Requirements for reports

1. The questions should be addressed in the same order they appear in the assignment. The text of the question MUST be retained and placed before each answer. The working language is English.

2. The answer to a particular question may take a form of a plot, formula etc followed by a brief explanation and **a conclusion**. All your conclusions MUST be justified numerically, i.e., by some computed quantities, plots, etc. The answers do not need to be lengthy but, again, they MUST be convincing in mathematical and statistical sense, i.e., in terms of some quantitative measures. Note that I pay much attention to the conclusions, so try to make it as clear as possible.

3. The due date for the assignment is December 21, 2019, 23:59.

4. Late submission: not possible.

5. The answer to a question MUST contain the most important fragments code in `R` or some other language which can be placed in the appendix of the report.

6. Failures to comply with the above rules may reduce your grade for the assignment.

Grading principle: 8 points maximum + 2 extra points for Problem 5.
For those who want more, consider additional home assignment on pairs trading, see LMS. Note that this additional assignment can improve only the part of the accumulated grade (i.e., earned over the semester) associated with the home assignments. It does not affect the grade for the in-class tests.

Assignment on Markov Chain Monte Carlo

1. Monte-Carlo integration. Let n be the length of your first+last name. Let $d = n \bmod 9$ be the remainder of the ratio $\frac{n}{9}$. If the resulting $d < 2$, then add 2 to d.

   (a) Generate two uniform random numbers (command `runif()`) such that $a \in [0, 0.5]$, $b \in [0.8, 2.0]$. Round them to two decimal places. Show these two numbers.

   (b) Compute the d-dimensional integral by the Monte Carlo method

$$\int\limits_{M_d} \underbrace{\frac{1}{\sqrt{2\pi}} e^{-Z_1^2/2} \frac{1}{\sqrt{2\pi}} e^{-Z_2^2/2} \ldots \frac{1}{\sqrt{2\pi}} e^{-Z_d^2/2}}_{d \text{ times}} dZ_1 \ldots dZ_d \tag{1}$$

   where $M_d = \{a < Z_1^2 + Z_2^2 + \ldots + Z_d^2 < b, \quad Z_i \geqslant 0,\ i = 1, \ldots, d\}$. Note that the volume of the d-dimensional ball of radius R is

$$V_d = \frac{\pi^{d/2} R^d}{\Gamma\left(\frac{d+2}{2}\right)}$$

   where $\Gamma(x)$ is the so-called gamma function such that $\Gamma(x+1) = x\Gamma(x)$, $\Gamma(1) = 1$, $\Gamma(1/2) = \sqrt{\pi}$. Use the symmetry arguments (consider fraction of the first quadrant (2D), octant (3D) etc in the total volume, by induction).

(c) Run your program for several sample sizes providing the error margins for every run. Make comments on that.

(d) Try to calculate integral (1) analytically. For that purpose, notice that if $Z_i \sim N(0,1)$ then $Z_1^2 + Z_2^2 + ... + Z_d^2 \sim \chi_d^2$. Use elementary probability rules for calculating $\Pr(a < \chi_d^2 < b)$ (command `pchisq`). Compare the result with your numerical calculations.

2. Accept-reject algorithm.

(a) Generate a random integer $d \in [1, 6]$. Show the respective line of code. Let d defined so be the number of distribution type in the list below:

i. the lognormal distribution, see [1], p.91, $\mu \in [0, 0.5]$, $\sigma \in [0, 0.5]$

ii. the Weibull distribution, see [1], p.90, $\alpha \in [0.5, 1.5]$, $\beta \in [4.5, 5.5]$

iii. the inverse Gaussian distribution, see [1], p.94, $\mu \in [0.3, 0.8]$, $\lambda \in [5.5, 6.5]$

iv. the logistic distribution, see [1], p.96, $\mu \in [-0.5, 0.5]$, $\sigma \in [0.5, 1.5]$

v. the double exponential distribution, see [1], p.100, $\lambda \in [0.5, 1.5]$

vi. the exponential distribution, see [1], p.85, $\theta \in [0.8, 1.2]$

(b) Parameters in the above list should be generated randomly from the uniform distribution, rounded to two decimal places and the respective lines of code should be presented.

(c) Take the uniform distribution as an envelope. A sketch of the target distribution may help choose the support ("width") of the uniform envelope as well as the scaling factor M. Clearly state the parameters of the envelope distribution.

(d) Obtain a histogram of random numbers generated from the target distribution. Compare it with theoretical plot of density.

(e) Compute the mean and variance of the generated sample, compare them with theoretical values.

3. Study of Markov chains.

(a) Introduce a 5×5 1-step transition matrix (T_1) of Markov chain. Explain the principle of the construct. Use your own matrix because two similar matrices in reports of two different students have zero chance to appear.

(b) Check whether the chain has the limiting distribution:

i. Theoretically, i.e., report whether the chain is ergodic

ii. By computing the n-step transition matrix T_n. Discuss the meaning of lines in T_n.

iii. Under eigenvector approach

iv. Comment on the convergence rate (how fast the chain reaches the stationary state)

(c) Build a histogram of the limiting distribution

(d) Suggest some diagnostics to ensure that you have reached the stationary mode

(e) Compute the mean and variance using the limiting distribution

(f) Compute the mean and variance over time and compare them with those obtained in item 3e. Point out the moment when you start averaging

(g) Comment on quality of mixing by inspecting the state evolution of the chain

(h) Check the autocorrelation of the chain. Discuss whether the memory length in the chain is consistent with the convergence rate.

