Directed Graphical Models

Xiaoke Zhang, Peng Chen

$School of Management$
$University of Science and Technology of China$

December 8, 2021

Table of Contents

Graphical Models

A graphical model is a family of probability distributions defined in terms of a directed or undirected graph.

- The nodes in the graph are identified with random variables, and can be divided into two categories: hidden nodes and observation nodes.
- The edges correspond to the conditional dependency or correlation relationship of the random variable, and can be directed or undirected.

Graphical models describe the complex conditional independence relationship between variables in a visual way, and can decompose a complex joint probability into the product of multiple simple conditional probabilities.

Graphical Models

The two most common forms of graphical models are:

- Undirected graphical models/ Markov random fields.
- Directed graphical models/ Bayesian network.

The usefulness of graphical models:

- In undirected graphical models, we want to know whether the two variables are directly related, which means if there are edge connections between nodes, or use the graph to infer the model parameters.
- In directed graphical models, similarly, there are two main purposes. First, learning the parameters or latent variables of the model via the directed graphical structure. Second, understanding the causality, which means determining the links with its direction from the observational data.

Directed graphical models

In directed graphical models, directed edges connects each node, indicating the conditional independence relationship between random variables. Let $\mathcal{G}(\mathcal{V}, \epsilon)$ be a directed acyclic graph, where $\mathcal{V}$ are the nodes and ϵ are the edges of the graph. Let $\{X_v : v \in \mathcal{V}\}$ be a collection of random variables indexed by the nodes of the graph. To each node $v \in \mathcal{V}$, let π_v denote the subset of indices of its parents. Then the joint probability density is

$$p\left(x_{\mathcal{V}}\right) = \prod_{v \in \mathcal{V}} p\left(x_v \mid x_{\pi_v}\right).$$

D-separation

We say an undirected path P is d-separated by a set of nodes E (containing the evidence) iff at least one of the following conditions hold:

- P contains a chain, $s \to m \to t$ or $s \leftarrow m \leftarrow t$, where $m \in E$.
- P contains a tent or fork, $s \swarrow^{m} \searrow t$, where $m \in E$.
- P contains a collider or v-structure, $s \searrow_{m} \swarrow t$, where m is not in E and nor is any descendant of m.

Next, we say that a set of nodes A is d-separated from a different set of nodes B given a third observed set E iff each undirected path from every node $a \in A$ to every node $b \in B$ is d-separated by E. Finally, we define the CI properties of a DAG as follows:

$$\mathbf{x}_A \perp \mathbf{x}_B \mid \mathbf{x}_E \iff \text{A is d-separated from B given E.}$$

Bayes ball algorithm

The Bayes ball algorithm is a simple way to see if A is d-separated from B given E. The idea is this. We "shade" all nodes in E, indicating that they are observed. We then place "balls" at each node in A, let them "bounce around" according to some rules, and then ask if any of the balls reach any of the nodes in B.

Bayes ball algorithm

The three main rules of Bayes ball algorithm:

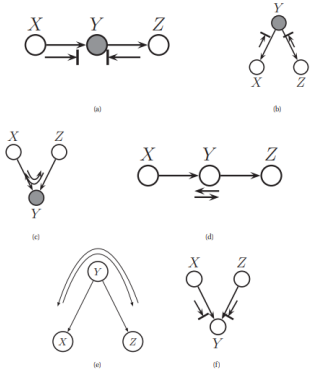


Figure: Bayes ball rules. A shaded node is one we condition on.

Bayes ball algorithm

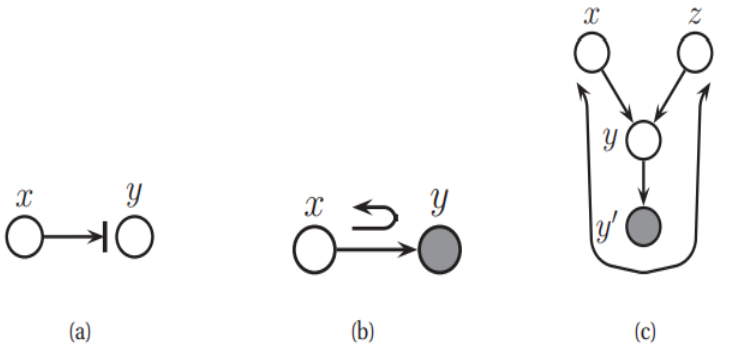


Figure: (a-b) Bayes ball boundary conditions. (c) Example of why we need boundary conditions. y' is an observed child of y, rendering y "effectively observed", so the ball bounces back up on its way from x to z.

A DGM

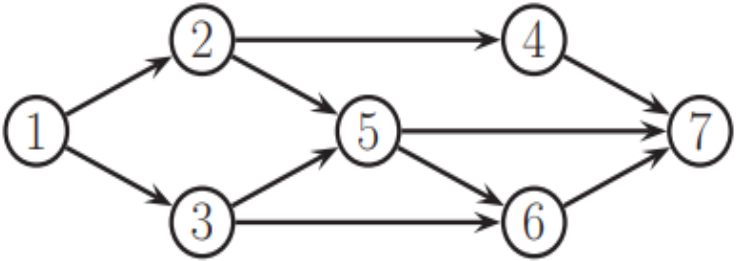


Figure: A DGM, among $2 \perp 6 \mid 1, 5$.

Bayesian network

- A Bayesian network (also known as a Bayes network, Bayes net, belief network, or decision network) is a probabilistic graphical model that represents a set of variables and their conditional dependencies via a directed acyclic graph (DAG).
- Given data x and parameter θ, a simple Bayesian network starts with a prior $p(\theta)$ and likelihood $p(x \mid \theta)$ to compute a posterior probability

$$p(\theta \mid x) \propto p(x \mid \theta)p(\theta).$$

- Often the prior on θ depends in turn on other parameters φ that are not mentioned in the likelihood, equipped with a prior $p(\varphi)$ on the newly introduced parameters φ is required, resulting in a posterior probability

$$p(\theta, \varphi \mid x) \propto p(x \mid \theta)p(\theta \mid \varphi)p(\varphi)$$

This is the simplest example of a hierarchical Bayes model.

Latent Dirichlet Allocation

- Latent Dirichlet Allocation (LDA) is a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each document is modeled as random mixtures over latent topics, and each topic is characterized by a distribution over words.
- In particular, it is assumed that the prior distribution of the topic distribution of the document is the Dirichlet distribution, and the prior distribution of the word distribution of the topic also is the Dirichlet distribution.

Latent Dirichlet Allocation

- A word is the basic unit of discrete data, defined to be an item from a vocabulary indexed by $\{w_1, \ldots, w_V\}$.
- A topic is characterized by a distribution over words. Suppose there are K topics $Z = \{z_1, \ldots, z_K\}$, and the words' probabilities are parameterized by a $K \times V$ matrix $\boldsymbol{\varphi}$, where φ_{ij} represents the probability of generating the word w_j in topic z_i.
- A document is a sequence of N_m words denoted by

$$\boldsymbol{w}_m = (w_{m1}, w_{m2}, \ldots, w_{mN_m}),$$

  where w_{mn} is the nth word in the sequence.
- A corpus is a collection of M documents denoted by

$$D = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_M\}.$$

Basic terms

Topic 77

word	prob.
MUSIC	.090
DANCE	.034
SONG	.033
PLAY	.030
SING	.026
SINGING	.026
BAND	.026
PLAYED	.023
SANG	.022
SONGS	.021
DANCING	.020
PIANO	.017
PLAYING	.016
RHYTHM	.015
ALBERT	.013
MUSICAL	.013

Topic 82

word	prob.
LITERATURE	.031
POEM	.028
POETRY	.027
POET	.020
PLAYS	.019
POEMS	.019
PLAY	.015
LITERARY	.013
WRITERS	.013
DRAMA	.012
WROTE	.012
POETS	.011
WRITER	.011
SHAKESPEARE	.010
WRITTEN	.009
STAGE	.009

Topic 166

word	prob.
PLAY	.136
BALL	.129
GAME	.065
PLAYING	.042
HIT	.032
PLAYED	.031
BASEBALL	.027
GAMES	.025
BAT	.019
RUN	.019
THROW	.016
BALLS	.015
TENNIS	.011
HOME	.010
CATCH	.010
FIELD	.010

