Midterm Exam

Instruction

Please complete your exam and upload your solutions on Bright Space. Maximum number of points is 15 while the sum of all the points of all questions is 18.

In [1]:
```
using Distributions
using Turing
using StatisticalRethinking
using Random, Plots, MCMCChains
```

Question 1 Markov chain (2 points)

In the class we had the following example for given probability transition matrics P: What is p(x(t+2)=3|p(x(t)=1)?

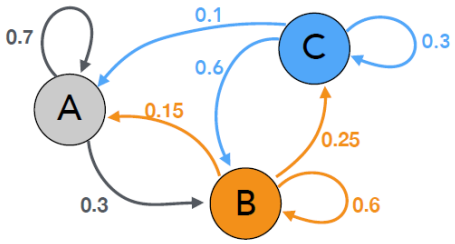
In [2]:
```
# Solution to the question given in the class
using LinearAlgebra, Statistics, StatsBase, Plots;

n, N = 5, 10^6
P= [0 0.5 0 0 0.5;
    0.5 0 0.5 0 0;
    0 0.5 0 0.5 0;
    0 0 0.5 0 0.5;
    0.5 0 0 0.5 0]

P2=P*P
P2[1,3]
```

Out[2]: 0.25

Now, let us consider the following problem:
For 3 discrete states shown below, there are transition probabilities to move to other states, as well as a probability to stay in the same state.



$$p(A \to A) = 0.7 \quad p(A \to B) = 0.3 \quad p(A \to C) = 0.0$$

$$p(B \to A) = 0.15 \quad p(B \to B) = 0.6 \quad p(B \to C) = 0.25$$

$$p(C \to A) = 0.1 \qquad p(C \to B) = 0.3 \quad p(C \to C) = 0.6$$

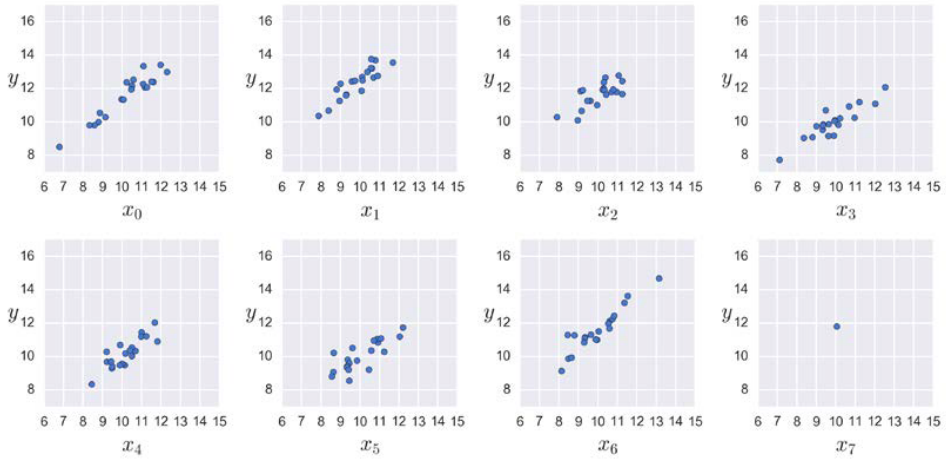
Transition probability matrix can be written as, $$P = \begin{bmatrix} 0.7 & 0.3 & 0.0 \\ 0.15 & 0.6 & 0.25 \\ 0.1 & 0.3 & 0.6 \end{bmatrix}$$

Now, let us assume at time n the system is in state 2 = B. What is the probability that it will be in state B again 3 time periods later (at time n+3)?

In [ ]:

Question 2 Hierarchical models (2 points)

The data obtained from 8 different groups is shown in figure below.



Group 7 contains only one data point. Write the following probabilistic models:

1. In case each group is considered separately.
2. In case all the data is piled together and groups are not considered.
3. Hierarchical model.

Assume prior distributions and prior coefficient values that make sense to you.

1. Draw your estimate for the regression for the graphs with x axis x_6 and x_7 for cases a) and b) and explain your drawing.

