

Admixture models and coordinate ascent variational inference

In this lecture we will derive a **coordinate ascent variational inference (CAVI)** algorithm to do approximate posterior inference in an **admixture model**.

More specifically, the model will be **latent Dirichlet allocation (LDA)**, the most well-known form of admixture model introduced independently by Pritchard, Stephens, Donnelly (2000) for statistical genetics and Blei, Ng, Jordan (2003) for topic modeling.

Admixture model

The data consist of "documents", or more generally groups of data points. For "document" $d=1 \dots D$:

$$\omega_d \equiv (\omega_{d1} \dots \omega_{dN_d})$$

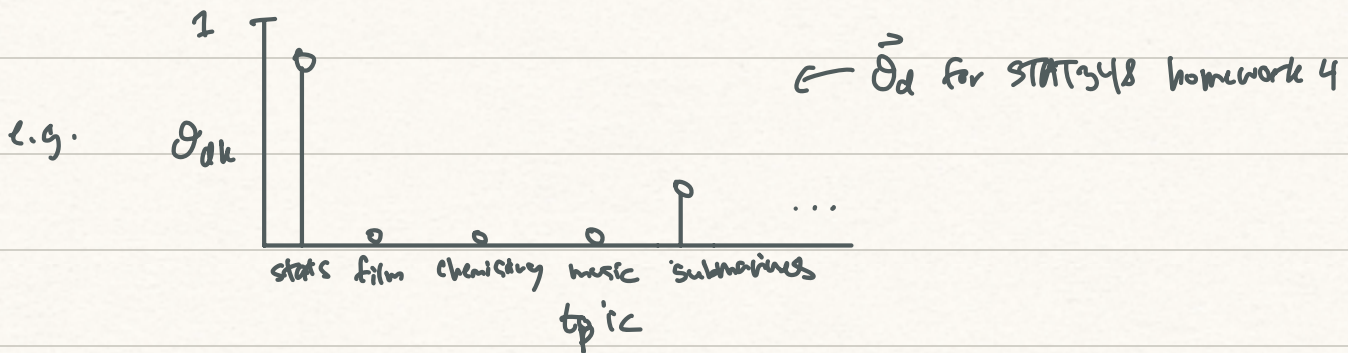
Each document is a group of N_d "word tokens", or more generally categorical data points. Each token i in document d is:

$$\omega_{di} \in \{1, \dots, V\}$$

\uparrow vocabulary

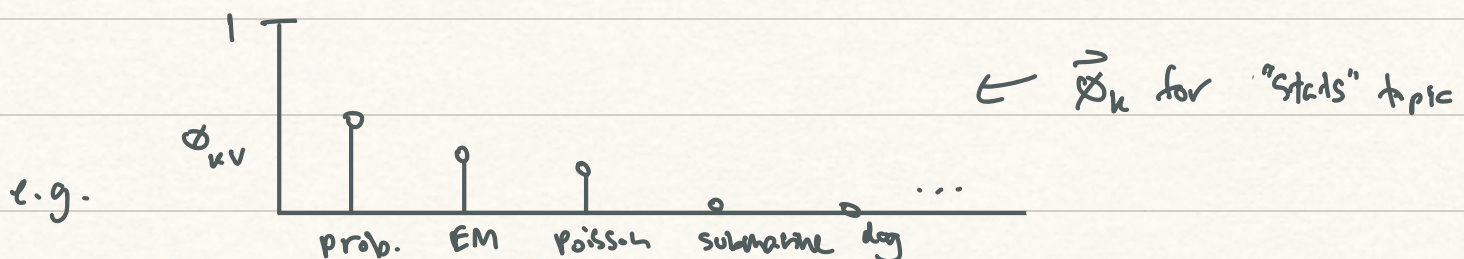
We assume that each document can be described by a mixture of **K topics**:

$$\theta_d \equiv (\theta_{d1} \dots \theta_{dK}), \quad \sum_k \theta_{dk} = 1$$



