Unsupervised Learning

E. Le Pennec

Fall 2023
Outline

1 Unsupervised Learning?
2 A First Glimpse
   • Clustering
   • Dimensionality Curse
   • Dimension Reduction
   • Generative Modeling
3 Dimension Reduction
   • Simplification
   • Reconstruction Error
   • Relationship Preservation
   • Comparing Methods?
4 Clustering
   • Prototype Approaches
   • Contiguity Approaches
   • Agglomerative Approaches
   • Other Approaches
   • Scalability
5 Generative Modeling
   • (Plain) Parametric Density Estimation
   • Latent Variables
   • Approximate Simulation
   • Diffusion Model
   • Generative Adversarial Network
6 References
Outline

1 Unsupervised Learning?

2 A First Glimpse
   - Clustering
   - Dimensionality Curse
   - Dimension Reduction
   - Generative Modeling

3 Dimension Reduction
   - Simplification
   - Reconstruction Error
   - Relationship Preservation
   - Comparing Methods?

4 Clustering
   - Prototype Approaches
   - Contiguity Approaches
   - Agglomerative Approaches
   - Other Approaches
   - Scalability

5 Generative Modeling
   - (Plain) Parametric Density Estimation
   - Latent Variables
   - Approximate Simulation
   - Diffusion Model
   - Generative Adversarial Network

6 References
Motivation

What is possible with data without labels?

- To group them?
- To visualize them in a 2 dimensional space?
- To generate more data?
**Marketing and Groups**

To group them?

- **Data:** Base of customer data containing their properties and past buying records
- **Goal:** Use the customers *similarities* to find groups.
- **Clustering:** propose an explicit *grouping* of the customers
- **Visualization:** propose a representation of the customers so that the groups are *visible*. (Bonus)
To visualize them?

- **Data**: Images of a single object
- **Goal**: Visualize the *similarities* between images.
- **Visualization**: propose a representation of the images so that similar images are *close*.
- **Clustering**: use this representation to cluster the images. (Bonus)
Images and Generation

To generate more data?

- **Data**: Images.
- **Goal**: Generate images similar to the ones in the dataset.
- **Generative Modeling**: propose (and train) a generator.
The classical definition of Tom Mitchell

A computer program is said to learn from experience $E$ with respect to some class of tasks $T$ and performance measure $P$, if its performance at tasks in $T$, as measured by $P$, improves with experience $E$. 
## Supervised Learning

### Experience, Task and Performance measure

- **Training data**: \( \mathcal{D} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\} \) (i.i.d. \( \sim \mathbb{P} \))
- **Predictor**: \( f : \mathcal{X} \rightarrow \mathcal{Y} \) measurable
- **Cost/Loss function**: \( \ell(f(X), Y) \) measure how well \( f(X) \) predicts \( Y \)
- **Risk**:

\[
\mathcal{R}(f) = \mathbb{E}[\ell(Y, f(X))] = \mathbb{E}_X \left[ \mathbb{E}_{Y|X}[\ell(Y, f(X))] \right]
\]

- Often \( \ell(f(X), Y) = \|f(X) - Y\|^2 \) or \( \ell(f(X), Y) = 1_{Y \neq f(X)} \)

### Goal

- Learn a rule to construct a **predictor** \( \hat{f} \in \mathcal{F} \) from the training data \( \mathcal{D}_n \) s.t. **the risk** \( \mathcal{R}(\hat{f}) \) is **small on average** or with high probability with respect to \( \mathcal{D}_n \).
Unsupervised Learning

Experience, Task and Performance measure

- **Training data**: \( D = \{X_1, \ldots, X_n\} \) (i.i.d. \( \sim P \))
- **Task**: ???
- **Performance measure**: ???

- No obvious task definition!

Tasks for this lecture

- **Dimension reduction**: construct a map of the data in a **low dimensional** space without **distorting** it too much.
- **Clustering (or unsupervised classification)**: construct a **grouping** of the data in **homogeneous** classes.
- **Generative modeling**: generate new conditional samples.
**Dimension Reduction**

- **Training data**: $\mathcal{D} = \{X_1, \ldots, X_n\} \in \mathcal{X}^n$ (i.i.d. $\sim \mathbb{P}$)
- Space $\mathcal{X}$ of possibly high dimension.

**Dimension Reduction Map**

- Construct a map $\Phi$ from the space $\mathcal{X}$ into a space $\mathcal{X}'$ of **smaller dimension**:
  \[ \Phi : \mathcal{X} \rightarrow \mathcal{X}' \quad \text{where} \quad X \mapsto \Phi(X) \]

- Map can be defined only on the dataset.

**Motivations**

- Visualization of the data
- Dimension reduction (or embedding) before further processing
Dimension Reduction

- Need to control the **distortion** between $\mathcal{D}$ and $\Phi(\mathcal{D}) = \{\Phi(X_1), \ldots, \Phi(X_n)\}$

### Distortion(s)

- **Reconstruction error:**
  - Construct $\tilde{\Phi}$ from $\mathcal{X}'$ to $\mathcal{X}$
  - Control the error between $X$ and its reconstruction $\tilde{\Phi}(\Phi(X))$

- **Relationship preservation:**
  - Compute a *relation* $X_i$ and $X_j$ and a *relation* between $\Phi(X_i)$ and $\Phi(X_j)$
  - Control the difference between those two *relations*.

- Lead to different constructions....
Clustering

- **Training data**: \( \mathcal{D} = \{X_1, \ldots, X_n\} \in \mathcal{X}^n \) (i.i.d. \( \sim \mathbb{P} \))
- Latent groups?

Clustering

- Construct a map \( f \) from \( \mathcal{D} \) to \( \{1, \ldots, K\} \) where \( K \) is a number of classes to be fixed:

\[
 f : \quad X_i \mapsto k_i
\]

- Similar to classification except:
  - no ground truth (no given labels)
  - label only elements of the dataset!

Motivations

- Interpretation of the groups
- Use of the groups in further processing
Clustering

- Need to define the **quality** of the cluster.
- No obvious measure!

### Clustering quality

- **Inner homogeneity**: samples in the same group should be similar.
- **Outer inhomogeneity**: samples in two different groups should be different.

- Several possible definitions of similar and different.
- Often based on the distance between the samples.
- Example based on the Euclidean distance:
  - **Inner homogeneity** = intra-class variance,
  - **Outer inhomogeneity** = inter-class variance.

**Beware**: choice of the number of clusters $K$ often complex!
Generative Modeling

- **Training data**: \( \mathcal{D} = \{ (X_1, Y_1), \ldots, (X_n, Y_n) \} \in (\mathcal{X} \times \mathcal{Y})^n \) (i.i.d. \( \sim \mathbb{P} \))
- Same kind of data than for supervised learning if \( \mathcal{Y} \neq \emptyset \).

### Generative Modeling

- Construct a map \( G \) from the product of \( \mathcal{Y} \) and a randomness source \( \Omega \) to \( \mathcal{X} \)
  \[ G : \mathcal{Y} \times \Omega \rightarrow \mathcal{X} \]
  \[ (Y, \omega) \mapsto X \]
- Unconditional model if \( \mathcal{Y} = \emptyset \).

### Motivation

- Generate plausible novel conditional samples based on a given dataset.

### Sample Quality

- Related to the proximity between the law of \( G(Y, \omega) \) and the law of \( X|Y \).
- Most classical choice is the Kullback-Leibler divergence.
Generative Modeling

Ingredients

- Generator $G_\theta(Y, \omega)$ and cond. density prob. $P_\theta(X|Y)$ (Explicit vs implicit link)
- Simple / Complex / Approximate estimation...

Some Possible Choices

<table>
<thead>
<tr>
<th>Probabilistic model</th>
<th>Generator</th>
<th>Estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>Simple (parametric)</td>
<td>Simple (ML)</td>
</tr>
<tr>
<td>Flow</td>
<td>Image of simple model</td>
<td>Simple (ML)</td>
</tr>
<tr>
<td>Factorization</td>
<td>Factorization of simple model</td>
<td>Simple (ML)</td>
</tr>
<tr>
<td>VAE</td>
<td>Simple model with latent var.</td>
<td>Simple (ML)</td>
</tr>
<tr>
<td>EBM</td>
<td>Arbitrary</td>
<td>Implicit (MCMC)</td>
</tr>
<tr>
<td>Diffusion</td>
<td>Continuous noise</td>
<td>Complex (ML/score/discrim.)</td>
</tr>
<tr>
<td>GAN</td>
<td>Discrete Noise with latent var.</td>
<td>Approximate (ML)</td>
</tr>
<tr>
<td></td>
<td>Implicit</td>
<td>Complex (Discrimination)</td>
</tr>
</tbody>
</table>

SOTA: Diffusion based approach!

Bonus Task: Representation Learning

- **General observation:** most data do not have a label!
- **Example:** The number of images on which someone has described the content of the image is a tiny fraction of the images online.
- Labeling is very expensive and time consuming
- A lot of information can be extracted from the structure of the data, before seeing any label.

How can we leverage the large quantity of un-labeled data?

- Learn relevant features (= representations) in an unsupervised fashion
- Use those features to solve a supervised task with a fraction of labeled data.

- **Semi-supervised framework**
- Very useful in practice, for images, time series, text.
Semi-supervised Framework

With representation learned in an unsupervised fashion + a simple linear model, one can achieve the same performance with 10% of data labeled than with a fully annotated dataset.

Complementary regularization based approaches also exist.
Unsupervised Learning is a Versatile Approach!

Except for generative modeling, the learner is always right

- A subjective measure of performance
- Subjective choices for the algorithmic constraints (e.g., the type of transformation of the data we allow for low-dimensional representation, type of groups in clustering)
  \[ \Rightarrow \text{Very difficult or impossible to tell which is the best method.} \]

Yet:
- Extremely important in practice:
  - 90-99% of the data is un-labeled!
  - the tasks themselves are fundamental
- Huge success in various fields (NLP, images...)

Unsupervised Learning?
Unsupervised Learning is a Versatile Approach!

**Lecture goals for the three main tasks**

- Discussing possible choices of measures of performance and algorithmic constraints
- Understand the correspondences between those choices and a variety of classical algorithms
- For the simplest algorithms (PCA, k-means), get a precise mathematical understanding of the learning process.
Outline

1. Unsupervised Learning?

2. A First Glimpse
   - Clustering
   - Dimensionality Curse
   - Dimension Reduction
   - Generative Modeling

3. Dimension Reduction
   - Simplification
   - Reconstruction Error
   - Relationship Preservation
   - Comparing Methods?

4. Clustering
   - Prototype Approaches
   - Contiguity Approaches
   - Agglomerative Approaches
   - Other Approaches
   - Scalability

5. Generative Modeling
   - (Plain) Parametric Density Estimation
   - Latent Variables
   - Approximate Simulation
   - Diffusion Model
   - Generative Adversarial Network

6. References
Outline

1 Unsupervised Learning?

2 A First Glimpse
   - Clustering
   - Dimensionality Curse
   - Dimension Reduction
   - Generative Modeling

3 Dimension Reduction
   - Simplification
   - Reconstruction Error
   - Relationship Preservation
   - Comparing Methods?

4 Clustering
   - Prototype Approaches
   - Contiguity Approaches
   - Agglomerative Approaches
   - Other Approaches
   - Scalability

5 Generative Modeling
   - (Plain) Parametric Density Estimation
   - Latent Variables
   - Approximate Simulation
   - Diffusion Model
   - Generative Adversarial Network

6 References
What’s a group?

- No simple or unanimous definition!
- Require a notion of similarity/difference...

Three main approaches

- A group is a set of samples similar to a prototype.
- A group is a set of samples that can be linked by contiguity.
- A group can be obtained by fusing some smaller groups...
Prototype Approach

- A group is a set of samples similar to a prototype.
- Most classical instance: $k$-means algorithm.
- Principle: alternate prototype choice for the current groups and group update based on those prototypes.

- Number of groups fixed at the beginning
- No need to compare the samples between them!
Contiguity Approach

- A group is the set of samples that can be linked by contiguity.
- Most classical instance: DBScan
- Principle: group samples by contiguity if possible (proximity and density)
- Some samples may remain isolated.
- Number of groups controlled by the scale parameter.

DBSCAN: Density-Based Spatial Clustering of Applications with Noise
A First Glimpse

Agglomerative Approach

- A group can be obtained by fusing some smaller groups...
- Hierarchical clustering principle: sequential merging of groups according to a best merge criterion
- Numerous variations on the merging criterion...
- Number of groups chosen afterward.

Source: upGrad
Choice of the method and of the number of groups

No method or number of groups is better than the others...

- Criterion not necessarily explicit!
- No cross validation possible
- Choice of the number of groups: a priori, heuristic, *based on the final usage*...
Outline

1. Unsupervised Learning?

2. A First Glimpse
   - Clustering
   - Dimensionality Curse
   - Dimension Reduction
   - Generative Modeling

3. Dimension Reduction
   - Simplification
   - Reconstruction Error
   - Relationship Preservation
   - Comparing Methods?

4. Clustering
   - Prototype Approaches
   - Contiguity Approaches
   - Agglomerative Approaches
   - Other Approaches
   - Scalability

5. Generative Modeling
   - (Plain) Parametric Density Estimation
   - Latent Variables
   - Approximate Simulation
   - Diffusion Model
   - Generative Adversarial Network

6. References
DISCLAIMER: Even if they are used everywhere beware of the usual distances in high dimension!

Dimensionality Curse

- Previous approaches based on distances.
- Surprising behavior in high dimension: everything is (often) as far away.
- Beware of categories...
A First Glimpse

**DISCLAIMER:** Even if they are used everywhere beware of the usual distances in high dimension!

### High Dimensional Geometry Curse

- **Folks theorem:** In high dimension, everyone is alone.
- **Theorem:** If $X_1, \ldots, X_n$ in the hypercube of dimension $d$ such that their coordinates are i.i.d then

\[
d^{-1/p} \left( \max \|X_i - X_j\|_p - \min \|X_i - X_j\|_p \right) = 0 + O_P \left( \sqrt{\frac{\log n}{d}} \right)
\]

\[
\frac{\min \|X_i - X_j\|_p}{\max \|X_i - X_j\|_p} = 1 + O_P \left( \sqrt{\frac{\log n}{d}} \right).
\]

- When $d$ is large, all the points are almost equidistant...
- Nearest neighbors are meaningless!