Figure: Three topics related to the word play

Basic terms

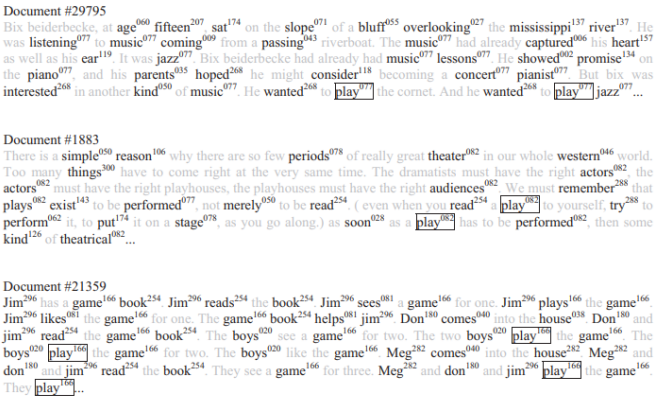


Figure: Three documents from the TASA corpus containing different senses of the word play

Dirichlet distibution

If a multivariate continuous random variable $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ has following probability density:

$$p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}{\prod_{i=1}^{K} \Gamma\left(\alpha_i\right)} \prod_{i=1}^{K} \theta_i^{\alpha_i - 1},$$

where $\sum_{i=1}^{K} \theta_i = 1, \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K), \theta_i \geq 0, \alpha_i > 0, i = 1, \ldots, K$. Then $\boldsymbol{\theta}$ is said to be a Dirichlet random variable with parameter $\boldsymbol{\alpha}$, and write as $\boldsymbol{\theta} \sim \mathrm{Dir}(\boldsymbol{\alpha})$.

Dirichlet distibution

- A K-dimensional Dirichlet random variable $\boldsymbol{\theta}$ can take values in the (K-1)-simplex, because $\theta_i \geq 0, \sum_{i=1}^{K} \theta_i = 1$.
- Dirichlet distribution belongs to the exponential distribution family.

$$p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) = exp\left\{\sum_{i=1}^{K}(\alpha_i - 1)\log\theta_i + \log\Gamma\left(\sum_{i=1}^{K}\alpha_i\right) - \sum_{i=1}^{K}\log\Gamma(\alpha_i)\right\},$$

  and $T_i(\boldsymbol{\theta}) = \log\theta_i, Q_i(\boldsymbol{\alpha}) = \alpha_i - 1, i = 1, \ldots, K$.
- Dirichlet distribution is the conjugate prior of multinomial distribution.

Generative process

Generative process for the corpus D:

- For each topic z_k $(k = 1, \ldots, K)$, generate the parameters

$$\boldsymbol{\varphi}_k \sim \mathrm{Dir}(\boldsymbol{\beta})$$

  as the words' distribution of the topic.
- For each document $\boldsymbol{w}_m$ $(m = 1, \ldots, M)$, generate the parameters $\boldsymbol{\theta}_m \sim \mathrm{Dir}(\boldsymbol{\alpha})$ as the topics' distribution of the document.
- For each word w_{mn} in the document $\boldsymbol{w}_m$, here $m = 1, \ldots, M$, and $n = 1, \ldots, N_m$:
  (a) Generate topic $z_{mn} \sim \mathrm{Mult}(\boldsymbol{\theta}_m)$ as the topic corresponding to the word.
  (b) Generate word $w_{mn} \sim \mathrm{Mult}(\boldsymbol{\varphi}_{z_{mn}})$.

Graphical model representation

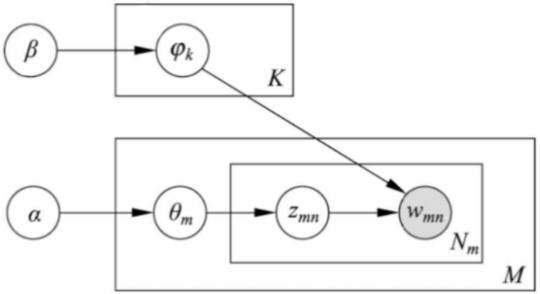


Figure: Graphical model representation of LDA. The boxes are "plates" representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document.

Unfolded graphical model representation

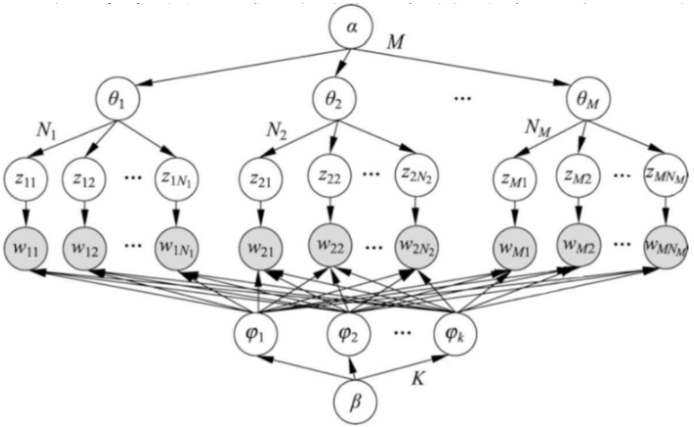


Figure: Unfolded graphical model representation of LDA

Outline

Variational inference

Let $\boldsymbol{x}$ be observed variables and $\boldsymbol{z}$ be latent variables or parameters. In variational inference, given a variational distribution family $\mathcal{Q}$, we need to find the one that is closest to the probability densities $p(\boldsymbol{z} \mid \boldsymbol{x})$ in the sense of KL divergence.
Theoretically, minimizing the KL divergence is equivalent to maximizing the evidence lower bound:

$$= \min_{q \in \mathcal{Q}} \mathrm{KL}(q(\boldsymbol{z}) \| p(\boldsymbol{z} \mid \boldsymbol{x})) = \max_{q \in \mathcal{Q}} \mathrm{ELBO}(q),$$

where $\mathbb{E}_q[\log p(\boldsymbol{z}, \boldsymbol{x})] - \mathbb{E}_q[\log q(\boldsymbol{z})]$.
Usually, assuming that the latent variables are mutually conditional independent, a generic member of the mean-field variational family is

$$q(\boldsymbol{z}) = \prod_{j=1}^{m} q_j\left(z_j\right).$$

Simplified LDA

Assuming that the parameters φ_k of the words' distribution of the topic z_k are not random, but fixed. And the number of words in different documents all are N.

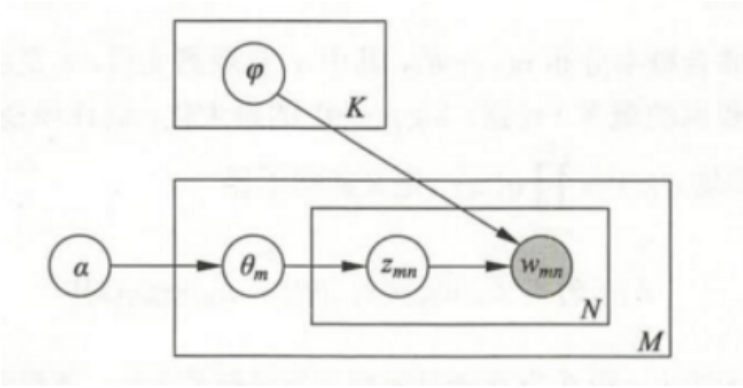


Figure: Simplified LDA

Joint distribution

For one document $\boldsymbol{w}$ (which is observable), the joint density function of $\boldsymbol{\theta}, \boldsymbol{w}, \boldsymbol{z}$ is

$$p(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w} \mid \boldsymbol{\alpha}, \boldsymbol{\varphi}) = p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) \prod_{n=1}^{N} p\left(z_n \mid \boldsymbol{\theta}\right) p\left(w_n \mid z_n, \boldsymbol{\varphi}\right),$$

where $\boldsymbol{\alpha}, \boldsymbol{\varphi}$ are parameters and $\boldsymbol{\varphi}$ is a $K \times V$ matrix which represents words' distribution of each topic.

