Cédric Beaulac

Inference on latent variable models

University of Toronto

Department of Statistical Science

1 Introduction

As the type of data set researcher acquires contains more observations, more variables and more complex structures, statisticians and computer scientists needs to establish more flexible models and think of creative tools to allow for inference under the curse of dimensionality and the big data problems. In the following set of notes we will present the technicalities behind graphical models and the multiple inferences tools that exist to fit those model to a data set.

We will begin by introducing graphical models, explaining their strength and some of their properties. We will then explain how exact computation of the total likelihood can be computed under certain conditions before introducing the Expectation-Maximisation algorithm as the cornerstone of inference on simple graphical model. Finally, we will discuss how the graphical structure can be used as a generative models and how dimensionality reduction is used in that sense.

When the posterior distribution of the latent variable is intractable, new inference methods must be used. Variational Bayesian Inference is a popular solution to this problem and recent development of Variationnal Inference for neural Network (VAE) made possible the inference on complex graphical models with flexible approaches. Sampling methods such has Importance Sampling or Monte Carlo Markov Chains (MCMC) are alternative appraches to solve problem when expectation with respect to Intractable distribution must be computed.

Finally, we will discuss some interesting applications of graphical models. Modelling the confounder variable has been already been discuss using those tools and can prove useful in causal inference. We will also be looking at modelling spatial correlation with graphical model and finally uses those model as generative models in constructing interesting, general and powerful Procedural Content Generator.

2 Graphical Models

2.1 Introduction

Even though the central topic of this document is inference on graphical models and piratical applications, let us begin be noting that graphical model are actually *artificial*. What is meant by *artificial* is the fact that they are actually never required as all of the probabilistic inference can always be solved no matter how the variable structure is represented.

That being said, *Probabilistic Graphical Models* have a lot to offer in the sense that they provide a very simple way to visualize the structure of a probabilistic model. Insights into the properties of the probabilistic model like conditional independence can be easily illustrated with a graph. Finally, the complex computations required for inference can also be expressed in terms of graphical manipulations.

In a probabilistic graphical models, each node represents a random variable (or a group of random variable) and the edges express some probabilistic relationship between the variables.

2.2 Bayesian Networks and Random Markov Fields

By simple application of the product rule of probability, we can write joint distributions as a product a conditional distributions :

$$p(x, y, z) = p(z|x, y)p(y|x)p(x) \tag{1}$$

A Bayesian Network is a directed graphical model that implies a *natural factorization* of the joint distribution. Here is graphical model associated with the factorization of equation 1 :

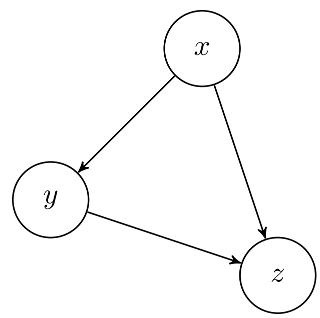


Figure 1: Graphical representation of the factorization in equation 1

Mathematically speaking, there exist multiple correct factorization for the joint $p(x, y, z)$ but the purpose of these model is to suggest a factorization and proceed with the inference according to that designated structure. Thus the factorization of interest of a joint distribution can easily be obtain by taking a quick look at the associated graph. In this situation for a graph with K nodes the joint distribution is given by :

$$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k|\mathrm{pa}(k)), \tag{2}$$

where $\mathbf{x} = \{x_1, ..., x_k\}$ and $\mathrm{pa}(k)$ denotes the set of parent nodes of x_k. Let us note that the directed graph we are considering are subject to and important restriction, there must be no direct cycles, i.e. we will be mostly looking at directed acyclic graphs.

A Markov Random field is an undirected graphical model. These will be left out for now.

2.3 Bayesian Network as generative models

It's interesting to use a Bayesian Network and its associated factorization (2) in order to generate observations from the joint distribution, this technique is called ancestral sampling.

The process is really simple. We begin by sampling from the marginal distribution of the random variables without parents. Then we proceed to successively sample from the conditional distribution in which the parent variables have been set to their sampled values. Being able to generate new observations can be useful in many scenarios that will be discussed later on.

