MLPG Open Assessment COM00032H

Exam No: Y3843100

March 19, 2020

1 Conditional independence in Bayesian networks

Independent pairs

$$\{(A,C),(A,E),(A,F),(B,C),(B,E),(B,F),(D,C),(D,E),$$

$$(D,F),(H,A),(H,B),(H,C),(H,D),(H,E),(H,F),(H,G)\}$$

Independent pairs conditioned on $Z = \{C, G\}$ There are no *conditionally* independent pairs yet

$$\{(H,A),(H,B),(H,D),(H,E),(H,F)\}$$

still remain unconditionally independent

Markov equivalent DAG Reverse edges but maintain immoralities, i.e. change B edges *fig 1*

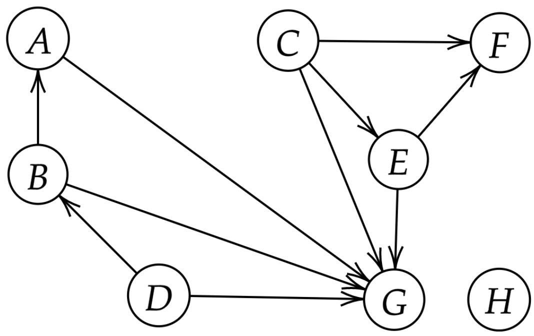


Figure 1: Question 1.3 Markov equivalent DAG

Non-Markov equivalent DAG Change immoralities. I.e. reverse all edges from G *fig 2*

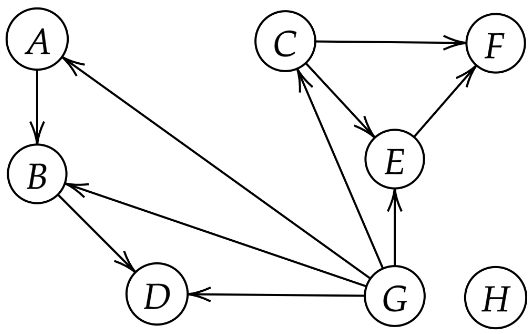


Figure 2: Question 1.4 Non-Markov equivalent DAG

2 House prices with STAN

2.1 A simple model (figs 4 and 3)

We can state that our estimated posteriors approximate the true distributions as all chains have converged $\hat{R} = 1$, which is also observed by the beta plots. The preference for MCMC sampling over variational inference was due to the fact that the size of our dataset and the length of this assessment permits the usage of the more computationally intensive method. In addition, the asymptotic correctness of the posterior justifies the larger computational expense [1]. Empirical tests show that the STAN default of 1000 iterations (2000 samples) and 4 chains is enough for convergence.

2.2 A less simple model (figs. 6 and 5)

Denoting that size has a positive effect on price does not affect performance. This is potentially due to the fact that the data already embodies this fact and explicitly stating it does not give us any new knowledge.

2.3 Two models (figs. 7,8)

Houses in 0 get cheaper with age, which was obfuscated in 2.2. Splitting also reduces noise. The higher certainty in our split models is also reflected by the superior `lp__`.

2.4 A compromise model (figs. 9, 10 11)

We create a hierarchical model [2] that has variant slopes, but maintains the intercept across locales $y_i = \alpha + \beta_{j[i]} x_i + \varepsilon_i$ with all $\varepsilon \sim N(0, \sigma^2)$ [3]. With it we are able to capture the Age difference and still share some global knowledge via the intercept.

3 VI vs MCMC

HMC posterior approximations start with a random set of parameters θ which are updated over a number of iterations according to a randomly sampled momentum vector ρ. The update follows a Hamiltonian process where $H(\rho, \theta) = -\log p(\rho \mid \theta) - \log p(\theta)$. The new state (ρ^*, θ^*) is evaluated against a Metropolis acceptance test $min(1, \exp(H(\rho, \theta) - H(\rho^*, \theta^*)))$. If not accepted, the previous θ is recovered and a new random sample is drawn for the next iteration [4]. STAN's VI approximates posteriors by using automatic differentiation (AD) to transform the variables into a co-ordinate space where it asserts that the new transformed density is Gaussian. Then it uses the transformed support to maximise the evidence lower bound (ELBO, which is in essence the analytical form of KL divergence between our approximation and the true distribution) [5]. The maximisation is done via gradient ascend. VI is faster and can be used on large datasets but specifically for STAN if a Gaussian mixture model is not appropriate for our problem, VI will struggle. MCMC is asymptotically precise yet is computationally intensive and multimodal distributions can confuse it.

