Studio 6: Continuous Data, Continuous Priors
18.05 Spring 2014
Jeremy Orloff and Jonathan Bloom

You should have downloaded studio6.zip and unzipped it into your 18.05 working directory.

NASDAQ Data

We have data from the NASDAQ stock exchange on trades in a certain stock on 4 days in March 2014. Here are the first 4 lines of the tradesdata0.csv

	Date	timeNumber	timeHHMMSS	Size	Price
1	20140303	0.3958333	93000	228	1206.366
2	20140303	0.3958449	93001	892	1206.516
3	20140303	0.3958565	93002	1343	1205.846
4	20140303	0.3958681	93003	855	1206.520

The data file, tradesdata0.csv is in studio6.zip

We processed this data to produce the data file for this class: studio5dataframe.csv

(If you're interested, the processing code is in studio6-prep.r)

Today's project: Model the rate at which trades come into the exchange.

Exporatory data analysis

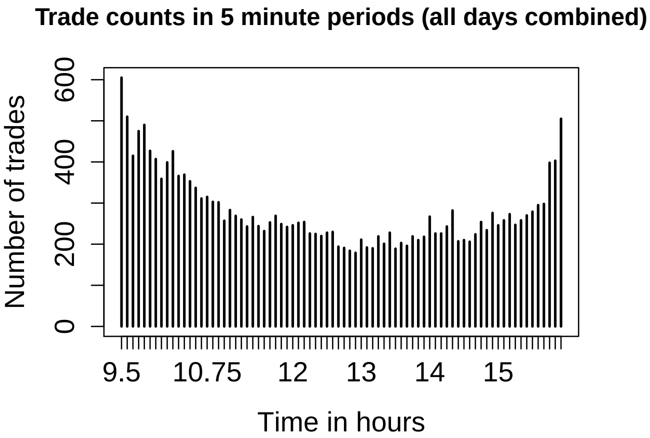
Real data analysis starts by *exploring* the data.

Some things to try are:

- Plot lists of data.
  This can help find glaring errors in the data:
  - on the wrong scale
  - missing
  - all 0
  - multiple modes
- Histograms
- Time plots
- Slice and dice data to find (suggested) patterns

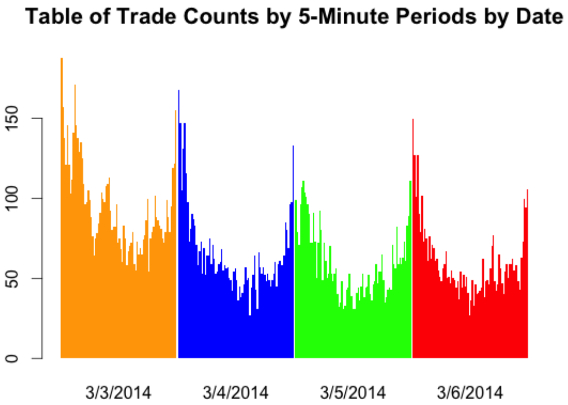
(See studio6-prep.r and studio6.r)

Exploration: number of trades vs. time of day



- More trades at beginning and end of day than in the middle.
- Note: 9.5 = 9:30 am, 16.0 = 4:00 pm

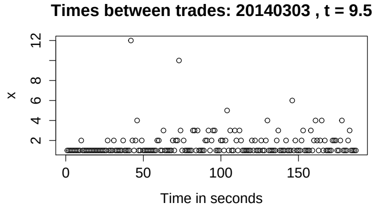
Exploration: number of trades vs. time of day II

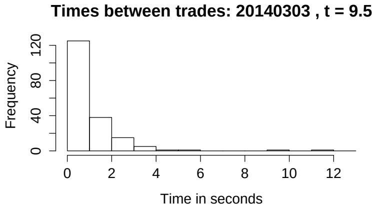


- More trades at beginning and end of *each* day
- Could the waiting time be exponentially distributed with a parameter that changes during the day?

Exploration: a single time slot

Code is in studio6.r, which also generates many more plots.