Variational distribution

The problematic coupling between $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$ arises due to the edges between $\boldsymbol{\theta}, \boldsymbol{z}$, and $\boldsymbol{w}$. By dropping these edges and the $\boldsymbol{w}$ nodes, and endowing the resulting simplified graphical model with free variational parameters, we obtain a family of distributions on the latent variables based on mean-field. It's characterized by

$$q(\boldsymbol{\theta}, \boldsymbol{z} \mid \boldsymbol{\gamma}, \boldsymbol{\eta}) = q(\boldsymbol{\theta} \mid \boldsymbol{\gamma}) \prod_{n=1}^{N} q\left(z_n \mid \boldsymbol{\eta}_n\right).$$

Variational distribution

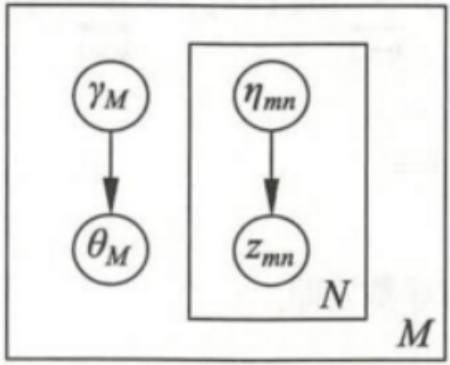


Figure: Variational distribution based on mean-field

ELBO

The ELBO of one document is

$$
\begin{aligned}
L(\boldsymbol{\gamma}, \boldsymbol{\eta}; \boldsymbol{\alpha}, \boldsymbol{\varphi}) &= \mathbb{E}_q[\log p(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w} \mid \boldsymbol{\alpha}, \varphi)] - \mathbb{E}_q[\log q(\boldsymbol{\theta}, \boldsymbol{z} \mid \boldsymbol{\gamma}, \boldsymbol{\eta})] \\
&= \mathbb{E}_q[\log p(\boldsymbol{\theta} \mid \boldsymbol{\alpha})] + \mathbb{E}_q[\log p(\boldsymbol{z} \mid \boldsymbol{\theta})] + \mathbb{E}_q[\log p(\boldsymbol{w} \mid \boldsymbol{z}, \boldsymbol{\varphi})] \\
&\quad - \mathbb{E}_q[\log q(\boldsymbol{\theta} \mid \boldsymbol{\gamma})] - \mathbb{E}_q[\log q(\boldsymbol{z} \mid \boldsymbol{\eta})].
\end{aligned}
$$

ELBO

Theorem

Based on the settings in the previous article, the ELBO of one document is

$$
\begin{aligned}
L(\boldsymbol{\gamma}, \boldsymbol{\eta}; \boldsymbol{\alpha}, \boldsymbol{\varphi}) = & \log \Gamma\left(\sum_{l=1}^{K} \alpha_l\right) - \sum_{k=1}^{K} \log \Gamma\left(\alpha_k\right) + \sum_{k=1}^{K}\left(\alpha_k - 1\right)\left[\Psi\left(\gamma_k\right) - \Psi\left(\sum_{l=1}^{K} \gamma_l\right)\right] \\
& + \sum_{n=1}^{N} \sum_{k=1}^{K} \eta_{nk}\left[\Psi\left(\gamma_k\right) - \Psi\left(\sum_{l=1}^{K} \gamma_l\right)\right] \\
& + \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{v=1}^{V} \eta_{nk} w_n^v \log \varphi_{kv} \\
& - \log \Gamma\left(\sum_{l=1}^{K} \gamma_l\right) + \sum_{k=1}^{K} \log \Gamma\left(\gamma_k\right) - \sum_{k=1}^{K}\left(\gamma_k - 1\right)\left[\Psi\left(\gamma_k\right) - \Psi\left(\sum_{l=1}^{K} \gamma_l\right)\right] \\
& - \sum_{n=1}^{N} \sum_{k=1}^{K} \eta_{nk} \log \eta_{nk}.
\end{aligned}
$$

A little trick

An exponential family has following density funciton

$$p(x \mid \eta) = h(x) \exp\left\{\eta^{\mathrm{T}} T(x) - A(\eta)\right\},$$

where θ is the natural parameter, $T(x)$ is the sufficient statistic, and moreover $A(\eta) = \log \int h(x) \exp\left\{\eta^{\mathrm{T}} T(x)\right\} \mathrm{d}x$. In fact $\frac{\mathrm{d}}{\mathrm{d}\eta} A(\eta) = E[T(X)]$,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\eta} A(\eta) &= \frac{\mathrm{d}}{\mathrm{d}\eta} \log \int h(x) \exp\left\{\eta^{\mathrm{T}} T(x)\right\} \mathrm{d}x \\
&= \frac{\int T(x) \exp\left\{\eta^{\mathrm{T}} T(x)\right\} h(x) \mathrm{d}x}{\int h(x) \exp\left\{\eta^{\mathrm{T}} T(x)\right\} \mathrm{d}x} \\
&= \int T(x) \exp\left\{\eta^{\mathrm{T}} T(x) - A(\eta)\right\} h(x) \mathrm{d}x \\
&= \int T(x) p(x \mid \eta) \mathrm{d}x = E[T(X)].
\end{aligned}$$

The first term

- The first term of the ELBO:

$$\begin{aligned}\mathbb{E}_q[\log p(\boldsymbol{\theta} \mid \boldsymbol{\alpha})] &= \mathbb{E}_{q(\boldsymbol{\theta},\boldsymbol{z}|\boldsymbol{\gamma},\boldsymbol{\eta})}[\log p(\boldsymbol{\theta} \mid \boldsymbol{\alpha})] \\ &= \mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\gamma})}[\log p(\boldsymbol{\theta} \mid \boldsymbol{\alpha})] \\ &= \sum_{k-1}^{K} \mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\gamma})} \log(\theta_k).\end{aligned}$$

Dirchlet distribution is an exponential family, the density function is

$$p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) = \exp\left\{\left(\sum_{k=1}^{K}(\alpha_k - 1)\log\theta_k\right) + \log\Gamma\left(\sum_{l=1}^{K}\alpha_l\right) - \sum_{k=1}^{K}\log\Gamma\left(\alpha_k\right)\right\}.$$

Correspondingly, the natural parameter $\eta_k = \alpha_k - 1$, the sufficient statistic $T(\theta_k) = \log\theta_k$, and $A(\boldsymbol{\alpha}) = \sum_{k=1}^{K}\log\Gamma\left(\alpha_k\right) - \log\Gamma\left(\sum_{l=1}^{K}\alpha_l\right)$.

The first term

Thus,

$$E_{p(\boldsymbol{\theta}|\boldsymbol{\alpha})}\left[\log\theta_k\right] = \frac{\mathrm{d}}{\mathrm{d}\alpha_k}A(\alpha) = \frac{\mathrm{d}}{\mathrm{d}\alpha_k}\left[\sum_{k=1}^{K}\log\Gamma\left(\alpha_k\right) - \log\Gamma\left(\sum_{l=1}^{K}\alpha_l\right)\right]$$

$$= \Psi\left(\alpha_k\right) - \Psi\left(\sum_{l=1}^{K}\alpha_l\right), \quad k = 1, 2, \cdots, K,$$

where Ψ is digamma function, $\Psi\left(\alpha_k\right) = \frac{\mathrm{d}}{\mathrm{d}\alpha_k}\log\Gamma\left(\alpha_k\right)$.

The second term

- The second term

$$
\begin{aligned}
E_q(\log p(\boldsymbol{z} \mid \boldsymbol{\theta})) &= \sum_{n=1}^{N} E_q\left[\log p\left(z_n \mid \boldsymbol{\theta}\right)\right] \\
&= \sum_{n=1}^{N} E_{q(\boldsymbol{\theta}, z_n \mid \boldsymbol{\gamma}, \boldsymbol{\eta})}\left[\log\left(z_n \mid \boldsymbol{\theta}\right)\right] \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} q\left(z_{nk} \mid \boldsymbol{\eta}\right) E_{q(\boldsymbol{\theta} \mid \boldsymbol{\gamma})}\left[\log \theta_k\right] \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} \eta_{nk}\left[\Psi\left(\gamma_k\right) - \Psi\left(\sum_{l=1}^{K} \gamma_l\right)\right],
\end{aligned}
$$

where η_{nk} represents the probability of the n-th word of the document generated by the k-th topic, and γ_k represents the parameter of dircchlet distribution of the k-th topic.

The third term

- The third term

$$
\begin{aligned}
E_q[\log p(\boldsymbol{w} \mid \boldsymbol{z}, \varphi)] &= \sum_{n=1}^{N} E_q\left[\log p\left(w_n \mid z_n, \varphi\right)\right] \\
&= \sum_{n=1}^{N} E_{q(z_n|\eta)}\left[\log p\left(w_n \mid z_n, \varphi\right)\right] \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} q\left(z_{nk} \mid \eta\right) \log p\left(w_n \mid z_{nk}, \varphi\right) \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{v=1}^{V} \eta_{nk} w_n^v \log \varphi_{kv}.
\end{aligned}
$$

If the n-th word of the document is the v-th word of the word set then $w_n^v = 1$, otherwise $w_n^v = 0$. ϕ_{kv} represents the probability of the k-th topic of the document generate the v-th word of the word set.

