Exam

D. Jason Koskinen
koskinen@nbi.ku.dk

Advanced Methods in Applied Statistics
Feb - Apr 2019

Photo by Howard Jackman

Info

- In submitting the solutions there is no need to rephrase the problem. “Solution for 1a” is sufficient.
- The submission format for explanations and plots is a PDF file. Also, include any and all software scripts used to establish your answer(s) and/or produce plots in a **separate** file(s).
    - The write-up submission should include the text “WriteUp” in the file name.
- Working in groups or any communication about the problems is prohibited. Using the internet as a resource is encouraged, but soliciting any help is also prohibited.
- Some questions have multiple parts. For full credit, all parts must be done.

Info

- The exam will be graded out of 100 possible points
  - It will count for 40% of the final course grade
- Submit all code used!! The software you write to complete the problem is **part** of the solution.
- Must be submitted by 14:00 CET Friday April 5, 2019 for full credit.
- The exam **MUST BE** electronically submitted via the Digital Exam website.
  - For catastrophic submission failures you can email the exam to Jason
- For any concerns, questions, or comments email Jason.

Starting points (5 pts.)

- On the first page of your write-up include your full name, date, name of this course, UCPH ID, and the title of your exam submission
- Also type out (please don't copy/paste) " I (your name here) expressly vow to uphold my scientific and academic integrity by working individually on this exam and soliciting no direct external help or assistance."
- Finding help/solutions online is fine. But, for example, posting to a forum and receiving assistance is not okay.
- Good luck!!!

Problem 1 (25 pts.)

- There is a file posted online which has 5 columns, each representing a physical observable of interest generated from some underlying function. There are 5000 entries, i.e. rows.
  - http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2019/data/Exam_2019/Exam_2019_Prob1.txt
- The variables/columns are independent distributions with **no** correlation to the data in the other columns
- Be mindful about accounting for truncated ranges, as well as likelihood functions that have periodic components which will create local minima/maxima
  - There is at least one column of data which is generated from a function with local minima/maxima

Lists of Distributions

- The data in each column is produced from functions ***similar to***, or potentially exactly the same as, f(x) or f(k) shown at right
- Note that the displayed functions may be unnormalized
  - Hint: Some will require a normalization to convert them to probability distribution functions
  - The functions f(x) have bounds on their parameters a, b, and c

$$-10 \le a \le 10$$
$$-10 \le b \le 10$$
$$4000 \le c \le 8000$$

$$f(x) \propto \begin{cases} \dfrac{1}{x+5}\sin(ax) \\ \sin(ax)+1 \\ \sin(ax^2) \\ \sin(ax+1)^2 \\ x\tan(x) \\ 1+ax+bx^2 \\ a+bx \\ \sin(ax)+ce^{bx}+1 \\ e^{-\frac{(x-a)^2}{2b^2}} \end{cases}$$

$$f(k) \propto \begin{cases} \binom{n}{k}p^k(1-p)^{n-k} & binomial \\ \dfrac{\lambda^k e^{-\lambda}}{k!} & poisson \\ \dfrac{-1}{\ln(1-p)}\dfrac{p^k}{k} & logarithmic \end{cases}$$

Problem 1a

- Use the separate data from columns 1, 2, and 3 to identify the function on the previous slide from which each was generated. Find the *best-fit values* and *uncertainties* on those values for the distribution using a *likelihood method* (either bayesian or maximum likelihood is fine)
  - E.g. if f(x)=sin(ax+b)*exp(-x+c)+x/k! were one of the functions, then find the best-fit values for a, b, c, and k and their uncertainties
  - Degeneracies exist, e.g. sin(x)=cos(a+x), which can produce functionally identical data distributions
  - Any function, with associated best-fit parameters which is **statistically compatible** with the data in the files will be accepted as a proper solution. Only one solution is necessary, but needs to be **justified** as statistically compatible.
- Data in column 1 and 2 have artificially truncated ranges
  - Column 1 is only sampled in the independent variable from 20 to 27
  - Column 2 is only sampled in the independent variable from -1 to 1

Problem 1b

- Plot the data and the corresponding best-fit function on the same plots
  - 3 separate 1-dimensional plots
  - Plot as a function of the independent variable
  - Histogram the data, and scale the best-fit function to be ‘reasonable’ so that the features of both the data and best-fit function can be visually compared

Problem 2 (15 pts.)

