# Probabilistic Modeling and Reasoning

Traiko Dinev <traiko.dinev@gmail.com>

August 22, 2019

*NOTE: This partially follows Probabilistic Modeling and Reasoning, a masters level course at the University of Edinburgh.*

*NOTE: Note this "summary" is NOT a reproduction of the course materials nor is it copied from the corresponding courses. It was entirely written and typeset from scratch.*



## 1 Probability Identities

Non-exhaustive list of identities useful for the rest of this cheatsheet:

$$\mathrm{P}(A, B) = \mathrm{P}(A \mid B)\ \mathrm{P}(B) \qquad \text{product rule}$$

$$\mathrm{P}(A) = \sum_B \mathrm{P}(A \mid B)\ \mathrm{P}(B) = \sum_B \mathrm{P}(A, B) \qquad \text{sum rule}$$

$$x \perp\!\!\!\perp y \iff \mathrm{P}(x, y) = \mathrm{P}(x)\ \mathrm{P}(y)$$

$$x_1 \perp\!\!\!\perp x_2 \perp\!\!\!\perp, \ldots, x_N \iff \mathrm{P}(x1, \ldots, x_N) = \prod_i \mathrm{P}(x_i)$$

We can store discrete distributions as tables of data. Conditional independece allows us to save space. Consider the conditional independence rules first:

$$\mathrm{P}(x, y \mid z) = \mathrm{P}(x \mid z)\ \mathrm{P}(y \mid z) \iff x \perp\!\!\!\perp y \mid z$$

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

Then to store $\mathrm{P}(x, y, z) = \mathrm{P}(x)\mathrm{P}(y)\mathrm{P}(z)$ we would need $dim(x) \times dim(y) \times dim(z)$ space. If they are all equal this means $k^{3d} - 1$ entries. The factorization allows us to have $3(k^d - 1)$ entries instead.

## 2 Directed Graphical Models

From the chain rule, we can factorize any distribution as:

$$\mathrm{P}(\mathbf{x}) = \prod_i \mathrm{P}(x_i \mid \pi_i), \quad \pi_i = \{x_1, \ldots, x_{i-1}\}$$

We can prove that the following holds true by induction:

$$\mathrm{P}(\mathbf{x}) = \prod_i \mathrm{P}(x_i \mid \pi_i) \iff x_i \perp\!\!\!\perp (\mathrm{pre}_i \setminus \pi_i) \mid \pi_i,\ \forall i$$

where $\pi_i$ is some subset of elements. This is to say that the factorization implies independence and a set of independences implies a factorization.

Thus we can visualize a distribution by drawing a DAG (directed acyclic graph) where the parents are the above $\pi_i$'s. Thus if:

$$\mathrm{P}(\mathbf{x}) = \mathrm{P}(x_1)\ \mathrm{P}(x_2)\ \mathrm{P}(x_3 \mid x_1, x_2)\ \mathrm{P}(x_4 \mid x_3)\ \mathrm{P}(x_5 \mid x_2)$$

then the DAG is in Figure 1.

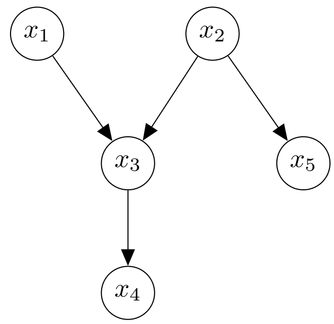



Figure 1: Simple DAG

A graph can be generated from a distribution and a (topological) ordering of the elements. A topological ordering is one where the parents come before the children. Note that different orderings may generate different graphs.

## 2.1 Examples

**Markov Models** (of order 1) or chains are a series of serial connections (Figure 2).

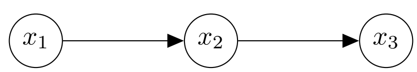



Figure 2: Markov Chain

**Hidden Markov Models** contain a markov chain that is not observed. Each hidden **h** influences an observable. **x**'s are often at different timesteps, making the chain represent a time series. Figure 3.