$n \leq 2^d$
Outline

1 Unsupervised Learning?

2 A First Glimpse
   • Clustering
   • Dimensionality Curse
   • Dimension Reduction
   • Generative Modeling

3 Dimension Reduction
   • Simplification
   • Reconstruction Error
   • Relationship Preservation
   • Comparing Methods?

4 Clustering
   • Prototype Approaches
   • Contiguity Approaches
   • Agglomerative Approaches
   • Other Approaches
   • Scalability

5 Generative Modeling
   • (Plain) Parametric Density Estimation
   • Latent Variables
   • Approximate Simulation
   • Diffusion Model
   • Generative Adversarial Network

6 References
Visualization and Dimension Reduction

- How to view a dataset in high dimension!
- High dimension: dimension larger than 2!
- **Projection** onto a 2D space.
Visualization and Dimension Reduction

- How to view a dataset in high dimension!
- High dimension: dimension larger than 2!
- Projection onto a 2D space.
Visualization and Dimension Reduction

- How to view a dataset in high dimension!
- High dimension: dimension larger than 2!
- Projection onto a 2D space.
Visualization and Dimension Reduction

- How to view a dataset in high dimension!
- High dimension: dimension larger than 2!
- **Projection** onto a 2D space.
A First Glimpse

Principal Component Analysis

Simple formula: \[ \tilde{X} = P(X - m) \]

How to chose \( P \)?

- Maximising the dispersion of the points?
- Allowing to well reconstruct \( X \) from \( \tilde{X} \)?
- Preserving the relationship between the \( X \) through those between the \( \tilde{X} \)?
Principal Component Analysis

Simple formula: \( \tilde{X} = P(X - m) \)

How to choose \( P \)?

- Maximising the dispersion of the points?
- Allowing to well reconstruct \( X \) from \( \tilde{X} \)?
- Preserving the relationship between the \( X \) through those between the \( \tilde{X} \)?

The 3 approaches yield the same solution!
Reconstruction Approaches

- Learn a formula to encode and one formula to decode.
- Auto-encoder structure

- Yields a formula for new points.
Reconstruction Approaches

- Learn a formula to encode and one formula to decode.
- Auto-encoder structure

Yields a formula for new points.
Reconstruction Approaches

- Learn a formula to encode and one formula to decode.
- Auto-encoder structure

Yields a formula for new points.
Relationship Preservation Approaches

- Based on the definition of the relationship notion (in both worlds).
- Huge flexibility

- Not always yields a formula for new points.
No Better Choice?

- Different criterion for different methods: impossible to use cross-validation.
- The larger the dimension the easier is to be faithful!
- In visualization, dimension 2 is the only choice.
- Heuristic criterion for the dimension choice: elbow criterion (no more gain), stability. . .

- Dimension Reduction is rarely used standalone but rather as a step in a predictive/prescriptive method.
- The dimension becomes an hyper-parameter of this method.
Representation Learning

- How to transform arbitrary objects into numerical vectors?
- Objects: Categorical variables, Words, Images/Sounds...

- The two previous dimension reduction approaches can be used (given possibly a first simple high dimensional representation)
Outline

1 Unsupervised Learning?

2 A First Glimpse
   - Clustering
   - Dimensionality Curse
   - Dimension Reduction
   - Generative Modeling

3 Dimension Reduction
   - Simplification
   - Reconstruction Error
   - Relationship Preservation
   - Comparing Methods?

4 Clustering
   - Prototype Approaches
   - Contiguity Approaches
   - Agglomerative Approaches
   - Other Approaches
   - Scalability

5 Generative Modeling
   - (Plain) Parametric Density Estimation
   - Latent Variables
   - Approximate Simulation
   - Diffusion Model
   - Generative Adversarial Network

6 References
Generative Modeling

- Generate new samples similar to the ones in an original dataset.
- Generation may be conditioned by a input.
- Key for image generation... and chatbot!
Heuristic: If we can estimate the (conditional) law $P$ of the data and can simulate it, we can obtain new samples similar to the input ones.

Estimation and Simulation

- How to estimate the density?
- How to simulate the estimate density?
- Other possibilities?
Simple Estimation and Simple Simulation

A First Glimpse

Parametric Model, Image and Factorization

- Use
  - a simple parametric model, ...
  - or the image of a parametric model (flow), ...
  - or a factorization of a parametric model (recurrent model)

as they are simple to estimate and to simulate.

- Estimation by Maximum Likelihood principle.
- Recurrent models are used in Large Language Models!
Latent Variable

- Generate first a (low dimensional) latent variable $Z$ from which $X$ is easy to sample.
- Estimation based on approximate Maximum Likelihood (VAE/ELBO)
- The latent variable can be generated by a simple method (or a more complex one...).
A First Glimpse

Complex Estimation and Complex Simulation

Monte Carlo Markov Chain

- Rely on much more complex probability model...
- which can only be simulated numerically.
- Often combined with noise injection to stabilizes the numerical scheme (Diffusion).

- Much more expensive to simulate than with Latent Variable approaches.
Generative Adversarial Network

- Bypass the density estimation problem, by transforming the problem into a competition between the generator and a discriminator.
- The better the generator, the harder it is for the generator to distinguish true samples from synthetic ones.
- No explicit density!
- Fast simulator but **unstable training**...
Outline

1. Unsupervised Learning?
2. A First Glimpse
   - Clustering
   - Dimensionality Curse
   - Dimension Reduction
   - Generative Modeling
3. Dimension Reduction
   - Simplification
   - Reconstruction Error
   - Relationship Preservation
   - Comparing Methods?
4. Clustering
   - Prototype Approaches
   - Contiguity Approaches
   - Agglomerative Approaches
   - Other Approaches
   - Scalability
5. Generative Modeling
   - (Plain) Parametric Density Estimation
   - Latent Variables
   - Approximate Simulation
   - Diffusion Model
   - Generative Adversarial Network
6. References
Dimension Reduction

- **Training data**: \( \mathcal{D} = \{X_1, \ldots, X_n\} \in \mathcal{X}^n \) (i.i.d. \( \sim \mathbb{P} \))
- Space \( \mathcal{X} \) of possibly high dimension.

**Dimension Reduction Map**

- Construct a map \( \Phi \) from the space \( \mathcal{X} \) into a space \( \mathcal{X}' \) of **smaller dimension**:
  \[
  \Phi : \mathcal{X} \to \mathcal{X}'
  \]
  \[
  X \mapsto \Phi(X)
  \]

**Criterion**

- Reconstruction error
- Relationship preservation
Outline

1 Unsupervised Learning?
2 A First Glimpse
   • Clustering
   • Dimensionality Curse
   • Dimension Reduction
   • Generative Modeling
3 Dimension Reduction
   • Simplification
   • Reconstruction Error
   • Relationship Preservation
   • Comparing Methods?
4 Clustering
   • Prototype Approaches
   • Contiguity Approaches
   • Agglomerative Approaches
   • Other Approaches
   • Scalability
5 Generative Modeling
   • (Plain) Parametric Density Estimation
   • Latent Variables
   • Approximate Simulation
   • Diffusion Model
   • Generative Adversarial Network
6 References
How to Simplify?

A Projection Based Approach

- Observations: \( X_1, \ldots, X_n \in \mathbb{R}^d \)
- Simplified version: \( \Phi(X_1), \ldots, \Phi(X_n) \in \mathbb{R}^d \) with \( \Phi \) an affine projection preserving the mean \( \Phi(X) = P(X - m) + m \) with \( P^T = P = P^2 \) and \( m = \frac{1}{n} \sum_i X_i \).

How to choose \( P \)?

- **Inertia criterion:**
  \[
  \max_P \sum_{i,j} \| \Phi(X_i) - \Phi(X_j) \|^2
  \]

- **Reconstruction criterion:**
  \[
  \min_P \sum_i \| X_i - \Phi(X_i) \|^2
  \]

- **Relationship criterion:**
  \[
  \min_P \sum_{i,j} |(X_i - m)^T (X_j - m) - (\Phi(X_i) - m)^T (\Phi(X_j) - m)|^2
  \]

- **Rk:** Best solution is \( P = I \)! Need to reduce the rank of the projection to \( d' < d \ldots \)
Inertia criterion

- **Heuristic**: a good representation is such that the projected points are far apart.

### Two views on inertia

- **Inertia**: 
  \[ I = \frac{1}{2n^2} \sum_{i,j} \|X_i - X_j\|^2 = \frac{1}{n} \sum_{i=1}^{n} \|X_i - m\|^2 \]

- 2 times the mean squared distance to the mean = Mean squared distance between individual

### Inertia criterion (Principal Component Analysis)

- **Criterion**: 
  \[ \max_P \sum_{i,j} \frac{1}{2n^2} \|PX_i - PX_j\|^2 = \max_P \frac{1}{n} \sum_i \|PX_i - m\|^2 \]

- **Solution**: Choose \( P \) as a projection matrix on the space spanned by the \( d' \) first eigenvectors of \( \Sigma = \frac{1}{n} \sum_i (X_i - m)(X_i - m)^\top \)
First Component of the PCA

- \( \tilde{X} = m + a^\top (X - m)a \) with \( \|a\| = 1 \)
- Inertia: \( \frac{1}{n} \sum_{i=1}^{n} a^\top (X_i - m)(X_i - m)^\top a \)

Principal Component Analysis: optimization of the projection

- Maximization of \( \tilde{l} = \frac{1}{n} \sum_{i=1}^{n} a^\top (X_i - m)(X_i - m)^\top a = a^\top \Sigma a \) with
  \[ \Sigma = \frac{1}{n} \sum_{i=1}^{n} (X_i - m)(X_i - m)^\top \text{ the empirical covariance matrix.} \]
- Explicit optimal choice given by the eigenvector of the largest eigenvalue of \( \Sigma \).
Principal Component Analysis: sequential optimization of the projection

- Explicit optimal solution obtained by the projection on the eigenvectors of the largest eigenvalues of $\Sigma$.
- Projected inertia given by the sum of those eigenvalues.

- Often fast decay of the eigenvalues: some dimensions are much more important than others.
- Not exactly the curse of dimensionality setting...
- Yet a lot of small dimension can drive the distance!
Reconstruction Criterion

- **Heuristic:** a good representation is such that the projected points are close to the original ones.

**Reconstruction Criterion**

- **Criterion:**
  \[
  \min_P \frac{1}{n} \sum_i \|X_i - (P(X_i - m) + m)\|^2 = \min_P \frac{1}{n} \sum_i ((I - P)(X_i - m))^2
  \]

- **Solution:** Choose \(P\) as a projection matrix on the space spanned by the \(d'\) first eigenvectors of \(\Sigma = \frac{1}{n} \sum_i (X_i - m)(X_i - m)^\top\)

- Same solution with a different heuristic!

- **Proof (Pythagora):**
  \[
  \sum_i \|X_i - m\|^2 = \sum_i \left( \|P(X_i - m)\|^2 + \|(I - P)(X_i - m)\|^2 \right)
  \]
PCA, Reconstruction and Distances

Close projection doesn’t mean close individuals!

- Same projections but different situations.
- Quality of the reconstruction measured by the angle with the projection space!
**Heuristic:** a good representation is such that the projected points scalar products are similar to the original ones.

**Relationship Criterion (Multi Dimensional Scaling)**

- **Criterion:** \( \min_P \sum_{i,j} |(X_i - m)^\top (X_j - m) - (\Phi(X_i) - m)^\top (\Phi(X_j) - m)|^2 \)

- **Solution:** Choose \( P \) as a projection matrix on the space spanned by the \( d' \) first eigenvectors of \( \Sigma = \frac{1}{n} \sum_i (X_i - m)(X_i - m)^\top \)

- Same solution with a different heuristic!
- Much more involved justification!
Dimension Reduction

- **PCA model:** \( X - m \simeq P(X - m) \)
  
  - **Prop:** \( P = VV^\top \) with \( V \) an orthonormal family in dimension \( d \) of size \( d' \).
  
- **PCA model with \( V \):** \( X - m \simeq VV^\top(X - m) \) where \( \tilde{X} = V^\top(X - m) \in \mathbb{R}^{d'} \)

- **Row vector rewriting:** \( X^\top - m^\top \simeq \tilde{X}^\top V^\top \)

---

### Matrix Rewriting and Low Rank Factorization

- **Matrix rewriting**

  \[
  \begin{bmatrix}
  X_1^\top - m^\top \\
  \vdots \\
  X_n^\top - m^\top
  \end{bmatrix}^{(n \times d)} \begin{bmatrix}
  \tilde{X}_1^\top \\
  \vdots \\
  \tilde{X}_n^\top
  \end{bmatrix}^{(n \times d')} \sim \begin{bmatrix}
  V^\top
  \end{bmatrix}^{(d' \times d)}
  \]

- **Low rank matrix factorization!** (Truncated SVD solution...)

---

**Link with SVD**
SVD Decomposition

- Any matrix $n \times d$ matrix $A$ can be decomposed as

$$A^{(n\times d)} = U^{(n\times n)} D^{(n\times d)} W^\top^{(d\times d)}$$

with $U$ and $W$ two orthonormal matrices and $D$ a diagonal matrix with decreasing values.

