Mixture Models and Hidden Variable Graphical Models

P. Restadh

April 23, 2021

KTH Royal Institute of Technology

Outline

Teeth - Better/cheaper dental care gives nicer teeth.
Scandinavia and Canada - Wealthy countries, good healthcare
Hockey - Expensive sport. Played more in Scandinavia and Canada.
So why does hockey players look like this?

Image removed because of copyright reasons. Google "hockey player missing teeth".

In fact there is a hidden subset of the population "People who are punched in the face a lot'.

Let $\mathcal{M} \subseteq \Delta_{r-1}$ be our statistical model. Each $\mathbf{p}^1, \ldots, \mathbf{p}^k \in \mathcal{M}$ and $\pi \in \Delta_{k-1}$.

$$\mathbb{P}(H = i) = \pi_i \qquad \text{and} \qquad \mathbb{P}(X \mid H = i) = \mathbf{p}^i$$

So we have

$$\mathbb{P}(X = j) = \sum \pi_i p_j^i$$

Definition (14.1.1)

The k**:th mixture model** of the model $\mathcal{M} \subseteq \Delta_{r-1}$ is the family of probability distributions

$$\mathrm{Mixt}^k(\mathcal{M}) := \left\{\pi_1 \mathbf{p}^1 + \cdots + \pi_k \mathbf{p}^k : \pi \in \Delta_{k-1}, \mathbf{p}^1, \ldots, \mathbf{p}^k \in \mathcal{M}\right\}.$$

Example

Let H be binary "infection status" (we can't observe because we don't have a test). X is symptoms, follows some distribution in $\mathcal{M}$. Then the distribution of X can be observed in $\mathrm{Mixt}^2(\mathcal{M})$.

Example

Consider $\mathcal{M}(X_1 \perp\!\!\!\perp X_2 \perp\!\!\!\perp X_3)$. Can be parameterized as $p_{ijn} = \alpha_i \beta_j \gamma_n$.

$$\mathrm{Mixt}^k(\mathcal{M}) = \left\{ p_{ijn} = \sum_{p=1}^{k} \pi_p \alpha_{pi} \beta_{pj} \gamma_{pn} \colon \pi \in \Delta_{k-1} \right\}$$

What if we want $\mathbf{p}^i$ from different models?

Definition (14.1.4)

Let $\mathcal{M}_1, \ldots, \mathcal{M}_k \subseteq \Delta_{r-1}$ be k statistical models. The **mixture model** $\mathcal{M}_1 * \cdots * \mathcal{M}_k$ is the family of probability distributions

$$\mathcal{M}_1 * \cdots * \mathcal{M}_k \coloneqq \left\{ \pi_1 \mathbf{p}^1 + \cdots + \pi_k \mathbf{p}^k \colon \pi \in \Delta_{k-1}, \forall i \ \mathbf{p}^i \in \mathcal{M}_i \right\}.$$

The algebraic side

Definition (14.1.4)

Let $V_1, \ldots, V_k \subseteq \mathbb{K}^r$ be algebraic varieties. The **join variety** is the variety

$$V_1 * \cdots * V_k := \overline{\left\{ \pi_1 \mathbf{p}^1 + \cdots + \pi_k \mathbf{p}^k \colon \sum_i \pi_i = 1 \text{ and } \mathbf{p}^i \in V_i \text{ for all } i \right\}}.$$

We also define $\mathrm{Sec}^k(V) = V * \cdots * V$ (k times).

Proposition

We have

$$\mathrm{Mixt}^k(\mathcal{M}) \subseteq \mathrm{Sec}^k(\overline{\mathcal{M}}) \cap \Delta_{\mathcal{R}}$$

Example

In the case of $k = 1$ we get

$$\mathcal{M} = \mathrm{Mixt}^1(\mathcal{M}) \subseteq \mathrm{Sec}^1(\overline{\mathcal{M}}) \cap \Delta_{\mathcal{R}} = \overline{\mathcal{M}} \cap \Delta_{\mathcal{R}}$$

Recall $\text{rank}_+(M) = \min\{k\colon M = \sum_{i=1}^{k} M_i\}$ where each M_i is non-negative with rank 1. In general $\text{rank}_+(M) \geq \text{rank}(M)$.

