Ensemble data assimilation methods
Experiments with the Lorenz 1963 model

Javier Amezcua, University of Reading

1 Introduction

In this practical we will explore the implementation of two variations of the ensemble Kalman filter. The first one is the perturbed-observations ensemble Kalman filter (EnKF; Burgers et al, 1998). The second one is a deterministic ensemble square root filter (Tippett et al, 2003), in particular the ensemble transform Kalman filter (Bishop et al, 2001; Wang et al, 2004). Our experiments will use the Lorenz (1963) model.

The computer instructions to run the experiments of this practical are contained in the file **Controls63Ens.m**. You will be asked to modify values of certain variables in these instructions. You do not need to modify any of the other files.

Controls63Ens.m is divided in cells (by using % % througout the file). An individual cell can be run either by hitting control+enter, or using the 'evaluate cell' command in the cell toolbar of the Matlab editor window. **Controls63Ens.m** contains instructions to do the following tasks:

- Run the Lorenz 1963 model
- Generate synthetic observations
- Perform data assimilation with EnKF and ETKF
- Compute diagnostics (e.g. root mean squared error)
- Produce and display graphs

2 Set up for the experiment

2.1 Nature run

The first part of the practical is to get a nature run, corresponding to cells 1.a and 1.b. The Lorenz 1963 model has 3 variables: $\{x, y, z\}$, hence the state vector is $\mathbf{x} = [x, y, z]^T$. The time evolution of these variables is described by the following 3 coupled ordinary differential equations:

$$\begin{aligned} \frac{dx}{dt} &= \sigma(y - x) \\ \frac{dy}{dt} &= x(\rho - z) - y \\ \frac{dz}{dt} &= xy - \beta z \end{aligned} \tag{1}$$

These equations have 3 parameters: $\{\sigma, \beta, \rho\}$. For some values of these parameters the model is chaotic, in particular we choose $\{\sigma = 10,\ \beta = 8/3,\ \rho = 28\}$. The model is integrated using the 4^{th} order Runge-Kutta method with $\Delta t = 0.01$ (adimensional units).

In the program, the default initial condition is $\mathbf{x}_0 = [-10, 10, 20]^T$ and the maximum time of integration is $t_{max} = 10$. You can integrate the model by running cell 1.a, and you can visualize the evolution of the variables by running cell 1.b. This will produce two plots: one showing the time evolution of each of the 3 state variables, and one displaying the trajectories of the variables in state space. This is shown in figure 1. You can rotate this last plot to look at the figure from different perspectives.

Let us start the experiments now:

a) Select different initial conditions for the model, e.g. $\{\mathbf{x}_0 = [-1, -1, 0]^T, \mathbf{x}_0 = [0, 20, 10]^T, \mathbf{x}_0 = [5, 5, 5]^T\}$ and different final times: $t_{max} = \{2, 5, 10, 30, 50\}$. What are the differences in the plot you get? In particular, what happens to the 3D plot as the final time of integration increases?

