tssl_lab3

October 2, 2023

1 TSSL Lab 3 - Nonlinear state space models and Sequential Monte Carlo

In this lab we will make use of a non-linear state space model for analyzing the dynamics of SARS-CoV-2, the virus causing covid-19. We will use an epidemiological model referred to as a Susceptible-Exposed-Infectious-Recovered (SEIR) model. It is a stochastic adaptation of the model used by the The Public Health Agency of Sweden for predicting the spread of covid-19 in the Stockholm region early in the pandemic, see Estimates of the peak-day and the number of infected individuals during the covid-19 outbreak in the Stockholm region, Sweden February – April 2020.

The background and details of the SEIR model that we will use are available in the document *TSSL Lab 3 Predicting Covid-19 Description of the SEIR model* on LISAM. Please read through the model description before starting on the lab assignments to get a feeling for what type of model that we will work with.

1.0.1 DISCLAIMER

Even though we will use a type of model that is common in epidemiological studies and analyze real covid-19 data, you should *NOT* read to much into the results of the lab. The model is intentionally simplified to fit the scope of the lab, it is not validated, and it involves several model parameters that are set somewhat arbitrarily. The lab is intended to be an illustration of how we can work with nonlinear state space models and Sequential Monte Carlo methods to solve a problem of practical interest, but the actual predictions made by the final model should be taken with a big grain of salt.

We load a few packages that are useful for solving this lab assignment.

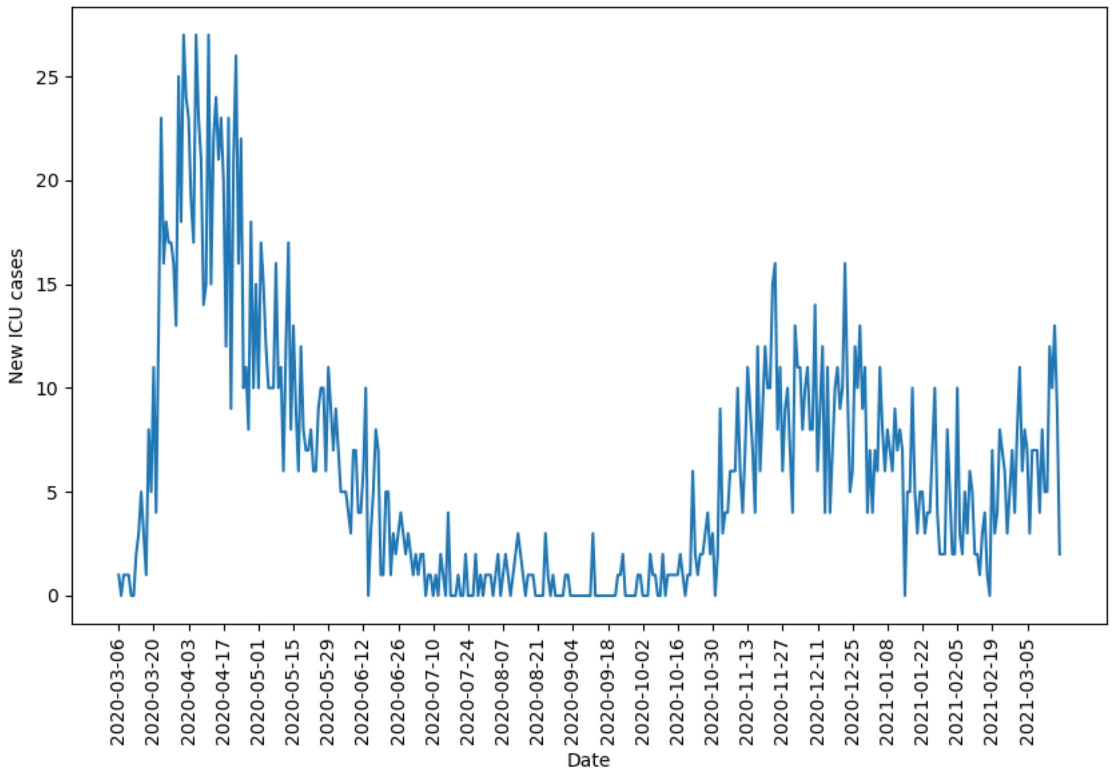
[1]:
```
import pandas  # Loading data / handling data frames
import numpy as np
import matplotlib.pyplot as plt
plt.rcParams["figure.figsize"] = (10,6)  # Increase default size of plots
```

1.1 3.1 A first glance at the data

The data that we will use in this lab is a time series consisting of daily covid-19-related intensive care cases in Stockholm from March 2020 to March 2021. As always, we start by loading and plotting the data.

[2]:

```
data=pandas.read_csv('SIR_Stockholm.csv',header=0)
y_sthlm = data['ICU'].values
u_sthlm = data['Date'].values
ndata = len(y_sthlm)
plt.plot(u_sthlm,y_sthlm)
plt.xticks(range(0, ndata, 14), u_sthlm[::14], rotation = 90)  # Show only one␣
 ↪tick per 2 week for clarity
plt.xlabel('Date')
plt.ylabel('New ICU cases')
plt.show()
```



[3]:

```
len(y_sthlm)
```

[3]: 378

Q0: What type of values can the observations y_t take? Is a Gaussian likelihood model a good choice if we want to respect the properties of the data?

A: The observations y_t are discrete in natrue and hence a Gaussian likelihood model would be a bad choice since the Gaussian distribution is continuous in nature therefore it is better to use a probability mass function from other distributions that support discrete data.

1.2 3.2 Setting up and simulating the SEIR model

In this section we will set up a SEIR model and use this to simulate a synthetic data set. You should keep these simulated trajectories, we will use them in the following sections.

[4]:
```python
from tssltools_lab3 import Param, SEIR

"""For Stockholm the population is probably roughly 2.5 million."""
population_size = 2500000

"""" Binomial probabilities (p_se, p_ei, p_ir, and p_ic) and the transmission␣
↪rate (rho)"""
pse = 0        # This controls the rate of spontaneous s->e transitions. It is␣
↪set to zero for this lab.
pei = 1 / 5.1  # Based on FHM report
pir = 1 / 5    # Based on FHM report
pic = 1 / 1000 # Quite arbitrary!
rho = 0.3      # Quite arbitrary!

""" The instantaneous contact rate b[t] is modeled as
  b[t] = exp(z[t])
  z[t] = z[t-1] + epsilon[t], epsilon[t] ~ N(0,sigma_epsilon^2)
"""
sigma_epsilon = .1

""" For setting the initial state of the simulation"""
i0 = 1000  # Mean number of infectious individuals at initial time point
e0 = 5000  # Mean number of exposed...
r0 = 0     # Mean number of recovered
s0 = population_size - i0 - e0 - r0  # Mean number of susceptible
init_mean = np.array([s0, e0, i0, 0.], dtype=np.float64)  # The last 0. is the␣
↪mean of z[0]

"""All the above parameters are stored in params."""
params = Param(pse, pei, pir, pic, rho, sigma_epsilon, init_mean,␣
↪population_size)

""" Create a model instance"""
model = SEIR(params)
```

Q1: Generate 10 different trajectories of length 200 from the model an plot them in one figure. Does the trajectories look reasonable? Could the data have been generated using this model?

For reproducibility, we set the seed of the random number generator to 0 before simulating the trajectories using np.random.seed(0)

Save these 10 generated trajectories for future use.