Question 3 Gaussian model (4 points)

Let us assume the following measurement equations for n measurements:

$$x_i = \mu + \epsilon_i$$
$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Also, let's assume a Gaussian prior for θ

$$\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$$

We are interested in determining the mean of the Gaussians with **known variance**. Let us draw N i.i.d. data points $X = \{x_1, x_2, \ldots, x_N\}$ from one dimensional Gaussian distribution $\mathcal{N}(x_i|\mu, \sigma^2)$. The mean of the Gaussians is $\mathcal{N}(\mu_N, \sigma_N^2)$

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$$

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\bar{x},\ \bar{x} = \frac{\sum_{n=1}^{N} x_n}{N}$$

a) Write the formula for the likelihood.

b) What would be the values of μ_N and σ_N^2 for non-informative prior (assume that σ_0^2 tends to infinity)? Are these values maximum likelihood estimates?

c) How would these formulas change if the variance is known but different for each measurement? So, instead of having the same σ^2 for N measurements, assume that we have $\sigma_1^2, \sigma_2^2, \ldots, \sigma_N^2$.

d) Posterior predictive distribution is Normal with $\mu_* = \mu_N$ and $\sigma_*^2 = \sigma_N^2 + \sigma^2$. Define posterior predictive distribution? Why is the variance of posterior predicting distribution larger than the estimated variance σ_N^2?

e) Now, assume that the σ^2 is unknown as well. How would you model the problem in this case? What would be the posterior distribution and why?

Question 4 (2 points)

a) Why would one want to generate multiple chains for the same parameter in MCMC sampling?

b1) What steps were performed in deriving black box variational inference to allow for using variational inference without the need to derive all the formulas for the proposal distribution q?

b2) Why is this variational inference called black box? What does it mean that nothing in the algorithm is model specific?

Question 5 (3 points)

Given a data set $D = \{(x_1, y_1), \ldots, (x_N, y_N)\}$, where $x_n \in \mathbb{R}^M$ and $y_n \in \{0, 1\}$. The probabilistic classification method known as *logistic regression* attempts to model these data as

$$p(y_n = 1|x_n) = \sigma(\theta^T x_n + b)$$

where $\sigma(x) = 1/(1 + e^{-x})$ is the *logistic function*. Let's introduce shorthand notation $\mu_n = \sigma(\theta^T x_n + b)$. So, for every input x_n, we have a model output μ_n and an actual data output y_n .

(a) Express $p(y_n|x_n)$ as a Bernoulli distribution in terms of μ_n and y_n.
(b) If furthermore is given that the data set is IID, show that the log-likelihood is given by

$$L(\theta) \triangleq \log p(D|\theta) = \sum_n \{y_n \log \mu_n + (1 - y_n) \log(1 - \mu_n)\}$$

(c) Show that the derivative of the log-likelihood is

$$\nabla_\theta L(\theta) = \sum_{n=1}^{N} \left(y_n - \sigma(\theta^T x_n + b)\right) x_n$$

Question 6 (3 points)

We will use an example from [2] called seven scientists. Seven scientists make a measurement of the same quantity. They get the answers x = {−27.020, 3.570, 8.191, 9.898, 9.603, 9.945, 10.056}.

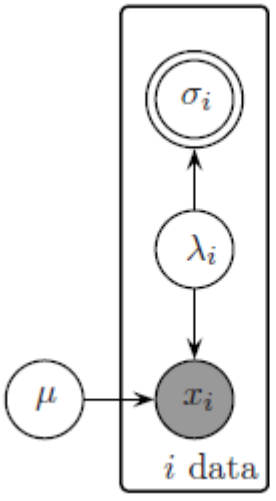
The code for models that piles everything together and for the hierarchical model shown in the figure below is given.

a) Analyze μ of both models. Why are the means and the variances different? Why does the hierarchical model provide variance around 10 even though there is a large outlier?

b) Comment on autocorrelation of the chains. You can use function autocorplot(chain).

c) What is the expected Waic?

d) We had a lecture on model checking. Is the Normal distribution appropriate for this data? How would one perform model checking?



$$\begin{aligned}
\mu &\sim \text{Gaussian}(0, 0.001) \\
\lambda_i &\sim \text{Gamma}(0.001, 0.001) \\
\sigma_i &\leftarrow 1/\sqrt{\lambda_i} \\
x_i &\sim \text{Gaussian}(\mu, \lambda_i)
\end{aligned}$$

Let us try first the implementation in which we assume that there is only one λ. In this case, we expect that the mean will be around the sample mean which is 3.46

In [5]:
```
using Distributions, Plots;
using Turing
```

```
using StatisticalRethinking
using Random, Plots, MCMCChains
```

In [6]:

```
μ0=0
σ0=1/0.001
α0=0.001
β0=1/0.001
x = [-27.020, 3.570, 8.191, 9.898, 9.603, 9.945, 10.056]

@model seven_scientists(x) = begin
    μ ~ Normal(μ0, σ0)
    l ~ Gamma(α0,β0)
    x ~ Normal(μ, 1/sqrt(l))
end;
```