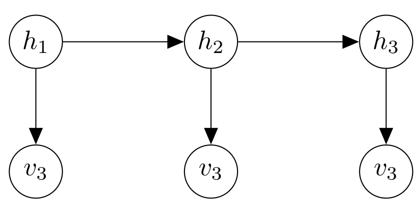



Figure 3: Hidden Markov Model

**Probabilistic PCA/ Independent Component Analysis** are methods that both use tha same graphical model. Here the latents (hiddens) variables are not connected. At the same time they influence all of the observables. Figure 4

## 2.2 D-Separation

The main reason for using graphical models is to more easily determinite independencies between variables. In a DAG a tool for using this is D-separation. We start by examining the three possible trail connections in a DAG.

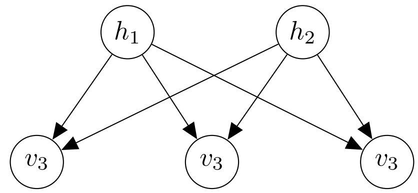



Figure 4: PPCA/ ICA graphical model

Note that these are not all possible connections between three elements, but rather *all possible connections when following a trail.*

**Serial Connections** are like Markov chains.

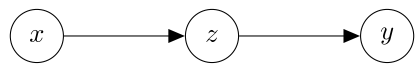



Figure 5: Serial Connecton

Importantly we have:

$$\mathrm{P}(x, z, y) = \mathrm{P}(x)\mathrm{P}(z \mid x)\mathrm{P}(y \mid z)$$
$$x \perp\!\!\!\perp y \mid z \qquad x \not\!\perp\!\!\!\perp y$$

This means that if we know the variable $z$, x and y, **and all their parents and children** that are not connected to $z$ are independent of each other.

**Diverging Connections**

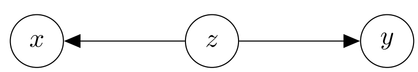



Figure 6: Diverging Connecton

The same property of independence holds true here:

$$\mathrm{P}(x, z, y) = \mathrm{P}(z)\mathrm{P}(x \mid z)\mathrm{P}(y \mid z)$$
$$x \perp\!\!\!\perp y \mid z \qquad x \not\!\perp\!\!\!\perp y$$

**Converging Connections (Colliders)**

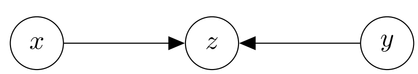



Figure 7: Collider

For colliders if we **do not know** $z$, $x$ and $y$ are independent:

$$\mathrm{P}(x, z, y) = \mathrm{P}(z)\mathrm{P}(x \mid z)\mathrm{P}(y \mid z)$$
$$x \perp\!\!\!\perp y \qquad x \not\!\perp\!\!\!\perp y \mid z$$

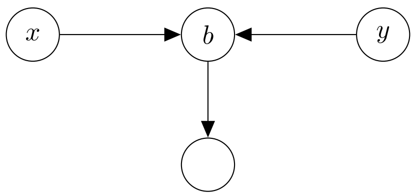
This is true since $\mathrm{P}(\cdot) =$...

**D-Separation** Sets $X$ and $Y$ are d-separated by $Z$ iff all trails are blocked by $Z$. One of the following needs to be true for a trail to be blocked:

**1)** Either $b$ is in a head-tail or tail-tail configuration


and $b$ is in $Z$.

**2)** $b$ is a part of a collider




and neither b or its descendents are in $z$. **Then** $X \perp\!\!\!\perp Y \mid Z$.

## 2.3 I-Maps

A graph is an I-map for a set of independencies $I$ iff all independencies asserted by the graph are part of $I$. A graph can thus have fewer independencies than the set. A fully-connected graph is a trivial I-map for all sets $I$.

## 2.4 Directed Local Markov Property

$$\mathbf{x}_i \perp\!\!\!\perp (\mathrm{pre}_i \setminus \mathrm{pa}_i) \mid \mathrm{pa}_i \leftrightarrow \mathbf{x}_i \perp\!\!\!\perp (\mathrm{nondesc}(\mathbf{x}_i) \setminus \mathrm{pa}_i) \mid \mathrm{pa}_i$$

This [todo] figure.

## 2.5 Gloabl directed Markov Property

All independencies asserted by D-separation.