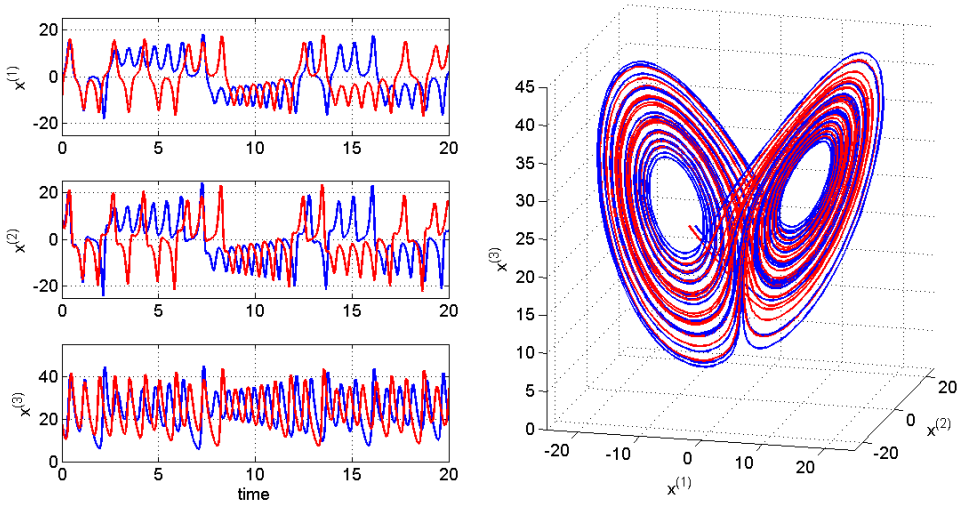

Figure 1: Nature run of the Lorenz 1963 model. The left side of the figure displays the time evolution of the 3 variables separately. The right side of the figure displays the trajectories in phase space (time is implicit in this figure). Two nearby initial points have been chosen; their different evolution highlights the chaotic nature of the model.

2.2 Observations

Now we will generate synthetic observations from the trajectory started at $\mathbf{x}_0 = [-10, 10, 20]$ and ran until $t_{max} = 40$. The instructions to run this part of the exercise are contained in cells 2.a, 2.b and 2.c. The observations are generated by simulation the observation equation:

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \boldsymbol{\eta} \tag{2}$$

where $\mathbf{H}$ is a linear observation operator and $\boldsymbol{\eta}$ is a random variable with distribution $N(\boldsymbol{\mu} = \mathbf{0}, \mathbf{R})$. The observation error covariance matrix is $\mathbf{R} = \sigma^2\mathbf{I}$, i.e. a diagonal matrix with the same variance for each observation.

We can play with 3 aspects of the observational network: how many of the 3 variables we observe (denoted in cell 2.a. as option 1, option 2 and option 3), how frequently do we observe (denoted by

the variable $freqobs$), and the variance of the observations (denoted by the variable $varobs$). One can visualize the synthetic observations with the plotting routines contained in cell 2.c.

c) Run the code with the following choices: $freqobs = \{8, 24, 40\}$ and $varobs = \{1, 2, 8\}$. How do you think the DA problem will differ for these different situations?

d) Can you design an observation operator which corresponds to only observing y?

3 Using the ensemble Kalman filter

3.1 State estimation

The instructions for this part of the practical are contained in cells 3.a, 3.b, 4.a and 4.b. Cell 3.a runs the data assimilation procedure and gets 4 variables: $\mathbf{X}^b$ and $\mathbf{X}^a$, which are the background and analysis ensembles, and $\bar{\mathbf{x}}^b$ and $\bar{\mathbf{x}}^b$ which are the background and analysis ensemble means. Cell 3.b displays the nature run, the observations, as well as the background and analysis values for all state variables. It should result obvious that background and assimilation differ only at observation times. This was not the case with 4DVar, which is a smoother and hence modifies the whole trajectory of the state variables between observation times.

To evaluate the performance of the data assimilation, one can use the root-mean squared error of the analysis mean with respect to the truth. For every time step, this is computed as:

$$RMSE(t) = \sqrt{\frac{\mathbf{d}(t)^T \mathbf{d}(t)}{3}} \tag{3}$$

where $\mathbf{d}(t) = \bar{\mathbf{x}}^a(t) - \mathbf{x}^{true}(t)$. This is computed by running cell 4.a and plotted by running 4.b. The spread (standard deviation) of the ensemble at anytime is also computed and displayed. Under ideal conditions, these two quantities (analysis RMSE and analysis ensemble spread) should be comparable.

After this description, let us start the experiments. We can play with 3 implementation aspects. The first is the type of ensemble Kalman filter: $met = \{SEnKF, ETKF\}$, ensemble size $M = \{3, 10, \ldots\}$, and inflation factor $\rho = \{0, 0.1, \ldots\}$, where inflation is applied as: $\mathbf{X}^b \rightarrow (1 + \rho)\mathbf{X}^b$.

Let us start experimenting with more ensemble members than variables. This often an unrealistic case in practice but it will illustrate the effect of ensemble size.

e) Choose $M = \{40, 20, 10\}$ and use both flavours of EnKF and different observed variables and frequency of observations (use the recommended values from the previous sections). What should you choose for ρ in this case?

f) After running cell 3.b., you will notice there are two plots showing the model evolution. What is the difference between the 2 of them? What happens as you decrease the ensemble size?

For the rest of this section let us set $M = 3$, the minimum possible size. In this case, this size is the same as the number of state variables.

g) First, keep $\rho = 0$. Try different number of observed variables, different frequencies of observations, and the 2 types of EnKF. How is the performance in this case? Is the filter performing well?

h) We will try to 'fix' the filter by introducing inflation. Starting from $\rho = 0$ (no inflation), increase the value until you find a 'decent' performance of the filter. This process is known as 'tuning'. Again, experiment with different observed variables, different frequencies of observations, and the two EnKF methods. Hint: inflation is very sensitive to the frequency of observations. For 'frequent' observations (every 8 model steps), start with $\rho = \{0.01, 0.02, \ldots\}$, for 'infrequent' observations (every 24 and 40 steps) try values such as $\rho = \{0.1, 0.2, \ldots\}$.

3.2 Parameter estimation

In the previous section we used EnKF and ETKF for state estimation. In many applications, the parameters of the model are not known with absolute certainty. Having incorrect values for the parameters can lead to very different evolutions of the model. This is shown in figure 2. The left panel shows the trajectory generated with the 'true' parameters, whereas the other 2 panels show trajectories generated 'wrong' parameters.

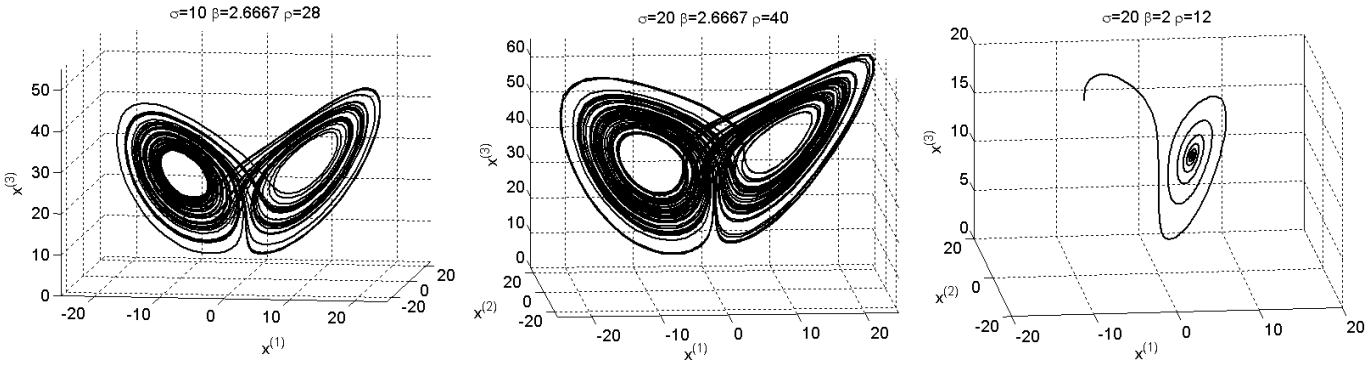

Figure 2: Trajectories of the Lorenz 1963 model with 3 different sets of parameters.

To estimate parameters we simply extend the state vector to: $\mathbf{x} = [x, y, z, \sigma, \beta, \rho]^T$. There are two important properties of the parameters. First, they are unobserved variables, but they can be updated via the sample covariance between them and the state variables. Second, the model evolution for them is just the identity, i.e. their values do not change during the forecast step, e.g. for ρ this would mean $\rho^b(t_{obs} = j) = \rho^a(t_{obs} = j - 1)$ (we will revisit this later).