Example (14.1.6)

As before we get

$$\text{Mixt}^k\left(\mathcal{M}_{X_1 \perp\!\!\!\perp X_2}\right) = \{P \in \Delta_{\mathcal{R}}\colon\ \text{rank}_+(P) \leq k\}\,.$$

Slightly more work

$$\text{Sec}^k\left(\overline{\mathcal{M}_{X_1 \perp\!\!\!\perp X_2}}\right) \cap \Delta_{\mathcal{R}} = \{P \in \Delta_{\mathcal{R}}\colon\ \text{rank}(P) \leq k\}\,.$$

Recommended: Proposition 14.1.8

Definition

A finite sequence $X_1, \ldots, X_n$ of R.V. on a countable state space is **exchangeable** if for all $\sigma \in S_n$ and $a_1, \ldots, a_n \in \mathcal{X}$,

$$\mathbb{P}\left(X_1 = a_1, \ldots, X_n = a_n\right) = \mathbb{P}\left(X_1 = a_{\sigma(1)}, \ldots, X_n = a_{\sigma(n)}\right)$$

An infinite sequence (X_i) of R.V. is exchangeable if each of its finite subsequences is exchangeable.

Formal way of saying "only the output matters, not the order".
Strictly weaker version of i.i.d.

Theorem (De Finetti)

Let $X_1, X_2, \dots$ be an infinite sequence of exchangeable random variables with state space $\{0,1\}$. Then there exists a unique probability measure μ on $[0,1]$ such that for all n and $a_1, \dots, a_n \in \{0,1\}$

$$\mathbb{P}(X_1 = a_1, \dots, X_n = a_n) = \int_0^1 \theta^k (1-\theta)^{n-k} d\mu(\theta),$$

where $k = \sum_{i=1}^{k} a_i$.

Example

Let $X_1, X_2, \dots$ be i.i.d $\sim \mathrm{Ber}(p)$. Then $\mu = \delta_p$.

> *Another way to say De Finettis's theorem is that an infinite exchangeable sequence of binary random variables is a mixture of i.i.d. Bernoulli random variables.*

Idea of proof.

Let us consider $X_1, \ldots, X_k$ exchangeable binary R.V. The distributions that parameterize k i.i.d Bernoulli are on the form $\mathcal{C}_k = \left\{ \left(\theta^{a_1+\cdots+a_k}(1-\theta)^{k-a_1-\cdots-a_k} \right) \right\}$ where we range the a_i's over the outcomes and θ is the parameter in $\mathrm{Ber}(\theta)$.
Let $EX_n \subseteq \Delta_{2^n-1}$ be the set of n exchangeable sequences of binary random variables. Conclude that "exchangeable" is a set of linear restrictions on EX_n, thus this is a polytope. Let $\pi_{n,k}$ be the action of computing the margin of all but k of the variables. Then $\pi_{n,k}(EX_n) \subseteq EX_k$ and as each permutation in S_n can be seen as a permutatoin in S_{n+1} we have $\pi_{n,k}(EX_n) \supseteq \pi_{n+1,k}(EX_{n+1})$.
Then the proof consists of showing

$$\lim_{n \to \inf} \pi_{n,k}(EX_n) = \mathrm{Mixt}^k(\mathcal{C}_k).$$

$\square$

Example (14.2.1)

Consider the claw tree, fig. 1. This can be parameterized as $p_{i_1 i_2 i_3 i_4} = \pi_{i_1}\alpha_{i_1 i_2}\beta_{i_1 i_3}\gamma_{i_1 i_2}$.
Thus $p_{i_2 i_3 i_4} = \sum_{i_1=1}^{r_1} \pi_{i_1}\alpha_{i_1 i_2}\beta_{i_1 i_3}\gamma_{i_1 i_2}$. This we recognize as $\mathrm{Mixt}^{r_1}(\mathcal{M}_{X_1 \perp\!\!\!\perp X_2 \perp\!\!\!\perp X_3})$.

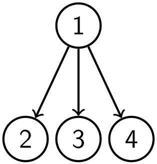

Figure 1: A claw tree.