## 2.6 Markov Blanket

By definition:

$$x \perp\!\!\!\perp (\mathrm{all} \setminus \mathbf{x} \setminus \mathrm{MB}(\mathbf{x})) \mid \mathrm{MB}(\mathbf{x})$$

And for DAGs we get:

$$\mathrm{MB}(\mathbf{x}) = \mathrm{pa}(\mathbf{x}) \ \cup \ \mathrm{children}(\mathbf{x}) \ \cup \ \mathrm{co\text{-}parents}(\mathbf{x})$$

# 3 Undirected Graphical Models

Firstly, we note the following. For non-negative functions $a$ and $b$:

$$x \perp\!\!\!\perp y \mid z \leftrightarrow \mathrm{P}(x,y,z) = a(x,z) \times b(y,z)$$
$$x \perp\!\!\!\perp y \leftrightarrow \mathrm{P}(x,y) = a(x) \times b(y)$$
$$\sum_{x,y,z} a(x,z)b(y,z) = 1$$
$$\textbf{if } p(x,y,z) = \frac{1}{Z}\phi_A(x,z)\phi_B(y,z), \quad Z = \sum_{x,y,z} \phi_A(x,z)\phi_B(y,z)$$

## 3.1 Gibbs Distribution

.. is a distribution that factorizes as:

$$\mathrm{P}(\mathbf{x}) = \frac{1}{Z}\prod_c \phi_C(\mathcal{X}_C), \qquad \mathcal{X}_C \subseteq \{x_1, \ldots, x_d\}$$

## 3.2 Energy-Based Model

If in the above $\phi_C(\mathcal{X}_C) = \exp(-\mathrm{E}_c(\mathcal{X}_c))$. Then:

$$\mathrm{P}(\mathbf{x}) = \frac{1}{Z} = \frac{1}{Z} exp\Big[-\sum_c \mathrm{E}_c(\mathcal{X}_c)\Big] = \frac{1}{Z}\prod_c \underbrace{\exp^{-\mathrm{E}_c(\mathcal{X}_c)}}_{\phi_c(\mathcal{X}_c)}$$

## 3.3 Undirected Graphs

Assuming a distribution (up to a constant) factorizes as:

$$\mathrm{P}(\mathbf{x}) \propto \phi_1(x_1,x_2,x_3)\phi_2(x_2,x_3,x_4)\phi_3(x_3,x_5)\phi_4(x_5,x_6)$$

then we visualize it as the following graph:

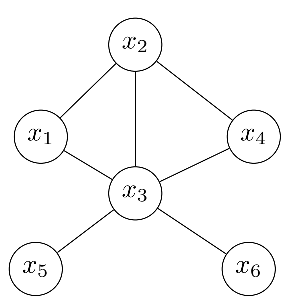



Figure 8: Undirected Graph

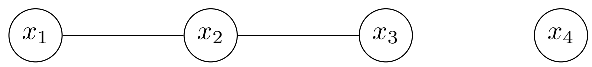
We form cliques for all variables in each factor $\phi_i$.

### 3.4 Independencies in Undirected Models

If

$$\mathrm{P}(\mathbf{x}) \propto \phi_1(x_1, x_2)\phi_2(x_2, x_3)\phi_3(x_4)$$

Then the corresponding graph is:




This directly implies that

$$x_4 \perp\!\!\!\perp x_1, x_2, x_3$$
$$x_1 \perp\!\!\!\perp x_3 \mid x_2$$

In other words, a trail is blocked if there are no paths between the two nodes. Thus D-separation is more easily done here.

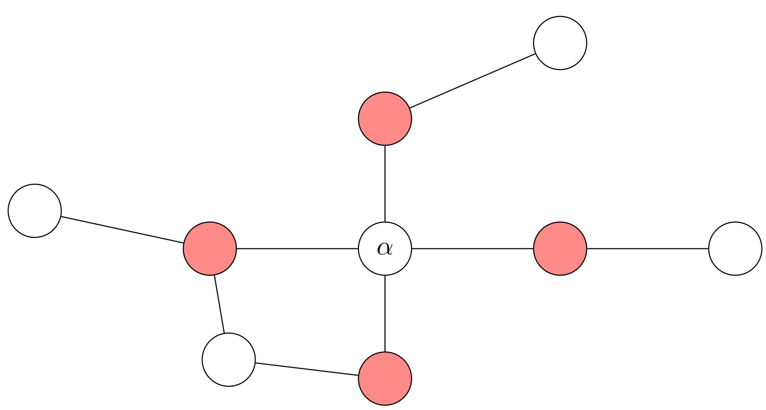
### 3.5 Graph → Distribution

Since we built undirected graphs by connecting cliques, it follows that given a graph we look at the maximum cliques to recover the distribution. **I-maps** are defined as before.