Each "topic" is then defined by a distribution over the vocabulary:

$$\phi_k \equiv (\phi_{k1} \dots \phi_{kV}), \quad \sum_v \phi_{kv} = 1$$



The generative process is as follows. We assume each word token was drawn from some topic:

for $d = 1 \dots D$:

for $i = 1 \dots N_d$:

$$z_{di} \sim \text{Categorical}(\theta_d)$$

$$w_{di} \sim \text{Categorical}(\phi_{z_{di}})$$

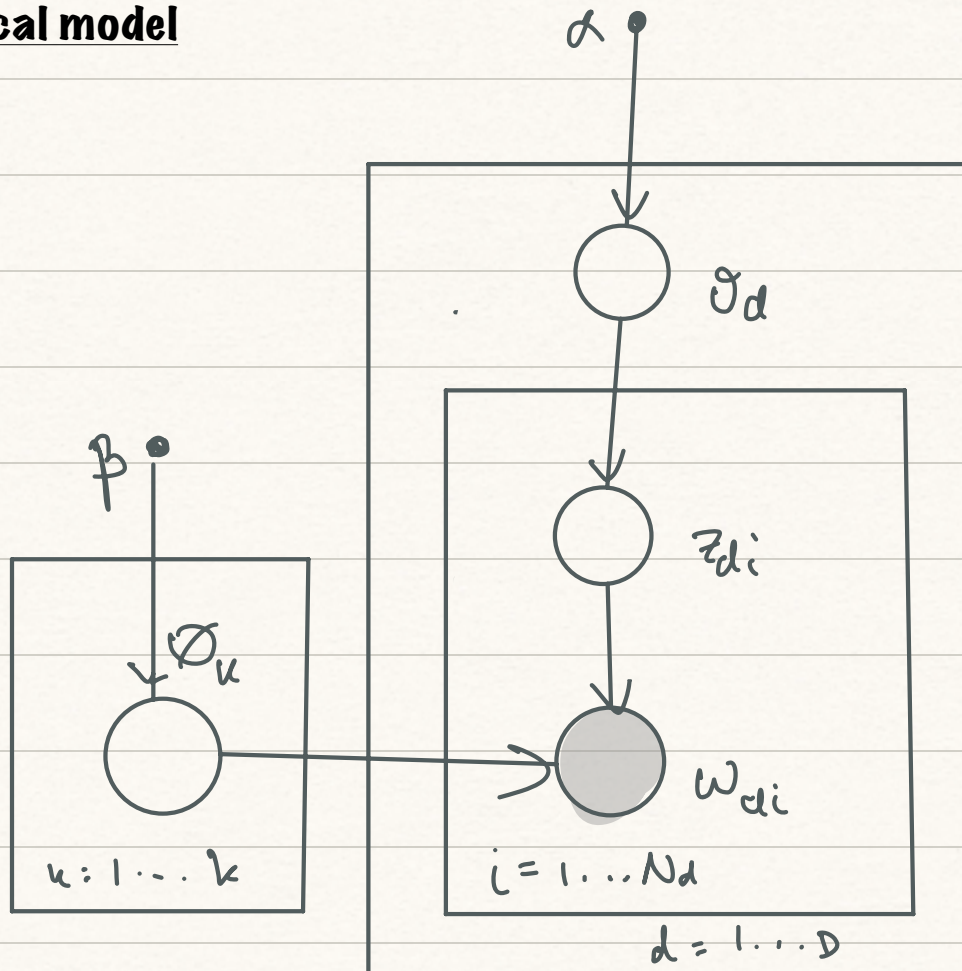
Each document is modeled by a mixture model, with mixture weights θ_d , but the whole corpus of documents is then said to be modeled by an **admixture model** or **mixed-membership mixture model**.