In [7]:

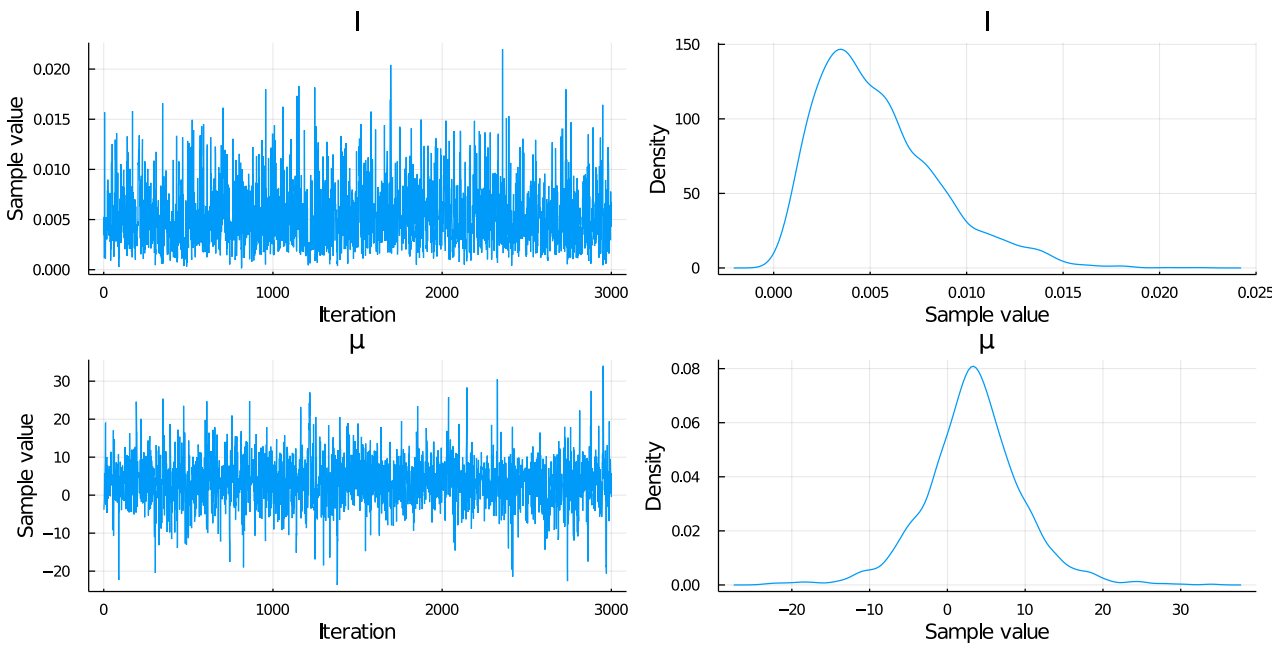
```
model = seven_scientists(x)
chain = sample(model, NUTS(0.65), 3000);
```

```
┌ Info: Found initial step size
│   ϵ = 0.4
└ @ Turing.Inference C:\Users\Miodrag Bolic\.julia\packages\Turing\O1Pn0\src\inference\h
mc.jl:195
Sampling: 100%|████████████████████████████████████████| Time: 0:00:00
```

In [8]:

```
plot(chain)
```

Out[8]:



In [9]:

```
summarystats(chain)
```

Out[9]:

Summary Statistics

parameters	mean	std	naive_se	mcse	ess	rhat
Symbol	Float64	Float64	Float64	Float64	Float64	Float64
l	0.0054	0.0031	0.0001	0.0001	1408.4424	1.0030
μ	3.3965	6.3316	0.1156	0.1910	1214.5668	0.9998

Now, let us look at the problem given above where the precision for each scientist is considered

separately.