SVD
Low Rank Approximation

- The best low rank approximation or rank $r$ is obtained by restriction of the matrices to the first $r$ dimensions:

$$
\begin{bmatrix}
A \\
(n \times d)
\end{bmatrix} \overset{\sim}{\rightarrow} \begin{bmatrix}
U_r \\
(n \times r)
\end{bmatrix} \begin{bmatrix}
D_{r,r} \\
(r \times r)
\end{bmatrix} \begin{bmatrix}
W_r^T \\
(r \times d)
\end{bmatrix}
$$

for both the operator norm and the Frobenius norm!

- PCA: Low rank approximation with Frobenius norm, $d' = r$ and

$$
\begin{bmatrix}
X_1^T - m^T \\
\vdots \\
X_n^T - m^T
\end{bmatrix} \leftrightarrow A, \\
\begin{bmatrix}
\tilde{X}_1^T \\
\vdots \\
\tilde{X}_n^T
\end{bmatrix} \leftrightarrow U_r D_{r,r}, \\
V^T \leftrightarrow W_r^T
$$
SVD Decompositions

- Recentered data:
  \[ R = \begin{pmatrix} X_1^T - m^T \\ \vdots \\ X_n^T - m^T \end{pmatrix} = UDW^T \]

- Covariance matrix:
  \[ \Sigma = R^T R = WD^T DW \]
  with \( D^T D \) diagonal.

- Gram matrix (matrix of scalar products):
  \[ G = RR^T = UDD^T U \]
  with \( DD^T \) diagonal.

- Those are the same \( U, W \) and \( D \), hence the link between all the approaches.
Outline

1 Unsupervised Learning?
2 A First Glimpse
   - Clustering
   - Dimensionality Curse
   - Dimension Reduction
   - Generative Modeling
3 Dimension Reduction
   - Simplification
   - Reconstruction Error
   - Relationship Preservation
   - Comparing Methods?
4 Clustering
   - Prototype Approaches
   - Contiguity Approaches
   - Agglomerative Approaches
   - Other Approaches
   - Scalability
5 Generative Modeling
   - (Plain) Parametric Density Estimation
   - Latent Variables
   - Approximate Simulation
   - Diffusion Model
   - Generative Adversarial Network
5 References
Reconstruction Error Approach

Goal

- Construct a map $\Phi$ from the space $\mathcal{X}$ into a space $\mathcal{X}'$ of smaller dimension:
  $\Phi: \mathcal{X} \rightarrow \mathcal{X}'$
  $X \mapsto \Phi(X)$

- Construct $\tilde{\Phi}$ from $\mathcal{X}'$ to $\mathcal{X}$

- Control the error between $X$ and its reconstruction $\tilde{\Phi}(\Phi(X))$

- Canonical example for $X \in \mathbb{R}^d$: find $\Phi$ and $\tilde{\Phi}$ in a parametric family that minimize
  \[
  \frac{1}{n} \sum_{i=1}^{n} \|X_i - \tilde{\Phi}(\Phi(X_i))\|^2
  \]
Principal Component Analysis

- $\mathcal{X} \in \mathbb{R}^d$ and $\mathcal{X}' = \mathbb{R}^{d'}$
- Affine model $\mathbf{X} \sim m + \sum_{l=1}^{d'} X^{(l)} V^{(l)}$ with $(V^{(l)})$ an orthonormal family.
- Equivalent to:
  \[ \Phi(\mathbf{X}) = V^\top (\mathbf{X} - m) \quad \text{and} \quad \tilde{\Phi}(\mathbf{X}') = m + V \mathbf{X}' \]
- Reconstruction error criterion:
  \[ \frac{1}{n} \sum_{i=1}^{n} \| \mathbf{X}_i - (m + V V^\top (\mathbf{X}_i - m)) \|_2^2 \]
- **Explicit solution:** $m$ is the empirical mean and $V$ is any orthonormal basis of the space spanned by the $d'$ first eigenvectors (the one with largest eigenvalues) of the empirical covariance matrix $\frac{1}{n} \sum_{i=1}^{n} (\mathbf{X}_i - m) (\mathbf{X}_i - m)^\top$. 
## Principal Component Analysis

### PCA Algorithm

- Compute the empirical mean $m = \frac{1}{n} \sum_{i=1}^{n} X_i$
- Compute the empirical covariance matrix $\frac{1}{n} \sum_{i=1}^{n} (X_i - m)(X_i - m)^\top$.
- Compute the $d'$ first eigenvectors of this matrix: $V^{(1)}, \ldots, V^{(d')}$
- Set $\Phi(X) = V^\top (X - m)$

- **Complexity:** $O(n(d + d^2) + d'd^2)$
- **Interpretation:**
  - $\Phi(X) = V^\top (X - m)$: coordinates in the restricted space.
  - $V^{(i)}$: influence of each original coordinates in the $i$th new one.

- **Scaling:** This method is not invariant to a scaling of the variables! It is custom to normalize the variables (at least within groups) before applying PCA.
Decathlon

Dimension Reduction
Swiss Roll
Principal Component Analysis

Decathlon

Decathlon Renormalized

Swiss Roll

\[ \triangle \text{Normalized automatisieren} \]
Multiple Factor Analysis

- PCA assumes $\mathcal{X} = \mathbb{R}^d$!
- How to deal with categorical values?
- MFA = PCA with clever coding strategy for categorical values.

Categorical value code for a single variable

- Classical redundant dummy coding:
  \[ X \in \{1, \ldots, V\} \mapsto P(X) = (1_{X=1}, \ldots, 1_{X=V})^T \]

- Compute the mean (i.e. the empirical proportions):
  \[ \bar{P} = \frac{1}{n} \sum_{i=1}^{n} P(X_i) \]

- Renormalize $P(X)$ by $1/\sqrt{(V - 1)\bar{P}}$:
  \[ P(X) = (1_{X=1}, \ldots, 1_{X=V}) \mapsto \left( \frac{1_{X=1}}{\sqrt{(V - 1)\bar{P}_1}}, \ldots, \frac{1_{X=V}}{\sqrt{(V - 1)\bar{P}_V}} \right) = P^r(X) \]

- $\chi^2$ type distance!
Multiple Factor Analysis

- PCA becomes the minimization of
  \[
  \frac{1}{n} \sum_{i=1}^{n} \left\| P^{r}(X_{i}) - (m + VV^{\top}(P^{r}(X_{i}) - m)) \right\|^2
  \]
  \[
  = \frac{1}{n} \sum_{i=1}^{n} \sum_{v=1}^{V} \left| 1_{X_{i}=v} - (m' + \sum_{l=1}^{d'} V^{(l)}^{\top}(P(X_{i}) - m')V^{(l,v)}) \right|^2 \frac{1}{(V - 1)P_{v}}
  \]

- Interpretation:
  - \( m' = P \)
  - \( \Phi(X) = V^{\top}(P^{r}(X) - m) \): coordinates in the restricted space.
  - \( V^{(l)} \) can be interpreted as a probability profile.

- Complexity: \( O(n(V + V^2) + d'V^2) \)

- Link with Correspondence Analysis (CA)
Multiple Factor Analysis

MFA Algorithm

- Redundant dummy coding of each categorical variable.
- Renormalization of each block of dummy variable.
- Classical PCA algorithm on the resulting variables

Interpretation as a reconstruction error with a rescaled $\chi^2$ metric.

Interpretation:
- $\Phi(X) = V^\top (P^r(X) - m)$: coordinates in the restricted space.
- $V^{(i)}$: influence of each modality/variable in the ith new coordinates.

Scaling: This method is not invariant to a scaling of the continuous variables! It is custom to normalize the variables (at least within groups) before applying PCA.
Multiple Factor Analysis

Dimension Reduction
Non Linear PCA

PCA Model

- PCA: Linear model assumption
  \[ X \simeq m + \sum_{l=1}^{d'} X_{\cdot, (l)}^{\prime} V^{(l)} = m + V X' \]
  
  - with
    - \( V^{(l)} \) orthonormal
    - \( X_{\cdot, (l)}^{\prime} \) without constraints.

- Two directions of extension:
  - Other constraints on \( V \) (or the coordinates in the restricted space): ICA, NMF, Dictionary approach
  - PCA on a non-linear image of \( X \): kernel-PCA

- Much more complex algorithm!
## Dimension Reduction

### ICA (Independent Component Analysis)

- Linear model assumption
  \[ X \approx m + \sum_{l=1}^{d'} X'(l) V(l) = m + VX' \]
- with
  - \( V(l) \) without constraints.
  - \( X'(l) \) independent

### NMF (Non Negative Matrix Factorization)

- (Linear) Model assumption
  \[ X \approx \sum_{l=1}^{d'} X'(l) V(l) = VX' \]
- with
  - \( V(l) \) non-negative
  - \( X'(l) \) non-negative.
Non Linear PCA

Dictionary

- (Linear) Model assumption
  \[ X \simeq m + \sum_{l=1}^{d'} X^{',(l)} V^{(l)} = m + V X' \]
  - with
    - \( V^{(l)} \) without constraints
    - \( X' \) sparse (with a lot of 0)

kernel PCA

- Linear model assumption
  \[ \Psi(X - m) \simeq \sum_{l=1}^{d'} X^{',(l)} V^{(l)} = V X' \]
  - with
    - \( V^{(l)} \) orthonormal
    - \( X'_l \) without constraints.
Non Linear PCA

Decathlon

Swiss Roll

ICA

NMF

Kernel PCA
# Auto Encoder

## Dimension Reduction

**Deep Auto Encoder**

- Construct a map $\Phi$ with a **NN** from the space $\mathcal{X}$ into a space $\mathcal{X}'$ of smaller dimension:

  $$\Phi : \mathcal{X} \rightarrow \mathcal{X}'$$

  $$\mathcal{X} \mapsto \Phi(\mathcal{X})$$

- Construct $\tilde{\Phi}$ with a **NN** from $\mathcal{X}'$ to $\mathcal{X}$

- Control the error between $\mathcal{X}$ and its reconstruction $\tilde{\Phi}(\Phi(\mathcal{X}))$:

  $$\frac{1}{n} \sum_{i=1}^{n} \| \mathcal{X}_i - \tilde{\Phi}(\Phi(\mathcal{X}_i)) \|^2$$

- Optimization by gradient descent.

- NN can be replaced by another parametric function...

---

*Image of the page containing the text.*
Deep Auto Encoder

Shallow Auto Encoder (PCA)

Deep Auto Encoder

Dimension Reduction
Outline

1. Unsupervised Learning?
2. A First Glimpse
   - Clustering
   - Dimensionality Curse
   - Dimension Reduction
   - Generative Modeling
3. Dimension Reduction
   - Simplification
   - Reconstruction Error
   - Relationship Preservation
   - Comparing Methods?
4. Clustering
   - Prototype Approaches
   - Contiguity Approaches
   - Agglomerative Approaches
   - Other Approaches
   - Scalability
5. Generative Modeling
   - (Plain) Parametric Density Estimation
   - Latent Variables
   - Approximate Simulation
   - Diffusion Model
   - Generative Adversarial Network
6. References
**Pairwise Relation**

- Different point of view!
- Focus on pairwise relation \( \mathcal{R}(X_i, X_j) \).

---

**Distance Preservation**

- Construct a map \( \Phi \) from the space \( \mathcal{X} \) into a space \( \mathcal{X}' \) of **smaller dimension**:

\[
\Phi: \mathcal{X} \rightarrow \mathcal{X}'
\]

\[
X \mapsto \Phi(X) = X'
\]

- such that

\[
\mathcal{R}(X_i, X_j) \sim \mathcal{R}'(X'_i, X'_j)
\]

- Most classical version (MDS):
  - Scalar product relation: \( \mathcal{R}(X_i, X_j) = (X_i - m)^\top (X_j - m) \)
  - Linear mapping \( X' = \Phi(X) = V^\top (X - m) \).
  - Euclidean scalar product matching:

\[
\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left| (X_i - m)^\top (X_j - m) - \Phi X'_i \Phi^\top X'_j \right|^2
\]

- \( \Phi \) often defined only on \( \mathcal{D} \)…
MultiDimensional Scaling

MDS Heuristic

- Match the scalar products:
  \[ \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left| (X_i - m)^\top (X_j - m) - X_i'\top X_j' \right|^2 \]

- Linear method: \( X' = U^\top (X - m) \) with \( U \) orthonormal

- **Beware:** \( X \) can be unknown, only the scalar products are required!

- Resulting criterion: minimization in \( U^\top (X_i - m) \) of
  \[ \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left| (X_i - m)^\top (X_j - m) - (X_i - m)^\top UU^\top (X_j - m) \right|^2 \]
  without using explicitly \( X \) in the algorithm...

- Explicit solution obtained through the eigendecomposition of the know Gram matrix \( (X_i - m)^\top (X_j - m) \) by keeping only the \( d' \) largest eigenvalues.
In this case, MDS yields the same result as the PCA (but with different inputs, distance between observation vs correlations)!

**Explanation:** Same SVD problem up to a transposition:

- **MDS**
  \[
  \begin{array}{l}
  \overline{X}_{(n)}^\top \overline{X}_{(n)} \sim \overline{X}_{(n)}^\top UU^\top \overline{X}_{(n)} \\
  \end{array}
  \]

- **PCA**
  \[
  \begin{array}{l}
  \overline{X}_{(n)} \overline{X}_{(n)}^\top \sim U^\top \overline{X}_{(n)} \overline{X}_{(n)}^\top U \\
  \end{array}
  \]

- **Complexity:** PCA $O((n + d')d^2)$ vs MDS $O((d + d')n^2)$...
MultiDimensional Scaling

Decathlon

Swiss Roll

PCA

MDS
Generalized MDS

- Preserving the scalar products amounts to preserve the Euclidean distance.
- Easier **generalization** if we work in terms of distance!

