THE UNIVERSITY OF READING

Data Assimilation Using Observers

Anne K. Griffith and Nancy K. Nichols

Numerical Analysis Report 11/94

DEPARTMENT OF MATHEMATICS

Data Assimilation Using Observers

Anne K. Griffith and Nancy K. Nichols

Numerical Analysis Report 11/94

Department of Mathematics
P.O.Box 220
University of Reading
Whiteknights
Reading
RG6 2AX
United Kingdom

Abstract

This report gives a brief introduction to data assimilation and discusses how this can be treated as an observer design problem. The particular observer design investigated here and ours to make the resulting observer system as robust as possible to perturbations in the model equations.

This observer is tested in the context of data assimilation for a simple discrete model. Issues investigated include the choice of eigenvalues to be assigned to the observer, a choice of a suitable observation matrix, and modifications for the case where observations occur less frequently.

Finally, the choice of the weighting matrix in the Cressman data assimilation scheme is compared to the feedback matrix of the observer system. This facilitates a theoretical evaluation of the Cressman scheme.

cknowledgements

The work described in this report was carried out under an EPSRC CASE studentship with the UK Meteorological Office. We would like to thank Andrew Lorenc of the Met Office for his help and support with this work.

C ntents

terms of observers.

In Section 2, theory for developing a robust dynamic observer for a simple model is put forward, and an algorithm for implementing the observer is given. Section 3 describes experiments carried out with such an observer, using the heat equation with the "theta method" discretisation as a model. Experiment 1 investigates the best eigenvalue assignment in the development of the observer, aiming for quick convergence of the observer to the true solution. Experiment 2 looks at a suitable choice of the observation matrix C and Experiment 3 investigates the performance of the observer when the observations are infrequent. In Section 4 the Cressman scheme, a simple "successive correction" method for data assimilation, is compared to the observer method, both theoretically and practically. Section 5 summarises the conclusions drawn from the study, and gives suggestions for future work.

2 Eigenvalue assignment

We consider the follo

We want to construct the feedback matrix G so that $\hat{\mathbf{w}}^k \to \mathbf{w}^k$ as $k \to \infty$, regardless of the true initial condition $\mathbf{w}^0$, which is unknown.
Subtracting (2.3) from (2.1) and using (2.2) we have:

$$E(\mathbf{w}^{k+1} - \hat{\mathbf{w}}^{k+1}) = A(\mathbf{w}^k - \hat{\mathbf{w}}^k) - GC(\mathbf{w}^k - \hat{\mathbf{w}}^k), \tag{2.4}$$

so defining $\mathbf{e}^k = \mathbf{w}^k - \hat{\mathbf{w}}^k$, we have the error equation

$$E\mathbf{e}^{k+1} = A\mathbf{e}^k - GC\mathbf{e}^k. \tag{2.5}$$

Hence, for $\mathbf{e}^k \to 0$ as $k \to \infty$, we require that the eigenvalues of $E^{-1}(A - GC)$ have modulus less than unity.

If E is invertible (as assumed) and $\mathcal{S}$ is observable, then we can construct G to do this; in fact we can choose G to assign any eigenvalues we wish to the system $\mathcal{S}'$ [3]. Since this "inverse eigenvalue problem" is not uniquely determined [6], we have a certain amount of freedom to choose the eigenvectors as well. We can use this freedom to make the system as robust to perturbations as possible.

In [6] it is shown that for a robust system we require $\mathrm{cond}(X)$ to be as small as possible, where X is the modal matrix whose columns are the right eigenvectors corresponding to our chosen eigenvalues. Our objective, then, is for a suitable eigen*structure* assignment.

2.2 Eigenstructure assignment - theory

Eigenvalue assignment

We suppose that the set of eigenvalues we wish to assign is

$$\Lambda = \{\lambda_1, \lambda_2, ..., \lambda_n\}; \tag{2.6}$$

where

$$\lambda_i \in \mathbb{C}, \quad |\lambda_i| < 1, \quad \text{and} \quad \lambda_i \in \Lambda \Rightarrow \bar{\lambda} \in \Lambda \qquad \text{for} \quad i = 1, ..., n. \tag{2.7}$$

We let $D = diag[\lambda_i]$ and let X be the modal matrix of right eigenvectors of $E^{-1}(A - GC)$ and Y be the modal matrix of $E^{-T}(A^T - C^T G^T)$. Then our problem is to choose G and X to satisfy

$$(A - GC)X = EXD, \tag{2.8}$$

or, equivalently, to choose Y and G^T to satisfy

$$(A^T - C^T G^T)Y = E^T Y D. \tag{2.9}$$

For our purposes, we work with equation (2.9).

If we calculate the QR decompostion of C^T, we find that

$$C^T = [\tilde{Q}_c, Q_c] \begin{bmatrix} R_o \\ 0 \end{bmatrix}, \tag{2.10}$$

where $\tilde{Q}_c$ is $n \times p$, Q_c is $n \times (n-p)$]

Eigenvalue assignment for robustness

The sensitivity of eigenvalue λ_i to perturbations in the components of A, E, C and G is given by

$$c_i = \frac{\|\mathbf{x}_i\|_2 \|E^T \mathbf{y}_i\|_2}{|\mathbf{y}_i^T E \mathbf{x}_i|}, \tag{2.16}$$

where $\mathbf{x}_i$ are the columns of X, and $\mathbf{y}_i^T$ the rows of Y^T (see [6]). If we scale $\mathbf{x}_i$ and $\mathbf{y}_i$ such that

$$\|E^T \mathbf{y}_i\|_2 = 1 \tag{2.17}$$

and

$$\left|\mathbf{y}_i^T E \mathbf{x}_i\right| = 1, \tag{2.18}$$

then to minimize c_i we must minimize $\|\mathbf{x}_i\|_2$. For the optimal conditioning we must minimize all the c_i together, and hence we must choose the columns of X to minimize

$$\nu = \sum_i c_i^2 = \sum_i \|\mathbf{x}_i\|_2^2 \equiv \|X\|_F^2, \tag{2.19}$$

where $\|.\|$

We want $\mathbf{y}_i$ to be as close to orthogonal as possible to this set. Calculating the QR decompostion gives

$$Y_{-i} = [\ \tilde{}_i, \mathbf{z}_i] \begin{bmatrix} \tilde{Y}_i \\ 0 \end{bmatrix}, \tag{2.29}$$

where $[\ \tilde{}_i, \mathbf{z}_i]$ is orthogonal, $\tilde{Y}_i$ is upper triangular and nonsingular, and $\mathbf{z}_i$ is an $n \times 1$ vector. This gives us the vector $\mathbf{z}_i$ which is orthogonal to Y_{-i}, but $\mathbf{z}_i$ may not be in S_i, which would violate condition **a**). Choosing $\mathbf{y}_i$ to be the orthogonal projection of $\mathbf{z}_i$ into S_i ensures that $\mathbf{y}_i$ is as orthogonal as possible to the set Y_{-i} whilst satisfying condition **a**). So, after normalization to ensure (2.17) holds, we take

$$\mathbf{y}_i = S_i S_i^T \mathbf{z}_i / \| E^T S_i S_i^T \mathbf{z}_i \|_2. \tag{2.30}$$

When all the columns have been modified in this way, the same procedure can then be repeated to modify the $\mathbf{y}_i$ again, until $\|(Y^T E)^{-1}\|_F$ reaches a local minimum. The feedback matrix G can then be calculated from (2.12), using the Y derived.