### 3.6 Local Markov Property

An edge is independent of all other edges given its neighbors. In the below graph the neighbors are colored in red. Formally, if $\mathrm{ne}(\alpha)$ are the neighbors of $\alpha$, then:

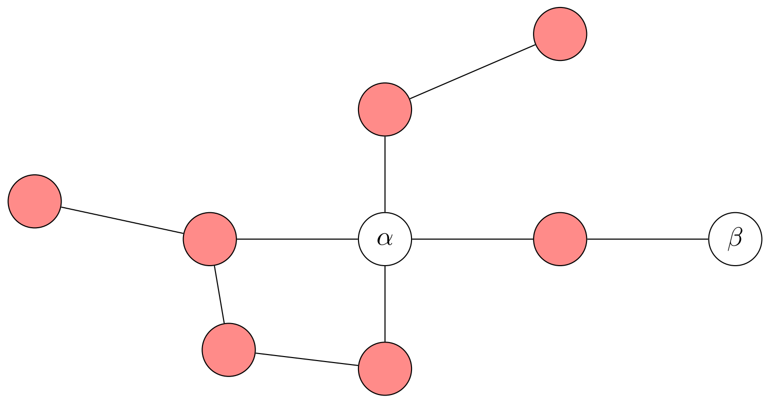
$$\alpha \perp\!\!\!\perp X \setminus (\alpha \ \cup \ \mathrm{ne}(\alpha)) \mid \mathrm{ne}(\alpha), \quad \forall \alpha \in X$$




### 3.7 Pairwise Markov Property

$$\alpha \perp\!\!\!\perp \beta \mid \underbrace{X}_{\text{all nodes}} \setminus \{\alpha, \beta\}$$

for all non-neighboring $\alpha, \beta \in X$.




### 3.8 Markov Blanket

For undirected graph, the markov blanket of $x$ is the neighbors of $x$.

N.B. All of the above properties (Markov properties) are equivalent for both directed and undirected graphical models. This means if one of them is true **for a distribution**, then all of them are true.

**TODO: Add Bishop plots with repetition boxes.**

### 3.9 Minimal I-map

- If an edge is removed, it ceases to be an imap.
- A graph is an I-map if P($\cdot$) factorizes over the graph.

**Undirected Models**

- For all $\mathbf{x}_i$ find $MB(\mathbf{x}_i)$ and connect.

**Directed Models**

- For all $\mathbf{x}_i$ find $\pi_i \subseteq \text{pre}_i$ such that $\mathbf{x}_i \perp\!\!\!\perp \{\text{pre}_i \setminus \pi_i\} \mid \pi_i$
- Set $\text{pa}_i = \pi_i$

## 4 Equivalence and Conversion Between Models

Two directed graphs $G_1$ and $G_2$ are $I$-equivalent if they have the same set of immoralities and the same skeleton. An immorality is a collider without covering edge. **Look for colliders that don't match.** Since serial (head-tail) and diverging (tail-tail) connections imply the same independencies, we only need to look for converging (head-head) connections that don't match.

## 4.1 Directed → Undirected

We have:

$$\mathrm{P}(\cdot) = \prod_i \mathrm{P}(\mathbf{x}_i \mid \mathrm{pa}_i) = \prod_i \underbrace{\phi_i(\mathbf{x}_i, \mathrm{pa}_i)}_{cliques}$$

- This is called **moralization** and we obtain a **moral graph**
- This it **NOT** an undirected I-map for the distribution. Most notably we can not represent collider independencies.

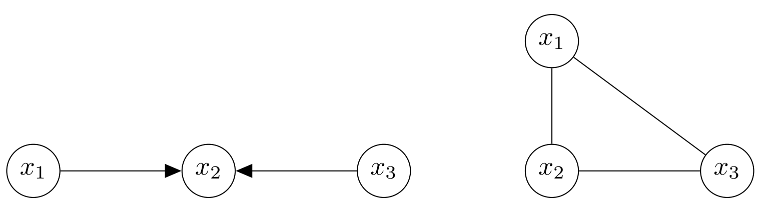
## 4.2 Undirected → Directed

(See example below)

- Choose an ordering.
- Read independencies off of the graph and find $\pi_i$ for each $i$.
- Connect $\pi_i \to \mathbf{x}_i$

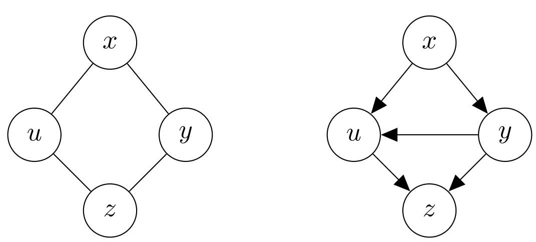
## 4.3 Non-equivalent Trails

**Colliders can not be represented by undirected graphs.**




In the moralized graph on the right the independence $\mathbf{x}_1 \perp\!\!\!\perp \mathbf{x}_3$ is lost.