- There is a file posted online (http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2019/data/Exam_2019/Exam_2019_Problem2.txt) with data.
  - The first column is the azimuth angle of the data point
  - The second column is the zenith angle of the data point
  - There are 100 paired data points in total
  - The values are in units of radian

Problem 2a

- Quantify whether the data is spherically isotropically distributed
  - Include any supporting plots, discussion, and numbers
  - A spherically isotropic distribution is uniform in the azimuth angle from 0 to 2π, and uniform in cos(zenith angle) from -1 to 1
  - Hint: you can use Monte Carlo generated pseudo-experiments to produce a test-statistic distribution of a spherically isotropic distribution.

Problem 2b

- Test whether the data fits the two following alternative hypotheses better than the isotropic hypothesis:
  - Hypothesis A: That 20% of the total sample is uniformly distributed from azimuth = {0.225π, 0.55π} and zenith = {0.30π, 1π) and the remaining 80% is fully isotropic
  - Hypothesis B: That 15% of the total sample is uniformly distributed from azimuth = {0π, 1π} and zenith = {0.5π, 1π) and the remaining 85% is fully isotropic
  - Report the two p-values: $H_{isotropic}$ versus H_A as well as $H_{isotropic}$ versus H_B

Problem 3 (15 pts.)

- Small problems
- It is acceptable to report non-integer values in 3b

Problem 3a

- The following data file has a list of test statistic values. For this test statistic higher values are **always** associated with worse agreement than lower values.
  - The file is at: http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2019/data/Exam_2019/Exam_2019_Problem3a.txt
- The file is a list of 3000 bootstrap test statistic samples. What is the critical value, i.e. threshold, of the test statistic that corresponds to a **one-sided** p-value of 4.55%?
- If the true distribution for the test statistics in the file is chi-squared distributed, does the test statistic threshold established with the bootstrap samples match the expected critical value from a chi-squared distribution with 5 degrees-of-freedom?
  - Quantitatively and qualitatively justify your answer.

Problem 3b

- As of 2019, only 4 rock climbers have ever successfully ascended a sport climb rated at 5.15c: Adam Ondra, Chris Sharma, Stefano Ghisolfi, and Alex Megos.
- In 2024, and in the absence of any other data, each climber will have a probability (p) which follows a beta distribution of completing an attempted 5.15d with the following values of the beta distribution:
  - Ondra: $\alpha=60$ $\beta=35$
  - Sharma: $\alpha=4$ $\beta=100$
  - Ghisolfi: $\alpha=40$ $\beta=35$
  - Megos: $\alpha=89$ $\beta=45$

Problem 3b (cont.)

- In 2024, each climber will attempt to climb an expected number of routes rated at 5.15d. The distribution of attempts will follow a Poisson distribution with a mean of 4 for each climber.
- What is the most likely number of total attempted 5.15d routes for the combined 4 climbers in 2024?

Problem 3b (cont.)

- Compute the likelihood distribution of Alex Megos succeeding k=2 times out of n=5 attempts, as a function of his probability of success p.
  - What is the likelihood value for p=0.35?
- Suppose Alex Megos has attempted 5 separate climbs of 5.15d routes in 2024, and succeeded twice. Estimate his posterior probability of succeeding in one given attempt.
  - Plot the prior, posterior, and likelihood on the same plot
  - The plots and distributions can be unnormalized as long as they are on a vertical scale that they can be visually compared

Problem 4 (30 pts.)

