

# Supplementary material: Relational learning and network modelling using infinite latent attribute models

March 31, 2014

## 1 Joint distribution tests

We give evidence for the correctness of our algorithm using the joint distribution testing methodology of Geweke [2004]. There are two ways to sample from the joint distribution,  $P(R, \theta)$  over parameters,  $\theta = \{\mathbf{Z}, \mathbf{C}, \mathbf{W}, \alpha, \gamma, \sigma\}$  and data,  $R$  defined by a probabilistic model such as ILA. The first we will refer to as “marginal-conditional” sampling, shown in Algorithm 1. Both steps here are straightforward: sampling from the prior followed by sampling from the likelihood model. The second way, referred to as “successive-conditional” sampling, is shown in Algorithm 2, where  $Q$  represents a single (or multiple) iteration(s) of our MCMC sampler. To validate our sampler we can then check, either informally or using hypothesis tests, whether the samples drawn from the joint  $P(R, \theta)$  in these two different ways appear to have come from the same distribution.

<b>Algorithm 1</b> Marginal conditional	<b>Algorithm 2</b> Successive conditional
1: <b>for</b> $m = 1$ to $M$ <b>do</b>	1: $\theta^{(1)} \sim P(\theta)$
2: $\theta^{(m)} \sim P(\theta)$	2: $R^{(1)} \sim P(R \theta^{(1)})$
3: $R^{(m)} \sim P(R \theta^{(m)})$	3: <b>for</b> $m = 2$ to $M$ <b>do</b>
4: <b>end for</b>	4: $\theta^{(m)} \sim Q(\theta \theta^{(m-1)}, R^{(m-1)})$
	5: $R^{(m)} \sim P(R \theta^{(m)})$
	6: <b>end for</b>

We apply this method to our ILA sampler with  $N = 10$ . We draw  $10^4$  samples using both the marginal-conditional and successive-conditional procedures. We look at various characteristics of the samples, including the number of features and the  $\alpha$ ,  $\gamma$  and bias hyperparameters. The distribution of the number of features under the successive-conditional sampler matches that under the marginal-conditional sampler almost perfectly as shown in Figure 1.

Both the histogram and the quantile-quantile plot show the similarity of the two distributions, with the straight line in the later indicating an almost perfect match. The deviation from a straight line in the upper corner of the qq-plot is a result of there being fewer samples available to estimate these quantiles accurately. Under the successive-conditional sampler the average number of features is 5.86 while under the marginal-conditional is 5.85 with standard deviations 4.83 and 4.86 respectively: a hypothesis test did not reject the null hypothesis that the means of the two distributions are equal. While this cannot completely guarantee correctness of the algorithm and code,  $10^4$  samples is a large number for such a small model and thus provides strong evidence that our algorithm is correct.

## 2 More trace and autocorrelation plots

The figures included here supplement Section 8.1 of the main paper. Figure 2 shows autocorrelation plots for ILA on the synthetic  $N = 90$  dataset, for the two chains in Figure 13 of the main paper, whereas Figure 3 shows autocorrelations for ILA on the NIPS dataset, corresponding to the chain in Figure 14 of the main paper. Figures 4 and 5 show traceplots and autocorrelation plots for single runs of the IRM and LFRM respectively, both on the NIPS dataset.

## 3 Raftery and Lewis convergence diagnostic

This convergence diagnostic proposed by Raftery and Lewis [1992] calculates the number of iterations required to estimate a particular quantile of the posterior distribution (with respect to some parameter of interest) to within  $\pm r$  of the true value,  $q_{\text{true}}$ , with probability,  $s$ . In other words, we wish to ensure that

$$Pr(q_{\text{true}} - r < q_{\text{infer}} < q_{\text{true}} + r) > s$$

where  $q_{\text{infer}}$  is the quantile estimated from our MCMC samples. The diagnostic estimates both the number of iterations,  $T$  and the number of burn-in iterations,  $S$ , necessary to satisfy this condition. It also provides a dependence factor interpreted as the proportional increase in the number of iterations needed to reach convergence taking into account dependence between the samples in the chain.