This method for improving the robustness of the system can not be guaranteed to converge to the minimum possible value of $\|(Y^T E)^{-1}\|_F$, but in practice it has been found to reduce its initial value significantly.

2.4 n algorithm for a robust observer

1) Calculate the QR decomposition of C^T into

$$C^T = \left[\tilde{Q}_c, Q_c\right] \begin{bmatrix} R_o \\ 0 \end{bmatrix}. \tag{2.31}$$

2) For each $i = 1, .., n$,
calculate the QR decomposition of $(A - \lambda_i E)Q_c$ into

$$(A - \lambda_i E)Q_c = \left[\tilde{S}_i, S_i\right] \begin{bmatrix} R_i \\ 0 \end{bmatrix}. \tag{2.32}$$

3) Choose columns from each of the S_i as columns of the first guess Y, in such a way that Y is invertible.

4) For $i = 1, ..., n$, modify the columns $\mathbf{y}_i$ of Y as follows:

4a) calculate the QR decomposition of $Y_{-i} = \{\mathbf{y}_1, .., \mathbf{y}_{i-1}, \mathbf{y}_{i+1}, .., \mathbf{y}_n\}$ into

$$Y_{-i} = [\tilde{\ }_i, \mathbf{z}_i] \begin{bmatrix} \tilde{Y}_i \\ 0 \end{bmatrix}. \tag{2.33}$$

4b) project the vector $\mathbf{z}_i$ into space S_i to satisfy condition **a**) and then normalize:

$$\mathbf{y}_i = S_i S_i^T \mathbf{z}_i / \| E^T S_i S_i^T \mathbf{z}_i \|_2. \tag{2.34}$$

5) repeat step 4 until $\|(Y^T E)^{-1}\|_F$ reaches a local minimum.

6) using the Y found, let the feedback matrix be G where

$$G^T = R_o^{-1} \tilde{Q}_c^T (A^T Y - E^T Y D) Y^{-1}. \tag{2.35}$$

.

Implementation of the method

In this section, the theory of Section 2 is tried out for a simple model, which is introduced in Section 3.1. Experiment 1 described in Section 3.2 investigates how the choice of the set of eigenvalues Λ affects the results. In Experiment 2 (Section 3.3), different forms for the observation matrix C are developed, and the effect that these different choices have on the results is examined. Finally, Experiment 3 (Section 3.4) looks at how the method may be modified if observations are not available at every timestep.

3.1 The theta method for the 1D heat equation

The 1D heat equation on $x \in [0,1]$ with a point heat source of strength $\frac{1}{3}$ at $x = \frac{1}{4}$ is,

$$w_t = \sigma w_{xx} + \frac{1}{3}\delta(x - \frac{1}{4}), \tag{3.1}$$

where δ is the Dirac delta function. For this equation, with initial and boundary conditions

$$w(x,$$

1.

$$s(x) \begin{cases} > 0 & \text{if} \quad x = \frac{1}{4} \\ = 0 & \text{if} \quad x \neq \frac{1}{4} \end{cases} \tag{3.6}$$

2.

$$\int_0^1 s(u)du = \frac{1}{3}. \tag{3.7}$$

If we choose the vector $\mathbf{s}$ so that its j^{th} component s_j is given by

$$s_j = \begin{cases} \frac{1}{3\Delta x} & \text{if} \quad j = \frac{J}{4} \\ 0 & \text{otherwise,} \end{cases} \tag{3.8}$$

then s_j is a good approximation to $s(j\Delta x)$ in (3.6) as $\Delta x \to 0$. Note also that

$$\sum_{j=1}^{J} s_j \Delta x = \frac{1}{3}, \tag{3.9}$$

where the left hand side is the rectangular rule approximation to the left hand side of (3.7), given that $s_j \approx s(j\Delta x)$ for Δx small. Hence in the limit as $\Delta x \to 0$, (3.7) is satisfied.

The discrete model

The discretisation can be written in matrix form as follows;

$$E\mathbf{w}^{n+1} = A\mathbf{w}^n + \mathbf{u} \tag{3.10}$$

where

$$\mathbf{w}^n = (w_1^n, w_2^n, ..., w_{J-1}^n)^T, \tag{3.11}$$

and where the vector $\mathbf{u}$ containing the boundary conditions and the source term has the form

$$\mathbf{u} = (w_a, 0, ..., \frac{\Delta t}{3\Delta x}, 0, ..., w_b)^T, \tag{3.12}$$

(the non-zero elements of $\mathbf{u}$ are u_j where $j = 1$, $j = J/4$ and $j = J-1$). The $(J-1) \times (J-1)$ tridiagonal matrices E and A are given by

$$E = \begin{pmatrix} (1+2\mu\theta) & -\mu\theta & & & \\ -\mu\theta & (1+2\mu\theta) & -\mu\theta & & \\ & \vdots & \vdots & \vdots & \\ & & \vdots & \vdots & \vdots \end{pmatrix}, \tag{3.13}$$

$$A = \begin{pmatrix} (1 - 2\mu(1 - \theta)) & \mu(1 - \theta) \\ \mu(1 - \theta) & (1 - \theta) \end{pmatrix}$$

3.2 Experiment 1: Changing the eigenvalues

As discussed in Section 2.1 the observer (2.3) we wish to construct will converge to the "true" solution provided that the eigenvalues of $E^{-1}(A - GC)$ have modulus less than unity. Apart from this restriction, we are free to choose this set Λ of eigenvalues as we like. The aim of Experiment 1 is to try out different choices of eigenvalues for the set Λ, and to see which choice gives the fastest convergence of the observer solution to the observations.

The choice of the observation matrix C used in this experiment is rather arbitrary; if there are p observations, then C is the $p \times 15$ matrix

$$C = \left(\right.$$

ues were reduced in modulus by 0.25, since both sets of eigenvalues overall had similar modulus. As would be expected, choosing large eigenvalues distributed between -1 and -0.5 gave worse results still; some 11 or more observations were needed for reasonable convergence in this case.

3.3 Experiment 2: Changing the observation matrix

The matrix C can be considered as an interpolation of the model values $\mathbf{w}^k$ from the grid points to the observation points. Choosing the matrix C therefore determines what linear combination of grid point values should be used as the model equivalents to each observation. The theory demands that C should be full row rank (ie, rank p) for constructing the observer. The matrix C used in Experiment 1 was chosen arbitrarily rather than using physical considerations. In Experiment 2, it is supposed that observations are available at anything from 1 to 14 observation points, which do not in general coincide with grid points. The aim of Experiment 2 is to develop an observation matrix which will represent a linear interpolation from the grid point positions to the observation positions. For this, it is at first supposed that the 14 observation positions sittheoinTl9tjcopj-Ttrix-atiol-lsT

Figure 2 illustrates the performance of the observer in the case $p = 5$, $\theta = 0$. The positions of the observations are marked with a + on the x-axis. Comparing Figures 2 and 1 shows that convergence to the observations is almost as fast as when the matrix C of Experiment 1 is used, in which the observation positions coincide with the grid point positions. This was the case for all values of θ tested, and for $p \leq 12$. If more than 12 observations were used, however, C was no longer of rank p, and the method failed.