**Closed loops can not be represented by directed graphs.**




Consider the left graph. Let's review the process of creating the directed graph:

- Choose an ordering: $x, y, u, z$.
- For each element, consider the parent set and read independencies off the directed graph.

- Start with $y \not\perp\!\!\!\perp x$, which implies the edge $x \to y$.
- Find the minimal set for which $u$ is independent of the parents $\pi_u$. This is $x, y$. Hence $x, y \to u$.
- For $z$ this set is $u, y$. Hence $u \to z$ and $y \to z$

**We no longer have the independence** $u \perp\!\!\!\perp y \mid x, z$.

# 5 Factor Graphs

$$\mathrm{P}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4) = \frac{1}{Z}\phi_1(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)\phi_2(\mathbf{x}_3, \mathbf{x}_4)\phi_3(\mathbf{x}_4) \tag{1}$$

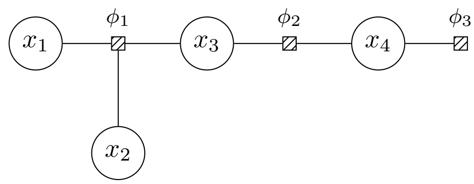



Figure 9: Factor Graph

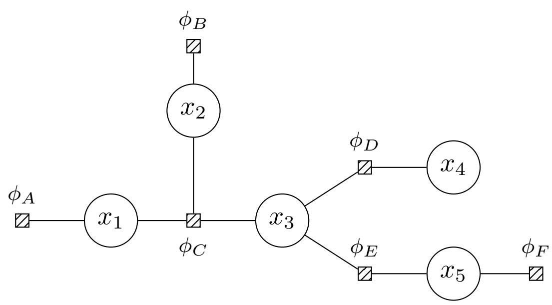
# 6 Exact Inference in Factor Graphs

Assume discrete variables. The task is to compute $\mathrm{P}(\mathbf{x}_k = k)$ for all $k$.

We can group terms in the leaves of the tree. Consider:

$$\mathrm{P}(\cdot) \propto \phi_A(\mathbf{x}_1)\phi_B(\mathbf{x}_2)\phi_C(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)\phi_D(\mathbf{x}_3, \mathbf{x}_4)\phi_E(\mathbf{x}_3, \mathbf{x}_5)\phi_F(\mathbf{x}_5) \tag{2}$$

which is:




We iteratively "eliminate" variables by summing or integrating them out. Firstly, we eliminate $\mathbf{x}_5$:

$$
\begin{aligned}
\mathrm{P}(\mathbf{x}_1, \ldots, \mathbf{x}_4) &= \sum_{\mathbf{x}_5} \mathrm{P}(\mathbf{x}_1, \ldots, \mathbf{x}_5) \\
&\propto \sum_{\mathbf{x}_5} \phi_A(\mathbf{x}_1)\phi_B(\mathbf{x}_2)\phi_C(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)\phi_D(\mathbf{x}_3, \mathbf{x}_4)\phi_E(\mathbf{x}_3, \mathbf{x}_5)\phi_F(\mathbf{x}_5) \\
&\propto \phi_A(\mathbf{x}_1)\phi_B(\mathbf{x}_2)\phi_C(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)\phi_D(\mathbf{x}_3, \mathbf{x}_4) \sum_{\mathbf{x}_5} \phi_E(\mathbf{x}_3, \mathbf{x}_5)\phi_F(\mathbf{x}_5) \\
&\propto \phi_A(\mathbf{x}_1)\phi_B(\mathbf{x}_2)\phi_C(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)\phi_D(\mathbf{x}_3, \mathbf{x}_4)\widetilde{\phi}_5(\mathbf{x}_3)
\end{aligned}
$$