To complete a Bayesian specification of the model, we place priors over the parameters:

$$\theta_d \stackrel{\text{iid}}{\sim} \text{Dir}(\alpha_1 \dots \alpha_k), \quad d = 1 \dots D$$

$$\phi_k \stackrel{\text{iid}}{\sim} \text{Dir}(\beta_1 \dots \beta_V), \quad k = 1 \dots k$$

The graphical model



Applications of LDA

- **Topic modeling:** the data is a corpus of text documents. The goal is to infer a description of the **thematic structure** of the corpus via the topics.
- **Population genetics:** each "document" is an individual. At each genetic locus, they can have 0, 1 or 2 mutations. We assume that an individual's genome can be described as an admixture over **K ancestral populations** ("topics").
- **Inference of cell types:** each cell has been profiled and is associated with a collection of "reads", where each read corresponds to a genetic locus that has been expressed. We assume that cells can be described as a mixture over cell archetypes, each of which is a distribution over genes.
- **Radioactive material detection:** each document is a censor, which is a collection of "reads" of different levels of rays. We assume each censor corresponds to a mixture over topics of rays, each of which corresponds to a different kind of material.

Complete conditionals, sufficient stats, "bag of words"

$$p(z_{di} = k | w_{di} = v, -) \propto p(z_{di} = k | \theta_d) p(w_{di} = v | z_{di} = k, \Phi)$$

$$\propto \theta_{dk} \phi_{kv}$$

$$\rightarrow = \frac{\theta_{dk} \phi_{kv}}{\sum_j \theta_{dj} \phi_{jv}}$$

$$p(\theta_d | -) \propto p(\theta_d | \alpha) \prod_i \prod_k p(z_{di} = k | \theta_d)^{1(z_{di} = k)}$$

$$\propto \text{Dir}(\theta_d; \alpha) \prod_i \prod_k \theta_{dk}^{1(z_{di} = k)}$$

$$\propto \prod_k \theta_{dk}^{\alpha_k - 1} \cdot \prod_k \theta_{dk}^{\sum_i 1(z_{di} = k)}$$

$$\rightarrow = \text{Dir}(\theta_d; \alpha_1 + y_{d1}, \dots, \alpha_K + y_{dK})$$

$$\text{where } y_{dj} = \sum_i 1(z_{di} = j)$$

$$p(\phi_k | -) \propto \text{Dir}(\phi_k; \beta) \prod_d \prod_i \prod_v \phi_{kv}^{1(z_{di}=k) 1(w_{di}=v)}$$

$$\propto \prod_v \phi_{kv}^{\beta_v - 1} \cdot \prod_v \phi_{kv}^{\sum_d \sum_i \dots}$$

$$\rightarrow = \text{Dir}(\phi_k; \beta_1 + y_{k1}, \dots, \beta_V + y_{kV})$$

$$\text{where } y_{kv} = \sum_d \sum_i 1(z_{di}=k) 1(w_{di}=v)$$

Notice that the sufficient statistics for the Dirichlet vectors are just the counts of indicators Z . These counts form two matrices:

$$(y_{kv})_{k,v} \in \mathbb{N}^{K \times V} \quad (y_{dk})_{d,k} \in \mathbb{N}^{D \times K}$$

Consider how we might sample the Z 's in a Gibbs sampler, and then compute these statistics:

for $d = 1 \dots D$:

for $v = 1 \dots V$:

for $i = 1 \dots N_d$:

if $w_{di} = v$:

$$z_{di} \sim \text{Cat}(p_{dv})$$

$$\text{where } p_{dv} = (p_{dv1} \dots p_{dvK}), \quad p_{dvk} = \frac{\phi_{kv} \phi_{dv}}{\sum_j \phi_{dj} \phi_{jv}}$$

The inner loop over the vocabulary $v = 1 \dots V$ is just a way to show that all indicators z_{di} for tokens $w_{di}=v$ are iid from the same complete conditional, described by the probability vector p_{dv} .