**Generalized MDS**

- Generalized MDS:
  - Distance relation: \( \mathcal{R}(X_i, X_j) = d(X_i, X_j) \)
  - Linear mapping \( X' = \Phi(X) = V^\top (X - m) \).
  - Euclidean matching:
    \[
    \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} |d(X_i, X_j) - d'(X'_i, X'_j)|^2
    \]

- Strong connection (but no equivalence) with MDS when \( d(x, y) = ||x - y||^2 \! \).
- **Minimization:** Simple gradient descent can be used (can be stuck in local minima).
- MDS: equivalent to PCA (but more expensive) if \( d(x, y) = \|x - y\|^2 \).
- ISOMAP: use a localized distance instead to limit the influence of very far point.

### ISOMAP

- For each point \( X_i \), define a neighborhood \( \mathcal{N}_i \) (either by a distance or a number of points) and let
  \[
  d_0(X_i, X_j) = \begin{cases} 
  +\infty & \text{if } X_j \notin \mathcal{N}_i \\
  \|X_i - X_j\|^2 & \text{otherwise}
  \end{cases}
  \]
- Compute the shortest path distance for each pair.
- Use the MDS algorithm with this distance.
ISOMAP

Decathlon

Swiss Roll
Random Projection

Random Projection Heuristic

- Draw at random $d'$ unit vector (direction) $U_i$.
- Use $X' = U^\top (X - m)$ with $m = \frac{1}{n} \sum_{i=1}^{n} X_i$

**Property:** If $X$ lives in a space of dimension $d''$, then, as soon as, $d' \sim d'' \log(d'')$, 

$$\|X_i - X_j\|^2 \sim \frac{d}{d'} \|X'_i - X'_j\|^2$$

- Do not really use the data!
Random Projection

Decathlon

Swiss Roll
t-Stochastic Neighbor Embedding

**SNE heuristics**

- From $X_j \in \mathcal{X}$, construct a set of conditional probability:
  \[
  P_{j|i} = \frac{e^{-\|X_i - X_j\|^2/2\sigma_i^2}}{\sum_{k \neq i} e^{-\|X_i - X_k\|^2/2\sigma_i^2}} \quad P_{i|i} = 0
  \]
  \[
  Q_{j|i} = \frac{e^{-\|X'_i - X'_j\|^2/2\sigma_i^2}}{\sum_{k \neq i} e^{-\|X'_i - X'_k\|^2/2\sigma_i^2}} \quad Q_{i|i} = 0
  \]
  is close from $P$.

- **t-SNE:** use a Student-t term $(1 + \|X'_i - X'_j\|^2)^{-1}$ for $X'_i$

- Minimize the Kullback-Leibler divergence $\left(\sum_{i,j} P_{j|i} \log \frac{P_{j|i}}{Q_{j|i}}\right)$ by a simple gradient descent (can be stuck in local minima).

- Parameters $\sigma_i$ such that $H(P_i) = -\sum_{j=1}^n P_{j|i} \log P_{j|i} = \text{cst.}$
t-Stochastic Neighbor Embedding

- Very successful/ powerful technique in practice
- Convergence may be long, unstable, or strongly depending on parameters.
- See this distill post for many impressive examples

Representation depending on t-SNE parameters
t-SNE

Decathlon

Swiss Roll
UMAP

- Topological Data Analysis inspired.

**Uniform Manifold Approximation and Projection**

- Define a notion of asymmetric scaled local proximity between neighbors:
  - Compute the $k$-neighborhood of $X_j$, its diameter $\sigma_i$ and the distance $\rho_i$ between $X_i$ and its nearest neighbor.
  - Define
    \[
    w_i(X_i, X_j) = \begin{cases} 
    e^{-(d(X_i, X_j) - \rho_i)/\sigma_i} & \text{for } X_j \text{ in the } k\text{-neighborhood} \\
    0 & \text{otherwise}
    \end{cases}
    \]
  - Symmetrize into a **fuzzy** nearest neighbor criterion
    \[
    w(X_i, X_j) = w_i(X_i, X_j) + w_j(X_j, X_i) - w_i(X_i, X_j)w_j(X_j, X_i)
    \]
  - Determine the points $X'_i$ in a low dimensional space such that
    \[
    \sum_{i \neq j} w(X_i, X_j) \log \left( \frac{w(X_i, X_j)}{w'(X'_i, X'_j)} \right) + (1 - w(X_i, X_j)) \log \left( \frac{1 - w(X_i, X_j)}{1 - w'(X'_i, X'_j)} \right)
    \]
- Can be performed by local gradient descent.
UMAP

Dimension Reduction

Decathlon

Swiss Roll
Graph based

Graph heuristic

- Construct a graph with weighted edges $w_{i,j}$ measuring the proximity of $X_i$ and $X_j$ ($w_{i,j}$ large if close and 0 if there is no information).
- Find the points $X'_i \in \mathbb{R}^{d'}$ minimizing

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} \|X'_i - X'_j\|^2$$

- Need of a constraint on the size of $X'_i$...
- Explicit solution through linear algebra: $d'$ eigenvectors with smallest eigenvalues of the Laplacian of the graph $D - W$, where $D$ is a diagonal matrix with $D_{i,i} = \sum_j w_{i,j}$.
- Variation on the definition of the Laplacian...
Graph

Dimension Reduction

Decathlon

Swiss Roll
Outline

1 Unsupervised Learning?
2 A First Glimpse
   • Clustering
   • Dimensionality Curse
   • Dimension Reduction
   • Generative Modeling
3 Dimension Reduction
   • Simplification
   • Reconstruction Error
   • Relationship Preservation
   • Comparing Methods?
4 Clustering
   • Prototype Approaches
   • Contiguity Approaches
   • Agglomerative Approaches
   • Other Approaches
   • Scalability
5 Generative Modeling
   • (Plain) Parametric Density Estimation
   • Latent Variables
   • Approximate Simulation
   • Diffusion Model
   • Generative Adversarial Network
6 References
How to Compare Different Dimensionality Reduction Methods?

- **Difficult!** Once again, the metric is very subjective.

**However, a few possible attempts**

- Did we preserve a lot of inertia with only a few directions?
- Do those directions *make sense* from an expert point of view?
- Do the low dimension representation *preserve* some important information?
- Are we better on *subsequent task*?
A Challenging Example: MNIST

**MNIST Dataset**
- Images of 28 × 28 pixels.
- No label used!
- 4 different embeddings.
A Challenging Example: MNIST

MNIST Dataset

- Images of $28 \times 28$ pixels.
- No label used!
- 4 different embeddings.
A Challenging Example: MNIST

**MNIST Dataset**
- Images of $28 \times 28$ pixels.
- No label used!
- 4 different embeddings.
- Quality evaluated by visualizing the true labels *not used to obtain the embeddings.*
- Only a few labels could have been used.
A Simpler Example: A 2D Set

Cluster Dataset

- Set of points in 2D.
- No label used!
- 3 different embeddings.
A Simpler Example: A 2D Set

Cluster Dataset
- Set of points in 2D.
- No label used!
- 3 different embeddings.
A Simpler Example: A 2D Set

Cluster Dataset
- Set of points in 2D.
- No label used!
- 3 different embeddings.
- Quality evaluated by stability...
Outline

1. Unsupervised Learning?
2. A First Glimpse
   - Clustering
   - Dimensionality Curse
   - Dimension Reduction
   - Generative Modeling
3. Dimension Reduction
   - Simplification
   - Reconstruction Error
   - Relationship Preservation
   - Comparing Methods?
4. Clustering
   - Prototype Approaches
   - Contiguity Approaches
   - Agglomerative Approaches
   - Other Approaches
   - Scalability
5. Generative Modeling
   - (Plain) Parametric Density Estimation
   - Latent Variables
   - Approximate Simulation
   - Diffusion Model
   - Generative Adversarial Network
6. References
Clustering

- **Training data**: $D = \{X_1, \ldots, X_n\} \in \mathcal{X}^n$ (i.i.d. $\sim \mathbb{P}$)
- Latent groups?

Clustering

- Construct a map $f$ from $D$ to $\{1, \ldots, K\}$ where $K$ is a number of classes to be fixed:

$$f : X_i \mapsto k_i$$

Motivations

- Interpretation of the groups
- Use of the groups in further processing

- Several strategies possible!
- Can use dimension reduction as a preprocessing.
Outline

1. Unsupervised Learning?
2. A First Glimpse
   - Clustering
   - Dimensionality Curse
   - Dimension Reduction
   - Generative Modeling
3. Dimension Reduction
   - Simplification
   - Reconstruction Error
   - Relationship Preservation
   - Comparing Methods?

4. Clustering
   - Prototype Approaches
     - Contiguity Approaches
     - Agglomerative Approaches
     - Other Approaches
     - Scalability

5. Generative Modeling
   - (Plain) Parametric Density Estimation
   - Latent Variables
   - Approximate Simulation
   - Diffusion Model
   - Generative Adversarial Network

6. References
Partitions Based

Partition Heuristic

- Clustering is defined by a partition in $K$ classes...
- that minimizes a homogeneity criterion.

K-Means

- Cluster $k$ defined by a center $\mu_k$.
- Each sample is associated to the closest center.
- Centers defined as the minimizer of $\sum_{i=1}^{n} \min_k \|X_i - \mu_k\|^2$

Iterative scheme (Loyd):
- Start by a (pseudo) random choice for the centers $\mu_k$
- Assign each samples to its nearby center
- Replace the center of a cluster by the mean of its assigned samples.
- Repeat the last two steps until convergence.
Partition Based Clusterings

K-means, step 0 - 4

Petal Width vs. Petal Length
Partition based

- Other schemes:
  - McQueen: modify the mean each time a sample is assigned to a new cluster.
  - Hartigan: modify the mean by removing the considered sample, assign it to the nearby center and recompute the new mean after assignment.

A good initialization is crucial!

- Initialize by samples.
- k-Mean++: try to take them as separated as possible.
- No guarantee to converge to a global optimum: repeat and keep the best result!

- Complexity: $O(n \times K \times T)$ where $T$ is the number of steps in the algorithm.
Partition based

- **k-Medoid**: use a sample as a center
  - PAM: for a given cluster, use the sample that minimizes the intra distance (sum of the squared distance to the other points)
  - Approximate medoid: for a given cluster, assign the point that is the closest to the mean.

**Complexity**

- **PAM**: $O(n^2 \times T)$ in the worst case!
- **Approximate medoid**: $O(n \times K \times T)$ where $T$ is the number of steps in the algorithm.

**Remark**: Any distance can be used... but the complexity of computing the centers can be very different.
K-Means

Clustering

$k = 4$

$k = 10$

$k = 10$
Clustering

Model Based

Model Heuristic

- Use a generative model of the data:

\[ P(X) = \sum_{k=1}^{K} \pi_k P_{\theta_k}(X|k) \]

where \( \pi_k \) are proportions and \( P_{\theta_k}(X|k) \) are parametric probability models.

- Estimate those parameters (often by a ML principle).

- Assign each observation to the class maximizing the a posteriori probability (obtained by Bayes formula)

\[
\frac{\hat{\pi}_k \hat{P}_{\theta_k}(X|k)}{\sum_{k'=1}^{K} \hat{\pi}_{k'} \hat{P}_{\theta_{k'}}(X|k')} \]

- Link with Generative model in supervised classification!
A two class example

- A mixture $\pi_1 f_1(X) + \pi_2 f_2(X)$
- and the posterior probability $\pi_i f_i(X) / (\pi_1 f_1(X) + \pi_2 f_2(X))$
- Natural class assignment!
Sub-population estimation

- A mixture \( \pi_1 f_1(X) + \pi_2 f_2(X) \)
- Two populations with a parametric distribution \( f_i \).
- Most classical choice: Gaussian distribution

Gaussian Setting

- \( X_1, \ldots, X_n \) independent
- \( X_i \sim N(\mu_1, \sigma_1^2) \) with probability \( \pi_1 \) or \( X_i \sim N(\mu_2, \sigma_2^2) \) with probability \( \pi_2 \)
- We don’t know the parameters \( \mu_i, \sigma_i, \pi_i \).
- We don’t know from which distribution each \( X_i \) has been drawn.
Clustering

Model Based

Maximum Likelihood

- Density:
  \( \pi_1 \Phi(X, \mu_1, \sigma_1^2) + \pi_2 \Phi(X, \mu_2, \sigma_2^2) \)

- log-likelihood:
  \[ L(\theta) = \sum_{i=1}^{n} \log \left( \pi_1 \Phi(X_i, \mu_1, \sigma_1^2) + \pi_2 \Phi(X_i, \mu_2, \sigma_2^2) \right) \]

- No straightforward way to optimize the parameters!

What if algorithm

- Assume we know from which distribution each sample has been sampled: \( Z_i = 1 \) if from \( f_1 \) and \( Z_i = 0 \) otherwise.

- log-likelihood:
  \[ \sum_{i=1}^{n} Z_i \log \Phi(X_i, \mu_1, \sigma_1^2) + (1 - Z_i) \log \Phi(X_i, \mu_2, \sigma_2^2) \]

- Easy optimization... but the \( Z_i \) are unknown!
**Model Based**

### What if algorithm

- Assume we know from which distribution each sample has been sampled: $Z_i = 1$ if from $f_1$ and $Z_i = 0$ otherwise.

- Log-likelihood:
  \[
  \sum_{i=1}^{n} Z_i \log \Phi(X_i, \mu_1, \sigma_1^2) + (1 - Z_i) \log \Phi(X_i, \mu_2, \sigma_2^2)
  \]

- Easy optimization... but the $Z_i$ are unknown!

### Bootstrapping Idea

- Replace $Z_i$ by its expectation given the current estimate.

- \( \mathbb{E}[Z_i] = P(Z_i = 1|\theta) \) (A posteriori probability)

- and iterate...

