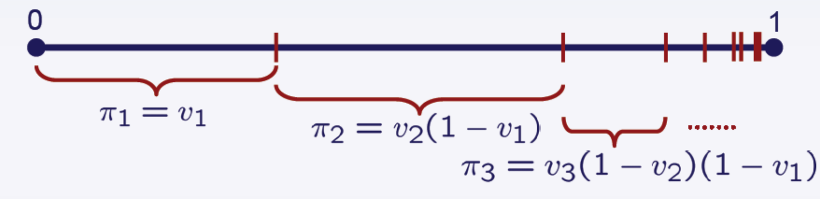
Variational Inference for Dirichlet Process Mixtures

David Blei and Michael Jordan

May 5, 2010

Twenty Thousand Feet view

- Given a model θ and data $\mathbf{x} = \{x_1, ...x_N\}$.
  - We want to learn the model $\hat{\theta}$.
  - Make predictions about a new data point x_{N+1}.
- Being Bayesians we want to
  - Estimate the posterior distribution over the model(parameters) $p(\theta|\mathbf{x})$
  - Estimate the predictive distribution $p(x_{N+1}|\mathbf{x}) = \int p(x_{N+1}|\theta)p(\theta|\mathbf{x})d\theta$
- Finally we go one step further and assume our parameters grow with data.

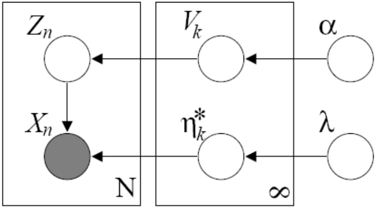


- $V_i \sim Beta(1, \alpha)$
- $\eta_i^* \sim G_0$
- $\pi_i(\mathbf{v}) = v_i \prod_{l=1}^{j-1}(1 - v_l)$
- $G = \sum_{i=1}^{\infty} \pi_i(\mathbf{v}) \delta \eta_i^*$

Dirichlet Process Mixtures

- Generalization of finite mixture models.
- A Dirichlet Process prior is placed over mixture components.
- Nonparametric, do not have to specify the number of components before hand.

DP Mixture Model



1. Draw $V_i \mid \alpha \sim \text{Beta}(1, \alpha), \quad i = \{1, 2, \ldots\}$
2. Draw $\eta_i^* \mid G_0 \sim G_0, \quad i = \{1, 2, \ldots\}$
3. For the nth data point:
   (a) Draw $Z_n \mid \{v_1, v_2, \ldots\} \sim \text{Mult}(\pi(\mathbf{v}))$.
   (b) Draw $X_n \mid z_n \sim p(x_n \mid \eta_{z_n}^*)$.

Posterior over the latent variables

- Let $\mathbf{W} = \{\mathbf{V}, \boldsymbol{\eta}^*, \mathbf{Z}\}$ and let $\theta = \{\alpha, \lambda\}$

Posterior over the latent variables

- Let $\mathbf{W} = \{\mathbf{V}, \boldsymbol{\eta}^*, \mathbf{Z}\}$ and let $\theta = \{\alpha, \lambda\}$
- Posterior over the latent variables takes the form:
- $p(\mathbf{w}|\mathbf{x}, \theta) = exp\{log(p(\mathbf{x}, \mathbf{w}|\theta)) - log \int p(\mathbf{x}, \mathbf{w}|\theta))d\mathbf{w}\}$

Posterior over the latent variables

- Let $\mathbf{W} = \{\mathbf{V}, \boldsymbol{\eta}^*, \mathbf{Z}\}$ and let $\theta = \{\alpha, \lambda\}$
- Posterior over the latent variables takes the form:
- $p(\mathbf{w}|\mathbf{x}, \theta) = exp\{log(p(\mathbf{x}, \mathbf{w}|\theta)) - log \int p(\mathbf{x}, \mathbf{w}|\theta))d\mathbf{w}\}$
  - The integral over the latent variables, makes exact computation of the posterior intractable.

Approximate Inference

- Posterior is intractable.
- Use either MCMC or approximate deterministic inference techniques.
- Here the authors present a mean field variational method.

