Supplement to "Principled Selection of Hyperparameters in the Latent Dirichlet Allocation Model"

Clint P. George[*]
Informatics Institute
University of Florida
clintpg@ufl.edu

Hani Doss[†]
Department of Statistics
University of Florida
doss@stat.ufl.edu

March 2, 2017

Abstract

This document provides details regarding some of the empirical results given in "Principled Selection of Hyperparameters in the Latent Dirichlet Allocation Model," by Clint P. George and Hani Doss.

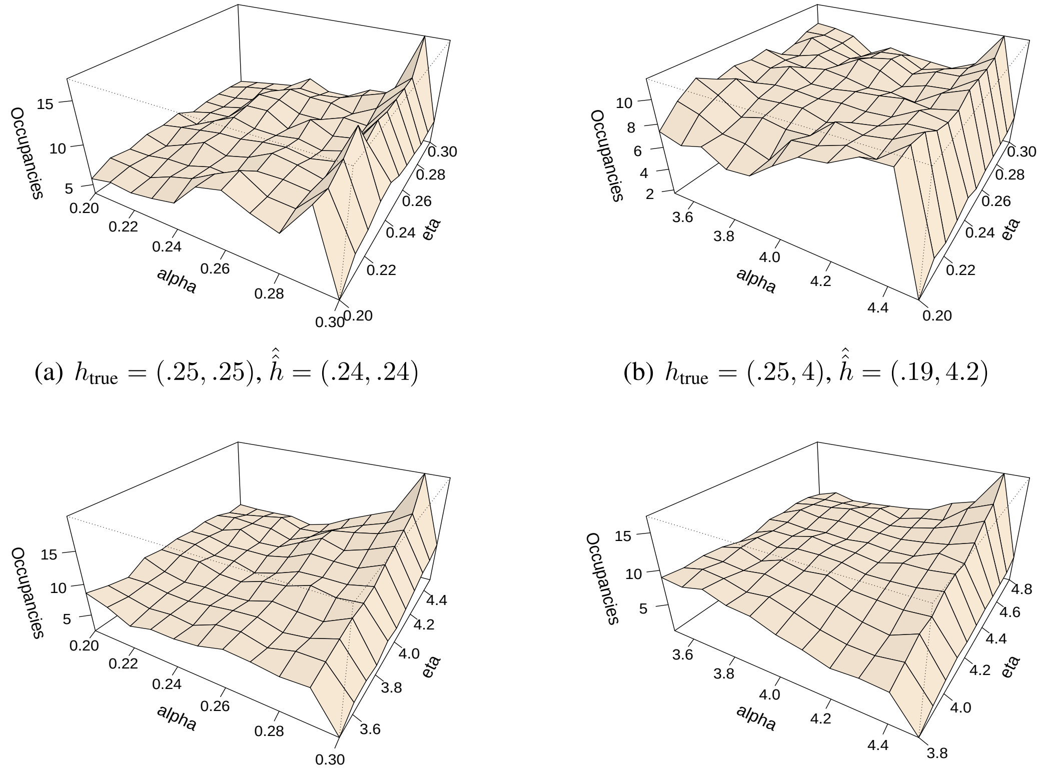
Throughout this document, figures and tables are labelled with the prefix "S"; so for example, the first figure is labelled Figure S-2. We do this in order to avoid confusion with the tables and figures of the main paper.

Occupancy Times for the Serial Tempering Chain Used to Create Figure 7

Recall that for the serial tempering chain to work well, it is necessary that the proportions of time spent in the different components of the mixture distribution be approximately equal, and the vector of these proportions is the main diagnostic for assessing convergence of the chain (Geyer, 2011). Figure S-1 provides plots that give the number of iterations that the serial tempering chain spent in each of the 121 components of the invariant distribution of the chain, for each of the four corpora used in Figure 7. The plots show that the distributions of the occupancy times are acceptably close to the uniform.

