in \mathbb{R}$ and $f : \mathbb{R} \to \mathbb{R}$ be defined as $f(x) = ax + b$, then $f(X) \sim N(\mu + b, a^2\sigma)$.

Multi-Variate Normal Distribution

This distribution plays a fundamental role in this project so we will discuss its properties in more detail.

This distribution is an extension of the uni-variate one when having a multi-variate random variable.

Definition 21. We say that a random vector $\boldsymbol{X} = (X_1, \ldots, X_p)$ follows a *multi-variate normal distribution* of parameters $\boldsymbol{\mu} \in \mathbb{R}^p$ and $\boldsymbol{\Sigma} \in \mathbb{M}_n(\mathbb{R})$, denoted as $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ if and only if its probability density function is

$$f(\boldsymbol{x}) = \frac{1}{\sqrt{det(2\pi\boldsymbol{\Sigma})}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}.$$

Where $\boldsymbol{\mu}$ is the mean vector of the distribution, and $\boldsymbol{\Sigma}$ the covariance matrix. The inverse matrix $\boldsymbol{\sigma}^{-1}$ is called *precision*. It also satisfies that

$$\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{X}], \quad \boldsymbol{\Sigma} = \mathbb{E}\left[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^T\right].$$

As $\boldsymbol{\Sigma}$ is a real symmetric matrix, it can be eigendecomposed

$$\boldsymbol{\Sigma} = \boldsymbol{E}\boldsymbol{\Delta}\boldsymbol{E}^T,$$

where $\boldsymbol{E}^T\boldsymbol{E} = \boldsymbol{I}$ and $\boldsymbol{\Delta} = diag(\lambda_1, \ldots, \lambda_n)$.

Using the transformation

$$\boldsymbol{y} = \boldsymbol{\Delta}^{\frac{1}{2}}\boldsymbol{E}^T(\boldsymbol{x} - \boldsymbol{\mu}),$$

we get that

$$(\boldsymbol{x} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}(\boldsymbol{x} - \boldsymbol{\mu}) = \boldsymbol{y}^T\boldsymbol{y}.$$

Using this, the multi-variate Normal Distribution reduces to a product of n uni-variate standard normal distributions.

Gamma Distribution

Another continuous distribution that we are going to use is the *Gamma distribution*.

Definition 22. We say that a continuous random variable X defined on $\mathbb{R}^+$ follows a *Gamma distribution* of parameters $\alpha, \beta > 0$, denoted as $X \sim Gamma(\alpha, \beta)$ if and only if its density function is

$$f(x) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)},$$

where Γ is the Gamma function defined as

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx.$$

The mean is given by $\mathbb{E}\Big[X\Big] = \frac{\alpha}{\beta}$.

Definition 23. The *inverse gamma distribution* is defined by the density function

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\beta/x}.$$

Beta Distribution

Definition 24. We say that a continuous random variable X defined on the interval $[0, 1]$ follows a *Beta distribution* of parameters $\alpha, \beta > 0$, denoted as $X \sim Beta(\alpha, \beta)$ if and only if its density function is

$$f(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1},$$

where $B(\alpha, \beta)$ is the *beta function* defined as

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1} (1-x)^{\beta-1} dx.$$

The mean is given by $\mathbb{E}\Big[X\Big] = \frac{\alpha}{\alpha+\beta}$.

Dirichlet Distribution

The Dirichlet distribution is a family of continuous multi-variate probability distributions parameterized by a vector $\boldsymbol{\alpha}$ of positive reals. It is a multi-variate generalization of the Beta Distribution.

Definition 25. We say that a continuous random multi-variate variable $\boldsymbol{X}$ with order $K \geq 2$, follows a *Dirichlet Distribution* with parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$, if and only if its density function is defined as

$$f(\boldsymbol{x}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} x_k^{\alpha_k - 1},$$

and it satisfies that

$$\sum_{k=1}^{K} x_k = 1 \text{ and } x_k > 0 \; \forall k = 1, \dots, K.$$

Where the normalization constant is the multi-variate beta function

$$B(\boldsymbol{\alpha}) = \frac{\prod_k \Gamma(\alpha_k)}{\Gamma\left(\sum_k \alpha_k\right)}.$$

When the vector parameter $\boldsymbol{\alpha}$ is filled with the same value α_0, the distribution is called Symmetric-Dirichlet with parameter α_0.

Proposition 2. *Let* $(X_0, \dots, X_n) \sim Dirichlet(\alpha_0, \dots, \alpha_n)$*, then* $X_0 \sim Beta(\alpha_0, \alpha_1 + \cdots + \alpha_n)$.

Proof. Following Farrow (2008), we can write the joint probability as

$$f(x_1, \dots, x_n) = f_1(x_1) f_2(x_2 \mid x_1) \dots f_{n-1}(x_{n-1} \mid x_1, \dots, x_{n-2}).$$

We do not need the last term because it is fixed given the others. In fact, let $A = \sum_i \alpha_i$, we can write it as

$$\left(\frac{\Gamma(A)}{\Gamma(\alpha_1)\Gamma(A-\alpha_1)} x_1^{\alpha_1 - 1}(1-x_1)^{A-\alpha_1-1}\right)\left(\frac{\Gamma(A-\alpha_1)}{\Gamma(\alpha_2)\Gamma(A-\alpha_1-\alpha_2)} \frac{x_2^{\alpha_2-1}(1-x_1-x_2)^{A-\alpha_2-\alpha_1-1}}{(1-\alpha_1)^{A-\alpha_1-1}}\right)$$

$$\dots \left(\frac{\Gamma(A-\alpha_1-\cdots-\alpha_{n-2})}{\Gamma(\alpha_{n-1})\Gamma(A-\alpha_1-\cdots-\alpha_{n-1})} \frac{x_{n-1}^{\alpha_n-1} x_n^{\alpha_n-1}}{(1-x_1-\cdots-x_{n-2})^{\alpha_{n-1}+\alpha_n-1}}\right).$$

From this, we get that

$$f_1(x_1) = \frac{\Gamma(A)}{\Gamma(\alpha_1)\Gamma(A-\alpha_1)} x_1^{\alpha_1-1}(1-x_1)^{A-\alpha_1-1} \implies X_1 \sim \text{Beta}(\alpha_1, A-\alpha_1).$$

Making the decomposition over any other X_j, results on $X_j \sim \text{Beta}(\alpha_j, A-\alpha_j)$. □

2.3 KULLBACK-LEIBLER DIVERGENCE

Definition 26. Let P and Q be two probability distributions over the same probability space $\mathcal{X}$, the *Kullback-Leibler divergence* $KL\Big(Q \mid P\Big)$ measures the "difference" between both distributions as

$$KL\Big(Q \mid P\Big) = \mathbb{E}_Q\Big[\log Q(x) - \log P(x)\Big].$$

The Kullback-Leibler divergence is defined if and only if for all $x \in \mathcal{X}$ such that $P(x) = 0$, then $Q(x) = 0$. In meassure terms, Q is absolutely continuous with respect to P.

Proposition 3. *The Kullback-Leibler divergence is always non-negative.*

Proof. As the logarithm is bounded by $x - 1$, we can bound $\log \frac{P(x)}{Q(x)}$

$$\log x \leq x - 1 \implies \frac{P(x)}{Q(x)} - 1 \geq \log \frac{P(x)}{Q(x)}.$$

Since probabilities are non-negative, we can multiply by $Q(x)$ in the last inequality

$$P(x) - Q(x) \geq Q(x) \frac{\log P(x)}{\log Q(x)} = Q(x) \log P(x) - Q(x) \log Q(x).$$

Now we integrate (sum in case of discrete variables) both sides

$$0 \geq \mathbb{E}_Q \Big[\log P(x) - \log Q(x) \Big] \implies \mathbb{E}_Q \Big[\log Q(x) - \log P(x) \Big] \geq 0.$$

□

As a result, the Kullback-Leibler divergence is 0 if and only if the two distributions are equal almost everywhere.

3

GRAPH THEORY

Definition 27. A *graph* $G = (V, E)$ is a set of vertices or nodes V and edges $E \subset V \times V$ between them. If V is a set of ordered pairs then the graph is called a *directed graph*, otherwise if V is a set of unordered pairs it is called an *undirected graph*.

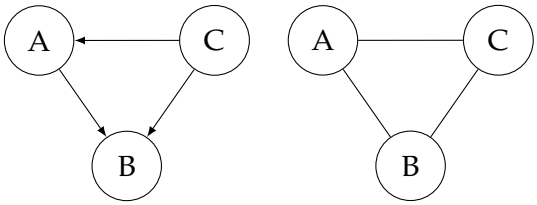


Figure 1: Example of directed and undirected graph, respectively.

Definition 28. In a directed graph $G = (V, E)$, a *directed path* $A \to B$ is a sequence of vertices $A = A_0, A_1, \ldots, A_{n-1}, A_n = B$ where $(A_i, A_{i+1}) \in E \ \forall i \in 0, \ldots, n-1$.

If G is a undirected graph, $A \to B$ is an *undirected path* if $\{A_i, A_{i+1}\} \in E \ \forall i \in 0, \ldots, n-1$

Definition 29. Let A, B be two vertices of a directed graph G. If $A \to B$ is a directed path and $B \not\to A$ (meaning there isn't a directed path from B to A), then A is called an *ancestor* of B and B is called a *descendant* of A.

For example, in the figure 1, C is an ancestor of B.

Definition 30. A *directed acyclic graph (DAG)* is a directed graph such that no directed path between any two nodes revisits a vertex.

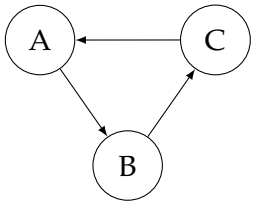


Figure 2: Example of graph which isn't a DAG.

As we can see in the figure 2, $A \to B \to C \to A \to B$ is a path from A to B that revisits A.

Now where are going to define some relations between nodes in a DAG.

Definition 31. The *parents* of a node A is the set of nodes B such that there is a directed edge from B to A. The same applies for the *children* of a node.

The *Markov blanket* of a node is composed by the node itself, its children, its parents and the parents of its children.

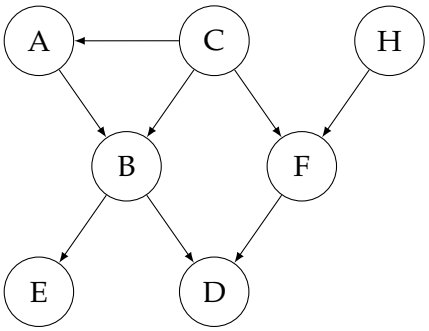


Figure 3: Directed acyclic graph

Definition 32. In a graph, the *neighbors* of a node are those directly connected to it.

We can use figure 3 to reflect on these definitions. The parents of B are $pa(B) = \{A, C\}$ and its children are $ch(B) = \{E, D\}$. Taking this into account, its neighbors are $ne(B) = \{A, C, E, D\}$ and its Markov blanket is $\{A, B, C, D, E, F\}$.

Definition 33. Let G be a DAG, U be a path between two vertex and $A \in U$

- A is called a *collider* if $\forall B \in ne(A) \cap U, (B, A) \in E$.
- A is called a *fork* if $\forall B \in ne(A) \cap U, (A, B) \in E$.

Notice, a vertex can be a collider for a path but not for others. A vertex is said to be a collider or a fork without any reference to the path when it is for any path that goes through it. This happens when the edge direction condition is satisfied for every neighbor.

For example in figure 3, D is a collider and C is a fork.

Definition 34. Let G be an undirected graph, a *clique* is a maximally connected subset of vertices. That is, all the members of the clique are connected to each others and there is no bigger clique that contains another.

Formally, $S \subset V$ is a *clique* if and only if $\forall A, B \in S,\ \{A, B\} \in E$ and $\nexists C \in V \backslash S$ such that $\forall A \in S,\ \{A, C\} \in E$.

Part II

STATISTICAL INFERENCE

Statistical inference is the process of using data analysis to deduce properties of an underlying distribution.

Bayesian inference is a method of statistical inference in which Bayes' theorem is used, it derives the *posterior probability* as a consequence of two antecedents: a *prior probability* and a *likelihood function* derived from a statistical model for the observed data.

Maximum likelihood estimation (MLE) is a method of estimating the parameters of a probability distribution by maximizing a *likelihood function*, so that under the assumed statistical model the observed data is most probable.

4

INTRODUCTION

Inferential statistical analysis infers properties of a population or dataset, using different techniques as testing hypotheses and deriving estimates. This analysis assumes that the observed data set is sampled from a larger population (Upton & Cook (2014)).

Descriptive statistics is solely concerned with properties of the observed data, and opposed to *inferential statistics* does not rest on the assumption that the data comes from a larger population.

In *machine learning* the procedure for deducing properties of the model is typically referred to as *training or learning* rather than inference, in contrast, using a model for prediction is referred to as *inference*.

A *statistical model* is a set of assumptions concerning the generation of the observed data and similar data (Cox (2006)). There are different levels of modeling assumptions, which differ on whether the process that generates the data, is fully, partially or minimally described by a family of probability distributions involving a finite amount of unknown parameters. This study's approach is *fully parametric*, which assumes this generation is fully described by those parameters.

Different schools or paradigms of statistical inference have become established (Bandyopadhyay & Forster (2011)). These paradigms are not mutually exclusive, and methods that work well under one paradigm often have attractive interpretations under other paradigms. In this chapter we are reviewing two of them: the *Bayesian paradigm* and the *likelihoodist paradigm*.

As purely Bayesian or likelihoodist methods are beyond the scope of this study, we are just reviewing the needed definitions to later understand related *variational methods*, a few example are analyzed in each section.

5

MAXIMUM LIKELIHOOD

Given a set of observations $\boldsymbol{x}$ and parameters $\boldsymbol{\theta}$ that model the obtained samples via $P(\boldsymbol{x} \mid \boldsymbol{\theta})$, *maximum likelihood estimation* (Rossi (2018)) is a *classical inference* method of estimating the maximum likelihood parameters, i.e, the ones that maximizes the likelihood $P(\boldsymbol{x} \mid \boldsymbol{\theta})$:

$$\boldsymbol{\theta}^{ML} = \arg\max_{\boldsymbol{\theta}} P(\boldsymbol{x} \mid \boldsymbol{\theta}).$$

This value symbolizes the *value of the parameter to which the data is most probable to be generated with*. There are several techniques for finding this value, for example, if the likelihood function is differentiable, its derivate can be used to determine the maxima.

Remark 1. Maximum likelihood estimation does not consider a probability distribution over the parameter, whereas Bayesian estimation does.

5.1 MAXIMUM LIKELIHOOD AND THE EMPIRICAL DISTRIBUTION

Consider a set of i.i.d random variables $\boldsymbol{X} = (X_1, \ldots, X_N)$ and their observations $\boldsymbol{x} = (x_1, \ldots, x_N)$, we are going to show the relation between the maximum likelihood and the Kullback-Leibler divergence of the empirical distribution and our model. The empirical distribution is defined as the distribution whose probability mass function Q is

$$Q(x) = \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}[x = x_n].$$

Where X is i.i.d with the rest of variables we are considering.

Proposition 4. *In case $P(x \mid \boldsymbol{\theta})$ is unconstrained, maximum likelihood distribution corresponds to the empirical distribution, i.e, $P(x \mid \boldsymbol{\theta}^{ML}) = Q(x)$.*

Proof. The Kullback-Leibler divergence between the empirical and our considered model $P(x \mid \boldsymbol{\theta})$ is:

$$KL\Big(Q \mid P\Big) = \mathbb{E}_Q\Big[\log Q(x)\Big] - \mathbb{E}_Q\Big[\log P(x \mid \boldsymbol{\theta})\Big].$$

Notice the term $\mathbb{E}_Q\Big[\log Q(\boldsymbol{x})\Big]$ is a constant and the log likelihood under Q takes the form

$$\mathbb{E}_Q\Big[\log P(x \mid \boldsymbol{\theta})\Big] = \frac{1}{N} \int_x \sum_{n=1}^{N} \mathbb{I}[x = x_n] \log P(x \mid \boldsymbol{\theta}) = \frac{1}{N} \sum_{n=1}^{N} \log P(x_n \mid \boldsymbol{\theta}).$$

As the logarithm is a strictly increasing function, maximizing the log likelihood equals to maximize the likelihood itself, in conclusion, it is equivalent to minimize the Kullback-Leibler divergence between the empirical distribution Q and our distribution P.

$$\arg\min_{\boldsymbol{\theta}} KL\Big(Q \mid P\Big) = \arg\min_{\boldsymbol{\theta}} -\mathbb{E}_Q\Big[\log P(\boldsymbol{x} \mid \boldsymbol{\theta})\Big] = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_Q\Big[\log P(\boldsymbol{x} \mid \boldsymbol{\theta})\Big] =$$
$$\arg\max_{\boldsymbol{\theta}} \frac{1}{N}\sum_{n=1}^{N} \log P(x_n \mid \boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \frac{1}{N}\sum_{n=1}^{N} P(x_n \mid \boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \sum_{n=1}^{N} P(x_n \mid \boldsymbol{\theta}) = \boldsymbol{\theta}^{ML}.$$

□

6

BAYESIAN INFERENCE

Bayesian inference considers a probability distribution over the set of parameters. This distribution is then governed by so called *hyper-parameters* $\boldsymbol{\alpha}$, these are typically omitted when using $P(\boldsymbol{\theta} \mid \boldsymbol{\alpha})$, where $P(\boldsymbol{\theta})$ is written instead.

Bayesian inference attempts to determine the *posterior distribution* $P(\boldsymbol{\theta} \mid \boldsymbol{x})$ using a *prior belief* $P(\boldsymbol{\theta})$ and the *likelihood function* $P(\boldsymbol{x} \mid \boldsymbol{\theta})$ on the basis of Bayes' theorem:

$$P(\boldsymbol{\theta} \mid \boldsymbol{x}) = \frac{P(\boldsymbol{x} \mid \boldsymbol{\theta})P(\boldsymbol{\theta})}{P(\boldsymbol{x})}.$$

6.1 EXAMPLE: DISCRETE PRIOR

For this example, consider a set of i.i.d variables $\boldsymbol{X} = (X_1, \ldots, X_N)$ with their corresponding set of observations be $\boldsymbol{x} = (x_1, \ldots, x_N)$, where each X models the results of coin-tossing experiment, let 1 symbolize *heads* and 0 *tails*.

Bayesian inference attempts to estimate the probability distribution of θ given $\boldsymbol{x}$. This parameter models the probability of the tossing resulting in heads as

$$P(x_n = 1 \mid \theta) = \theta \quad \forall n \in \{1, \ldots, N\}.$$

The joint probability takes the form:

$$P(\boldsymbol{x}, \theta) = P(\theta) \prod_{n=1}^{N} P(x_n \mid \theta).$$

We want to calculate the posterior distribution:

$$P(\theta \mid \boldsymbol{x}) = \frac{P(\boldsymbol{x} \mid \theta)P(\theta)}{P(\boldsymbol{x})},$$

to do so, we need to specify the prior distribution $P(\theta)$. For now, we are using a discrete variable that verifies:

$$P(\theta = 0.2) = 0.1, \quad P(\theta = 0.5) = 0.7 \quad \text{and} \quad P(\theta = 0.8) = 0.2.$$

This means that we have a 70% belief that the coin is fair, a 10% belied that is biased to tails and 20% that is biased to heads. Let n_h be the number of heads in our observed data and n_t the number of tails, mathematically:

$$n_h = \#\{x \in \boldsymbol{x} \ : \ x = 1\} \quad \text{and} \quad n_t = \#\{x \in \boldsymbol{x} \ : \ x = 0\}.$$

Given that $x_n \mid \theta$ is a Bernoulli trail $\forall n \in \{1, \dots, N\}$, the posterior is:

$$P(\theta \mid \boldsymbol{x}) = \frac{P(\theta)}{P(\boldsymbol{x})} \theta^{n_h} (1-\theta)^{n_t}.$$

Suppose that $n_h = 2$ and $n_t = 8$, the posterior might be calculated up to a normalization factor:

$$P(\theta = 0.2 \mid \boldsymbol{x}) = \frac{1}{P(\boldsymbol{x})} \times 0.1 \times 0.2^2 \times 0.8^8 = \frac{1}{P(\boldsymbol{x})} \times 6.71 \times 10^{-4},$$
$$P(\theta = 0.5 \mid \boldsymbol{x}) = \frac{1}{P(\boldsymbol{x})} \times 0.7 \times 0.5^2 \times 0.5^8 = \frac{1}{P(\boldsymbol{x})} \times 6.83 \times 10^{-4},$$
$$P(\theta = 0.8 \mid \boldsymbol{x}) = \frac{1}{P(\boldsymbol{x})} \times 0.2 \times 0.2^2 \times 0.8^8 = \frac{1}{P(\boldsymbol{x})} \times 3.27 \times 10^{-7}.$$

We can compute the normalizing factor as

$$P(\boldsymbol{x}) = \sum_{\theta \in \{0.2, 0.5, 0.8\}} P(\boldsymbol{x}, \theta) = 6.71 \times 10^{-4} + 6.83 \times 10^{-4} + 3.27 \times 10^{-7} = 0.00135.$$

Therefore, the posterior is

$$P(\theta = 0.2 \mid \boldsymbol{x}) = 0.4979,$$
$$P(\theta = 0.5 \mid \boldsymbol{x}) = 0.5059,$$
$$P(\theta = 0.8 \mid \boldsymbol{x}) = 0.00024.$$

6.2 EXAMPLE: CONTINUOUS PRIOR

In the previous example, we have used a discrete prior for the parameter distribution, a continuous prior might be chosen instead. Suppose an uniform prior distribution:

$$P(\theta) = k \implies \int_0^1 P(\theta) d\theta = k = 1$$

due to normalization.

Using the previous calculations we have

$$P(\theta \mid \boldsymbol{x}) = \frac{1}{P(\boldsymbol{x})} \theta^{n_h} (1-\theta)^{n_t},$$

where

$$P(\boldsymbol{x}) = \int_0^1 \theta^{n_h} (1-\theta)^{n_t} d\theta.$$

This implies that

$$P(\theta \mid \boldsymbol{x}) = \frac{\theta^{n_h} (1-\theta)^{n_t}}{\int_0^1 u^{n_h} (1-u)^{n_t} du} \implies \theta \mid \boldsymbol{x} \sim Beta(n_h + 1, n_t + 1).$$

A Beta distribution could be also considered as the prior distribution:

$$\theta \sim \text{Beta}(\alpha, \beta) \implies P(\theta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha - 1} (1-\theta)^{\beta - 1}.$$

in this case, the posterior is:

$$P(\theta, \boldsymbol{x}) = \frac{1}{B(\alpha + n_h, \beta + n_t)} \theta^{\alpha + n_h - 1} (1-\theta)^{\beta + n_t - 1} \implies \theta \mid \boldsymbol{x} \sim Beta(n_h + \alpha, n_t + \beta)$$

6.3 UTILITY

The Bayesian posterior says nothing about how to benefit from the beliefs it represents, in order to do this we need to specify the utility of each decision.