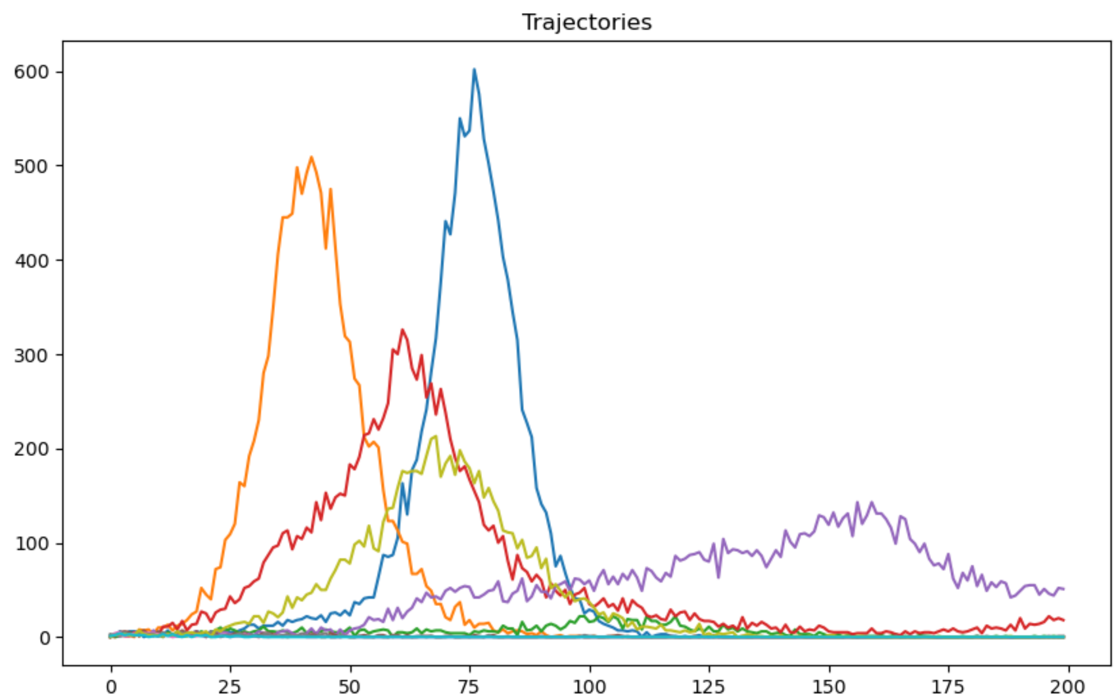
(hint: The SEIR class has a simulate method)

Ans: The parameter, pic, which is the probability that an infectious individual enters the ICU and rho, the probability of exposure for each encounter with an infectious individual have been set arbitrarily, it becomes a gross simplification of reality. So, we can say that the data was not generated using this model and its parameters.

[5]:
```
np.random.seed(0)
# help(model.simulate)
trajectories = model.simulate(T = 200, N = 10)
alpha = trajectories[0]
y_obs = trajectories[1]
```

[32]:
```
for i in range(10):
    plt.plot(y_obs[0,i,:])
plt.title('Trajectories')
```

[32]: Text(0.5, 1.0, 'Trajectories')



1.3 3.3 Sequential Importance Sampling

Next, we pick out one trajectory that we will use for filtering. We use simulated data to start with, since we then know the true underlying SEIR states and can compare the filter results with the ground truth.

Q2: Implement the **Sequential Importance Sampling** algorithm by filling in the following functions.

The **exp_norm** function should return the normalized weights and the log average of the unnormalized weights. For numerical reasons, when calculating the weights we should “normalize” the log-weights first by removing the maximal value.

Let $\bar{\omega}_t = \max(\log \omega_t^i)$ and take the exponential of $\log \tilde{\omega}_t^i = \log \omega_t^i - \bar{\omega}_t$. Normalizing $\tilde{\omega}_t^i$ will yield the normalized weights!

For the log average of the unnormalized weights, care has to be taken to get the correct output, $\log(1/N \sum_{i=1}^N \tilde{\omega}_t^i) = \log(1/N \sum_{i=1}^N \omega_t^i) - \bar{\omega}_t$. We are going to need this in the future, so best to implement it right away.

(hint: look at the SEIR model class, it contains all necessary functions for propagation and weighting)

[7]:
```
from tssltools_lab3 import smc_res

def exp_norm(logwgt):
    """
    Exponentiates and normalizes the log-weights.

    Parameters
    ----------
    logwgt : ndarray
        Array of size (N,) with log-weights.

    Returns
    -------
    wgt : ndarray
        Array of size (N,) with normalized weights, wgt[i] = exp(logwgt[i])/
 ↪sum(exp(logwgt)),
        but computed in a /numerically robust way/!
    logZ : float
        log of the normalizing constant, logZ = log(sum(exp(logwgt))),
        but computed in a /numerically robust way/!
    """
    c_t = np.max(logwgt)
    scaled_wgt = logwgt - c_t
    scaled_norm_wgt = np.exp(scaled_wgt)/np.sum(np.exp(scaled_wgt))
    logZ = np.log((1/logwgt.shape[0])*np.sum(np.exp(scaled_wgt))) + c_t
    return scaled_norm_wgt, logZ

def ESS(wgt):
    """
    Computes the effective sample size.

    Parameters
    ----------
    wgt : ndarray
        Array of size (N,) with normalized importance weights.
```

```
    Returns
    -------
    ess : float
        Effective sample size.
    """
    ess = np.sum(wgt)**2/np.sum(wgt**2)
    return ess

def sis_filter(model, y, N):
    d = model.d
    n = len(y)

    # Allocate memory
    particles = np.zeros((d, N, n), dtype = float)  # All generated particles
    logW = np.zeros((1, N, n))  # Unnormalized log-weight
    W = np.zeros((1, N, n))  # Normalized weight
    alpha_filt = np.zeros((d, 1, n))  # Store filter mean
    N_eff = np.zeros(n)  # Efficient number of particles
    logZ = 0.  # Log-likelihood estimate

    # Filter loop
    for t in range(n):
        # Sample from "bootstrap proposal"
        if t == 0:
            particles[:, :, 0] = model.sample_state(N=N) # Initialize from p(alpha_1)
            logW[0, :, 0] = model.log_lik(y[0], particles[:, :, 0]) # Compute weights

        else:
            particles[:, :, t] = model.sample_state(particles[:, :, t-1], N = N) # Propagate according to dynamics
            logW[0, :, t] = model.log_lik(y[t], particles[:, :, t]) + logW[0, :, t-1] # Update weights

        # Normalize the importance weights and compute N_eff
        W[0, :, t], _ = exp_norm(logW[0, :, t])
        N_eff[t] = ESS(W[0,:,t])

        # Compute filter estimates
        alpha_filt[:, 0, t] = np.sum(particles[:, :, t]*W[0, :, t],axis=1)


    return smc_res(alpha_filt, particles, W, logW=logW, N_eff=N_eff)
```

Q3: Choose one of the simulated trajectories and run the SIS algorithm using $N = 100$ particles.

Show plots comparing the filter means from the SIS algorithm with the underlying truth of the Infected, Exposed and Recovered.

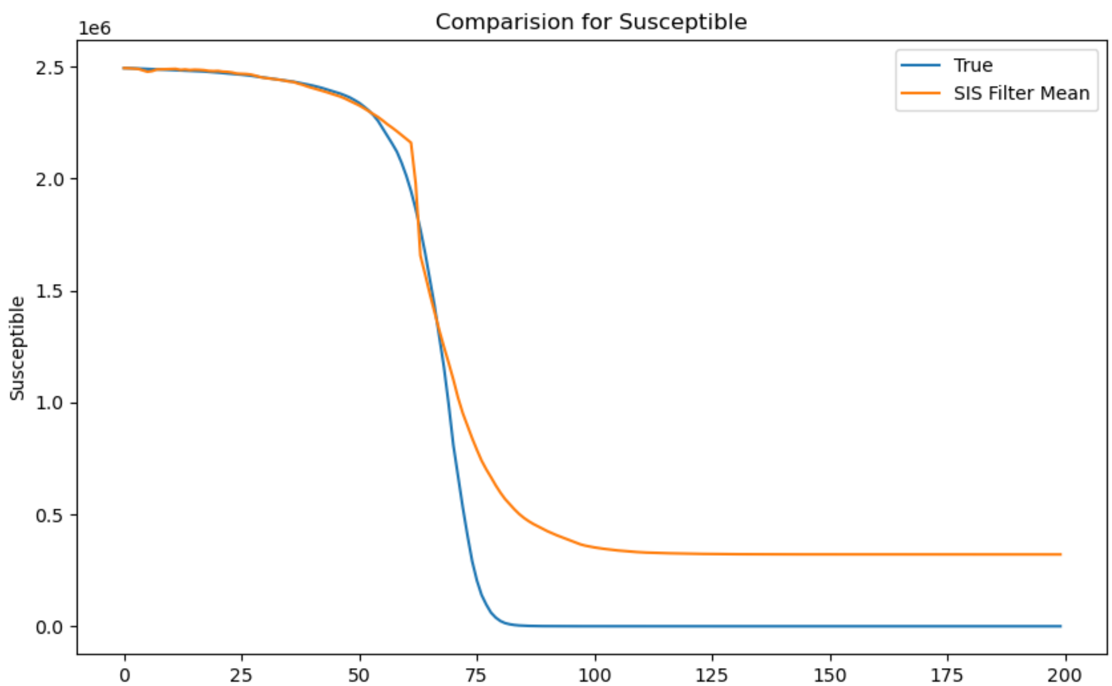
Also show a plot of how the ESS behaves over the run.