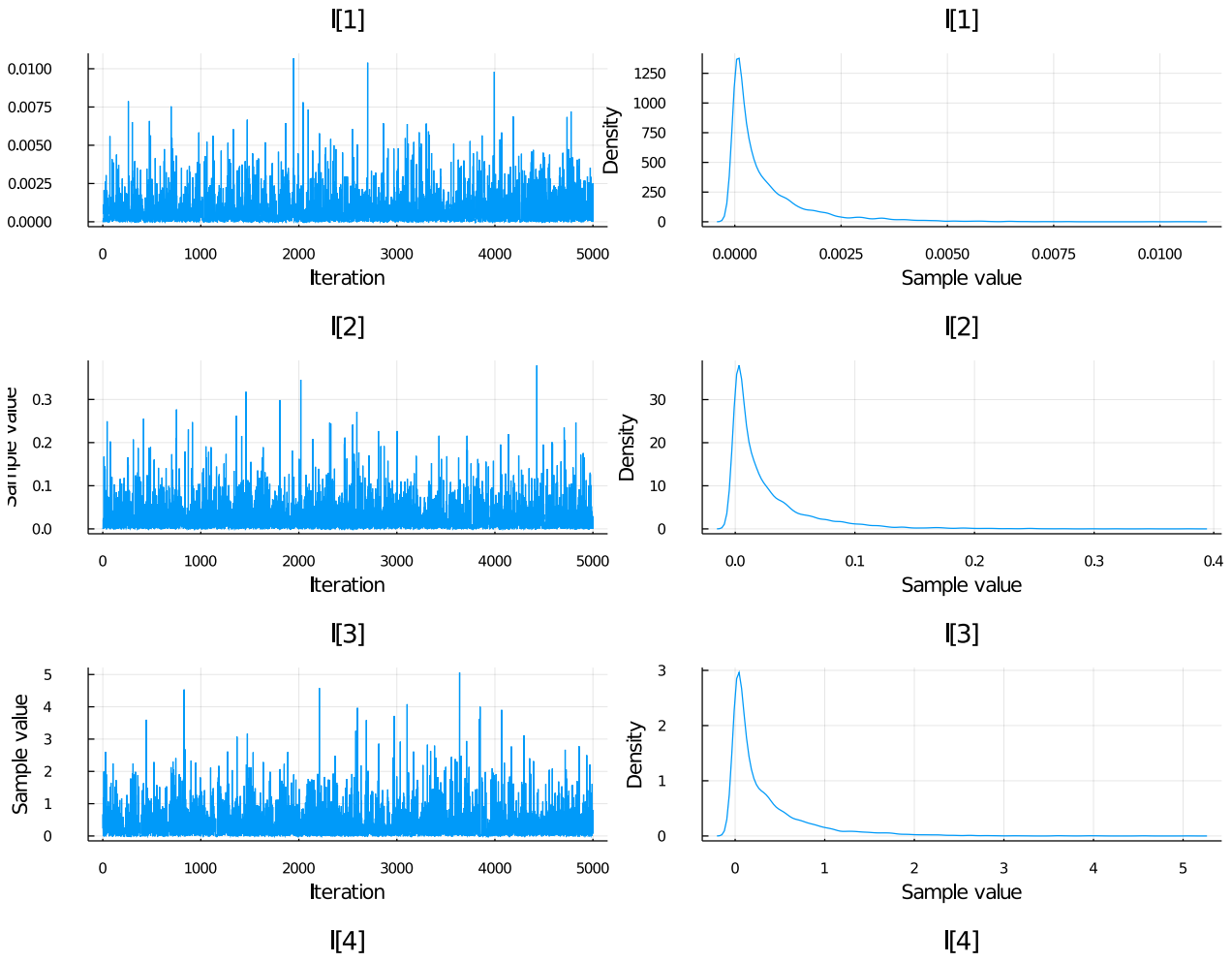
In [10]:
```
@model seven_scientists1(x) = begin
    # Our prior belief about the probability of heads in a coin.
    μ ~ Normal(μ0, σ0)
    N = length(x)
    l = Vector{Real}(undef, N)
    for j in 1:N
        # Heads or tails of a coin are drawn from a Bernoulli distribution.
        l[j] ~ Gamma(α0,β0)
        #l[j] ~ Uniform(0,30)
        x[j] ~ Normal(μ, 1/sqrt(l[j]))
    end
end;
```