With this idea we define an utility function over the parameters

$$U(\theta, \theta_{true}) = \alpha \mathbb{I}[\theta = \theta_{true}] - \beta \mathbb{I}[\theta \neq \theta_{true}],$$

where $\alpha, \beta \in \mathbb{R}$. This symbolizes the gains or looses of choosing the parameter θ, when the true value of the parameter is supposed to be θ_{true}. Therefore, the expected utility of a parameter θ_0 is calculated as

$$U(\theta = \theta_0) = \int_{\theta_{true}} U(\theta = \theta_0, \theta_{true}) P(\theta = \theta_{true} \mid \boldsymbol{x}).$$

We might as well define an utility function over the previous example:

$$U(\theta, \theta_{true}) = 10 \mathbb{I}[\theta = \theta_{true}] - 20 \mathbb{I}[\theta \neq \theta_{true}],$$

where we interpret that the loss of choosing the wrong parameter is twice as important as the gains from doing it right.

The expected utility of the decision that the parameter is $\theta = 0.2$ in our discrete example would be

$$\begin{aligned}
U(\theta = 0.2) &= U(\theta = 0.2, \theta_{true} = 0.2) P(\theta_{true} = 0.2 \mid \boldsymbol{x}) \\
&\quad + U(\theta = 0.2, \theta_{true} = 0.5) P(\theta_{true} = 0.5 \mid \boldsymbol{x}) \\
&\quad + U(\theta = 0.2, \theta_{true} = 0.8) P(\theta_{true} = 0.8 \mid \boldsymbol{x}) \\
&= 10 \times 0.4979 - 20 \times 0.5059 - 20 \times 0.00024 \\
&= -5.1438, \\
U(\theta = 0.5) &= -4.9038, \\
U(\theta = 0.8) &= -20.0736.
\end{aligned}$$

Given this, if we had to make a decision for the parameter, we could choose the value with the highest utility. Other approaches like the mode or mean (continuous posterior) of the distribution are possible.

6.4 MAXIMUM A POSTERIORI ESTIMATION

Maximum a posteriori probability estimation is a Bayesian inference method of estimating the mode of the posterior distribution. In contrast to maximum likelihood estimation, it employs an augmented optimization objective which incorporates a prior distribution.

Definition 35. *Maximum A Posteriori (MAP)* refers to the value of the parameter $\boldsymbol{\theta}$ that better fits the data:

$$\boldsymbol{\theta}^{MAP} = \arg\max_{\boldsymbol{\theta}} P(\boldsymbol{x} \mid \boldsymbol{\theta}) P(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} P(\boldsymbol{\theta} \mid \boldsymbol{x}).$$

Remark 2. Maximum likelihood estimation is a particular case of maximum a posterior estimation with a flat (constant) prior.

Remark 3. MAP estimation can be seen as a limiting case of Bayesian estimation under the 0–1 utility function:

$$U(\boldsymbol{\theta}, \boldsymbol{\theta}_{true}) = \mathbb{I}[\boldsymbol{\theta} = \boldsymbol{\theta}_{true}],$$

using this, the expected utility of a parameter $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ is

$$U(\boldsymbol{\theta} = \boldsymbol{\theta}_0) = \int_{\boldsymbol{\theta}_{true}} \mathbb{I}[\boldsymbol{\theta}_{true} = \boldsymbol{\theta}_0] P(\boldsymbol{\theta} = \boldsymbol{\theta}_{true} \mid \boldsymbol{x}) = P(\boldsymbol{\theta}_0 \mid \boldsymbol{x}).$$

This means that the maximum utility decision is to take the value $\boldsymbol{\theta}_0$ with the highest posterior value.

Part III

VARIATIONAL INFERENCE

Variational Bayesian methods consist of intractable integrals approximation techniques arising in inference and machine learning problems. They are commonly applied in complex models consisting of *observed variables*, *unknown parameters* and *latent variables*. The relation between these random variables might be described by a *graphical model*.

Variational Bayesian inference solves the inference problem by creating an equivalent *optimization problem* and approaching its solution through machine learning techniques.

7

INTRODUCTION

As typical in Bayesian inference, the parameters and latent variables are grouped together as "hidden variables". *Variational Bayesian methods* or simply *variational methods* are primarily used for two purposes:

1. Perform statistical inference over the unobserved variables by providing an analytical approximation to the posterior probability of them.
2. To derive a lower bound for the marginal likelihood of the observed data (i.e. marginal probability of the data over the unobserved variables). This is typically used for performing model selection, where a higher marginal likelihood for a given model indicates a better fit of the data by that model and hence a greater probability that the model in question was the one that generated the data.

The considered elements of a variational Bayesian model are: a set of observed variables $\boldsymbol{X} = \{X_1, \ldots, X_N\}$ among with hidden variables $\boldsymbol{Z} = \{Z_1, \ldots, Z_M\}$. Their corresponding set of observations $\boldsymbol{x} = (x_1, \ldots, x_N)$, $\boldsymbol{z} = (z_1, \ldots, z_M)$, where the latter denotes a possible configuration for the hidden variables.

The samples are governed by the joint distribution $P(\boldsymbol{x}, \boldsymbol{z})$. Inference consists of learning the posterior distribution of the hidden variables $P(\boldsymbol{z} \mid \boldsymbol{x})$, given the dataset $\boldsymbol{x}$.

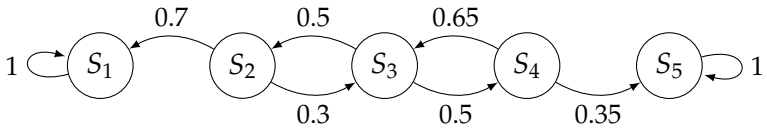
In *classical inference* the conditional is calculated as

$$P(\boldsymbol{z} \mid \boldsymbol{x}) = \frac{P(\boldsymbol{x}, \boldsymbol{z})}{\int_{\boldsymbol{z}} P(\boldsymbol{x}, \boldsymbol{z})} ,$$

but for many models this integral is computationally hard to solve.

As we have already reviewed using parameters, *Bayesian inference* derives the posterior probability $P(\boldsymbol{z} \mid \boldsymbol{x})$ as a consequence of a *prior probability* $P(\boldsymbol{z})$ and a *likelihood function* $P(\boldsymbol{x} \mid \boldsymbol{z})$. Other methods as *Markov chain Monte Carlo (MCMC)* and *variational Bayesian inference* try a different approach when solving the given inference problem: on one hand, *variational inference*, is a machine learning method whose main goal is to approximate probability distributions (Jordan *et al.* (1999); Wainwright & Jordan (2008)). On the other hand, *MCMC* approximates the posterior distribution using a Markov chain. Let us briefly introduce the main idea behind this method, we need to introduce two concepts: *Markov chain* and *MCMC*.

A *Markov Chain* is formally defined as a stochastic process, i.e, a family of random variables, that satisfies the *Markov property* also known as the memoryless property: *the conditional probability distribution of future states of the process (conditional on both present and past values) depends only on the present state*. To fully understand it, imagine a system with a number of possible states $S_1, \ldots, S_5$ and the probabilities of going from one state to another stated in the following diagram.



Consider a sequence of random variables X_t that symbolize the current state at the step t. The Markov property means that the probability of moving to the next state depends only on the present one, i.e,

$$P(X_{n+1} = x \mid X_1 = x_1 \ldots, X_n = x_n) = P(X_{n+1} = x \mid X_n = x_n).$$

We need two concepts to define the process of MCMC:

- **Ergodic Markov chain**. A Markov chain where it exists a number $N \in \mathbb{N}$ such that any state can be reached from any other state in any number of steps less or equal than N.
- **Stationary Distribution**. The probability distribution to which the process converges over time.

In MCMC, an ergodic Markov chain over the latent variables $\boldsymbol{Z}$ is considered, whose stationary distribution is the posterior $P(\boldsymbol{z} \mid \boldsymbol{x})$, samples are taken from the chain to approximate the posterior with them.

In contrast, *variational inference* exchanges the inference problem with an optimization one. It fixes a family of distributions $\mathcal{Q}$ over the latent variables $\boldsymbol{Z}$ and find the element that minimizes its Kullback-Leibler divergence with the posterior $P(\boldsymbol{z} \mid \boldsymbol{x})$:

$$Q^{opt} = \arg\min_{Q \in \mathcal{Q}} KL\Big(Q(\boldsymbol{z}) \mid P(\boldsymbol{z} \mid \boldsymbol{x})\Big).$$

These Q distributions are typically referred as *variational distributions* of the optimization problem.

Compared to *Markov Chain Monte Carlo (MCMC)*, variational inference tends to be faster and scale easier to large data (Blei *et al.* (2017)), it has been applied to different problems such as computer vision, computational neuroscience and document analysis (Blei (2014)).

Analyzing the Kullback-Leibler divergence, it may be decomposed in the following way:

$$\begin{aligned}
KL\Big(Q(\boldsymbol{z}) \mid P(\boldsymbol{z} \mid \boldsymbol{x})\Big) &= \mathbb{E}_{Q(\boldsymbol{z})}\Big[\log Q(\boldsymbol{z})\Big] - \mathbb{E}_{Q(\boldsymbol{z})}\Big[\log P(\boldsymbol{z} \mid \boldsymbol{x})\Big] \\
&= \mathbb{E}_{Q(\boldsymbol{z})}\Big[\log Q(\boldsymbol{z})\Big] - \mathbb{E}_{Q(\boldsymbol{z})}\Big[\log P(\boldsymbol{x}, \boldsymbol{z}) - \log P(\boldsymbol{x})\Big]. \\
&= \mathbb{E}_{Q(\boldsymbol{z})}\Big[\log Q(\boldsymbol{z})\Big] - \mathbb{E}_{Q(\boldsymbol{z})}\Big[\log P(\boldsymbol{x}, \boldsymbol{z})\Big] + \mathbb{E}_{Q(\boldsymbol{z})}\Big[\log P(\boldsymbol{x})\Big]. \\
&= \mathbb{E}_{Q(\boldsymbol{z})}\Big[\log Q(\boldsymbol{z})\Big] - \mathbb{E}_{Q(\boldsymbol{z})}\Big[\log P(\boldsymbol{x}, \boldsymbol{z})\Big] + \log P(\boldsymbol{x}).
\end{aligned}$$

Although the Kullback-Leibler divergence cannot be computed as long as $P(\boldsymbol{z} \mid \boldsymbol{x})$ is unknown, we can optimize an equivalent objective: we can use its positiveness to set the following lower bound to the evidence, defined as *evidence lower bound* or *ELBO*.

$$\log P(\boldsymbol{x}) \geq \underbrace{-\mathbb{E}_{Q(\boldsymbol{z})}\Big[\log Q(\boldsymbol{z})\Big]}_{Entropy} + \underbrace{\mathbb{E}_{Q(\boldsymbol{z})}\Big[\log P(\boldsymbol{x}, \boldsymbol{z})\Big]}_{Energy} = \text{ELBO}(Q).^{1}$$

Minimizing the Kullback-Leibler divergence is equivalent to maximize the ELBO as equality holds if and only if $Q(\boldsymbol{z}) = P(\boldsymbol{z} \mid \boldsymbol{x})$. The ELBO may be written as

$$\begin{aligned}\text{ELBO}(Q) &= \mathbb{E}_{Q(\boldsymbol{z})}\Big[\log P(\boldsymbol{z})\Big] + \mathbb{E}_{Q(\boldsymbol{z})}\Big[\log P(\boldsymbol{x} \mid \boldsymbol{z})\Big] - \mathbb{E}_{Q(\boldsymbol{z})}\Big[\log Q(\boldsymbol{z})\Big] \\ &= \mathbb{E}_{Q(\boldsymbol{z})}\Big[\log P(\boldsymbol{x} \mid \boldsymbol{z})\Big] - KL\Big(Q(\boldsymbol{z}) \mid P(\boldsymbol{z})\Big),\end{aligned}$$

where it is expressed as the sum of the log likelihood of the observations and the Kullback-Leibler divergence between the prior $P(\boldsymbol{z})$ and $Q(\boldsymbol{z})$.

The *expectation maximization algorithm* and *coordinate ascent variational inference* are two algorithms designed to optimize this lower bound in order to solve the optimization problem we are focusing.

1 Energy and Entropy terms come from a statistical physics terminology (Barber (2007))

8

EXPECTATION MAXIMIZATION

The *expectation maximization (EM)* algorithm (McLachlan & Krishnan (2007); Bishop (2006)) is a partially non-Bayesian, likelihoodist iterative method to find maximum likelihood estimates of parameters in statistical models where the model depends on latent variables.

The algorithm consist of a two step iteration where the first optimizes the variational distribution element Q and the second optimizes the set of parameters $\boldsymbol{\theta}$. A fully Bayesian version would consider a probability distribution over the parameter, where the distinction between the two steps disappears. In that case, as many steps as latent variables (including the parameters) are needed per iteration, where each variable is optimized at a time. For *graphical models* this is easy to compute as each variable's new variational distribution depends only on its *Markov blanket*, so local *message passing* can be used for efficient inference (Chapter 19.2).

Using the same notation as the previous characters, EM's iterative procedure increases the ELBO for the parametric marginal $\log P(\boldsymbol{x} \mid \boldsymbol{\theta})$,

$$\log P(\boldsymbol{x} \mid \boldsymbol{\theta}) \geq \underbrace{-\mathbb{E}_{Q(\boldsymbol{z})}\Big[\log Q(\boldsymbol{z})\Big]}_{\text{Entropy}} + \underbrace{\mathbb{E}_{Q(\boldsymbol{z})}\Big[\log P(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta})\Big]}_{\text{Energy}},$$

and the marginal likelihood $P(\boldsymbol{x} \mid \boldsymbol{\theta})$ itself. Which means that aims for the value of $\boldsymbol{\theta}$ to which the dataset better fits the model, i.e, the maximum likelihood parameter.

The EM algorithm can be viewed as two alternating maximization steps, that is, as an example of *coordinate ascent* (Neal & Hinton (1998)). Consider the above ELBO as a function of Q and $\boldsymbol{\theta}$:

$$ELBO(Q, \boldsymbol{\theta}) = -\mathbb{E}_{Q(\boldsymbol{z})}\Big[\log Q(\boldsymbol{z})\Big] + \mathbb{E}_{Q(\boldsymbol{z})}\Big[\log P(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta})\Big].$$

Then, the EM algorithm consists on:

- For fixed $\boldsymbol{\theta}$, choose Q such as:

$$Q^{new} = \arg\max_{Q} ELBO(Q, \boldsymbol{\theta}),$$

  which is equivalent to:

$$Q^{new}(\boldsymbol{z}) = P(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta}).$$

- For fixed Q, choose $\boldsymbol{\theta}$ such as:

$$\boldsymbol{\theta}^{new} = \arg\max_{\boldsymbol{\theta}} ELBO(Q, \boldsymbol{\theta}).$$

Algorithm 1: Expectation Maximization Algorithm

Data: A dataset $\boldsymbol{x}$ and a distribution $P(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta})$.
Result: Approximation of the maximum likelihood parameter.
Initialize $\boldsymbol{\theta}^{(0)}$;
while *Convergence stop criteria* **do**
  $Q^{(t)} = P(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta}^{(t-1)})$;
  $\boldsymbol{\theta}^{(t)} = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{Q^{(t)}(\boldsymbol{z})}\Big[\log P(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta})\Big]$;
end
return θ;

As Q does not depend on the parameter, this is equivalent to maximize the energy term:

$$\boldsymbol{\theta}^{new} = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{Q(\boldsymbol{z})}\Big[\log P(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta})\Big].$$

In the specific situation where $\boldsymbol{x}$ consists on N observations of the same variable X and each observation of X is related with a single hidden variable Z, for example, a mixture distribution, the following considerations might be taken:

- Observations are treated as observations of i.i.d variables $X_1, \ldots, X_N$ and $Z_1, \ldots, Z_N$.
- The variational distribution Q now factorizes over the hidden variables as it is known that they are independent from each other:

  $$Q(\boldsymbol{z}) = \prod_{n=1}^{N} Q(z_n).$$

  Notice that the same letter Q is being used for each variable Z_n, this is just notational, in practice there must be a variational distribution for each of these i.i.d variables.
- The model distribution P factorizes over the variables as:

  $$P(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta}) = \prod_{n=1}^{N} P(x_n, z_n \mid \boldsymbol{\theta}).$$

- The lower bound is then written as

  $$\begin{aligned}\log P(x_1, \ldots, x_N \mid \boldsymbol{\theta}) &= \sum_{n=1}^{N} \log P(x_n \mid \boldsymbol{\theta}) \\ &\geq \sum_{n=1}^{N} -\mathbb{E}_{Q(z_n)}\Big[\log Q(z_n)\Big] + \mathbb{E}_{Q(z_n)}\Big[\log P(x_n, z_n \mid \boldsymbol{\theta})\Big].\end{aligned}$$

  Where equality holds if and only if $Q(z_n) = P(z_n \mid x_n, \boldsymbol{\theta})\ \forall n = 1, \ldots, N$.

The procedure to optimize the parameter consists in two steps:

- **E-step**. For fixed θ, find the distributions that maximize the above bound, i.e, choose $Q^{new}(z_n) = P(z_n \mid x_n, \theta)\ \forall n = 1, \ldots, N$.
- **M-step**. For a fixed distribution Q, find the parameter θ that maximizes the bound:

  $$\theta^{new} = \arg\max_{\theta} \sum_{n=1}^{N} \mathbb{E}_{Q(\boldsymbol{z})}\Big[\log P(z_n, x_n \mid \theta)\Big].$$

Example 2. Consider a single variable X with $Dom(X) = \mathbb{R}$ and a single hidden variable Z with $dom(Z) = \{1, 2\}$. Let the conditional probability be:

$$P(x \mid z, \theta) = \frac{1}{\sqrt{\pi}} e^{-(x-\theta z)^2},$$

and $P(Z = 1) = P(Z = 2) = 0.5$. Suppose an observation $x = 2.75$ we want to optimize the parameter θ in the marginal

$$P(X = 2.75 \mid \theta) = \sum_{z \in \{1,2\}} P(X = 2.75 \mid z, \theta) P(z) = \frac{1}{2\sqrt{\pi}} \left(e^{-(2.75-\theta)^2} + e^{-(2.75-2\theta)^2} \right).$$

Then using a distribution $Q(z)$, the lower bound given to the log likelihood is

$$\log P(x \mid \theta) \geq -Q(1) \log Q(1) - Q(2) \log Q(2) - \mathbb{E}_Q \left[(x - \theta z)^2 \right] + \text{const.}$$

The M-step can be done analytically, noticing that due to the negative sign we want to minimize $\mathbb{E}_Q \left[(x - \theta z)^2 \right]$

$$\frac{d}{d\theta} \mathbb{E}_Q \left[(x - \theta z)^2 \right] = \mathbb{E}_Q \left[2xz + 2\theta z^2 \right] = 2x \mathbb{E}_Q \left[z \right] + 2\theta \mathbb{E}_Q \left[z^2 \right] = 0 \iff \theta = \frac{x \mathbb{E}_Q \left[z \right]}{\mathbb{E}_Q \left[z^2 \right]},$$

$$\frac{d^2}{d^2\theta} \mathbb{E}_Q \left[(x - \theta z)^2 \right] = 2 \mathbb{E}_Q \left[z^2 \right] \geq 0,$$

so the new parameter optimal parameter is

$$\theta_{new} = x \frac{\mathbb{E}_Q \left[z \right]}{\mathbb{E}_Q \left[z^2 \right]}.$$

The E-step would set $Q_{new}(z) = P(z \mid x, \theta)$, in this case

$$Q^{new}(Z = 1) = \frac{P(X = 2.75 \mid Z = 1, \theta) P(Z = 1)}{P(X = 2.75)} = \frac{e^{-(2.75-\theta)}}{e^{-(2.75-\theta)} + e^{-(2.75-2\theta)}},$$
$$Q^{new}(Z = 2) = \frac{P(X = 2.75 \mid Z = 2, \theta) P(Z = 2)}{P(X = 2.75)} = \frac{e^{-(2.75-2\theta)}}{e^{-(2.75-\theta)} + e^{-(2.75-2\theta)}}.$$

8.1 EM INCREASES THE MARGINAL LIKELIHOOD

It is clear that the EM algorithm does increase the lower bound in each iteration but also the marginal likelihood.

Proposition 5. *The log likelihood $P(x \mid \boldsymbol{\theta})$ is not decreased in each iteration of the EM algorithm, i.e, if $\boldsymbol{\theta}^{old}$ and $\boldsymbol{\theta}^{new}$ are two consecutive values of the parameter, EM verifies:*

$$\log P(\boldsymbol{x} \mid \boldsymbol{\theta}^{new}) - \log P(\boldsymbol{x} \mid \boldsymbol{\theta}^{old}) \geq 0.$$

Proof. As a result of the E-step Q is set to

$$Q(\boldsymbol{z}) = P(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta}^{old}).$$

So the lower bound in terms of $\boldsymbol{\theta}^{old}$ and $\boldsymbol{\theta}^{new}$ is

$$ELBO(\boldsymbol{\theta}^{new} \mid \boldsymbol{\theta}^{old}) = \underbrace{-\mathbb{E}_{P(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\theta}^{old})}\Big[\log P(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta}^{old})\Big]}_{\text{Entropy}} + \underbrace{\mathbb{E}_{P(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\theta}^{old})}\Big[\log P(\boldsymbol{z}, \boldsymbol{x} \mid \boldsymbol{\theta}^{new})\Big]}_{\text{Energy}}.$$

From the definition of the Kullback-Leibler divergence we get that

$$\log P(\boldsymbol{x} \mid \boldsymbol{\theta}^{new}) = ELBO(\boldsymbol{\theta}^{new} \mid \boldsymbol{\theta}^{old}) + KL\Big(P(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta}^{old}) \mid P(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta}^{new})\Big).$$

We could use $\boldsymbol{\theta}^{old}$ in the above formula getting

$$\log P(\boldsymbol{x} \mid \boldsymbol{\theta}^{old}) = ELBO(\boldsymbol{\theta}^{old} \mid \boldsymbol{\theta}^{old}) + KL\Big(P(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta}^{old}) \mid P(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta}^{old})\Big) = ELBO(\boldsymbol{\theta}^{old} \mid \boldsymbol{\theta}^{old}).$$

So we can compute the difference between the log likelihood between two consecutive iterations as

$$\begin{aligned}\log P(\boldsymbol{x} \mid \boldsymbol{\theta}^{new}) - \log P(\boldsymbol{x} \mid \boldsymbol{\theta}^{old}) = {} & ELBO(\boldsymbol{\theta}^{new} \mid \boldsymbol{\theta}^{old}) - ELBO(\boldsymbol{\theta}^{old} \mid \boldsymbol{\theta}^{old}) \\ & + KL\Big(P(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta}^{old}) \mid P(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta}^{new})\Big).\end{aligned}$$

Where we know the last term is always positive. About the difference of bounds, the M-step ensures the new parameter makes the lower bound higher or equal to the current one, so that difference is also positive.

$$\log P(\boldsymbol{x} \mid \boldsymbol{\theta}^{new}) - \log P(\boldsymbol{x} \mid \boldsymbol{\theta}^{old}) \geq 0.$$

□

8.2 EXAMPLE: BINOMIAL MIXTURE

In this section we are using a coin-flipping experiment as an example to show limitations of classical maximum likelihood inference due to the presence of hidden variables. The EM algorithm is then used to surpass these limitations. The example consists in a *mixture distribution* based on Do & Batzoglou (2008).