2.4 Conditional Independence

An important property that is inherent to certain factorization is conditional independence. We say that x is conditionally independent of y given z if

$$p(x, y|z) = p(x|z)p(y|z). \tag{3}$$

The following model :

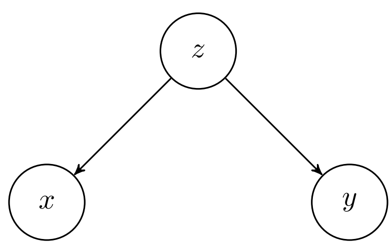


Figure 2: Illustration of conditional independence within a graph.

leads to the following factorization :

$$p(x, y, z) = p(x|z)p(y|z)p(z) \tag{4}$$

and naturally express this property as we can easily demonstrate :

$$
\begin{aligned}
p(x, y|z) &= \frac{p(x, y, z)}{p(z)} \\
&= \frac{p(x|z)p(y|z)p(z)}{p(z)} \\
&= p(x|z)p(y|z).
\end{aligned} \tag{5}
$$

On thing to notice is that the graphical representation of the probabilistic model can be used to visualize the conditional independence. More rigorously, the concept of D-separation aims at defining conditional independence based of the graphical structure of a model.

Let's consider a general directed graph in which X, Y and Z are non-intersecting sets of nodes and we would like to use the graphical representation to understand if there exist some conditional independence between the sets of nodes we are interested in. Quickly, if all path from any nodes of X to any nodes from Y are *blocked* by some nodes in Z then we will claim that X is conditionally independent of Y given Z. In other words, X and Y are D-separated given Z if no information can flow form X to Y without crossing Z.

A simple example of d-separation if provided by the concept of i.i.d. data set. If we are interested in estimation μ the mean of the distribution of an observed sample $\mathbf{S} = \{x_1, ..., x_N\}$. It is important to understand the observations are independent given the distribution. Graphically speaking, the joint $p(\mathbf{S}, \mu)$ can be defined by a prior $p(\mu)$ over the parameter and a set of independent conditional distribution $p(x_n|\mu)$, $n = 1, .., N$, i.e $p(S|\mu) = \prod_{n=1}^{N} p(x_n|\mu)$. This can graphically represented as :

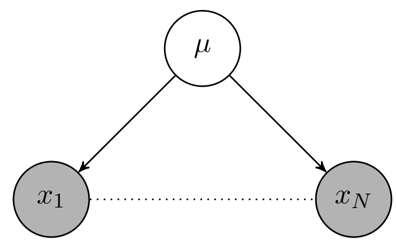


Figure 3: Graphical representation of an i.i.d. data set with parameter μ

This gives us the perfect opportunity to introduce some graphical notations. It is typical to use represent observed variables as shaded nodes and latent variables as white one. Latent variable are unobserved variables, often they are of interest to us as the graphical representation will be used to defined their relationship with the observed data. A *plate* is also typically used to represent a set of unlinked variables. Therefore, the figure above would typically be depicted as :

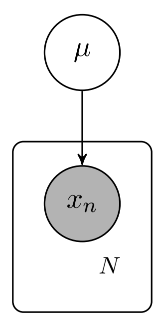


Figure 4: Graphical representation of an i.i.d. data set with parameter μ using the plate notation.

In the following sections we will introduce models with latent variables as they are the main motivation for most of the following content of this set of notes.

3 Latent Variable Models

3.1 Discrete Latent Variables

We will begin by introducing models with discrete latent variable. Latent variables are unobserved variable that are in place to allow for a more flexible observation distribution. As the models are including more variable to allow for greater flexibility, the uses of a graphical representation for the model becomes more important as a visualization tool.

To begin, we will introduce a classic discrete latent variable model, the mixture model. By allowing the observations to be distributed according to a linear superposition of various models, we greatly increases the number of shape that the distribution $p(\mathbf{x})$ can take. Let's define $\mathbf{z}$ as a discrete latent variable representing the various component of the mixture. Then we assume a distribution of $\mathbf{x}$ given $\mathbf{z}$ and thus obtain the following model :

$$p(x, z|\theta) = p_(x|z, \theta)p(z|\theta), \tag{6}$$

where θ represents the set of parameters for the distributions. In a fully Bayesian set up, these parameters are considered random variables and prior are established on those variables leading to the following model :

$$p(x, z, \theta) = p(x|z, \theta)p(z|\theta)p(\theta). \tag{7}$$

For now, let's stick to a frequentist approach. Assuming $\mathbf{z}$ can take K different values, to find the marginal distribution over $\mathbf{x}$ we have to integrate out $\mathbf{z}$:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k p(\mathbf{x}|\theta_k), \tag{8}$$

where $\pi_k = p(z = k)$, the probability that the observation was generated form the kth component.

