Statistical data analysis 2, 2024/2025, project 1

Introduction

Gibbs sampling can be applied when the joint distribution is not known explicitly, but the conditional distribution of each variable is known. The Gibbs sampling algorithm is used to generate an instance from the distribution of each variable in turn, conditional on the current values of the other variables. Gibbs sampling is particularly well-adapted to sampling the posterior distribution of a Bayesian network, since Bayesian networks are typically specified as a collection of conditional distributions, the *Local Probability Distributions* (LPDs).

The Gibbs Sampler

Let X_i be binary random variables with finite state space $\{T, F\}$ for *true* and *false*. A Gibbs sampler now runs a Markov chain on $X = (X_1, ..., X_n)$. For convenience of notation, we denote the set $(X_1, ..., X_{i-1}, X_{i+1}, ..., X_n)$ as $X_{\setminus i}$ and evidence $\mathbf{e} = (e_1, ..., e_m)$. The evidence variables represent the observed values. For the example in Figure 1, $X =$ (Extra Government Funding, Economic Environment) and $e =$ (Stable Job = T, Mortgage Loan = T). Then, the following method gives one possible way of creating a Gibbs sampler:

1. Initialise:

   (a) Instantiate X_i to one of its possible values x_i, $1 \leq i \leq n$.

   (b) Let $x^{(0)} = (x_1, ..., x_n)$.

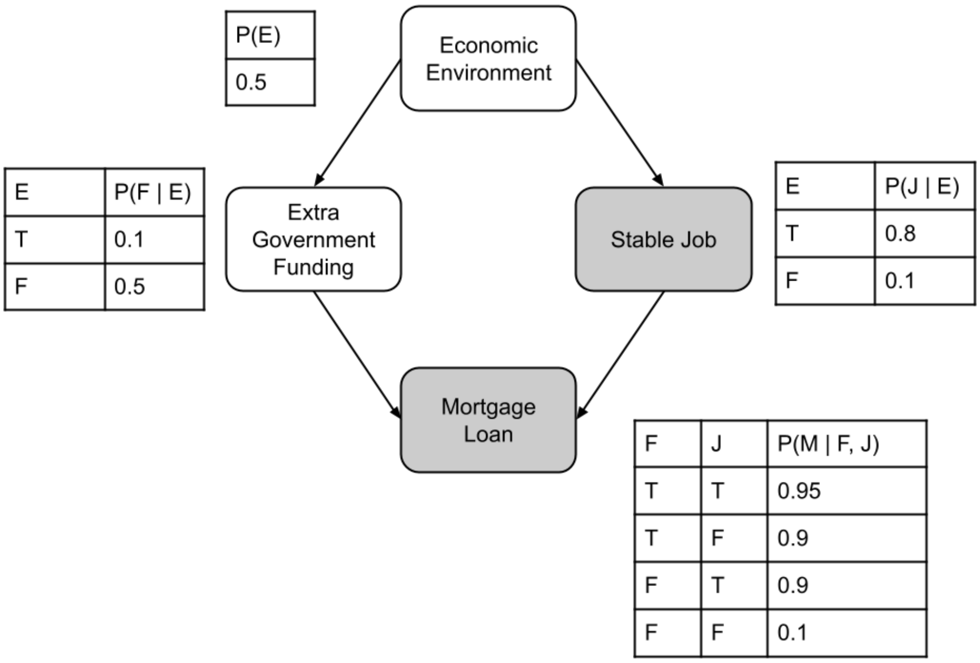
2. For $t = 1, 2, ...$:

   (a) Pick an index i, $1 \leq i \leq n$ uniformly at random (other schedules are also possible, e.g. iterate cyclically over all the indices).

   (b) Draw $x_i^{(t)}$ from $\mathbb{P}(X_i | x_{\setminus i}^{(t-1)}, \mathbf{e})$.

   (c) Set $x_{\setminus i}^{(t)} = x_{\setminus i}^{(t-1)}$.

The sampler generates a sequence of samples $x^{(0)}, x^{(1)}, ...$ from the Markov chain over all possible states. The stationary distribution of the Markov chain is the joint distribution $\mathbb{P}(X_1, ...X_n | \mathbf{e})$. Thus, samples are usually taken from the Markov chain after a certain number of iterations, allowing enough time for the chain to reach the stationary distribution (sometimes referred to as the *burn-in* time). In order to yield independent samples from the distribution $\mathbb{P}(X_1, ...X_n | \mathbf{e})$, the set of samples are usually thinned out, e.g., by picking every 100-th sample after burn-in.



