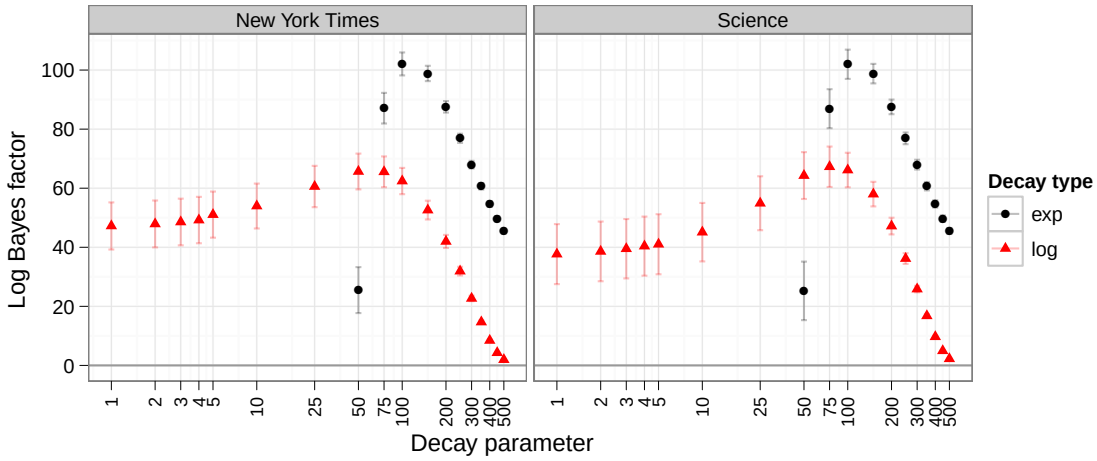

Figure 4: Bayes factors of the distance dependent CRP versus the traditional CRP on documents from *Science* and the *New York Times*. The black line at 0 denotes an equal fit between the traditional CRP and distance dependent CRP, while positive values denote a better fit for the distance dependent CRP. Also illustrated are standard errors across documents.

5.1 Language modeling

We evaluated the fully-observed distance dependent CRP models on two data sets: a collection of 100 OCR'ed documents from the journal *Science* and a collection of 100 world news articles from the *New York Times*. We modeled each document independently. We assess sampler convergence visually, examining the autocorrelation plots of the log likelihood of the state of the chain (Robert and Casella, 2004).

We compare models by estimating the Bayes factor, the ratio of the probability under the distance dependent CRP to the probability under the traditional CRP (Kass and Raftery, 1995). For a decay function f, this Bayes factor is

$$BF_{f,\alpha} = p(w_{1:N} \mid \text{dist-CRP}_{f,\alpha}) / p(w_{1:N} \mid \text{CRP}_{\alpha}). \tag{10}$$

A value greater than one indicates an improvement of the distance dependent CRP over the traditional CRP. Following Geyer and Thompson (1992), we estimate this ratio with a Monte Carlo estimate from posterior samples.

Figure 4 illustrates the average log Bayes factors across documents for various settings of the exponential and logistic decay functions. The logistic decay function always provides a better model than the traditional CRP; the exponential decay function provides a better model at certain settings of its parameter. (These curves are for the hierarchical setting with the base distribution over terms G_0 unobserved; the shapes of the curves are similar in the non-hierarchical settings.)