The most used mixture model is by far the mixture of Gaussian where the conditional distributions $p_{\theta_k}(x)$ are Gaussian. In this case, the components differs in their parameters $\theta_k = (\mu_k, \Sigma_k)$. It is typical to define $\mathbf{z}$ as a k-dimensional binary random variable where one particular element z_k is set to 1 and all other elemetns are equal to 0. The parameter of this distribution is π, define such that $\pi_k = p(z_k = 1)$, with $0 \leq \pi \leq 1$ and $\sum_1^K \pi_k = 1$. Finally the marginal and conditional distributions are :

$$p(\mathbf{z}|\pi) = \prod_{k=1}^{K} \pi_k^{z_k}, \tag{9}$$

$$p(\mathbf{x}|\mathbf{z}, \mu, \Sigma) = N(\mathbf{x}|\mu_k, \Sigma_k), \tag{10}$$

which leads to the well known marginal distribution :

$$p(\mathbf{x}|\pi, \mu, \Sigma) = \sum_{\mathbf{z}} p(\mathbf{z}|\pi) p(\mathbf{x}|\mathbf{z}, \mu, \Sigma) = \sum_{k=1}^{K} \pi_k N(\mathbf{x}|\mu_k, \Sigma_k), \tag{11}$$

Suppose we have a data set $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ and we wish to model these observations with a mixture of Gaussian. The graphical model of such modelling would look like :

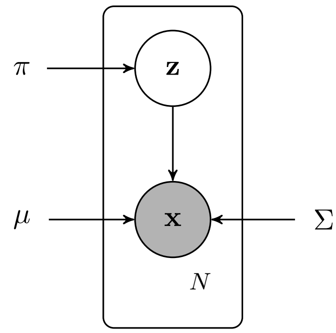


Figure 5: Graphical representation of an i.i.d. data set being fit to a mixture of Gaussian in a frequentist set up (the parameters are fixed unknown scalar, not random variables).

We would like a technique that can efficiently estimate the value of all the parameters. A classic statistical approach is to select the set of parameters that maximises the likelihood of the observed data :

$$\begin{aligned} p(\mathbf{X}|\pi, \mu, \Sigma) &= \prod_{n=1}^{N} p(\mathbf{x}_n|\pi, \mu, \sigma) \\ &= \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k N(\mathbf{x}|\mu_k, \Sigma_k) \\ \Rightarrow \ln p(\mathbf{X}|\pi, \mu, \Sigma) &= \sum_{n=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_k N(\mathbf{x}|\mu_k, \Sigma_k) \right) \end{aligned} \tag{12}$$

which is of course impossible to maximize with simple tools because of the summation inside the logarithmic. This is why we will use a special decomposition of the likelihood function. This decomposition is the center-piece of inference on graphical models.

3.1.1 The ELBO-KL decomposition

Let us demonstrate to popular likelihood decomposition that we will use. Notice that these equations hold for any distribution $q(\mathbf{Z})$ and we've dropped the parameters θ

for readability :

$$
\begin{aligned}
\ln p(\mathbf{X}) &= \ln\left(p(\mathbf{X},\mathbf{Z})/p(\mathbf{Z}|\mathbf{X})\right) \\
&= \ln\left(p(\mathbf{X},\mathbf{Z})\right) - \ln\left(p(\mathbf{Z}|\mathbf{X})\right) \\
&= \ln\left(p(\mathbf{X},\mathbf{Z})\right) - \ln\left(p(\mathbf{Z}|\mathbf{X})\right) + \ln q(\mathbf{Z}) - \ln q(\mathbf{Z}) \\
&= \ln\left(\frac{p(\mathbf{X},\mathbf{Z})}{q(\mathbf{Z})}\right) - \ln\left(\frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})}\right) \\
\Rightarrow \mathbf{E}_{q(\mathbf{Z})}[\ln p(\mathbf{X}|\theta)] &= \mathbf{E}_{q(\mathbf{Z})}\left[\ln\left(\frac{p(\mathbf{X},\mathbf{Z}|\theta)}{q(\mathbf{Z})}\right)\right] - \mathbf{E}_{q(\mathbf{Z})}\left[\ln\left(\frac{p(\mathbf{Z}|\mathbf{X},\theta)}{q(\mathbf{Z})}\right)\right] \\
\Rightarrow \ln p(\mathbf{X}|\theta) &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln\left(\frac{p(\mathbf{X},\mathbf{Z}|\theta)}{q(\mathbf{Z})}\right) - \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln\left(\frac{p(\mathbf{Z}|\mathbf{X},\theta)}{q(\mathbf{Z})}\right) \\
&= \mathcal{L}(q,\theta) + KL(q||p).
\end{aligned}
\tag{13}
$$