The number of tokens in document d that take value v are:

$$y_{dv} = \sum_{i=1}^{N_d} 1(w_{di}=v)$$

Therefore, equivalently, we can sample the sums of indicators from a multinomial.

for $d = 1 \dots D$

for $v = 1 \dots V$

$$(y_{dv1} \dots y_{dvK}) \sim \text{Multinomial}(y_{dv}, p_{dv})$$

where $y_{dvk} = \sum_i 1(w_{di} = v) 1(z_{di} = k)$

"The number of tokens in document d of word type v assigned to topic k "

The sufficient statistics for the parameters are then just contractions of this 3-dimensional array:

$$y_{dk} = \sum_v y_{dvk}$$

$$y_{kv} = \sum_d y_{dvk}$$

What have we learned? To fit LDA to a corpus of word tokens, we only need the document-word counts:

$$p(\Theta, \Phi, z \mid (w_d)_d) = p(\Theta, \Phi, z \mid (y_{dv})_{d,v})$$

In other words, the generative process encodes the "bag of words" assumption. The order or the unique identity of the different word tokens does not matter. All that matters is how many tokens of each word type v are in each document d : y_{dv} .

This is a very bad assumption about how text is actually generated. Nevertheless, when applied to documents, LDA learns interpretable latent structure that can be useful at a high level.

CAVI for LDA

We will make the simplifying "mean-field assumption" that the q -distribution is a factorized family. This will facilitate the optimization problem:

$$q(\Theta, \Phi, z) \approx p(\Theta, \Phi, z | w)$$

We will make the simplifying "mean-field assumption" that the q -distribution is a factorized family. This will facilitate the optimization problem:

$$q(\Theta, \Phi, z) = \prod_d q(\theta_d) \prod_k q(\phi_w) \prod_d \prod_i q(z_{di})$$

Recall from last time the ELBO objective function we will seek to maximize:

$$\underset{q}{\operatorname{argmin}} \operatorname{KL}(q(\dots) \parallel p(\dots | w)) = \underset{q}{\operatorname{argmax}} \underbrace{\mathbb{E}_q \left[\log \frac{p(\dots)}{q(\dots)} \right]}_{= \operatorname{ELBO}(q)}$$

Since q factorizes, it will be convenient to maximize the ELBO coordinate-wise (i.e., updating each factor, holding all other factors fixed).

$$\text{e.g. } q^*(\theta_d) = \underset{q(\theta_d)}{\operatorname{argmax}} \operatorname{ELBO}(q)$$

For any factor, the optimal update takes the following form, using θ_d as an example:

$$q^*(\theta_d) \propto \exp \left(\mathbb{E}_{q_{\setminus \theta_d}} [\log p(\theta_d | -)] \right)$$

where the expectation is with respect to all factors in q except the given one, and the term inside is the complete conditional of the given variable under the true model.

This general form for the optimal update is given by Bishop (2006). Let's confirm that it is true. Consider the KL from any other setting of the density $q(\theta_d)$ to the (supposedly) optimal one:

$$\begin{aligned} \operatorname{KL}(q(\theta_d) \parallel q^*(\theta_d)) &= \mathbb{E}_{q(\theta_d)} [\log q(\theta_d) - \log q^*(\theta_d)] \\ &\propto \mathbb{E}_{q(\theta_d)} [\log q(\theta_d)] - \mathbb{E}_{q(\theta_d)} [\mathbb{E}_{q_{\setminus \theta_d}} [\log p(\theta_d | -)]] \\ &\propto -H(q(\theta_d)) - \mathbb{E}_q [\log p(\Theta, \Phi, z | w)] \\ &\propto -\operatorname{ELBO}(q) \end{aligned}$$

So indeed to maximize the ELBO wrt $q(\theta_d)$ you minimize the KL to the stated optimal update.

As we've seen before with EM, the optimal update takes the form of a geometric mean. This means that if the complete conditional is an exponential family, the optimal update to the factor is also a member of that same exponential family:

$$\text{if } p(\theta_d | -) = f(\theta_d; \eta_d(\dots)) \propto h(\theta_d) \exp(\ell(\theta_d)^T \eta_d(\dots))$$

where the complete conditional's natural parameter is potentially a function of data and other latent variables.

$$\text{then } q^*(\theta_d) = f(\theta_d; \mathbb{E}_q[\eta_d(\dots)])$$

where the natural parameter of this factor is the expectation under the variational distribution (not including this factor) of the complete conditional's natural parameter.