Table 3 in the Appendix gives a different set of the observation positions. The same linear interpolation is used, but C no longer has the neat structure of (3.20) since the observations are ordered at random. Figure 3 shows the results obtained in this case with $p = 5$ and $\theta = 0$. The convergence to the observ

3.4 Experiment 3: Less frequent observations

We now consider the situation where observations are not available at every timestep. This is an important consideration in the context of data assimilation where in practice there will not in general be a complete set of data at every model timestep. Experiment 3a examines the behaviour of the observer after a supply of frequent observations runs out. In Experiments 3b and 3c it is supposed that observations are available every second, fourth or eighth timestep, and two modifications to the data assimilation scheme are considered for dealing with this. The observ

model.

In this case, the observations are generated from a model run with one value of μ, and the numerical model uses another. So, the observer model not only starts with the wrong initial conditions, but also contains in itself model error, due to the incorrect values of μ. The observer can correct for the wrong initial conditions by driving the model solution to the true solution, but when the observer is switched off, the model solution is expected to drift away from the true solution.

As usual, plots were done for $t = 0.25$, $t = 0.5$, $t = 0.75$ and $t = 1$. However, if observations were available every other timestep, so that the timestep length was doubled, then these plots showed the solutions at 10, 20 , 30 and 40 timesteps. If the observations were available every fourth timestep, the plots showed the solutions at 5, 10, 15 and 20 timesteps. Experiments 1 and 2 show that the observer needs enough timesteps to "settle", so the results at 10 timesteps were not expected to be very good.

Even so, when observations were available every other timestep the results were pleasing, (see Figure 7); on the whole the observer solution converged quite quickly to the true solutions. Generally, the results improved as p increased. When observations were available only every fourth timestep, the results were significantly poorer, (see Figure 8), and if the observations were available every eighth timestep, then satisfactory results were found only for $p \geq 10$.

This approach to the problem of infrequent data is rather unsatisfactory because of its limited practical application, since models generally run with the largest timestep feasible anyway. It also has the disadvantage that p needs to be

4 Data assimilation using successive correction

Some of the earliest attempts at data assimilation in the late 1950s used an approach known as "successive correction". Since then, this conceptionally simple approach has been developed into schemes which are sometimes quite sophisticated. The basic idea is to modify the model solution in the light of the observations. In its simplest form, this means adding some proportion of the difference between an observation and its model counterpart to the model solution at all grid points within some "radius of influence" of the observation. In the Cressman scheme [2], the proportion to be added to a particular grid point depends on its distance to the observation. If w_{ij} is the weight or proportion of the correction to grid point i with respect to observation point j, then

$$w_{ij} = \begin{cases} \frac{1}{\rho_i}\frac{R-d_{ij}}{R+d_{ij}} & \text{if} \quad d_{ij} < R \\ 0 & \text{if} \quad d_{ij} \geq R \end{cases} \tag{4.1}$$

Here R is the radius of influence, and d_{ij} is the distance of grid point i to observation j, and ρ_i is the number of nonzero entries in row i. In [2] the correction stage of each model timestep is repeated several times with successively smaller values of R. Here just one value of R is used since this is only a small scale problem.

The weight w_{ij} forms the ij^{th} element of the Cressman weighting matrix W. This section discusses how W relates to the observer feedback matrix G, and compares the performance of the Cressman scheme to the observer for the same model as used in Section 3.

4.1 Comparison of the observer and successive correction techniques

Suppose the true state of the atmosphere is described by the discrete linear time invariant system

$$\mathcal{S}: \qquad E\mathbf{w}^{k+1} = A\mathbf{w}^k + B\mathbf{u}^k \tag{4.2}$$

and that we have observations $\mathbf{y}^k$ of the state $\mathbf{w}^k$ given by

$$\mathbf{y}^k = C\mathbf{w}^k. \tag{4.3}$$

In a general sucessive correction method, each model timestep involves two stages: a model update, and then a correction. Writing $\tilde{\mathbf{w}}^{k+1}$ for the updated model state,

and $\hat{\mathbf{w}}^{k+1}$ for the corrected model state, we have

Stage 1: model update

$$E\tilde{\mathbf{w}}^{k+1} = A\hat{\mathbf{w}}^{k} + B\mathbf{u}^{k}, \tag{4.4}$$

Stage 2: correction

$$\hat{\mathbf{w}}^{k+1} = \tilde{\mathbf{w}}^{k+1} + W(\mathbf{y}^{k+1} - C\tilde{\mathbf{w}}^{k+1}). \tag{4.5}$$

Substituting (4.2) - (4.4) into (4.5) gives

$$\hat{\mathbf{w}}^{k+1} = E^{-1}(A\hat{\mathbf{w}}^{k} + B$$

4.2 Experiments with the Cressman scheme

The Cressman scheme as described in (4.1) was implemented for the system (4.2) to compare its performance with the observer G developed in Section 2. In this experiment, the matrices E, A, B, C; the input $\mathbf{u}$ and the initial condition $\mathbf{w^0}$ were chosen as in Section 3.1. The Cressman scheme was tested for different values of θ and p, and for different values of R, the radius of influence. As before, plots were produced comparing the "Cressman solution" (solid line) with the original numerical solution (dashed line) and the "true solution" (plotted ooo), at 20, 40, 60, and 80 timesteps.

For all values of θ, the success of the scheme for driving the model solution to the true solution depended strongly on the number of observations used. For $R = 0.3$, using just one observation had almost no impact on the solution, and using 3 observations gave some improvement to the n

tions. Reformulating data assimilation schemes in terms of observers introduces the possibility of using results from control theory to carry out analysis on these schemes. This could perhaps explain some of the successes and failings of the different schemes. The experimen

APPENDIX

Table 1: System eigenvalues

The eigenvalues of $E^{-1}A$ defined in equation (3.10) for $\theta = 0$, $\theta = 0.5$ and $\theta = 1$ are:

$\theta = 0$	$\theta = 0.5$	$\theta = 1$
0.7156	0.9878	0.9879
0.6045	0.9524	0.9026
0.8125	0.8977	0.9535
0.8921	0.8286	0.8421
0.9513	0.7510	0.7786
0.9877	0.6701	0.7168
0.4849	0.5904	0.6600
0.3600	0.5152	0.6098
0.2351	0.4467	0.5666
0.1151	0.2241	0.6305
0.0044	0.2379	0.5011
−0.0925	0.3353	0.4410
−0.1221	0.2934	0.4604
−0.2313	0.3865	0.4482
−0.2677	0.2610	0.4770

List f Figures

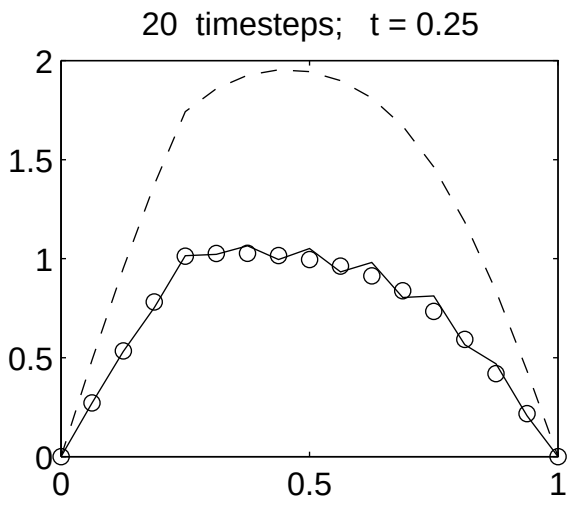
Key to the figures:

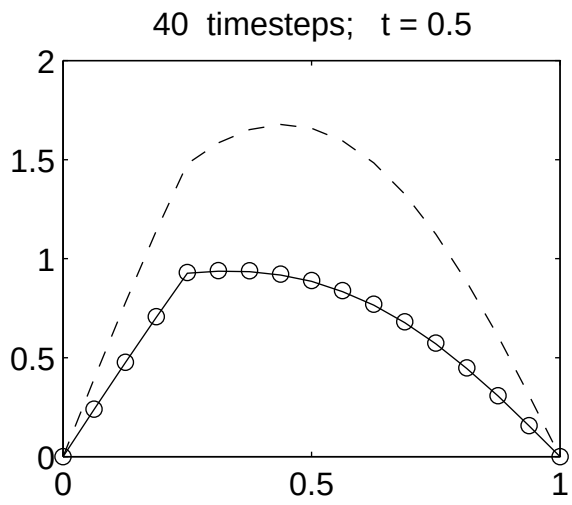
Observer solution

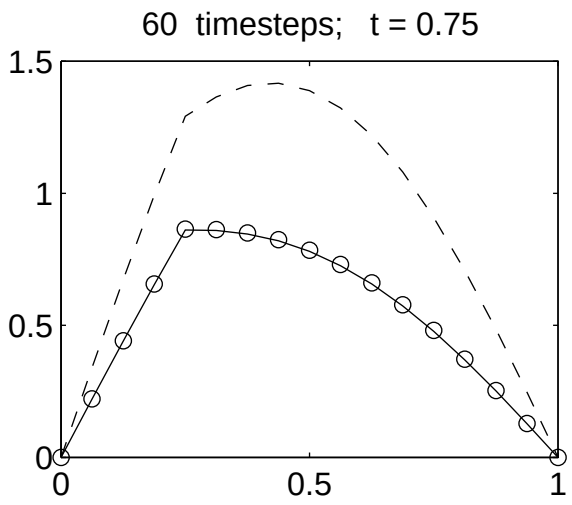
\- - - Numerical solution with no observer

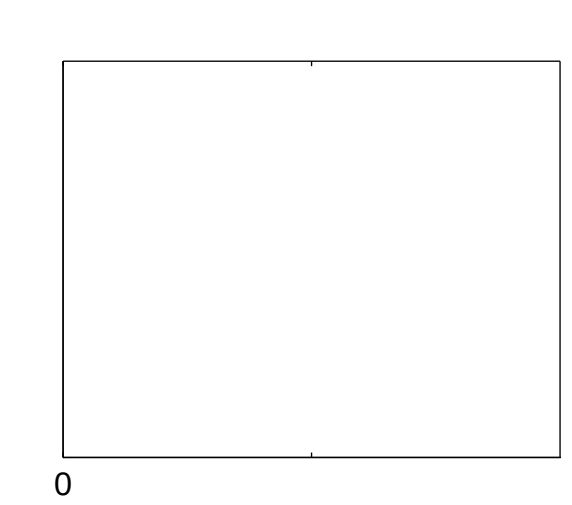
o o o True solution

\+ + + Positions of the observations


20 timesteps; t = 0.25
2
1.5
1
0.5
0
0
0.5
1


40 timesteps; t = 0.5
2
1.5
1
0.5
0
0
0.5
1


60 timesteps; t = 0.75
1.5
1
0.5
0
0
0.5
1


0

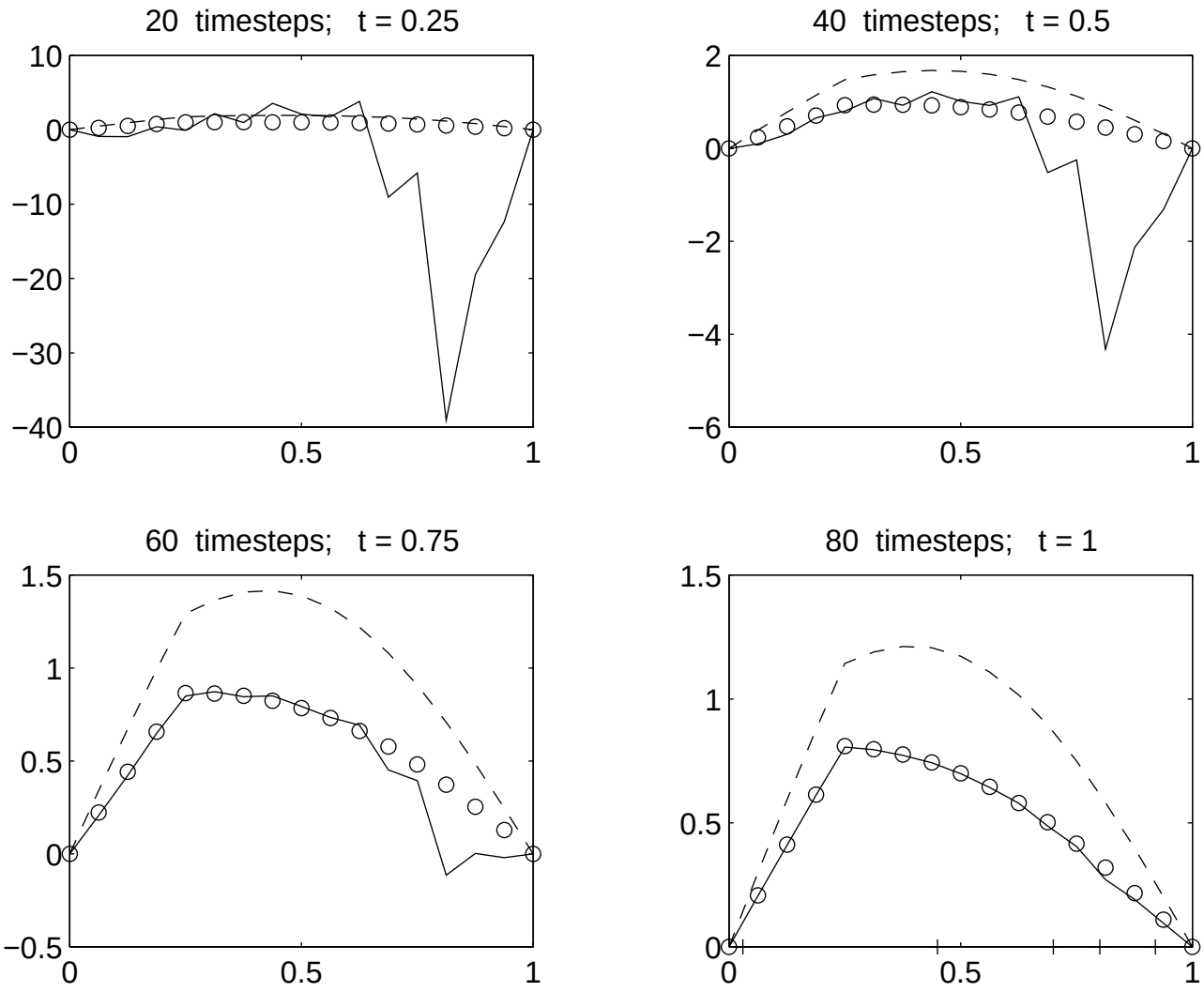
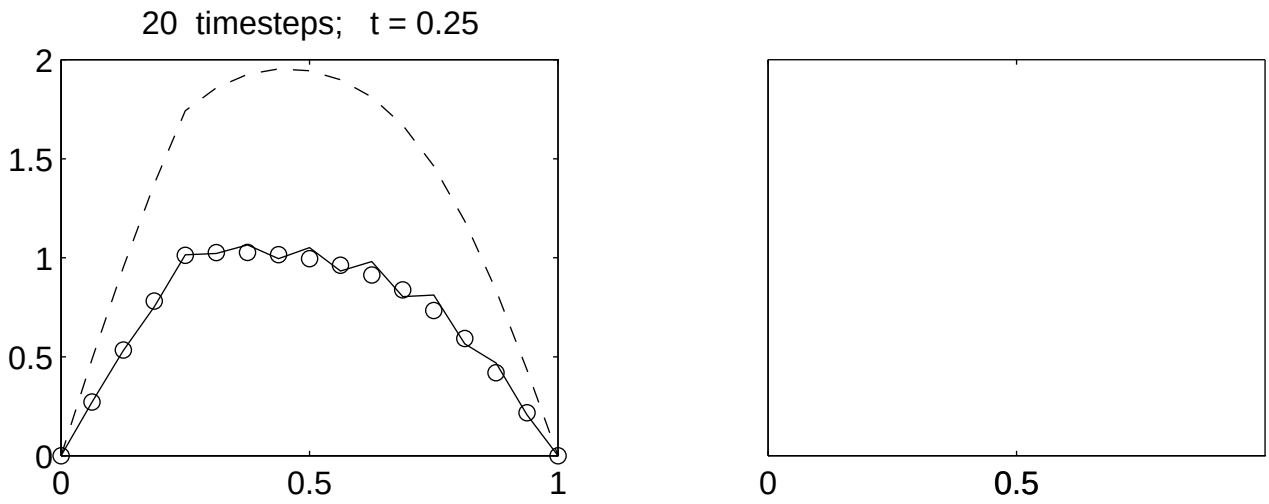