Notice that since the KL divergence is greater or equal than 0, then, $\mathcal{L}(q)$ is a lower bound for the likelihood. It is defined as the evidence lower bound (ELBO) or as the variational lower bound. Almost all techniques for inference on graphical models is based upon the maximisation of this lower bound.

3.1.2 The EM algorithm : Maximization of the variational lower bound for tractable posterior.

The Expectation-Maximization algorithm is an iterative procedure that slowly increases the value of the variational lower bound with two distinct steps. We will see the general idea and it how the algorithm intends to maximize the likelihood and then we will see how this procedure leads to a maximization of the variational lower bound.

We've already discussed the issue of maximizing the likelihood of the observed data set in models with latent variables. Remember that :

$$\ln p(\mathbf{X}|\theta) = \ln\left(\sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\theta)\right)$$

and thus maximizing the likelihood is analytically impossible. For now, let's assume that the complete-data set contains both $\mathbf{X}$ and $\mathbf{Z}$ and that the complete log likelihood $\ln p(\mathbf{X},\mathbf{Z}|\theta)$ is straight forward to maximize. Since we only observed $\mathbf{X}$, the only information we have about $\mathbf{Z}$ is through the posterior distribution of the latent $p(\mathbf{Z}|\mathbf{X},\theta)$. Therefore we cannot directly use the complete-data log likelihood and instead we will compute the expectation of the complete log likelihood under the posterior distribution with the current set of parameters:

$$\mathbf{E}_{p(\mathbf{Z}|\mathbf{X},\theta^{old})}[\ln p(\mathbf{X},\mathbf{Z}|\theta)] = \mathcal{Q}(\theta,\theta^{old}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\theta^{old}) \ln p(\mathbf{X},\mathbf{Z}|\theta) \tag{14}$$

which is the $\mathbf{E}$ step of the EM algorithm. Then, we will proceed at maximizing $\mathcal{Q}(\theta,\theta^{old})$ with respect to θ, the $\mathbf{M}$ step. For a mixture of Gaussian, the results are quiet simple :

$$p(z_n = k|\mathbf{x}_n,\theta^{old}) = \gamma(z_k) = \frac{\pi_k N(\mathbf{x}_n|\mu_k,\Sigma_k)}{\sum_{j=1}^{K} \pi_j N(\mathbf{x}_n|\mu_j,\Sigma_j)} \tag{15}$$

and then optimizing $\mathcal{Q}(\theta,\theta^{old})$ is easier and leads to the following estimates :

$$\begin{aligned}
\mu_k^{new} &= \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\mathbf{x}_n \\
\Sigma^{new} &= \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(\mathbf{x}_n-\mu_k^{new})(\mathbf{x}_n-\mu_k^{new})^T \\
\pi_k^{new} &= \frac{N_k}{N}
\end{aligned}$$

where $N_k = \sum_{n=1}^{N} \gamma(z_{nk})$.

Now we are going to use the ELBO-KL decomposition to demonstrate how this technique succeed at maximizing the likelihood and we will motivate the need for other techniques.

In section 3.1.1, we've demonstrated that $\mathcal{L}(q, \theta)$ is a lower bound for the log-likelihood of the observed data $\ln p(\mathbf{X}|\theta)$ defined as a function of the parameters θ and a distribution over the latent variables $q(\mathbf{Z})$. here we will now demonstrate how every step of the EM algorithm increase $\mathcal{L}(q, \theta)$. In the **E** step, we maximize $\mathcal{L}(q, \theta)$ with respect to $q(\mathbf{Z})$ while in the **M**, we then maximize $\mathcal{L}(q, \theta)$ with respect to θ.

The **E** step consist of considering the effect of **Z** through the posterior distribution $p(\mathbf{Z}|\mathbf{X}, \theta)$ under the current set of parameters, and then compute the expectation of the complete log-likelihood under that posterior distribution. This in fact maximize $\mathcal{L}(q, \theta)$ with respect to $q(\mathbf{Z})$ by setting it to $p(\mathbf{Z}|\mathbf{X}, \theta)$. Since $\mathcal{L}(q, \theta) = \ln p(\mathbf{X}|\theta) - KL(q||p)$ we see that by setting $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \theta)$, the KL divergence vanishes which effectively maximize $\mathcal{L}(q, \theta)$. This also highlights one of the main assumption necessary to use an EM algorithm, we need to be able to compute $p(\mathbf{Z}|\mathbf{X}, \theta)$ which is impossible for many choices of prior $p(\mathbf{Z})$ and conditional distribution $p(\mathbf{X}|\mathbf{Z})$.

In the following **M** step, we maximize $\mathcal{L}(q, \theta)$ with respect to the parameters θ. This will surely increases the value of $\mathcal{L}(q, \theta)$, unles it is already at a maximum, but it will also increases the value of the KL divergence term. Since we keep $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \theta^{old})$ fixed and update the parameters the two distribution compared in this term will be different and thus the divergence will become greater than 0 again. Since both the ELBo and KL divergence term are going to increase by updating the parameter to θ^{new} the increase in log-likelihood is greater than the increase in the lower bound. Let's actually see what happens when we substitute $q(\mathbf{Z})$ by $p(\mathbf{Z}|\mathbf{X}, \theta^{old})$ in the lower bound :

$$
\begin{aligned}
\mathcal{L}(q,\theta) &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln\left(\frac{p(\mathbf{X},\mathbf{Z}|\theta)}{q(\mathbf{Z})}\right) \\
\Rightarrow \mathcal{L}(p(\mathbf{Z}|\mathbf{X},\theta^{old}),\theta) &= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\theta^{old}) \ln\left(\frac{p(\mathbf{X},\mathbf{Z}|\theta)}{p(\mathbf{Z}|\mathbf{X},\theta^{old})}\right) \\
&= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\theta^{old}) \ln p(\mathbf{X},\mathbf{Z}|\theta) - \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\theta^{old}) p(\mathbf{Z}|\mathbf{X},\theta^{old}) \\
&= \mathcal{Q}(\theta,\theta^{old}) + \text{const.}
\end{aligned}
$$

Thus the **M** step maximizes the the expectation of the complete-data log-likelihood under the posterior distribution of the latent as intended.

In this section, we've briefly introduced discrete latent variables models and an example of such model with the mixture of Gaussian. We've also introduced a parameter inference algorithm, the EM algorithm, that works under some mild condition, the necessity to compute the posterior distribution of the latent variable. Finally, we've introduce a decomposition of the observed data log-likelihood that gives us a general framework for inference on latent variable models. In the next sections we will discuss the variational lower bound optimization with new tools and under a very general framework, but before we will quickly discuss continuous latent variables models.

3.2 Continuous Latent Variables

An important motivation for continuous latent variable models is dimensionality reduction. As we will see, many data sets have the property that the data points all lie close to a manifold of much lower dimension that the original data space. Being able to fully grasp the statistical property of a data set of large dimension using a set of continuous latex variables of lower dimension would then be very useful to analyse these complex data set.

A standard dimensionality reduction procedure is the principal component analysis (PCA) which we will introduce in this section. Then we will formulate the dimensionality reduction problem in probabilistic set up and discuss inference for those models. Finally we will introduce auto-encoders.

3.3 Principal Component Analysis

Consider a data set contained in a matrix $\mathbf{X}$ with N row and D column. Our goal is to project this data onto a space of much lower dimensionality $M << D$ while maximizing the variance on the projected space, so to capture as best as possible the variability within the original space.

Let's begin with the case where we try to project the data onto a 1-dimensional space. We will try to define a unit vector $\mathbf{u}_1$ representing the orientation containing th highest variability in the data set. Each data point $\mathbf{x}_n$ is projected onto that one dimensional space via $\mathbf{u}_1^T\mathbf{x}_n$ and so the mean of the projected data is $\mathbf{u}_1^T\bar{\mathbf{x}}$. This implies that the variance of the projected data set is given by $\mathbf{u}_1^T C\mathbf{u}_1$ where C is the covariance matrix of the D-dimension data set.