- Data was taken to examine what variables (or combinations of variables) might be used to identify when a patient will miss their scheduled appointment, i.e. a 'No-show'.
- Create a classifier which separates patients that are likely to have a 'No-show' from those that are not likely to 'No-show'
  - Consider 'No-show=True' as the signal or real positive, and the 'No-show=False' as the background or real negative
- The data set has been divided:
  - Training/Testing data set is at:
    - http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2019/data/Exam_2019/Exam_2019_Prob4_TrainData.csv
    - http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2019/data/Exam_2019/Exam_2019_Prob4_TestData.csv
  - The 'blind' analysis data set is at http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2019/data/Exam_2019/Exam_2019_Prob4_BlindData.csv
    - Only used in problem 4c
  - Include **ALL input files** when submitting your solution

Problem 4 (cont.)

- There are many possible features (i.e. variables) to use, but we will restrict the classification algorithm to only use the following:

```
features_to_train = ['Gender',
                     'ScheduledDay',
                     'AppointmentDay',
                     'Age',
                      'TimeDifference',
                      'Neighbourhood',
                    'Diabetes',
                    'Alcoholism',
                     'Handcap',
                    'SMS_received',
                    'R1'
                    ]
```

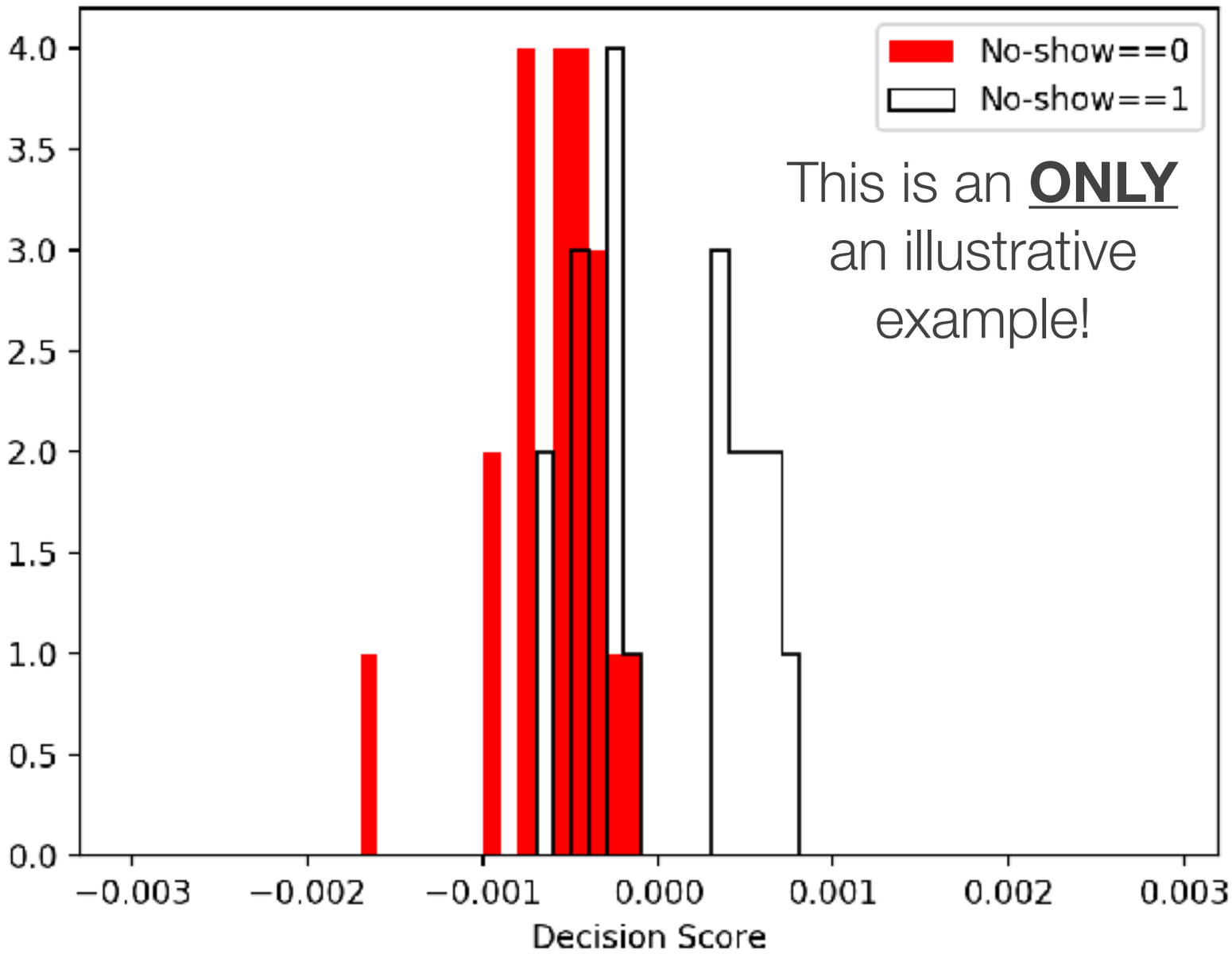
- The above features should be the only variables (besides “ID”) which are in the file(s)