The instructions to run this part of the practical are found in the file **ControlsL63Pe.m**. As before, this file is divided in cells. Cells 1.a and 1.b. contain the evolution of the model.

i) We will consider the values $\{\sigma = 10, \beta = 8/3, \rho = 28\}$ as the 'true parameters' for our next experiments. However, run the model with the following sets of parameters: $\{\sigma = 9, \beta = 3, \rho = 30\}$, $\{\sigma = 6, \beta = 7, \rho = 14\}$, $\{\sigma = 0, \beta = 0, \rho = 0\}$. How do the trajectories differ? Which cases will be more challenging for parameter estimation?

The next part of **ControlsL63Pe.m** are cells 2.a and 2.b. These are exactly as in **ControlsL63En.m**: they generate observations and display plots of these observations. For the moment, let us choose the following configuration: observe x and z only, with a frequency of $freqobs = 12$ model steps, and an observational variance $\sigma^2 = 2$. You can come back and experiment with different values at the end of the practical if you have time left.

The parameter estimation experiments are controlled by cell 3.a and 3.b. In this case, we will only experiment using ETKF with $M = 20$ ensemble members, and no inflation $\rho = 0$. Again, you can experiment with EnKF and other ensemble sizes if you have time left at the end of the practical. You will notice a parameter called α. For the moment leave it as $\alpha = 0$, its function will become clear later. Cell 3.b. displays the evolution of the nature, background, and analysis values for the state variables and the parameters. Cells 4.a and 4.b compute and plot diagnostics (analysis RMSE and analysis ensemble spread).

i) Run the ETKF with parameter estimation of cell 3.a. You will need initial guesses for the parameters. First try $\{\sigma = 9, \beta = 3, \rho = 30\}$. How do the trajectories of the state variables look? What about the value of the parameters? Now, do the same for the initial values $\{\sigma = 6, \beta = 7, \rho = 14\}$ and $\{\sigma = 0, \beta = 0, \rho = 0\}$. What happens in this case?

As you observed, the parameter estimation can result challenging when our guess from the parameters is far from the true ones, or when the behaviour of the model changes drastically after a critical value of one or more parameters. We can help the filter if we add a random kick to the parameters in the forecast step, allowing more possible values are explored. For example, for the parameter ρ this would mean: $\rho^f(t_{obs} = j) = \rho^a(t_{obs} = j - 1) + \epsilon$, where ϵ is a univariate random variable with distribution $N(0, \alpha^2)$.

j) Run the parameter estimation routine for $\{\sigma = 6, \beta = 7, \rho = 14\}$ and $\{\sigma = 0, \beta = 0, \rho = 0\}$ by increasing the value of α progressively: $\alpha = \{0, 0.1, \ldots, 1, \ldots, 10\}$. What happens to the estimated model trajectory and parameter values? What happens for very large values of α?

4 References

C. H. Bishop, B. J. Etherton and S. J. Majumdar, 2001. Adaptive sampling with the ensemble transform Kalman filter. Part I: Theoretical aspects. *Mon. Wea. Rev.*, **129** (3), 420-436.

G. Burgers, P. J. van Leeuwen and G. Evensen, 1998. Analysis scheme in the ensemble Kalman filter. *Mon. Wea. Rev.*, **126**, 1719-1724.

R. E. Kalman, 1960. A new approach to linear filtering and prediction problems. *Trans. of the ASME, Jour. of Bas. Engin. D*, **82**, 35-45.

E. N. Lorenz, 1963. Deterministic non-periodic flow. *J. Atmos. Sci.*, **20**, 130-141.

M. K. Tippett, J. L. Anderson, C. H. Bishop, T. M. Hamill, J. S. Whitaker, 2003. Ensemble square-root filters. *Mon. Wea. Rev.*, **131**, 1485–1490.

X. Wang, C. H. Bishop and S. J. Julier, 2004. Which is better, an ensemble of positive-negative pairs or a centered spherical simplex ensemble? *Mon. Wea. Rev.*, **132** (7), 1590-1605.