- Can be proved to be a good idea!
EM Algorithm

- (Random) initialization: $\mu_i^0, \sigma_i^0, \pi_i^0$.
- Repeat:
  - Expectation (Current a posteriori probability):
    \[
    E_t[Z_i] = P(Z_i = 1|\theta^t) = \frac{\pi_1^t \Phi(X_i, \mu_1^t, (\sigma_1^t)^2)}{\pi_1^t \Phi(X_i, \mu_1^t, (\sigma_1^t)^2) + \pi_2^t \Phi(X_i, \mu_2^t, (\sigma_2^t)^2)}
    \]
  - Maximization of
    \[
    \sum_{i=1}^{n} E_t[Z_i] \log \Phi(X_i, \mu_1, \sigma_1^2) + E_t[1 - Z_i] \log \Phi(X_i, \mu_2, \sigma_2^2)
    \]
    to obtain $\mu_i^{t+1}, \sigma_i^{t+1}, \pi_i^{t+1}$. 
Model Based

- Large choice of parametric models.

**Gaussian Mixture Model**

- Use

\[
\mathbb{P}_{\theta_k}(\vec{X}|k) \sim N(\mu_k, \Sigma_k)
\]

with \( N(\mu, \Sigma) \) the Gaussian law of mean \( \mu \) and covariance matrix \( \Sigma \).

- Efficient optimization algorithm available (EM)
- Often some constraint on the covariance matrices: identical, with a similar structure.
- Strong connection with \( K \)-means when the covariance matrices are assumed to be the same multiple of the identity.
Probabilistic latent semantic analysis (PLSA)

- Documents described by their word counts $w$
- Model:

$$P(w) = \sum_{k=1}^{K} \pi_k P_{\theta_k}(w|k)$$

with $k$ the (hidden) topic, $\pi_k$ a topic probability and $P_{\theta_k}(w|k)$ a multinomial law for a given topic.

- Clustering according to

$$P(k|w) = \frac{\hat{\pi}_k P_{\hat{\theta}_k}(w|k)}{\sum_{k'} \hat{\pi}_{k'} P_{\hat{\theta}_{k'}}(w|k')}$$

- Same idea than GMM!
- Bayesian variant called LDA.
Model Based

Parametric Density Estimation Principle

- Assign a probability of membership.
- Lots of theoretical studies...
- Model selection principle can be used to select $K$ the number of classes (or rather to avoid using a nonsensical $K$...):
  - AIC / BIC / MDL penalization
  - Cross Validation is also possible!

- Complexity: $O(n \times K \times T)$
Gaussian Mixture Models

$k = 4$

$k = 10$

$k = 10$
Outline

1. Unsupervised Learning?
2. A First Glimpse
   - Clustering
   - Dimensionality Curse
   - Dimension Reduction
   - Generative Modeling
3. Dimension Reduction
   - Simplification
   - Reconstruction Error
   - Relationship Preservation
   - Comparing Methods?
4. Clustering
   - Prototype Approaches
   - Contiguity Approaches
   - Agglomerative Approaches
   - Other Approaches
   - Scalability
5. Generative Modeling
   - (Plain) Parametric Density Estimation
   - Latent Variables
   - Approximate Simulation
   - Diffusion Model
   - Generative Adversarial Network
6. References
Density Heuristic

- Cluster are connected dense zone separated by low density zone.
- Not all points belong to a cluster.

- Basic bricks:
  - Estimate the density.
  - Find points with high densities.
  - Gather those points according to the density.

- Density estimation:
  - Classical kernel density estimators...

- Gathering:
  - Link points of high density and use the resulted component.
  - Move them toward top of density hill by following the gradient and gather all the points arriving at the same summit.
Clustering (Non Parametric) Density Based

Concepts

2 paramètres:

- **Eps**: rayon maximum de voisinage
- **MinPts**: nb minimum de pts pour que \( V(p) \) soit un voisinage de taille Eps du pts \( p \)

Exemple avec MinPts = 4 et Eps = 1cm

\( x \) est densité-atteignable depuis \( y \) s’il existe une chaîne de points (de longueur quelconque) partant de \( y \) et allant jusqu’à \( x \) et telle que le point \( p_{i+1} \) est densité atteignable depuis \( p_i \).

\( x \) appartient à \( V(y) \), \( p_2 \) appartient à \( V(p_1) \), \( x \) appartient à \( V(p_2) \).

\( x \) et \( y \) sont densité-connectés s’il existe un point \( z \) tel que \( x \) soit atteignable depuis \( z \) et \( y \) soit atteignable depuis \( z \).

Une classe \( C \) doit vérifier les 2 conditions suivantes :

1) Si un point \( x \) appartient à \( C \) alors tout point atteignable depuis \( x \) appartient à \( C \).
2) Tous les points d’une classe sont densité-connectés.

Les points pleins appartiennent à une même classe, les points vides à une autre.

Le point \( z \) appartient à deux classes. Par convention, on l’affecte à la première classe à laquelle il est affecté.

Examples

- **DBSCAN**: link point of high densities using a very simple kernel.
- **PdfCLuster**: find connected zone of high density.
- **Mean-shift**: move points toward top of density hill following an evolving kernel density estimate.

Complexity: \( O(n^2 \times T) \) in the worst case.

Can be reduced to \( O(n \log(n) T) \) if samples can be encoded in a tree structure (n-body problem type approximation).
DBSCAN

Clustering

$\epsilon = .45$

$\epsilon = .2$

$\epsilon = .1$
Outline

1. Unsupervised Learning?
2. A First Glimpse
   - Clustering
   - Dimensionality Curse
   - Dimension Reduction
   - Generative Modeling
3. Dimension Reduction
   - Simplification
   - Reconstruction Error
   - Relationship Preservation
   - Comparing Methods?

4. Clustering
   - Prototype Approaches
   - Contiguity Approaches
   - Agglomerative Approaches
   - Other Approaches
   - Scalability

5. Generative Modeling
   - (Plain) Parametric Density Estimation
   - Latent Variables
   - Approximate Simulation
   - Diffusion Model
   - Generative Adversarial Network

5. References
Agglomerative Clustering

Agglomerative Clustering Heuristic

- Start with very small clusters (a sample by cluster?)
- Sequential merging of the most similar clusters...
- according to some *greedy* criterion $\Delta$.

- Generates a hierarchy of clustering instead of a single one.
- Need to select the number of cluster afterwards.
- Several choices for the merging criterion...
- Examples:
  - Minimum Linkage: merge the closest cluster in term of the usual distance
  - Ward’s criterion: merge the two clusters yielding the less inner inertia loss (k-means criterion)
### Agglomerative Clustering

#### Algorithm

- Start with \((C_i^{(0)}) = \{X_i\}\) the collection of all singletons.
- At step \(s\), we have \(n - s\) clusters \((C_i^{(s)})\):
  - Find the two most similar clusters according to a criterion \(\Delta\):
    \[
    (i, i') = \arg\min_{(j, j')}(\Delta(C_j^{(s)}, C_{j'}^{(s)}))
    \]
  - Merge \(C_i^{(s)}\) and \(C_{i'}^{(s)}\) into \(C_i^{(s+1)}\)
  - Keep the \(n - s - 2\) other clusters \(C_{i''}^{(s+1)} = C_{i''}^{(s)}\)
- Repeat until there is only one cluster.

- Complexity: \(O(n^3)\) in general.
- Can be reduced to \(O(n^2)\)
  - if only a bounded number of merging is possible for a given cluster,
  - for the most classical distances by maintaining a nearest neighbors list.
Merging criterion based on the distance between points

- **Minimum linkage:**
  \[ \Delta(C_i, C_j) = \min_{X_i \in C_i} \min_{X_j \in C_j} d(X_i, X_j) \]

- **Maximum linkage:**
  \[ \Delta(C_i, C_j) = \max_{X_i \in C_i} \max_{X_j \in C_j} d(X_i, X_j) \]

- **Average linkage:**
  \[ \Delta(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{X_i \in C_i} \sum_{X_j \in C_j} d(X_i, X_j) \]

- Clustering based on the proximity...
Agglomerative Clustering

Merging criterion based on the inertia (distance to the mean)

- Ward’s criterion:
  \[
  \Delta(C_i, C_j) = \sum_{x_i \in C_i} \left( d^2(x_i, \mu_{C_i \cup C_j}) - d^2(x_i, \mu_{C_i}) \right) 
  + \sum_{x_j \in C_j} \left( d^2(x_j, \mu_{C_i \cup C_j}) - d^2(x_j, \mu_{C_j}) \right)
  \]

- If \( d \) is the Euclidean distance:
  \[
  \Delta(C_i, C_j) = \frac{2|C_i||C_j|}{|C_i| + |C_j|} d^2(\mu_{C_i}, \mu_{C_j})
  \]

- Same criterion than in the \( k \)-means algorithm but greedy optimization.
Outline

1 Unsupervised Learning?
2 A First Glimpse
   ● Clustering
   ● Dimensionality Curse
   ● Dimension Reduction
   ● Generative Modeling
3 Dimension Reduction
   ● Simplification
   ● Reconstruction Error
   ● Relationship Preservation
   ● Comparing Methods?
4 Clustering
   ● Prototype Approaches
   ● Contiguity Approaches
   ● Agglomerative Approaches
   ● Other Approaches
   ● Scalability
5 Generative Modeling
   ● (Plain) Parametric Density Estimation
   ● Latent Variables
   ● Approximate Simulation
   ● Diffusion Model
   ● Generative Adversarial Network
6 References
Clustering

**Grid based**

**Grid heuristic**

- Split the space in pieces
- Group those of high density according to their proximity

- Similar to density based estimate (with partition based initial clustering)
- Space splitting can be fixed or adaptive to the data.
- Examples:
  - STING (Statistical Information Grid): Hierarchical tree construction plus DBSCAN type algorithm
  - AMR (Adaptive Mesh Refinement): Adaptive tree refinement plus $k$-means type assignment from high density leaves.
  - CLIQUE: Tensorial grid and 1D detection.

- Linked to Divisive clustering (DIANA)
Graph based

- Spectral clustering: dimension reduction + k-means.
- Message passing: iterative local algorithm.
- Graph cut: min/max flow.

- Kohonen Map,
- ...
Outline

1. Unsupervised Learning?

2. A First Glimpse
   - Clustering
   - Dimensionality Curse
   - Dimension Reduction
   - Generative Modeling

3. Dimension Reduction
   - Simplification
   - Reconstruction Error
   - Relationship Preservation
   - Comparing Methods?

4. Clustering
   - Prototype Approaches
   - Contiguity Approaches
   - Agglomerative Approaches
   - Other Approaches
   - Scalability

5. Generative Modeling
   - (Plain) Parametric Density Estimation
   - Latent Variables
   - Approximate Simulation
   - Diffusion Model
   - Generative Adversarial Network

6. References
Scalability

**Large dataset issue**
- When \( n \) is large, a \( O(n^\alpha \log n) \) with \( \alpha > 1 \) is not acceptable!
- How to deal with such a situation?

**Beware:** Computing all the pairwise distance requires \( O(n^2) \) operations!

**Ideas**
- Sampling
- Online processing
- Simplification
- Parallelization
Sampling

**Sampling heuristic**

- Use only a subsample to construct the clustering.
- Assign the other points to the constructed clusters afterwards.

- Requires a clustering method that can assign new points (partition, model...)
- Often repetition and choice of the best clustering

**Example:**
- CLARA: K-medoid with sampling and repetition

**Two-steps algorithm:**
- Generate a large number $n'$ of clusters using a fast algorithm (with $n' \ll n$)
- Cluster the clusters with a more accurate algorithm.
Online heuristic

- Modify the current clusters according to the value of a single observation.

- Requires compactly described clusters.
- Examples:
  - Add to an existing cluster (and modify it) if it is close enough and create a new cluster otherwise (*k*-means without reassignment)
  - Stochastic descent gradient (GMM)
- May lead to far from optimal clustering.
Simplification

Simplification heuristic

- Simplify the algorithm to be more efficient at the cost of some precision.
- Algorithm dependent!
- Examples:
  - Replace groups of observation (preliminary cluster) by the (approximate) statistics.
  - Approximate the distances by cheaper ones.
  - Use n-body type techniques.
Parallelization

Parallelization heuristic

- Split the computation on several computers.

- Algorithm dependent!
- Examples:
  - Distance computation in $k$-means, parameter gradient in model based clustering
  - Grid density estimation, Space splitting strategies
- Classical batch sampling not easy to perform as partitions are not easily merged...
Outline

1. Unsupervised Learning?
2. A First Glimpse
   - Clustering
   - Dimensionality Curse
   - Dimension Reduction
   - Generative Modeling
3. Dimension Reduction
   - Simplification
   - Reconstruction Error
   - Relationship Preservation
   - Comparing Methods?
4. Clustering
   - Prototype Approaches
   - Contiguity Approaches
   - Agglomerative Approaches
   - Other Approaches
   - Scalability
5. Generative Modeling
   - (Plain) Parametric Density Estimation
   - Latent Variables
   - Approximate Simulation
   - Diffusion Model
   - Generative Adversarial Network
6. References
Generative Modeling

- **Training data**: \( \mathcal{D} = \{ (X_1, Y_1), \ldots, (X_n, Y_n) \} \in (\mathcal{X} \times \mathcal{Y})^n \) (i.i.d. \( \sim \mathbb{P} \))
- Same kind of data than for supervised learning if \( \mathcal{Y} \neq \emptyset \).

**Generative Modeling**

- Construct a map \( G \) from the product of \( \mathcal{Y} \) and a randomness source \( \Omega \) to \( \mathcal{X} \)
  
  \[ G : \mathcal{Y} \times \Omega \rightarrow \mathcal{X} \quad (Y, \omega) \mapsto X \]

- Unconditional model if \( \mathcal{Y} = \emptyset \)...

**Motivation**

- Generate plausible novel conditional samples based on a given dataset.