The experiment consist of randomly choosing one of a pair of coins A and B with unknown biases, θ_A and θ_B. Let 1 denote *heads* and 0 denote *tails*. The selected coin is tossed M times, repeating this whole procedure N times. In short, the experiment is governed by two parameters:

$$P(A = 1) = \theta_A \quad \text{and} \quad P(B = 1) = \theta_B,$$

Maximum likelihood training attempts to infer the value of $\boldsymbol{\theta} = (\theta_A, \theta_B)$ that maximizes the likelihood.

Given that the maximum likelihood distribution is the empirical distribution:

$$\theta_A^{ML} = \frac{\text{Heads of coin } A}{\text{Total flips of coin } A} \quad \text{and} \quad \theta_B^{ML} = \frac{\text{Heads of coin } B}{\text{Total flips of coin } B}.$$

Consider now that the identity of the coin that is being flipped is unknown. Let X be the random variable modeling the number of heads in M flips and Z the coin being flipped. In this situation, using the empirical distribution is not possible as the identity of the coin is unknown. In this kind of situations, the EM algorithm performs maximum likelihood training while dealing with the hidden variable.

The conditional $x_n \mid z_n, \boldsymbol{\theta}$ follows a Bernoulli distribution:

$$x_n \mid z_n, \boldsymbol{\theta} \sim B(M, \theta_{z_n}) \implies P(x_n \mid z_n, \boldsymbol{\theta}) = \binom{M}{x_n} \theta_{z_n}^{x_n} (1 - \theta_{z_n})^{M - x_n} \; \forall n = 1, \dots, N,$$

and the probability of choosing a coin is $P(z_n = A) = P(z_n = B) = 0.5 \; \forall n = 1, \dots, N$.

The lower bound is:

$$\sum_{n=1}^{N} \log P(x_n \mid \boldsymbol{\theta}) \geq \sum_{n=1}^{N} -\mathbb{E}_{Q(z_n)}\Big[\log Q(z_n)\Big] + \mathbb{E}_{Q(z_n)}\Big[\log P(z_n, x_n \mid \boldsymbol{\theta})\Big],$$

The **E-step** consists on setting (given a fixed θ, which is not considered a random variable),

$$\begin{aligned} Q(z_n) = P(z_n \mid x_n, \boldsymbol{\theta}) &= \frac{P(x_n \mid z_n, \boldsymbol{\theta}) P(z_n)}{P(x_n)} \\ &= \frac{0.5 \; \theta_{z_n}^{x_n} (1 - \theta_{z_n})^{M - x_n}}{\int_{z_n} P(x_n, z_n \mid \boldsymbol{\theta})} \\ &= \frac{P(z_n) \theta_{z_n}^{x_n} (1 - \theta_{z_n})^{M - x_n}}{\int_{z_n, \boldsymbol{\theta}} P(x_n \mid z_n, \boldsymbol{\theta}) P(z_n)} \\ &= \frac{\theta_{z_n}^{x_n} (1 - \theta_{z_n})^{M - x_n}}{\theta_A^{x_n} (1 - \theta_A)^{M - x_n} + \theta_B^{x_n} (1 - \theta_B)^{M - x_n}} \quad \forall n = 1, \dots, N. \end{aligned}$$

On the other hand, the **M-step** consists on setting

$$\begin{aligned} \boldsymbol{\theta}^{new} &= \arg\max_{\theta} \sum_{n=1}^{N} \mathbb{E}_{Q(z_n)}\Big[\log P(x_n, z_n \mid \boldsymbol{\theta})\Big] \\ &= \arg\max_{\theta} \sum_{n=1}^{N} \mathbb{E}_{Q(z_n)}\Big[\log P(x_n \mid z_n, \boldsymbol{\theta}) P(z_n)\Big] \\ &= \arg\max_{\theta} \sum_{n=1}^{N} \mathbb{E}_{Q(z_n)}\Big[\log P(x_n \mid z_n, \boldsymbol{\theta})\Big] \end{aligned}$$

where in the last equality we used that $P(z_n) = 0.5$ is constant. The expectation is written as

$$\begin{aligned} \mathbb{E}_{Q(z_n)}\Big[\log P(x_n \mid z_n, \boldsymbol{\theta})\Big] = Q(z_n = A)\Big(x_n \log \theta_A + (M - x_n) \log (1 - \theta_A)\Big) \\ + Q(z_n = B)\Big(x_n \log \theta_B + (M - x_n) \log (1 - \theta_B)\Big). \end{aligned}$$

As θ_A and θ_B are separated in each term, the optimum can be found separately, the term affected by θ_A is

$$\sum_{n=1}^{N} Q(z_n = A)\Big(x_n \log \theta_A + (M - x_n) \log (1 - \theta_A)\Big),$$

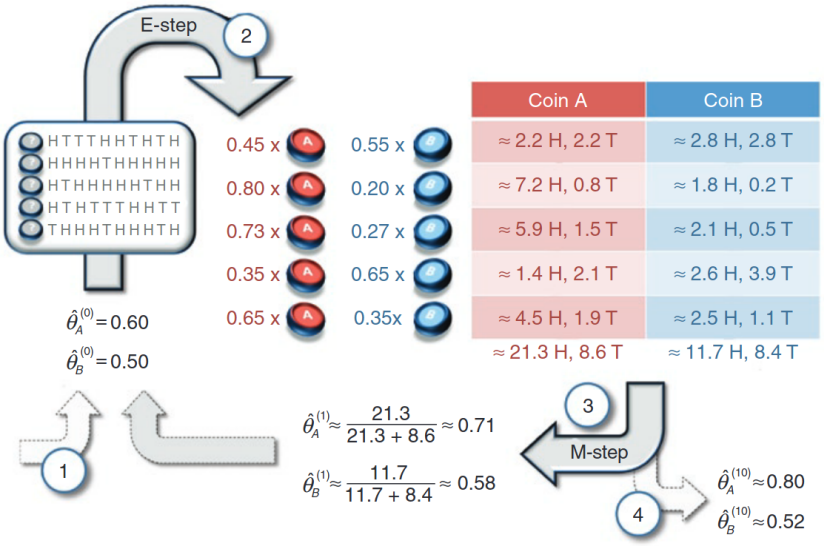


Figure 4: EM algorithm step in Mixture Do & Batzoglou (2008)

deriving and setting to 0 we get the following maxima.

$$\sum_{n=1}^{N} Q(z_n = A)\left(\frac{x_n}{\theta_A} - \frac{M - x_n}{1 - \theta_A}\right) = 0 \iff \theta_A = \sum_{n=1}^{N} Q(z_n = A)\frac{x_n}{M}.$$

The same argument is valid for θ_B.

We are know using the figure 4 to set the example in a numerical context. We are following the given data points in the diagram.

1. We are considering a prior $\boldsymbol{\theta} = (0.6, 0.5)$, a set of $N = 5$ samples and $M = 10$ flips per sample.
2. In the **E-step**, we calculate each $Q(z_n)$.

$$Q(z_1 = A) = \frac{0.6^5 0.4^5}{0.6^5 0.4^5 + 0.5^5 0.5^5} \approx 0.45 \implies Q(z_1 = B) \approx 0.55,$$
$$\vdots$$
$$Q(z_5 = A) \approx 0.65 \implies Q(z_5 = B) \approx 0.35.$$

3. The **M-step** is then:

$$\theta_A = \sum_{n=1}^{5} Q(z_n)\frac{x_n}{10} \approx \frac{21.3}{21.3 + 8.6} \approx 0.71,$$
$$\theta_B \approx 0.58.$$

4. Step 4 represents a possible solution after 10 iterations.

8.3 PARTIAL STEPS

Making a partial M-step consist on not using the optimal parameter for the energy term, but using one with just higher energy. Finding this values can be easier than finding the optimal

one and convergence still follows as the only requirement to make the likelihood increase was to increase the lower bound.

On the other hand, when studying the increase on the likelihood, we supposed that the optimal E-step was being used. It cannot be guarantee that a partial step would increase the likelihood in this case, even though it does increase the lower bound.

Another important factor is, that the EM algorithm assumes that the energy term is possible to calculate, which may not be. As an approach to solve this situation, we can set a class of distributions $\mathcal{Q}$, and minimize the Kullback-Leibler divergence between $P(\boldsymbol{z} \mid \boldsymbol{x}, \theta)$ and a distribution $Q \in \mathcal{Q}$, so we pick a distribution such that

$$Q = \arg\min_{Q \in \mathcal{Q}} KL\Big(Q(\boldsymbol{z}) \mid P(\boldsymbol{z} \mid \boldsymbol{x}, \theta)\Big).$$

An extreme case is to choose $\mathcal{Q}$ as delta functions, where the energy term is now a constant, and the optimal setting is

$$Q(z_n) = \delta(z_n, z_n^{opt}), \quad z_n^{opt} = \arg\max_z P(z, x_n \mid \theta).$$

This is called *Viterbi training* and does not guarantee that the log likelihood is being increased in each iteration.

9

MEAN-FIELD VARIATIONAL INFERENCE

9.1 THE MEAN-FIELD VARIATIONAL FAMILY

The *mean-field variational family* $\mathcal{Q}$ is defined as the family of distributions where the variables are mutually independent, i.e, any $Q \in \mathcal{Q}$ verifies

$$Q(\boldsymbol{z}) = \prod_{m=1}^{M} Q_m(z_m),$$

where $\boldsymbol{Z} = \{Z_1, \ldots, Z_M\}$ is the considered set of variables. The mean-field family is commonly used to model the family of distributions over the latent variables in our optimization problem. Notice that each *factor* Q_m can be different and this family does not depend on the observed data.

The mean-field family can capture any marginal of the latent variables but not correlation between them, as it assumes they are independent. For example, consider a two dimensional Gaussian distribution where a high percentage of the density is inside the blue ellipse shown in the following figure. Any mean-field approximation would factorize as a product of two Gaussian distributions, condensing its density in a circle as shown in purple.

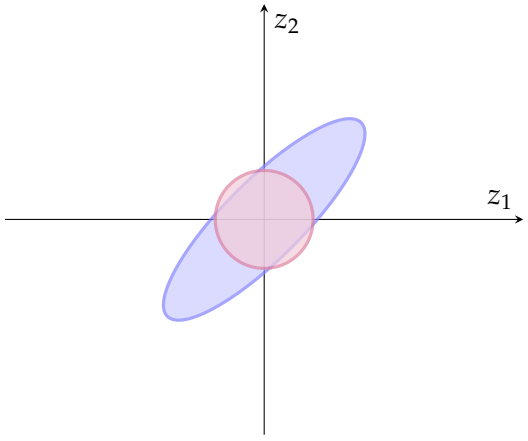


Figure 5: Mean-field family distribution (purple) approximating a Gaussian distribution (blue)

Notice the parametric form of each factor Q_m is not specified and the appropriate configuration depends on the variable. For example, a continuous variable might have a Gaussian factor and a categorical variable have a categorical factor.

Algorithm 2: Coordinate Ascent Variational Inference

Data: A distribution $P(\boldsymbol{x}, \boldsymbol{z})$ with a dataset $\boldsymbol{x}$.
Result: A distribution of the mean-field family $Q(\boldsymbol{z}) = \prod_{m=1}^{M} Q_m(z_m)$
Initialize $Q(\boldsymbol{z})$;
while *Convergence stop criteria* **do**
  for $m \in 1, \dots, M$ **do**
    $\boldsymbol{z}_{\setminus m} = (z_1, \dots, z_{m-1}, z_{m+1}, \dots, z_M)$;
    Set $Q_m(z_m) \propto \exp \mathbb{E}_{Q_{\setminus m}}\Big[\log P(z_m \mid \boldsymbol{z}_{\setminus m}, \boldsymbol{x})\Big]$;
  end
  Compute $ELBO(Q)$; // Used for convergence criteria.
end
return Q;

9.2 CAVI ALGORITHM

In this section, we describe a widely used algorithm to solve the optimization problem we discussed in the previous section using the mean-field family. It is *coordinate ascent variational inference* or *CAVI* (also known as *Variational Bayes*) and its procedure is to iteratively optimize each factor of the mean-field family distribution, while fixing the others. With this, the ELBO reaches a local optimum.

CAVI might be seen as a generalization of the EM algorithm, where model parameters are considered random hidden variables.

Let $\boldsymbol{x}$ be the given observations of the observed variables. CAVI iterates fixing all hidden variable but one at a time and maximizing its contribution to the ELBO. Consider the m^{th} variable Z_m, denoting by $\setminus m$ full set of indexes without the m^{th}, then $\boldsymbol{Z}_{\setminus m}$ is the full set of variables without the focused one. Let the factors $Q_n, n \neq m$ be fixed.

The contribution of Z_m to the ELBO is (summarizing other factors in the constant term):

$$
\begin{aligned}
ELBO(Q) &= \mathbb{E}_Q\Big[\log P(\boldsymbol{x}, \boldsymbol{z})\Big] - \mathbb{E}_Q\Big[\log Q(\boldsymbol{z})\Big] \\
&\overset{1}{=} \mathbb{E}_{Q_m}\Big[\mathbb{E}_{Q_{\setminus m}}\Big[\log P(\boldsymbol{x}, \boldsymbol{z})\Big]\Big] - \mathbb{E}_{Q_m}\Big[\log Q_m(z_m)\Big] + \text{ const.} \\
&\overset{2}{=} \mathbb{E}_{Q_m}\Big[\mathbb{E}_{Q_{\setminus m}}\Big[\log P(z_m \mid \boldsymbol{z}_{\setminus m}, \boldsymbol{x}) + \log P(\boldsymbol{z}_{\setminus m}, \boldsymbol{x})\Big]\Big] - \mathbb{E}_{Q_m}\Big[\log Q_m(z_m)\Big] + \text{ const.} \\
&\overset{3}{=} \mathbb{E}_{Q_m}\Big[\mathbb{E}_{Q_{\setminus m}}\Big[\log P(z_m \mid \boldsymbol{z}_{\setminus m}, \boldsymbol{x})\Big]\Big] - \mathbb{E}_{Q_m}\Big[\log Q_m(z_m)\Big] + \text{ const.} \\
&\overset{4}{=} -KL\Big(Q_m(z_m) \mid \exp \mathbb{E}_{Q_{\setminus m}}\Big[\log P(z_m \mid \boldsymbol{z}_{\setminus m}, \boldsymbol{x})\Big]\Big) + \text{ const..}
\end{aligned}
$$

1. The expectations in the ELBO formula are separated. The logarithm factorizes as $\log Q(z) = \sum_{m=1}^{M} \log Q_m(z_m)$. The constant term comes from $\mathbb{E}_{Q_{\setminus m}}\Big[\log Q_{\setminus m}(z_{\setminus m})\Big]$.
2. P is separated as $P(\boldsymbol{z}, \boldsymbol{x}) = P(z_m \mid \boldsymbol{z}_{\setminus m}, \boldsymbol{x}) P(\boldsymbol{z}_{\setminus m}, \boldsymbol{x}) \implies \log P(\boldsymbol{z}, \boldsymbol{x}) = \log P(z_m \mid \boldsymbol{z}_{\setminus m}, \boldsymbol{x}) + \log P(\boldsymbol{z}_{\setminus m}, \boldsymbol{x})$.
3. $\mathbb{E}_{Q_m}\Big[\mathbb{E}_{Q_{\setminus m}}\Big[\log P(\boldsymbol{z}_{\setminus m}, \boldsymbol{x})\Big]\Big] = \mathbb{E}_{Q_{\setminus m}}\Big[\log P(\boldsymbol{z}_{\setminus m}, \boldsymbol{x})\Big]$ is constant.
4. Applied Kullback-Leibler definition.

Maximizing the ELBO is equivalent to minimize the given Kullback-Leibler divergence and this divergence is zero when Q_m^{new} is:

$$Q_m^{new}(z_m) \propto \exp \mathbb{E}_{Q_{\setminus m}}\Big[\log P(z_m \mid \boldsymbol{z}_{\setminus m}, \boldsymbol{x})\Big].$$

Notice that the proportionality restriction is enough to fully determine the distribution as its integral is normalized. Equivalently, the distribution is proportional to

$$Q_m^{new}(z_m) \propto \exp \mathbb{E}_{Q_{\setminus m}}\Big[\log P(z_m, \boldsymbol{z}_{\setminus m}, \boldsymbol{x})\Big]. \tag{1}$$

As the ELBO is generally a non-convex function and the CAVI algorithm converges to a local optimum, the initialization values of the algorithm play an important role on its performance. The convergence criteria is usually a threshold for the ELBO.

9.3 CAVI AS AN EM GENERALIZATION

As CAVI considers parameters as hidden variables, we need to specify a variational distribution for them. We are choosing the distribution that summarizes the information in the optimal point, let $\boldsymbol{\theta}_{opt}$ be the optimal value of $\boldsymbol{\theta}$:

$$Q(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_{opt}).$$

The variational distribution factorizes as

$$Q(\boldsymbol{z}, \boldsymbol{\theta}) = Q(\boldsymbol{z})Q(\boldsymbol{\theta}).$$

The lower bound takes the form

$$\begin{aligned}
\log P(\boldsymbol{x} \mid \boldsymbol{\theta}) &\geq \mathbb{E}_{Q(\boldsymbol{z},\boldsymbol{\theta})}\Big[\log P(\boldsymbol{x}, \boldsymbol{z}, \theta)\Big] - \mathbb{E}_{Q(\boldsymbol{z},\boldsymbol{\theta})}\Big[\log Q(\boldsymbol{z}, \boldsymbol{\theta})\Big] \\
&= \mathbb{E}_{Q(\boldsymbol{z})}\Big[\mathbb{E}_{Q(\boldsymbol{\theta})}\Big[\log P(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\theta})\Big]\Big] - \mathbb{E}_{Q(\boldsymbol{z})}\Big[\log Q(\boldsymbol{z})\Big] - \mathbb{E}_{Q(\boldsymbol{\theta})}\Big[\log Q(\boldsymbol{\theta})\Big] \\
&= \mathbb{E}_{Q(\boldsymbol{z})}\Big[\log P(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\theta}_{opt})\Big] - \mathbb{E}_{Q(\boldsymbol{z})}\Big[\log Q(\boldsymbol{z})\Big] + \text{ const.}
\end{aligned}$$

The CAVI update can be seen as an iterative two step process. Firstly, given a fixed $Q(\boldsymbol{z})$, as the distribution class of $Q(\boldsymbol{\theta})$ is fixed, optimizing it is equivalent to find the optimal parameter $\boldsymbol{\theta}_{opt}$.

$$\begin{aligned}
\boldsymbol{\theta}_{opt} &= \arg\max_{\boldsymbol{\theta}} \Big(\mathbb{E}_{Q(\boldsymbol{z})}\Big[\log P(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\theta})\Big]\Big) \\
&= \arg\max_{\boldsymbol{\theta}} \Big(\mathbb{E}_{Q(\boldsymbol{z})}\Big[\log P(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta})P(\boldsymbol{\theta})\Big]\Big) \\
&= \arg\max_{\boldsymbol{\theta}} \Big(\mathbb{E}_{Q(\boldsymbol{z})}\Big[\log P(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta})\Big] + \log P(\boldsymbol{\theta})\Big).
\end{aligned}$$

If we take a flat prior, this term is equivalent to the M-step. Secondly, given a fixed parameter $\boldsymbol{\theta}$, the update is

$$Q(\boldsymbol{z}) \propto \exp \mathbb{E}_{Q(\boldsymbol{\theta})}\Big[\log P(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta})\Big] = P(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta}_{opt}),$$

which is the E-step of the EM algorithm.

10

EXPONENTIAL FAMILY

The exponential family (Koopman (1936)) is a parametric set of probability distributions of a certain form. This form is chosen based on some useful algebraic properties and generality, Appendix A shows some examples of common distributions in the exponential family.

Let X be a random variable and $\boldsymbol{\theta}$ a set of parameters. A family of distributions is said to belong the exponential family if its probability distribution has the form

$$P(x \mid \boldsymbol{\theta}) = h(x) \exp\Big(\sum_{i=1}^{S} \eta_i(\boldsymbol{\theta}) T_i(x) - \psi(\boldsymbol{\theta}) \Big),$$

where $h(x)$, $T_i(x)$, $\eta_i(\boldsymbol{\theta})$ and $\psi(\boldsymbol{\theta})$ are known functions such that h is called a *base measure*, $\eta_i(\boldsymbol{\theta})$ are called the *distribution parameters*, $T_i(x)$ the *test statistics* and ψ is the *log normalizer* as it ensures logarithmic normalization due to

$$\begin{aligned} 1 &= \int_x h(x) \exp\Big(\sum_{i=1}^{S} \eta_i(\boldsymbol{\theta}) T_i(x) - \psi(\boldsymbol{\theta}) \Big) \\ &= \int_x e^{-\psi(\boldsymbol{\theta})} h(x) \exp\Big(\sum_{i=1}^{S} \eta_i(\boldsymbol{\theta}) T_i(x) \Big) \\ &= e^{-\psi(\boldsymbol{\theta})} \int_x h(x) \exp\Big(\sum_{i=1}^{S} \eta_i(\boldsymbol{\theta}) T_i(x) \Big), \end{aligned}$$

so ψ verifies

$$\psi(\boldsymbol{\theta}) = \log \int_x h(x) \exp\Big(\sum_{i=1}^{S} \eta_i(\boldsymbol{\theta}) T_i(x) \Big).$$

Naming $\boldsymbol{\eta}$ and $\boldsymbol{T}$ the corresponding vector functions, the parameters can always be transformed as $\boldsymbol{\theta}^{new} = \boldsymbol{\eta}(\boldsymbol{\theta})$, in which case we say the distribution into its *canonical form* (notice ψ has changed but we are not differentiating it from the previous one since it is fully determined by the other functions):

$$P(x \mid \boldsymbol{\theta}) = h(x) e^{\boldsymbol{\theta}^T \boldsymbol{T}(x) - \psi(\boldsymbol{\theta})}.$$

In this form, it is easier to see that $\boldsymbol{T}$ is the sufficient statistic for $\boldsymbol{\theta}$. This is a consequence of *Fisher–Neyman factorization theorem* which says that $\boldsymbol{T}$ *is sufficient for* $\boldsymbol{\theta}$ *if and only if the probability distribution* P *can be factored into a product such that one factor,* h*, does not depend on* $\boldsymbol{\theta}$ *and the other factor, which does depend on* $\boldsymbol{\theta}$*, depends on* x *only through* $\boldsymbol{T}$*.*

An important property of the exponential family is that they have *conjugate priors*, this is said when the posterior distribution is in the same probability distribution family as the prior distribution, they are then called *conjugate distributions*, and the prior is called a *conjugate prior* of the likelihood distribution.