Proposition (14.2.2)

Let $\mathcal{M} \subseteq \mathbb{R}^m \times \mathrm{PD}_m$ be an algebraic exponential family with vanishing ideal $I = I(\mathcal{M}) \subseteq \mathbb{R}[\mu, \Sigma]$. Let $H \sqcup O = [m]$ be a partition of the index labeling into hidden variables H and observed variables O. The hidden variable model consists of all marginal distributions on the variables X_O for a distribution with parameters in $\mathcal{M}$. The vanishing ideal of the hidden variable model is the elimination ideal

$$I \cap \mathbb{R}\left[\mu_O, \Sigma_{O,O}\right].$$

Finding generators is relatively easy in the Gaussian case[1].

[1] Given appropriate assumptions, see Proposition 14.2.5.

Remark

From here everything is Gaussian.

Example

Recall *parametrized Gaussian directed graphical models*. $G = (V, D)$ is DAG, each edge $v \to u$ was given weight λ_{vu}.

$$X_v = \varepsilon_v + \sum_{u \in \mathrm{pa}_G(v)} \lambda_{uv} X_u$$

where ε were independent ($\sim \mathcal{N}(0, \Omega)$, Ω diagonal). Then $X = (X_v)_{v \in V}$ was multivariate Gaussian with covariance matrix

$$\Sigma = (I - \Lambda)^{-T} \Omega (I - \Lambda)^{-1}$$

Mixed graphs

$G = (V, D, B)$ is a **mixed graph**, V is set of vertices, D directed edges, B bidirected edges.

Example

$$\begin{tikzcd} 1 \arrow[r] & 2 \arrow[r] \arrow[r, bend left, leftrightarrow] & 3 \end{tikzcd}$$

Compare to

$$\begin{tikzcd} & & 4 \arrow[dl] \arrow[d] & \\ 1 \arrow[r] & 2 \arrow[r] & 3 \end{tikzcd}$$

So let $G = ([m], D, B)$ be a mixed graph.

$$\mathrm{PD}(B) := \{\Omega \in \mathrm{PD}_m \colon \omega_{ij} = 0 \text{ if } i \neq j \text{ and } i \leftrightarrow j \notin B\}$$
$$\mathbb{R}^D := \left\{\Lambda \in \mathbb{R}^{m \times m} \colon \lambda_{ij} = 0 \text{ if } i \to j \notin D\right\}$$

Let $\varepsilon \sim \mathcal{N}(0, \Omega)$ for some $\Omega \in \mathrm{PD}(B)$. Then we can define

$$X_v = \varepsilon_v + \sum_{u \in \mathrm{pa}_G(v)} \lambda_{uv} X_u$$

Proposition (14.2.8)

Let $G = (V, D, B)$ be a mixed graph. Let $\Omega \in \mathrm{PD}(B)$, $\varepsilon \sim \mathcal{N}(0, \Omega)$, and $\Lambda \in \mathbb{R}^D$. Then the random vector X is a multivariate normal random variable with covariance matrix

$$\Sigma = (I - \Lambda)^{-T} \Omega (I - \Lambda)^{-1}.$$

Definition

Let $G = (V, D, B)$ be a mixed graph. The **linear structural equation model** $\mathcal{M}_G \subseteq \mathrm{PD}_m$ consists of all covariance matrices

$$\mathcal{M}_G := \left\{ (I - \Lambda)^{-T} \Omega (I - \Lambda)^{-1} \colon \Omega \in \mathrm{PD}(B), \Lambda \in \mathbb{R}^D \right\}$$

Definition (14.2.10)

Let $G = (V, D, B)$ be a mixed graph. Let G^{sub} be the directed graph obtained from G whose vertex set is $V \cup B$ and with edge set

$$D \cup \{ b \to i \colon i \leftrightarrow j \in B \} .$$

The resulting graph G^{sub} is the **bidirected subdivision** of G.

Proposition (14.2.11)

Let $G = (V, D, B)$ be a mixed graph with m vertices and G^{sub} the bidirected subdivision. Let $\mathcal{M}_G \subseteq \mathrm{PD}_m$ be the linear structural equation model associated to G and $\mathcal{M}'_{G^{sub}}$ be the Gaussian graphical model associated to the directed graph G^{sub} where all variables B are hidden variables. Then $\mathcal{M}'_{G^{sub}} \subseteq \mathcal{M}_G$ and $I(\mathcal{M}'_{G^{sub}}) = I(\mathcal{M}_G)$.