(hint: In the model we use the S, E, I as states, but S will be much larger than the others. To calculate R, note that $S + E + I + R = Population$)

[8]:
```
sis = sis_filter(model, y = y_obs[0,0,:], N = 100)
```

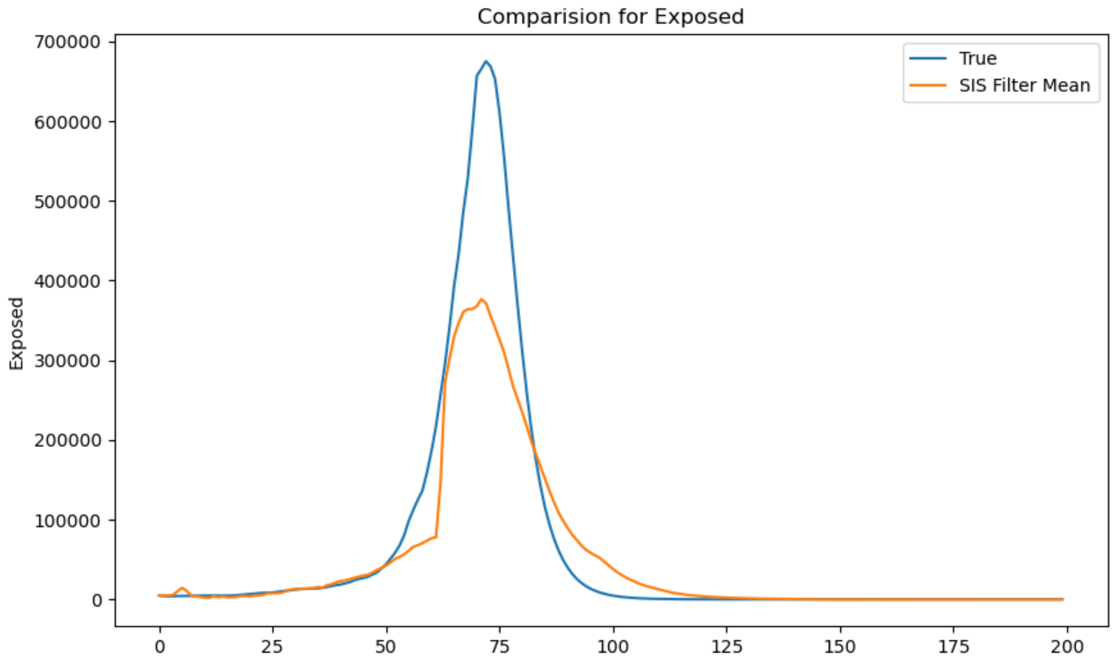
[9]:
```
plt.plot(alpha[0,0,:], label = 'True')
plt.plot(sis.alpha_filt[0,0,:], label = 'SIS Filter Mean')
plt.title('Comparision for Susceptible')
plt.ylabel('Susceptible')
plt.legend()
```

[9]:
```
<matplotlib.legend.Legend at 0x1d348679fd0>
```



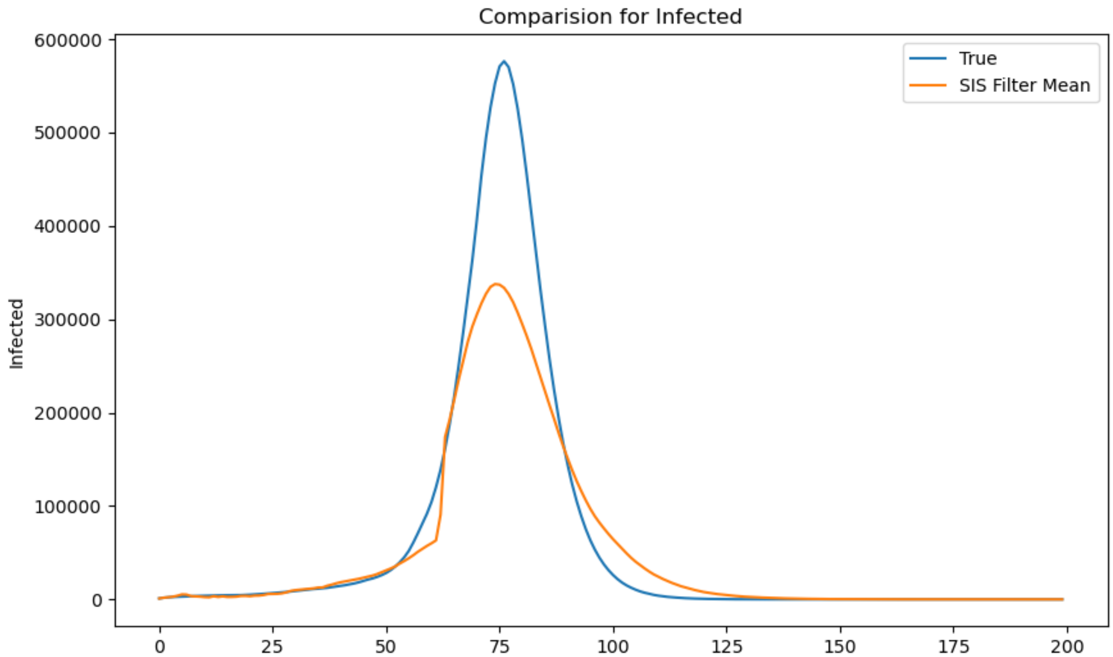
[10]:
```
plt.plot(alpha[1,0,:], label = 'True')
plt.plot(sis.alpha_filt[1,0,:], label = 'SIS Filter Mean')
plt.title('Comparision for Exposed')
plt.ylabel('Exposed')
plt.legend()
```