Problem 4a

- Make a single plot with overlaid histograms using **all events** from the test file versus the test statistics; separated into ‘No-show==1’ and ‘No-show==0’
  - Separate the two populations and plot the No-show==1 patients in ***black*** and No-show==0 in ***red***

Problem 4a (example)

- Example here is an illustration for only 20 No-show==0 entries and 20 No-show==1 entries, your plot may look **very** different



Problem 4b

- Rank the variables starting with most important to least important
  - Provide the ranked feature list, include some quantitative metric which you use for the ranking
- Discuss how to identify and avoid overtraining in supervised machine learning algorithms

Problem 4c

- Using the same classifier developed in Problem 4a, run the classifier over all the entries on the blind sample
  - http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2019/data/Exam_2019/Exam_2019_Prob4_BlindData.csv
  - Results will be graded on the **classification accuracy**
  - The new data file has a unique ID number for every patient
    - Produce a text file which contains **only** the IDs which your classifier classifies as **No-show==1** (last_name.AMAS_Exam_2019.Problem4.NoShowTrue.txt)
    - Produce a text file which contains **only** the IDs which your classifier classifies as **No-show==0** (last_name.AMAS_Exam_2019.Problem4.NoShowFalse.txt)
    - The file names **MUST BE EXACT**. For two submissions from Jason Koskinen these would be "koskinen.AMAS_Exam_2019.Problem4.NoShowFalse.txt" and "koskinen.AMAS_Exam_2019.Problem4.NoShowTrue.txt"
  - Basic text files. No Microsoft Word documents, Adobe PDF, or any other extraneous text editor formats. Only a single ID number per line in the text file that can be easily read by numpy.loadtxt().
    - One entry per line and no commas, brackets, parenthesis, etc.