Variational Inference

- $log p(\mathbf{x}|\theta) = log \int p(\mathbf{w}, \mathbf{x}|\theta) d\mathbf{w}$

Variational Inference

- $log p(\mathbf{x}|\theta) = log \int p(\mathbf{w}, \mathbf{x}|\theta) d\mathbf{w}$
- $log p(\mathbf{x}|\theta) = log \int q_\nu(\mathbf{w}) \frac{p(\mathbf{w}, \mathbf{x}|\theta) d\mathbf{w}}{q_\nu(\mathbf{w})}$

Variational Inference

- $log p(\mathbf{x}|\theta) = log \int p(\mathbf{w}, \mathbf{x}|\theta) d\mathbf{w}$
- $log p(\mathbf{x}|\theta) = log \int q_\nu(\mathbf{w}) \frac{p(\mathbf{w}, \mathbf{x}|\theta) d\mathbf{w}}{q_\nu(\mathbf{w})}$
- $= log \mathbb{E}_q \frac{p(\mathbf{W}, \mathbf{x}|\theta)}{q_\nu(\mathbf{W})}$
- $\geq \mathbb{E}_q log \frac{p(\mathbf{W}, \mathbf{x}|\theta)}{q_\nu(\mathbf{W})}$

Variational Inference

- $logp(\mathbf{x}|\theta) = log \int p(\mathbf{w}, \mathbf{x}|\theta) d\mathbf{w}$
- $logp(\mathbf{x}|\theta) = log \int q_\nu(\mathbf{w}) \frac{p(\mathbf{w}, \mathbf{x}|\theta) d\mathbf{w}}{q_\nu(\mathbf{w})}$
- $= log \mathbb{E}_q \frac{p(\mathbf{W}, \mathbf{x}|\theta)}{q_\nu(\mathbf{W})}$
- $\geq \mathbb{E}_q log \frac{p(\mathbf{W}, \mathbf{x}|\theta)}{q_\nu(\mathbf{W})}$
- $= \mathbb{E}_q[logp(\mathbf{W}, \mathbf{x}|\theta)] - \mathbb{E}_q[logq_\nu(\mathbf{W})] \equiv \mathcal{L}(q)$

Variational Inference

- $KL(q_\nu(\mathbf{w})||p(\mathbf{w}|\mathbf{x},\theta)) = logp(\mathbf{x}|\theta) - (\mathbb{E}_q[logp(\mathbf{W},\mathbf{x}|\theta)] - \mathbb{E}_q[logq_\nu(\mathbf{W})])$

Variational Inference

- $KL(q_\nu(\mathbf{w})||p(\mathbf{w}|\mathbf{x},\theta)) = logp(\mathbf{x}|\theta) - (\mathbb{E}_q[logp(\mathbf{W},\mathbf{x}|\theta)] - \mathbb{E}_q[logq_\nu(\mathbf{W})])$
- Equivalently, $log(p(\mathbf{x}|\theta)) \geq E_q[logp(\mathbf{W},\mathbf{x}|\theta))] - E_q[logq_\nu(\mathbf{W})]$
  - With the bound being tight when $q_\nu(\mathbf{w}) = p(\mathbf{w}|\mathbf{x},\theta)$

Variational Inference

- $KL(q_\nu(\mathbf{w})||p(\mathbf{w}|\mathbf{x},\theta)) = logp(\mathbf{x}|\theta) - (\mathbb{E}_q[logp(\mathbf{W},\mathbf{x}|\theta)] - \mathbb{E}_q[logq_\nu(\mathbf{W})])$
- Equivalently, $log(p(\mathbf{x}|\theta)) \geq E_q[logp(\mathbf{W},\mathbf{x}|\theta))] - E_q[logq_\nu(\mathbf{W})]$
  - With the bound being tight when $q_\nu(\mathbf{w}) = p(\mathbf{w}|\mathbf{x},\theta)$
- $\underset{\nu}{argmax}\mathcal{L}(q) \Leftrightarrow \underset{\nu}{argmin}KL(q_\nu(\mathbf{w})||p(\mathbf{w}|\mathbf{x},\theta))$

Variational Inference

- $KL(q_\nu(\mathbf{w})||p(\mathbf{w}|\mathbf{x},\theta)) = logp(\mathbf{x}|\theta) - (\mathbb{E}_q[logp(\mathbf{W},\mathbf{x}|\theta)] - \mathbb{E}_q[logq_\nu(\mathbf{W})])$
- Equivalently, $log(p(\mathbf{x}|\theta)) \geq E_q[logp(\mathbf{W},\mathbf{x}|\theta))] - E_q[logq_\nu(\mathbf{W})]$
  - With the bound being tight when $q_\nu(\mathbf{w}) = p(\mathbf{w}|\mathbf{x},\theta)$
- $\underset{\nu}{argmax}\mathcal{L}(q) \Leftrightarrow \underset{\nu}{argmin}KL(q_\nu(\mathbf{w})||p(\mathbf{w}|\mathbf{x},\theta))$
- $q_\nu(\mathbf{w})$ is the *variational distribution* and ν is the corresponding variational parameter.
- Note that the marginal probability of the data has no variational parameter.