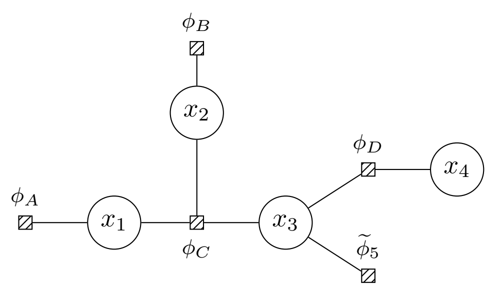
The idea is that we have reduced the factors above. Numerically, we would have the following:

$$
\widetilde{\phi}_5(\mathbf{x}_3) = \begin{cases} a & \mathbf{x}_3 = 1 \\ \ldots & \\ z & \mathbf{x}_3 = N \end{cases}
$$

For each value of the factor. This is pre-computing all of the values, which in this case costs $O(N^2)$. For each value of $\mathbf{x}_3$ we need to sum over $\mathbf{x}_5$.

If we keep on doing this, the total cost will be greatly reduced from the $O(N^4)$ that is needed to sum over $\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5$.

The above operation is represented by the following **reduced** factor graph:




## 6.1 Message Passing

**TODO: Graphs**

$$
\mu_{\phi \to \mathbf{x}}(\mathbf{x}) = \sum_{\mathbf{x}_1, \ldots, \mathbf{x}_j} \phi(\mathbf{x}_1, \ldots, \mathbf{x}_j, x) \prod_{i=1}^{j} \mu_{\mathbf{x}_i \to \phi}(\mathbf{x}_i)
$$

$$
\mu_{\mathbf{x} \to \phi}(\mathbf{x}) = \prod_{i=1}^{j} \mu_{\phi_i \to \mathbf{x}}
$$

$$
\mathrm{P}(\mathbf{x}) \propto \prod_{i=1}^{j} \mu_{\phi_i \to \mathbf{x}}(\mathbf{x})
$$

$$
\mathrm{P}(\mathbf{x}_1, \ldots, \mathbf{x}_j) \propto \phi(\mathbf{x}_1, \ldots, \mathbf{x}_j) \prod_{i=1}^{j} \mu_{\mathbf{x}_i \to \phi}(\mathbf{x}_i)
$$

# 7 Inference for Markov Chains

**TODO: Just scan this..**

## 7.1 $\alpha$-recursion

## 7.2 Smoothing

## 7.3 $\alpha$-$\beta$ Recursion

# 8 Model-Based Learning

- Probabilistic model: table for a distribution $\mathrm{P}(\mathbf{x})$
- Statistical model: set of probabilistic models: $\{\mathrm{P}(\mathbf{x};\theta)\}$
- Bayesian model: prior on theta, replace "parametrized by" with "conditioned on" $\mathrm{P}(\mathbf{x}) = \int \mathrm{P}(\mathbf{x} \mid \theta)\ \mathrm{P}(\theta)$. To get the probability distribution of $\mathbf{x}$ as above, we need to integrate over all possible values of $\theta$.

# 9 Moment Matching

Assume we have a parametric statistical model:

$$\mathrm{P}(\mathbf{x};\theta) = \frac{\mathrm{P}^*(\mathbf{x};\theta)}{Z(\theta)}$$

where $Z(\theta) = \int \mathrm{P}^*(\mathbf{x};\theta)\ \mathrm{d}\mathbf{x}$ is the partition function. Maximum likelihood returns:

$$\theta^* = \arg\max_\theta L(\theta) = \arg\max_\theta \sum_i \log \mathrm{P}(\mathbf{x}_i;\theta)$$

the last being true for i.i.d. data only. This directly implies that at the maximum, the derivative is 0:

$$\nabla_\theta L(\theta)|_{\theta^*} = 0$$

We define the moments to be:

$$m(\mathbf{x};\theta) = \nabla_\theta \log \mathrm{P}^*(\mathbf{x};\theta)$$

Then:

$$
\begin{aligned}
\nabla_\theta L(\theta) &= \nabla_\theta \sum_i \log \mathrm{P}(\mathbf{x}_i;\theta) = \nabla_\theta \sum_i \log \frac{\mathrm{P}^*(\mathbf{x}_i;\theta)}{Z(\theta)} \\
&= \left[\nabla_\theta \sum_i \log \mathrm{P}^*(\mathbf{x}_i;\theta)\right] - \nabla_\theta n \times \log Z(\theta) \\
&= \sum_i \nabla_\theta \mathrm{P}^*(\mathbf{x}_i;\theta) - n\nabla_\theta \log Z(\theta) \\
&= \sum_i m(\mathbf{x}_i;\theta) - n\nabla_\theta \log Z(\theta)
\end{aligned}
$$

then for the second term we have:

$$
\begin{aligned}
\nabla_\theta \log Z(\theta) &= \frac{1}{Z(\theta)}\nabla_\theta Z(\theta) = \frac{1}{Z(\theta)}\nabla_\theta \int \mathrm{P}^*(\mathbf{x};\theta)\mathrm{d}\mathbf{x} = \\
&= \frac{\int \nabla_\theta \mathrm{P}^*(\mathbf{x};\theta)\mathrm{d}\mathbf{x}}{Z(\theta)} = \\
&= \frac{\int \nabla_\theta \big[\log \mathrm{P}^*(\mathbf{x};\theta)\big]\mathrm{P}^*(\mathbf{x};\theta)\mathrm{d}\theta}{Z(\theta)} = \\
&= \int \nabla_\theta \big[\log \mathrm{P}^*(\mathbf{x};\theta)\big]\mathrm{P}(\mathbf{x};\theta)\mathrm{d}\mathbf{x} \\
&= \int m(\mathbf{x};\theta)\mathrm{P}(\mathbf{x};\theta)\mathrm{d}\mathbf{x}
\end{aligned}
$$

which implies that:

$$
\nabla_\theta L(\theta) = \sum_i m(\mathbf{x}_i;\theta) - n\int m(\mathbf{x};\theta)\mathrm{P}(\mathbf{x};\theta)\mathrm{d}\theta
$$

which means that at $\theta^*$, where $\nabla_\theta L(\theta) = 0$:

$$
\frac{1}{n}\sum_i m(\mathbf{x}_i;\theta) = \int m(\mathbf{x};\theta)\mathrm{P}(\mathbf{x};\theta)\mathrm{d}\theta
$$

This proves that maximumum likelihood can be thought of as moment matching. We are matching the empirical moments to their expectation over the normalized data distribution $\mathrm{P}(\mathbf{x};\theta)$. Note we don't have access to $\mathrm{P}(\mathbf{x};\theta)$, but only to $\mathrm{P}^*(\mathbf{x};\theta)$, the unnormalized distribution.

# 10 Bayesian Inference

This:

$$
\begin{aligned}
\mathrm{P}(\theta \mid \mathcal{D}) &= \frac{\mathrm{P}(\theta;\mathcal{D})}{\mathrm{P}(\mathcal{D})} = \frac{\mathrm{P}(\mathcal{D} \mid \theta)\mathrm{P}(\theta)}{\mathrm{P}(\mathcal{D})} \\
&\propto \mathrm{Likelihood}(\theta)\mathrm{P}(\theta) \propto \underbrace{\left[\prod_i \mathrm{P}(\mathbf{x}_i \mid \theta)\right]}_{\text{for i.i.d. data}} \mathrm{P}(\theta)
\end{aligned}
$$

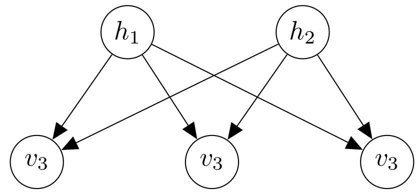



Figure 10: FA/ICA graphical model

# 11 Factor Analysis

Both FA and ICA have the same (parametrized) graphical model. See Figure 10. We have $D$ visible and $H$ hidden units:

$$\mathbf{h} \sim \mathcal{N}(\mathbf{0}, \mathbb{I}) \qquad H < D$$
$$\mathrm{P}(\mathbf{v} \mid \mathbf{h}; \theta) = \mathcal{N}(\mathbf{v}; \mathbf{F}\mathbf{h} + \mathbf{c}, \psi)$$

where:

$$\theta = \{ \underbrace{\mathbf{c}}_{\text{mean of } \mathbf{x}}, \mathbf{F} = \underbrace{(f_1, \ldots, f_H)}_{\text{factors of size D; F is DxH}}, \Psi = \underbrace{\mathrm{diag}(\psi_1, \ldots, \psi_D)}_{\text{noise variance}} \}$$

