ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

D-BSSE
Department of Biosystems
Science and Engineering

# Statistical Models in Computational Biology

Jack Kuipers
David Dreifuss
Xiang Le Guo
Rudolf Schill

Due 6th of April 2023
**Please submit your project with the filename** `Lastname(s)_Project6.pdf`.

**Problem 15: Monte Carlo estimation of an expected value** **(2 points)**

Let $X$ be a random variable with probability density (or mass) function $f$. Furthermore, let $g$ be a differentiable function whose domain contains the support $\mathcal{X}$ of $X$. Then $g(X)$ is in turn a random variable, and we seek its expected value

$$\mathbb{E}[g(X)] = \int_{\mathcal{X}} g(x) f(x)\, \mathrm{d}x.$$

Assume that the integral (or the sum) cannot be evaluated using analytical methods. One way to address this problem is to draw independent samples $X_1, \ldots, X_N$ from the distribution of $X$. By the law of large numbers, the expectation $\mathbb{E}[g(X)]$ can then be approximated by the sum

$$\widehat{g}(\mathbf{X}) := \frac{1}{N} \sum_{i=1}^{N} g(X_i),$$

where $\mathbf{X} = (X_1, \ldots, X_N)$. Now, $\widehat{g}(\mathbf{X})$ is itself a random variable, and $g(X_1), \ldots, g(X_N)$ are also independent. Show[1] that $\mathbb{E}[\widehat{g}(\mathbf{X})] = \mathbb{E}[g(X)]$ and $\mathrm{Var}(\widehat{g}(\mathbf{X})) = \frac{\mathrm{Var}(g(X))}{N}$, where $\mathrm{Var}(\cdot)$ is the variance. Think about if those results also apply if $X_1, ..., X_N$ are generated from a MCMC sampler (no need for demonstration here).

**Problem 16(data analysis): Sampling in the Rain Network** **(8 points)**

The network shown in Figure 1 is an example of a Bayesian network that models the relationship between the variables *Cloudy* ($C$), *Rain* ($R$), *Sprinkler* ($S$) and *Wet grass* ($W$). In particular, it encodes the fact that if the grass is wet, either the sprinkler is on or it's raining, which are both influenced by whether it's cloudy or not. Each variable is then a binary variable taking values T (True) or F (False). The local probability distributions defining the Bayesian network are indicated in the figure.

Consider the task of estimating the probability of rain, given that the sprinkler is on and the grass is wet:

$$P(R = \mathsf{T} \mid S = \mathsf{T}, W = \mathsf{T}).$$

In this exercise we will do this via sampling, and compare with the analytical result[2]. One approach is to use Gibbs sampling to sample from $P(R, C \mid S = \mathsf{T}, W = \mathsf{T})$ and marginalising over $C$.

[1] *Hint:* Use Bienaymé's identity, which states that the variance of a sum of *pairwise independent* random variables equals the sum of the individual variances of the random variables. This is because the covariance between any pair of independent random variables is zero, meaning that their covariance cancels.

[2] In general, analytical solutions are not available for interesting models.

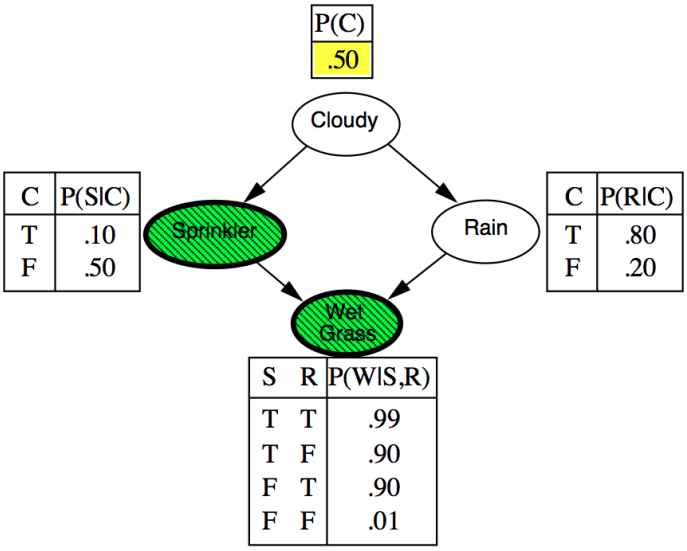


Figure 1: The *Rain Bayesian network*. Green nodes indicate that they were observed to be true.

We will assess whether the Gibbs sampler has reached the stationary distribution in two manners. First, we will plot the relative frequencies of $R = \mathsf{T}$ and $C = \mathsf{T}$ up to each iteration, for multiple independent runs of the sampler (or chains). A more formal method is the *Gelman and Rubin* multiple sequence diagnostic test [1], which compares the variance within the chains with the variance across the chains. The R package `coda` provides convergence diagnostics for MCMC samplers, including the *Gelman and Rubin* test.

(a) Derive the expressions for $P(C = \mathsf{T} \mid R = \mathsf{T}, S = \mathsf{T}, W = \mathsf{T})$, $P(C = \mathsf{T} \mid R = \mathsf{F}, S = \mathsf{T}, W = \mathsf{T})$, $P(R = \mathsf{T} \mid C = \mathsf{T}, S = \mathsf{T}, W = \mathsf{T})$ and $P(R = \mathsf{T} \mid C = \mathsf{F}, S = \mathsf{T}, W = \mathsf{T})$ and compute their values. (1 point)

(b) Implement the Gibbs sampler for the Bayesian network in Figure 1 and draw 100 samples from the joint posterior probability distribution $P(R, C \mid S = \mathsf{T}, W = \mathsf{T})$. (1 point)

(c) Estimate the marginal probability of rain, given that the sprinkler is on and the grass is wet $P(R = \mathsf{T} \mid S = \mathsf{T}, W = \mathsf{T})$ from the 100 samples. (1 point)

(d) We expect adjacent members from a Gibbs sampling sequence to be positively correlated, and we can quantify this with the auto-correlation function. The lag-$k$ auto-correlation $\rho_k$ is the correlation between every draw and its $k$th neighbouring sample. The effective sample size (ESS) is a statistic aiming to quantify how much the auto-correlation reduces the amount of information contained in the $n$ samples of the chain. For a single chain, it can be estimated as:

$$\widehat{\mathrm{ESS}} = \frac{n}{1 + 2\sum_{k=1}^{\infty} \rho_k}$$

Use the R function `acf()` to provide plots and estimates of the auto-correlation functions for the samples of both variables *Rain* and *Cloudy*. Provide estimates of the ESS. (1 point)

(e) Draw 50,000 samples instead of 100 using the Gibbs sampler.

(f) Plot the relative frequencies of $R = \mathsf{T}$ and $C = \mathsf{T}$ 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)

(g) Apply the *Gelman and Rubin* test and plot the potential scale reduction factor changes over the iterations using `gelman.plot()` from the `coda` package. This factor measures the ratio

of the variances within and between independent runs of the sampler and should be close to $1.0$ for stationary distributions. Suggest a *burn-in* time based on this plot. (1 point)

(h) Re-estimate $P(R = \mathsf{T} \mid S = \mathsf{T}, W = \mathsf{T})$ based on samples obtained after the suggested *burn-in* time. (1 point)

(i) Compute the probability $P(R = \mathsf{T} \mid S = \mathsf{T}, W = \mathsf{T})$ analytically. Compare with (c) and (h) and comment on your results. (1 point)

## References

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