[*] Research supported by the International Center for Automated Research at the UF Levin College of Law

[†] Research supported by NSF Grant DMS-11-06395 and NIH grant P30 AG028740



(a) $h_{\text{true}} = (.25, .25)$, $\hat{\hat{h}} = (.24, .24)$

(b) $h_{\text{true}} = (.25, 4)$, $\hat{\hat{h}} = (.19, 4.2)$

(c) $h_{\text{true}} = (4, .25)$, $\hat{\hat{h}} = (4.2, .27)$

(d) $h_{\text{true}} = (4, 4)$, $\hat{\hat{h}} = (4.9, 4.2)$

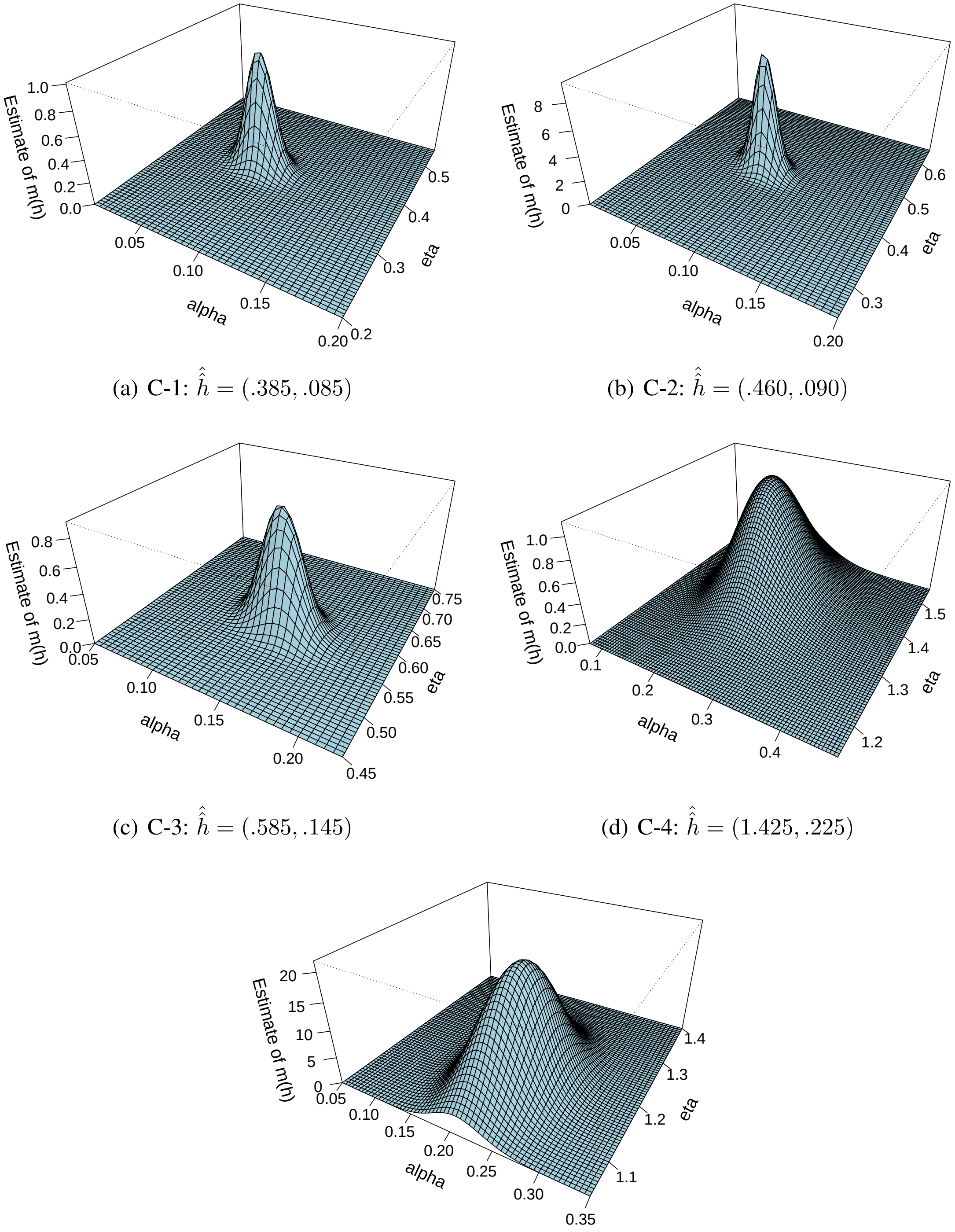
Figure S-1: Number of iterations (in units of 100) that the final serial tempering chain spent at each of the hyperparameter values $h_1, \dots, h_{121}$ in the subgrid, for the four synthetic corpora used in Figure 7.