Proposition 6. *Let X be a random variable and $\boldsymbol{\theta}$ a set of parameters. Suppose an exponential family likelihood:*

$$P(x \mid \boldsymbol{\theta}) = h(x)e^{\boldsymbol{\theta}^T \boldsymbol{T}(x) - \psi(\boldsymbol{\theta})}.$$

and prior with hyper-parameters $\boldsymbol{\alpha}, \gamma$:

$$P(\boldsymbol{\theta} \mid \boldsymbol{\alpha}, \gamma) \propto e^{\boldsymbol{\theta}^T \boldsymbol{\alpha} - \gamma\psi(\boldsymbol{\theta})}.$$

Then, the posterior is in the same parametric family as the prior with

$$P(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{\alpha}, \gamma) = P(\boldsymbol{\theta} \mid \boldsymbol{\alpha} + \boldsymbol{T}(x), \gamma + 1).$$

Proof.

$$P(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{\alpha}, \gamma) \propto P(x \mid \boldsymbol{\theta}) P(\boldsymbol{\theta} \mid \boldsymbol{\alpha}, \gamma) \propto \exp\Big(\boldsymbol{\theta}^T [\alpha + \boldsymbol{T}(x)] - [\gamma + 1]\psi(\boldsymbol{\theta})\Big).$$

□

10.1 LATENT VARIABLE AND CONDITIONALLY CONJUGATE MODELS

One important case of exponential family are *latent variable models* or *LVMs*. In this models, the following assumptions are made:

1. There set of i.i.d random variables $\boldsymbol{X} = (X_1, \ldots, X_N)$ and a set of observations $\boldsymbol{x} = (x_1, \ldots, x_N)$.

2. Both global $\boldsymbol{\theta}$ and local hidden variables $\boldsymbol{Z} = (Z_1, \ldots, Z_N)$ govern the data. We refer to *global hidden variables* when they affect the whole distribution and *local hidden variables* when they affect only to a subset, in this case each Z_n affects only x_n. Given this, the joint probability is
$$P(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{z}) = P(\boldsymbol{\theta}) \prod_{n=1}^{N} P(z_n, x_n \mid \boldsymbol{\theta}).$$

3. The n^{th} observation x_n and the n^{th} local hidden variable z_n are conditionally independent, given the global variables $\boldsymbol{\theta}$, of all other observations and local hidden variables,
$$P(x_n, z_n \mid \boldsymbol{\theta}, \boldsymbol{x}_{\setminus n}, \boldsymbol{z}_{\setminus n}) = P(x_n, z_n \mid \boldsymbol{\theta}).$$

Remark 4. These models are widely used to discover patterns in data sets (Blei (2014)). LVMs include popular models like *Latent Dirichlet Allocation* models used to uncover the hidden topics in text corpora (Blei *et al.* (2003)), mixture of Gaussian models to discover hidden clusters in data (Bishop (2006)) and probabilistic principal component analysis for dimensionality reduction (Tipping & Bishop (1999)).

One important case of LVMs and exponential family models are *conditionally conjugate models*, where the following assumptions are made.

1. The prior for the global latent variable $P(\boldsymbol{\theta})$ is in the exponential family with an hyper-parameter $\alpha = [\alpha_1, \alpha_2]$, where α_1 is a vector and α_2 is a scalar, and statistics that concatenate the global latent variable and its log normalizer $[\boldsymbol{\theta}, -\psi(\boldsymbol{\theta})]$,
$$P(\boldsymbol{\theta}) = h(\boldsymbol{\theta}) \exp\Big(\alpha^T [\boldsymbol{\theta}, -\psi(\boldsymbol{\theta})] - \psi(\alpha)\Big).$$
2. Each local term $P(z_n, x_n \mid \boldsymbol{\theta})$ is in the exponential family of the form
$$P(z_n, x_n \mid \boldsymbol{\theta}) = h(z_n, x_n) \exp\Big(\boldsymbol{\theta}^T T(z_n, x_n) - \psi(\boldsymbol{\theta})\Big).$$
3. The complete conditional of a local latent variable verifies
$$P(z_n \mid \boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{z}_{\setminus n}) = P(z_n \mid x_n, \boldsymbol{\theta})$$
and is also in the exponential family
$$P(z_n \mid x_n, \boldsymbol{\theta}) = h(z_n) \exp\Big(\eta(\boldsymbol{\theta}, x_n)^T T(z_n) - \psi(\boldsymbol{\theta}, x_n)\Big).$$

Using Proposition 6, the posterior $P(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{z})$ is in the same family with parameter
$$\bar{\alpha} = [\alpha_1 + \sum_{n=1}^{N} T(z_i, x_i), \alpha_2 + N]^T.$$

A step by step reasoning would be:

$$\begin{aligned}
P(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{z}) = \frac{P(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta}) P(\boldsymbol{\theta})}{P(\boldsymbol{x}, \boldsymbol{z})} &\propto P(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta}) P(\boldsymbol{\theta}) = P(\boldsymbol{\theta}) \prod_{n=1}^{N} h(x_n, z_n) \exp\Big(\boldsymbol{\theta}^T T(z_n, x_n) - \psi(\boldsymbol{\theta})\Big) \\
&\propto h(\boldsymbol{\theta}) \exp\Big(\alpha^T [\boldsymbol{\theta}, -\psi(\boldsymbol{\theta})]\Big) \prod_{n=1}^{N} h(x_n, z_n) \exp\Big(\boldsymbol{\theta}^T T(z_n, x_n) - \psi(\boldsymbol{\theta})\Big) \\
&\propto h(\boldsymbol{\theta}) \exp\Big(\alpha^T [\boldsymbol{\theta}, -\psi(\boldsymbol{\theta})] + \sum_{n=1}^{N} \boldsymbol{\theta}^T T(z_n, x_n) - \psi(\boldsymbol{\theta})\Big) \\
&\propto h(\boldsymbol{\theta}) \exp\Big(\alpha_1^T \boldsymbol{\theta} - \alpha_2^T \psi(\boldsymbol{\theta}) - N\psi(\boldsymbol{\theta}) + \sum_{n=1}^{N} T(z_n, x_n)^T \boldsymbol{\theta}\Big) \\
&\propto h(\boldsymbol{\theta}) \exp\Big(\big(\alpha_1 + \sum_{n=1}^{N} T(z_n, x_n)\big)^T \boldsymbol{\theta} - (\alpha_2 + N)^T \psi(\boldsymbol{\theta})\Big).
\end{aligned}$$

10.2 CAVI IN CONDITIONALLY CONJUGATE MODELS

Set aside the conditionally conjugate models and consider the following situation where we fit a distribution $Q(\boldsymbol{z}) = \prod_{n=1}^{N} Q_n(z_n)$ in the mean-field family, using an exponential family distribution for the marginal $P(z_n \mid \boldsymbol{z}_{\setminus n}, \boldsymbol{x})$:
$$P(z_n \mid \boldsymbol{z}_{\setminus n}, \boldsymbol{x}) = h(z_n) \exp\Big(\eta_n(\boldsymbol{z}_{\setminus n}, \boldsymbol{x})^T T(z_n) - \psi(\boldsymbol{z}_{\setminus n}, \boldsymbol{x})\Big).$$

The update of the CAVI algorithm is then given by

$$\begin{aligned} Q(z_n) &\propto \exp \mathbb{E}_{Q(\boldsymbol{z}_{\setminus n})}\Big[\log P(z_n \mid \boldsymbol{z}_{\setminus n}, \boldsymbol{x})\Big] \\ &= h(z_n) \exp\Big(\mathbb{E}_{Q_{\boldsymbol{z}_{\setminus n}}}\Big[\eta_n(\boldsymbol{z}_{\setminus n}, \boldsymbol{x})\Big]^T T(z_n) - \mathbb{E}_{Q_{\boldsymbol{z}_{\setminus n}}}\Big[\psi(\boldsymbol{z}_{\setminus n}, \boldsymbol{x})\Big]\Big) \\ &\propto h(z_n) \exp\Big(\mathbb{E}_{Q_{\boldsymbol{z}_{\setminus n}}}\Big[\eta_n(\boldsymbol{z}_{\setminus n}, \boldsymbol{x})\Big]^T T(z_n)\Big) = h(z_n) e^{v_n^T T(z_n)}, \end{aligned} \tag{2}$$

where

$$v_n = \mathbb{E}_{Q_{\boldsymbol{z}_{\setminus n}}}\Big[\eta_n(\boldsymbol{z}_{\setminus n}, \boldsymbol{x})\Big].$$

Summarizing, the factor is in the exponential family with the same base measure h and updating it is equivalent to setting the distribution parameter η to the expected one of the complete conditional $\mathbb{E}\Big[\eta_n(\boldsymbol{x}, \boldsymbol{z}_{\setminus n})\Big]$. This expression facilitates deriving CAVI algorithm for many complex models.

Going back to the conditionally conjugate models, our variational distribution is in the mean-field family with $Q(\boldsymbol{\theta} \mid \lambda)$ where λ is called the *global variational parameter*, and for each local variable, the distribution is $Q(z_n \mid \gamma_n)$, where γ_n is called a *local variational parameter*:

$$Q(z_n \mid \gamma_n) \propto h(z_n) e^{{\gamma_n}^T T(z_n)},$$

$$Q(\boldsymbol{\theta} \mid \lambda) = h(\boldsymbol{\theta}) \exp\Big(\lambda^T [\boldsymbol{\theta}, -\psi(\boldsymbol{\theta})] - \psi(\lambda)\Big).$$

CAVI iteratively updates each local variational parameter and the global variational parameter.

Following the steps done in 2, the local variational parameter update is

$$\gamma_n^{new} = \mathbb{E}_{Q(\boldsymbol{\theta}|\lambda)}\Big[\eta(\boldsymbol{\theta}, x_n)\Big],$$

In this case, the local hidden variable conditional does not depend on the other local hidden variables, neither other data-points.

The global variational parameter update is calculated as

$$\begin{aligned} Q^{new}(\boldsymbol{\theta} \mid \lambda) &\propto \exp \mathbb{E}_{Q(\boldsymbol{z})}\Big[\log P(\boldsymbol{\theta} \mid \boldsymbol{z}, \boldsymbol{x})\Big] \propto h(\boldsymbol{\theta}) \exp \mathbb{E}_{Q(\boldsymbol{z})}\Big[\big[\lambda_1 + \sum_{n=1}^N T(x_n, z_n), \lambda_2 + N\big]^T [\boldsymbol{\theta}, \psi(\boldsymbol{\theta})]\Big] \\ &\propto h(\boldsymbol{\theta}) \exp\left(\Big[\lambda_1 + \sum_{n=1}^N \mathbb{E}_{Q(\boldsymbol{z})}\Big[T(x_n, z_n)\Big], \lambda_2 + N\Big]^T \Big[\boldsymbol{\theta}, -\psi(\boldsymbol{\theta})\Big]\right) \\ &= Q(\boldsymbol{\theta}, \lambda^{new}). \end{aligned}$$

Then, the variational parameter updated is

$$\lambda^{new} = \left[\lambda_1 + \sum_{n=1}^N \mathbb{E}_{Q(z_n|\gamma_n)}\Big[T(x_n, z_n)\Big], \lambda_2 + N\right].$$

We can compute the ELBO at each iteration up to a constant that does not depend on the variational parameters,

$$ELBO = \Big(\lambda_1 + \sum_{n=1}^{N} \mathbb{E}_{Q(z_n|\gamma_n)}\Big[T(x_n, z_n)\Big]\Big)^T \mathbb{E}_{Q(\boldsymbol{\theta}|\lambda)}\Big[\boldsymbol{\theta}\Big] - (\lambda_2 + N)\mathbb{E}_{Q(\boldsymbol{\theta}|\lambda)}\Big[\psi(\boldsymbol{\theta})\Big]$$
$$+ \lambda^T \mathbb{E}_{Q(\boldsymbol{\theta},\lambda)}\Big[T(\boldsymbol{\theta})\Big] - \psi(\lambda) + \sum_{n=1}^{N} \gamma_n^T \mathbb{E}_{Q(z_n,\gamma_n)}\Big[z_n\Big] - \psi(\gamma_n) + \text{ const.}$$

The calculations are the following:

$$\begin{aligned} ELBO(Q(\boldsymbol{\theta}, \boldsymbol{z})) &= \mathbb{E}_{Q(\boldsymbol{\theta},\boldsymbol{z})}\Big[\log P(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{x})\Big] - \mathbb{E}_{Q(\boldsymbol{\theta},\boldsymbol{z})}\Big[\log Q(\boldsymbol{\theta}, \boldsymbol{z})\Big] \\ &= \mathbb{E}_{Q(\boldsymbol{z}|\boldsymbol{\gamma})}\Big[\mathbb{E}_{Q(\boldsymbol{\theta}|\lambda)}\Big[\log P(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{x})\Big]\Big] - \mathbb{E}_{Q(\boldsymbol{\theta},\boldsymbol{z})}\Big[\log Q(\boldsymbol{\theta}, \boldsymbol{z})\Big] \\ &= \mathbb{E}_{Q(\boldsymbol{\theta}|\lambda)}\Big[\log P(\boldsymbol{\theta})\Big] + \mathbb{E}_{Q(\boldsymbol{z}|\boldsymbol{\gamma})}\Big[\mathbb{E}_{Q(\boldsymbol{\theta}|\lambda)}\Big[\sum_{n=1}^{N} \log P(z_n, x_n \mid \boldsymbol{\theta})\Big]\Big] \\ &\quad - \mathbb{E}_{Q(\boldsymbol{\theta},\boldsymbol{z})}\Big[\log Q(\boldsymbol{\theta}, \boldsymbol{z})\Big] \end{aligned}$$

The first term is

$$\begin{aligned} \mathbb{E}_{Q(\boldsymbol{\theta}|\lambda)}\Big[\log P(\boldsymbol{\theta})\Big] &= \mathbb{E}_{Q(\boldsymbol{\theta}|\lambda)}\Big[\lambda_1 \boldsymbol{\theta} - \lambda_2 \psi(\boldsymbol{\theta}) - \psi(\lambda)\Big] \\ &= \lambda_1 \mathbb{E}_{Q(\boldsymbol{\theta}|\lambda)}\Big[\boldsymbol{\theta}\Big] - \lambda_2 \mathbb{E}_{Q(\boldsymbol{\theta}|\lambda)}\Big[\psi(\boldsymbol{\theta})\Big] - \psi(\lambda). \end{aligned}$$

The middle term is

$$\mathbb{E}_{Q(\boldsymbol{z}|\boldsymbol{\gamma})}\Big[\mathbb{E}_{Q(\boldsymbol{\theta}|\lambda)}\Big[\log P(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{x})\Big]\Big] = \mathbb{E}_{Q(\boldsymbol{z}|\boldsymbol{\gamma})}\Big[\mathbb{E}_{Q(\boldsymbol{\theta}|\lambda)}\Big[\log P(\boldsymbol{\theta}) + \sum_{n=1}^{N} \log P(z_n, x_n \mid \boldsymbol{\theta})\Big]\Big]$$
$$= \mathbb{E}_{Q(\boldsymbol{z}|\boldsymbol{\gamma})}\Big[\mathbb{E}_{Q(\boldsymbol{\theta}|\lambda)}\Big[\log P(\boldsymbol{\theta})\Big]\Big] + \mathbb{E}_{Q(\boldsymbol{z}|\boldsymbol{\gamma})}\Big[\mathbb{E}_{Q(\boldsymbol{\theta}|\lambda)}\Big[\sum_{n=1}^{N} \log P(z_n, x_n \mid \boldsymbol{\theta})\Big]\Big]$$
$$= \Big(\lambda_1 + \sum_{n=1}^{N} \mathbb{E}_{Q(z_n|\gamma_n)}\Big[T(x_n, z_n)\Big]\Big)^T \mathbb{E}_{Q(\boldsymbol{\theta}|\lambda)}\Big[\boldsymbol{\theta}\Big] - (\lambda_2 + N)\mathbb{E}_{Q(\boldsymbol{\theta}|\lambda)}\Big[\eta(\boldsymbol{\theta})\Big].$$

The last term is

$$\begin{aligned} \mathbb{E}_{Q(\boldsymbol{\theta},\boldsymbol{z})}\Big[\log Q(\boldsymbol{\theta}, \boldsymbol{z})\Big] &= \mathbb{E}_{Q(\boldsymbol{\theta})}\Big[\log Q(\boldsymbol{\theta})\Big] + \mathbb{E}_{Q(\boldsymbol{z})}\Big[\sum_{n=1}^{N} \log Q(z_n)\Big] \\ &= \mathbb{E}_{Q(\boldsymbol{\theta})}\Big[\lambda^T T(\boldsymbol{\theta}) - \psi(\lambda)\Big] + \sum_{n=1}^{N} \mathbb{E}_{Q(z_n)}\Big[\gamma_n^T z_n - \psi(\gamma_n)\Big] \\ &= \lambda^T \mathbb{E}_{Q(\boldsymbol{\theta},\lambda)}\Big[T(\boldsymbol{\theta})\Big] - \psi(\lambda) + \sum_{n=1}^{N} \gamma_n^T \mathbb{E}_{Q(z_n,\gamma_n)}\Big[z_n\Big] - \psi(\gamma_n). \end{aligned}$$

11

EXAMPLE: GAUSSIAN MIXTURE

11.1 MODEL STATEMENT

A Gaussian mixture model is a latent variable model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters.

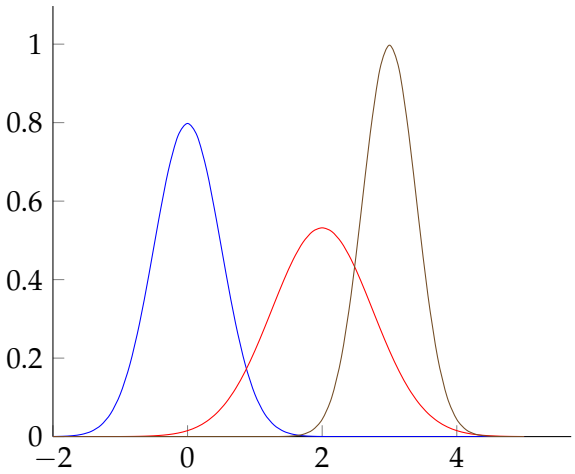


Figure 6: One-dimensional Gaussian mixture with 3 clusters.

The following elements are being considered (using Bishop (2006) notation):

- K mixture components and N observations.
- A set of i.i.d real valued random variables $\boldsymbol{X} = (X_1, \ldots, X_N)$ and a corresponding set of observations $\boldsymbol{x} = (x_1, \ldots, x_n)$.
- The cluster assignment latent variables $\boldsymbol{Z} = (Z_1, \ldots, Z_N)$, where each z_n is a vector full of zeros but a position with a one, indicating the cluster to which x_n belongs.
- We choose a Dirichlet distribution over the mixing coefficients $\boldsymbol{\pi}$

$$\boldsymbol{\pi} \sim \text{Symmetric-Dirichlet}(\alpha_0) \implies P(\boldsymbol{\pi}) \propto \prod_{k=1}^{K} \pi_k^{\alpha_0 - 1}.$$

The hyper-parameter α_0 is the effective prior of each mixture component. Then $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)$ are the mixture weights, i.e, prior probability of a particular component k.

$$P(\boldsymbol{z} \mid \boldsymbol{\pi}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{n,k}} \implies (Z_n \mid \boldsymbol{\pi}) \sim \text{Categorical}(\boldsymbol{\pi}).$$

- $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$ and $\boldsymbol{\Lambda} = (\Lambda_1, \dots, \Lambda_K)$ are the distribution parameters of each observation full conditional

$$(\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \sim \prod_{n=1}^{N} \prod_{k=1}^{K} \mathcal{N}(\mu_k \mid \Lambda_k)^{z_{n,k}}.$$

- The prior governing $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$ is an independent Gaussian, Inverse-Gamma distribution with hyper parameters m_0, β_0, w_0 and v_0:

$$P(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = P(\boldsymbol{\mu} \mid \boldsymbol{\Lambda}) P(\boldsymbol{\Lambda}) \text{ where } \begin{cases} (\boldsymbol{\mu} \mid \boldsymbol{\Lambda}) & \sim \prod_{k=1}^{K} \mathcal{N}(m_0, \beta_0 \Lambda_k) \\ \boldsymbol{\Lambda} & \sim \prod_{k=1}^{K} \text{Inverse-Gamma}(w_0, v_0) \end{cases}.$$

The joint probability factorizes as

$$P(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = P(\boldsymbol{x} \mid \boldsymbol{z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) P(\boldsymbol{z} \mid \boldsymbol{\pi}) P(\boldsymbol{\pi}) P(\boldsymbol{\mu} \mid \boldsymbol{\Lambda}) P(\boldsymbol{\Lambda}).$$

Bishop (2006) gives the explicit update for the CAVI algorithm, in the following section we summarize the needed steps to reach the update for one factor of the variational distribution.