[10]:
```
<matplotlib.legend.Legend at 0x1d348704040>
```



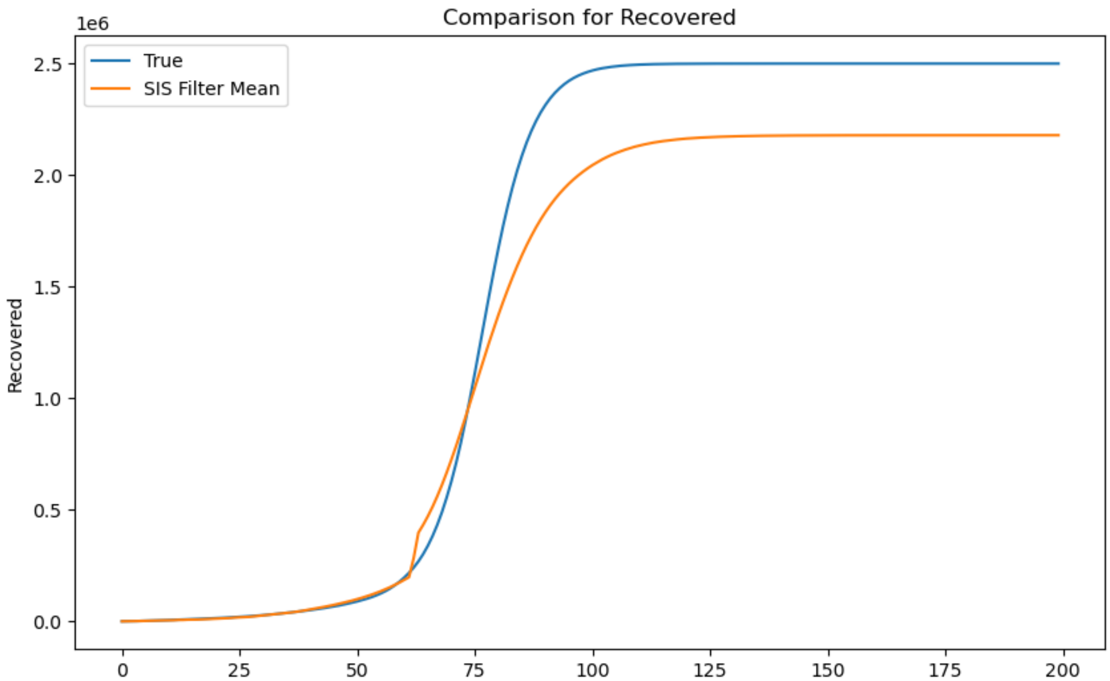
[11]:
```
plt.plot(alpha[2,0,:], label = 'True')
plt.plot(sis.alpha_filt[2,0,:], label = 'SIS Filter Mean')
plt.title('Comparision for Infected')
plt.ylabel('Infected')
plt.legend()
```

[11]: <matplotlib.legend.Legend at 0x1d348795550>



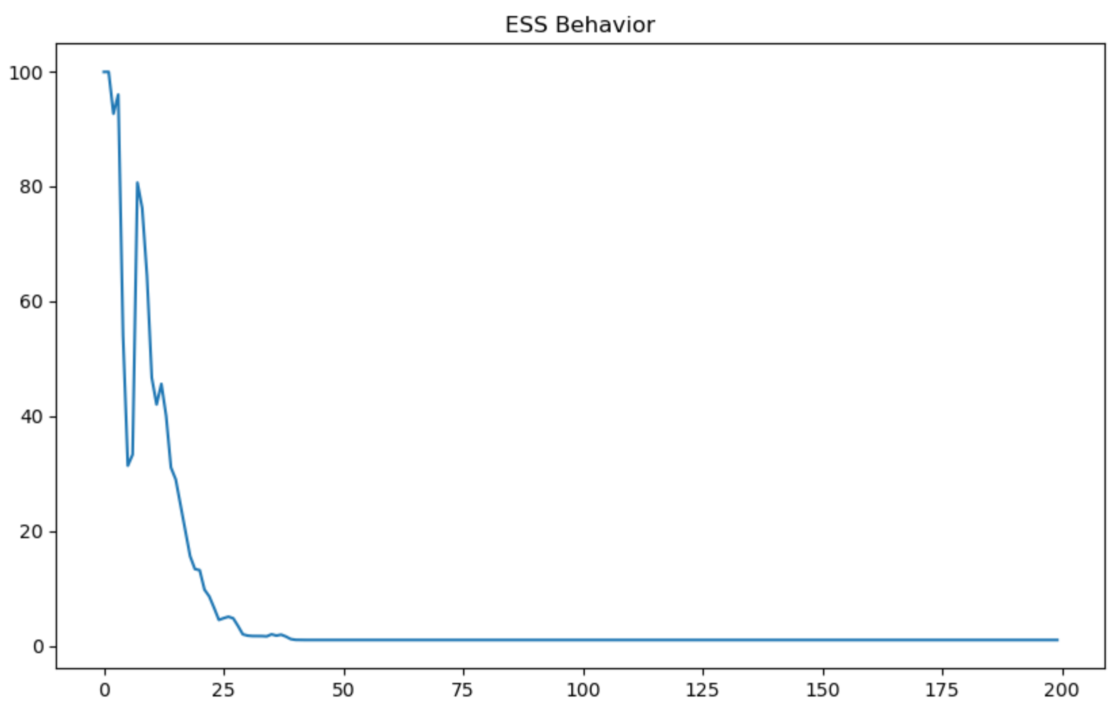
[12]:
```python
s_true = alpha[0,0,:]
e_true = alpha[1,0,:]
i_true = alpha[2,0,:]
r_true = population_size - (s_true + e_true + i_true)

s_sis = sis.alpha_filt[0,0,:]
e_sis = sis.alpha_filt[1,0,:]
i_sis = sis.alpha_filt[2,0,:]
r_sis = population_size - (s_sis + e_sis + i_sis)

plt.plot(r_true, label = 'True')
plt.plot(r_sis, label = 'SIS Filter Mean')
plt.title('Comparison for Recovered')
plt.ylabel('Recovered')
plt.legend()
```

[12]:
```
<matplotlib.legend.Legend at 0x1d3488ccca0>
```



[13]:
```
plt.plot(sis.N_eff)
plt.title('ESS Behavior')
```

[13]: Text(0.5, 1.0, 'ESS Behavior')



1.4 3.4 Sequential Importance Sampling with Resampling

Pick the same simulated trajectory as for the previous section.

Q4: Implement the **Sequential Importance Sampling with Resampling** or **Bootstrap Particle Filter** by completing the code below.

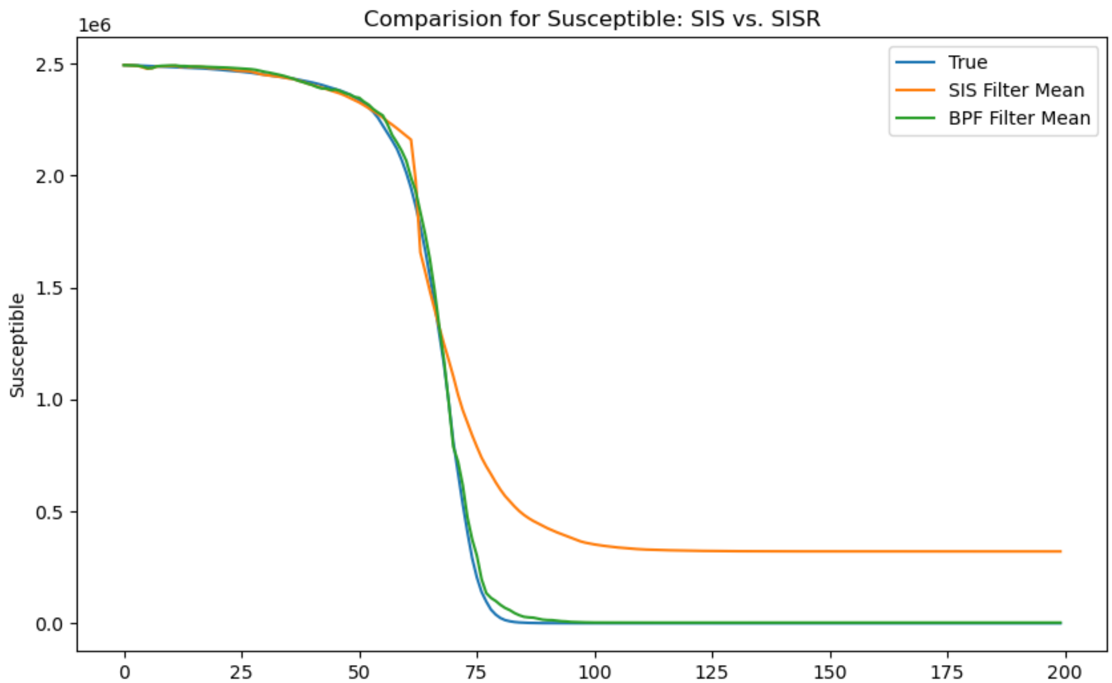
[14]:
```
def bpf(model, y, numParticles):
    d = model.d
    n = len(y)
    N = numParticles

    # Allocate memory
    particles = np.zeros((d, N, n), dtype = float)  # All generated particles
    logW = np.zeros((1, N, n))  # Unnormalized log-weight
    W = np.zeros((1, N, n))  # Normalized weight
    alpha_filt = np.zeros((d, 1, n))  # Store filter mean
    N_eff = np.zeros(n)  # Efficient number of particles
    logZ = 0.  # Log-likelihood estimate

    # Filter loop
    for t in range(n):
        # Sample from "bootstrap proposal"
        if t == 0:  # Initialize from prior
            particles[:, :, 0] = model.sample_state(N=N)
        else:  # Resample and propagate according to dynamics
            ind = np.random.choice(N, N, replace=True, p=W[0, :, t-1])
            resampled_particles = particles[:, ind, t-1]
            # print(resampled_particles.shape)
            particles[:, :, t] = model.sample_state(resampled_particles, N = N)

        # Compute weights
        logW[0, :, t] = model.log_lik(y[t], particles[:, :, t])
        W[0, :, t], logZ_now = exp_norm(logW[0, :, t])
        logZ +=  logZ_now # Update log-likelihood estimate
        N_eff[t] =  ESS(W[0,:,t])

        # Compute filter estimates
        alpha_filt[:, 0, t] = np.sum(particles[:, :, t]*W[0, :, t],axis=1)

    return smc_res(alpha_filt, particles, W, N_eff = N_eff, logZ = logZ)
```