- Plot of data doesn't set off alarms
- Histogram resembles that of an exponential distribution

Board question: Bayesian updating

- Fix the date as March 4, 2014 (20140304).
- For each 5 minute time slot we'll assume the wait time between trades follow an exponential($1/\theta$) distribution. (θ is then the mean wait time.)
- studio6.r shows how to get the list of wait times for any 5 minute time slot.

1. Outline the mathematics needed to do Bayesian updating starting from a uniform prior on θ in the range [0, 8].

2. Outline a plan to write code in R to do the updating for each time slot in turn.

Solution for problem 1

The prior is $f(\theta) = 1/8$ on [0, 8].

The likelihood for wait time x_1 is

$$f(x_1 \,|\, \theta = \frac{\mathrm{e}^{-x_1/\theta}}{\theta}$$

The posterior is

$$f(\theta \,|\, x_1) = \frac{f(x_1 \,|\, \theta) f(\theta)}{T},$$

where T = total probability $= \int_0^8 f(x_1 \,|\, \theta) f(\theta)\, d\theta$.

For subsequent data points, x_2 etc., the formulas are the same except the prior is always the previous posterior.

• We could multiply the likelihoods for each x_i together to get the likelihood of all the data and then update all at once.

Code outline for problem 2

Updating a single day/time slot (Do this for each time slot on March 4.)
(i) Get the list of waiting times for that day/time slot.

(ii) Discretize θ in [0,8]:

`thetaRange = seq(0,8,dtheta)`, where `dtheta = 0.02`

(iii) For the data point x the likelihood array is

$$\texttt{likelihood} = \texttt{exp}(-\texttt{x}/\texttt{thetaRange})/\texttt{thetaRange}$$

(iv) For each data point x_j do numerical Bayesian updating by:

```
prior = posterior  # Previous posterior becomes new prior.
unnormPosterior = prior*likelihood
posterior = unnormPosterior/(dtheta*sum(unnormPosterior))
```

Code outline continued

• Note: We could also compute the likelihood of all the data and update all at once.

Details on normalizing priors and posteriors

Since priors and posteriors are functions of θ:

• Numerically they are lists of length `length(thetaRange)`.

• They are normalized so that the numerical intergral

$$\mathtt{sum}(\mathtt{f}(\mathtt{thetaRange}) * \mathtt{dtheta}) = 1$$

• For example the pdf $f(\theta) = c\theta^2$ is given numerically by

```
f = thetaRange^2/sum(thetaRange^2*dtheta).
```

• The uniform prior is given numerically by

$$\mathtt{uniformPrior} = \mathtt{rep}(1, \mathtt{n})/(\mathtt{n} * \mathtt{dtheta}),$$

where `n = length(thetaRange)`

R: Bayesian updating

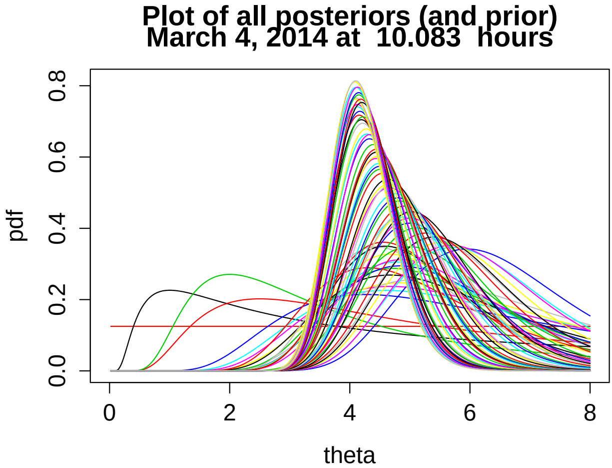
3(a) Implement your coding plan. Make sure that the final posterior for each timeslot is saved for later use.

3(b) For each posterior find the MAP estimate (value of θ that maximizes the posterior) and make a plot of MAP vs. time slot.

(Hint: get help on the R function `which.max`.)