11.2 VARIATIONAL DISTRIBUTION AND CAVI UPDATE

We consider a variational distribution in the mean-field family,

$$Q(\boldsymbol{z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = Q(\boldsymbol{z}) Q(\boldsymbol{\pi}) \prod_{k=1}^{K} Q(\mu_k) Q(\Lambda_k).$$

Let us consider the update for $Q(\boldsymbol{z})$, using the update given in 1:

$$Q^{new}(\boldsymbol{z}) \propto \exp \mathbb{E}_{Q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})} \Big[\log P(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \Big].$$

Which implies

$$\begin{aligned} \log Q^{new}(\boldsymbol{z}) &= \mathbb{E}_{Q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})} \Big[\log P(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \Big] + \text{const.} \\ &= \mathbb{E}_{Q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})} \Big[\log \left(P(\boldsymbol{x} \mid \boldsymbol{z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) P(\boldsymbol{z} \mid \boldsymbol{\pi}) P(\boldsymbol{\pi}) P(\boldsymbol{\mu} \mid \boldsymbol{\Lambda}) P(\boldsymbol{\Lambda}) \right) \Big] + \text{const.} \\ &= \mathbb{E}_{Q(\boldsymbol{\pi})} \Big[\log P(\boldsymbol{z} \mid \boldsymbol{\pi}) \Big] + \mathbb{E}_{Q(\boldsymbol{\mu}, \boldsymbol{\Lambda})} \Big[\log P(\boldsymbol{x} \mid \boldsymbol{z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \Big] + \text{const.} \end{aligned}$$

Where

$$\begin{aligned} \mathbb{E}_{Q(\boldsymbol{\pi})} \Big[\log P(\boldsymbol{z} \mid \boldsymbol{\pi}) \Big] &= \mathbb{E}_{Q(\boldsymbol{\pi})} \Big[\log \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{n,k}} \Big] = \mathbb{E}_{Q(\boldsymbol{\pi})} \Big[\sum_{n=1}^{N} \sum_{k=1}^{K} z_{n,k} \log \pi_k \Big] \\ &= \sum_{n=1}^{N} \sum_{k=1}^{K} z_{n,k} \mathbb{E}_{Q(\boldsymbol{\pi})} \Big[\log \pi_k \Big]. \end{aligned}$$

Following a similar reasoning with the other expectation, we get that defining

$$\log \rho_{n,k} = \mathbb{E}_{Q(\boldsymbol{\pi})}\Big[\pi_k\Big] + \frac{1}{2}\mathbb{E}_{\boldsymbol{\Lambda}}\Big[\log \|\Lambda_k\|\Big] - \frac{D}{2}\log(2\pi) - \frac{1}{2}\mathbb{E}_{\mu_k,\Lambda_k}\Big[(x_n - \mu_k)^T \Lambda_k (x_n - \mu_k)\Big],$$

where D is the dimensionality of the data x. We obtain

$$Q^{new}(\boldsymbol{z}) \propto \prod_{n=1}^{N}\prod_{k=1}^{K} \rho_{n,k}^{z_{n,k}}.$$

Part IV

GRAPHICAL MODELS

A *graphical model* is a statistical model for which a graph expresses the conditional dependence structure between random variables.

Commonly, they provide a graph-based representation for encoding a multi-dimensional distribution representing a set of independences that hold in the specific distribution. The most commonly used are *Bayesian networks* and *Markov random fields*, which differ in the set of independences they can encode and the factorization of the distribution that they include.

12

INTRODUCTION (WIP)

Probabilistic graphical models are diagrammatic representations of probability distributions that represent the dependence/independence relation between the considered variables. Their use presents the following advantages (Bishop (2006)):

1. They allow a simple visualization of probabilistic models.
2. Insights into the model's properties can be obtained form the graph.
3. Complex computations, required in inference task, are simplified using the graph structure.

In probabilistic graphical models, each node represents a random variable and links represents relations between them. The graph encodes how the joint probability distribution of the considered variables factorizes. The different classes of graphical models differ in how they represent these relations and how the distribution is factorized.

We begin discussing *Bayesian networks*, also known as *directed graphical models*, where links between variables are indicated by a directed arrow. The other major class of graphical models are *Markov random fields* or *undirected graphical models* in which links have no directional meaning. The former are useful to express casual relations between the variables whereas the latter express soft constraints between the variables.

Consider a set of variables $\boldsymbol{X} = (X_1, \ldots, X_N)$, the possible ways these variables can interact is extremely large, for binary variables, the joint distribution table would take $O(2^N)$ space, which is unpractical when the amount of variables scales. When dealing with this such large distributions it is a common practice to factorize the joint probability in a graphical model, reducing the needed resources to deal with the inference problem.

13

BAYESIAN NETWORKS

Given a set of variables $\boldsymbol{X} = (X_1, \ldots, X_N)$, *Bayesian networks* might be defined either as a probability distribution of a certain form or a DAG whose nodes represent these variables and links an independence constraint. Both ideas are present in the following definition.

Definition 36. A *belief network or Bayesian network* is a pair (G, P) formed by a DAG G and joint probability distribution P such that, there is a correspondence between variables and nodes such as:

$$P(x_1, \ldots, x_N) = \prod_{n=1}^{N} P(x_n \mid pa(x_n)).$$

Remark 5. A Bayesian network might be given as a distribution from which the DAG can be constructed or a DAG which represents the distribution. For example in Figure 7, given the DAG one could easily define the joint distribution and conversely.

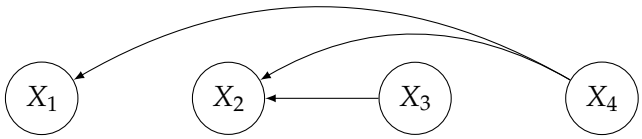


Figure 7: Bayesian Network factorizing $P(x_1, x_2, x_3, x_4) = P(x_1 \mid x_4)P(x_2 \mid x_3, x_4)P(x_3)P(x_4)$

Any probability distribution can be written as a Bayesian Network, even though it may end up been a fully-connected "cascade"[1] DAG, which means that each variable X_n is a parent of any X_m with $m > n$. This is because any distribution satisfies:

$$P(x_1, \ldots, x_N) = P(x_1) \prod_{n=2}^{N} P(x_n \mid x_1, \ldots, x_{n-1})$$

Bayesian Networks are good for encoding *conditional independence* over the variables, but are not for encoding dependence. For example, with the following network

$$P(x, y) = P(y \mid x)P(x).$$

represented as $x \to y$ in a DAG, it may appear to encode dependence between both variables but the conditional $P(y \mid x)$ could happen to equal $P(y)$, giving $P(x, y) = P(x)P(y)$.

1 This term comes from the visual structure of the graph.

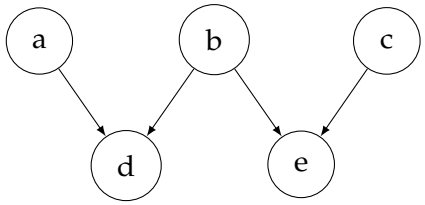


Figure 8: D-separation example

How could we check if two variables are conditionally independent given a Bayesian network? For example in Figure 3, $X_1 \perp\!\!\!\perp X_2 \mid X_4$ as

$$P(x_2|x_4) = \frac{1}{P(x_4)} \int_{x_1,x_3} P(x_1, x_2, x_3, x_4) = \frac{1}{P(x_4)} \int_{x_1,x_3} P(x_1|x_4)P(x_2|x_3, x_4)P(x_3)P(x_4)$$
$$= \int_{x_3} P(x_2|x_3, x_4)P(x_3),$$

$$P(x_1, x_2|x_4) = \frac{1}{P(x_4)} \int_{x_3} P(x_1, x_2, x_3, x_4) = \frac{1}{P(x_4)} \int_{x_3} P(x_1|x_4)P(x_2|x_3, x_4)P(x_3)P(x_4)$$
$$= P(x_1|x_4) \int_{x_3} P(x_2|x_3, x_4)P(x_3) = P(x_1|x_4)P(x_2|x_4).$$

13.1 D-SEPARATION AND D-CONNECTION

Now we are going to define two central concepts to determine conditional independence in any Bayesian network, these are *d-connection* and *d-separation*.

Definition 37. Let G be a DAG where $\boldsymbol{X}, \boldsymbol{Y}$ and $\boldsymbol{Z}$ are disjoint sets of vertices. We say that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are *d-connected* by $\boldsymbol{Z}$ if and only if there exists an undirected path U from any vertex in $\boldsymbol{X}$ to any vertex in $\boldsymbol{Y}$ such that:

- For any collider C, itself or any it's descendants is in $\boldsymbol{Z}$.
- No non-collider on U is on $\boldsymbol{Z}$.

Definition 38. Let G be a DAG where $\boldsymbol{X}, \boldsymbol{Y}$ and $\boldsymbol{Z}$ are disjoint sets of vertices. $\boldsymbol{X}$ and $\boldsymbol{Y}$ are *d-separated* by $\boldsymbol{Z}$ if and only if they are not d-connected by $\boldsymbol{Z}$ in G

For example, in Figure 8, d d-separates a and c (e is a collider in the path that is not in $\{d\}$), and $\{d, e\}$ d-connect them.

Theorem 2 (Pearl & Dechter (2013))**.** *Let G be a DAG where $\boldsymbol{X}, \boldsymbol{Y}$ and $\boldsymbol{Z}$ are disjoint sets of vertices. If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are d-separated by $\boldsymbol{Z}$, then they are independent conditional on $\boldsymbol{Z}$ in all probability distributions that G may represent.*

The Bayes Ball algorithm (Shachter (2013)) provides a linear time complexity algorithm that computes conditional independent using this theorem.

In cases where the Bayesian networks contains i.i.d nodes that are essentially the same but repeated a number of times, the *plate notation* is commonly used to represent this nodes in a compacted manner (Figure 9).

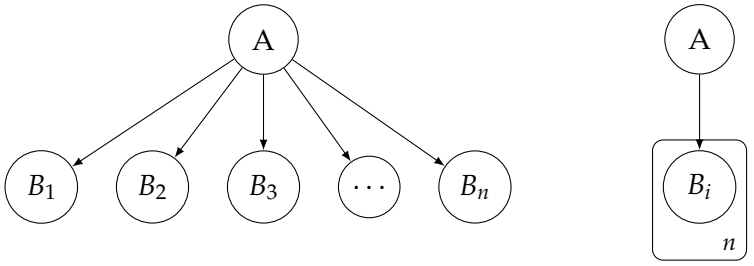


Figure 9: Plate notation example. Standard notation on the left and plate on the right.

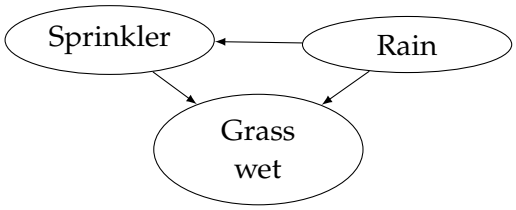
Example 3. In this example we are modeling three discrete random variables: Sprinkler (S), Rain (R) and Grass wet (G).

The joint probability function is:

$$P(s, r, g) = P(s|r)P(g|s, r)P(r)$$

The following DAG illustrates the Bayesian network among with the probability tables we are using.

	Sprinkler	
Rain	T	F
F	0.4	0.6
T	0.01	0.99



Rain	
T	F
0.2	0.8

		Grass wet	
Sprinkler	Rain	T	F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

This model can answer questions about the presence of a cause given the presence of an effect. For example, What is the probability that it has being raining given the grass is wet?

$$P(R = T|G = T) = \frac{P(G = T, R = T)}{P(G = T)} = \frac{\sum_s P(G = T, R = T, s)}{\sum_{r,s} P(G = T, r, s)}$$

Using the expression of the joint probability among with the tables we can compute every term. For example:

$$\begin{aligned} P(G = T, R = T, S = T) &= P(S = T|R = T)P(G = T|R = T, S = T)P(R = T) \\ &= 0.01 * 0.99 * 0.2 = 0.00198 \end{aligned}$$

14

MARKOV RANDOM FIELDS

Whereas Bayesian networks are represented by an acyclic directed graph, *Markov random fields* are undirected graphs that may be cyclic. Thus, this kind of graphical models can represent certain dependencies that Bayesian networks cannot, like cyclic ones. *Markov random fields* are commonly used to model low-to mid-level tasks in image processing and computer vision (Li (2009)).

Definition 39. Given an undirected graph $G = (V, E)$, a set of random variables $\boldsymbol{X} = (X_1, \ldots, X_N)$ *form a Markov random field over* G if they satisfy the, so called, Markov properties (Barber (2007)):

- **Pairwise Markov property**. Any two non-adjacent variables are conditionally independent given all other variables:

$$X_n \perp\!\!\!\perp X_m \mid \boldsymbol{X}_{\setminus n,m} \quad \forall n, m \in \{1, \ldots, N\} \quad n \neq m.$$

- **Local Markov property**. A variable is conditionally independent over all other variables given its neighbors:

$$X_n \perp\!\!\!\perp \boldsymbol{X}_{\setminus ne(X_n)} \mid \boldsymbol{X}_{ne(X_n)} \quad \forall n \in \{1, \ldots, N\}.$$

- **Global Markov property**. Any two subsets of variables are conditionally independent given a separating subset (any path from one set to the other passes trough this one). Let A and B be two subset of indexes and S a separating subset:

$$\boldsymbol{X}_A \perp\!\!\!\perp \boldsymbol{X}_B \mid \boldsymbol{X}_S.$$

The Global Markov property is stronger than the Local Markov property, which is stronger than the Pairwise one. However, these properties are equivalent for positive distributions ($P(x) > 0$), this is Theorem 4.4 in Koller & Friedman (2009).

As these Markov properties can be difficult to establish, a commonly used class of Markov random fields are those who factorize as product of potentials over the graph's cliques.

Definition 40. A *potential* ϕ is a non-negative function. It is worth to mention that a probability distribution is a special case of a potential.

Let $\boldsymbol{X}$ be a set of random variables, G an undirected graph whose nodes are $\boldsymbol{X}$ and $\boldsymbol{X}_c, c \in \{1, \ldots, C\}$ be the maximal cliques of G. If the joint probability distribution P can be factorized as:

$$P(x_1, \ldots, x_n) = \frac{1}{Z} \prod_{c=1}^{C} \phi_c(\boldsymbol{X}_c).$$

where Z is a constant that ensures normalization. Figure 10 shown an example in this class of Markov random fields.

Markov random fields factorize if any of the following conditions is fulfilled:

- The distribution is positive, this is shown in Hammersley–Clifford theorem (Grimmett (1973)).
- The graph is *chordal*, which means that any cycle of four of more vertices has a chord, an edge that is not part of the cycle but connects two of its vertices.

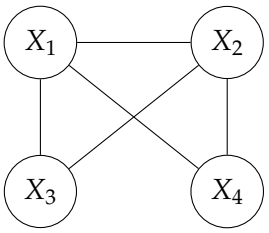


Figure 10: Markov Network $P(x_1, x_2, x_3, x_4) = \phi(x_1, x_2, x_3)\phi(x_2, x_3, x_4)/Z$

Part V

BAYESIAN NETWORKS LEARNING

15

MAXIMUM LIKELIHOOD TRAINING

Remember that, in case $P(x \mid \theta)$ is unconstrained, the maximum likelihood distribution corresponds to the empirical distribution.

For a Belief Network we know there is the following constraint

$$P(x_1, \ldots, x_N \mid \theta) = \prod_{n=1}^{N} P(x_n \mid pa(x_n), \theta).$$

In this case N variables are being considered so the empirical distribution counts the number of occurrences of a configuration of these variables.

We now aim to minimize the Kullback-Leibler divergence between the empirical distribution $Q(x_1, \ldots, x_N)$ and $P(x_1, \ldots, x_N \mid \theta)$ in order to get the Maximum Likelihood value:

$$\begin{aligned} KL\Big(Q \mid P\Big) &= -\mathbb{E}_Q\Big[\sum_{n=1}^{N} \log P(x_n \mid pa(x_n), \theta)\Big] + \mathbb{E}_Q\Big[\sum_{n=1}^{N} \log Q(x_n \mid pa(x_n))\Big] \\ &= -\sum_{n=1}^{N} \mathbb{E}_Q\Big[\log P(x_n \mid pa(x_n), \theta)\Big] + \sum_{n=1}^{N} \mathbb{E}_Q\Big[\log Q(x_n \mid pa(x_n))\Big]. \end{aligned}$$

We might use Proposition 1 on $\log P(x_n \mid pa(x_n), \theta)$ and $Q(x_n, pa(x_n))$, resulting:

$$\begin{aligned} KL\Big(Q \mid P\Big) &= \sum_{n=1}^{N} \mathbb{E}_{Q(x_n, pa(x_n))}\Big[\log Q(x_n \mid pa(x_n))\Big] - \mathbb{E}_{Q(x_n, pa(x_n))}\Big[\log P(x_n \mid pa(x_n), \theta)\Big] \\ &= \sum_{n=1}^{N} \mathbb{E}_{Q(x_i, pa(x_n))}\Big[KL\Big(Q(x_n \mid pa(x_n)) \mid P(x_n \mid pa(x_n), \theta)\Big)\Big]. \end{aligned}$$

The optimal setting is then

$$P(x_n \mid pa(x_n), \theta) = Q(x_n \mid pa(x_n)),$$

in terms of the initial data it is to set $P(x_n \mid pa(x_n))$ to the number of times the state appears in it.

16

BAYESIAN TRAINING

A Bayesian approach where we set a distribution over the parameters is an alternative to Maximum Likelihood training of a Bayesian Network, as we did in the coin tossing example **??**.

Consider the following scenario where a disease D and two habits A and B are being studied. Consider the following i.i.d variables $\{A_1, \ldots, A_N\}$, $\{B_1, \ldots, B_N\}$ and $\{D_1, \ldots, D_N\}$ governed by the parameters θ_A, θ_B and θ_D as shown in figure 11. Let $N = 7$ be the number of observations of the variables as shown in the table below and $\mathcal{D} = \{(a_n, b_n, d_n),\ n = 1, \ldots, N\}$ the set of observations.

All the variables are binary satisfying

$$P(A_n = 1 \mid \theta_A) = \theta_A, \quad P(B_n = 1 \mid \theta_B) = \theta_B, \quad P(D_n = 1 \mid A_n = 0, B_n = 1, \theta_D) = \theta_1,$$

$$\theta_D = (\theta_0, \theta_1, \theta_2, \theta_3).$$

Where we are using a binary to decimal transformation between the states of A and B and the sub-index of θ.

Summarizing, we are considering a Bernoulli distribution on A_n, B_n and D_n conditioned on the others.

The graph gives the following joint probability distribution over A_n, B_n and D_n:

$$P(a_n, b_n, d_n, \theta_A, \theta_B, \theta_D) = P(d_n \mid a_n, b_n, \theta_D) P(a_n \mid \theta_A) P(b_n \mid \theta_A).$$

We need to specify a prior and since dealing with multidimensional continuous distributions is computationally problematic, it is usual to use univariate distributions.

16.1 GLOBAL AND LOCAL PARAMETER INDEPENDENCE

A convenient assumption is that the prior factorizes, this is usually called *global parameter independence*:

$$P(\theta_A, \theta_B, \theta_D) = P(\theta_A) P(\theta_B) P(\theta_D).$$

Using this, the joint probability factorizes as

$$P(\mathcal{D}, \theta_A, \theta_B, \theta_D) = P(\theta_A) P(\theta_B) P(\theta_D) \prod_n P(a_n \mid \theta_A) P(b_n \mid \theta_B) P(d_n \mid a_n, b_n, \theta_D).$$

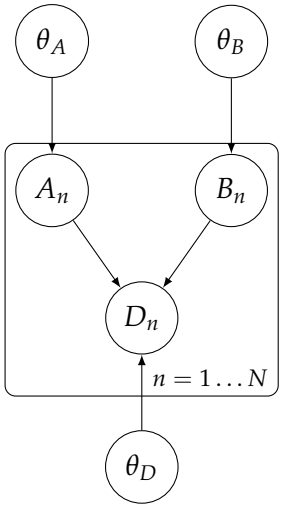


Figure 11: Bayesian parameter model for the relation between A, B, D

A	B	D
1	1	1
1	0	0
0	1	1
0	1	0
1	1	1
0	0	0
1	0	1

Table 1: Set of observations, where 1 means true and 0 means false.

Therefore, learning corresponds to making inference,

$$\begin{aligned} P(\theta_A, \theta_B, \theta_D \mid \mathcal{D}) &= \frac{P(\mathcal{D} \mid \theta_a, \theta_B, \theta_D) P(\theta_A, \theta_B, \theta_D)}{P(\mathcal{D})} = \frac{P(\mathcal{D} \mid \theta_A, \theta_B, \theta_D) P(\theta_A) P(\theta_B) P(\theta_D)}{P(\mathcal{D})} \\ &= \frac{1}{P(\mathcal{D})} P(\theta_A) P(\theta_B) P(\theta_D) \prod_n P(a_n \mid \theta_A) P(b_n \mid \theta_B) P(d_n \mid a_n, b_n, \theta_D) \\ &= P(\theta_A \mid \mathcal{D}_A) P(\theta_B \mid \mathcal{D}_B) P(\theta_D \mid \mathcal{D}). \end{aligned}$$

Where $\mathcal{D}_A$ is the subset of observations related to the variable A. If we further assume that $P(\theta_D)$ factorizes as $P(\theta_D) = P(\theta_0)P(\theta_1)P(\theta_2)P(\theta_3)$, which is called *local parameter independence*, $P(\theta_D \mid \mathcal{D})$ factorizes as

$$P(\theta_D \mid \mathcal{D}) = P(\theta_0 \mid \mathcal{D}) P(\theta_1 \mid \mathcal{D}) P(\theta_2 \mid \mathcal{D}) P(\theta_3 \mid \mathcal{D}).$$

16.2 LEARNING BINARY VARIABLES

The simplest cases to continue are $P(\theta_A \mid \mathcal{D}_A)$ and $P(\theta_B \mid \mathcal{D}_B)$ since they require only a uni-variate prior distribution $P(\theta_A)$ or $P(\theta_B)$. The procedure is shown using θ_A and it is analogous when using θ_B.

The posterior is

$$P(\theta_A \mid \mathcal{D}_A) = \frac{1}{P(\mathcal{D}_A)} P(\theta_A) \theta_A^{\#(a=1)} (1 - \theta_A)^{\#(a=0)}.$$

The most convenient choice for the prior is a Beta distribution as conjugacy will hold:

$$\theta_A \sim \text{Beta}(\alpha_A, \beta_A) \implies P(\theta_A) = \frac{1}{B(\alpha_A, \beta_A)} \theta_A^{\alpha_A - 1} (1 - \theta_A)^{\beta_A - 1}.$$

Therefore, it follows that

$$\theta_A \mid \mathcal{D}_A \sim \text{Beta}(\alpha_A + \#(A = 1), \beta_A + \#(A = 0)).$$

The marginal is then

$$\begin{aligned} P(A=1 \mid \mathcal{D}_A) &= \frac{P(A=1,\mathcal{D}_A)}{P(\mathcal{D}_A)} = \int_{\theta_A} \frac{P(A=1,\mathcal{D}_A,\theta_A)}{P(\mathcal{D}_A)} = \int_{\theta_A} \frac{P(A=1 \mid \mathcal{D}_A,\theta_A)P(\mathcal{D}_A,\theta_A)}{P(\mathcal{D}_A)} \\ &= \int_{\theta_A} \frac{P(A=1 \mid \mathcal{D}_A,\theta_A)P(\theta_A \mid \mathcal{D}_A)P(\mathcal{D}_A)}{P(\mathcal{D}_A)} \\ &= \int_{\theta_A} P(\theta_A \mid \mathcal{D}_A)\theta_A = \mathbb{E}\Big[\theta_A \mid \mathcal{D}_A\Big] \\ &= \frac{\alpha_A + \#(A=1)}{\alpha_A + \#(A=1) + \beta_A + \#(A=0)}. \end{aligned}$$

Where the last equality is given by the expected value of a Beta distribution.