(i) Give an example of non-convergent Markov chain. Page 181 of [2] may provide a clue. Check that such a matrix has no limiting distribution, by methods of item 3b.

4. Google PageRank algorithm. Generate a random integer $d \in [1, 6]$. Generate a random $d \times d$ matrix $\mathbf{L}$ filled with integers 0 and 1. Interpret $\mathbf{L}$ in the sense of the PageRank algorithm, see [3, p.576].

   (a) Find the ranking $\boldsymbol{p}$ of the pages.

   (b) What makes you think that the iterative process converges? Explain in terms of Markov chains.

   (c) Draw and explain a sketch similar to Figure 14.46 of [3], illustrating the result of item 4a.

5. **Optional, the remaining 2 points**. The Metropolis-Hastings (MH) algorithm by a decoding example.

   (a) Read a .txt document containing a large text in English (the larger the better and not too ancient), e.g., from this source http://www.gutenberg.org/ebooks/1661 or from elsewhere.

   (b) Compute the matrix with probabilities of two successive letters ignoring punctuation, spaces and non-letter symbols (may be a tough task).

   (c) Using the MH algorithm decode the sequence "yonpnretic" – a permuted meaningful word.

   (d) Show the entire process in detail.

6. The Gibbs sampler.

   (a) Generate n numbers from the normal distribution $N(\mu, \sigma^2)$ where μ, σ^2 are the mean and variance (must be unique for each student)

   (b) Use the normal prior $N(\mu_0, \sigma_0^2)$ for the mean and the inverse gamma prior $\mathrm{IG}(a, b)$ with density

$$f(x) = \frac{b^a}{\Gamma(a)} \frac{e^{-\frac{b}{x}}}{x^{a+1}} \tag{2}$$

   for variance.

   (c) Write down the posterior distribution.

   (d) Derive the conditional distributions for μ and σ^2 from the posterior.

   (e) Choose the reasonable parameters μ_0, σ_0^2, a and b for the priors. Plot density (2) (without the x-independent scaling factor) to facilitate choice of a, b (trace the maximum of distribution (2)).

   (f) Write a program implementing the Gibbs sampler algorithm.

   (g) Plot the initial dynamics of the 2D Markov chain (first 100-200 steps).

   (h) Plot the cumulative mean against time for both parameters. Based on that, estimate the burn-in period (i.e., non-stationary phase).

(i) Plot the joint and marginal distributions of μ, σ using Markov chain values after the burn-in period. Compute the mean values and standard deviations using these distributions. Compare your findings with the original parameters and time averages of item 6h. Also, compare the computed mean with that of the original data set.

(j) Check the chain for quality of mixing.

(k) Increase the data size n, re-run your program and describe what happens to the marginal distributions.

(l) Make your priors "wider" and "narrower" about true parameters, re-run your program and investigate the marginal distributions. Is there any effect for big data size? For small data size? For very small data size?

See pp. 202-203 of [4] for examples.

Assignment on the Kalman filter

Use `dlm` package for this task. See Chapter 2 of [5].

5. Download the archive and pick up the data file `DLMn.dat` where `n` is your number in the Excel grading list on this course. Read it with the command `read.csv("DLMn.csv",header=F,sep=',')`.

6. Consider the random walk plus noise model ($t = 0, 1, 2, ...$)

$$Y_t = \mu_t + v_t, \quad v_t \sim N(0, V)$$
$$\mu_t = \mu_{t-1} + w_t, \quad w_t \sim N(0, W)$$

with $\mu_0 \sim N(a, b)$. Observations Y_t and state evolution vector μ_t are stored in the first and second columns of the loaded data frame. The standard deviation of the state is $M(1, 3)$ and that of the observations is $M(1, 4)$ where M is the data frame. $M(1, 5)$ and $M(1, 6)$ entries contain the mean a and standard deviation b for the initial state μ_0.

7. Compute the filtering state estimates and plot them together with the actual observations and actual states. Explain the purpose of filtering.

8. Compute the one-step ahead forecasts for observations and plot them together with the actual observations

9. Compute the smoothing state estimates and plot them. Explain the purpose of smoothing.

References

[1] Z. Karian, E. Dudewicz, Handbook of Fitting Statistical Distributions with R, Taylor & Francis, 2010.
URL https://books.google.ru/books?id=jR0XAAAACAAJ

[2] E. Suess, B. Trumbo, Introduction to Probability Simulation and Gibbs Sampling with R, Use R!, Springer New York, 2010.
URL https://books.google.ru/books?id=R-_N4qWm_IAC

[3] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition, Springer Series in Statistics, Springer New York, 2009.
URL https://books.google.com/books?id=tVIjmNS3Ob8C

[4] C. Robert, G. Casella, Introducing Monte Carlo Methods with R, Use R!, Springer, 2010.
URL https://books.google.ru/books?id=WIjMyiEiHCsC

[5] G. Petris, S. Petrone, P. Campagnoli, Dynamic Linear Models with R, Use R!, Springer New York, 2009.
URL https://books.google.ru/books?id=VCt3zVq8TO8C