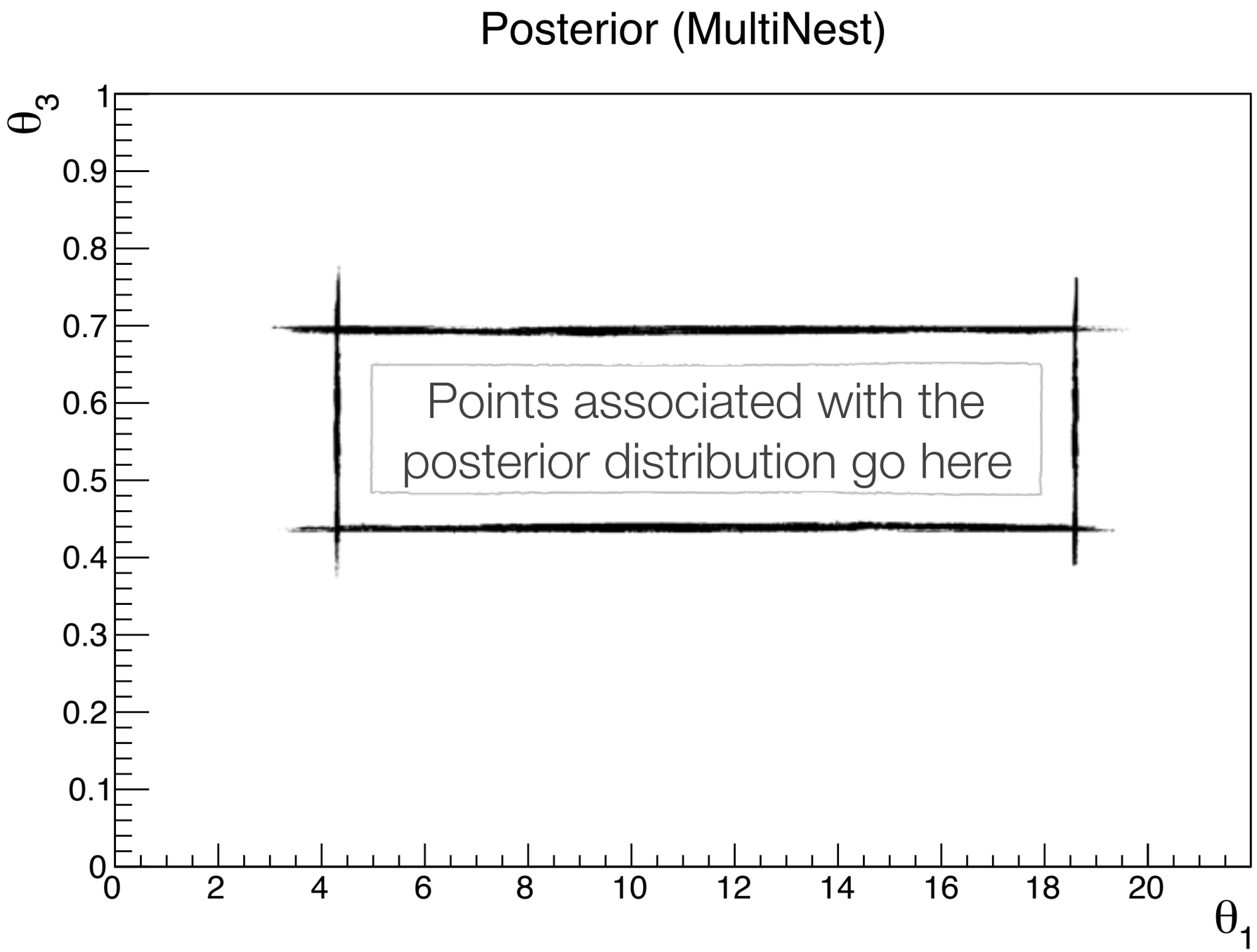
Problem 5 (15 pts.)

- With the function below as the defined quasi-likelihood in 3-dimensions use MultiNest, or some other nested sampling bayesian algorithm, to plot the 2-D posterior distribution for the parameters θ_1 and θ_3, i.e. scatter-point plot for θ_3 vs. θ_1 (empty example on a following slide)

$$\mathcal{L}(\theta_1, \theta_2, \theta_3) = 3\left(\cos(\theta_1)\cos(\theta_2) + \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\theta_3-\mu)^2}{2\sigma^2}} \cos(\theta_1/2) + 3 \right)$$

- The range should be restricted for θ_1 and θ_2 to 0-7π and for θ_3 from 0-3. Also, set μ=0.68 as the true mean of the normal distribution and σ^2=0.04
- What are the best-fit values for θ_1, θ_2 ,and θ_3 that you find from maximizing the above function, i.e. when you generate the posterior distribution?

Problem 5


Posterior (MultiNest)
θ_3
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
2
4
6
8
10
12
14
16
18
20
θ_1
Points associated with the posterior distribution go here

Problem 5 (cont.)

- The posterior distribution is proportional to the output of the quasi-likelihood. Make two separate raster scan plots in 2-D of the output from the likelihood over the same ranges as for the previous plot. Essentially, map out the likelihood (or ln-likelihood) landscape.
    - For the scan of θ_2 vs. θ_1, fix θ_3 to the best-fit point found from MultiNest, i.e. an unchanging value. Similarly, for the scan of θ_3 vs. θ_1, fix θ_2 to the best-fit point.
- Does the posterior distribution match the raster scan plots? Discuss why it should, or why it should not.