The fourth term

- The fourth term

Similar to the first term,

$$E_q[\log q(\theta \mid \gamma)] = \log \Gamma\left(\sum_{l=1}^{K} \gamma_l\right) - \sum_{k=1}^{K} \log \Gamma\left(\gamma_k\right) + \sum_{k=1}^{K} \left(\gamma_k - 1\right)\left[\Psi\left(\gamma_k\right) - \Psi\left(\sum_{l=1}^{K} \gamma_l\right)\right].$$

The last term

- The last term

$$
\begin{aligned}
E_q[\log q(\boldsymbol{z} \mid \eta)] &= \sum_{n=1}^{N} E_q\left[\log q\left(z_n \mid \eta\right)\right] \\
&= \sum_{n=1}^{N} E_{q(z_n|\eta)}\left[\log q\left(z_n \mid \eta\right)\right] \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} q\left(z_{nk} \mid \eta\right) \log q\left(z_{nk} \mid \eta\right) \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} \eta_{nk} \log \eta_{nk}.
\end{aligned}
$$

Variational parameter estimation (E step)

Considering about maximizing ELBO $L(\boldsymbol{\gamma}, \boldsymbol{\eta}; \boldsymbol{\alpha}, \boldsymbol{\varphi})$ about η_{nk}, η_{nk} satisfies the constraint $\sum_{l=1}^{K} \eta_{nl} = 1$. The largrange function for constrained optimization problems involving η_{nk} is

$$\begin{aligned} L_{[\eta_{nk}]} =& \eta_{nk} \left[\Psi(\gamma_k) - \Psi\left(\sum_{l=1}^{K} \gamma_l \right) \right] \\ &+ \eta_{nk} \log \varphi_{kv} - \eta_{nk} \log \eta_{nk} + \lambda_n \left(\sum_{l=1}^{K} \eta_{nl} - 1 \right). \end{aligned}$$

Let the partial be 0, then the estimation of η_{nk} is

$$\eta_{nk} \propto \varphi_{kv} \exp\left(\Psi(\gamma_k) - \Psi\left(\sum_{l=1}^{K} \gamma_l \right) \right).$$

Variational parameter estimation (E step)

Next, considering about maximizing ELBO $L(\boldsymbol{\gamma}, \boldsymbol{\eta}; \boldsymbol{\alpha}, \boldsymbol{\varphi})$ about γ_k.

$$
\begin{aligned}
L_{[\gamma_k]} = & \sum_{k=1}^{K} (\alpha_k - 1) \left[\Psi(\gamma_k) - \Psi\left(\sum_{l=1}^{K} \gamma_l \right) \right] \\
& + \sum_{n=1}^{N} \sum_{k=1}^{K} \eta_{nk} \left[\Psi(\gamma_k) - \Psi\left(\sum_{l=1}^{K} \gamma_l \right) \right] \\
& - \log \Gamma\left(\sum_{l=1}^{K} \gamma_l \right) + \log \Gamma(\gamma_k) - \sum_{k=1}^{K} (\gamma_k - 1) \left[\Psi(\gamma_k) - \Psi\left(\sum_{l=1}^{K} \gamma_l \right) \right].
\end{aligned}
$$

Let the partial be 0, then the estimation of γ_k is

$$
\gamma_k = \alpha_k + \sum_{n=1}^{N} \eta_{nk}.
$$

Variational parameter estimation (E step)

(1) initialize $\eta_{nk}^{(0)} := 1/K$ for all i and n
(2) initialize $\gamma_K := \alpha_K + N/k$ for all k
(3) **repeat**
(4) for $n = 1$ to N
(5) for $k = 1$ to K
(6) $\eta_{nk}^{(t+1)} := \varphi_{kv} \exp\left[\Psi\left(\gamma_k^{(t)}\right) - \Psi\left(\sum_{l=}^{K} \gamma_l^{(t)}\right)\right]$
(7) normalize $\eta_{nk}^{(t+1)}$ to sum to 1 .
(8) $\gamma^{(t+1)} := \alpha + \sum_{n=1}^{N} \eta_n^{(t+1)}$
(9) **until** convergence

Model parameter estimation (M step)

Given a corpus of documents $D = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_M\}$, and the model parameter estimation is performed based on the corpus. To maximize with respect to $\boldsymbol{\varphi}$, we isolate terms and add Lagrange multipliers:

$$L_{[\boldsymbol{\varphi}]} = \sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{v=1}^{V}\eta_{mnk}w_{mn}^{v}\log\varphi_{kv} + \sum_{k=1}^{K}\lambda_k\left(\sum_{v=1}^{V}\varphi_{kv} - 1\right).$$

Let the partial be 0, then the estimation of φ_{kv} is

$$\varphi_{kv} = \sum_{m=1}^{M}\sum_{n=1}^{N}\eta_{mnk}w_{mn}^{v},$$

where η_{mnk} is the probability of the n-th word generated by the k-th topic in the m-th document. $w_{mn}^{v} = 1$ the n-th word is the v-th word in word set in the m-th document, otherwise $w_{mn}^{v} = 0$.

Model parameter estimation (M step)

Based on the corpus, to maximize with respect to $\boldsymbol{\alpha}$, the largrange function involving $\boldsymbol{\alpha}$ is

$$L_{[\alpha]} = \sum_{m=1}^{M} \left\{ \log \Gamma \left(\sum_{l=1}^{K} \alpha_l \right) - \sum_{k=1}^{K} \log \Gamma \left(\alpha_k \right) + \sum_{k=1}^{K} \left(\alpha_k - 1 \right) \left[\Psi \left(\gamma_{mk} \right) - \Psi \left(\sum_{l=1}^{K} \gamma_{ml} \right) \right] \right\}.$$

Taking the derivative with respect to α_k gives

$$\frac{\partial L}{\partial \alpha_k} = M \left[\Psi \left(\sum_{l=1}^{K} \alpha_l \right) - \Psi \left(\alpha_k \right) \right] + \sum_{m=1}^{M} \left[\Psi \left(\gamma_{mk} \right) - \Psi \left(\sum_{l=1}^{K} \gamma_{ml} \right) \right].$$

Model parameter estimation (M step)

This derivative depends on α_l, where $l \neq k$, and we therefore must use an iterative method to find the maximal α. In particular, the Hessian is

$$\frac{\partial^2 L}{\partial \alpha_k \partial \alpha_l} = M \left[\Psi' \left(\sum_{l=1}^{K} \alpha_l \right) - \delta(k,l) \Psi' (\alpha_k) \right],$$

and thus we can invoke the linear-time Newton-Raphson algorithm. The Newton-Raphson optimization technique finds a stationary point of a function by iterating

$$\boldsymbol{\alpha}_{\text{new}} = \boldsymbol{\alpha}_{\text{old}} - H(\boldsymbol{\alpha}_{\text{old}})^{-1} g(\boldsymbol{\alpha}_{\text{old}}).$$

Outline

Motivation

- Consider that we have a sequence of observations $x_{1:T} = (x_1, x_2, \cdots, x_T),\ T \geq 1$, where there is some natural order of the data.
- For each index $t = 1, \cdots, T$, we are interested in inferring some non-observed/hidden quantity of interest $z_t \in \mathcal{Z}$ where $\mathcal{Z}$ is a finite set.
- In the case of part-of-speech tagging (POST), observations $(x_1, x_2, \cdots, x_T)$ are the T words in a document. We are interested in inferring their corresponding tags $(z_1, z_2, \cdots, z_T)$.

PRON VB ADV ADJ PREP DET ADJ NOUN PREP DET ADJ COORD ADJ NOUN

Nothing is so painful to the human mind as a great and sudden change.

Robot Localisation

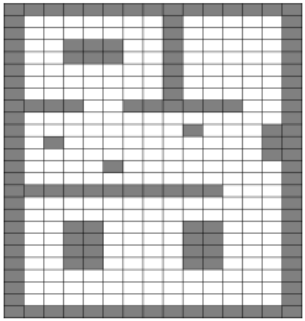


Figure: Robot localization: Hidden state Z_t is the position of the robot on the grid, and observation X_t is the (non-)detection of an obstacle. The objective is to calculate the probability in real time $P(Z_t = z_t | X_{1:t} = x_{1:t})$ that the robot is in a given cell.

Motivation

- By modelling the joint probability

$$\begin{aligned} p\left(z_{1:T}, x_{1:T}\right) :&= \mathbb{P}\left(Z_{1:T} = z_{1:T}, X_{1:t} = x_{1:t}\right) \\ &= \underbrace{\mathbb{P}\left(X_{1:T} = x_{1:T} \mid Z_{1:T} = z_{1:T}\right)}_{\text{Likelihood}} \underbrace{p\left(Z_{1:T}\right)}_{\text{Prior}}, \end{aligned}$$

  we can capture complex dependencies between the hidden states and the observations.
- Then we can calculate the posterior mode or maximum a posteriori (MAP) estimate

$$\hat{z}_{1:T} = \underset{z_{1:T} \in \mathcal{Z}^T}{\arg\max} \quad p\left(z_{1:T} \mid x_{1:T}\right)$$

- The search space has $|\mathcal{X}|^T$ elements and grows exponentially fast with T.