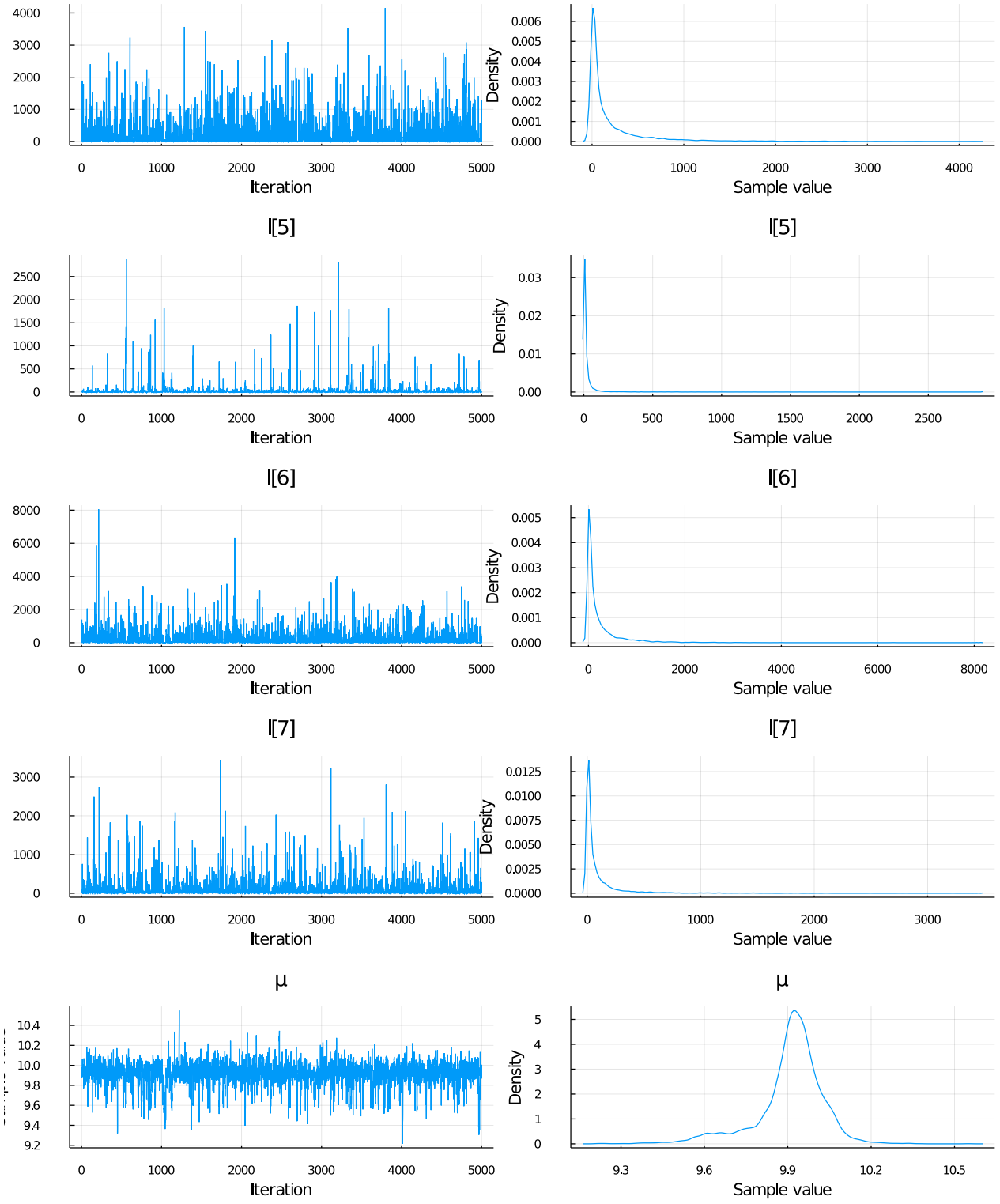
In [11]:
```
# Sampling from the model and indering the posterior distribution
model = seven_scientists1(x)
chain = sample(model, NUTS(0.65), 5000);
```

```
┌ Info: Found initial step size
└   ϵ = 0.4
 @ Turing.Inference C:\Users\Miodrag Bolic\.julia\packages\Turing\O1Pn0\src\inference\hmc.jl:195
Sampling: 100%|████████████████████████████████████████| Time: 0:00:03
```

In [12]:
```
plot(chain)
```

Out[12]:





In [13]:
```
summarystats(chain)
```

Out[13]:

Summary Statistics

parameters	mean	std	naive_se	mcse	ess	rhat
Symbol	Float64	Float64	Float64	Float64	Float64	Float64
l[1]	0.0007	0.0010	0.0000	0.0000	4121.9071	1.0002
l[2]	0.0258	0.0362	0.0005	0.0005	4762.8142	1.0001
l[3]	0.3380	0.4846	0.0069	0.0076	4833.9051	0.9998
l[4]	205.6215	387.5155	5.4803	6.4840	3416.3375	0.9998
l[5]	28.7381	125.7128	1.7778	3.1015	1710.6435	0.9998

l[6]	234.7892	460.0865	6.5066	8.9224	2640.4689	1.0000
l[7]	97.2919	230.5359	3.2603	4.0441	2694.7646	0.9998
μ	9.9090	0.1233	0.0017	0.0043	820.9406	1.0006