Since we want the orientation that produce a maximum of variability we will have to maximize $\mathbf{u}_1^T C\mathbf{u}_1$ with respect to $\mathbf{u}_1$ while including the normalization condition that $\mathbf{u}_1^T\mathbf{u}_1 = 1$, therefore we must maximize :

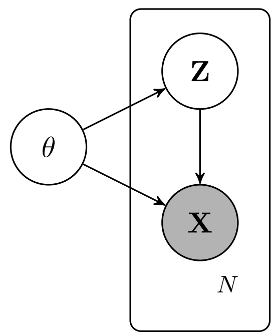
$$\mathbf{u}_1^T C\mathbf{u}_1 + \lambda_1(1 - \mathbf{u}_1^T\mathbf{u})$$

which leads to $C\mathbf{u}_1 = \lambda_1\mathbf{u}_1$. This implies that $\mathbf{u}_1$ must be a eigenvector of C and that λ_1 must be a eigenvalue. By left-multiplying by $\mathbf{u} - 1^T$ we observe that $\mathbf{u}_1^T C\mathbf{u}_1 = \lambda_1$ and therefore to obtain the highest variability λ_1 must be the biggest eigenvalue and therefore $\mathbf{u}_1$ is the eigenvector associated with the largest eigenvalue.

We could show in a similar manner that if we'd like to project the data on a M dimensional space using matrix U of size $D \times M$ we would have to construct it using the eigenvectors associated with the M largest eigenvalues.

3.4 Probabilistic PCA

A more general framework would be to assume that there exist a latent variable $\mathbf{Z}$ of dimension M that contributed in generating our D-dimensional observations $\mathbf{X}$ and we would like to fit the right parameters for the observed and latent variables distributions.



Again, maximizing the variational lower bound is one way to proceed. Instead, we will quickly introduce a set up where we can retrieve the classical PCA results using a probabilistic modelling.

Assume a Gaussian prior distribution over the latent variables $\mathbf{z}$:

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})$$

and a Gaussian conditional distribution for the observed variable $\mathbf{x}$:

$$p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|\mathbf{W}\mathbf{z} + \mu, \sigma^2\mathbf{I}).$$

The first thing to notice is that with these distributions modelling choices, we have the ability to extract exact form for the marginal distribution of the observed data $p(\mathbf{x})$ and for the posterior distribution of the latent variables $p(\mathbf{z}|\mathbf{x})$.

$$
\begin{aligned}
p(\mathbf{x}) &= \mathcal{N}(\mathbf{x}|\mu, \mathbf{W}\mathbf{W}^T + \sigma^2 I) \\
p(\mathbf{z}|\mathbf{x}) &= \mathcal{N}(\mathbf{z}|\mathbf{M}^{-1}\mathbf{W}^T(\mathbf{x} - \mu), \sigma^{-2}\mathbf{M})
\end{aligned}
$$

where $\mathbf{M} = \mathbf{W}^T\mathbf{W} + \sigma^2\mathbf{I}$.

Since we have access to all of those distributions, exact log-likelihood maximization is actually possible and here are the following parameters estimates :

$$
\begin{aligned}
\mu_{ML} &= \bar{\mathbf{x}} \\
\mathbf{W}_{ML} &= \mathbf{U}(\mathbf{L} - \sigma^2\mathbf{I})^{1/2}\mathbf{R} \\
\sigma^2_{ML} &= \frac{1}{D - M} \sum_{i=M+1}^{D} \lambda_i
\end{aligned}
$$

where $\mathbf{U}$ is a $D \times M$ matrix whose columns are given by the of size M of the eigenvector associated with M largest eigenvalues. $\mathbf{L}$ is a $M \times M$ diagonal matrix for which non-zero elements are the corresponding eigenvalues and $\mathbf{R}$ s an arbitrary $M \times M$ orthogonal matrix. We also notice the nice intuitive properties of σ^2_{ML}, it is the average variance associated with the deleted dimensions.