For the two chains presented in Figure 13 of the main paper ( $N = 90$  synthetic dataset, ILA logistic), the values of this diagnostic for  $q_{\text{true}} = 0.025$ ,  $r = 0.005$  and  $s = 0.95$  are shown in Table 1. The small burn-in values reported suggest that both chains find a posterior mode almost immediately for each monitored parameter. For the  $\gamma$  in the first chain, the first 24 iterations should be discarded, and an estimated 33,000 additional iterations performed in order to estimate the 2.5% percentile of the posterior distribution to within 0.005 accuracy and 95% confidence. For the remaining variables and chains, 33,000

Table 1: Raftery and Lewis Diagnostic results for the first chain of Figure 13 of the main paper (ILA logistic,  $N = 90$  synthetic dataset). A large number of iterations are required to accurately estimate the posterior over the CRP hyperparameter  $\gamma$ .

parameter	q	r	s	Burn-in (S)	Total (T)	Dependence factor
Chain 1						
$\alpha$	0.025	0.005	0.95	3	4235	1.1305
$\gamma$				24	33214	8.8665
bias				4	4983	1.3302
$\alpha$	0.025	0.01	0.95	3	1061	1.1323
$\gamma$				24	8324	8.8837
bias				4	1249	1.3330
Chain 2						
$\alpha$	0.025	0.005	0.95	3	4322	1.1538
$\gamma$				14	14543	3.8823
bias				5	5763	1.5384
$\alpha$	0.025	0.01	0.95	3	1083	1.1558
$\gamma$				14	3647	3.8922
bias				5	1445	1.5422

iterations are more than sufficient to estimate this quantile to the required precision, with only the first three to four samples needing to be discarded. The diagnostics are similar for the second chain. However, if we relax the precision required for the estimate of this quantile to  $\pm 0.01$  the number of iterations needed decreases to under 9000 (excluding burn-in). The dependence factors are reasonable ( $< 5$ ) with only the dependence factor of the  $\gamma$  parameter having a value close to 9. While this suggests a large number of iterations are required if such stringent estimation conditions are to be met, we believe this is a result of the distinct isolated posterior modes present for the synthetic example, which are probably not representative of more complex real world networks.

This intuition is supported by our results for the NIPS dataset, for which the Raftery and Lewis metrics for all three models are presented in Table 2. ILA and LFRM require more iterations in order to be accurate (around 9000 and 7000 respectively), compared to IRM (around 3000) a conclusion that agrees with the discussion in Section 8.1 of the main paper. However, these numbers are much smaller than for the synthetic dataset.

## 4 Gelman and Rubin diagnostic

Five chains were used for this convergence test, each starting from different initial values sampled from the prior, making them over-dispersed with respect to the posterior distribution. We want to test whether all five chains converge

Table 2: Raftery and Lewis Diagnostic results for IRM, LFRM and ILA on the NIPS dataset. LFRM and ILA require a similar number of iterations, considerably more than IRM for the same level of accuracy.

model	parameter	q	r	s	Burn-in (S)	Total (T)	Dependence factor
ILA	$\alpha$	0.025	0.01	0.95	5	1473	1.5720
	$\gamma$				10	2588	2.7620
	bias				34	8790	9.3810
IRM	$\alpha$	0.025	0.01	0.95	14	3310	3.5326
	$\beta_1$				6	1757	1.8751
	$\beta_2$				6	1608	1.7161
LFRM	$\alpha$	0.025	0.01	0.95	2	893	0.9530
	bias				25	6703	7.1537

Table 3: Gelman diagnostics for five ILA logistic MCMC chains run on the  $N = 90$  synthetic dataset. These values being close to 1 suggests all the chains converge to the same target distribution.

parameter	potential scale factor
$\alpha$	1.0021
$\gamma$	1.0059
bias	1.0007

to the same target distribution. Failure could indicate the presence of a multimodal posterior distribution (different chains converge to different local modes) or the need to run a longer chain. The test proposed by Gelman and Rubin [1992], is based on a comparison of the within and between chain variances for each parameter to give the *potential scale reduction factor*. A large factor ( $\gg 1$ ) indicates that the between-chain variance is substantially greater than the within-chain variance, so that longer simulation is needed. We ran our model on a synthetic dataset with  $N = 90$  nodes for five different seeds. The scale factor is given in the Table 3. All the potential scale factors are close to one and indicate that each of the chains has stabilized, and they are likely to have converged to the same mode of the target distribution. A similar conclusion is drawn for the NIPS dataset as shown in Table 4.

## References

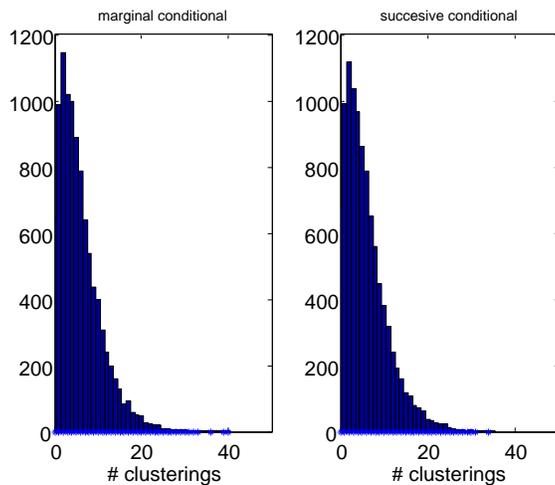
- A. Gelman and D.B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7:457–511, 1992.

Table 4: Gelman diagnostics for five chains of ILA, IRM and LFRM on the NIPS dataset. In all cases the potential scale factors are not very large, although the values for ILA and LFRM are somewhat larger than for IRM, suggesting the latter is mixing a little better.

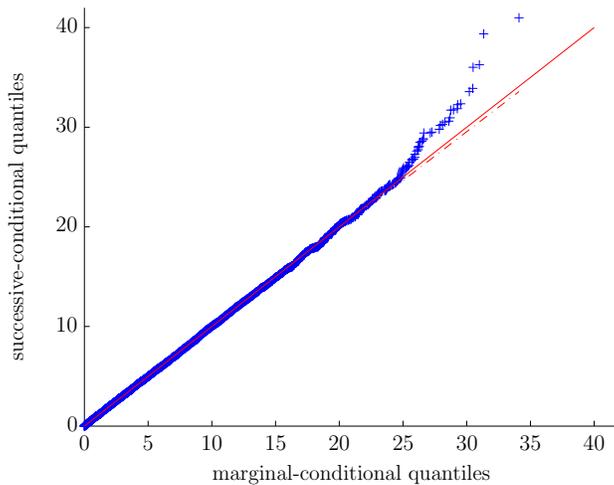
model	parameter	potential scale factor
ILA	$\alpha$	1.4003
	$\gamma$	1.4151
	bias	1.1769
IRM	$\alpha$	1.0909
	$\beta_1$	1.0946
	$\beta_2$	1.2302
LFRM	$\alpha$	1.0266
	bias	1.4539

J. Geweke. Getting it right. *JASA*, 99(467):799–804, 2004.

Adrian E. Raftery and Steven Lewis. How many iterations in the Gibbs sampler?  
 In *Bayesian Statistics 4*. Oxford University Press, 1992.



(a) Empirical distributions on the number of features



(b) qq-plot of the two empirical distributions.

Figure 1: Comparing the distribution of the number of features under the marginal-conditional and successive-conditional samplers. Figure (a) shows the empirical distribution over the number of clusterings (features) under the marginal-conditional and successive-conditional sampler respectively. Figure (b) shows the qq-plot of the two empirical distributions. The agreement of the two distributions is evidence for the correctness of our MCMC sampler for ILA.

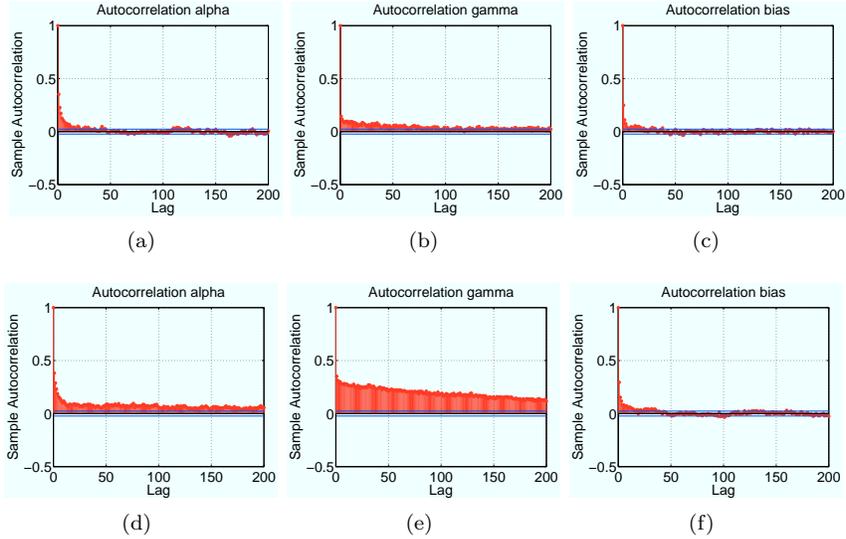


Figure 2: Autocorrelation plots for the ILA parameters  $\alpha$ ,  $\gamma$  and bias, for synthetic dataset with  $N = 90$ . The two rows correspond to the same two runs as Figure 13 of the main paper. The autocorrelations generally decay rapidly apart from for  $\gamma$  in the second run, which is a consequent of that chain having explored two modes.

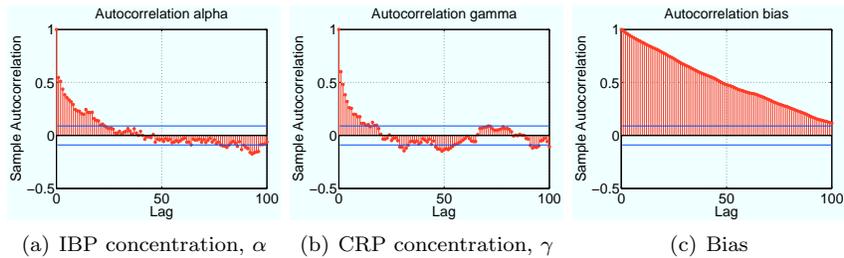


Figure 3: Autocorrelations plots of the  $\alpha$ ,  $\gamma$  and bias parameters of ILA for the NIPS dataset ( $N = 234$ ). On this more complex real world dataset distinct modes are not apparent but there is significant autocorrelation, particularly for the bias parameter, for up to 100 iterations.

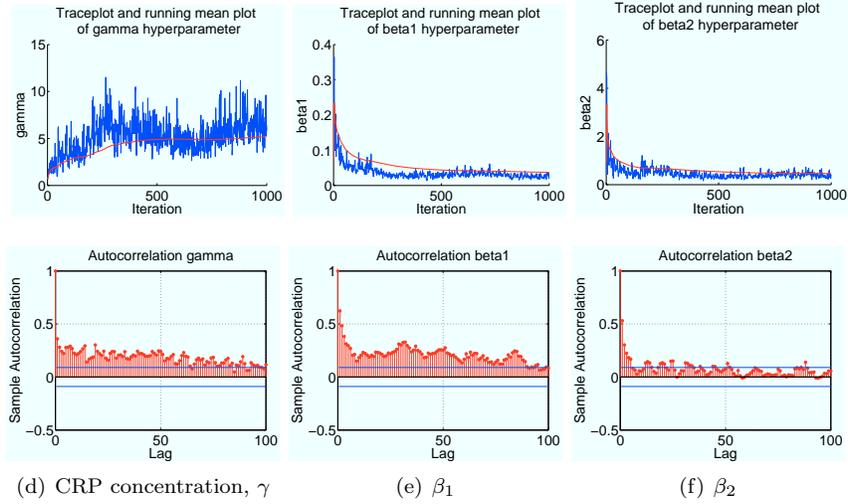


Figure 4: Traceplots and autocorrelation plots of the  $\gamma$  and weight hyperparameters  $\beta_1$  and  $\beta_2$  of IRM for the NIPS dataset. For IRM burn-in is slower than for ILA or LFRM but the autocorrelations decay rapidly.

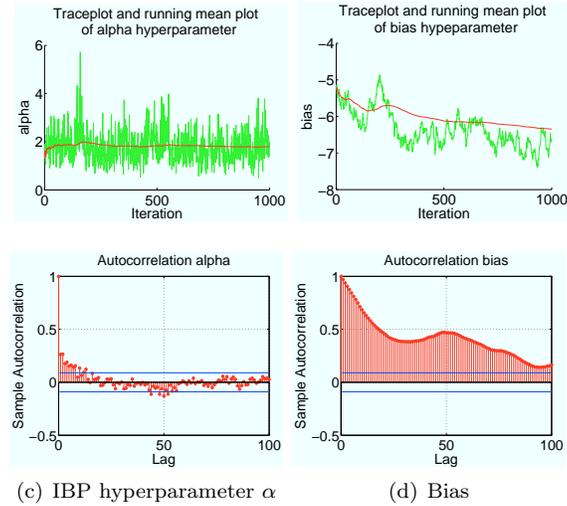


Figure 5: Traceplots and autocorrelation plots of the  $\alpha$  and bias parameters of LFRM for the NIPS dataset. As for ILA the bias parameter has significant autocorrelation up to a delay of around 100.