Q5: Use the same simulated trajectory as above and run the BPF algorithm using $N = 100$ particles. Show plots comparing the filter means from the Bootstrap Particle Filter algorithm with the underlying truth of the Infected, Exposed and Recovered. Also show a plot of how the ESS behaves over the run. Compare this with the results from the SIS algorithm.

[15]:
```
bpf_sisr = bpf(model, y_obs[0,0,:], 100)
```

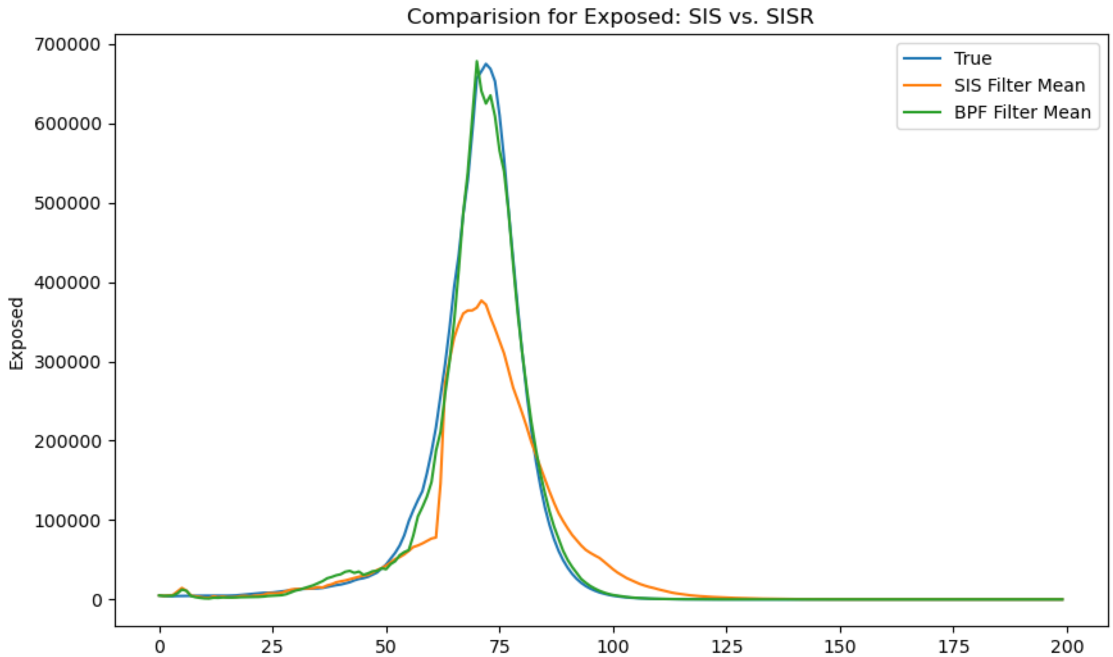
[41]:
```
plt.plot(alpha[0,0,:], label = 'True')
plt.plot(sis.alpha_filt[0,0,:], label = 'SIS Filter Mean')
plt.plot(bpf_sisr.alpha_filt[0,0,:], label = 'BPF Filter Mean')
plt.title('Comparision for Susceptible: SIS vs. SISR')
plt.ylabel('Susceptible')
plt.legend()
```

[41]: <matplotlib.legend.Legend at 0x1d352411be0>



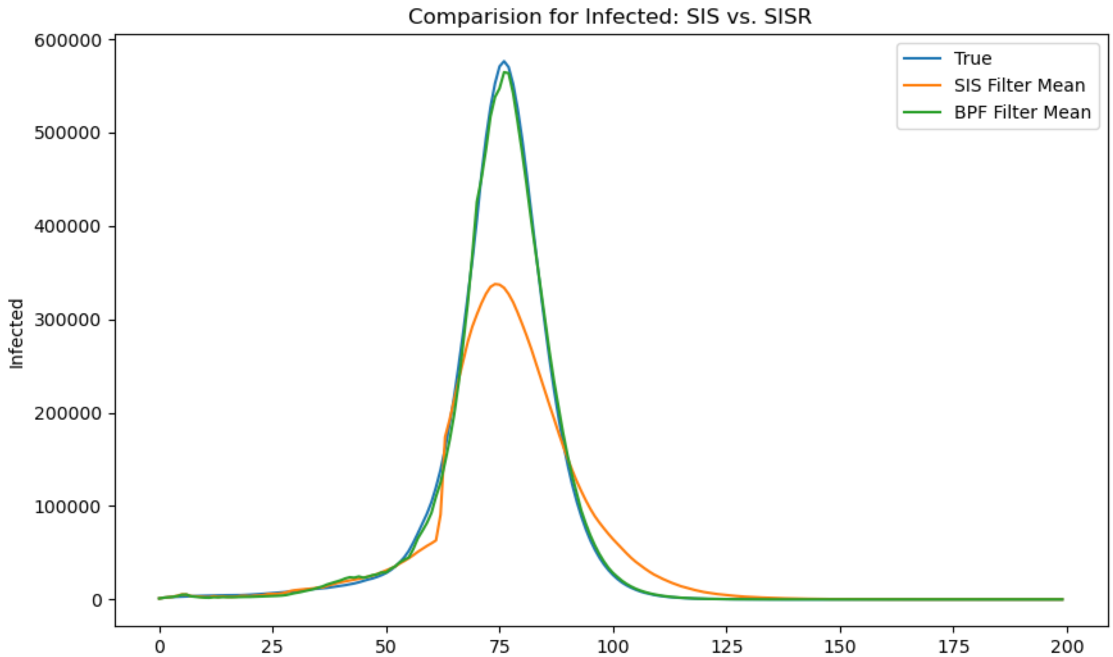
[42]:
```
plt.plot(alpha[1,0,:], label = 'True')
plt.plot(sis.alpha_filt[1,0,:], label = 'SIS Filter Mean')
plt.plot(bpf_sisr.alpha_filt[1,0,:], label = 'BPF Filter Mean')
plt.title('Comparision for Exposed: SIS vs. SISR')
plt.ylabel('Exposed')
plt.legend()
```