Figure 3: Experiment 2; $p = 5$, $\theta = 0$. Second set of observation positions (unordered)


20 timesteps; t = 0.25
2
1.5
1
0.5
0
0
0.5
1
0
0.5

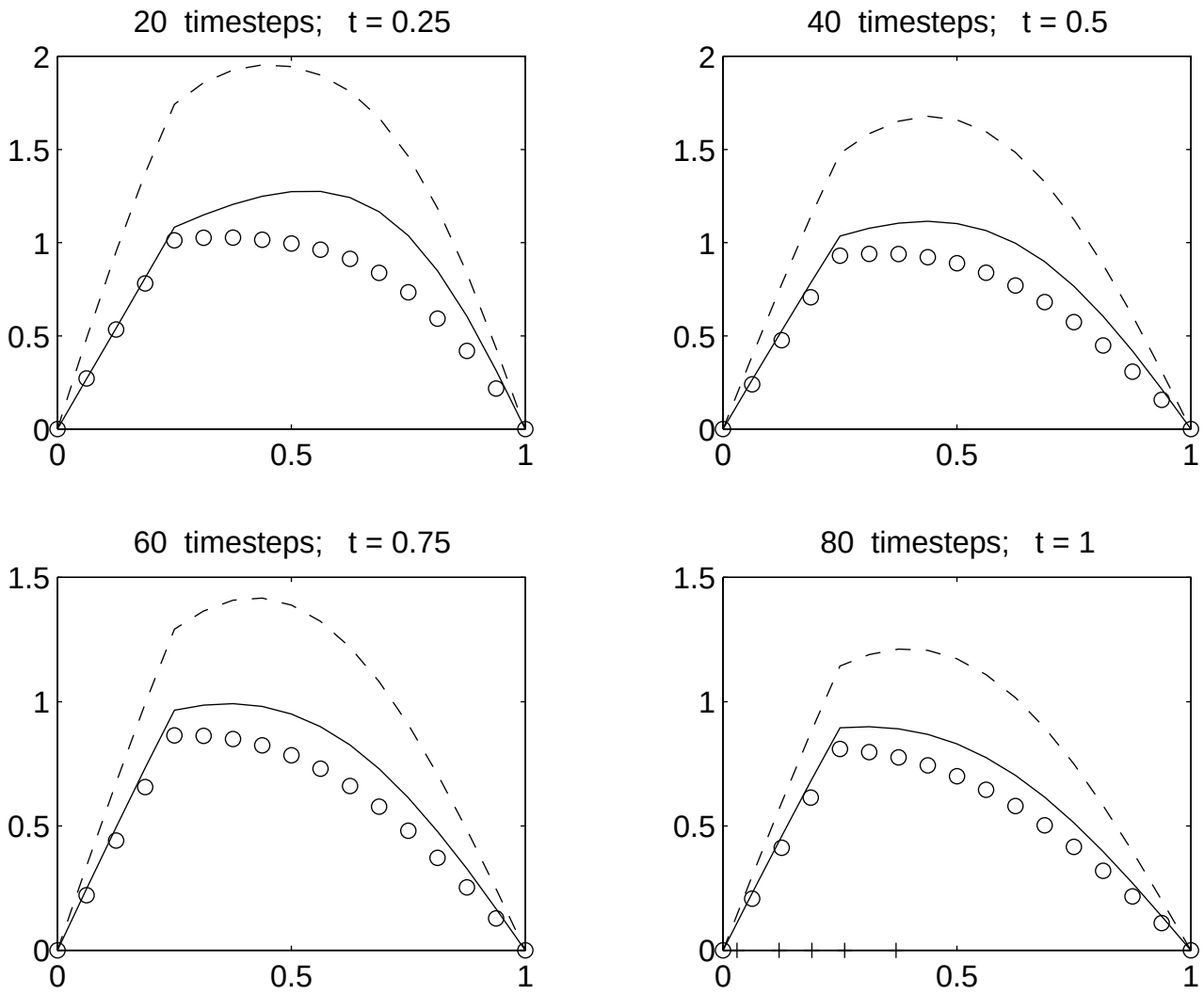
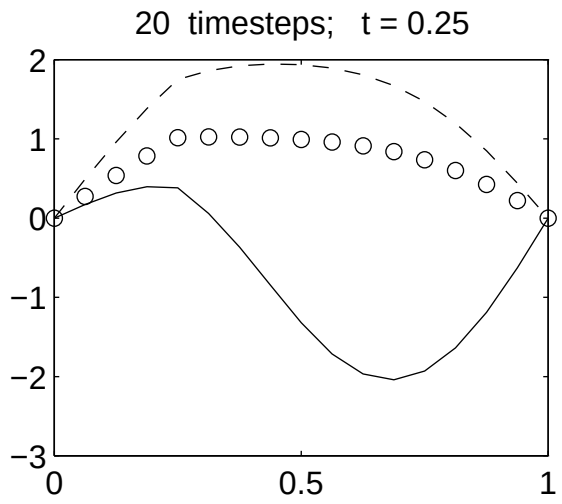
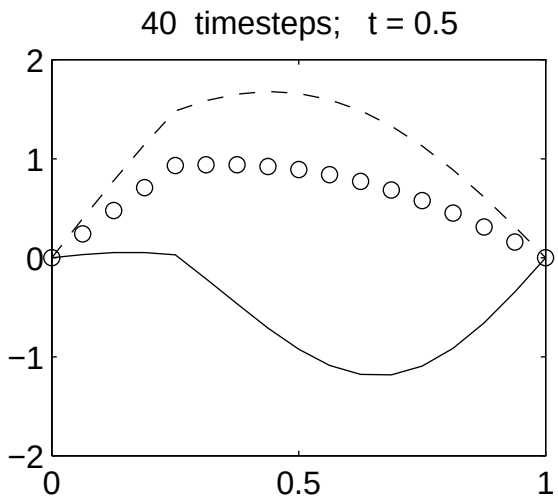
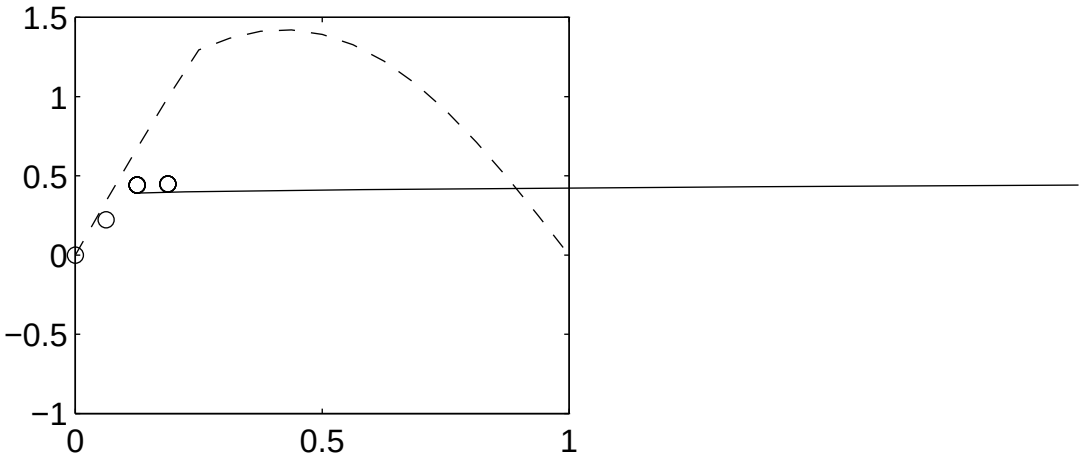
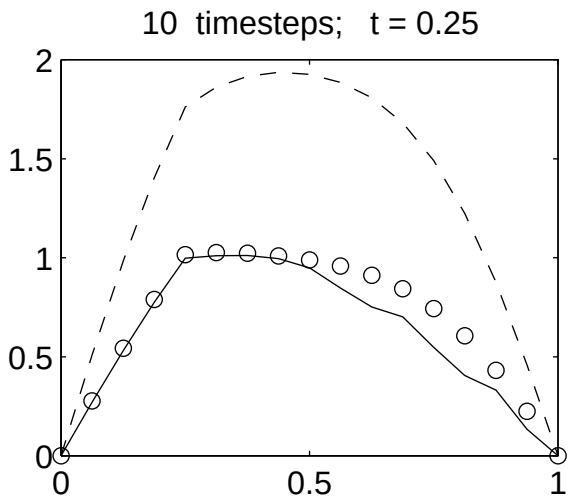