Table 3: Details Regarding Computation of $\widehat{M_\zeta}(h)$ and its Argmax, and Convergence Diagnostics

For each of the nine corpora in Table 3, we computed (i) the estimate $\widehat{M_\zeta}(h)$ for h over a grid, using the method described in Section 2, and (ii) an estimate of the standard error of $\widehat{M_\zeta}(h)$ for each h in the grid. Figures S-2 and S-4 show plots of $\widehat{M_\zeta}(h)$, and also give $\hat{\hat{h}} = \arg\max_h \widehat{M_\zeta}(h)$, for the nine corpora. Figures S-3 and S-5 give plots of the standard errors of $\widehat{M_\zeta}(h)$ for the nine corpora, and these indicate that the accuracy of $\widehat{M_\zeta}(h)$ is acceptable over the entire h-range for all nine cases.

Figure S-6 gives the distributions of the occupancy times for corpora C-1 and C-9 (these are the least complex and most complex, respectively). The figures show that these distributions are acceptably close to the uniform. Plots for the other corpora (not shown) are similar.

The actual values of $\hat{\hat{h}}_{\text{VEM}}$, $\hat{\hat{h}}_{\text{GEM}}$, and $\hat{\hat{h}}_{\text{ST}}$ for the nine corpora are given in Table S-1.