**Sample Quality**

- Related to the proximity between the law of \( G(Y, \omega) \) and the law of \( X|Y \).
- Most classical choice is the Kullback-Leibler divergence.
Generative Modeling

**Ingredients**
- Generator $G_\theta(Y, \omega)$ and cond. density prob. $P_\theta(X|Y)$ (Explicit vs implicit link)
- Simple / Complex / Approximate estimation...

**Some Possible Choices**

<table>
<thead>
<tr>
<th>Probabilistic model</th>
<th>Generator</th>
<th>Estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>Simple (parametric)</td>
<td>Explicit</td>
</tr>
<tr>
<td>Flow</td>
<td>Image of simple model</td>
<td>Explicit</td>
</tr>
<tr>
<td>Factorization</td>
<td>Factorization of simple model</td>
<td>Explicit</td>
</tr>
<tr>
<td>VAE</td>
<td>Simple model with latent var.</td>
<td>Explicit</td>
</tr>
<tr>
<td>EBM</td>
<td>Arbitrary</td>
<td>Implicit (MCMC)</td>
</tr>
<tr>
<td>Diffusion</td>
<td>Continuous noise</td>
<td>Implicit (MCMC)</td>
</tr>
<tr>
<td>GAN</td>
<td>Discrete Noise with latent var.</td>
<td>Explicit</td>
</tr>
<tr>
<td></td>
<td>Implicit</td>
<td>Explicit</td>
</tr>
</tbody>
</table>

- **SOTA:** Diffusion based approach!

Generators

\[ \tilde{X} = G(Y, \omega) \]

- Small abuse of notations...
- More an algorithm than a map!

Generators

- One step: \( \omega \sim \tilde{Q}(\cdot|Y) \) and \( \tilde{X} = G(Y, \omega) \).
- Several steps:
  - \( \omega_0 \sim \tilde{Q}_0(\cdot|Y) \) and \( \tilde{X}_0 = G_0(\omega_0) \)
  - \( \omega_{t+1} \sim \tilde{Q}_{t+1}(\cdot|Y, \tilde{X}_t) \) and \( \tilde{X}_{t+1} = G_{t+1}(Y, \tilde{X}_k, \omega_{t+1}) \)
- Fixed or variable number of steps.
- Fixed or variable dimension for \( \tilde{X}_t \) and \( \omega_t \)...

- \( \tilde{Q} \) (or \( \tilde{Q}_t \)) should be easy to sample.
- Most of the time, parametric representations for \( \tilde{Q} \) (or \( \tilde{Q}_t \)) and \( G \) (or \( G_t \)).
Outline

1. Unsupervised Learning?
2. A First Glimpse
   - Clustering
   - Dimensionality Curse
   - Dimension Reduction
   - Generative Modeling
3. Dimension Reduction
   - Simplification
   - Reconstruction Error
   - Relationship Preservation
   - Comparing Methods?
4. Clustering
   - Prototype Approaches
   - Contiguity Approaches
   - Agglomerative Approaches
   - Other Approaches
   - Scalability
5. Generative Modeling
   - (Plain) Parametric Density Estimation
   - Latent Variables
   - Approximate Simulation
   - Diffusion Model
   - Generative Adversarial Network
6. References
Warmup: Density Estimation and Generative Modeling

\[ X \sim P \text{ with } dP(x) = p(x)d\lambda \quad \rightarrow \quad \tilde{X} \sim \tilde{P} \text{ with } d\tilde{P}(x) = \tilde{p}(x)d\lambda \]

Heuristic

- Estimate \( \bar{p} \) by \( \tilde{p} \) from an i.i.d. sample \( X_1, \ldots, X_n \).
- Simulate \( \tilde{X} \) having a law \( \tilde{P} \).

By construction, if \( \tilde{p} \) is close from \( p \), the law of \( \tilde{X} \) will be close from the law of \( X \).

Issue: How to do it?

- How to estimate \( \tilde{p} \)? Parametric, non-parametric? Maximum likelihood? Other criteria?
- How to simulate \( \tilde{P} \)? Parametric? One-step? Multi-step? Iterative?
Warmup: Parametric Density Estimation

\[ X \sim P(\cdot) \text{ with } dP(x) = p(x) d\lambda \quad \longrightarrow \quad \tilde{X} \sim \tilde{P}_\theta \text{ with } d\tilde{P}_\theta(x) = \tilde{p}_\theta(x) d\lambda \]

### Maximum Likelihood Approach

- Select a family \( \tilde{P} \) and estimate \( p \) by \( \tilde{p}_\theta \) from an i.i.d. sample \( X_1, \ldots, X_n \).
- Simulate \( \tilde{X} \) having a law \( \tilde{P}_\theta \).
- By construction, if \( \tilde{p}_\theta \) is close from \( p \), the law of \( \tilde{X} \) will be close from the law of \( X \).

### Issue: How to do it?

- Which family \( \tilde{P} \)?
- How to simulate \( \tilde{P}_\theta \)? Parametric? Iterative?

- Corresponds to \( \omega \sim \tilde{P}_\theta \) and \( \tilde{X} = G(\omega) = \omega \)
Conditional Density Est. and Generative Modeling

\[ X \mid Y \sim P(Y \cdot \mid Y) \text{ with } dP(x \mid Y) = p(x \mid Y) d\lambda \]

\[ \rightarrow \tilde{X} \mid Y \sim \tilde{P}(\cdot \mid Y) \text{ with } d\tilde{P}(x \mid Y) = \tilde{p}(x \mid Y) d\lambda \]

**Heuristic**

- Estimate \( p \) by \( \tilde{p} \) from an i.i.d. sample \((X_1, Y_1), \ldots, (X_n, Y_n)\).
- Simulate \( \tilde{X} \mid Y \) having a law \( \tilde{P}(\cdot \mid Y) \).

By construction, if \( \tilde{p} \) is close from \( p \), the law of \( \tilde{X} \mid Y \) will be close from the law of \( X \mid Y \).

**Issue: How to do it?**

- How to estimate \( \tilde{p} \)? Parametric, non-parametric? Maximum likelihood? Other criteria?
- How to simulate \( \tilde{P} \)? Parametric? One-step? Multi-step? Iterative?
$X|Y \sim P(\cdot|Y)$ with $dP(x|Y) = p(x|Y)d\lambda$

$\rightarrow \tilde{X}|Y \sim \tilde{P}_{\tilde{\theta}(Y)}$ with $d\tilde{P}_{\tilde{\theta}(Y)}(x) = \tilde{p}_{\tilde{\theta}(Y)}(x)d\lambda$

**Maximum Likelihood Approach**

- Select a family $\tilde{P}$ and estimate $p$ by $\tilde{p}_{\tilde{\theta}}$ from an i.i.d. sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ where $\tilde{\theta}$ is now a function of $Y$.
- Simulate $\tilde{X}|Y$ having a law $\tilde{P}_{\tilde{\theta}(Y)}$

- If $\tilde{p}_{\tilde{\theta}}$ is close from $p$, the law of $\tilde{X}|Y$ will be close from the law of $X|Y$.

**Issue: How to do it?**

- Which family $\tilde{P}$? Which function family for $\tilde{\theta}$?
- How to simulate $\tilde{P}_{\tilde{\theta}(Y)}$? Parametric? Iterative?

- Corresponds to $\omega \sim \tilde{Q}(\cdot|Y) = \tilde{P}_{\tilde{\theta}(Y)}$ and $\tilde{X} = G(Y, \omega) = \omega$
Direct Parametric Conditional Density Estimation

\[ \omega \sim \tilde{Q}_{\tilde{\theta}(Y)} \sim \tilde{q}_{\tilde{\theta}(Y)}(x) d\lambda \quad \text{and} \quad \tilde{X}|Y = G(Y, \omega) = \omega \]

**Estimation**

- By construction,
  \[ dP(\tilde{X}|Y) = \tilde{q}_{\tilde{\theta}(Y)}(x) d\lambda \]
- Maximum Likelihood approach:
  \[ \tilde{\theta} = \arg\max_{\theta} \sum_{i=1}^{n} \log \tilde{q}_{\tilde{\theta}(Y_i)}(X_i) \]

**Simulation**

- \( \tilde{P} \) has been chosen so that this distribution is easy to sample...
- Possible families: Gaussian, Multinomial, Exponential model...
- Possible parametrizations for \( \tilde{\theta} \): linear, neural network...
- Limited expressivity!
Invertible Transform

$$\omega \sim \tilde{Q}_{\tilde{\theta}(Y)} \sim \tilde{q}_{\tilde{\theta}(Y)}(x)d\lambda \quad \text{and} \quad \tilde{X}|Y = G(\omega) \text{ with a given } G \text{ invertible.}$$

**Estimation**

- By construction,
  $$d\tilde{P}(G^{-1}(\tilde{X})|Y) = \tilde{q}_{\tilde{\theta}(y)}(G^{-1}(x))d\lambda$$
- Maximum Likelihood approach:
  $$\tilde{\theta} = \text{argmax}_{\theta} \sum_{i=1}^{n} \log \tilde{q}_{\tilde{\theta}(Y_i)}(G^{-1}(X_i))$$

**Simulation**

- $\tilde{Q}$ has been chosen so that this distribution is easy to sample...
- Possible transform $G$: Change of basis, known transform...
Flow

\[ \omega \sim \tilde{Q}_{\theta(Y)} = \tilde{q}_{\theta(Y)}(x) \, d\lambda \quad \text{and} \quad \tilde{X} \mid Y = G_{\theta_G(Y)}(\omega) \text{ with } G_{\theta} \text{ invertible.} \]

Estimation

- By construction,
  \[ d\tilde{P}(\tilde{X} \mid Y) = |\text{Jac}_{G^{-1}}(x)| \tilde{q}_{\theta(Y)}(G_{\theta_G(Y)}^{-1}(x)) \, d\lambda \]
  where \( \text{Jac}_{G_{\theta_G(Y)}^{-1}}(x) \) is the Jacobian of \( G_{\theta_G(Y)}^{-1} \) at \( x \)
- Maximum Likelihood approach:
  \[ \tilde{\theta}, \tilde{\theta}_G = \arg\max_{\theta, \theta_G} \sum_{i=1}^{n} \left( \log |\text{Jac}_{G_{\theta_G(Y)}}^{-1}(x)| + \log \tilde{q}_{\theta(Y_i)}(G_{\theta_G(Y)}^{-1}(X_i)) \right) \]

Simulation

- \( \tilde{Q} \) has been chosen so that this distribution is easy to sample...
- Often, in practice, \( \tilde{\theta}(Y) \) is independent of \( Y \)...
- Main issue: \( G_{\theta} \), its inverse and its Jacobian should be easy to compute.
Possible Flows

$G_\theta$?

- Main issue: $G_\theta$, its inverse and its Jacobian should be easy to compute.

Flow Models

- Composition

$$G_\theta = G_{\theta_T} \circ G_{\theta_{T-1}} \circ G_{\theta_1} \circ G_{\theta_0}$$

$$|\text{Jac}G_\theta^{-1}| = \prod |\text{Jac}G_{\theta_i}^{-1}|$$

- Real NVP

$$G_\theta(x) = \begin{pmatrix} x_1 \\ \vdots \\ x_{d'} \\ x_{d'} e^{s_{d''}}(x_1,...,d') + t_d(x_1,...,d') \\ \vdots \\ x_d e^{s_d}(x_1,...,d') + t_d(x_1,...,d') \end{pmatrix} \rightarrow G^{-1}_\theta(x) = \begin{pmatrix} x_1 \\ \vdots \\ x_{d'} \\ (x_{d'} + t_d(x_1,...,d')) e^{-s_{d''}}(x_1,...,d') + \\ \vdots \\ (x_d + t_d(x_1,...,d')) e^{-s_d}(x_1,...,d') \end{pmatrix}$$

$$|\text{Jac}G(x)^{-1}| = \prod_{d''=d'+1}^d e^{-s_{d''}(x_1,...,d')}$$

- Combined with permutation along dimension or invertible transform across dimension.

- Not that much flexibility...
Factorization

\[ \omega_0 \sim \tilde{Q}_0(\cdot|Y) \text{ and } \tilde{X}_0 = G_0(\omega_0) \]
\[ \omega_{t+1} \sim \tilde{Q}_{t+1}(\cdot|Y, (\tilde{X}_l)_{l \leq t}) \text{ and } \tilde{X}_{t+1} = G_{t+1}(Y, (\tilde{X}_l)_{l \leq t}, \omega_{t+1}) \]
\[ \tilde{X} = (\tilde{X}_0, \ldots, \tilde{X}_{d-1}) \]

Factorization

- Amounts to use a factorized representation
  \[ \tilde{P}(\tilde{X}|Y) = \prod_{0 \leq k < d} \tilde{P}(\tilde{X}_k|Y, (\tilde{X}_l)_{l \leq t}) \]

- \( \tilde{Q}_t \) and \( G_k \) can be chosen as in the plain density estimation case as the \( X_i(t) \) are observed.

Estimation

- \( d \) generative models to estimate instead of one.