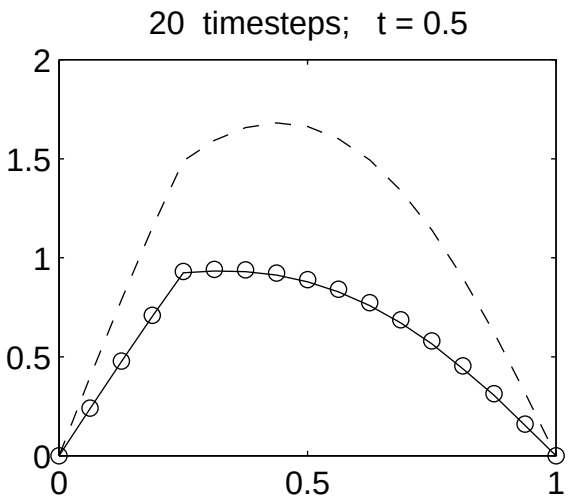
Figure 5: Experiment 3a; $p = 5$, $\theta = 0$. Observations available for the first 10 timesteps

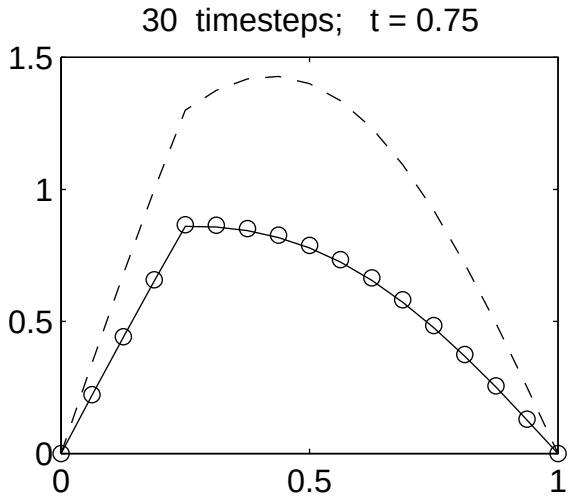

20 timesteps; t = 0.25
2
1
0
−1
−2
−3
0
0.5
1

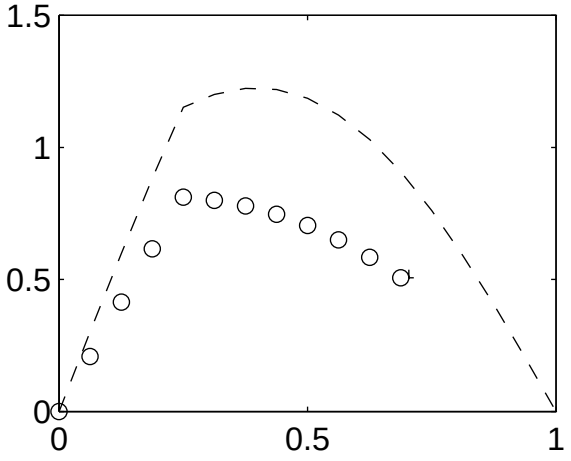

40 timesteps; t = 0.5
2
1
0
−1
−2
0
0.5
1


1.5
1
0.5
0
−0.5
−1
0
0.5
1


10 timesteps; t = 0.25
2
1.5
1
0.5
0
0
0.5
1


20 timesteps; t = 0.5
2
1.5
1
0.5
0
0
0.5
1


30 timesteps; t = 0.75
1.5
1
0.5
0
0
0.5
1


1.5
1
0.5
0
0
0.5
1

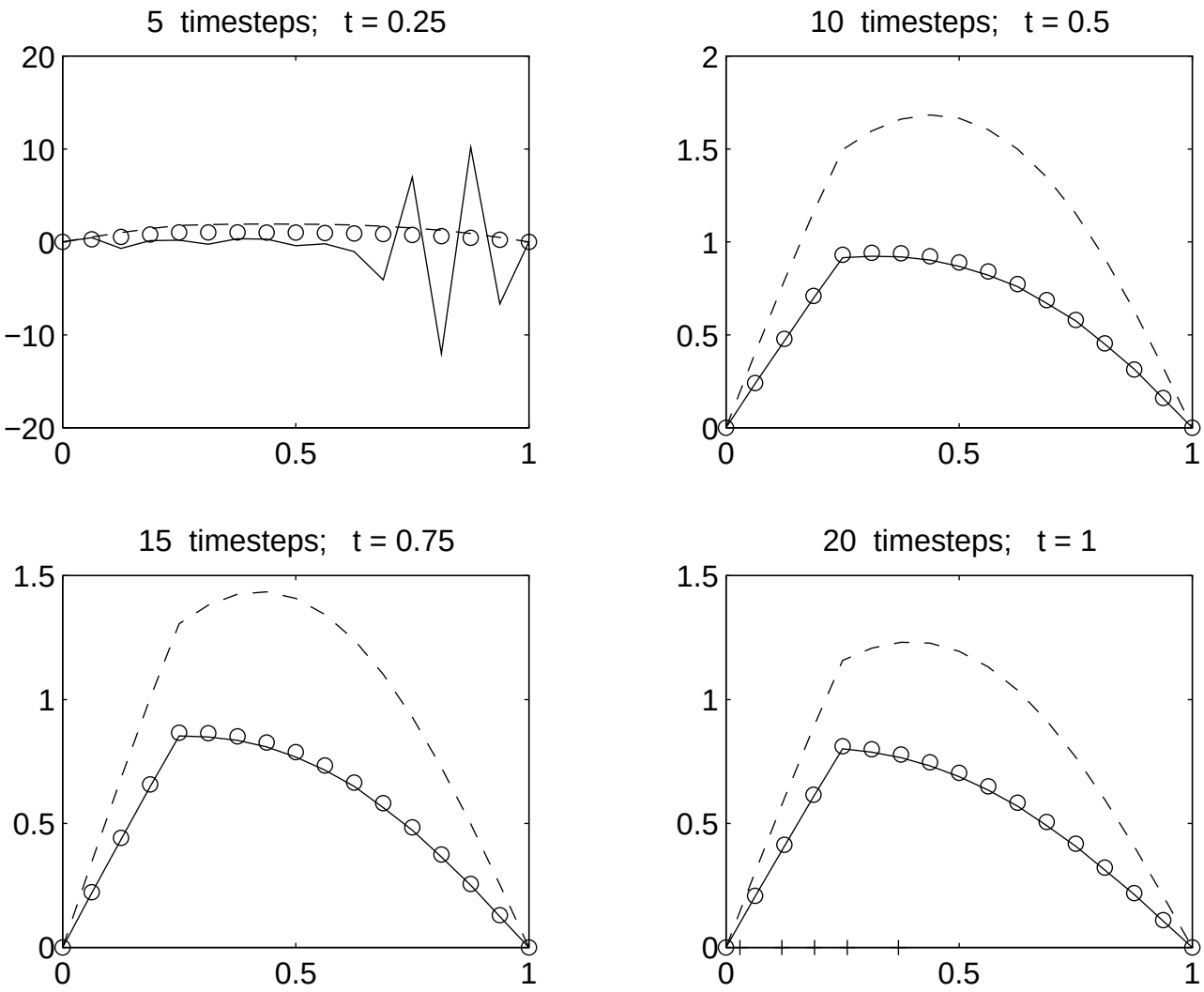


Figure 8: Experiment 3b; $p = 5$, $\theta = 1$. Observations available every fourth timestep