(a) C-1: $\hat{\hat{h}} = (.385, .085)$

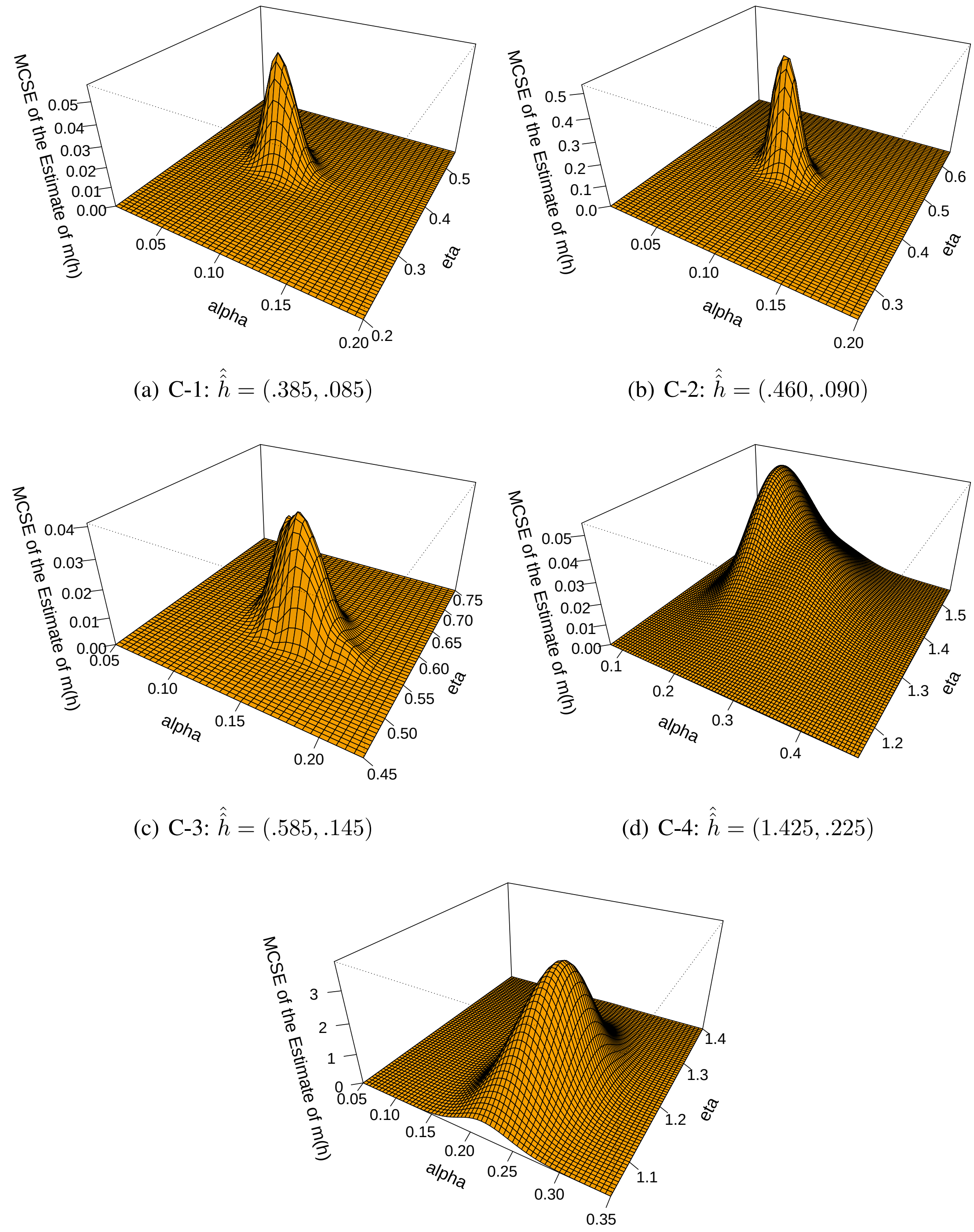
(b) C-2: $\hat{\hat{h}} = (.460, .090)$

(c) C-3: $\hat{\hat{h}} = (.585, .145)$

(d) C-4: $\hat{\hat{h}} = (1.425, .225)$

(e) C-5: $\hat{\hat{h}} = (1.165, .225)$

Figure S-2: Plots of $\widehat{M_\zeta}(h)$ for the five 20Newsgroups corpora.