To be clear, this says two things: 1) that the optimal factor is in the same exponential family as the complete conditional, and 2) that its natural parameter takes the form above. To be explicit about this we will introduce notation for the **variational parameters** (which until now have been implicit):

$$\textcircled{1} \quad q^*(\theta_d) = q(\theta_d; \tilde{\eta}_d) \equiv f(\theta_d; \tilde{\eta}_d)$$

$$\textcircled{2} \quad \tilde{\eta}_d = \mathbb{E}_q[\eta_d(\dots)]$$

we are using a tilde to distinguish the **variational parameter** from the parameter of the complete conditional in the true model.

In the case, the complete conditional is Dirichlet. The natural parameter of the Dirichlet is just its shape parameters (the usual parameters), so:

$$\begin{aligned} q^*(\theta_d) &= \text{Dir}(\theta_d; \alpha_1 + \mathbb{E}_q[y_{d1}], \dots, \alpha_k + \mathbb{E}_q[y_{dk}]) \\ &\equiv \text{Dir}(\theta_d; \tilde{\alpha}_{d1}, \dots, \tilde{\alpha}_{dk}) \end{aligned}$$

Let's take a look at the expectations:

$$\begin{aligned} \mathbb{E}_q[y_{dk}] &\equiv \mathbb{E}_{q_{\setminus \theta_d}} \left[\sum_i 1(z_{id} = k) \right] \\ &= \sum_{i=1}^{N_d} q(z_{id} = k) \end{aligned}$$

So we see that this depends on the factors $q(z_{id})$.

$$q^*(z_{id} = k) \propto \exp \left(\mathbb{E}_q \left[\log p(z_{id} = k | w_{id} = v, -) \right] \right) \\ \propto \exp \left(\mathbb{E}_q \left[\log \frac{\vartheta_{dk} \phi_{kv}}{\sum_j \vartheta_{dj} \phi_{jv}} \right] \right)$$

The normalizer of the probability does not depend on k (the assumed value of z_{id}):

$$\propto \exp \left(\mathbb{E}_q \left[\log \vartheta_{dk} \phi_{kv} \right] \right) \\ \propto \exp \left(\mathbb{E}_{q(\vartheta_d)} \left[\log \vartheta_{dk} \right] + \mathbb{E}_{q(\phi_{kv})} \left[\log \phi_{kv} \right] \right) \\ \propto G_q[\vartheta_{dk}] \cdot G_q[\phi_{kv}]$$

Here we are introducing new notation for the **geometric expected value**.

$$\text{where } G_q[x] \triangleq \exp \left(\mathbb{E}_{x \sim q} [\log x] \right)$$

Putting it altogether:

$$q^*(z_{id} = k) \equiv \tilde{p}_{avk} = \frac{G_q[\vartheta_{dk}] G_q[\phi_{kv}]}{\sum_j G_q[\vartheta_{dj}] G_q[\phi_{jv}]}$$

Do we know these geometric expected values? In this case, the relevant q -distribution is a Dirichlet (as we showed in the last derivation). Both the arithmetic and geometric expected values under a Dirichlet are known in closed form:

$$(x_1 \dots x_k) \sim \text{Dir}(\alpha_1 \dots \alpha_k)$$

$$\mathbb{E}[x_k] = \frac{\alpha_k}{\sum_j \alpha_j}, \quad G[x_k] = \frac{\exp(\psi(\alpha_k))}{\sum_j \exp(\psi(\alpha_j))}$$

where $\psi(\cdot)$ is digamma function.

CAVI for LDA:

While ELBO not converged:

$$q(\theta_d) \quad || \quad \begin{aligned} \tilde{\alpha}_{dk} &= \alpha_k + \mathbb{E}_q[Y_{dk}] & \forall d, k \\ \zeta_q[\theta_{dk}] &= \exp(\psi(\tilde{\alpha}_{dk}) - \sum_j \psi(\alpha_{dj})) & \forall d, k \end{aligned}$$

$$q(\phi_u) \quad || \quad \begin{aligned} \tilde{\beta}_{kv} &= \beta_v + \mathbb{E}_q[Y_{kv}] & \forall k, v \\ \zeta_q[\phi_{kv}] &= \exp(\psi(\tilde{\beta}_{kv}) - \sum_j \psi(\tilde{\beta}_{jv})) & \forall k, v \end{aligned}$$