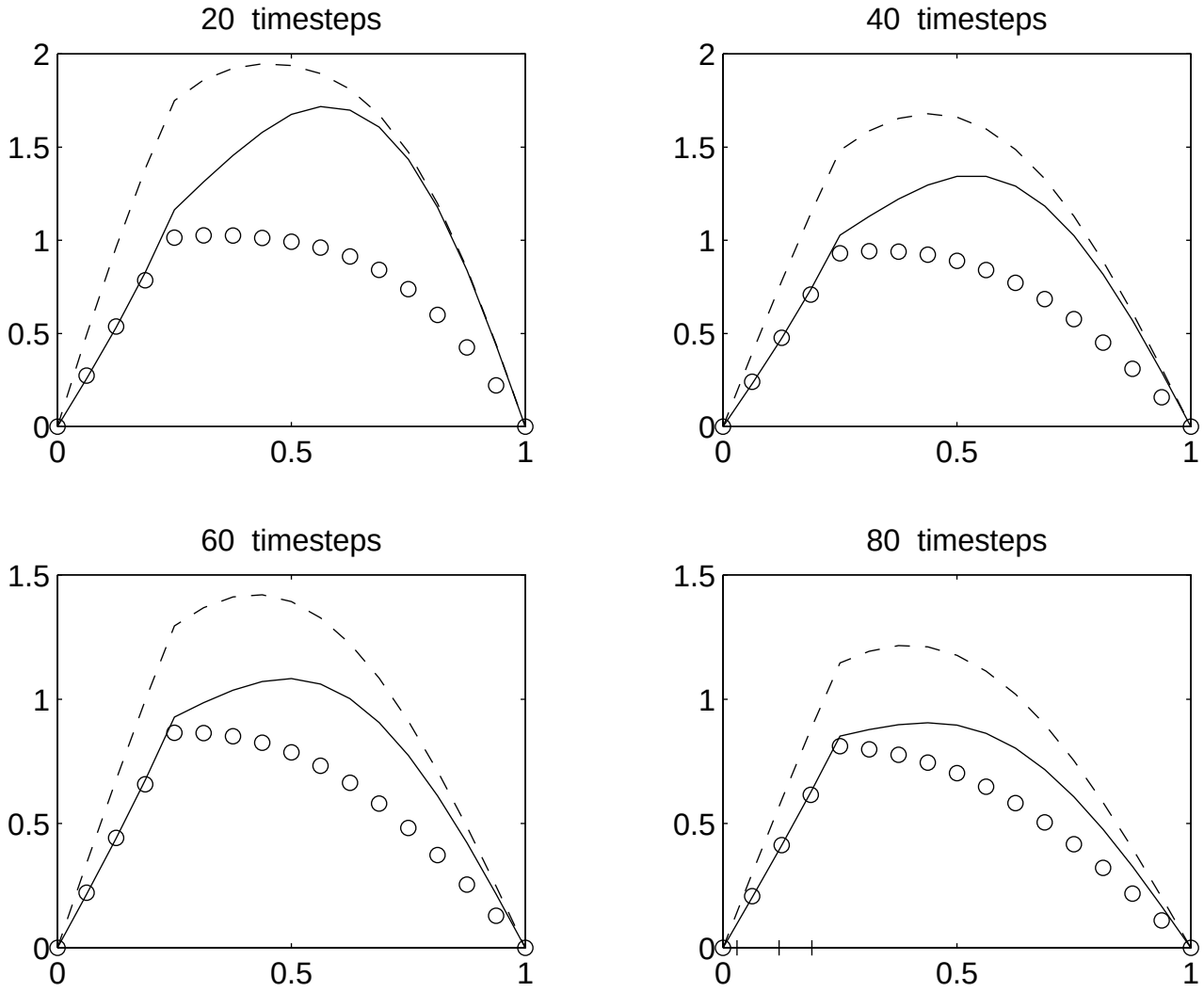


Figure 9: Cressman sc

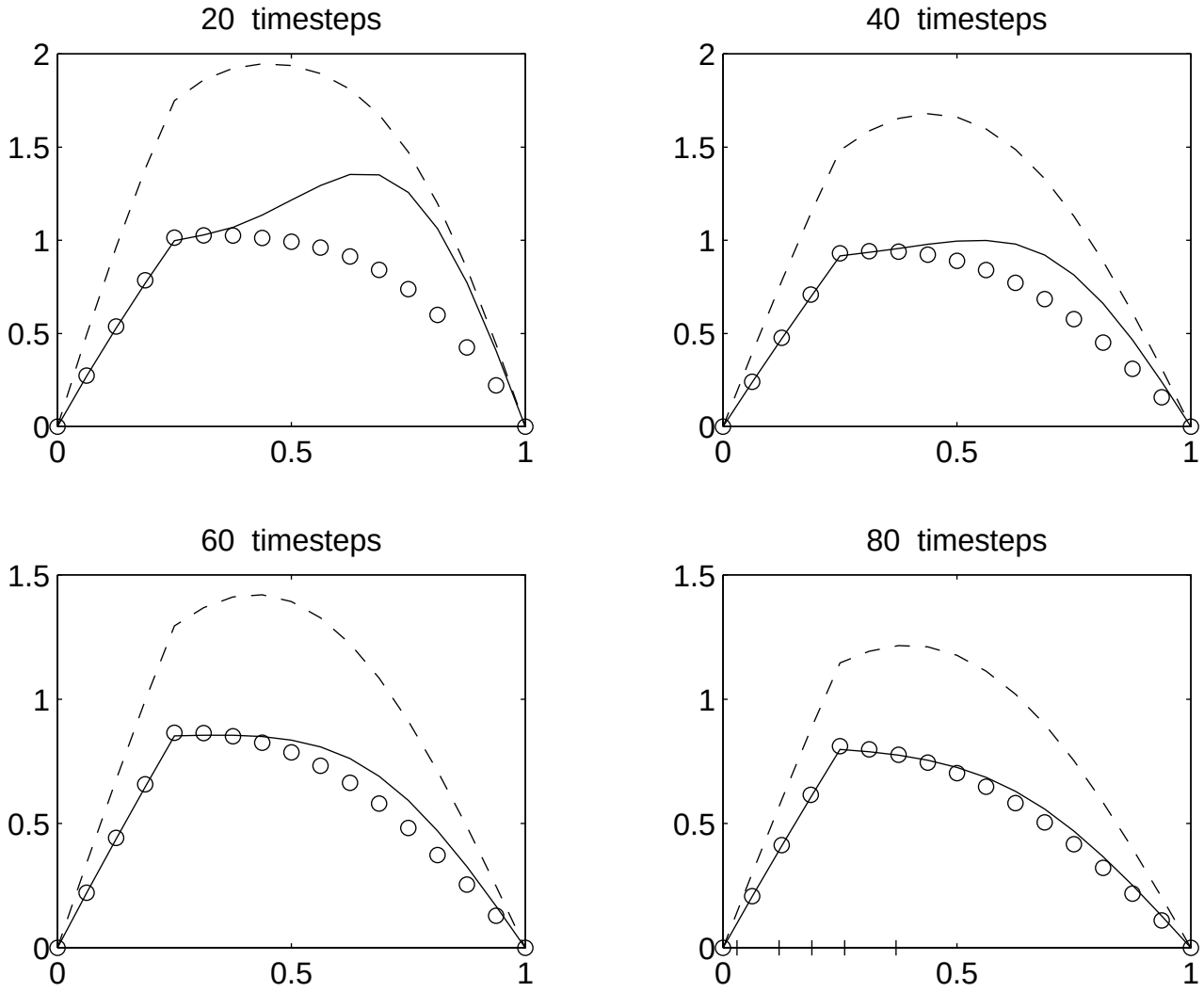


Figure 10: Cressman scheme; $p = 5$, $\theta = 0.5$, $R = 0.3$

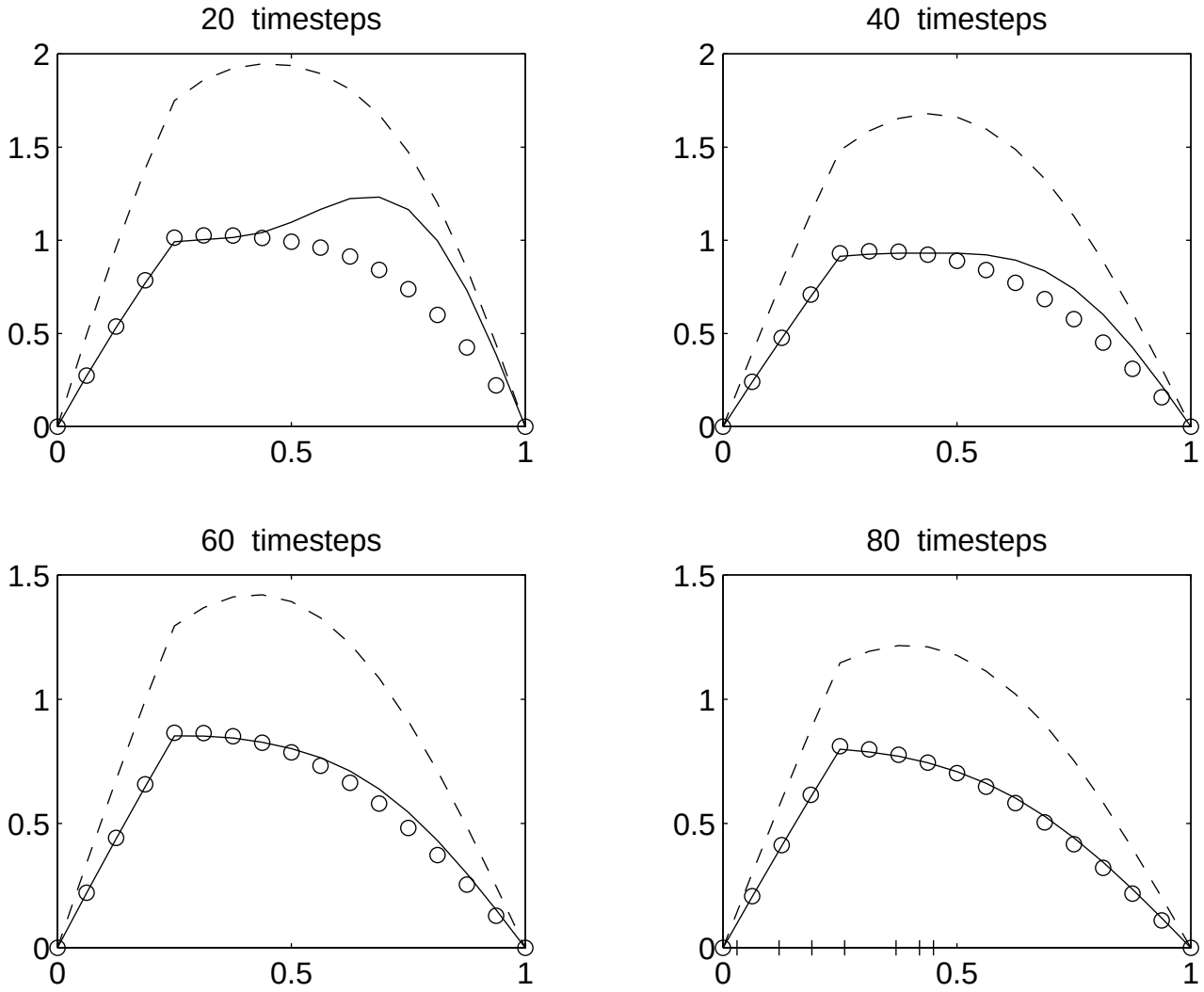


Figure 11: Cressman scheme; $p = 7$, $\theta = 0.5$, $R = 0.3$