(a) C-1: $\hat{\hat{h}} = (.385, .085)$

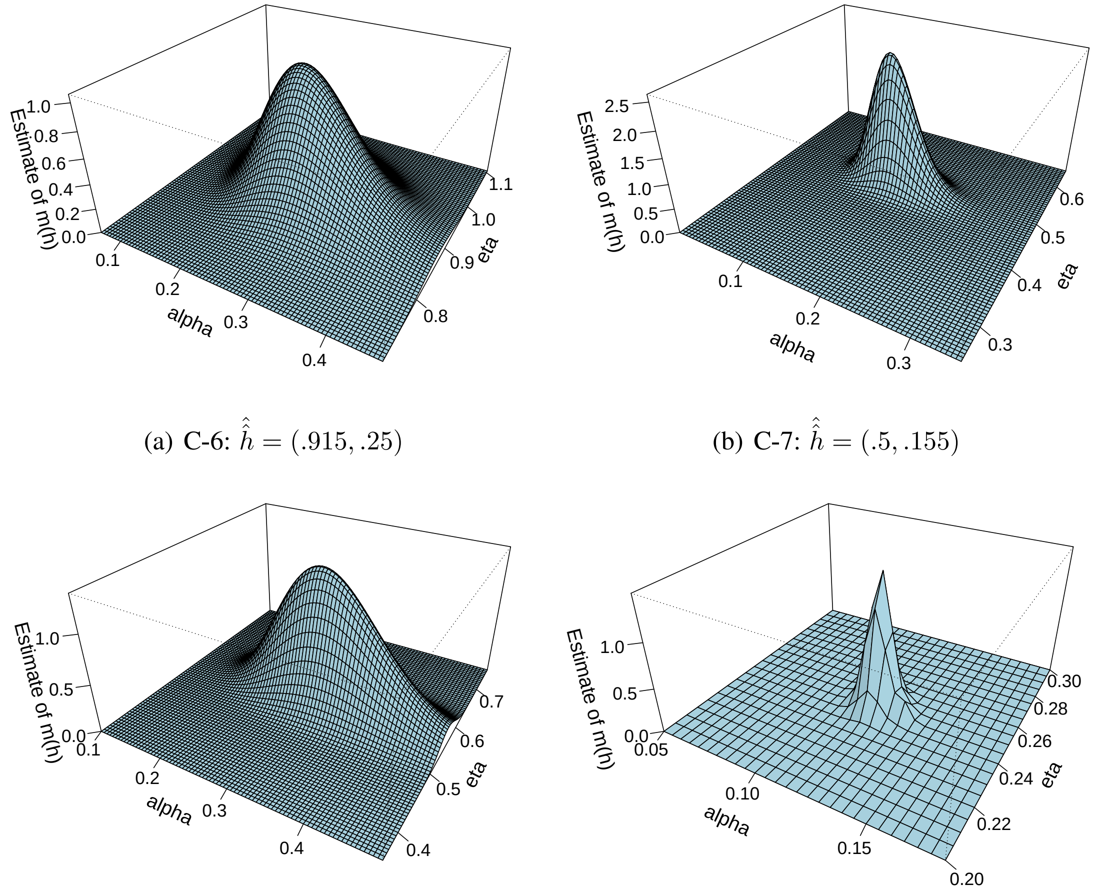
(b) C-2: $\hat{\hat{h}} = (.460, .090)$

(c) C-3: $\hat{\hat{h}} = (.585, .145)$

(d) C-4: $\hat{\hat{h}} = (1.425, .225)$

(e) C-5: $\hat{\hat{h}} = (1.165, .225)$

Figure S-3: Monte Carlo standard error (MCSE) of $\widehat{M_\zeta}(h)$ for the five 20Newsgroups corpora.



(a) C-6: $\hat{\hat{h}} = (.915, .25)$

(b) C-7: $\hat{\hat{h}} = (.5, .155)$

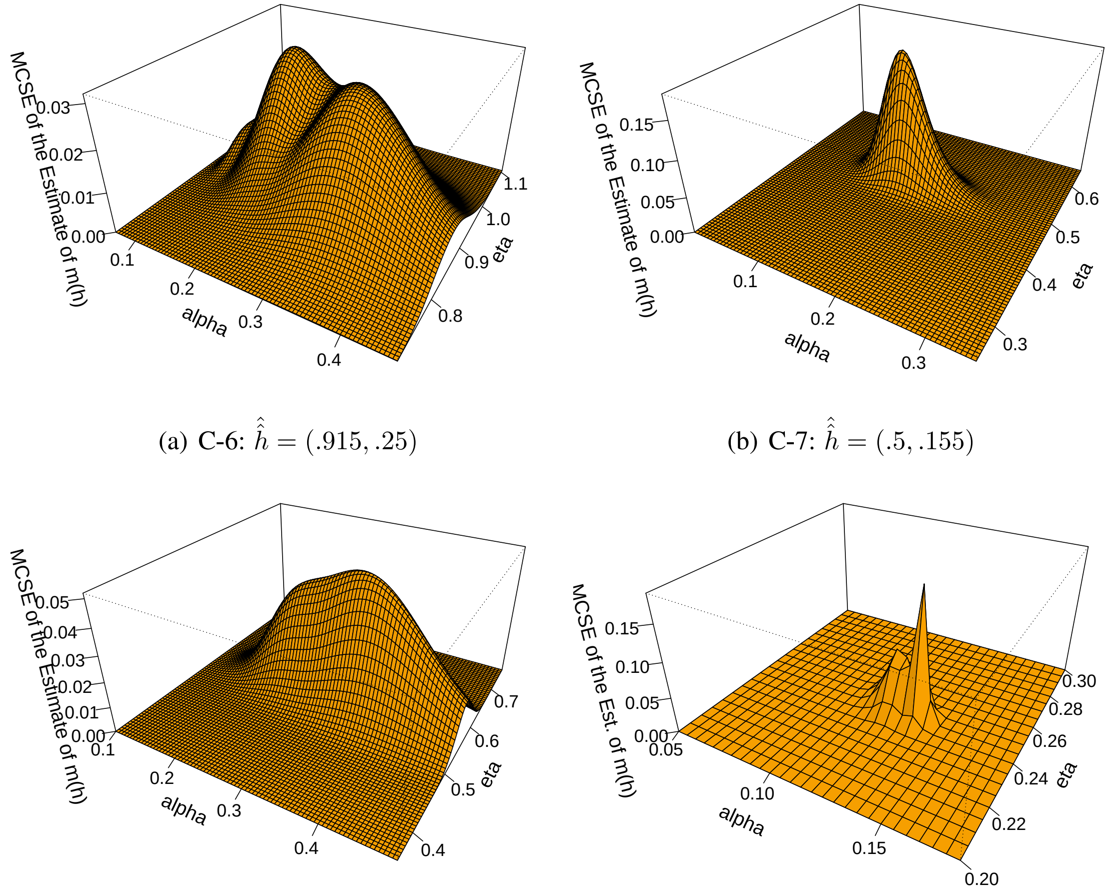
(c) C-8: $\hat{\hat{h}} = (.57, .31)$

(d) C-9: $\hat{\hat{h}} = (.250, .120)$

Figure S-4: Plots of $\widehat{M_\zeta}(h)$ for corpora C-6, C-7, C-8, and C-9.