Motivation

- One way to simplify this model is to assume independence between the pairs $(Z_t, X_t), (Z_\tau, X_\tau)$ for any $t \neq \tau$. In this case, the MAP estimation reduces to solving independently

$$\hat{z}_t = \underset{z_t \in \mathcal{Z}}{\operatorname{argmax}} \, \mathbb{P}(X_t = x_t \mid Z_t = z_t) \, \mathbb{P}(Z_t = z_t), \quad t = 1, \ldots, T$$

  which has a linear complexity $T|\mathcal{X}|$.
- In the POST task, this independence is clearly inappropriate for *mind* and *change*.

Motivation

- A full model $p(z_{1:T}, x_{1:T})$ is practically useless as the estimate cannot be calculated.
- A much simpler statistical model which assumes independence across time allows to compute the estimate, but is too simplistic to address the task.
- Hidden Markov Model (HMM) offers a trade-off between the ability to capture dependencies and the tractability of the estimation algorithms.

Outline

Recap: Discrete Markov chain

- A process is called a Markov chain if for any $t \geq 0$ and any $z_0, \ldots, z_{t+1} \in \mathcal{Z}$,

$$\mathbb{P}\left(Z_{t+1} = z_{t+1} \mid Z_t = z_t, \ldots, Z_0 = z_0\right) = \mathbb{P}\left(Z_{t+1} = z_{t+1} \mid Z_t = z_t\right).$$

- The Markov chain is said to be homogeneous if $\mathbb{P}\left(Z_{t+1} = j \mid Z_t = i\right)$ does not depend on t, and we write

$$A_{i,j} := \mathbb{P}\left(X_{t+1} = j \mid X_t = i\right) \quad i, j \in \mathcal{X}.$$

- For $z_0 \in \mathcal{Z}$, denote $\mu_{z_0} = \mathbb{P}\left(Z_0 = z_0\right)$ the probability of the initial state Z_0.

Characterization

- State sequence $Z_{0:T}$ is a homogeneous Markov chain taking values in (discrete) state space $\mathcal{Z}$ with transition matrix (A_{ij}).
- Observation sequence $X_{1:T}$ taking values in (discrete) observation space $\mathcal{X}$.
- Conditional independent:

$$\mathbb{P}\left(X_1 = x_1, \ldots, X_T = x_T \mid Z_0 = z_0, \ldots, Z_T = z_T\right) = \prod_{t=1}^{T} \mathbb{P}\left(X_t = x_t \mid Z_t = z_t\right).$$

- Homogeneous HMM: $g_z(x) := \mathbb{P}\left(X_t = x \mid Z_t = z\right)$ is independent of t.

Characterization

- Joint probability of complete data:

$$\mathbb{P}\left(Z_{0:T} = z_{0:T}, X_{0:T} = x_{0:T}\right) = \mu_{z_0} \prod_{t=1}^{T} g_{z_t}\left(x_t\right) A_{z_{t-1}, z_t}.$$

- Graphical representation:

$$\begin{array}{ccccccccccc} Z_0 & \longrightarrow & Z_1 & \longrightarrow & Z_2 & \longrightarrow & \ldots & \longrightarrow & Z_{T-1} & \longrightarrow & Z_T \\ & & \downarrow & & \downarrow & & & & \downarrow & & \downarrow \\ & & X_1 & & X_2 & & \ldots & & X_{T-1} & & X_T \end{array}$$

Outline

Inference in HMM

- Assume parameters $\mu_{z_0}, A, g_z(x)$ and observation sequence $(x_1, \cdots, x_T)$ are known.
- Filtering:

$$p\left(z_t \mid x_{1:t}\right).$$

- Prediction:

$$p\left(z_t \mid x_{1:s}\right), \quad s < t.$$

- Smoothing:

$$p\left(z_t \mid x_{1:s}\right), \quad s > t.$$

- Observed likelihood:

$$p\left(x_{1:T}\right).$$

- Most likely state path:

$$\arg\max_{z_{0:T}} p\left(z_{0:T} \mid x_{1:T}\right).$$

Forward Filtering

- We are interested in the conditional probability $p(z_t \mid x_{1:t})$ given the observations up to time t.
- Normalization:

$$p(z_t \mid x_{1:t}) = \frac{p(z_t, x_{1:t})}{p(x_{1:t})} =: \frac{\alpha_t(z_t)}{\sum_z \alpha_t(z)}.$$

- Forward recursion:

$$\alpha_t(z_t) = p(x_t \mid z_t) \sum_{z_{t-1}} p(z_t \mid z_{t-1}) \alpha_{t-1}(z_{t-1}) \text{ with } \alpha_0(z_0) = p(z_0).$$

- Also, the likelihood $p\left(x_{1:T}\right)$ can be computed from the α-recursion

$$p\left(x_{1:T}\right) = \sum_{z \in \mathcal{Z}} \alpha_T(z).$$

Forward Filtering

- Consider $\mathcal{Z} = \{1, 2, \cdots, K\}$.
- Computation cost: $O(T|\mathcal{Z}|^2)$.
- Sometimes, we need to normalize α_t to avoid numerical underflow/overflow.

Algorithm 1 Forward α-recursion

- For $i = 1, \ldots, K$, set $\alpha_0(i) = \mu_i$
- For $t = 1, \ldots, T$
  - For $j = 1, \ldots, K$, set $\alpha_t(j) = g_j(x_t) \sum_{i=1}^{K} A_{i,j} \alpha_{t-1}(i)$

Forward-backward Smoothing

- We are now interested in the conditional probability $p(z_t \mid x_{1:T})$ given all the observations.
- Normalization:

$$p(z_t \mid x_{1:T}) = \frac{p(z_t, x_{1:T})}{p(x_{1:T})} = \frac{p(z_t, x_{1:t})p(x_{t+1:T} \mid z_t)}{p(x_{1:T})} =: \frac{\alpha_t(z_t)\beta_t(z_t)}{\sum_z \alpha_t(z)\beta_t(z)}.$$

- Backward recursion:

$$\beta_t(z_t) = \sum_{z_{t+1} \in \mathcal{Z}} \beta_{t+1}(z_{t+1})\, p(x_{t+1} \mid z_{t+1})\, p(z_{t+1} \mid z_t) \text{ with } \beta_T(z_T) = 1.$$

Forward-backward Smoothing

- Computation cost: $O(T|\mathcal{Z}|^2)$.

Algorithm 2 Backward β-recursion

- For $j = 1, \ldots, K$, set $\beta_T(j) = 1$
- For $t = 1, \ldots, T$
  - For $i = 1, \ldots, K$, set $\beta_{t-1}(i) = \sum_{j=1}^{K} g_j(x_t) A_{i,j} \beta_t(j)$

Maximum a Posterior Estimation

- We are interested in the maximum a posterior estimate

$$\hat{z}_{0:T} = \text{argmax}_{z_{0:T}} p(z_{0:T} \mid x_{1:T}) = \text{argmax}_{z_{0:T}} p(z_{0:T}, x_{1:T}).$$

- Direct optimization is unfeasible as the number of different state paths is $|\mathcal{Z}|^{T+1}$.
- The MAP estimate can be calculated efficiently using the Viterbi algorithm, which uses a backward-forward recursion.

Viterbi algorithm

$$
\begin{aligned}
&\max_{z_{0:T}} \; p\left(z_{0:T}, x_{1:T}\right) \\
=&\max_{z_{0:T}} p\left(z_0\right) \prod_{t=1}^{T} p\left(z_t \mid z_{t-1}\right) p\left(x_t \mid z_t\right) \\
=&\max_{z_{0:T-1}} \left\{ \left[p\left(z_0\right) \prod_{t=1}^{T-1} p\left(z_t \mid z_{t-1}\right) p\left(x_t \mid z_t\right) \right] \max_{z_T} p\left(z_T \mid z_{T-1}\right) p\left(x_T \mid z_T\right) \right\} \\
=&\max_{z_{0:T-1}} \left\{ \left[p\left(z_0\right) \prod_{t=1}^{T-1} p\left(z_t \mid z_{t-1}\right) p\left(x_t \mid z_t\right) \right] m_{T-1}\left(z_{T-1}\right) \right\}.
\end{aligned}
$$

Viterbi algorithm

- For $t = T-1, \ldots, 1$, let $m_T(z_T) = 1$ and

$$m_{t-1}(z_{t-1}) = \max_{z_{t:T}} \left\{ \prod_{k=t}^{T} p(z_k \mid z_{k-1}) \, p(x_k \mid z_k) \right\}.$$

- Backward recursion: for $t = T-1, \ldots, 1$

$$m_{t-1}(z_{t-1}) = \max_{z_t} p(x_t \mid z_t) \, p(z_t \mid z_{t-1}) \, m_t(z_t).$$

- Actually, we combine the prior (given z_{t-1}) and likelihood (given $x_{t:T}$) to compute the most probable tail path $z_{t:T}$.

Viterbi algorithm

- After back recursion, we have $p(z_0)m_0(z_0) = \max_{z_{1:T}} p(z_{0:T}, x_{1:T})$.
- Then,

$$\hat{z}_0 = \arg\max_{z_0} \left(\max_{z_{1:T}} p\left(z_0, z_{1:T}, x_{1:T}\right) \right) = \arg\max_{z_0} m_0\left(z_0\right) p\left(z_0\right).$$

- Forward recursion,

$$\hat{z}_t = \arg\max_{z_t} \left(m_t\left(z_t\right) p\left(x_t \mid z_t\right) p\left(z_t \mid \hat{z}_{t-1}\right) \right)$$

Viterbi algorithm

- First, perform a backward path which computes messages $m_t, t = T, \cdots, 0$.
- Then, perform a forward path to return the estimates $\widehat{z}_t, t = 0, \cdots, T$.
- Profiling procedure:

$$
\begin{array}{lccccccccc}
\text{Backward} & m_0(z_0) & \longleftarrow & m_1(z_1) & \longleftarrow & \cdots & \longleftarrow & m_{T-1}(z_{T-1}) & \longleftarrow & m_T(z_T) \\
 & \downarrow & & \downarrow & & & & \downarrow & & \downarrow \\
\text{Forward} & \widehat{z}_0 & \longrightarrow & \widehat{z}_1 & \longrightarrow & \cdots & \longrightarrow & \widehat{z}_{T-1} & \longrightarrow & \widehat{z}_T
\end{array}
$$

Outline

Learning in HMM

- We have considered several inference problems assuming the parameters A, μ, and g of the HMM are known.
- We are now interested in learning these parameters in two cases
  - The fully observed case: we have a dataset where the hidden states $(z_0, z_1 \cdots, z_T)$ are known.
  - The unsupervised case: all we have is the data $(x_1 \cdots, x_T)$ and the hidden states are not observed.
- Here, we only consider the estimation of transition matrix A.

Learning in HMM

- If the hidden states $(z_0, z_1, \cdots, z_T)$ are known, A can be fitted using maximum likelihood.
- Let $n_{i,j} = \sum_{t=1}^{T} I\left(z_t = j, z_{t-1} = i\right)$ be the number of transitions between state i and state j.
- The MLE of $A_{i,j}$ is

$$\widehat{A}_{i,j} = \frac{n_{i,j}}{\sum_{\ell \in \mathcal{Z}} n_{i,\ell}}.$$

Learning in HMM

- If the hidden states are unknown, finding the MLE is much more challenging as we want to optimize

$$\widehat{A} = \arg\max_{A} \log p(x_{1:T}; A).$$

- The Baum-Welch algorithm, a special case of the Expectation-Maximization algorithm, can be used to find the MLE efficiently.
  - E step

$$Q\left(A; A^{(k-1)}\right) = \mathbb{E}\left[\log p\left(Z_{0:T}, x_{1:T}; A\right) \mid x_{1:T}, A^{(k-1)}\right].$$

  - M step

$$A^{(k)} = \arg\max_{A} Q\left(A; A^{(k-1)}\right).$$

Learning in HMM

- E-step (calculate the Q function):
  - prior:

$$p\left(z_{0:T}; A\right) = \mu_{z_0} \prod_{i,j \in \mathcal{Z}} A_{i,j}^{n_{i,j}}.$$

  - log joint probability:

$$\log p\left(z_{0:T}, x_{1:T}; A\right) = \log \mu_{z_0} + \sum_{i,j \in \mathcal{Z}} n_{i,j} \log A_{i,j} + \sum_{t=1}^{T} \log g_{z_t}\left(x_t\right).$$

  - Q function:

$$\begin{aligned} Q\left(A; A^*\right) &= \mathbb{E}\left[\log p\left(Z_{0:T}, x_{1:T}; A\right) \mid x_{1:T}, A^*\right] \\ &= \sum_{i,j \in \mathcal{Z}} \mathbb{E}\left[N_{i,j} \mid x_{1:T}, A^*\right] \log A_{i,j} + C. \end{aligned}$$

    where $N_{i,j} = \sum_{t=1}^{T} I\left(Z_t = j, Z_{t-1} = i\right)$.

Learning in HMM

- E-step (calculate the Q function):
  - expected counts

$$\mathbb{E}\left[N_{i,j} \mid x_{1:T}, A^*\right] = \sum_{t=1}^{T} \mathbb{P}\left(Z_t = j, Z_{t-1} = i \mid x_{1:T}; A^*\right).$$

  - the terms $p\left(z_t, z_{t-1} \mid x_{1:T}; A^*\right)$ can be obtained from the forward-backward recursion

$$p\left(z_t, z_{t-1} \mid x_{1:T}; A^*\right) \propto \alpha_{t-1}\left(z_{t-1}\right) p\left(x_t \mid z_t\right) p\left(z_t \mid z_{t-1}\right) \beta_t\left(z_t\right).$$

- M-step:

$$\begin{aligned} A_{i,j}^{(k)} &= \arg\max_{A_{i,j}} \mathbb{E}\left[N_{i,j} \mid x_{1:T}, A^{(k-1)}\right] \log A_{i,j} \\ &= \frac{\mathbb{E}\left[N_{i,j} \mid x_{1:T}, A^{(k-1)}\right]}{\sum_{\ell} \mathbb{E}\left[N_{i,\ell} \mid x_{1:T}, A^{(k-1)}\right]}. \end{aligned}$$

Outline

Roots of Factor Analysis in Causal Discovery

- The roots of factor analysis go back to work by Charles Spearman just over a century ago (Spearman, 1904).
- He was trying to discover the hidden structure of human intelligence.
- His observation was that school children's grades in different subjects were all correlated with each other.
- He explained as follows: the reason grades are all correlated is that performance in these subjects is correlated with something else, a general or common factor, which he named "general intelligence".

What & Why

- What is Factor Analysis?
  - Factor analysis is a theory driven statistical data reduction technique used to explain covariance among observed random variables in terms of fewer unobserved random variables named factors.
- Why Factor Analysis?
  - Testing of theory: Explain covariation among multiple observed variables by mapping variables to latent constructs (called "factors").
  - Understanding the structure underlying a set of measures: Gain insight to dimensions.

Example

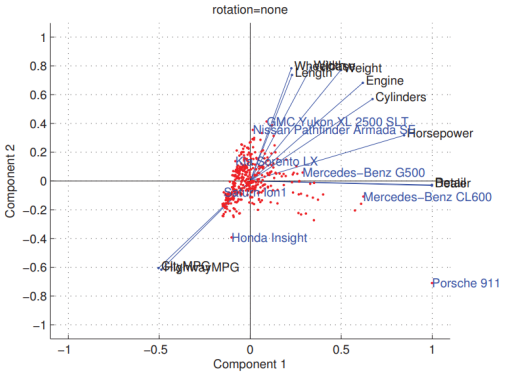


Figure: 2D projection of 2004 cars data based on factor analysis.

Assumption

- We assume that the data point $\boldsymbol{x}_i$ (observed variable) is obtained by linear projection of the p-dimensional $\boldsymbol{z}_i$ (latent variable) onto a d-dimensional space by projection matrix $\boldsymbol{\Lambda} \in \mathbb{R}^{d \times p}$, then applying some linear translation, and finally adding a Gaussian noise $\boldsymbol{\epsilon} \in \mathbb{R}^d$ with covariance matrix $\boldsymbol{\Psi} \in \mathbb{R}^{d \times d}$.
- Note that as the noises in different dimensions are independent, the covariance matrix $\boldsymbol{\Psi}$ is diagonal.

Outline

Factor Analysis Model

- The prior distribution of the latent variable is Gaussian

$$\mathbb{P}(\boldsymbol{z}_i) = \mathcal{N}(\boldsymbol{z}_i \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0),$$

where $\boldsymbol{\mu}_0 \in \mathbb{R}^p$ and $\boldsymbol{\Sigma}_0 \in \mathbb{R}^{p \times p}$ are the mean and the covariance matrix of $\boldsymbol{z}_i$.