$$q(z_{id}) \quad || \quad \begin{aligned} \tilde{p}_{dvk} &\propto \zeta_q[\theta_{dk}] \zeta_q[\phi_{kv}] & \forall d, v, k \\ \mathbb{E}_q[Y_{dk}] &= \sum_v Y_{dv} \tilde{p}_{dvk} \\ \mathbb{E}_q[Y_{kv}] &= \sum_d Y_{dv} \tilde{p}_{dvk} \end{aligned}$$

compute ELBO(q)

Computing the ELBO:

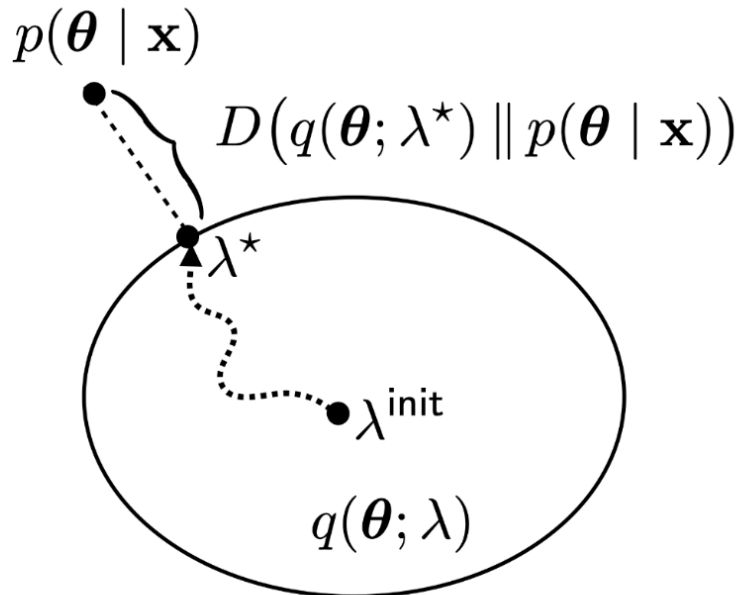
$$\begin{aligned} & \mathbb{E}_q \left[\log \frac{p(\theta, \phi, z, w)}{q(\theta, \phi, z)} \right] \\ &= \mathbb{E}_q \left[\log p(w | -) \right] + \mathbb{E}_q \left[\log \frac{p(\theta, \phi, z)}{q(\theta, \phi, z)} \right] \\ &= \mathbb{E}_q \left[\log p(w | -) \right] + \mathbb{E}_q \left[\log \frac{p(\theta)}{q(\theta)} \cdot \frac{p(\phi)}{q(\phi)} \cdot \frac{p(z | \theta)}{q(z)} \right] \end{aligned}$$

$$\begin{aligned}
\mathbb{E}_q \left[\log \frac{p(\mathbf{y})}{q(\mathbf{y})} \right] &= \sum_d \mathbb{E}_q \left[\log \frac{p(\mathcal{D}_d)}{q(\mathcal{D}_d)} \right] \\
&= - \sum_d \text{KL}(q(\mathcal{D}_d) \parallel p(\mathcal{D}_d)) \\
&= - \sum_d \text{KL}(D(\mathcal{D}_d; \tilde{\lambda}_d) \parallel D(\mathcal{D}_d; \lambda))
\end{aligned}$$

The KL between two Dirichlets is known in closed form. The other terms can be simplified in similar ways and will end up involving simple expectations under the variational distribution.

What is CAVI doing?

Popping up a level, CAVI is finding the member of a simple parametric family that is closest in KL to the exact posterior. It does so by maximizing the ELBO coordinate-wise. Here's the picture:



In this case, θ represents all latent variables parameters, \mathbf{x} is the data, and λ are the variational parameters. λ^* represents the value of the variational parameters that brings $q(\dots)$ closest in KL to the exact posterior. Since q is a factorized family though, there may always be a gap.