Mean Field Variational Inference

- Further assume that the variational distribution factorizes as $q_\nu(\mathbf{w}) = \prod_{i=1}^{M} q_{\nu\, m}(w_m)$
- Now,

$$log(p(\mathbf{x}|\theta)) \geq \mathbb{E}_q[logp(\mathbf{W}, \mathbf{x}|\theta)] - \mathbb{E}_q[logq_\nu(\mathbf{W})] \quad (1)$$

- $$log(p(\mathbf{x}|\theta)) \geq \mathbb{E}_q[logp(\mathbf{W}|\mathbf{x}, \theta) + log(p(\mathbf{x}|\theta))] - \mathbb{E}_q[logq_\nu(\mathbf{W})] \quad (2)$$

- $$log(p(\mathbf{x}|\theta)) \geq log(p(\mathbf{x}|\theta)) + \mathbb{E}_q[logp(\mathbf{W}|\mathbf{x}, \theta)] - \sum_{m=1}^{M} \mathbb{E}_q[logq_{\nu\, m}(W_m)] \quad (3)$$

Coordinate Ascent

- Optimize with respect to ν_i holding all ν_j, $j \neq i$ constant.

Coordinate Ascent

- Optimize with respect to ν_i holding all ν_j, $j \neq i$ constant.
- If $p(w_i | \mathbf{w}_{-i}, \mathbf{x}, \theta)$ is an exponential family distribution

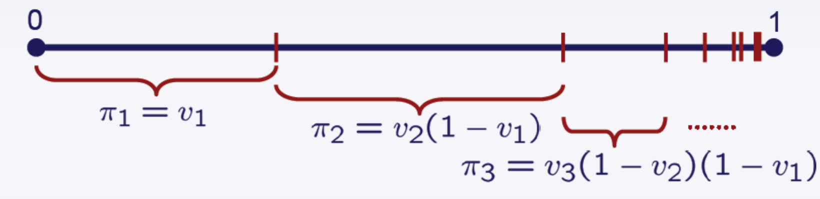
Coordinate Ascent

- Optimize with respect to ν_i holding all ν_j, $j \neq i$ constant.
- If $p(w_i|\mathbf{w}_{-i}, \mathbf{x}, \theta)$ is an exponential family distribution
- Then the corresponding variational parameter ν_i which optimizes the KL divergence has a closed form solution.

Nonparametrics

- The treatment so far has been general.
- It applies to parametric cases just as much as it does to nonparametrics.
- Further innovations required to apply it to nonparametric cases.

Back to stick breaking



- *If any* $v_t = 1$, $\pi_j(v) = 0, \forall j > t$

Variational approximation for the DP mixture

- Recall, $\mathbf{W} = \{\mathbf{V}, \boldsymbol{\eta}^*, \mathbf{Z}\}$ and $\theta = \{\alpha, \lambda\}$

$$log(p(\mathbf{x}|\theta)) \geq \mathbb{E}_q[logp(\mathbf{W}, \mathbf{x}|\theta)] - \mathbb{E}_q[logq_\nu(\mathbf{W})]$$

Variational approximation for the DP mixture

- Recall, $\mathbf{W} = \{\mathbf{V}, \boldsymbol{\eta}^*, \mathbf{Z}\}$ and $\theta = \{\alpha, \lambda\}$

$$
\begin{aligned}
log(p(\mathbf{x}|\alpha, \lambda)) \quad \geq \quad & \mathbb{E}_q[logp(\mathbf{V}|\alpha)] + \mathbb{E}_q[logp(\boldsymbol{\eta}^*|\lambda)] \\
& + \sum_{n=1}^{N} (\mathbb{E}_q[logp(Z_n|\mathbf{V})] + \mathbb{E}_q[logp(x_n|Z_n)]) \\
& - \mathbb{E}_q[logq(\mathbf{V}, \boldsymbol{\eta}^*, \mathbf{Z})]
\end{aligned}
$$

Variational approximation for the DP mixture

- Recall, $\mathbf{W} = \{\mathbf{V}, \boldsymbol{\eta}^*, \mathbf{Z}\}$ and $\theta = \{\alpha, \lambda\}$

$$
\begin{aligned}
log(p(\mathbf{x}|\alpha, \lambda)) \quad \geq \quad & \mathbb{E}_q[logp(\mathbf{V}|\alpha)] + \mathbb{E}_q[logp(\boldsymbol{\eta}^*|\lambda)] \\
& + \sum_{n=1}^{N} (\mathbb{E}_q[logp(Z_n|\mathbf{V})] + \mathbb{E}_q[logp(x_n|Z_n)]) \\
& - \mathbb{E}_q[logq(\mathbf{V}, \boldsymbol{\eta}^*, \mathbf{Z})]
\end{aligned}
$$

- Truncate by setting $q(v_T = 1) = 1$.
- *We are truncating the variational distribution. The model is still Nonparametric.*

Variational approximation for the DP mixture II

- The variational distribution now becomes
  $$q(\mathbf{v}, \boldsymbol{\eta}^*, \mathbf{z}) = \prod_{t=1}^{T-1} q_{\gamma_t}(v_t) \prod_{t=1}^{T} q_{\tau_t}(\eta_t^*) \prod_{n=1}^{N} q_{\phi_n}(z_n)$$
- $q_{\gamma_t}(v_t)$ are chosen to be beta distributions, $q_{\tau_t}(\eta_t^*)$ are some distributions in the exponential family and $q_{\phi_n}(z_n)$ are multinomial distributions.
- The variational parameters are

$$\nu = \{\gamma_1, ..., \gamma_{T-1}, \tau_1, ..., \tau_T, \phi_1, ..., \phi_N\} \tag{4}$$

Some Gory Details

$$
\begin{aligned}
log(p(\mathbf{x}|\alpha,\lambda)) \geq \; & \mathbb{E}_q[logp(\mathbf{V}|\alpha)] + \mathbb{E}_q[logp(\boldsymbol{\eta}^*|\lambda)] \\
& + \sum_{n=1}^{N} (\mathbb{E}_q[logp(Z_n|\mathbf{V})] + \mathbb{E}_q[logp(x_n|Z_n)]) \\
& - \mathbb{E}_q[\prod_{t=1}^{T-1} q_{\gamma_t}(v_t) \prod_{t=1}^{T} q_{\tau_t}(\eta_t^*) \prod_{n=1}^{N} q_{\phi_n}(z_n)]
\end{aligned}
$$

Some Gory Details

$$
\begin{aligned}
log(p(\mathbf{x}|\alpha,\lambda)) \geq\ & \mathbb{E}_q[logp(\mathbf{V}|\alpha)] + \mathbb{E}_q[logp(\boldsymbol{\eta}^*|\lambda)] \\
& + \sum_{n=1}^{N}(\mathbb{E}_q[logp(Z_n|\mathbf{V})] + \mathbb{E}_q[logp(x_n|Z_n)]) \\
& - \mathbb{E}_q[\prod_{t=1}^{T-1} q_{\gamma_t}(v_t) \prod_{t=1}^{T} q_{\tau_t}(\eta_t^*) \prod_{n=1}^{N} q_{\phi_n}(z_n)]
\end{aligned}
$$

Some Gory Details II

- In $\boldsymbol{\pi}(\mathbf{v})$ was finite dimensional $p(Z_n = t|\mathbf{V}) = \prod_{t=1}^{T} (\pi_t)^{\mathcal{I}(Z_n=t)}$

Some Gory Details II

- In $\boldsymbol{\pi}(\mathbf{v})$ was finite dimensional $p(Z_n = t|\mathbf{V}) = \prod_{t=1}^{T} (\pi_t)^{\mathcal{I}(Z_n=t)}$
- In our infinite dimensional case $p(Z_n = t|\mathbf{V}) =$
  $\prod_{t=1}^{\infty} (V_t \prod_{l=1}^{t-1}(1 - V_l))^{\mathcal{I}(Z_n=t)} = \prod_{t=1}^{\infty}(V_t)^{\mathcal{I}(Z_n=t)}(\prod_{l=1}^{t-1}(1 - V_l))^{\mathcal{I}(Z_n=t)}$

Some Gory Details II

- In $\boldsymbol{\pi}(\mathbf{v})$ was finite dimensional $p(Z_n = t|\mathbf{V}) = \prod_{t=1}^{T} (\pi_t)^{\mathcal{I}(Z_n=t)}$
- In our infinite dimensional case $p(Z_n = t|\mathbf{V}) = \prod_{t=1}^{\infty} (V_t \prod_{l=1}^{t-1}(1 - V_l))^{\mathcal{I}(Z_n=t)} = \prod_{t=1}^{\infty}(V_t)^{\mathcal{I}(Z_n=t)}(\prod_{l=1}^{t-1}(1 - V_l))^{\mathcal{I}(Z_n=t)}$
- Equivalently,

$$p(Z_n|\mathbf{V}) = \prod_{t=1}^{\infty} V_t^{\mathcal{I}(Z_n=t)}(1 - V_t)^{\mathcal{I}(Z_n>t)} \tag{5}$$

Some Gory Details II

- In $\boldsymbol{\pi}(\mathbf{v})$ was finite dimensional $p(Z_n = t|\mathbf{V}) = \prod_{t=1}^{T} (\pi_t)^{\mathcal{I}(Z_n=t)}$
- In our infinite dimensional case $p(Z_n = t|\mathbf{V}) = \prod_{t=1}^{\infty} (V_t \prod_{l=1}^{t-1}(1 - V_l))^{\mathcal{I}(Z_n=t)} = \prod_{t=1}^{\infty}(V_t)^{\mathcal{I}(Z_n=t)}(\prod_{l=1}^{t-1}(1 - V_l))^{\mathcal{I}(Z_n=t)}$
- Equivalently,

$$p(Z_n|\mathbf{V}) = \prod_{t=1}^{\infty} V_t^{\mathcal{I}(Z_n=t)}(1 - V_t)^{\mathcal{I}(Z_n>t)} \tag{5}$$

- With the truncation at T, we have

$$\begin{aligned} \mathbb{E}_q[logp(Z_n|\mathbf{V})] &= \sum_{t=1}^{T} q(Z_n > t)\mathbb{E}_q[log(1 - V_t)] \\ &\quad + q(Z_n = t)\mathbb{E}_q[logV_t] \end{aligned}$$

(6)

Comparison with Collapsed and Truncated Gibbs sampling

- Collapsed - Analogous to parametric cases. Integrating over G and η^* leads to a Polya Urn Scheme.
- Ben will talk more about this.

Truncated Gibbs Sampling

- Issue of sampling from the infinite dimensional quantity V.
- Solution: Truncate V to some fixed quantity T.
- Unlike truncating in the variational case, the true distribution is truncated.
- The truncated process $\simeq$ DP when the truncation level is large relative to the number of data points.

Experimental Setup

- The model - DP mixture of Gaussians, with fixed covariance.
- Toy problem - Each dataset contians 100 train and test points, with data dimensionality varying from 5 to 50.
- Each dimensionality has 10 synthetic datasets.

Results

Dim	Mean held out log probability (Std err)		
	Variational	Collapsed Gibbs	Truncated Gibbs
5	-147.96 (4.12)	-148.08 (3.93)	-147.93 (3.88)
10	-266.59 (7.69)	-266.29 (7.64)	-265.89 (7.66)
20	-494.12 (7.31)	-492.32 (7.54)	-491.96 (7.59)
30	-721.55 (8.18)	-720.05 (7.92)	-720.02 (7.96)
40	-943.39 (10.65)	-941.04 (10.15)	-940.71 (10.23)
50	-1151.01 (15.23)	-1148.51 (14.78)	-1147.48 (14.55)

Table 1: Average held-out log probability for the predictive distributions given by variational inference, TDP Gibbs sampling, and the collapsed Gibbs sampler.

Convergence Time Comparison

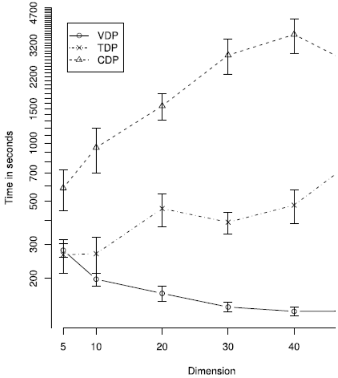
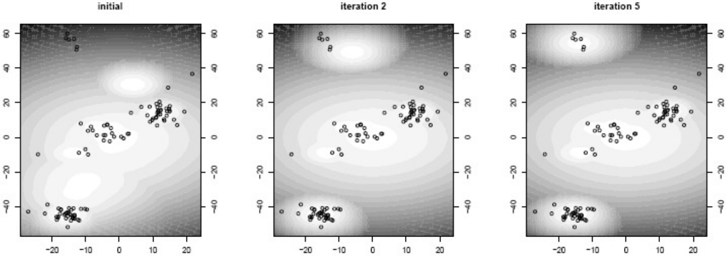


Figure 3: Mean convergence time and standard error across ten data sets per dimension for variational inference, TDP Gibbs sampling, and the collapsed Gibbs sampler.

Model Selection



- Truncation level was set at 20.
- Only 5 mixture components get used.

Large Scale applicability

- Clusters 5000 real world images.
- Each image is represented as 192 dimensional vectors.
- Convergence in 4 hours $\simeq$ 16 iterations of Gibbs sampling.

Coordinate Ascent

- Optimize with respect to ν_i holding all ν_j, $j \neq i$ constant.

Coordinate Ascent

- Optimize with respect to ν_i holding all ν_j, $j \neq i$ constant.
- Terms containing ν_i are

$$l_i = \mathbb{E}_q[logp(W_i|\mathbf{W_{-i}}, \mathbf{x}, \boldsymbol{\theta})] - \mathbb{E}_q[logq_{\nu i}(W_i)] \qquad (7)$$

Coordinate Ascent

- Optimize with respect to ν_i holding all ν_j, $j \neq i$ constant.
- Terms containing ν_i are

$$l_i = \mathbb{E}_q[logp(W_i|\mathbf{W_{-i}}, \mathbf{x}, \boldsymbol{\theta})] - \mathbb{E}_q[logq_{\nu i}(W_i)] \quad (7)$$

- If,

$$p(w_i|\mathbf{w}_{-1}, \mathbf{x}, \theta) = h(w_i)exp\{g(\mathbf{w}_{-1}, \mathbf{x}, \theta)^T w_i - a(g(\mathbf{w}_{-1}, \mathbf{x}, \theta))\} \quad (8)$$

- then,

$$\nu_i = \mathbb{E}_q[g(\mathbf{w}_{-1}, \mathbf{x}, \theta)] \quad (9)$$

Predicitve Distribution

$$p(x_{N+1} \mid \mathbf{x}, \alpha, \lambda) = \int \left(\sum_{t=1}^{\infty} \pi_t(\mathbf{v}) p(x_{N+1} \mid \eta_t^*) \right) dP(\mathbf{v}, \boldsymbol{\eta}^* \mid \mathbf{x}, \lambda, \alpha).$$

$$p(x_{N+1} \mid \mathbf{x}, \alpha, \lambda) \approx \sum_{t=1}^{T} \mathrm{E}_q \left[\pi_t(\mathbf{V})\right] \mathrm{E}_q \left[p(x_{N+1} \mid \eta_t^*)\right]$$

Truncated Gibbs Sampling

1. For $n \in \{1, \ldots, N\}$, independently sample Z_n from

$$p(z_n = k \mid \mathbf{v}, \boldsymbol{\eta}^*, \mathbf{x}) = \pi_k(\mathbf{v}) p(x_n \mid \eta_k^*),$$

2. For $k \in \{1, \ldots, K\}$, independently sample V_k from $\text{Beta}(\gamma_{k,1}, \gamma_{k,2})$, where

$$\begin{aligned} \gamma_{k,1} &= 1 + \sum_{n=1}^{N} \mathbf{1}\left[z_n = k\right] \\ \gamma_{k,2} &= \alpha + \sum_{i=k+1}^{K} \sum_{n=1}^{N} \mathbf{1}\left[z_n = i\right]. \end{aligned}$$

This step follows from the conjugacy between the multinomial distribution and the truncated stick-breaking construction, which is a generalized Dirichlet distribution (Connor and Mosimann 1969).

3. For $k \in \{1, \ldots, K\}$, independently sample η_k^* from $p(\eta_k^* \mid \tau_k)$. This distribution is in the same family as the base distribution, with parameters

$$\begin{aligned} \tau_{k,1} &= \lambda_1 + \sum_{i \neq n} \mathbf{1}\left[z_i = k\right] x_i \\ \tau_{k,2} &= \lambda_2 + \sum_{i \neq n} \mathbf{1}\left[z_i = k\right]. \end{aligned} \tag{27}$$