- Since the noise $\boldsymbol{\epsilon}$ is Gaussian, the data point $\boldsymbol{x}_i$ has a conditional Gaussian distribution given the latent variable,

$$\mathbb{P}(\boldsymbol{x}_i \mid \boldsymbol{z}_i) = \mathcal{N}(\boldsymbol{\Lambda}\boldsymbol{z}_i + \boldsymbol{\mu}, \boldsymbol{\Psi}),$$

where $\boldsymbol{\mu}$, which is the translation vector, is the mean of data $\{\boldsymbol{x}_i\}_{i=1}^n$

$$\boldsymbol{\mu} := \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \in \mathbb{R}^d.$$

Factor Analysis Model

- The marginal distribution of $\boldsymbol{x}_i$ is:

$$
\begin{aligned}
&\mathbb{P}\left(\boldsymbol{x}_i \mid \boldsymbol{\Lambda}, \boldsymbol{\mu}, \boldsymbol{\Psi}\right) \\
&= \int \mathbb{P}\left(\boldsymbol{x}_i \mid \boldsymbol{z}_i, \boldsymbol{\Lambda}, \boldsymbol{\mu}, \boldsymbol{\Psi}\right) \mathbb{P}\left(\boldsymbol{z}_i \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right) d\boldsymbol{z}_i \\
&= \mathcal{N}\left(\boldsymbol{\Lambda}\boldsymbol{\mu}_0 + \boldsymbol{\mu}, \boldsymbol{\Psi} + \boldsymbol{\Lambda}\boldsymbol{\Sigma}_0\boldsymbol{\Lambda}^\top\right) \\
&= \mathcal{N}\left(\widehat{\boldsymbol{\mu}}, \boldsymbol{\Psi} + \widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{\Lambda}}^\top\right)
\end{aligned}
$$

where $\widehat{\boldsymbol{\mu}} := \boldsymbol{\Lambda}\boldsymbol{\mu}_0 + \boldsymbol{\mu} \in \mathbb{R}^d$, $\widehat{\boldsymbol{\Lambda}} := \boldsymbol{\Lambda}\boldsymbol{\Sigma}_0^{\frac{1}{2}} \in \mathbb{R}^{d \times d}$.

Factor Analysis Model

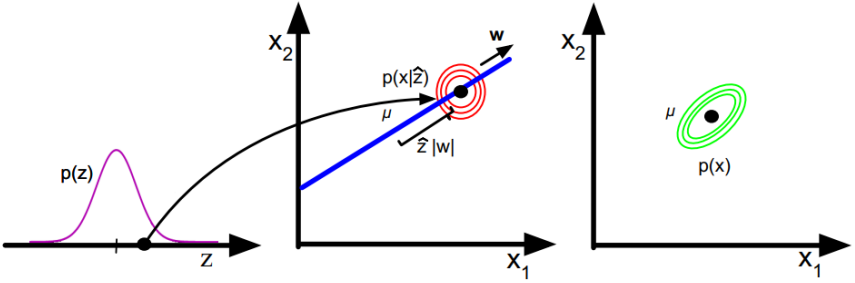
- As the mean $\widehat{\boldsymbol{\mu}}$ and covariance $\widehat{\boldsymbol{\Lambda}}$ are needed to be learned, we can absorb $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ into $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$ and assume that $\boldsymbol{\mu}_0 = \mathbf{0}$ and $\boldsymbol{\Sigma}_0 = \boldsymbol{I}$.
- Thus, the factor analysis model can be simplied as follows:

$$\begin{aligned} \boldsymbol{x}_i &:= \boldsymbol{\Lambda}\boldsymbol{z}_i + \boldsymbol{\mu} + \boldsymbol{\epsilon} \\ \mathbb{P}\left(\boldsymbol{z}_i\right) &= \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \\ \mathbb{P}(\boldsymbol{\epsilon}) &= \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}) \end{aligned} \quad .$$

- data point = linear function of latent variable plus Gaussian noise.

Factor Analysis Model

- data point = linear function of latent variable plus Gaussian noise.



Unidentifiability

- FA is unidentifiable.
- To see this, suppose $\mathbf{R}$ is an arbitrary orthogonal rotation matrix, satisfying $\mathbf{R}\mathbf{R}^T = \mathbf{I}$. Let us define $\tilde{\mathbf{\Lambda}} = \mathbf{\Lambda}\mathbf{R}$, then

$$\begin{aligned}\mathrm{cov}[\mathbf{x}] &= \tilde{\mathbf{\Lambda}}\mathbb{E}\left[\mathbf{z}\mathbf{z}^T\right]\tilde{\mathbf{\Lambda}}^T + \mathbb{E}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\right] \\ &= \mathbf{\Lambda}\mathbf{R}\mathbf{R}^T\mathbf{\Lambda}^T + \mathbf{\Psi} = \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi}\end{aligned}$$

Unidentifiability

- To ensure a unique solution, we need to remove $d \times (d-1)$ degrees of freedom.
  - Forcing $\mathbf{\Lambda}$ to be orthonormal.
  - Forcing $\mathbf{\Lambda}$ to be lower triangular.
  - Sparsity promoting priors on the weights.
  - Choosing an informative rotation matrix.
  - Use of non-Gaussian priors for the latent factors.
- While $\mathbf{\Lambda}$ is unidentifiable, the column space of which is uniquely determined.

Joint and Marginal Distributions in Factor Analysis

- By the property of Gaussian distribution and the corresponding moment, we can derive the joint distribution of complete data:

$$\begin{bmatrix} \boldsymbol{x}_i \\ \boldsymbol{z}_i \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi} & \boldsymbol{\Lambda} \\ \boldsymbol{\Lambda}^\top & \boldsymbol{I} \end{bmatrix} \right).$$

- The marginal distribution of data point $\boldsymbol{x}_i$ is:

$$\mathbb{P}\left(\boldsymbol{x}_i\right) = \mathbb{P}\left(\boldsymbol{x}_i \mid \boldsymbol{\Lambda}, \boldsymbol{\mu}, \boldsymbol{\Psi}\right) = \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi}\right).$$

- The posterior distribution of latent variable given data is:

$$q\left(\boldsymbol{z}_i\right) = \mathbb{P}\left(\boldsymbol{z}_i \mid \boldsymbol{x}_i\right) = \mathcal{N}\left(\boldsymbol{\mu}_{z|x}, \boldsymbol{\Sigma}_{z|x}\right),$$

where

$$\boldsymbol{\mu}_{z|x} := \boldsymbol{\Lambda}^\top \left(\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi}\right)^{-1} \left(\boldsymbol{x}_i - \boldsymbol{\mu}\right), \qquad \boldsymbol{\Sigma}_{z|x} := \boldsymbol{I} - \boldsymbol{\Lambda}^\top \left(\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi}\right)^{-1} \boldsymbol{\Lambda}.$$

Joint and Marginal Distributions in Factor Analysis

- If data $\{\boldsymbol{x}_i\}_{i=1}^n$ are centered, i.e. $\boldsymbol{\mu} = \mathbf{0}$, then we have:

$$\mathbb{P}(\boldsymbol{x}_i \mid \boldsymbol{\Lambda}, \boldsymbol{\Psi}) = \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Psi} + \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top\right)$$
$$\mathbb{P}(\boldsymbol{x}_i \mid \boldsymbol{z}_i, \boldsymbol{\Lambda}, \boldsymbol{\Psi}) = \mathcal{N}(\boldsymbol{\Lambda}\boldsymbol{z}_i, \boldsymbol{\Psi}).$$

- In some works, people center the data as a pre-processing to factor analysis.

Expectation Maximization for Factor Analysis

- We have already computed $\mathbb{P}(\boldsymbol{x}_i \mid \boldsymbol{\theta}) = \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi}\right)$, the optimal $\boldsymbol{\mu}$ is the sample mean $\frac{1}{n}\sum_{i=1}^n \boldsymbol{x_i}$.
- Assume the mean is known from now on, and we only need to estimate $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$.
- E-Step:

$$q^{(t)}(\boldsymbol{z}_i) \leftarrow \mathbb{P}(\boldsymbol{z}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(t-1)}) = \mathcal{N}\left(\boldsymbol{\mu}_{z|x}^{(t-1)}, \boldsymbol{\Sigma}_{z|x}^{(t-1)}\right).$$

- M-Step:

$$\boldsymbol{\Lambda}^{(t)} \leftarrow \arg\max_{\boldsymbol{\Lambda}} \sum_{i=1}^n \mathbb{E}_{\sim q^{(t)}(\boldsymbol{z}_i)}\left[\log \mathbb{P}(\boldsymbol{x}_i, \boldsymbol{z}_i \mid \boldsymbol{\Lambda}, \boldsymbol{\Psi})\right]$$

$$\boldsymbol{\Psi}^{(t)} \leftarrow \arg\max_{\boldsymbol{\Psi}} \sum_{i=1}^n \mathbb{E}_{\sim q^{(t)}(\boldsymbol{z}_i)}\left[\log \mathbb{P}(\boldsymbol{x}_i, \boldsymbol{z}_i \mid \boldsymbol{\Lambda}, \boldsymbol{\Psi})\right].$$

Outline

Covariance Learning

- Sparse covariance matrix assumption is not reasonable in many applications.
  - Financial returns depend on common equity market risks.
  - Housing prices depend on economic health and locality.
- The covariance matrix is dense due to the presence of the common factor.
- A natural extension is that the covariance matrix is sparse conditioning on the common factors.