For $P(d \mid a,b)$ the situation is more complex, the simplest approach is to specify a Beta prior for each of the components of θ_D. Focus on θ_2, notice the parameters α and β we used before now do depend on A and B, these are called *hyperparameters*.

$$\theta_2 \sim Beta\Big(\alpha_D(1,0) + \#(D=1,A=1,B=0),\ \beta_D(1,0) + \#(D=0,A=1,B=0)\Big).$$

Repeating the procedure we used with A we get that

$$P(D=1 \mid A=1,B=0,\mathcal{D}) = \frac{\alpha_D(1,0) + \#(D=1,A=1,B=0)}{\alpha_D(1,0) + \beta_D(1,0) + \#(A=1,B=0)}.$$

In case we had no preference, we could set all hyperparameters to the same value, where a complete ignorance prior would correspond to set them to 1.

Once we get to this situation there are two limit possibilities depending on the amount of data we got.

- **No data**. The marginal probability corresponds to the prior, which in the last case is

  $$P(D=1 \mid A=1,B=0,\mathcal{D}) = \frac{\alpha_D(1,0)}{\alpha_D(1,0) + \beta_D(1,0)}.$$

  Note that equal hyperparameters would give a result of 0.5.

- **Infinite data**. When infinite data is available, the marginal is generally dominated by it, this corresponds to the Maximum Likelihood solution.

  $$P(D=1 \mid A=1,B=0,\mathcal{D}) = \frac{\#(D=1,A=1,B=0)}{\#(A=1,B=0)}.$$

  This happens unless the prior has a pathologically strong effect.

Consider the data given in the table in figure 11, and equal parameters and hyperparameters 1 and a prior belief that any setting is equally probable, i.e, $P(A=1) = 0.5$.

We may illustrate the different results that are obtained using using Bayesian inference and Maximum likelihood training. The former is

$$P(A=1 \mid \mathcal{D}) = \frac{1 + \#(A=1)}{2+N} = \frac{5}{9} \approx 0.556.$$

and the latter is $4/7 = 0.571$. In conclusion, the Bayesian result is more prudent than this one, which fits in with our prior belief.

16.3 LEARNING DISCRETE VARIABLES

The natural generalization to more than two states is using a Dirichlet distribution as prior, assuming i.i.d data, local and global prior independence. We are considering two different scenarios, firstly one where the variable has no parents, as the case for A and B in the previous example. Secondly, we will consider a variable with a non void set of parents.

16.3.1 *No parents*

Consider a variable $X \sim Categorical(\theta)$ with $Dom(X) = \{1, \dots, I\}$ and $\theta = (\theta_1, \dots, \theta_I)$ so that

$$P(x) = \prod_{i=1}^{I} \theta_i^{\mathbb{I}[x=i]} \text{ with } \sum_{i=1}^{I} \theta_i = 1.$$

So that the posterior (considering N observations of the variable $\mathcal{D} = (x_1, \dots, x_N)$) is

$$P(\theta \mid \mathcal{D}) = \frac{1}{P(\mathcal{D})} P(\theta) \prod_{n=1}^{N} \prod_{i=1}^{I} \theta_i^{\mathbb{I}[x_n=i]} = \frac{1}{P(\mathcal{D})} P(\theta) \prod_{i=1}^{I} \theta_i^{\sum_n \mathbb{I}[x_n=i]}.$$

Then, assuming a prior $\theta \sim Dirichlet(\boldsymbol{u})$ with $\boldsymbol{u} = (u_1, \dots, u_I)$

$$P(\theta) = \frac{1}{B(\boldsymbol{u})} \prod_{i=1}^{I} \theta_i^{u_i - 1} \implies P(\theta \mid \mathcal{D}) = \frac{1}{B(\boldsymbol{u}) P(\mathcal{D})} \prod_{i=1}^{I} \theta_i^{u_i - 1 + \sum_n \mathbb{I}[x_n=i]}.$$

Therefore, defining $\boldsymbol{c} = (\sum_{n=1}^{N} \mathbb{I}[x_n = i])_{i=1,\dots,I}$, the posterior follows a Dirichlet distribution

$$\theta \mid \mathcal{D} \sim Dirichlet(\boldsymbol{u} + \boldsymbol{c}).$$

It may not be easy to decompose $B(\boldsymbol{u} + \boldsymbol{c})$ as $B(\boldsymbol{u})P(\mathcal{D})$, but using that it ensures normalization of the Beta distribution we may use that

$$\int_\theta P(\theta \mid \mathcal{D}) = 1 \implies \int_\theta \prod_{i=1}^{I} \theta_i^{u_i + c_i - 1} = B(\boldsymbol{u} + \boldsymbol{c}) = B(\boldsymbol{u}) P(\mathcal{D}).$$

Remark 6. Summarizing the above information, we just proved that the Dirichlet distribution is the conjugate prior of the Categorical distribution.

The marginal is then given by

$$\begin{aligned} P(X = i \mid \mathcal{D}) &= \int_\theta P(X = i \mid \theta) P(\theta \mid \mathcal{D}) = \int_\theta \theta_i P(\theta \mid \mathcal{D}) \\ &= \int_{\theta_i} \theta_i P(\theta_i \mid \mathcal{D}) = \mathbb{E}\Big[\theta_i \mid \mathcal{D}\Big]. \end{aligned}$$

Where we used that

$$\int_{\theta_{j \neq i}} \theta_i P(\theta \mid \mathcal{D}) = \theta_i \prod_{k \neq j} P(\theta_k \mid V) \int_{\theta_j} P(\theta_j \mid \mathcal{D}) = \theta_i \prod_{k \neq j} P(\theta_k \mid \mathcal{D})$$

.

As we already know from Proposition 2, the uni-variate marginal of a Dirichlet distribution is a Beta distribution of parameters

$$\theta_i \mid \mathcal{D} \sim \text{Beta}(u_i + c_i, \sum_{j \neq i} u_j + c_j).$$

Using the expectation of a Beta distribution we get that

$$P(X = i \mid \mathcal{D}) = \frac{u_i + c_i}{\sum_j u_j + c_j}.$$

16.3.2 *Parents*

Consider now that X has a set of parent variables $pa(X)$, in this case, we want to compute the marginal given a state of its parents and the data:

$$P(X = i \mid pa(X) = \boldsymbol{j}, \mathcal{D}).$$

We are using the following notation for the parameters

$$P(X = i \mid pa(X) = \boldsymbol{j}, \theta) = \theta_{i,\boldsymbol{j}}, \quad \boldsymbol{\theta_j} = (\theta_{1,\boldsymbol{j}}, \dots, \theta_{I,\boldsymbol{j}}).$$

Local independence means that

$$P(\boldsymbol{\theta}) = \prod_j P(\boldsymbol{\theta}_j).$$

As we did before, we consider a Dirichlet prior

$$\boldsymbol{\theta_j} \sim Dirichlet(\boldsymbol{u_j}),$$

the posterior is then

$$\begin{aligned}
P(\boldsymbol{\theta} \mid \mathcal{D}) &= \frac{P(\boldsymbol{\theta})P(\mathcal{D} \mid \boldsymbol{\theta})}{P(\mathcal{D})} = \frac{1}{P(\mathcal{D})}\Big(\prod_{\boldsymbol{j}} P(\boldsymbol{\theta}_j)\Big)P(\mathcal{D} \mid \boldsymbol{\theta}) \\
&= \frac{1}{P(\mathcal{D})}\Big(\prod_{\boldsymbol{j}} \frac{1}{B(\boldsymbol{u_j})} \prod_i \theta_{i,\boldsymbol{j}}^{u_{i,\boldsymbol{j}}-1}\Big)P(\mathcal{D} \mid \boldsymbol{\theta}) \\
&= \frac{1}{P(\mathcal{D})}\Big(\prod_{\boldsymbol{j}} \frac{1}{B(\boldsymbol{u_j})} \prod_i \theta_{i,\boldsymbol{j}}^{u_{i,\boldsymbol{j}}-1}\Big)\Big(\prod_n \prod_{\boldsymbol{j}} \prod_i \theta_{i,\boldsymbol{j}}^{\mathbb{I}[x_n = i, pa(x_n) = \boldsymbol{j}]}\Big) \\
&= \frac{1}{P(\mathcal{D})} \prod_{\boldsymbol{j}} \frac{1}{B(\boldsymbol{u_j})} \prod_i \theta_{i,\boldsymbol{j}}^{u_{i,\boldsymbol{j}}-1+\#(X=i, pa(X)=\boldsymbol{j})}.
\end{aligned}$$

Naming $\boldsymbol{v_j} = \boldsymbol{u_j} + \#(X = i, pa(X) = \boldsymbol{j})$ and using the same argument as we did in the *no parents* case with the normalization constants, the posterior is

$$(\boldsymbol{\theta} \mid \mathcal{D}) \sim \prod_j \text{Dirichlet}(\boldsymbol{v_j}).$$

Denoting $v_{i,j}$ the components of $\boldsymbol{v_j}$, the marginal is

$$P(X = i, pa(X) = \boldsymbol{j}, \mathcal{D}) = \frac{v_{i,j}}{\sum_i v_{i,j}}.$$

Notice all the above has been done using a fixed variable X, so that all the parameters depend on that variable.

We can define the data likelihood under a model, usually called the *model likelihood* using the same calculations as we did with $P(\boldsymbol{\theta} \mid \mathcal{D})$ but applied to all the variables of the model

$$\begin{aligned} P(\mathcal{D} \mid \mathcal{M}) &= \int_{\boldsymbol{\theta}} P(\boldsymbol{\theta}) P(\mathcal{D} \mid \boldsymbol{\theta}, \mathcal{M}) = \prod_x \prod_{\boldsymbol{j}} \frac{1}{B(\boldsymbol{u}_j)} \int_{\boldsymbol{\theta}} \prod_i \theta_{i,\boldsymbol{j}}^{u_{i,\boldsymbol{j}} - 1 + \#(x=i, pa(x)=\boldsymbol{j})} \\ &= \prod_x \prod_{\boldsymbol{j}} \frac{B(\boldsymbol{v}_j)}{B(\boldsymbol{u}_j)}. \end{aligned}$$

17

STRUCTURE LEARNING

So far, we have assumed the Bayesian Network was an hypothesis of the problem, however, this structure is not always given and may be learned also form the data.

Even considering complete data (no missing observations), there are some problems that need to be taken into account.

- The number of Bayesian networks is exponential over the number of variables so a brute force algorithm would not be viable.
- Testing dependencies requires a large amount of data. Therefore a threshold must be set to measure when a dependence is significant.
- A Bayesian network or a Markov network may not be enough to represent the data due to the existence of unobserved variables.

17.1 PC ALGORITHM

An approach to learn the structure is the PC algorithm, it starts with a complete graph G and tries to remove as many links as possible studying the independence of the variables.

The algorithm 3 iterates over neighborhoods, from smaller to bigger ones. It chooses a linked pair of variables $(X, Y) \in E$ and a subset $S_{XY} \subset ne(X)$, verifying $Y \notin S_{XY}$. If $X \perp\!\!\!\perp Y \mid S$, then the link is removed and S_{XY} is stored.

The main idea behind the algorithm is that a set of independencies is faithful to a graph if there is no link between two nodes X and Y if and only if there exists a subset of $ne(X)$ such that they are independent given this subset.

This gives the skeleton of the Bayesian Network, no more edges will be removed or added. The directed graph may be constructed following two rules rule:

1. For any undirected link $X - Y - Z$, if $Y \notin S_{XZ}$ then set $X \to Y \leftarrow Z$ (we are creating a collider for that path).
2. The rest of links may oriented arbitrarily not creating cycles or colliders.

The reasoning behind this is the d-separation theorem 2, if $Y \notin S_{XZ}$ then $X \not\perp\!\!\!\perp Z \mid Y$ so Y must d-connect them, to get this we set it as a collider. On the other hand, if $Y \in S_{XZ}$ and $X \perp\!\!\!\perp Z \mid S_{XZ}$ then we want S_{XZ} to d-separate them, that is, using any configuration that doesn't create a collider in S_{XZ}.

Algorithm 3: PC Algorithm

Data: Complete undirected graph G, with vertices $\mathcal{V}$
Result: G with removed links
$i = 0$;
while *all nodes have* $\leq i$ *neighbors* **do**
  for $X \in \mathcal{V}$ **do**
    for $Y \in ne(x)$ **do**
      if $\exists S \subset ne(X) \backslash Y$ *such that* $\#S = i$ *and* $X \perp\!\!\!\perp Y \mid S$ **then**
        Remove $X - Y$ from G;
        $S_{XY} = S$;
      end
    end
  end
  $i = i + 1$;
end

17.2 INDEPENDENCE LEARNING

Our main concern now is, given three variables X, Y and Z, measure $X \perp\!\!\!\perp Y \mid Z$. One approach is to measure the empirical *conditional mutual information* of the variables.

Definition 41. Given two random variables X, Y, we define their *mutual information* as the Kullback-Leibler divergence of their joint distribution and the product of their marginals,

$$MI(X;Y) = KL\Big(P_{X,Y} \mid P_X P_Y\Big).$$

Definition 42. Given three random variables X, Y, Z we define the *conditional mutual information* of X and Y over Z as

$$MI(X;Y \mid Z) = \mathbb{E}_Z\Big[KL\Big(P_{X,Y|Z} \mid P_{X|Z} P_{Y|Z}\Big)\Big].$$

Where $MI(X;Y \mid Z) \geq 0$ and $MI(X;Y \mid Z) = 0 \iff P_{X,Y|Z} = P_{X|Z} P_{Y|Z} \iff X \perp\!\!\!\perp Y \mid Z$. We can estimate this using the empirical distributions, however, this *empirical* mutual information will be typically greater than 0 even when $X \perp\!\!\!\perp Y \mid Z$, therefore a threshold must be established.

A Bayesian approach would consist on comparing the model likelihood under independence and dependence hypothesis. That is, computing the model likelihood for the below joint distributions assuming both local and global parameter independence

$$P_{indep}(x,y,z) = P(x \mid z, \theta_1) P(y \mid z, \theta_2) P(z \mid \theta_3) P(\theta_1) P(\theta_2) P(\theta_3),$$

$$P_{dep}(x,y,z) = P(x,y,z \mid \theta) P(\theta).$$

18

MISSING VARIABLES

Until this moment we have assumed that the data we are given is completed but in practice this data is not in two different ways. There may be unobserved or *hidden* variables that affect the visible ones, and there may be *missing* information, that is, states of visible variables that are missing.

Think about the example with the disease and the two habits we used in the last section, missing data would be a row in the table where some entry is missing, for example $x_3 = \{D = 1, A = 1\}$, where we know that this person got the disease and had habit A but we have no information about habit B, this is an example of *missing* data.

One approach to handle this situation would be marginalizing over that variable:

$$P(x_3 \mid \theta) = \int_b P(d_3, a_3, b \mid \theta) = P(a_3 \mid \theta_A) \int_b P(b \mid \theta_B) P(d_3 \mid b, a_3, \theta_D).$$

This leads to a non-factorized form of the posterior which is computationally difficult to handle, notice the problem is not conceptual but computational. Using the marginal to handle missing information does not always lead to this situation, in fact, marginalizing over a collider (D in our example) would mean loosing that variable as the integral simply equals 1.

$$P(x_3 \mid \theta) = \int_d P(d, a_3, b_3 \mid \theta) = P(a_3 \mid \theta_A) P(b_3 \mid \theta_B) \int_d P(d \mid b_3, a_3, \theta_D) = P(a_3 \mid \theta_A) P(b_3 \mid \theta_B);$$

There are three main types of missing data:

- **Missing completely at random (MCAR)**. If the events that lead to any particular data to be missing is independent from both the observed and the unobserved variables, and occur at random.
- **Missing at random (MAR)**. When the absence is not random but can be explained with the observed variables.
- **Missing not at random (MNAR)**. The missing data is related with the reason why it is missing. For example, skipping a question in a survey for being ashamed of the answer.

To express this mathematically, split the variables $\mathcal{X}$ into visible $\mathcal{X}_{vis}$ and hidden $\mathcal{X}_{hid}$, let M be a variable denoting that the state of the hidden variables is known (0) or unknown (1). So the difference between the three types resides on how $P(M = 1 \mid x_{vis}, x_{hid}, \theta)$ simplifies.

When data is *missing at random*, we assume that we can explain the missing information with the visible one, so the probability of being missing only depends on the visible data, that is

$$P(M = 1 \mid x_{vis}, x_{hid}, \theta) = P(M = 1 \mid x_{vis}),$$

so that,

$$P(x_{vis}, M = 1 \mid \theta) = P(M = 1 \mid x_{vis})P(x_{vis} \mid \theta)$$

Assuming the data is *missing completely at random* is stronger, as we are supposing that there is no reason behind the missing data, so that it being missing is independent from the visible and hidden data:

$$P(M = 1 \mid x_{vis}, x_{hid}, \theta) = P(M = 1),$$

so now

$$P(x_{vis}, M = 1 \mid \theta) = P(M = 1)P(x_{vis} \mid \theta).$$

In both cases we may simply use the marginal $P(x_{vis} \mid \theta)$ to assess parameters as $P(x_{vis}, M = 1 \mid \theta)$ does not depend on the missing variables.

In case data is *missing not at random*, no independence assumption is made over the probability of the data being unknown, meaning it depends on both the visible and the hidden information. From now on, we will assume missing information is either MAR or MCAR, even though this could lead to a misunderstanding of the problem as in the following simple example.

Example 4. Consider a situation where data is obtained from a survey where people are asked to choose between 3 options A, B and C. Assume that no one chose option C because they are ashamed of the answer, and the answers are uniform between A, B and not answering.

Normalizing the missing information would lead to setting $P(A \mid \mathcal{V}) = 0.5 = P(B \mid \mathcal{V})$ and $P(C \mid \mathcal{V}) = 0$ when the reasonable result is that not answering equals to choosing C so that $P(A \mid V) = P(B \mid \mathcal{V}) = P(C \mid \mathcal{V}) = \frac{1}{3}$

19

VARIATIONAL INFERENCE IN BAYESIAN NETWORKS

In this section we come back to variational inference, in this case we are reviewing two main algorithms, *expectation maximization* and *variational message passing*.

19.1 EXPECTATION MAXIMIZATION

Let $\boldsymbol{X} = (\boldsymbol{V}, \boldsymbol{H}) = \{X_1, \dots, X_M\}$ be the set of variables partitioned in visible and hidden. Let $\{\boldsymbol{v}^1, , \dots, \boldsymbol{v}^N, \}$ be the set of observations and $\{\boldsymbol{h}_1, \dots, \boldsymbol{h}_N\}$ a set of possible values of the hidden variables in each observation.

As we are not using a single variable anymore, the variational distribution is now over the set of hidden variables conditioned on the visible $Q(\boldsymbol{H} \mid \boldsymbol{V})$. As the E-step is the same as in the general case, we are focusing on the M-step, where Q_{old} is fixed.

The *energy term* in a Bayesian Network is

$$\sum_{n=1}^{N} \mathbb{E}_{Q_{old}(\boldsymbol{h}^n|\boldsymbol{v}^n)}\Big[\log P(\boldsymbol{x}^n \mid \theta)\Big] = \sum_{n=1}^{N}\sum_{i=0}^{M} \mathbb{E}_{Q_{old}(\boldsymbol{h}^n|\boldsymbol{v}^n)}\Big[\log P(x_i^n \mid pa(x_i^n), \theta)\Big].$$

It is useful to use the following notation that defines a conditional distribution of the hidden variable when the visible one equals v^n.

$$Q^n(\boldsymbol{x}) = Q^n(\boldsymbol{v}, \boldsymbol{h}) = Q_{old}(\boldsymbol{h}^n \mid \boldsymbol{v}^n)\mathbb{I}(\boldsymbol{v} = \boldsymbol{v}^n).$$

We can define the mixture distribution

$$Q(\boldsymbol{x}) = \frac{1}{N}\sum_{n=1}^{N} Q^n(\boldsymbol{x}).$$

Then we have that

$$\begin{aligned}
\mathbb{E}_{Q(\boldsymbol{x})}\Big[\log P(\boldsymbol{x} \mid \theta)\Big] &= \int_{\boldsymbol{x}} Q(\boldsymbol{x}) \log P(\boldsymbol{x} \mid \theta) = \int_{\boldsymbol{x}} \frac{1}{N}\sum_{n=1}^{N} Q^n(\boldsymbol{x}) \log P(\boldsymbol{x} \mid \theta) \\
&= \frac{1}{N}\int_{\boldsymbol{x}} \sum_{n=1}^{N} Q_{old}(\boldsymbol{h}^n \mid \boldsymbol{v}^n)\mathbb{I}[\boldsymbol{v} = \boldsymbol{v}^n] \log P(\boldsymbol{x} \mid \theta) \\
&= \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}_{Q_{old}(\boldsymbol{h}^n|\boldsymbol{v}^n)}\Big[\log P(\boldsymbol{h}^n, \boldsymbol{v}^n \mid \theta)\Big] \\
&= \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}_{Q_{old}(\boldsymbol{h}^n|\boldsymbol{v}^n)}\Big[\log P(\boldsymbol{x}^n \mid \theta)\Big],
\end{aligned}$$

equals the energy term. Then, using the Belief Network structure

$$\begin{aligned}\mathbb{E}_{Q(x)}\Big[\log P(\boldsymbol{x} \mid \theta)\Big] &= \sum_{i=1}^{M} \mathbb{E}_{Q(\boldsymbol{x})}\Big[\log P(x_i \mid pa(x_i, \theta))\Big] \\ &= \sum_{i=1}^{M} \mathbb{E}_{Q(x_i, pa(x_i))}\Big[\log P(x_i \mid pa(x_i, \theta))\Big] \\ &= \sum_{i=1}^{M} \mathbb{E}_{Q(pa(x_i))}\Big[\mathbb{E}_{Q(x_i|pa(x_i))}\Big[\log P(x_i \mid pa(x_i), \theta)\Big]\Big].\end{aligned}$$

We add a constant to the last term so it comes with the structure of a Kullback-Leibler Divergence (notice its sign has changed)

$$\sum_{i=1}^{M} \mathbb{E}_{Q(pa(x_i))}\Big[\mathbb{E}_{Q(x_i|pa(x_i))}\Big[\log Q(x_i \mid pa(x_i))\Big]\Big] - \mathbb{E}_{Q(x_i|pa(x_i))}\Big[\log P(x_i \mid pa(x_i), \theta)\Big] =$$

$$= \sum_{i=1}^{M} E_{Q(pa(x_i))}\Big[KL\Big(Q(x_i \mid pa(x_i)) \mid P(x_i \mid pa(x_i), \theta)\Big)\Big].$$

So maximizing the energy term is equivalent to minimize the above formula, that is, setting

$$P^{new}(x_i \mid pa(x_i), \theta) = Q(x_i \mid pa(x_i)).$$

So the first observation is that θ is not needed in order to maximize the energy term due to the Belief network structure.