P(E)
0.5

E	P(F | E)
T	0.1
F	0.5

E	P(J | E)
T	0.8
F	0.1

F	J	P(M | F, J)
T	T	0.95
T	F	0.9
F	T	0.9
F	F	0.1

Figure 1: The Mortgage Bayesian Network

Sampling from the Mortgage Network

The network shown in Figure 1 is an example of a Bayesian network that models the relationship between the variables *Economic Environment* (E), *Stable Job* (J), *Extra Government Funding* (F) and *Mortgage Loan* (M). The idea is that if the young person got the mortgage loan to buy a house, either they had a stable job or it was because the government allocated extra funding to make such loans more accessible. Whether extra funding is available or the person has a stable job is influenced by the state of the economic environment, which can be positive (T) or negative (F).
Now consider the task of estimating the probability of extra funding being allocated, given that the person has a stable job and has received the mortgage loan,

$$\mathbb{P}(F = T|J = T, M = T),$$

through sampling. Since *Stable Job* and *Mortgage Loan* are observed, the Gibbs sampler needs to draw samples from the probability distribution $\mathbb{P}(F, E|J = T, M = T)$, which can be derived by marginalization. For the sampler, you will need to compute the two conditional probability distributions of *Extra Government Funding* and *Economic Environment* given all other variables (as mentioned above, *Stable Job* and *Mortgage Loan* are fixed to T, indicated by green colour in Figure 1). This can be done using the *Local Probability Distributions* (LPDs) and the Markov blanket of the variable of interest.

In particular, perform the following tasks: (7 points).

1. Derive the formulas for $\mathbb{P}(E = T|F = T, J = T, M = T), \mathbb{P}(E = T|F = F, J = T, M = T), \mathbb{P}(F = T|E = T, J = T, M = T)$ and $\mathbb{P}(F = T|E = F, J = T, M = T)$ **up to**

the normalization constants in the denominators, and compute their values by renormalizing the two possible values for each conditional probability distribution. (2 points)

2. Implement the Gibbs sampler sketched above for the Bayesian network in Figure 1 and draw 100 samples from the joint probability distribution $\mathbb{P}(F, E|J = T, M = T)$. (4 points)

3. Estimate the marginal probability of extra government funding, given that the person has a stable job and received a mortgage loan $\mathbb{P}(F = T|J = T, M = T)$ from the 100 samples. (1 point)

Convergence Diagnostics

Now, determine whether the Gibbs sampler has reached the stationary distribution. An *adhoc* way is to plot the relative frequencies of $F = T$ and $E = T$ up to each iteration t against t, for two independent runs of the sampler. Then, the *burn-in* time can be set to the point where the two sequences converge. In addition, several formal tests are available to test the stationarity of the sampler after a given time. Here, we consider the *Gelman and Rubin* multiple sequence diagnostic test [1]. In this method, multiple sequences are simulated and the idea is that after convergence the behaviour of all chains should be approximately the same. In particular, the variance within the chains should be the same as the variance across the chains.

Perform the following tasks: (8 points)

4. Now draw 50 000 samples instead of 100 using the Gibbs sampler.

5. Provide the plot of the relative frequencies of $F = T$ and $E = T$ up to each iteration t against t, for two independent runs of the sampler. Suggest a *burn-in* time based on this plot. (1 point)

6. Apply the *Gelman* test and plot potential scale reduction factor changes over the iterations. Roughly speaking, this factor measures the ratio of variances within and between independent runs of the sampler. Thus, for a stationary distribution, this factor should be close to 1.0. Suggest a *burn-in time* based on this plot. (2 points)

7. Investigate the auto-correlation among the samples. We expect adjacent members from a Gibbs sampling sequence to be positively correlated, and we can quantify the amount of this correlation by using the autocorrelation function. The lag-k auto-correlation ρ_k is the correlation between every draw and its k-th neighbouring samples. Provide plots for both variables *Extra Government Funding* and *Economic Environment*. Suggest an interval for drawing approximately independent samples. (2 points)

8. Re-estimate $\mathbb{P}(F = T|J = T, M = T)$ based on 100 samples obtained after the suggested *burn-in* time and *thinning-out*. Compare with (3) and comment on your results. (1 point)

9. Compute the probability $\mathbb{P}(F = T|J = T, M = T)$ analytically and compare it to the sampling estimate. In real-world applications, sampling is performed because it is

usually not possible to easily compute the probabilities analytically. However, since the Bayesian network in Figure 1 is only a small network with discrete variables, the analytical approach is possible. (2 points)

References

[1] Gelman, A and Rubin, DB, *Inference from iterative simulation using multiple sequences*, Statistical Science, 7, 457-511, 1992.

Submitting

Use the Jupyter Notebook for both code and report. Submit two files: `.ipynb` containing your code, all responses and comments, and `.pdf` or `.html` file exported from the `.ipynb` file (make sure all important output of cells is visible). Submit those two files through Moodle.

Deadline: 07.01.2025, 23:59.

Contact

If you have any questions, comments, or difficulties, email `m.bochenek@uw.edu.pl` or `b.domzal@mimuw.edu.pl` or ask in person.