- Simple generator by construction.
Sequence and Markov Model

\[ \omega_{t+1} \sim \tilde{Q}(\cdot | Y, (\tilde{X}_l)_{t \geq l \geq t-o}) \quad \text{and} \quad \tilde{X}_{t+1} = G(Y, (\tilde{X}_l)_{t \geq l \geq t-o}, \omega_{t+1}) \]

\[ \tilde{X} = (\tilde{X}_0, \ldots, \tilde{X}_{d-1}) \]

Sequence and Markov Models

- **Sequence**: sequence of *similar* objects with a translation invariant structure.
- Translation invariant probability model of finite order (memory) \( o \).
- Requires an initial padding of the sequence.
- Faster training as the parameters are shared for all \( t \).
- Model used in Text Generation!
Outline

1. Unsupervised Learning?
2. A First Glimpse
   - Clustering
   - Dimensionality Curse
   - Dimension Reduction
   - Generative Modeling
3. Dimension Reduction
   - Simplification
   - Reconstruction Error
   - Relationship Preservation
   - Comparing Methods?
4. Clustering
   - Prototype Approaches
   - Contiguity Approaches
   - Agglomerative Approaches
   - Other Approaches
   - Scalability
5. Generative Modeling
   - (Plain) Parametric Density Estimation
   - Latent Variables
   - Approximate Simulation
   - Diffusion Model
   - Generative Adversarial Network
6. References
Generative Modeling

\[ \omega_0 \sim \tilde{Q}_0(\cdot | Y) \text{ and } \tilde{X}_0 = G_0(\omega_0) \]
\[ \omega_1 \sim \tilde{Q}_1(\cdot | Y, X_0) \text{ and } \tilde{X}_1 = G_1(Y, \omega_0) \]
\[ \tilde{X} = \tilde{X}_1 \]

- Most classical example:
  - Gaussian Mixture Model with \( \tilde{X}_0 = \omega_0 \sim \mathcal{M}(\pi) \) and \( \tilde{X} = \omega_1 \sim \mathcal{N}(\mu_{\tilde{X}_0}, \Sigma_{\tilde{X}_0}) \).

**Estimation**

- Still a factorized representation
  \[ \tilde{P}(\tilde{X}_1, \tilde{X}_0 | Y) = \tilde{P}_0(\tilde{X}_0 | Y) \tilde{P}_1(\tilde{X}_1 | Y, \tilde{X}_0) \]

  but only \( \tilde{X}_1 \) is observed.

- **Much more complex estimation!**

- Simple generator by construction provided that the \( \tilde{Q}_t \) are easy to simulate.
Generative Modeling

Log Likelihood and ELBO

\[
\log \tilde{p}(\tilde{X} \mid Y) = \log \mathbb{E}_{\tilde{P}(\tilde{X}_0 \mid Y, \tilde{X})} \left[ \tilde{p}(\tilde{X}, \tilde{X}_0 \mid Y) \right]
\]

\[
= \sup_{R(\cdot \mid Y, \tilde{X})} \mathbb{E}_{R(\cdot \mid Y, \tilde{X})} \left[ \log \tilde{p}(\tilde{X}, \tilde{X}_0 \mid Y) - \log r(\tilde{X}_0 \mid Y, \tilde{X}) \right]
\]

- Need to integrate over \( \tilde{X}_0 \) using the conditional law \( \tilde{P}(\tilde{X}_0 \mid Y, \tilde{X}) \), which may be hard to compute.

Evidence Lower BOund

- Using \( \log \tilde{p}(\tilde{X} \mid Y) = \mathbb{E}_{R(\cdot \mid Y, \tilde{X})} \left[ \log \left( \tilde{p}(\tilde{X}, \tilde{X}_0 \mid Y)/\tilde{p}(\tilde{X}_0 \mid Y, \tilde{X}) \right) \right] \),

\[
\log \tilde{p}(\tilde{X} \mid Y) = \mathbb{E}_{R(\cdot \mid Y, \tilde{X})} \left[ \log \tilde{p}(\tilde{X}, \tilde{X}_0 \mid Y) - \log r(\tilde{X}_0 \mid Y, \tilde{X}) \right]
\]

\[
- \text{KL}_{\tilde{X}_0} \left( R(\tilde{X}_0 \mid Y, \tilde{X}), \tilde{P}(\tilde{X}_0 \mid Y, \tilde{X}) \right)
\]

- Equality is obtained for \( R(\cdot \mid Y, \tilde{X}) = \tilde{P}(\tilde{X}_0 \mid Y, \tilde{X}) \).

- Maximization over \( \tilde{P} \) and \( R \) instead of only over \( \tilde{P} \)…
ELBO and Stochastic Gradient Descent

$$\sup_{\tilde{P}} E_{\tilde{X}, Y} \left[ \log \tilde{p}(\tilde{X}|Y) \right] = \sup_{\tilde{P}, R} E_{\tilde{X}, Y, \tilde{X}_0 \sim R(\cdot|Y, \tilde{X})} \left[ \log \tilde{p}(\tilde{X}, \tilde{X}_0|Y) - \log r(\tilde{X}_0|Y, \tilde{X}) \right]$$

$$= \sup_{\tilde{P}, R} E_{\tilde{X}, Y, \tilde{X}_0 \sim R(\cdot|Y, \tilde{X})} \left[ \log \tilde{p}(\tilde{X}|Y, \tilde{X}_0) \right]$$

$$+ \underbrace{E_{\tilde{X}, Y, \tilde{X}_0 \sim R(\cdot|Y, \tilde{X})} \left[ \log \tilde{p}(\tilde{X}_0|Y) - \log r(\tilde{X}_0|Y, \tilde{X}) \right]}_{E_{\tilde{X}, Y} \left[ KL(R(\cdot|Y, \tilde{X}), \tilde{P}(\tilde{X}_0|Y)) \right]}$$

- Parametric models for $\tilde{P}(X_0|Y)$, $\tilde{P}(X|Y, X_0)$ and $R(\tilde{X}_0|\tilde{X}, Y)$.

**Stochastic Gradient Descent**

- Sampling on $(X, Y, \tilde{X}_0 \sim R)$ for $E_{\tilde{X}, Y, \tilde{X}_0 \sim R(\cdot|Y, \tilde{X})} \left[ \nabla \log \tilde{p}(\tilde{X}|Y, \tilde{X}_0) \right]$

- Sampling on $(X, Y)$ for $E_{\tilde{X}, Y} \left[ \nabla KL(R(\cdot|Y, \tilde{X}), \tilde{P}(\tilde{X}_0|Y)) \right]$ if closed formula.

- Reparametrization trick for the second term otherwise...
Reparametrization Trick

Define a random variable $Z$ as the image by a parametric map $G$ of a random variable $\omega$ of fixed distribution $Q$.

Most classical case: Gaussian...

Allow to compute the derivative the expectation of a function of $Z$ through a sampling of $\omega$.

Application for ELBO:

- $\tilde{X}_0 = G_R(Y, \tilde{X}, \omega_R)$ with $\omega_R \sim Q(\cdot \mid Y, \tilde{X})$ a fixed probability law.
- Sampling on $\omega$ to approximate:

$$\nabla \mathbb{E}_{\tilde{X}, Y, \tilde{X}_0 \sim R(\cdot \mid Y, \tilde{X})} \left[ \log \tilde{p}(\tilde{X}_0 \mid Y) - \log r(\tilde{X}_0 \mid Y, \tilde{X}) \right]$$

$$= \mathbb{E}_{\tilde{X}, Y, \omega_0 \sim Q(\cdot \mid Y, \tilde{X})} \left[ \nabla \log \tilde{p}(G_R(Y, \tilde{X}, \omega_0) \mid Y) - \nabla \log r(G_R(Y, \tilde{X}, \omega_0) \mid Y, \tilde{X}) \right]$$
\( Z \sim \mathcal{N}(\mu, \sigma^2) \)  
\( Z = \mu + \sigma \tilde{Z} \)  
\( \tilde{Z} \sim \mathcal{N}(0, \sigma) \)

\[
\mathbb{E}_Z \left[ B(Z) \right] = \mathbb{E}_Z \left[ B(\mu + \sigma \tilde{Z}) \right]
\]

\[
\nabla \left. \mathbb{E}_Z \left[ C(Z) \right] \right|_{\mu} = \nabla \left. \mathbb{E}_Z \left[ C(\mu + \sigma \tilde{Z}) \right] \right|_{\mu}
\]

\[
\mathcal{W} \left[ \int \mathcal{L}(\tilde{z}) \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{(z-\mu)^2}{2\sigma^2}} \, dz \right] = \mathbb{E}_Z \left[ \nabla \mathcal{L}(\mu + \sigma \tilde{Z}) \right]
\]

\[
\mathcal{W} = \int \frac{1}{z - \mu} - e^{-\frac{(z-\mu)^2}{2\sigma^2}} \, dz = \frac{1}{\sqrt{2\pi \sigma^2}} \int e^{-\frac{(z-\mu)^2}{2\sigma^2}} \, dz = \frac{1}{\sqrt{2\pi \sigma^2}} \int e^{\frac{-\left( \frac{z-\mu}{\sigma} \right)^2}{2}} \, dz \]

\[
\mathcal{W} = \frac{1}{\sqrt{2\pi \sigma^2}} \left[ \sum \mathcal{W}(\nu + \sigma \tilde{Z}) \right]
\]
Variational Auto Encoder

Generation: \[ \tilde{X}_0 \sim \tilde{P}(\cdot | Y) \xrightarrow{\text{decoder}} \tilde{X} \sim \tilde{P}(\cdot | Y, X_0) \]

Training: \[ X \sim P(\cdot | Y) \xrightarrow{\text{encoder}} X_0 \sim R(\cdot | Y, X) \xrightarrow{\text{decoder}} \tilde{X} \sim \tilde{P}(\cdot | Y, X_0) \]

Variational Auto Encoder

- Training structure similar to classical autoencoder...but matching on distributions rather than samples.
- Encoder interpretation of the approximate posterior \( R(\cdot | Y, X) \).
- Implicit *low* dimension for \( X_0 \).
Latent Variables

\[
\omega_0 \sim \tilde{Q}_0(\cdot|Y) \quad \text{and} \quad \tilde{X}_0 = G_0(\omega_0)
\]

\[
\omega_{t+1} \sim \tilde{Q}_{t+1}(\cdot|Y, X_k) \quad \text{and} \quad \tilde{X}_{t+1} = G_{t+1}(Y, \tilde{X}_t, \omega_{t+1})
\]

\[
\tilde{X} = \tilde{X}_T
\]

Latent Variables

- Deeper hierarchy is possible...
- ELBO scheme still applicable using decoders \(R_i\)

\[
R_i(\tilde{X}_i|Y, \tilde{X}_{i+1}) \simeq \tilde{P}(\tilde{X}_i|Y, \tilde{X}_{i+1})
\]
Outline

1. Unsupervised Learning?
2. A First Glimpse
   - Clustering
   - Dimensionality Curse
   - Dimension Reduction
   - Generative Modeling
3. Dimension Reduction
   - Simplification
   - Reconstruction Error
   - Relationship Preservation
   - Comparing Methods?
4. Clustering
   - Prototype Approaches
   - Contiguity Approaches
   - Agglomerative Approaches
   - Other Approaches
   - Scalability
5. Generative Modeling
   - (Plain) Parametric Density Estimation
   - Latent Variables
   - Approximate Simulation
   - Diffusion Model
   - Generative Adversarial Network
6. References
Energy Based Model and MCMC Simulator

\[ d\tilde{P}(\tilde{X} \mid Y) \propto e^{u(\tilde{X}, Y)} d\lambda \]

\[ \rightarrow \omega_{t+1} \sim \tilde{Q}_u(\cdot \mid Y, \tilde{X}_t) \quad \text{and} \quad \tilde{X}_{t+1} = G_u(Y, X_t, \omega_{t+1}) \]

\[ \tilde{X} \approx \lim \tilde{X}_t \]

- Explicit conditional density model up to normalizing constant

\[ Z(Y) = \int e^{u(x, Y)} d\lambda(x) \]

### Simulation

- Several MCMC schemes to simulate the law without knowing \( Z(Y) \)

### Estimation

- Not so easy as \( Z(Y) \) depends a lot on \( u \) itself.
MCMC Simulation - Metropolis-Hastings

\[ X_{t+1/2} \sim Q_u(\cdot \mid Y, \tilde{X}_t) \]

\[ X_{t+1} = \begin{cases} 
X_{t+1/2} & \text{with proba } \alpha_t \\
X_t & \text{with proba } 1 - \alpha_t 
\end{cases} \]

with \( \alpha_t = \min \left( 1, \frac{e^{u(X_{t+1/2}, Y)} Q_u(\tilde{X}_t \mid Y, \tilde{X}_{t+1/2})}{e^{u(X_t, Y)} Q_u(\tilde{X}_{t+1/2} \mid Y, \tilde{X}_t)} \right) \)

**Metropolis Hastings**

- Most classical algorithm.
- Convergence guarantee under reversibility of the proposal.
- Main issue is the choice of this proposal \( \tilde{Q} \).
- Many enhanced versions exist!

Monte Carlo Markov Chain
MCMC Simulation - Langevin

\[ \tilde{X}_{t+1/2} = X_t + \gamma_t \nabla \tilde{X} u(\tilde{X}_t, Y) + \sqrt{2 \gamma_t} \omega_t \quad \text{with } \omega_t \sim \mathcal{N}(0, \text{Id}) \]

\[ \tilde{X}_{t+1} = \begin{cases} \tilde{X}_{t+1/2} \quad \text{with proba } \alpha_t \\ \tilde{X}_t \quad \text{with proba } 1 - \alpha_t \end{cases} \quad \text{with } \alpha_t = \min \left( 1, \frac{e^{u(\tilde{X}_{t+1/2}, Y)} \tilde{Q}_u(\tilde{X}_t | Y, \tilde{X}_{t+1/2})}{e^{u(\tilde{X}_t, Y)} \tilde{Q}_u(\tilde{X}_{t+1/2} | Y, \tilde{X}_t)} \right) \]

- Connection with a SDE:

\[ \frac{d\tilde{X}}{dt} = \nabla \tilde{X} u(\tilde{X}, Y) + \sqrt{2} dB_t \]

where \( B_t \) is a Brownian Motion.