Idea of proof.

Let $M \in \mathcal{M}'_{G^{sub}}$ be given by Λ' and Ω'. Choose $\Lambda = \Lambda'_{V,V}$. Construct $\Omega = (\omega_{ij})$ via letting

$$\omega_{ij} = \omega'_{bb}\lambda_{bi}\lambda_{bj} \text{ and } \omega_{ii} = \omega'_{ii} + \sum_{b=i\to j\in B} \omega'_{bb}\lambda_{bi}^2$$

for each bidirected edge $b = i \leftrightarrow j$. Then apply trek rule. □

Definition (14.2.12)

Let $G = (V, D, B)$ be a mixed graph. A **trek** from i to j in G consists of either

1. a directed path P_L ending in i and a directed path P_R ending in j which have the same source, or
2. a directed path P_L ending in i and a directed path P_R ending in j such that the sources of P_L and P_R are connected by a bidirected edge.

Let $\mathcal{T}_G(i, j)$ denote the set of all treks in G connecting i and j.

To each trek T we associate a monomial m_T which is the product of all λ_{st} over all edges appearing in T times ω_{st}, where s and t are the sources of P_L and P_R.

Proposition (14.2.13, Trek rule)

Let $G = (V, D, B)$ *be a mixed graph. Let* $\Omega \in \mathrm{PD}(B)$, $\Lambda \in \mathbb{R}^D$, *and* $\Sigma = (I - \Lambda)^{-T}\Omega(I - \Lambda)^{-1}$. *Then*

$$\sigma_{ij} = \sum_{T \in \mathcal{T}_G(i,j)} m_T.$$

Example (14.2.14)

$$\begin{tikzcd} 1 \arrow[r] & 2 \arrow[r] \arrow[r, bend left, leftrightarrow] & 3 \end{tikzcd}$$

The following slides has propositions that where not talked about in the lecture.

Proposition (14.1.8)

Let $V \subseteq \mathbb{P}^{r-1}$ and suppose that $\operatorname{Sec}^k(V)$ is not a linear space. Then $\operatorname{Sec}^{k-1}(V)$ is in the singular locus of $\operatorname{Sec}^k(V)$.

This propostition tells how the sequence

$$\operatorname{Sec}^1(V) \subseteq \operatorname{Sec}^2(V) \subseteq \dots$$

behaves. Thus we, in some sense, get a upper bound on

$$\operatorname{Mixt}^1(\mathcal{M}) \subseteq \operatorname{Mixt}^2(\mathcal{M}) \subseteq \dots$$

behaves, by the proposition on slide 8.

Definition (14.2.4)

Let $G = (V, D)$ be a directed acyclic graph, and let $H \sqcup O = [m]$ be a partition of the index labeling into hidden variables H and observed variables O. The hidden variables are said to be **upstream** of the observed variables if there is no edge $o \to h$, where $o \in O$ and $h \in H$.

If this is the case we introduce the following two-dimensional grading on $\mathbb{R}[\Sigma]$, associated to this partition of the variables:

$$\deg \sigma_{ij} = \begin{pmatrix} 1 \\ \#(\{i\} \cap O) + \#(\{j\} \cap O) \end{pmatrix}.$$

Proposition (14.2.5)

Let $G = (V, D)$ be a directed acyclic graph, and let $H \sqcup O = [m]$ be a partition of the index labeling into hidden variables H and observed variables O, where the H variables are upstream. The ideal $J_G \subseteq \mathbb{R}[\Sigma]$ is homogeneous with respect to the upstream grading (defined above). In particular, any homogeneous generating set of J_G in this grading contains as a subset a generating ser of the vanishing ideal of the hidden variable model $J_G \cap \mathbb{R}[\Sigma_{O,O}]$.

As mentioned in the lecture, this tells us that we easily can find a representation of $J_G \cap \mathbb{R}[\Sigma_{O,O}]$. However, it requires that we can find a "nice" set of generators to J_G.