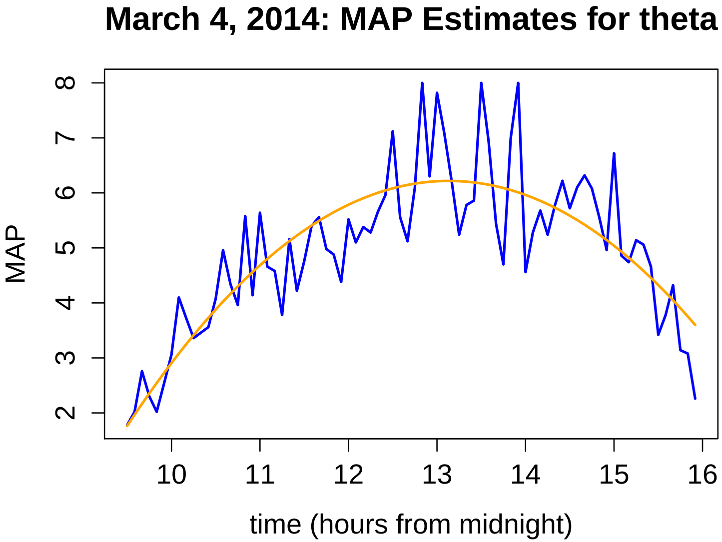
3(c) Redo (a) and (b) with the quadratic prior $c(4-\theta)^2$ on [0, 8].

One time slot

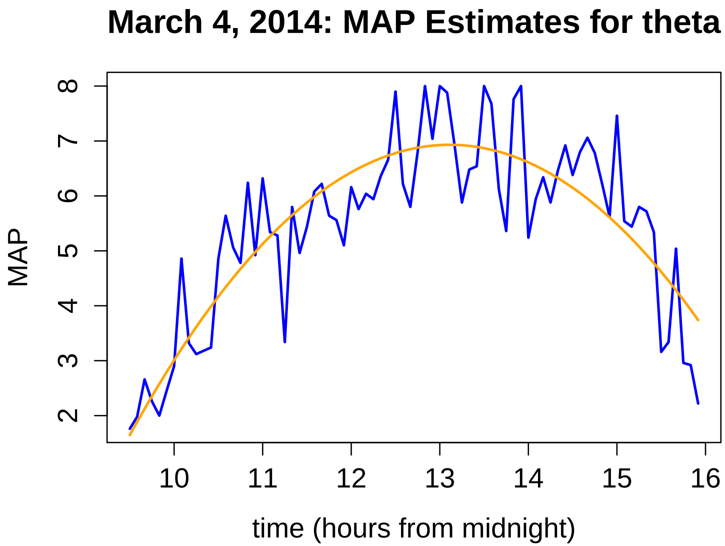


θ is the paramater of the exponential$(1/\theta)$ distribution for waiting time between trades. It is the mean waiting time between trades.

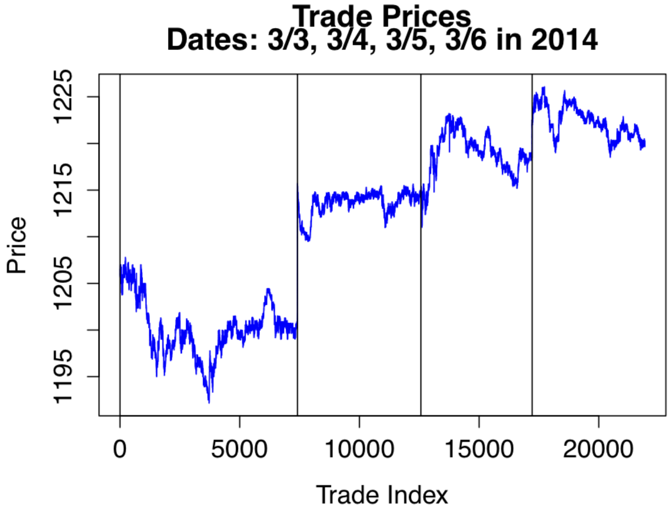
MAP Estimates for θ for all time slots (uniform prior)



MAP Estimates for θ for all time slots (quadratic prior)



Price vs trade number (a bonus picture)



The trades are listed in chronological order. The horizontal axis is the trade number .