Even though we've retrieve the same projecting matrix, having built it upon a probabilistic set up offers a great deal of benefits. The probabilistic set up under a full Bayesian treatment will give us tools to automatically find the dimensionality of the principal subspace, the model can be run as a generative model to provide samples from the distribution, the existence of a likelihood function allow for direct comparison with other probabilistic model, we can derive en EM algorithm for efficient computation and to include mixtures of probabilistic PCA models and much more.

3.5 Auto-Encoder

Consider a Neural Network taking as input D-dimensional observations, and that outputs also a D-dimensional vector. The most simple network has only one hidden layer. If we set the dimensionality of the hidden layer to $M < D$ then we have :

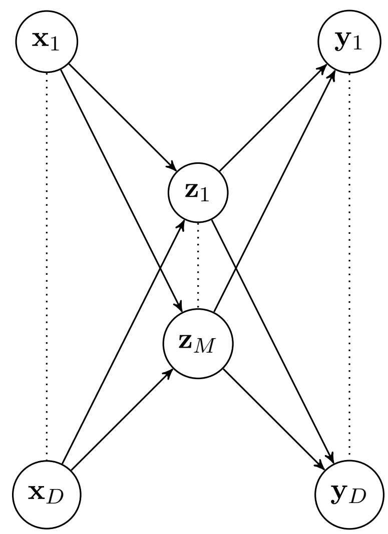


Figure 6: A simple Auto-Encoder

The middle layer containing the hidden unit can be perceived as our latent variables and since the number of hidden nodes is smaller than the number of inputs, a perfect reconstruction should not be possible. We will therefore find the optimal parameters $\mathbf{w}$ of the network so that it minimizes an error function that captures the degree of mismatch between the input vector and the output of the network. A simple error function could a sum-of-squares error :

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} ||y(\mathbf{x}_n, \mathbf{w}) - \mathbf{x}_n||^2 \tag{16}$$

If the hidden nodes have linear activations functions, it can be shows that the error function has a unique global minimum and this minimum performs a projection onto the M-dimensional space spanned by the M largest principal components.

So far, the uses of a Network structure resulted in the same solution that we've obtained with different approach. One way to utilize the Network structure is by adding layer with non-linear activation functions.

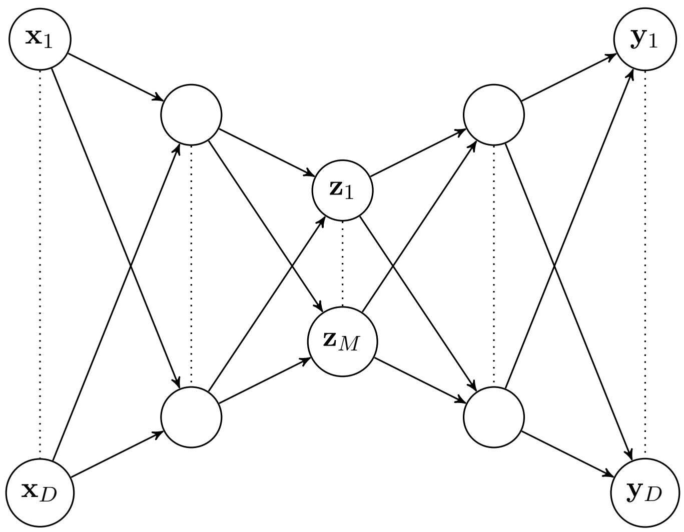


Figure 7: An Auto-Encoder with three hidden layer.

By creating a network with non-linear activation function we allow for *curved* orientation for the M-dimensional space. It allows for more interesting combination of the inputs and will hopefully grasp even more information about the original D-dimensional observations.

The optimization problem is now a lot more complicated. We will need new tools to solves that system, variational inference technique adapted for auto-encoder have been recently developed, we will address those in the next section.

4 Variational Inference

In section 3.1.2 we've explained the EM algorithm, an algorithm that uses the expectation of the complete data log-likelihood computed with respect to the posterior distribution of the latent variables to approximate the log-likelihood of the observed data. The algorithm then maximize this approximation with respect to the parameters. We've shown how this procedure succeed at maximizing the log-likelihood of the data. We've also noted that in order to use the EM algorithm, we have to be able to compute the posterior over the latent variables $p(\mathbf{z}|mathbfx)$.