Spiked Incoherent Low-Rank Models

- Modern high-dimensional factor models intends to decompose a large matrix $\mathbf{\Sigma}$ as follows:

$$\mathbf{\Sigma} = \mathbf{L} + \mathbf{S}.$$

- The decomposition has the following properties:
  - **Low-rank** (for $\mathbf{L}$): achieves dimension reductions.
  - **Spikedness**: helps seperate $\mathbf{L}$ from $\mathbf{S}$.
  - **Incoherence** (Pervasiveness): excludes matrices being low-rank and sparse simultaneously.

Factor Model

- Suppose that we have p-dimensional vector $\mathbf{X}$ of measurements, whose dependence is driven by K factors $\mathbf{f}$.

$$\mathbf{X} = \mathbf{a} + \mathbf{B}\mathbf{f} + \mathbf{u}, \quad \mathrm{E}\mathbf{u} = 0, \quad \mathrm{cov}(\mathbf{f}, \mathbf{u}) = 0,$$

  where $\mathbf{a}$ is an intercept term, $\mathbf{u}$ is the idiosyncratic component that is uncorrelated with the common factors $\mathbf{f}, \mathbf{B} = (\mathbf{b}_1, \cdots, \mathbf{b}_p)^T$ are the $p \times K$ factor loading matrix.
- The factor model induces a covariance structure for $\mathbf{\Sigma} = \mathrm{Var}(\mathbf{X})$:

$$\mathbf{\Sigma} = \mathbf{B}\mathbf{\Sigma}_f\mathbf{B}^T + \mathbf{\Sigma}_u, \quad \mathbf{\Sigma}_f = \mathrm{Var}(\mathbf{f}), \quad \mathbf{\Sigma}_u = \mathrm{Var}(\mathbf{u}).$$

- It is frequently assumed that $\mathbf{\Sigma}_u$ is sparse, since common dependence has already been taken out, i.e., $\mathbf{\Sigma}$ is conditionally sparse.

Identifiability Condition

- Since $\mathbf{X} = \mathbf{a} + \mathbf{B}\mathbf{f} + \mathbf{u} = \mathbf{a} + (\mathbf{B}\mathbf{H})(\mathbf{H}^{-1}\mathbf{f}) + \mathbf{u}$, we choose $\mathbf{H}$ so that the columns of $(\mathbf{B}\mathbf{H})$ are orthogonal and $\mathrm{Var}(\mathbf{H}^{-1}\mathbf{f}) = \mathbf{I}_p$.
- Since $\mathbb{E}\mathbf{X} = \mathbf{a} + \mathbf{B}\mathbb{E}\mathbf{f}$, we assume that $\mathbb{E}\mathbf{f} = 0$.
- Identifiability Condition: $\mathbf{B}^T\mathbf{B}$ is diagonal, $\mathbb{E}\mathbf{f} = 0$ and $\mathrm{cov}(\mathbf{f}) = \mathbf{I}_p$.

Covariance Decomposition

- Under the identifiability condition, we have the covariance structure

$$\mathbf{\Sigma} = \mathbf{B}\mathbf{B}^T + \mathbf{\Sigma}_u.$$

- This matrix admits a low-rank + sparse structure, with the first matrix being of rank K.
- This leads naturally to the following penalized least-squares problem: minimize with respect to $\mathbf{\Theta}$ and $\mathbf{\Gamma}$

$$\|\mathbf{S} - \mathbf{\Theta} - \mathbf{\Gamma}\|_F^2 + \lambda\|\mathbf{\Theta}\|_* + \lambda\nu \sum_{i \neq j} |\gamma_{ij}|$$

where $\mathbf{S}$ is the sample covariance matrix and $\|\mathbf{\Theta}\|_*$ is the nuclear norm (summation of the singular values of which) that encourages the low-rankness and the second penalty encourages sparseness of $\mathbf{\Gamma}$.

Extract Latent Factors

- Fan et al. (2013) proposed the Principal Orthogonal ComplEment Thresholding (POET) estimator.
- First, obtain an estimator $\widehat{\mu}$ and $\widehat{\mathbf{\Sigma}}$, e.g., the sample mean and covariance matrix and compute the eigen-decomposition:

$$\widehat{\mathbf{\Sigma}} = \sum_{j=1}^{p} \widehat{\lambda}_j \widehat{\boldsymbol{\xi}}_j \widehat{\boldsymbol{\xi}}_j^T$$

- Then, keep the top K eigenvalues and their corresponding eigenvectors and apply the thresholding procedure to the remaining.

Extract Latent Factors

- Thus obtaining the estimators for the factor loading matrix and the latent factors

$$\widehat{\mathbf{B}} = \left(\widehat{\lambda}_1^{1/2}\widehat{\boldsymbol{\xi}}_1, \cdots, \widehat{\lambda}_K^{1/2}\widehat{\boldsymbol{\xi}}_K\right), \quad \widehat{\mathbf{f}}_i = \operatorname{diag}\left(\widehat{\lambda}_1, \cdots, \widehat{\lambda}_K\right)^{-1} \widehat{\mathbf{B}}^T \left(\mathbf{X}_i - \widehat{\boldsymbol{\mu}}\right).$$

- Obtain $\widehat{\boldsymbol{\Sigma}}_u = \sum_{j=K+1}^{p} \widehat{\lambda}_j \widehat{\boldsymbol{\xi}}_j \hat{\boldsymbol{\xi}}_j^T$ and its regularized estimator $\widehat{\boldsymbol{\Sigma}}_{u,\lambda}^{\tau}$ by using the thresholding estimator.
- Estimate the covariance matrix $\boldsymbol{\Sigma} = \operatorname{Var}(\mathbf{X})$ by

$$\widehat{\boldsymbol{\Sigma}}_\lambda = \widehat{\mathbf{B}}\widehat{\mathbf{B}}^T + \widehat{\boldsymbol{\Sigma}}_{u,\lambda}^{\tau} = \sum_{j=1}^{K} \widehat{\lambda}_j \widehat{\boldsymbol{\xi}}_j \widehat{\boldsymbol{\xi}}_j^T + \widehat{\boldsymbol{\Sigma}}_{u,\lambda}^{\tau}$$

Select Number of Factors

- Three more recent methods to choose the number of factors K.
- For a pre-determined parameter $k_{\max}$, the **eigenvalue ratio estimator** is

$$\widehat{K}_1 = \underset{j \le k_{\max}}{\operatorname{argmax}} \frac{\lambda_j(\widehat{\Sigma})}{\lambda_{j+1}(\widehat{\Sigma})}.$$

- For a given $\delta > 0$ and pre-determined integer $k_{\max}$, the **eigenvalues difference estimator** is

$$\widehat{K}_2(\delta) = \max\left\{ j \le k_{\max} : \lambda_j(\widehat{\mathbf{\Sigma}}) - \lambda_{j+1}(\widehat{\mathbf{\Sigma}}) \ge \delta \right\}.$$

Select Number of Factors

- Define

$$V(k) = \frac{1}{np} \min_{\mathbf{B}\in\mathbb{R}^{p\times k},\mathbf{F}\in\mathbb{R}^{n\times k}} \left\| \mathbf{X} - \mathbf{1}_n \overline{\mathbf{X}}^T - \mathbf{F}\mathbf{B}^T \right\|_F^2 = p^{-1} \sum_{j>k} \lambda_j(\mathbf{S}),$$

- The **information criterion estimator** $\widehat{K}_3$ is to find the best $k \le k_{\max}$ such that the following penalized version of $V(k)$ is minimized:

$$PC(k) = V(k) + k\widehat{\sigma}^2 g(n,p),$$

where $g(n,p)$ is chosen and $\widehat{\sigma}^2$ is any consistent estimate of $(np)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{d} \mathrm{E}u_{ji}^2$.

References I