19.2 VARIATIONAL MESSAGE PASSING

"However, Monte Carlo methods are computationally very intensive, and also suffer from difficulties in diagnosing convergence"

"Expectation propagation is limited to certain classes of model for which the required expectations can be evaluated, is also not guaranteed to converge in general, and is prone to finding poor solutions in the case of multi-modal distributions."

"VMP allows variational inference to be applied automatically to a large class of Bayesian networks, without the need to derive application-specific update equations"

In this section, we review the *Variational Message Algorithm* or *VMA* as a variational Bayes application to Bayesian networks where the exponential family is considered using a message passing procedure between the nodes of a graphical model.

The full set of variables is $\mathcal{X} = X_1 \ldots, X_N$, where we are considering both hidden variables $\mathcal{H} = \{H_1, \ldots, H_J\}$ and visible ones $\mathcal{V} = \{V_1, \ldots, V_I\}$. A variational distribution Q in the mean-field family

$$Q(\mathcal{H}) = \prod_{j=1}^{J} Q_j(h_j).$$

The optimized factor for a fixed term $Q(h_j)$ is (as shown in 1) given by

$$\log Q_j^{new}(h_j) = \mathbb{E}_{Q_{\setminus j}}\Big[\log P(\mathcal{V}, \mathcal{H})\Big] + \text{const.}$$

Using the Bayesian network structure, the update is given by

$$\log Q_j^{new}(h_j) = \mathbb{E}_{Q_{\setminus j}}\Big[\sum_{n=1}^{N} \log P(x_n \mid pa(x_n))\Big] + \text{const.}$$

The contribution of H_j to the given formula lies on the terms $P(h_j \mid pa(h_j))$ and the conditionals of all its children, let $cp(X, Y)$ denote the co-parents of Y with X, $pa(X)\backslash\{Y\}$, then, adding all other terms to the constant value,

$$\log Q_j^{new}(h_j) = \mathbb{E}_{Q_{\setminus j}}\Big[\log P(h_j \mid pa(h_j))\Big] + \sum_{X_k \in ch(H_j)} \mathbb{E}_{Q_{\setminus j}}\Big[\log P(x_k \mid h_j, cp(x_k, h_j))\Big] + \text{const.}$$

This shows how the update of a hidden variable only depends on its Markov blanket. The optimization of Q_j is therefore computed as the sum of a term involving H_j and its parent nodes, along with a term for each children. This terms can be interpreted as "messages" from the corresponding nodes.

The exact form of the messages will depend on the functional form of the conditional distributions in the model. Important simplifications to the variational update equations occur when the conditional distribution of a node given its parents is in the exponential family. Sub-indexes will be used to denote parents, children and co-parents.

Consider a variable X and $Y \in pa_X$, such as Y is a hidden variable. Then, both distributions belong to the exponential family:

$$P(y \mid pa_y) = h_Y(pa_Y) \exp\Big(\boldsymbol{\eta}_Y(pa_y)^T \boldsymbol{T}_Y(y) - \psi_Y(pa_y)\Big).$$

$$P(x \mid y, cp_{x,y}) = h_X(y, cop_{x,y}) \exp\Big(\boldsymbol{\eta}_X(y, cop_{x,y})^T \boldsymbol{T}_X(x) - \psi_X(y, cp_{x,y})\Big).$$

Remark 7. Given a variable X in the exponential family, if we know the natural parameter vector $\boldsymbol{\eta}$, then we can find the expectation of the statistic function with respect to the distribution. Defining $\bar{\phi}$ as a reparameterisation of ϕ in terms of $\boldsymbol{\eta}$:

$$\begin{aligned} P(x \mid \theta) &= h(x) \exp\Big(\boldsymbol{\eta}(\theta)^T \boldsymbol{T}(x) + \psi(\theta)\Big) \\ &= h(x) \exp\Big(\boldsymbol{\eta}(\theta)^T \boldsymbol{T}(x) + \bar{\psi}(\boldsymbol{\eta}(\theta))\Big). \end{aligned}$$

Integrating with respect to X,

$$1 = \int_x h(x) \exp\Big(\boldsymbol{\eta}(\theta)^T \boldsymbol{T}(x) + \bar{\psi}(\boldsymbol{\eta}(\theta))\Big).$$

Differentiating with respect to $\boldsymbol{\eta}$:

$$\frac{d}{d\theta} 1 = 0 = \int_x \frac{d}{d\theta} h(x) \exp\Big(\boldsymbol{\eta}(\theta)^T \boldsymbol{T}(x) + \bar{\psi}(\boldsymbol{\eta}(\theta))\Big) = \int_x P(x \mid \theta)\Big[\boldsymbol{T}(x) + \frac{\bar{\psi}(\boldsymbol{\eta}(\theta))}{d\boldsymbol{\eta}}\Big].$$

What implies that

$$\mathbb{E}_P\Big[\boldsymbol{T}(x)\Big] = -\frac{\bar{\psi}(\boldsymbol{\eta}(\theta))}{d\boldsymbol{\eta}}$$

The distribution $P(Y \mid pa_Y)$ can be though as a prior over Y, and $P(X \mid pa_X) = P(X \mid Y, cop_{X,Y})$ as a contribution to the likelihood of Y.

Conjugacy requires that these two conditionals have the same functional form with respect to Y, so the latter has to be rewritten in terms of $\boldsymbol{T}_Y(y)$ by defining functions $\boldsymbol{\eta}_{X,Y}$ and λ as

$$\log P(x \mid y, cp_{x,y}) = \boldsymbol{\eta}_{XY}(x, cp_{x,y})^T \boldsymbol{T}_Y(y) + \lambda(x, cp_{x,y}).$$

Example 5. If X is Gaussian distributed with mean μ (Gaussian distributed) and standard deviation σ, let $\tau = 1/\sigma^2$ be its precision, the log conditional is

$$\log P(x \mid \mu, \tau) = \begin{pmatrix} \mu\tau & -\tau/2 \end{pmatrix}^T \begin{pmatrix} x \\ x^2 \end{pmatrix} - \frac{\mu^2\tau}{2} + \frac{1}{2}\big(\log\tau - \log 2\pi\big).$$

We may rewrite it the conditional as

$$\log P(x \mid \mu, \tau) = \begin{pmatrix} \tau x & -\tau/2 \end{pmatrix} \begin{pmatrix} \mu \\ \mu^2 \end{pmatrix} + \frac{1}{2}\big(\log\tau - \tau x^2 - \log 2\pi\big),$$

where

$$\boldsymbol{\eta}_{X,\mu}(x,\tau) = \begin{pmatrix} \tau x \\ -\tau/2 \end{pmatrix}^T, \quad \boldsymbol{T}_\mu(\mu) = \begin{pmatrix} \mu \\ \mu^2 \end{pmatrix}.$$

From this results, it can be seen that $\log P(x \mid y, cp_y)$ is linear in $\boldsymbol{T}_X(x)$ and $\boldsymbol{T}_Y(y)$, and, by the same reasoning, linear in any sufficient statistic of any parent of X. This is a general result for any variable X in this kind of models: *For any variable X, the log conditional under its parents must be multi-linear of the statistics of X and its parents.*[1]

Returning to the variational update for a node Y:

$$\log Q^{new}(y) = \mathbb{E}_{Q_{\setminus Y}}\Big[logP(y \mid pa_Y)\Big] + \sum_{X_k \in ch(Y)} \mathbb{E}_{Q_{\setminus Y}}\Big[\log P(x_k \mid pa_{X_k}))\Big] + \text{const.},$$

where the expectations are over the variational distribution of all other hidden variables and can be calculated in terms of $\boldsymbol{T}_Y(y)$:

$$\begin{aligned}
\log Q^{new}(y) &= \mathbb{E}_{Q_{\setminus Y}}\Big[\log(h_Y(pa_y)) + \boldsymbol{\eta}_Y(pa_y)^T \boldsymbol{T}_Y(y) + \psi_Y(pa_y)\Big] \\
&\quad + \sum_{X_k \in ch(Y)} \mathbb{E}_{Q_{\setminus Y}}\Big[\boldsymbol{\eta}_{X_k,Y}(x, cp(x_k))^T \boldsymbol{T}_Y(y) + \lambda_k(x_k, cp_{x_k,y})\Big] + \text{const.} \\
&= \left[\mathbb{E}_{Q_{\setminus Y}}\Big[\boldsymbol{\eta}_Y(pa_Y)^T\Big] + \sum_{X_k \in ch(Y)} \mathbb{E}_{Q_{\setminus Y}}\Big[\boldsymbol{\eta}_{X_k,Y}(x, cp_{x_k,y})^T\Big]\right]^T \boldsymbol{T}_Y(y) \\
&\quad + \log h_Y(pa_y) + \text{const.}
\end{aligned}$$

It follows that $Q^{new}(y)$ is in the exponential family of the same form as $P(y \mid pa_y)$ but with parameter function

$$\boldsymbol{\eta}_Y^{new} = \mathbb{E}_{Q_{\setminus Y}}\Big[\boldsymbol{\eta}_Y(pa_y)\Big] + \sum_{X_k \in Ch(Y)} \mathbb{E}_{Q_{\setminus Y}}\Big[\boldsymbol{\eta}_{x_k,Y}(x_k, cp_{x_k,y})\Big].$$

1 A function is multi-linear if it depends linearly with respect each variable.

As the expectations of $\boldsymbol{\eta}_Y$ and $\boldsymbol{\eta}_{X_k,Y}$ are multi-linear functions of the expectations of the statistic functions of their corresponding variables, it is possible to reparameterize these functions in terms of these expectations

$$\bar{\boldsymbol{\eta}}_Y(\{\mathbb{E}[\boldsymbol{T}_{X_k}(x_k)]\}_{X_k \in pa_Y}) = \mathbb{E}\left[\boldsymbol{\eta}_Y(pa_Y)\right]$$
$$\bar{\boldsymbol{\eta}}_{X_k,Y}(\mathbb{E}[\boldsymbol{T}_{X_k}(x_k)], \{\mathbb{E}[\boldsymbol{T}_{X_j}(x_j)]\}_{X_j \in cp_{X_k}}) = \mathbb{E}\left[\boldsymbol{\eta}_{X_k,cp_{X_k}}(x_k, cp_{x_k})\right]$$

19.3 VARIATIONAL MESSAGE PASSING ALGORITHM

The message from a parent node Y to a child node X is the expectation under Q of its statistic vector

$$\boldsymbol{m}_{Y \to X} = \mathbb{E}\left[\boldsymbol{T}_Y(y)\right].$$

The message from a child node X to a parent node Y is:

$$\boldsymbol{m}_{X \to Y} = \boldsymbol{\eta}_{\bar{X},Y}(\mathbb{E}\left[\boldsymbol{T}_X(x)\right], \{\boldsymbol{m}_{X_k \to X}\}_{X_k \in cp_Y})$$

which relied on having received all messages from all the co-parents. If a node X is observed, the messages defined above are defined as $\boldsymbol{T}_A(a)$ instead of $\mathbb{E}\left[\boldsymbol{T}_A(a)\right]$.

Example 6. If X is Gaussian distributed and Y, β are its parents, the messages are:

$$\boldsymbol{m}_{X \to Y} = \begin{pmatrix} \mathbb{E}\left[\beta\right]\mathbb{E}\left[X\right] \\ -\mathbb{E}\left[\beta\right]/2 \end{pmatrix}, \quad \boldsymbol{m}_{X \to \beta} = \begin{pmatrix} -\frac{1}{2}\left(\mathbb{E}\left[x^2\right] - 2\mathbb{E}\left[x\right]\mathbb{E}\left[y\right] + \mathbb{E}\left[y^2\right]\right) \\ \frac{1}{2} \end{pmatrix}.$$

And the messages from X to any of its child nodes is

$$\begin{pmatrix} \mathbb{E}\left[\beta\right]\mathbb{E}\left[x\right] \\ -\mathbb{E}\left[x^2\right] \end{pmatrix}.$$

When a node Y has received all messages from its parents and children, we can compute the updated parameter $\bar{\eta}_Y^{new}$ as

$$\eta_Y^{new} = \bar{\boldsymbol{\eta}}_Y(\{\boldsymbol{m}_{X_k \to Y}\}_{X_k \in pa_Y}) + \sum_{X_k \in ch_Y} \boldsymbol{m}_{X_k \to Y}.$$

Winn & Bishop (2005)

Part VI

COMMONLY STUDIED LATENT VARIABLE MODELS

20

GAUSSIAN MIXTURE

One of the most studied models is the Gaussian Mixture we reviewed in Chapter 11, its elements were

- A corresponding set of observations $\boldsymbol{x} = \{x_1, \ldots, x_n\}$.
- The cluster assignment latent variables $\boldsymbol{Z} = \{Z_1, \ldots, Z_N\}$.
- The mixture weights $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_K\}$, i.e, prior probability of a particular component k.
- Each normal distribution $\mathcal{N}(\mu_k, \Lambda_k)$.

The joint probability factorizes as

$$P(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = P(\boldsymbol{x} \mid \boldsymbol{z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) P(\boldsymbol{z} \mid \boldsymbol{\pi}) P(\boldsymbol{\pi}) P(\boldsymbol{\mu} \mid \boldsymbol{\Lambda}) P(\boldsymbol{\Lambda}).$$

We are now in situation to give the explicit Bayesian network for this model:

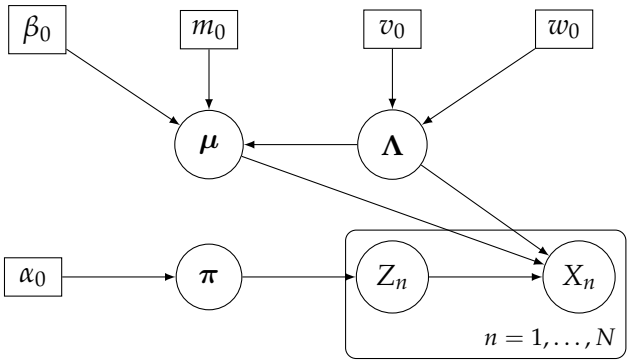


Figure 12: Gaussian mixture model. Squares represent hyper-parameters.

21

LATENT DIRICHLET ALLOCATION

Latent Dirichlet allocation or *LDA* is a conditionally conjugate model (figure 13) in natural language processing that allows set of observations to be explained by unobserved groups that explain why some parts of the data are similar.

For example, observations may be words in a document, which is a mixture of a small number of topics and each word's presence is attributable to one of the document's topics. Learning corresponds to extract information as the set of topics, their associated word probabilities, the topic of each word, and the particular topic mixture of each document.

The considered elements are (using Hoffman *et al.* (2013) and Blei *et al.* (2003) notation):

- K number of topics, V number of words in the vocabulary, M number of documents, N_d number of words in document d and N total number of words.
- $\boldsymbol{\beta} = \{\beta_1, \dots, \beta_K\}$, where β_k is the distribution of words in topic k. Each component $\beta_{k,n}$ is the probability of the n^{th} word in topic k.
- Each document d is associated with a vector of topic proportions θ_d, which is a $K-1$ simplex. Then each component $\theta_{d,k}$ is the probability of topic k in document d. Denote $\boldsymbol{\theta} = \{\theta_1, \dots, \theta_K\}$.
- Each word in each document is assumed to be related with a single topic. The variable $Z_{d,n}$ indexes the topic of the n^{th} word in the d^{th} document.

LDA model assumes that each document is generated with the following generative process:

1. Draw topics from a Dirichlet distribution, for each $k = 1, \dots, K$:

$$P(\beta_k) = \frac{1}{B(\eta)} \prod_{v=1}^{V} \beta_{k,v}^{\eta-1} \implies \beta_k \sim \text{Symmetric-Dirichlet}_V(\eta)$$

2. For each document $d = 1, \dots, D$:
   a) Draw topic proportions,

$$P(\theta_d) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} \theta_{d,k}^{\alpha-1} \implies \theta_d \sim \text{Symmetric-Dirichlet}_K(\alpha)$$

   b) For each word in the document $n = 1, \dots, N_d$:

i. Draw a topic,

$$P(z_{d,n} \mid \theta_d) = \prod_{k=1}^{K} \theta_{d,k}^{\mathbb{I}[z_{d,n}=k]} \implies (Z_{d,n} \mid \theta_d) \sim \text{Categorical}(\theta_d).$$

ii. Draw a word form the topic

$$P(w_{d,n} \mid z_{d,n}, \boldsymbol{\beta}) = \prod_{v=1}^{V} \beta_{z_{d,n},v}^{\mathbb{I}[w_{d,n}=v]} \implies (W_{d,n} \mid Z_{d,n}, \boldsymbol{\beta}) \sim \text{Categorical}(\beta_{Z_{d,n}})$$

The joint probability distribution is then

$$\begin{aligned} P(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w}, \boldsymbol{\beta}) &= P(\boldsymbol{\beta}) \prod_{d=1}^{D} P(\theta_d \mid \alpha) \prod_{n=1}^{N_d} P(z_{d,n} \mid \theta_d) P(w_{d,n} \mid z_{d,n}, \boldsymbol{\beta}) \\ &= \Big(\prod_{k=1}^{K} P(\beta_k) \Big) \prod_{d=1}^{D} P(\theta_d) \prod_{n=1}^{N_d} P(z_{d,n} \mid \theta_d) P(w_{d,n} \mid z_{d,n}, \boldsymbol{\beta}) \end{aligned}$$

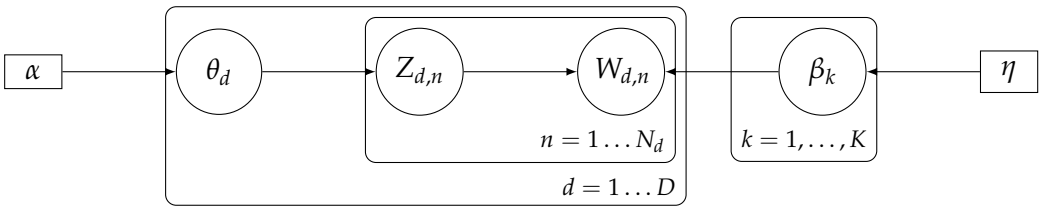


Figure 13: Latent Dirichlet Allocation model. Squares represent hyper-parameters

We can compute the posterior distributions for this model, starting with the complete conditional distribution of the local hidden variables. They only depend on other variables in the local context (same document) and the global variables

$$\begin{aligned} P(z_{d,n} \mid w_{d,n}, \theta_d, \boldsymbol{\beta}) &= \frac{P(z_{d,n}, w_{d,n}, \theta_d, \boldsymbol{\beta})}{\int_{z_{n,d}} P(z_{d,n}, w_{d,n}, \theta_d, \boldsymbol{\beta})} = \frac{P(z_{d,n} \mid \theta_d) P(w_{d,n} \mid z_{d,n}, \boldsymbol{\beta})}{\int_{z_{d,n}} P(z_{d,n} \mid \theta_d) P(w_{d,n} \mid z_{d,n}, \boldsymbol{\beta})} \\ &= \frac{\prod_{k=1}^{K} \theta_{d,k}^{\mathbb{I}[z_{d,n}=k]} \prod_{v=1}^{V} \beta_{z_{d,n},v}^{\mathbb{I}[w_{d,n}=v]}}{\int_{z_{d,n}} \prod_{k=1}^{K} \theta_{d,k}^{\mathbb{I}[z_{d,n}=k]} \prod_{v=1}^{V} \beta_{z_{d,n},v}^{\mathbb{I}[w_{d,n}=v]}} = \frac{\theta_{d,z_{d,n}} \beta_{z_{d,n},w_{d,n}}}{\sum_{k=1}^{K} \theta_{d,k} \beta_{k,w_{d,n}}} \end{aligned}$$

Naming

$$\gamma_{d,n} = \frac{\theta_{d,z_{d,n}} \beta_{z_{d,n} w_{d,n}}}{\sum_{k=1}^{K} \theta_{d,k} \beta_{k,w_{d,n}}}$$

and $\boldsymbol{\gamma} = \{\gamma_{d,n}\}_{d=1,\ldots,D \; n=1,\ldots,N_d}$, we get

$$(Z_{d,n} \mid w_{d,n}, \theta_d, \boldsymbol{\beta}) \sim \mathit{Categorical}(\boldsymbol{\gamma}).$$

On the other hand, the complete conditional of the topic proportions θ_d is only affected by the topic appearances, since $z_{d,n}$ is an indicator vector, the k^{th} element of the parameter to

this Dirichlet is the sum of the hyperparameter α and the number of words assigned to topic k in document d:

$$P(\theta_d \mid \boldsymbol{z_d}, \boldsymbol{w_d}, \boldsymbol{\beta}) = \frac{P(\theta_d, \boldsymbol{z_d}, \boldsymbol{w_d}, \boldsymbol{\beta})}{\int_{\theta_d} P(\theta_d, \boldsymbol{z_d}, \boldsymbol{w_d}, \boldsymbol{\beta})} = \frac{P(\theta_d) \prod_{n=1}^{N_d} P(z_{d,n} \mid \theta_d)}{\int_{\theta_d} P(\theta_d) \prod_{n=1}^{N_d} P(z_{d,n} \mid \theta_d)}$$
$$\propto \prod_{k=1}^{K} \theta_{d,k}^{\alpha-1} \prod_{n=1}^{N_d} \theta_{d,k}^{\mathbb{I}[z_{d,n}=k]} = \prod_{k=1}^{K} \theta_{d,k}^{\alpha-1+\sum_{n=1}^{N_d} \mathbb{I}[z_{d,n}=k]}.$$

The complete conditional depends only on the topic assignments

$$(\theta_d \mid z_d) \sim \text{Dirichlet}_K\Big(\alpha + \sum_{n=1}^{N_d} \mathbb{I}[z_{d,n} = 1], \ldots, \alpha + \sum_{n=1}^{N_d} \mathbb{I}[z_{d,n} = K]\Big).$$

In words, the probability of a topic in document d is updated with the number of times that topic appears in the document. The complete conditional of a topic depends on the words and topics assignments of the entire collection

Using a similar reasoning, the words distribution in a topic k, β_k, is updated with the number of appearances in all documents of the given topic.

$$(\beta_k \mid \boldsymbol{w}, \boldsymbol{z}) \sim \text{Dirichlet}_V\Big(\eta + \sum_{d=1}^{D} \sum_{n=1}^{N_d} \mathbb{I}[z_{d,n} = k]\mathbb{I}[w_{d,n} = 1], \ldots, \eta + \sum_{d=1}^{D} \sum_{n=1}^{N_d} \mathbb{I}[z_{d,n} = k]\mathbb{I}[w_{d,n} = V]\Big).$$

22

PROBABILISTIC PRINCIPAL COMPONENTS ANALYSIS

LVMs have usually been restricted to the exponential family because, in this case, inference is feasible. But recent advances in variational inference have enabled LVMs to be extended with neural networks. For example, *Variational Auto-encoders* or *VAE* (Kingma & Welling (2013)) are the most influential models combining both concepts.