Langevin

- If \( \gamma_t = \gamma \), Metropolis-Hasting algorithm.
- With \( \tilde{X}_{t+1} = \tilde{X}_{t+1/2} \), convergence toward an approximation of the law.
- Connection with SGD with decaying \( \alpha_t \)
EBM Estimation

\[ X \mid Y \sim P(\cdot \mid Y) \longrightarrow \tilde{X} \mid Y \sim \tilde{P}(\cdot \mid Y) \quad \text{with} \quad d\tilde{P}(x \mid Y) = \tilde{p}(x \mid Y) d\lambda \propto e^{u(x, Y)} d\lambda \]

- Intractable log-likelihood:
  \[ \log \tilde{p}(\tilde{x} \mid Y) = u(\tilde{x}, Y) - \log Z(u) \]

Estimation

- **Contrastive:** simulate some \( \tilde{P} \) at each step and use
  \[ \nabla \log \tilde{p}(\tilde{x} \mid Y) = \nabla u(\tilde{x}, Y) - \nabla \log Z(u) = \nabla u(\tilde{x}, Y) - \mathbb{E}_{\tilde{P}} \left[ \nabla u(X, Y) \right] \]

- **Noise contrastive:** learn to discriminate \( W = X \) from
  \( W = X' \sim R(\cdot \mid Y) \sim e^{r(x, Y)} d\lambda \) with the parametric approximation
  \[ \mathbb{P}(W = X \mid Y) \simeq \frac{e^{u(x, Y)}}{e^{u(x, Y)} + Z(Y) e^{r(x, Y)}} \]

- **Score based:** learn directly
  \[ s(\cdot \mid Y) = \nabla \tilde{X} u(\cdot \mid Y) \simeq \nabla X \log p(\cdot \mid Y). \]
Score Based Method

\[
\mathbb{E}\left[\|\nabla_X \log p(X|Y) - s(X|Y)\|^2\right] = \mathbb{E}\left[\frac{1}{2} \|s(x|Y)\|^2 + \text{tr} \nabla_X s(X, Y)\right] + \text{cst}.
\]

Score Based Method

- Non trivial formula based on partial integration.
- Hard to use in high dimension

\[
X_\sigma = X + \sigma \epsilon \quad \rightarrow \quad \mathbb{E}\left[\|\nabla_X \log p_\sigma(X_\sigma|Y) - s_\sigma(W, Y)\|^2\right] 
\]

\[
= \mathbb{E}\left[\|\nabla_X \log p_\sigma(X_\sigma|X, Y) - s_\sigma(X_\sigma, Y)\|^2\right] + \text{cst}.
\]

Noisy Score

- Connection to denoising through Tweedie formula

\[
\mathbb{E}[X|X_\sigma] = X_\sigma + \sigma^2 \nabla_X \log p_\sigma(X_\sigma|X, Y) \quad \text{and thus} \quad s_\sigma(X_\sigma, Y) \sim \frac{\mathbb{E}[X|X_\sigma] - X_\sigma}{\sigma^2}
\]
Better Exploration with Annealing and Noisy Score

**Annealing**
- Simulate a sequence of $\tilde{X}_T$ starting with $T$ large and decaying to 1.

$$X_\sigma = X + \sigma \epsilon \quad \rightarrow \quad \mathbb{E} \left[ \| \nabla_X \log p_\sigma (X_\sigma | Y) - s_\sigma (W, Y) \|^2 \right]$$

$$= \mathbb{E} \left[ \| \nabla_X \log p_\sigma (X_\sigma | X, Y) - s_\sigma (X_\sigma, Y) \|^2 \right] + \text{cst.}$$

**Noisy Score**
- Simulate a noisy sequence of $\tilde{X}_\sigma$ with $\sigma$ decaying to 0.
Outline

1 Unsupervised Learning?

2 A First Glimpse
   - Clustering
   - Dimensionality Curse
   - Dimension Reduction
   - Generative Modeling

3 Dimension Reduction
   - Simplification
   - Reconstruction Error
   - Relationship Preservation
   - Comparing Methods?

4 Clustering
   - Prototype Approaches
   - Contiguity Approaches
   - Agglomerative Approaches
   - Other Approaches
   - Scalability

5 Generative Modeling
   - (Plain) Parametric Density Estimation
   - Latent Variables
   - Approximate Simulation
   - Diffusion Model
   - Generative Adversarial Network

5 References
Noisy Model: Forward and Reverse Views

Forward: \[ X_{t+1} = X_t + \sigma_t \omega_t \rightarrow X_{t+1} \mid X_0 \sim N(X_0, s_t^2 = \sum_{t' \leq t} \sigma_{t'}^2) \]

Reverse: \[ \tilde{X}_T \sim N(0, s_T^2) \rightarrow \tilde{X}_{t-1} = \tilde{X}_t + \gamma_t u_{s_t^2}(X_t) + \sqrt{2\gamma_t} \omega_t \]

Noisy Model

- Construct a sequence of noisy version \( X_t \).
- Each \( X_t \) is easily sampled from \( X_0 \) so that the scores \( u_{s_t^2} \) can be estimated.
- Use a simplistic sequential Langevin approach to obtain \( \tilde{X}_0 \sim \tilde{P}(X) \) from \( \tilde{X}_T \sim N(0, s_T^2) \).

- Lot of approximations everywhere.
- Dependency on \( Y \) has been removed for sake of simplicity.
Diffusion with a Forward Point of View

Forward:

\[ X_{t+\delta_t} = (1 + \alpha_t \delta_t)X_t + \sqrt{2\beta_t \delta_t} \omega_t \]

\[ \rightarrow dX(t) = \alpha(t)X(t)dt + \sqrt{2\beta(t)}dB(t) \]

Forward diffusion

- Generalization of noisy model:

\[ X(t)|X(0) = N \left( X(0) \exp \int_0^t \alpha(u)du, \int_0^t 2\beta(u) \exp \left( \int_0^t \alpha(v)dvdu \right) \right) \]

Reverse:

\[ dX(t) = (-2\beta(t) \nabla_X \log P(X, t) - \alpha(t)X(t)) dt + \sqrt{2\beta(t)}dB(t) \]

\[ \rightarrow X_{t-\delta_t} = (1 - \alpha_t \delta_t)X_t + 2\beta_t \nabla_X \log p(X, t) \delta_t + \sqrt{2\beta_t \delta_t} \omega_t \]

Reverse diffusion

- Allow to sample back in time \( X_t|X_T \).
- Quite involved derivation...but Langevin type scheme in the end.
Noise Conditioned Score and Denoising Diffusion

$$\alpha_t = 0 \rightarrow \alpha(t) = 0 \rightarrow X(t)|X(0) = N \left( X(0), 2 \int_0^t \beta(u) du \right)$$

### Noise Conditioned Score

- Direct extension of noisy model.
- Better numerical scheme but numerical explosion for $X(t)$.

\[
(1 + \alpha_t \delta_t) = \sqrt{1 - 2\beta_t \delta_t} \approx 1 - \beta_t \delta_t
\]

$$\rightarrow X(t)|X(0) = N \left( X(0) \int_0^t e^{-\beta(u)} du, 2 \left( 1 - \int_0^t e^{-\beta(u)} \right) \right)$$

### Denoising Diffusion Probabilistic Model

- Explicit decay of the dependency on $P(X)$ and control on the variance.
- Better numerical results.

- Scores $\nabla_X \log p(X, t)$ estimated using the denoising trick as $X(t)|X(0)$ is explicit.
- Choice of $\beta(t)$ has a numerical impact.
A Possible Shortcut?

Forward (SDE): \[ dX(t) = \alpha(t)X(t)dt + \sqrt{2}\beta(t)dB_t \]

Backward (ODE): \[ dX(t) = (-2\beta(t)\nabla_X \log P(X, t) - \alpha(t)X(t))dt \]

Deterministic Reverse Equation

- If \( X(T) \) is initialized with the law resulting from the forward distribution, the marginal of the reverse diffusion are the right ones.
- No claim on the trajectories...but irrelevant in the generative setting.
- Much faster numerical scheme...

- Stability results on the score estimation error and the numerical scheme exist for both the stochastic and deterministic case.
Connection between Diffusion and VAE

\[ X \sim P \xleftarrow{Q(X_1|X)} X_1 \xleftarrow{Q(X_2|X_1)} X_1 \ldots \xleftarrow{Q(X_{t+1}|X_t)} \ldots X_{T-1} \xleftarrow{Q(X_T|X_{T-1})} X_T \sim P_T \]

- Generation of \( X \) from \( X_T \) using \( P(X_t|X_{t+1}) \).
- Encoder/Forward diffusion: \( Q(X_{t+1}|X_t) \).

**Variational Auto-Encoder**

- \( P_T \) is chosen as Gaussian.
- Both generative \( P(X_t|X_{t+1}) \) and encoder \( Q(X_{t+1}|X_t) \) have to be learned.

**Approximated Diffusion Model**

- \( Q(X_{t+1}|X_t) \) is known and \( P_T \) is approximately Gaussian.
- Generative \( P(X_t|X_{t+1}) \) has to be learned.

- Denoising tricks can be obtained as an ELBO starting from \( Q(X_{t+1}|X_t) = Q(X_{t+1}|X_t,X) \ldots \)
Outline

1. Unsupervised Learning?
2. A First Glimpse
   - Clustering
   - Dimensionality Curse
   - Dimension Reduction
   - Generative Modeling
3. Dimension Reduction
   - Simplification
   - Reconstruction Error
   - Relationship Preservation
   - Comparing Methods?
4. Clustering
   - Prototype Approaches
   - Contiguity Approaches
   - Agglomerative Approaches
   - Other Approaches
   - Scalability
5. Generative Modeling
   - (Plain) Parametric Density Estimation
   - Latent Variables
   - Approximate Simulation
   - Diffusion Model
   - Generative Adversarial Network
6. References
Generative Adversarial Network

$$\omega \sim \tilde{Q}(\cdot | Y) \text{ and } \tilde{X} = G(Y, \omega)$$

Non density based approach

- Can we optimize $G$ without thinking in term of density (or score)?

$$(\overline{X}, Z, Y) = \begin{cases} (X, 1, Y) & \text{with proba } 1/2 \\ (G(Y, \omega), 0, Y) & \text{otherwise} \end{cases}$$

GAN Approach

- Can we guess $Z$ with a discriminator $D(\overline{X}, Y)$?
- No if $G$ is perfect!
GAN Program

$$\max_G \min_D \mathbb{E}[\bar{X}, Y] \ell(D(\bar{X}, Y)$$

$$= \max_G \min_D \left( \frac{1}{2} \mathbb{E}_{X,Y}[\ell(D(X, Y), 1)] + \frac{1}{2} \mathbb{E}_\omega[\ell(D(G(Y, \omega), Y), 0)] \right)$$

 Discrimination

- Similar idea than the noise contrastive approach in EBM.

- If $\ell$ is a convexification of the $\ell^{0/1}$ loss then the optimal classifier is given by

$$D(\bar{X}|Y) = \begin{cases} 1 & \text{if } p(\bar{X}|Y) > \tilde{p}(\bar{X}|Y) \\ 0 & \text{otherwise.} \end{cases}$$

- If $\ell$ is the log-likelihood

$$\max_G \min_D \mathbb{E}[\bar{X}, Y] \ell(D(\bar{X}, Y) = \max_G \log_2 -\mathbb{E}_Y \left[ JK L_{1/2}(p(\cdot|Y), \tilde{p}(\cdot|Y)) \right]$$

- Direct (approximate) optimization using only samples (with the reparameterization trick).
Extensions to $f$ Divergences

$$D_f(P, Q) = \int f \left( \frac{p(x)}{q(x)} \right) q(x) = \sup_T \mathbb{E}_{X \sim P}[T(X)] - \mathbb{E}_{G \sim Q}[f^*(T(G))]$$

**$f$-GAN**

- Optimization of
  $$\min_G \sup_T \left( \mathbb{E}_{X,Y}[T(X)] - \mathbb{E}_{\omega,Y}[f^*(T(G(Y,\omega)))] \right)$$

- Direct (approximate) optimization using only samples (with the reparametrization trick).

- Direct extension of the previous scheme.

- $T$ is not a discriminator but there is an explicit link when $f(u) = \log(u)$. 
Wasserstein GAN

\[ W(P, Q) = \inf_{\xi \in \Pi(P, Q)} \mathbb{E}_{(p, q) \sim \xi} [\|p - q\|] \]

\[ = \frac{1}{K} \sup_{\|f\|_L \leq K} \mathbb{E}_{X \sim P}[f(X)] - \mathbb{E}_{G \sim Q}[f(G)] \]

Wasserstein GAN

- Optimization of

\[ \min_G \sup_{\|f\|_L \leq 1} \mathbb{E}_{X \sim P}[f(X)] - \mathbb{E}_{Z}[f(G(Z))] \]

- Direct (approximate) optimization using only samples (with the reparametrization trick).

- More stability but hard to optimize on all the 1-Lipschitz functions.
Outline

1. Unsupervised Learning?
2. A First Glimpse
   - Clustering
   - Dimensionality Curse
   - Dimension Reduction
   - Generative Modeling
3. Dimension Reduction
   - Simplification
   - Reconstruction Error
   - Relationship Preservation
   - Comparing Methods?
4. Clustering
   - Prototype Approaches
   - Contiguity Approaches
   - Agglomerative Approaches
   - Other Approaches
   - Scalability
5. Generative Modeling
   - (Plain) Parametric Density Estimation
   - Latent Variables
   - Approximate Simulation
   - Diffusion Model
   - Generative Adversarial Network
6. References

References
References


Ch. Bouveyron, G. Celeux, B. Murphy, and A. Raftery. *Model-Based Clustering and Classification for Data Science.* Cambridge University Press, 2019


D. Foster. *Generative Deep Learning (2nd ed.)* O’Reilly, 2023

Licence and Contributors

Creative Commons Attribution-ShareAlike (CC BY-SA 4.0)

- You are free to:
  - **Share:** copy and redistribute the material in any medium or format
  - **Adapt:** remix, transform, and build upon the material for any purpose, even commercially.
- Under the following terms:
  - **Attribution:** You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
  - **ShareAlike:** If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original.
  - **No additional restrictions:** You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

Contributors

- **Main contributor:** E. Le Pennec