This means that the model is:

$$\begin{aligned} \mathbf{v} &= \mathbf{F}\mathbf{h} + \mathbf{c} + \epsilon \qquad \epsilon \sim \mathcal{N}(0, \Psi) \\ &= \sum_{i=1}^{H} \mathbf{f}_i h_i + \mathbf{c} + \epsilon \end{aligned}$$

## 11.1 An important identity

$$\text{if } \mathbf{x} \sim \mathcal{N}(\mu_x, \mathbf{C}_x); \qquad \mathbf{z} \sim \mathcal{N}(\mu_z, \mathbf{C}_z)$$
$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{z}, \text{ then } \qquad y \sim \mathcal{N}(\mathbf{A}\mu_x + \mu_z, \mathbf{A}\mathbf{C}_x\mathbf{A}^T + C_z)$$

## 11.2 Factor Rotation Problem

$$\begin{aligned} \mathbf{v} &= \mathbf{F}\mathbf{h} + \mathbf{c} + \epsilon = \mathbf{F}(\mathbf{R}\mathbf{R}^T)\mathbf{h} + \mathbf{c} + \epsilon \\ &= (\mathbf{F}\mathbf{R})(\mathbf{R}^T\mathbf{h}) + \mathbf{c} + \epsilon = (\mathbf{F}\mathbf{R})\bar{\mathbf{h}} + \mathbf{c} + \epsilon \end{aligned}$$

Which means that:

$$\mathrm{P}(\bar{\mathbf{h}}) = \mathcal{N}(\bar{\mathbf{h}}; 0, \mathbb{I})$$

since $\mathbf{R}$ is an orthogonal matrix.

# 12 Independent Component Analysis (ICA)

ICA uses the same model as FCA but makes fewer assumptions. The joint of the hidden is:

$$\mathrm{P}_h(\mathbf{h}) = \prod_i p_h(h_i)$$

with the observed (visible) following:

$$\mathbf{v} = \mathbf{F}\mathbf{h} + \mathbf{c} + \epsilon$$

# 13 Intractable Likelihood Functions

We'll look at three scenarios. Firstly, unobserved varibles. Assuming the following likelihood form:

$$\mathrm{L}(\theta) = \mathrm{P}(\mathcal{D};\theta) = \int_u p(\underbrace{\mathbf{u}}_{\text{Unobserved}}, \mathcal{D};\theta)\ \mathrm{d}\mathbf{u}$$

This is intractable, since we need to integrate over the observed, which are ... unobserved. We can use monte-carlo integration (below) if we know their distribution. However, it turns out that the gradient of the likelihood we can compute in a more convenient way:

$$\nabla_\theta l(\theta) = \mathbb{E}_{\mathbf{u}\sim \mathrm{P}(\mathbf{u}\ \mid\ \mathcal{D};\theta)}\Big[\nabla_\theta \log \mathrm{P}(\mathbf{u}, \mathcal{D};\theta) \mid \mathcal{D};\theta\Big]$$

It is in most cases that we can compute the joint $p(\mathbf{u}, \mathcal{D};\theta)$ and conditional $p(\mathbf{u} \mid \mathcal{D};\theta)$. We can sample from the conditional and use Monte-Carlo integration.

TODO: proof

## 13.1 Intractable Partition Function

If the partition $Z(\theta)$ is intractable, we generally have the distribution up to a constant:

$$\mathrm{P}(\mathbf{x};\theta) = \frac{\mathrm{P}^*(\mathbf{x};\theta)}{Z(\theta)}$$

Then we can use moment matching.

$$\begin{aligned}\nabla_\theta l(\theta) &= \sum_i m(\mathbf{x}_i;\theta) - n\int m(\mathbf{x};\theta)\mathrm{P}(\mathbf{x};\theta)\mathrm{d}\theta \\ &\propto \frac{1}{n}\sum_i m(\mathbf{x}_i;\theta) - \mathbb{E}\big[m(\mathbf{x};\theta)\big]\end{aligned}$$

where $m(\mathbf{x};\theta) = \nabla_\theta \mathrm{P}^*(\mathbf{x};\theta)$

# 14 Score Matching

# 15 Sampling and Monte Carlo

# 16 Sampling from Continuous Distributions

## 16.1 Rejection Sampling

## 16.2 Ancestral Sampling

## 16.3 Gibbs Sampling

# 17 Variational Inference