In this section, we will introduce a general framework for inference on latent variables models that relies on maximizing the variational lower bound. Remember the ELBO-KL decomposition demonstrated in section 3.1.1 :

$$\begin{aligned}\ln p(\mathbf{X}) &= \int_{\mathbf{Z}} q(\mathbf{Z}) \ln \left(\frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right) d\mathbf{Z} - \int_{\mathbf{Z}} q(\mathbf{Z}) \ln \left(\frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})} \right) d\mathbf{Z} \\ &= \mathcal{L}(q) + KL(q||p),\end{aligned} \tag{17}$$

where we've assumed we are now in a fully Bayesian set up were the parameter θ are random variables and therefore we've included them in the set of latent variables $\mathbf{z}$.

We've already shown that $\mathcal{L}(q)$ is a lower bound for the log-likelihood of the observed data and that the choice of $q(\mathbf{Z})$ that maximizes this lower bound is the posterior distribution $p(\mathbf{Z}|\mathbf{X})$. Since we cannot compute it, we instead consider a restricted family of distributions $q(\mathbf{Z})$ and then seek the member of this family that maximize $\mathcal{L}(q)$.

On way to restrict the family of approximating distribution could be by using a parametric distribution $q(\mathbf{Z}|\omega)$ and then try to optimize $Lq(\mathbf{Z}|\omega)$ as a function of ω.

4.1 Factorized Distributions

A popular restricted family of approximate distribution are factorized distributions. Considered a partition of the elements of $\mathbf{Z}$ into disjoin groups that we denotes $\mathbf{Z}_j$, where $j = 1, ..., J$. We then assume that these groups are independent and therefore we can factorize the q distribution :

$$q(\mathbf{Z}) = \prod_{j=1}^{J} q_j(\mathbf{Z}_j) \tag{18}$$

Among all the distributions that takes the form of 18, the goal is find the one that maximizes the lower bound $\mathcal{L}(q)$. In order to optimize $\mathcal{L}(q)$ with respect to all of the distributions $q_j(\mathbf{Z}_j$ we will optimize the lower bound with respect to each of the factors in turn. Let us now replace substitute $q(\mathbf{Z})$ in the lower bound and analyse it with respect to one factor $q_i(\mathbf{Z}_i)$ which we denote by q_i for simplicity :

$$\begin{aligned}
\mathcal{L}(q) &= \int_{\mathbf{Z}} q(\mathbf{Z}) \ln\left(\frac{p(\mathbf{X},\mathbf{Z})}{q(\mathbf{Z})}\right) \mathrm{d}\mathbf{Z} \\
&= \int_{\mathbf{Z}} \prod_{j=1}^{J} q_j \ln\left(\frac{p(\mathbf{X},\mathbf{Z})}{\prod_{j=1}^{J} q_j}\right) \mathrm{d}\mathbf{Z} \\
&= \int_{\mathbf{Z}} \prod_{j=1}^{J} q_j \left(\ln p(\mathbf{X},\mathbf{Z}) - \sum_{j=1}^{J} \ln q_j\right) \mathrm{d}\mathbf{Z} \\
&= \int_{\mathbf{Z}_i} q_i \left(\int_{\mathbf{Z}_{j\neq i}} \ln p(\mathbf{X},\mathbf{Z}) \prod_{j\neq i} q_j \mathrm{d}\mathbf{Z}_{j\neq i}\right) \mathrm{d}\mathbf{Z}_i - \int_{\mathbf{Z}_i} q_i \ln q_i \mathrm{d}\mathbf{Z}_i + \mathrm{const} \\
&= \int_{\mathbf{Z}_i} q_i \mathbf{E}_{j\neq i}[\ln p(\mathbf{X},\mathbf{Z})] \mathrm{d}\mathbf{Z}_i - \int_{\mathbf{Z}_i} q_i \ln q_i \mathrm{d}\mathbf{Z}_i + \mathrm{const}
\end{aligned} \tag{19}$$

where $\mathbf{E}_{j\neq i}[\ln p(\mathbf{X},\mathbf{Z})] = \int_{\mathbf{Z}_{j\neq i}} \ln p(\mathbf{X},\mathbf{Z}) \prod_{j\neq i} q_j \mathrm{d}\mathbf{Z}_{j\neq i}$. Now if we fix $\{q_{j\neq i}\}$ and maximize $\mathcal{L}(q)$ this is done by recognizing that the result of equation 19 is a negative Kullback-Leibler divergence between q_i and