[42]: <matplotlib.legend.Legend at 0x1d3524d4fd0>



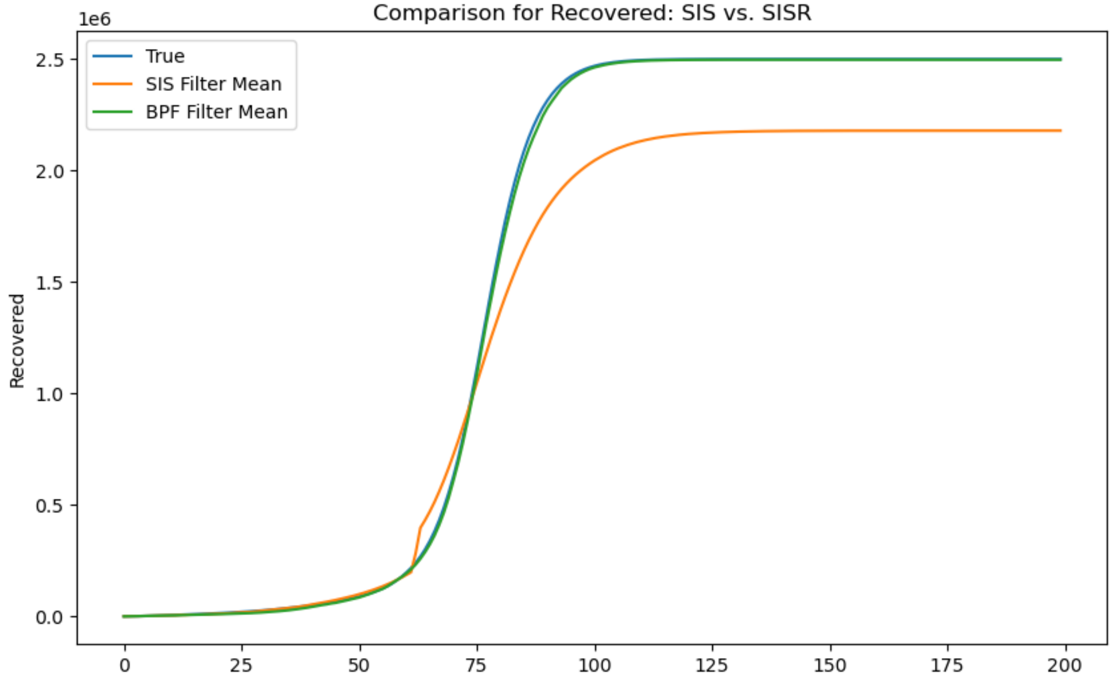
[43]:
```
plt.plot(alpha[2,0,:], label = 'True')
plt.plot(sis.alpha_filt[2,0,:], label = 'SIS Filter Mean')
plt.plot(bpf_sisr.alpha_filt[2,0,:], label = 'BPF Filter Mean')
plt.title('Comparision for Infected: SIS vs. SISR')
plt.ylabel('Infected')
plt.legend()
```

[43]: <matplotlib.legend.Legend at 0x1d352543f70>



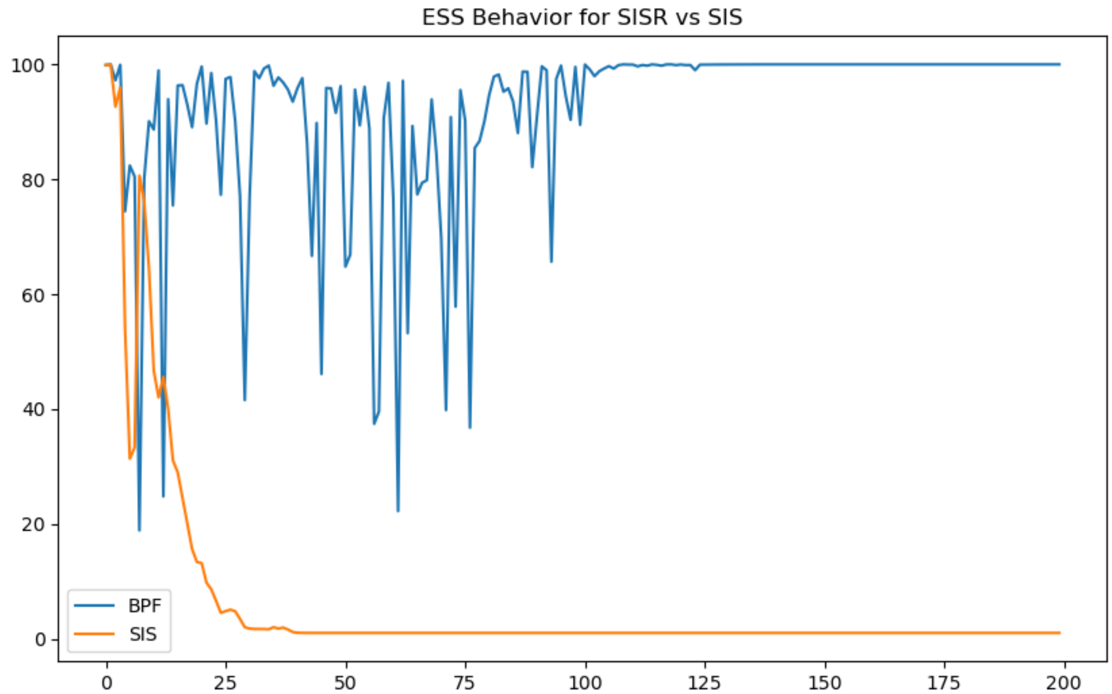
[44]:
```
s_sisr = bpf_sisr.alpha_filt[0,0,:]
e_sisr = bpf_sisr.alpha_filt[1,0,:]
i_sisr = bpf_sisr.alpha_filt[2,0,:]
r_sisr = population_size - (s_sisr + e_sisr + i_sisr)

plt.plot(r_true, label = 'True')
plt.plot(r_sis, label = 'SIS Filter Mean')
plt.plot(r_sisr, label = 'BPF Filter Mean')
plt.title('Comparison for Recovered: SIS vs. SISR')
plt.ylabel('Recovered')
plt.legend()
```

[44]: <matplotlib.legend.Legend at 0x1d3527b5bb0>



[45]:
```
plt.plot(bpf_sisr.N_eff, label = 'BPF')
plt.plot(sis.N_eff, label = 'SIS')
plt.title('ESS Behavior for SISR vs SIS')
plt.legend()
```

[45]: <matplotlib.legend.Legend at 0x1d3531f1a60>



From the above plot for ESS, with respect to SIS we can see that the weights degenerate quickly and after 25 time steps, the ESS becomes 1 meaning that all the weights except one have degenerated to zero. Whereas, with respect the ESS for SISR, the weights avoid degenerating because before we propagate the particles and calculate their weights, we first resample and only the good particles are propagated forward after which the weights are calculated.

1.5 3.5 Estimating the data likelihood and learning a model parameter

In this section we consider the real data and learning the model using this data. For simplicity we will only look at the problem of estimating the ρ parameter and assume that others are fixed.

You are more than welcome to also study the other parameters.

Before we begin to tweak the parameters we run the particle filter using the current parameter values to get a benchmark on the log-likelihood.

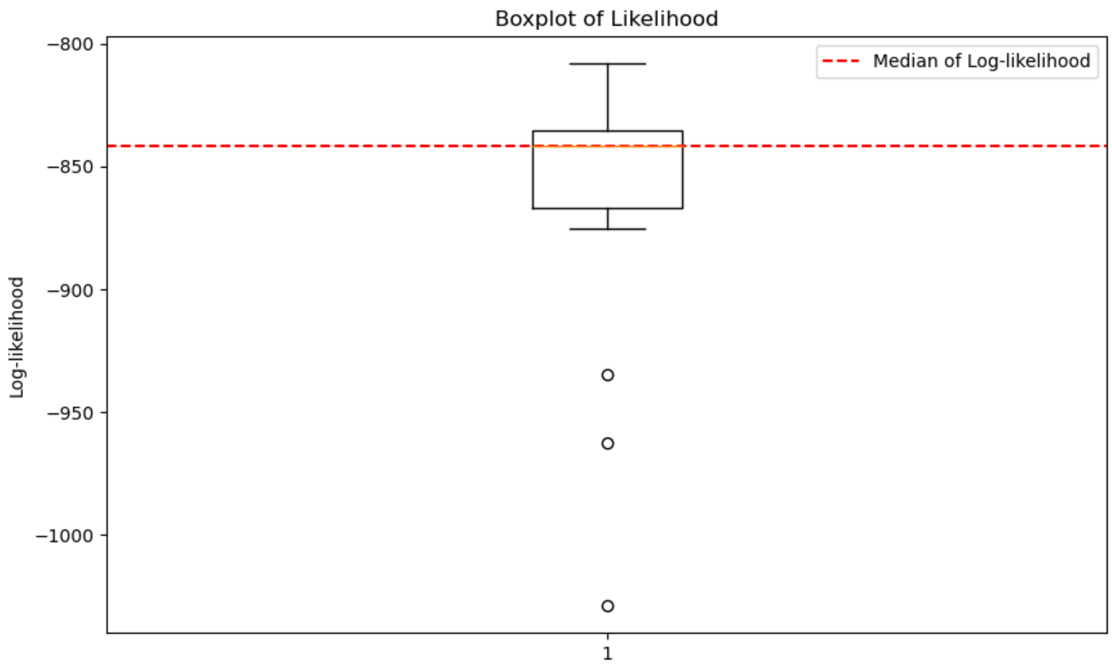
Q6: Run the bootstrap particle filter using $N = 200$ particles on the real dataset and calculate the log-likelihood. Rerun the algorithm 20 times and show a box-plot of the log-likelihood.

