Collapsed Segmented LDA (CSLDA):

$$p(S,Z,W|\alpha,\beta,\eta,\sigma) = \left[\prod_k \frac{\Gamma(W\beta)}{\Gamma(\beta)^W \cdot \Gamma(N_k + W\beta)} \prod_w \Gamma(N_{kw}+\beta)\right]$$

$$\times \left[\prod_d \mathcal{N}\left(s_d \,\middle|\, \eta^T \cdot \frac{N_{dk}}{N_d}, \sigma\right) \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K \cdot \Gamma(N_d + K\alpha)} \prod_k \Gamma(N_{dk}+\alpha)\right]$$

Where we used $\frac{N_{dk}}{N_d} \equiv \bar{Z}_d$

$$\log p(S,Z,W|\alpha,\beta,\eta,\sigma) =$$

$$\left[\sum_k \log\Gamma(W\beta) - W\log\Gamma(\beta) - \log\Gamma(N_k + W\beta) + \sum_w \log\Gamma(N_{kw}+\beta)\right]$$

$$+\left[\sum_d \underbrace{-\log\sigma - \frac{1}{2}\log(2\pi) - \frac{\left(s_d - \eta^T \cdot \frac{N_{dk}}{N_d}\right)^2}{2\sigma^2}}_{\text{Normal distribution}} + \right.$$

$$\left. \log\Gamma(K\alpha) - K\log\Gamma(\alpha) - \log\Gamma(N_d + K\alpha) + \sum_k \log\Gamma(N_{dk}+\alpha)\right]$$

Maximum a posteriori (MAP) estimate for the η hyperparameter:

$$\nabla_{\eta_k} \log p(S,Z,W|\alpha,\beta,\eta,\sigma) = \sum_d \frac{\frac{N_{dk}}{N_d}\left(s_d - \eta^T \frac{N_{d\cdot}}{N_d}\right)}{\sigma^2} =$$

$$= \sum_d \frac{s_d \frac{N_{dk}}{N_d}}{\sigma^2} - \sum_d \frac{\frac{N_{dk}}{N_d}\left(\eta^T \frac{N_{d\cdot}}{N_d}\right)}{\sigma^2} = 0$$

$$\Rightarrow \sum_d s_d \frac{N_{dk}}{N_d} = \sum_d \frac{N_{dk}}{N_d}\left(\sum_{k'} \eta_{k'} \frac{N_{dk'}}{N_d}\right) = \sum_d \frac{N_{dk}}{N_d}\left(\eta_k \frac{N_{dk}}{N_d} + \sum_{k' \neq k} \eta_{k'} \frac{N_{dk'}}{N_d}\right)$$

$$\Rightarrow \sum_d s_d \frac{N_{dk}}{N_d} = \eta_k \sum_d \left(\frac{N_{dk}}{N_d}\right)^2 + \sum_d \left(\frac{N_{dk}}{N_d} \sum_{k' \neq k} \eta_{k'} \frac{N_{dk'}}{N_d}\right)$$

$$\Rightarrow \sum_d \left(s_d \frac{N_{dk}}{N_d} - \frac{N_{dk}}{N_d} \sum_{k' \neq k} \eta_{k'} \frac{N_{dk'}}{N_d}\right) = \eta_k \sum_d \left(\frac{N_{dk}}{N_d}\right)^2$$

$$\Rightarrow \sum_d \frac{N_{dk}}{N_d}\left(s_d - \sum_{k' \neq k} \eta_{k'} \frac{N_{dk'}}{N_d}\right) = \eta_k \sum_d \left(\frac{N_{dk}}{N_d}\right)^2$$

$$\Rightarrow \eta_k = \frac{\sum_d \frac{N_{dk}}{N_d}\left(s_d - \sum_{k' \neq k} \eta_{k'} \frac{N_{dk'}}{N_d}\right)}{\sum_d \left(\frac{N_{dk}}{N_d}\right)^2}$$

Trying to apply the previous formula as an update rule for η does not converge. Instead, the following update can be used:

$$\eta_k^{new} \leftarrow (1-\gamma)\eta_k^{old} + \gamma \frac{\sum_d \frac{N_{dk}}{N_d}\left(s_d - \sum_{k' \neq k} \eta_{k'} \frac{N_{dk'}}{N_d}\right)}{\sum_d \left(\frac{N_{dk}}{N_d}\right)^2 + \epsilon}$$

With $1 \gg \gamma > 0$ in order for the previous series to converge and $1 \gg \epsilon > 0$ is a smoothing constant.

Gibbs sampler:

$$\begin{aligned} p(z_{di} = k | Z^{\setminus i}, S, W, \alpha, \beta, \eta, \sigma) &\propto p(z_{di} = k, Z_{-i}, S, W, \alpha, \beta, \eta, \sigma) \\ &\propto \left[\prod_{k'} \frac{\prod_w \Gamma(N_{k'w}^{\setminus i} + 1(k' = k \wedge w = w_{di}) + \beta)}{\Gamma(N_{k'}^{\setminus i} + 1(k' = k) + W\beta)} \right] \times \\ \mathcal{N}\left(s_d \,\middle|\, \eta^T \cdot \frac{N_{dk'}^{\setminus i} + 1(k' = k)}{N_d}, \sigma \right) &\prod_{k'} \Gamma(N_{dk'}^{\setminus i} + 1(k' = k) + \alpha) \end{aligned}$$

Dataset

- Number of movies ≈ 700
- Distribution of movie scores:

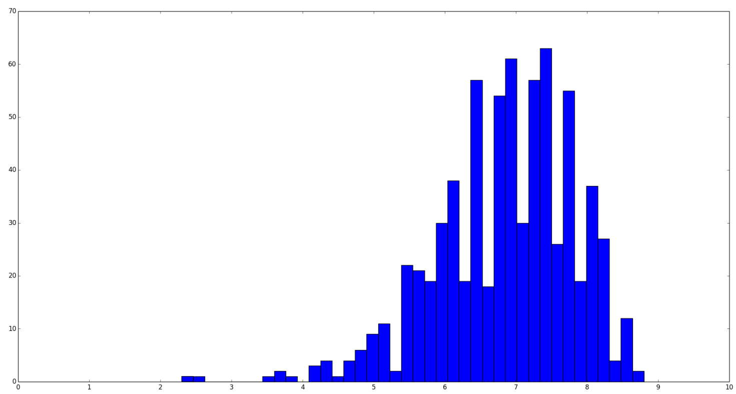

Figure 1: Moo

- Average number of tokens within a movie summary ≈ 75
- Average number of tokens within a movie script ≈ 18000

Results with 5 movies in the training set and 5 movies in the testing set (5 burn-in, 3 skip, 5 samples):

Perplexity values:

K		5	10	20
Using scores?	Yes	10059	7503	10938
	No	9297	10180	9663

Inverse accuracy values:

K		5	10	20
Using scores?	Yes	964	703	1133
	No	915	1143	960