4 Hidden Markov models

Appendices

A Tables and Plots

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
alpha	-23.43	0.11	3.51	-30.6	-25.63	-23.4	-21.21	-16.6	948	1.0
beta_L	58.26	0.07	2.27	53.82	56.75	58.28	59.75	62.65	1035	1.0
beta_A	0.1	8.2e-4	0.03	0.04	0.08	0.1	0.12	0.16	1253	1.0
beta_S	0.67	6.7e-4	0.02	0.63	0.66	0.67	0.68	0.71	1038	1.0
sigma	9.52	0.02	0.73	8.24	9.01	9.46	9.98	10.99	1618	1.0
lp__	-244.6	0.06	1.66	-248.7	-245.5	-244.3	-243.4	-242.5	717	1.0

Figure 3: Question 2.1 posterior table summary

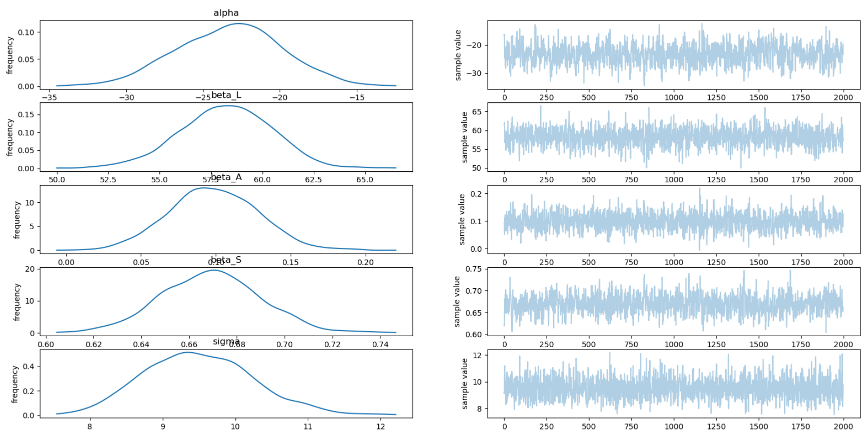


Figure 4: Question 2.1 plot summary

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
alpha	-23.09	0.11	3.43	-29.71	-25.33	-23.07	-20.8	-16.37	1020	1.0
beta_L	58.27	0.06	2.29	53.61	56.77	58.26	59.83	62.66	1245	1.0
beta_A	0.1	8.0e-4	0.03	0.04	0.08	0.1	0.12	0.16	1288	1.0
beta_S	0.67	6.4e-4	0.02	0.63	0.65	0.67	0.68	0.71	1065	1.0
sigma	9.48	0.02	0.7	8.17	8.99	9.43	9.92	10.95	1632	1.0
lp__	-245.0	0.06	1.6	-248.8	-245.8	-244.7	-243.8	-242.9	775	1.0

Figure 5: Question 2.2 posterior table summary

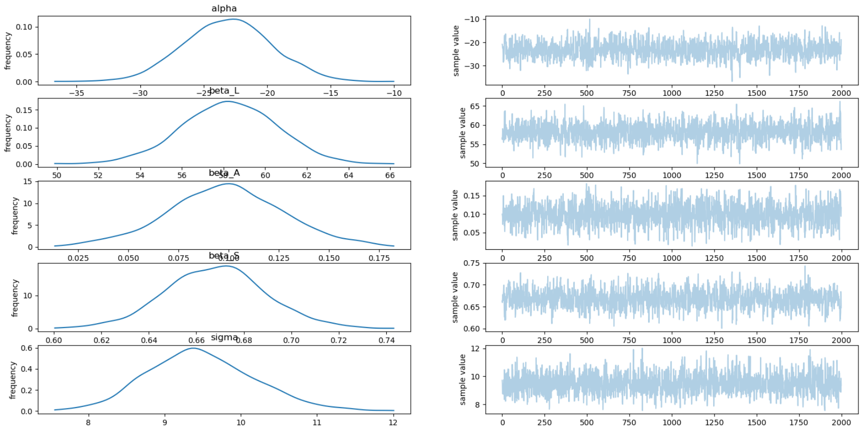
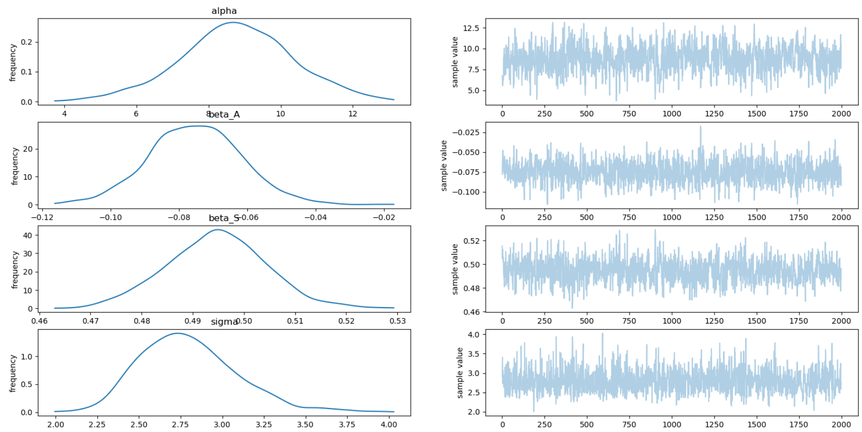


Figure 6: Question 2.2 plot summary

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
alpha	8.75	0.05	1.48	5.78	7.8	8.78	9.71	11.61	874	1.0
beta_A	-0.07	3.7e-4	0.01	-0.1	-0.08	-0.07	-0.07	-0.05	1238	1.0
beta_S	0.49	3.0e-4	9.6e-3	0.47	0.49	0.49	0.5	0.51	1002	1.0
sigma	2.82	9.2e-3	0.31	2.29	2.6	2.79	3.0	3.5	1146	1.0
lp__	-75.57	0.06	1.44	-79.08	-76.28	-75.24	-74.5	-73.79	667	1.0

(a) posterior table summary

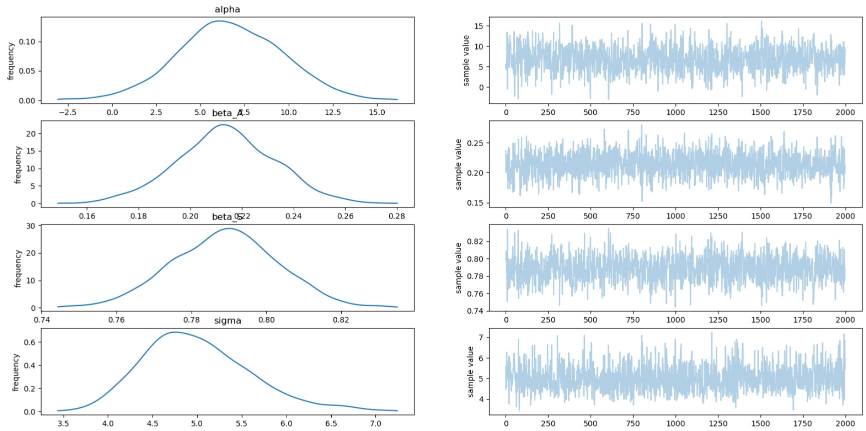


(b) plot summary

Figure 7: Question 2.3 Locale 0

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
alpha	6.62	0.11	2.97	0.84	4.67	6.5	8.52	12.78	787	1.0
beta_A	0.21	5.9e-4	0.02	0.18	0.2	0.21	0.23	0.25	1133	1.0
beta_S	0.79	4.8e-4	0.01	0.76	0.78	0.79	0.8	0.82	896	1.0
sigma	5.0	0.02	0.6	4.0	4.58	4.93	5.38	6.31	1447	1.0
lp__	-82.27	0.06	1.49	-86.0	-83.02	-81.92	-81.14	-80.4	731	1.0

(a) posterior table summary



(b) plot summary

Figure 8: Question 2.3 Locale 1

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
alpha	7.79	0.05	1.58	4.73	6.69	7.84	8.83	10.94	905	1.0
beta_A[1]	-0.07	4.3e-4	0.02	-0.1	-0.08	-0.07	-0.06	-0.04	1653	1.0
beta_A[2]	0.21	3.5e-4	0.01	0.18	0.2	0.21	0.22	0.24	1531	1.0
beta_S[1]	0.5	3.5e-4	0.01	0.48	0.49	0.5	0.51	0.52	1112	1.0
beta_S[2]	0.78	2.8e-4	9.0e-3	0.77	0.78	0.78	0.79	0.8	1033	1.0
sigma_beta_A	2.29	1.28	33.77	0.1	0.19	0.33	0.68	7.13	701	1.0
sigma_beta_S	15.39	7.2	189.6	0.43	0.83	1.45	3.29	52.89	694	1.01
sigma_y	3.86	7.2e-3	0.3	3.33	3.65	3.84	4.05	4.52	1703	1.0
y_hat[1]	95.94	0.02	1.04	93.83	95.28	95.95	96.63	97.94	1742	1.0
y_hat[...]										
lp__	-167.3	0.09	2.31	-172.9	-168.6	-166.9	-165.6	-163.9	620	1.0

Figure 9: Question 2.4 posterior table summary

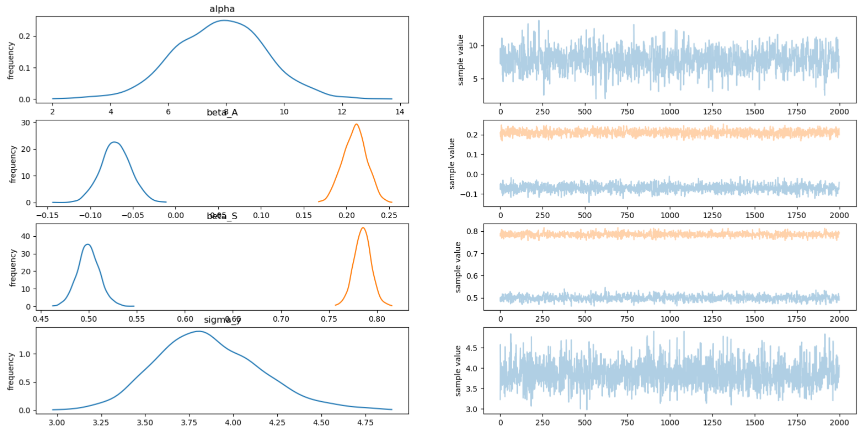


Figure 10: Question 2.4 plot summary

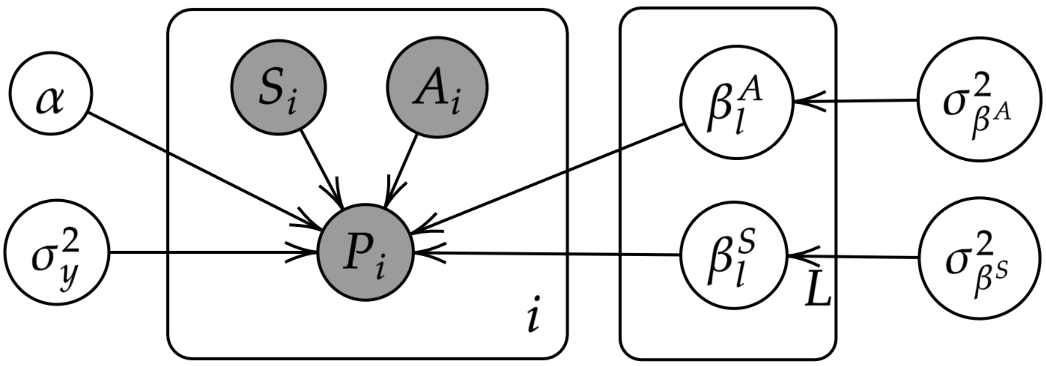


Figure 11: Question 2.4 model as Bayesian network

References

[1] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 859–877, 2017. DOI: 10.1080/01621459.2017.1285773. eprint: https://doi.org/10.1080/01621459.2017.1285773. [Online]. Available: https://doi.org/10.1080/01621459.2017.1285773.

[2] A. Gelman, "Multilevel (hierarchical) modeling: What it can and cannot do," *Technometrics*, vol. 48, no. 3, pp. 432–435, 2006. DOI: 10.1198/004017005000000661. eprint: https://doi.org/10.1198/004017005000000661. [Online]. Available: https://doi.org/10.1198/004017005000000661.

[3] *Stan case studies: Radon*, https://mc-stan.org/users/documentation/case-studies/radon.html, Accessed: 17-03-2020.

[4] M. Betancourt and M. Girolami, "Hamiltonian monte carlo for hierarchical models," *Current trends in Bayesian methodology with applications*, vol. 79, no. 30, pp. 2–4, 2015.

[5] A. Kucukelbir, R. Ranganath, A. Gelman, and D. Blei, "Automatic variational inference in stan," in *Advances in neural information processing systems*, 2015, pp. 568–576.