[21]:
```
bpf_logliks = []
for ii in range(20):
    bpf_ll = bpf(model, y_sthlm, 200)
    bpf_logliks.append(bpf_ll.logZ)
```

[22]:
```
plt.boxplot(bpf_logliks)
```

```
plt.axhline(y = np.median(bpf_logliks), color = 'r', linestyle = '--', label =␣
↪'Median of Log-likelihood')
plt.ylabel('Log-likelihood')
plt.title('Boxplot of Likelihood')
plt.legend()
```

[22]: <matplotlib.legend.Legend at 0x1d346698580>



Q7: Make a grid of the ρ parameter in the interval $[0.1, 0.9]$. Use the bootstrap particle filter to calculate the log-likelihood for each value. Run the bootstrap particle filter using $N = 200$ multiple times (at least 20) per value and use the average as your estimate of the log-likelihood. Plot the log-likelihood function and mark the maximal value.

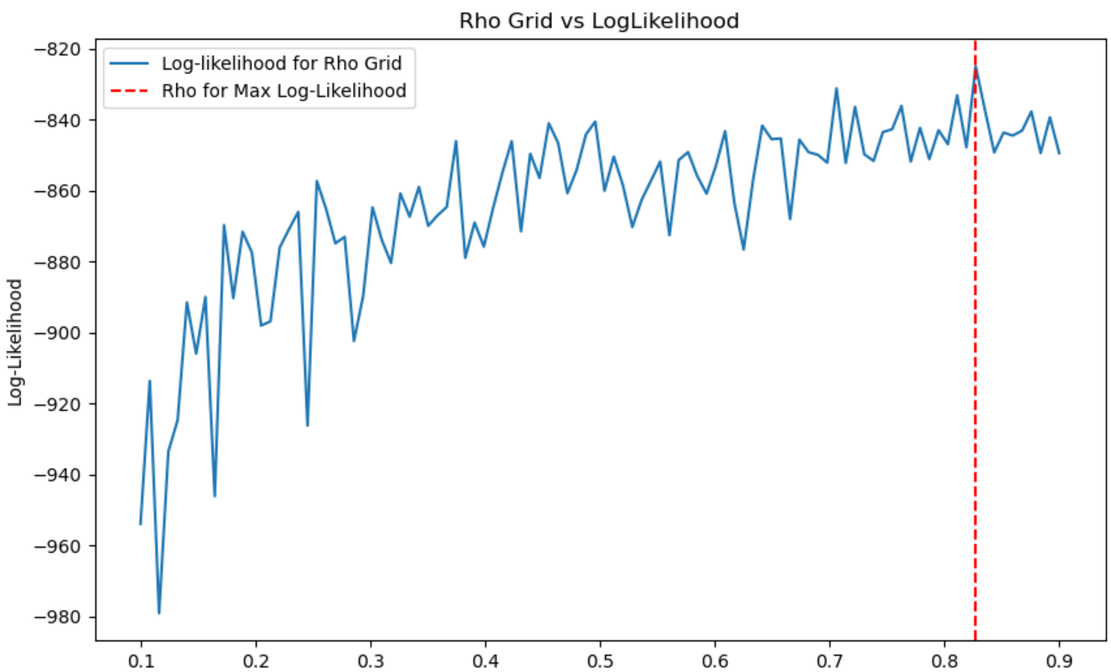
(hint: use np.linspace to create a grid of parameter values)

[23]:
```
rho_grid = np.linspace(0.1, 0.9, num = 100)

opt_loglik = []
for rho_i in rho_grid:
    params = Param(pse, pei, pir, pic, rho_i, sigma_epsilon, init_mean,␣
↪population_size)
    model_rho = SEIR(params)
    loglik = []
    for jj in range(20):
        bpf_rho = bpf(model_rho, y_sthlm, 200)
        loglik.append(bpf_rho.logZ)
```

```
    opt_loglik.append(np.mean(loglik))
```

[24]:
```
plt.plot(rho_grid, opt_loglik, label = 'Log-likelihood for Rho Grid')
opt_rho = rho_grid[np.argmax(opt_loglik)]
plt.axvline(x = opt_rho, color = 'r', linestyle = '--', label = 'Rho for Max␣
 ↪Log-Likelihood')
plt.ylabel('Log-Likelihood')
plt.title('Rho Grid vs LogLikelihood')
plt.legend()
```

[24]: <matplotlib.legend.Legend at 0x1d34b461a60>



[46]:
```
print('The  optimal value of the parameter rho is:', opt_rho)
```

```
The  optimal value of the parameter rho is: 0.8272727272727273
```

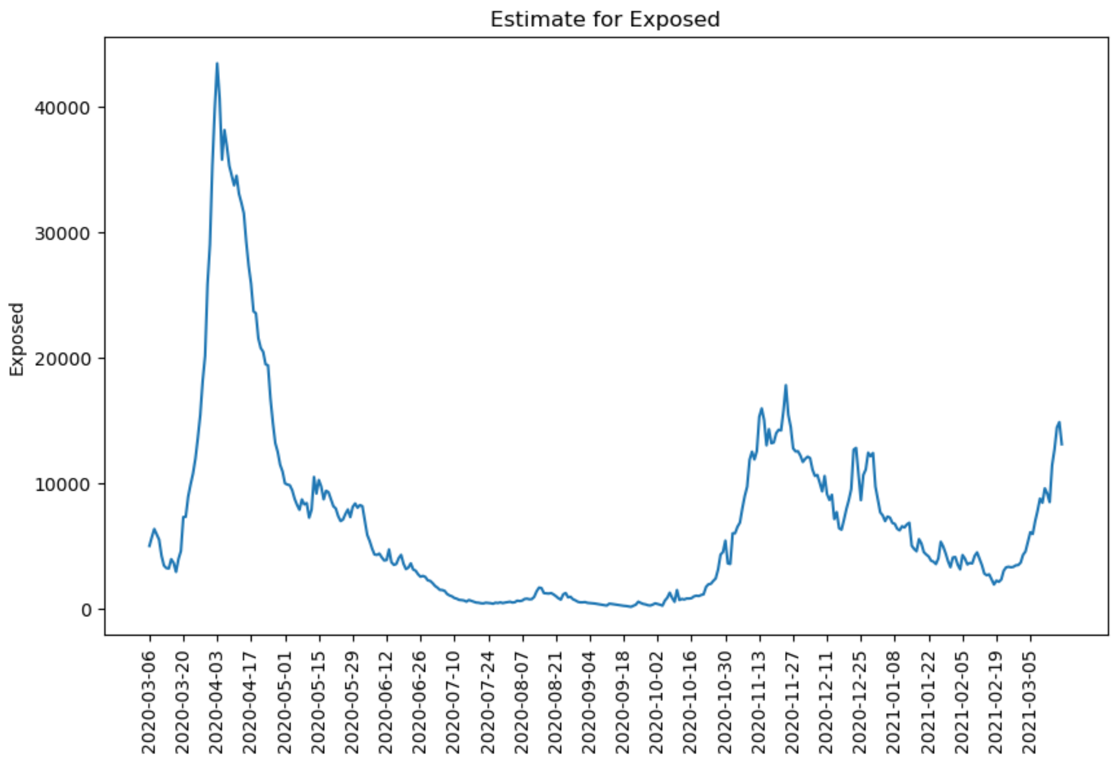
Q8: Run the bootstrap particle filter on the full dataset with the optimal ρ value. Present a plot of the estimated Infected, Exposed and Recovered states.