VAEs extent the classical technique of *principal components analysis* for data representation in lower-dimensional spaces. Suppose we have a D-dimensional representation of a data point x and z is its latent K-dimensional representation ($K < D$). PCA computes an affine transformation $\boldsymbol{W}$, represented by a $K \times D$ matrix.

A probabilistic view of PCA can be modeled with an LVM (Tipping & Bishop (1999)), with the following elements:

- $\boldsymbol{X} = \{X_1, \ldots, X_N\}$ i.i.d $\mathbb{R}^D$-valued random variables and the corresponding observations $\boldsymbol{x} = \{x_1, \ldots, x_N\}$.
- $\boldsymbol{Z} = \{Z_1, \ldots, Z_N\}$ i.i.d latent $\mathbb{R}^K$-valued random variables, where Z_n models the K-dimensional representation of x_n.
- A global latent $K \times D$-dimensional random variable $\boldsymbol{W}$.
- A noise hyper-parameter σ^2.

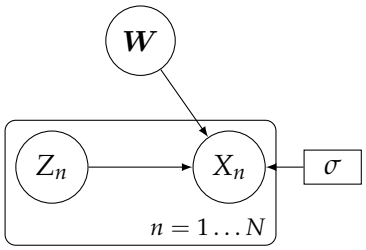


Figure 14: Probabilistic PCA model

We assume the priors are normally distributed:

$$Z_n \sim N_K(0, I) \quad \forall n = 1, \ldots, N \quad \text{and} \quad \boldsymbol{W} \sim N_{K \times D}(0, I).$$

The data points are considered generated via a projection,

$$(X_n \mid Z_n, \boldsymbol{W}) \sim N(\boldsymbol{W}^T Z_n, \sigma^2 I) \quad \forall n = 1, \ldots, N.$$

The probabilistic model extends the classical one in the way that the latter assumes the noise is infinitesimally small, i.e, $\sigma^2 \to 0$. The *expectation-maximization algorithm* (Section 8) is commonly used to solve this variational inference problem.

22.1 ARTIFICIAL NEURAL NETWORKS

An *artificial neural network* or *ANN* with L hidden layers can be seen as a deterministic non-linear function f parameterized by a set of matrix $\boldsymbol{W} = \{\boldsymbol{W}_0, \ldots, \boldsymbol{W}_L\}$ and non-linear activation functions $\{r_0, \ldots, r_L\}$. Given an input x the output y is calculated has

$$h_0 = r_0(\boldsymbol{W}_0^T x), \quad \ldots, \quad h_l = r_l(\boldsymbol{W}_l^T h_{l-1}) \quad \ldots \quad y = r_L(\boldsymbol{W}_L^T h_{L-1}).$$

Deep neural networks or *DNNs* are ANNs where the number of hidden layers is higher. Commonly, any neural network with more that 2 hidden layers is considered deep. Given a dataset $\{(x_1, y_1), \ldots, (x_N, y_N)\}$ and a loss function $l(y, y^*)$ that defines how well the output $y^* = f_{\boldsymbol{W}}(x)$ returned by the model matches the real output y, learning reduces to the optimization problem

$$\boldsymbol{W}^{opt} = \arg\min_{\boldsymbol{W}} \sum_{n=1}^{N} l(y_n, f_{\boldsymbol{W}}(x_n)).$$

This problem is usually solved by applying a variant of the stochastic gradient descent method, which involves the computation of the gradient of the loss function with respect to the parameters of the network. The algorithm for computing this gradient is known as *back-propagation*, which is based on a recursive application of the chain-rule of derivates. This can be implemented using the computational graph on the network.

The main idea of a computational graph is to express a deterministic function, as is the case of a neural network, using an acyclic directed graph. It is composed of input, output and operation nodes, where model data and parameters are shown as input nodes.

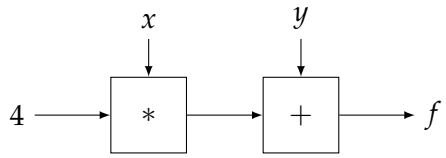


Figure 15: Computational graph example of function $f(x, y) = 4x + y$

22.2 NON-LINEAR PCA

Non-linear PCA or NLPCA, extends the classical PCA where the relation between the low dimensional space and the observed data is governed by a DNN instead of a linear transformation. It can be seen as a non-linear probabilistic PCA model.

The model is quite similar to the one presented for the PCA, the difference comes from the conditional distribution of X, that depends on Z through a fully-connected ANN with a single hidden layer.

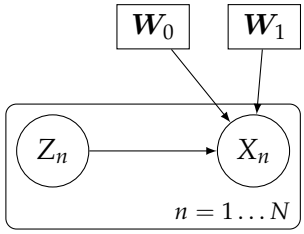


Figure 16: Non-linear PCA model. $\boldsymbol{W}_0$ and $\boldsymbol{W}_1$ together with two activation functions r_0, r_1 represent an ANN.

The prior of the latent variable is a centered Gaussian

$$Z_n \mid N(0, I) \quad \forall n \in 1, \dots, N$$

Let D be the dimension of the data X and K the dimension of the hidden variable Z.Let f a single hidden layer ANN with input dimension Z and output dimension D. Where the output is the mean value of the normal distribution under X.

As we are considering a single hidden layer, let $\boldsymbol{W}_0$ and $\boldsymbol{W}_1$ be the matrixes governing that ANN and r_0, r_1 the activation functions, the ANN f is

$$f(z_n) = r_1(\boldsymbol{W_1}(r_0(\boldsymbol{W}_0(z_n)))),$$

and the data-points are then generated as

$$(X_n \mid Z_n) \sim N(f(Z_n), I) \quad \forall n \in 1, \dots, N.$$

Where no noise is being considered this time.

22.3 VARIATIONAL AUTO-ENCODER

Similarly to the models PCA and NLPCA, a *variational autoencoder* or VAE, allows to perform dimensionality reduction. However a VAE will contain a neural network in the P model (decoder) and another one in the variational model Q (encoder).

The P model has the same structure as the nonlinear PCA. On the other hand, the distribution Q is defined with a reverse ANN, with input dimension D and output dimension K.

The neural networks can be defined to give an output of double the dimension, that is, the encoder would give a point in with $2K$ components, so the first K are used for the mean and the last K sor the variance of the normal distribution.

Part VII

CASE STUDY

23

INFERPY USAGE

`InferPy` (Cózar *et al.* (2019)) is a high-level API for probabilistic modeling written in `Python` and capable of running on top of `Edward` and `Tensorflow`. `InferPy`'s API is strongly inspired by `Keras` and it has a focus on enabling flexible data processing, easy-to-code probabilistic modeling, scalable inference and robust model validation.

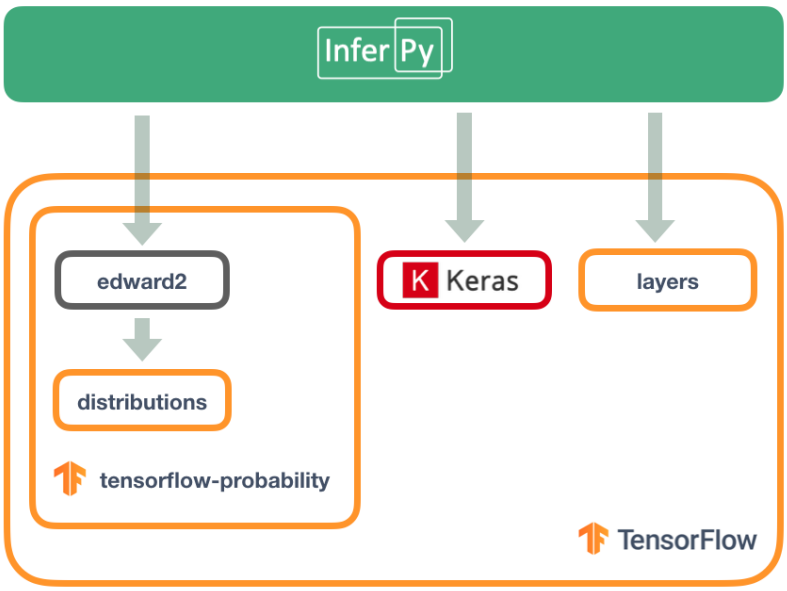


Figure 17: InferPy architecture.

The main features of `InferPy` are:

- Allows a simple definition over probabilistic models containing or not neural networks.
- All models that can be defined using `Edward2` can also be defined in `InferPy`, whose probability distributions are mainly inherited from `tensorflow-probability`.
- All the models parameters can be defined using the standard `Python` types (`Numpy` compatibility).
- `InferPy` relies on top of `Edward`'s inference engine, therefore, it includes all the inference algorithms avaible in that package.
- An additional advantage of using `Edward` and `TensorFlow` as inference engine is that all the parallelization details are hidden to the user. Moreover, the same code will run either in CPUs or GPUs.

23.1 INSTALLATION

`InferPy` has the following requirements:

- `Python` $\geq$ 3.5 and $<$ 3.8.
- `Tensorflow` $\geq$ 1.12.1 and $<$ 2.0.
- `Tensorflow-probability` 0.7.0.
- `NetworkX` $\geq$ 2.2.0 and $<$ 3.0.

`InferPy` is available at `Pip` and can be installed with the following command:

```
pip install inferpy
```

23.2 USAGE GUIDE WITH PCA

In this section we are constructing a PCA inference model with `InferPy`. Programming and `Python` knowledge are assumed, the guidance will be over the API usage.

Let us remember the variables we had in the model, consider $X_1, \dots, X_N$ i.i.d $\mathbb{R}^D$ variables, these are the observed ones. Then there are $Z_1, \dots, Z_N$ hidden variables that correspond to the hidden representation of the data in $\mathbb{R}^K$. There is also a global hidden variable $\boldsymbol{W}$ modeling the linear transformation from one space to the other. We are also considering another global variable $\boldsymbol{W}_0$ that will allow the model to generate non centered points.

The hidden variables prior is a centered gaussian and the data is supposed to be generated as

$$X_n \mid z_n, \boldsymbol{w}, \boldsymbol{w_0} \sim \mathcal{N}(\boldsymbol{w}^T z_n + \boldsymbol{w_0}, I)$$

No noise is being considered in this example.

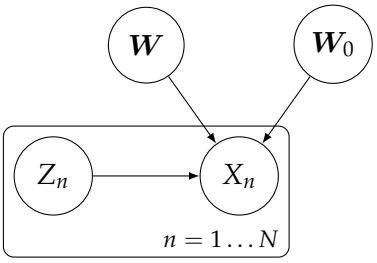


Figure 18: Probabilistic PCA model. No noise considered.

The database we are using is `Mnist`, which is a largely used dataset on a set of handwritten digits.

The full model definition would be the following:

```
@inf.probmodel
def pca(k,d):
    w =  inf.Normal(loc=tf.zeros([k,d]),
                    scale=1, name="w")                    # shape = [k,d]
    w0 = inf.Normal(loc=tf.zeros([d]),
                     scale=1, name="w0")                   # shape = [d]
    with inf.datamodel():
```

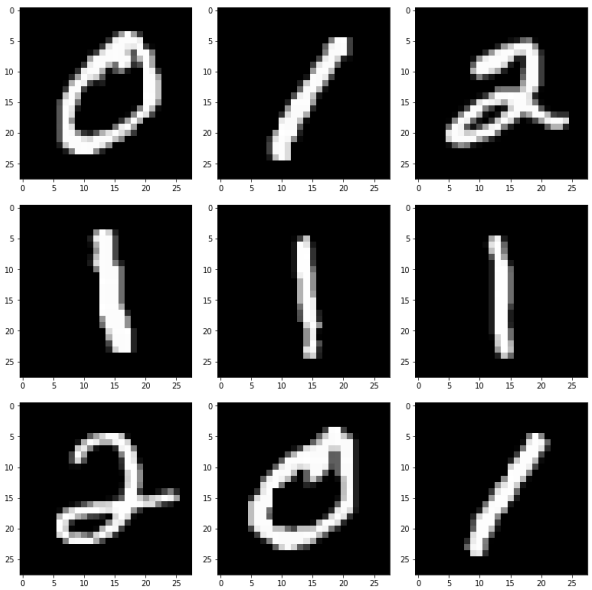


Figure 19: Mnist dataset example.

```
z = inf.Normal(tf.zeros([k]),1, name="z")          # shape = [N,k]
x = inf.Normal(np.dot(z,w) + w0, 1, name="x")      # shape = [N,d]
```

Models are defined by decorating a `Python` functions with `@inf.probmodel`. In this case we are defining a model with 2 parameters:

```
@inf.probmodel
def pca(k, d):
```

Now we declare the global hidden variables, we use the distributions inside `InferPy` which is assumed to be imported as `inf`, in this case `inf.Normal`. This distribution has three arguments, the mean (`loc`), the standard deviation (`scale`), and the name of the variable, this last parameter is needed as its how this model and the variational will communicate.

If either the mean or the standard desviation are given as an array, a constant value on the other parameter would be interpreted as a constant array of the corresponding size.

```
w =  inf.Normal(loc=tf.zeros([k,d]), scale=1, name="w")
```

Here we are defining a $K \times D$ Gaussian distributed variable with mean 0 and standard deviation 1, named "w".

We have reached the point where we have to define a set of i.i.d random variables, one for each observation, `InferPy` gives a explicit syntaxis for this, variables defined inside `with inf.datamodel(size)` are replicated and i.i.d of each other. The size can be omitted as it is calculated from the dataset. This is how we declare the hidden local variables in our model

```
with inf.datamodel():
    z = inf.Normal(tf.zeros([k]),1, name="z")
```

Our generative model is now defined, now we might define the variational one:

```
@inf.probmodel
def Q(k,d):
```

```
qw_loc = inf.Parameter(tf.zeros([k,d]), name="qw_loc")
qw_scale = tf.math.softplus(inf.Parameter(tf.ones([k,d]), name="qw_scale"))
qw = inf.Normal(qw_loc, qw_scale, name="w")

qw0_loc = inf.Parameter(tf.ones([d]), name="qw0_loc")
qw0_scale = tf.math.softplus(inf.Parameter(tf.ones([d]), name="qw0_scale"))
qw0 = inf.Normal(qw0_loc, qw0_scale, name="w0")

with inf.datamodel():
    qz_loc = inf.Parameter(np.zeros([k]), name="qz_loc")
    qz_scale = tf.math.softplus(inf.Parameter(tf.ones([k]), name="qz_scale"))
    qz = inf.Normal(qz_loc, qz_scale, name="z")
```

In this case, among the variables, we must define each parameter. Let us focus on "w" whose variational distribution is a Gaussian distribution, we must define two parameters, one for the mean and other for the standard deviation.

```
qw_loc = inf.Parameter(tf.zeros([k,d]), name="qw_loc")
qw_scale = tf.math.softplus(inf.Parameter(tf.ones([k,d]), name="qw_scale"))
qw = inf.Normal(qw_loc, qw_scale, name="w")
```

As the standard deviation value must not reach a negative value, we use a `softplus` function to smoothly approximate a rectifier function.

The same argument is applied to the rest of variables.

When both models are defined and instantiated in a variable, we need to create an inference object:

```
VI = inf.inference.VI(qmodel)
```

By default, the inference object uses el ELBO and loss function and uses `AdamOptimizer` from `TensorFlow`. An amount of `epochs` for the fitting might be set.

Given a training dataset `X_train`, the model is trained indicating which set of observations corresponds to each observed variable using the following syntax:

```
pmodel.fit({"x": X_train}, VI)
```

Once the model is trained, a variable posterior can be taken as

```
pmodel.posterior("z").parameters()
```

We can also take a sample from the parameter posterior

```
post = {"z": pmodel.posterior("z").sample()}
```

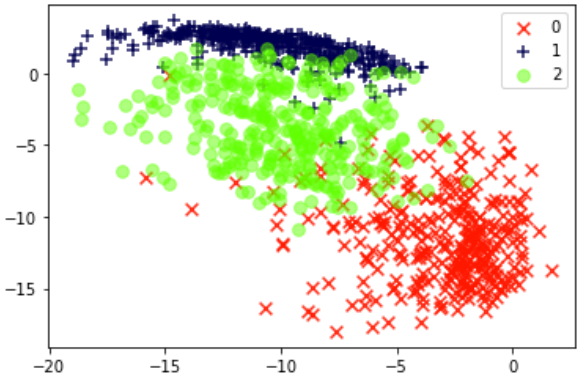


Figure 20: PCA posterior sample of MNIST.

24

GAUSSIAN MIXTURE

Part VIII

ANNEXES

A

DISTRIBUTIONS IN THE EXPONENTIAL FAMILY (WIP)

Distribution	Parameter set $\boldsymbol{\theta}$	Base measure h	Parameter function $\boldsymbol{\eta}$	Statistics $\boldsymbol{T}$
Bernoulli	p	1	$\log \frac{p}{1-p}$	x
Binomial	p	$\binom{n}{x}$	$\log \frac{p}{1-p}$	x
Categorical	$p_1, \ldots, p_K$	1	$\begin{pmatrix} \log p_1 \\ \vdots \\ \log p_K \end{pmatrix}$	$\begin{pmatrix} [x=1] \\ \vdots \\ [x=k] \end{pmatrix}$
Gaussian	μ, σ^2	$\frac{1}{\sqrt{2\pi}}$	$\begin{pmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{pmatrix}$	$\begin{pmatrix} x \\ x^2 \end{pmatrix}$
Beta	α, β	$\frac{1}{x(1-x)}$	$\begin{pmatrix} \alpha \\ \beta \end{pmatrix}$	$\begin{pmatrix} \log x \\ \log(1-x) \end{pmatrix}$
Dirichlet	$\alpha_1, \ldots, \alpha_K$	1	$\begin{pmatrix} \alpha_1 - 1 \\ \vdots \\ \alpha_K - 1 \end{pmatrix}$	$\begin{pmatrix} \log x_1 \\ \vdots \\ \log x_K \end{pmatrix}$

Figure 21: Some distributions in the exponential family

BIBLIOGRAPHY

Bandyopadhyay, Prasanta S, & Forster, Malcolm R. 2011. Philosophy of statistics: An introduction. *Pages 1–50 of: Philosophy of statistics*. Elsevier.

Barber, David. 2007. *Bayesian Reasoning and Machine Learning*. Cambridge University Press.

Bishop, Christopher M. 2006. *Pattern recognition and machine learning*. springer.

Blei, David M. 2014. Build, compute, critique, repeat: Data analysis with latent variable models. *Annual Review of Statistics and Its Application*, **1**, 203–232.

Blei, David M, Ng, Andrew Y, & Jordan, Michael I. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, **3**(Jan), 993–1022.

Blei, David M, Kucukelbir, Alp, & McAuliffe, Jon D. 2017. Variational inference: A review for statisticians. *Journal of the American statistical Association*.

Cox, David Roxbee. 2006. *Principles of statistical inference*. Cambridge university press.

Cózar, Javier, Cabañas, Rafael, Salmerón, Antonio, & Masegosa, Andrés R. 2019. InferPy: Probabilistic Modeling with Deep Neural Networks Made Easy. *arXiv preprint arXiv:1908.11161*.

Do, Chuong B, & Batzoglou, Serafim. 2008. What is the expectation maximization algorithm? *Nature biotechnology*, **26**(8), 897–899.

Farrow, Malcolm. 2008. MAS3301 Bayesian Statistics.

Grimmett, Geoffrey R. 1973. A theorem about random fields. *Bulletin of the London Mathematical Society*, **5**(1), 81–84.

Hoffman, Matthew D, Blei, David M, Wang, Chong, & Paisley, John. 2013. Stochastic variational inference. *The Journal of Machine Learning Research*, **14**(1), 1303–1347.

Jordan, Michael I, Ghahramani, Zoubin, Jaakkola, Tommi S, & Saul, Lawrence K. 1999. An introduction to variational methods for graphical models. *Machine learning*.

Kingma, Diederik P, & Welling, Max. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Koller, Daphne, & Friedman, Nir. 2009. *Probabilistic Graphical Models, Principles and Techniques*. The MIT Press.

Koopman, Bernard Osgood. 1936. On distributions admitting a sufficient statistic. *Transactions of the American Mathematical society*, **39**(3), 399–409.

Li, Stan Z. 2009. *Markov random field modeling in image analysis*. Springer Science & Business Media.

McLachlan, Geoffrey J, & Krishnan, Thriyambakam. 2007. *The EM algorithm and extensions*. Vol. 382. John Wiley & Sons.

Neal, Radford M, & Hinton, Geoffrey E. 1998. A view of the EM algorithm that justifies incremental, sparse, and other variants. *Pages 355–368 of: Learning in graphical models*. Springer.

Pearl, Judea, & Dechter, Rina. 2013. Identifying Independences in Casual Graphs with Feedback.

Rossi, Richard J. 2018. *Mathematical statistics: an introduction to likelihood based inference*. John Wiley & Sons.

Shachter, Ross D. 2013. Bayes-Ball: The Rational Pastime.

Tipping, Michael E, & Bishop, Christopher M. 1999. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **61**(3), 611–622.

Upton, Graham, & Cook, Ian. 2014. *A dictionary of statistics 3e*. Oxford university press.

Wainwright, Martin J., & Jordan, Michael I. 2008. *Graphical Models, Exponential Families and Variational Inference*. Now Publishers Inc.

Winn, John, & Bishop, Christopher M. 2005. Variational message passing. *Journal of Machine Learning Research*, **6**(Apr), 661–694.