Corpus	$\hat{\hat{h}}_{\text{VEM}}$	$\hat{\hat{h}}_{\text{GEM}}$	$\hat{\hat{h}}_{\text{ST}}$
C-1	$(.378, .037)$	$(.386, .082)$	$(.385, .085)$
C-2	$(.593, .038)$	$(.436, .117)$	$(.460, .090)$
C-3	$(.862, .021)$	$(.586, .146)$	$(.585, .145)$
C-4	$(1.722, .037)$	$(1.425, .229)$	$(1.425, .225)$
C-5	$(1.408, .044)$	$(1.106, .280)$	$(1.165, .225)$
C-6	$(1.147, .210)$	$(.915, .253)$	$(.915, .250)$
C-7	$(.710, .056)$	$(.445, .181)$	$(.500, .155)$
C-8	$(.841, .162)$	$(.570, .315)$	$(.570, .310)$
C-9	$(.170, .120)$	$(.235, .130)$	$(.250, .120)$

Table S-1: Values of $\hat{\hat{h}}_{\text{VEM}}$, $\hat{\hat{h}}_{\text{GEM}}$, and $\hat{\hat{h}}_{\text{ST}}$ for all nine corpora. The GEM and ST estimates are fairly similar.



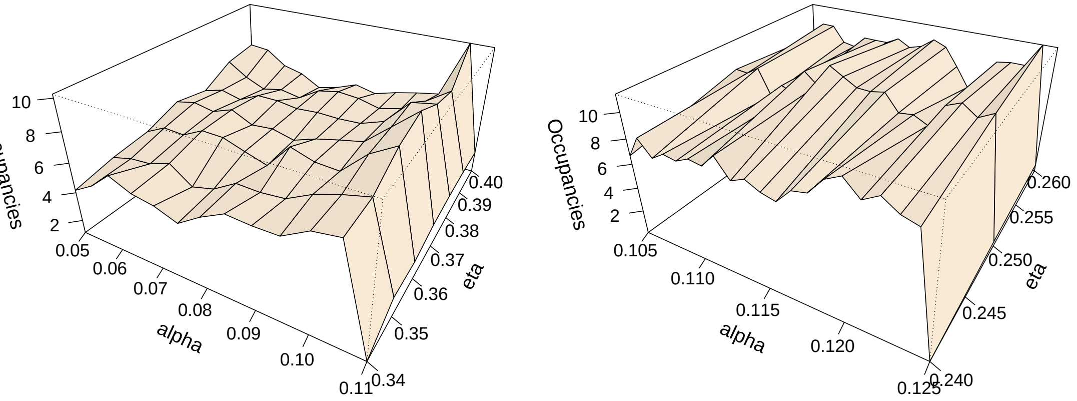
(a) C-6: $\hat{\hat{h}} = (.915, .25)$

(b) C-7: $\hat{\hat{h}} = (.5, .155)$

(c) C-8: $\hat{\hat{h}} = (.57, .31)$

(d) C-9: $\hat{\hat{h}} = (.250, .120)$

Figure S-5: Monte Carlo standard error (MCSE) of $\widehat{M_\zeta}(h)$ for corpora C-6, C-7, C-8, and C-9.



(a) C-1

(b) C-9

Figure S-6: Plots of the number of iterations (in units of 100) that the final serial tempering chain spent at each of the hyperparameter values $h_1, \dots, h_J$ in the subgrid, for corpora C-1 and C-9.

References

Geyer, C. J. (2011). Importance sampling, simulated tempering, and umbrella sampling. In *Handbook of Markov Chain Monte Carlo* (S. P. Brooks, A. E. Gelman, G. L. Jones and X. L. Meng, eds.). Chapman & Hall/CRC, Boca Raton, 295–311.