[25]:
```
params_opt = Param(pse, pei, pir, pic, opt_rho, sigma_epsilon, init_mean,␣
 ↪population_size)
model_opt_rho = SEIR(params)
bpf_opt_rho = bpf(model_opt_rho, y_sthlm, 200)
```

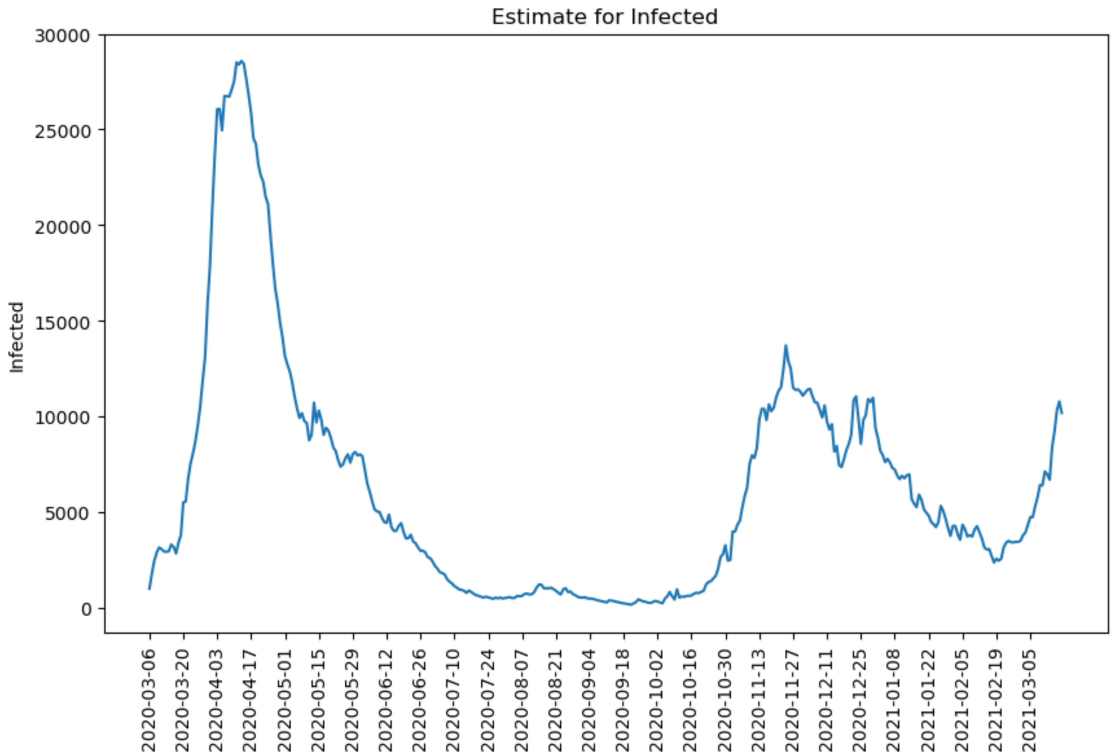
[26]:
```
bpf_opt_rho.logZ
```

[26]: -856.6048166306243

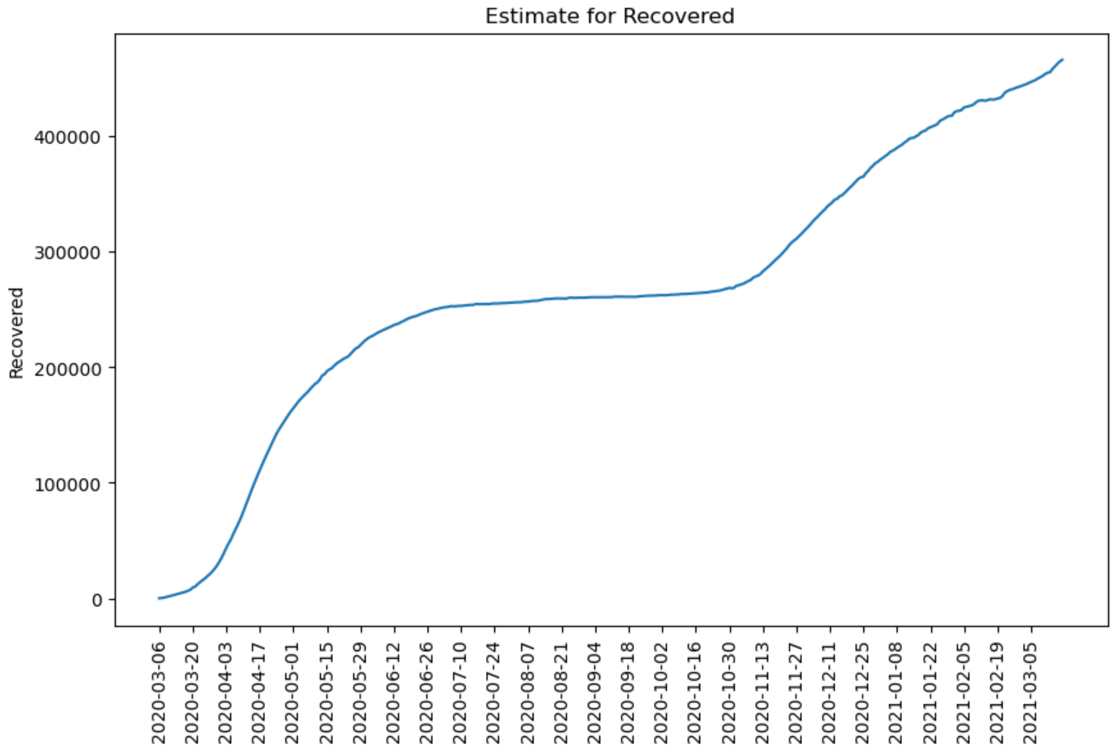
[27]:
```
plt.plot(u_sthlm, bpf_opt_rho.alpha_filt[1,0,:])
plt.title('Estimate for Exposed')
plt.ylabel('Exposed')
plt.xticks(range(0, ndata, 14), u_sthlm[::14], rotation = 90)
plt.show()
```



[28]:
```
plt.plot(u_sthlm, bpf_opt_rho.alpha_filt[2,0,:])
plt.title('Estimate for Infected')
plt.ylabel('Infected')
plt.xticks(range(0, ndata, 14), u_sthlm[::14], rotation = 90)
plt.show()
```



[29]:
```
s_est = bpf_opt_rho.alpha_filt[0,0,:]
e_est = bpf_opt_rho.alpha_filt[1,0,:]
i_est = bpf_opt_rho.alpha_filt[2,0,:]
r_est = population_size - (s_est + e_est + i_est)

plt.plot(u_sthlm, r_est)
plt.title('Estimate for Recovered')
plt.ylabel('Recovered')
plt.xticks(range(0, ndata, 14), u_sthlm[::14], rotation = 90)
plt.show()
```


Estimate for Recovered
Recovered
400000
300000
200000
100000
0
2020-03-06
2020-03-20
2020-04-03
2020-04-17
2020-05-01
2020-05-15
2020-05-29
2020-06-12
2020-06-26
2020-07-10
2020-07-24
2020-08-07
2020-08-21
2020-09-04
2020-09-18
2020-10-02
2020-10-16
2020-10-30
2020-11-13
2020-11-27
2020-12-11
2020-12-25
2021-01-08
2021-01-22
2021-02-05
2021-02-19
2021-03-05

[ ]: