

Data assimilation into biomass dynamics models: A Monte Carlo simulation experiment

**Al-Amin M. Ussif¹, Leif K. Sandal¹, Stein I. Steinshamn²
and Steven L. Morey³**

Abstract

In this paper, we use a variational data assimilation method to fit biomass dynamics models to simulated data. The method is the variational adjoint technique in which a cost function measuring the distance between the data and the model solution is minimized. This approach is a deterministic procedure in which the model is repeatedly solved and the solution compared to the observations or measurements in order to find the parameters of the model that give predictions which are as close as possible to the data. We will briefly review some of the methods commonly used in fisheries management and compare them with the method in this paper. Twin experiments are used to evaluate the performance of the algorithm. The parameter estimates have reflected the true values well and the deviations from the true parameters are not substantial compared with the errors in the simulated data.

¹ Norwegian School of Economics and Business Administration, Bergen, Norway

² Foundation for Research in Economics and Business Administration, Bergen, Norway

³ Center for Ocean-Atmospheric Prediction Studies, Florida State University, Tallahassee, Florida USA

Introduction

Several methods of fitting biomass dynamics models to data have been developed and used in the literature of both fisheries stock assessment and bioeconomics (Ludwig and Walters 1989; Polacheck et al. 1993; Bjørndal and Conrad 1995). In fisheries stock assessment literature, three approaches have often been used: The so-called Equilibrium Estimators (EE), Process-Error Estimators (PE) and Observation-Error Estimators (OE). Other methods are, the Total Least Squares (TLS) and the Approximate Likelihood (AL) methods (Ludwig and Walters 1988). The approaches in, for example, Polacheck et al. are in fact ways of introducing errors in the model or measurements but are strictly speaking not estimators. The estimators used are the least squares and maximum likelihood. These methods have all been in line with the conventional ordinary regression analysis (Russell 1980).

This paper has two main goals. First, it develops dynamic resource models for the management of natural resource stock such as fish. Second, the variational adjoint method (VAM) is used to fit the model dynamics to simulated data. The VAM is a data assimilation technique for fitting dynamical models to data. It has recently been applied to biological and ecosystem models (Lawson et al. 1995; Spitz et al. 1998). The method also provides a more accurate and efficient way of evaluating the gradient of the cost function compared with the finite difference approach (Huiskes 1998). Huiskes (1998) is a recent application of the VAM to Virtual Population Analysis. It can be applied to both linear and nonlinear models and does not require that nonlinear models be transformed into linear forms. Data assimilation methods are undoubtedly becoming more and more popular in areas such as biological systems, ecosystem modeling, fisheries stock assessment, etc.

The goal of data assimilation is to improve the model itself or its forecasts by extracting information from available observations or measurements. Thus, our purpose in this paper is

to show an example of the use of the VAM to improve the biomass dynamics models that are very much used by resource economists and others. While the VAM is applicable to more complex and large models, we will use two-dimensional models to illustrate our method. The biomass dynamics models used in most fisheries stock assessment are thus simpler and less informative than the one used in this analysis. The effort variable is normally assumed to be an exogenous variable. In this paper, an updating scheme for the harvest function is also developed. This will enable us to make future forecasts of both the stock and the appropriate harvest amount.

This paper will discuss and briefly compare the VAM with the conventional methods that are common in the literature. To our knowledge no such comparison exists. Using simulated measurements, the performance of the algorithm will be evaluated and an analysis of the bias of the parameter estimates will also be performed.

Discussion of the Methods

In this section, we discuss the different methods that have commonly been employed in fisheries management. We will compare and contrast these methods with the VAM.

Both the conventional methods (CM) and VAM are cast in least squares or maximum likelihood context. The VAM is, however, more advanced in that it can be applied to both linear and nonlinear models (Lawson et al. 1995). It also provides a more straightforward approach for estimating both the initial and boundary conditions of the dynamics. For more complex and large models, e.g., systems of ordinary or partial differential equations, this technique is preferred.

The CMs are basically ordinary regression methods, which are in a way “static” or “plug-in” methods. That is, one does not actually solve the model but puts in data into the models in

order to find the estimates. In the VAM, the model is solved and the predictions compared with the data until parameter estimates that give model predictions that are as close as possible to the data are found.

In the EE, the stock is assumed to be in equilibrium. This is rarely the case in practice. The PE assume that observations are made without error and the error occurs in the change in population biomass while the OE assume that the dynamics are deterministic and that the error occurs in the relationship between the stock and the index of abundance. The variational method in its general form accommodates all types of error. In the VAM, the model dynamics are assumed to be deterministic and exact while the data are assumed to contain errors. This approach corresponds to the OE in fisheries stock assessment literature.

A major difference between the approach in this paper and the earlier works is that, while we provide a more advanced and efficient algorithm for fitting dynamical systems to data, the former used ordinary regression methods but provided methods of analyzing error in the various components of the system, e.g. the observations or the model. While the focus of most previous research has been on single equation models, the application here is in a system of equations context.

Note that the VAM may be more appropriately described as deterministic trend modeling.

Consider the simple case of fitting $dx/dt = ax, x(0) = x_0$ to a set of data. It is easy to see that this is equivalent to fitting an exponential trend model: $x(t) = x_0 \exp(at)$, where x_0 and a are the parameters. This example can be solved using manual trial-error or “brute-force”. That is, individual parameters can be manually adjusted until the model fits the data. Such approach is subjective and expensive even with models that have few parameters. When the model dynamics become more complex and large, the VAM provides an automated and more efficient alternative. For straightforward problems, the CMs work equally well. However, for

more complex dynamics where analytical solutions are unattainable, the CMs are limited or very inefficient. The VAM handles this situation more efficiently. It must be noted that if the model is simple and not complicated enough these methods are equally efficient since the difference in regard to computational time is negligible. One of the disadvantages of the VAM is that it does not provide direct means of calculating error estimates. However, statistical methods such as the bootstrap (Effron and Tibshirani 1993) could be used to find the standard errors.

The Model

The biomass dynamics model takes the form

$$\frac{dB}{dt} = rB^\theta \left(1 - \frac{B}{K}\right) - h, B(0) = B_0 \quad (1)$$

where $B(t)$ and $h(t)$ are the biomass and the catch respectively, θ is a constant, r is the intrinsic growth rate and K is the carrying capacity. This form of the growth law is known as the modified logistic. For the most part of this analysis, the surplus production will be assumed to be the logistic function $\theta = 1$, (Schaefer 1957). The exponent in the surplus growth makes the model more nonlinear in the parameters. Including this parameter among the set to be estimated makes the estimation more challenging. We will nevertheless attempt to estimate this parameter as well.

To completely describe the system, we construct models that characterize the dynamics of the catch or harvest. Two forms of the equations will be used. That is, the dynamics are

$$\frac{dh}{dt} = G(B, h), h(0) = h_0,$$

$$G = \rho_0 h + \alpha, \quad (2a)$$

$$G = \rho_0 h + \rho_1 B h + \gamma \quad (2b)$$

Where ρ_0, ρ_1 and α are constants. Note that the parameters in (2a) must have opposite signs. The dynamics of the harvest can assume various forms and may depend on several factors. This relationship may be linear or nonlinear. The correct form of this dynamical equation is not known, however, several parametric forms exist. It has been the rule in bioeconomics that this equation is obtained based on the assumption of optimal behavior of managers. This leads to the maximization of a particular objective function, e.g. the discounted net revenues (Sandal and Steinshamn 1997c). Such frameworks seek to answer the question of what should happen (normative theories) rather than the more difficult and interesting question of what does happen (descriptive theories). This paper develops models that may be useful in seeking answers to the latter. Equations (2a-b) provide schemes that update the catch.

We notice that equation (2a) is a simple exponential trend model and the analytical solution is: $h = h_0 \exp(\rho_0 t) + \frac{\alpha}{\rho_0} (\exp(\rho_0 t) - 1)$. The estimation problem in a more general setting

involves finding the initial condition(s) and the model parameters. Equation (2b) is however, more complex and highly coupled. The difference between these two model dynamics is that the first does not contain stock dependence whereas the other does. Thus, equation (2b) provides a more sophisticated dynamics. It is easy to see that, equation (2a) is nested in (2b).

Note that these models are used as examples and can be replaced with any other model. This technique can equally be employed to estimate the initial conditions of the models but this will not be the focal point here. It is easy to see that equations (2a-b) are quite general.

Embedded in the dynamical equation (2a) are three special cases: (i) if we assume that, $\alpha \equiv 0$, then, the equation reduces to a pure exponential growth model for the harvest, (ii) if we now assume that, $\rho_0 \equiv 0$, we see that the harvest function is simply a straight line with an intercept. Finally if both parameters are zero, this yields a constant harvesting rule. Any of these cases may be observed in practice either in commercial or recreational fishing.

The models developed here are quite different from the ones used in most fisheries stock assessment. Here the focus is on modeling the harvest function (see also, Sandal and Steinshamn 1997b) whereas the CMs focused on modeling effort. Modeling effort can be quite problematic because it is highly subjective to different definitions and also it is rather more difficult to measure. Effort comprises of capital investments, labor input, experience, etc. These quantities are very difficult to measure in practice. Note that because equation (2a) is independent of B, the system can be estimated on an equation-by-equation basis. However, it may be more appropriate to fit them as a system. For the empirical analysis, we will conduct three sets of experiments using three combinations of equation (1) and (2). In the first i.e. Model 1 we use equations (1) and (2a) with the Schaefer logistic function, i.e. $\theta = 1$. Model 2 is a combination of (1) and (2b) using the Schaefer logistic as well. In the last case (Model 3), we use the same combination as in Model 1 but this time making θ a parameter to be estimated. This can pose a lot of difficulties because it introduces very strong nonlinearity in the parameters.

The cost function

In the VAM a quadratic cost function measuring the distance between the data and their model equivalents is minimized. The cost function is

$$J = \sum_{n=1}^N (B_n^{obs} - B_n^{mod})^2 + \sum_{n=1}^N (h_n^{obs} - h_n^{mod})^2 \quad (3)$$

Where the superscripts *obs* and *mod* denote the observations or measurements and the model counterparts respectively. We remind readers that, the stock biomass is not directly observable but methods of estimating the stock size such as the catch per unit effort (CPUE) and Virtual Population Analysis (VPA) exist. This function is minimized subject to equations (1) and (2a-b). The fact that the model counterparts of the data are not usually expressed as explicit functions of the independent variables makes the problem of minimizing the cost function more complicated. To calculate the gradients, two approaches are often used: the finite difference approach and the VAM. In the former, each parameter is perturbed in turn and the cost function calculated in order to evaluate the gradients. This requires $m+1$ calculation of the cost function where m is the number of parameters. The VAM however requires only one calculation of the cost function. This provides a huge computational time saving when analyzing large systems.

To derive the adjoint system, we construct the discrete equivalent of the Lagrange function L :

$$L = J + \sum_{n=1}^{N-1} \lambda_n (B_{n+1} - [B_n + (rB_n(1 - B_n/K) - h_n)\Delta t]) + \sum_{n=1}^{N-1} \gamma_n (h_{n+1} - [h_n + G(B_n, h_n)\Delta t]) \quad (4)$$

where λ, γ are the Lagrangian multipliers, N is the time horizon and Δt is the time step. The procedure for minimizing this function is quite straightforward. The algorithm is:

1. Choose a first guess for the parameters.
2. Integrate the model equations (1)-(2).
3. Calculate the misfits and hence the cost function.
4. Integrate the adjoint equation and calculate the gradients.
5. Find the minimum of the cost function using an iterative procedure.

The details of the derivation of the adjoint code are not presented here to avoid cumbersome technicalities. A schematic representation of the steps is however provided in Figure 1.

The data assimilation procedure consists of three components: the forward model, the adjoint model and the optimization procedure as shown in Figure 1. This should help convey the procedure in a straightforward fashion. Interested readers are referred to Lawson et al. (1995) and Evensen et al. (1999). Note that Evensen et al. gives the continuous version of the derivation.

Results: Twin-experiments

To illustrate the VAM using biomass dynamics models, several twin-experiments are performed, i.e., artificial data are generated from the model itself using known values of the parameters and the initial conditions. To test the robustness of the estimates we used data with nonconstant variance, i.e., the data is generated by the formula: $x_t^{obs} = x_t^{mod}(1 + \varepsilon_t)$ where $\varepsilon_t \rightarrow U(a, b)$, and U denotes uniform. That is, the model is solved and a given percentage of random noise is added to the model predictions. The model is then fit to the data. The goal is to show that for a reasonable amount of noise in the data, the VAM is capable of recovering the parameters to a sufficiently high level of precision.

In the Monte Carlo simulations, values of the parameters are chosen such that the simulated stock biomass and harvest are both increasing. This simulates a practical situation where the stock is being rebuilt to a higher level following overexploitation. Several experiments are performed. The results² of the experiments for each percentage level of noise are presented in TABLES 1-3.

The parameter recovery has generally been good. In the first experiment, clean data, i.e., no noise of any type is added, are used and the parameters were recovered to several orders of magnitude. The results are reported to four decimal places (TABLES 1-3, column 2). The accuracy did not increase substantially when the convergence criterion was increased. In the subsequent experiments, different levels of noise are added. The noise level was increased in steps of 5% to a maximum of 20% and the VAM is used to retrieve the parameters. For each level of noise, about 500 simulations were conducted and the means calculated. The VAM has been very efficient in retrieving all the parameters in Model 1-2 for all levels of noise. However, the recovery has not been very good for Model 3 due to the high nonlinearity in the parameter θ . Estimates of the parameters in Model 3 deteriorated very quickly with increasing noise level. Even at a level of 5%, the recovery became very difficult. The iterative procedure had serious problems with convergence. Another point of concern is the fact that the quality of the estimates of the other parameters declined very rapidly and became very unreliable. It therefore seems that better and more reliable estimates may be obtained by using the best reasonable guess available from other sources. Note that the results are reported for 5% level of noise only for Model 3.

The percentage deviation of the parameter estimates for Model 1-2 in all cases has been less than the percent noise in the data. Since one of the aims of the experiment is to test the

² The results are not very sensitive to the initial guesses of the parameters. This was tested using different starting values.

robustness of the method, the level of noise was increased to see when the parameters are no longer reasonable. It was discovered that when the level exceeds 40 percent, the deviation becomes substantial. Estimates of the bias for the parameters are also provided. To quantify the level of bias, 500 simulations were performed for each percent level of noise and the bias was calculated (Rows 5-8 TABLES 1-2). The largest bias is in the estimate of ρ and the lowest is in the estimate of r .

Conclusion

In this paper, the VAM is introduced and discussed. It is briefly compared with the traditional regression methods commonly applied in the literature. The similarities of the methods have been outlined and the differences are also carefully explained. It was pointed out that, while these methods produce the same results in specific cases, the VAM is superior in many ways. For example, it is a more suitable method when the model counterparts of the data are not easily expressed as explicit functions of the independent variables. In fitting large models to large amount of data, the efficiency gain in regard to computational time is very significant compared with the usual finite difference approximations.

The results of the parameter estimation are promising as shown in TABLE 1-2. For the level of noise used, all the parameters were retrieved reasonably well with their correct signs. Estimates of the bias show that for a reasonable amount of noise, the parameter estimates are not too far away from the true values. The bias increase with increasing level of noise but the increase is less than linear for a noise level of about 20%. The conclusion is that, while we believe that data assimilation methods can successfully be applied to more complex and realistic bioeconomic models, further research needs to be done in this area. Improving the

models will enhance the ability of researchers to answer questions that interest managers and fishers most.

Reference

1. Bjørndal, T. and Conrad, J.M. 1987. The Dynamics of An Open Access Fishery. *Can. J. of Economics.* 74-85.
2. Efron, B. and Tibshirani, R.J. 1993. *An Introduction to the Bootstrap.* Chapman and Hall.
3. Evensen, G., D.P. Dee and J. Schroeter. 1998. Parameter Estimation in Dynamical Models. NATO ASI, *Ocean Modeling and Parameterizations* edited by E.P.Chassignet and J. Verron, 1998.
4. Huiskes, M.J. 1998. Virtual Population Analysis with the Adjoint Method. *Fisheries Stock Assessment Models, Alaska Sea Grant College Program.* AK-SG-98-01.
5. Lawson, L.M., Spitz, H.Y., Hofman, E.E. and Long, R.B. 1995. A Data Assimilation Technique Applied to Predator-Prey Model. *Bulletin of Mathematical Biology.* Vol. 57:593-617.
6. Ludwig D. and Walters C.J. 1989. A Robust Method for Parameter Estimation from Catch and Effort Data. *Can. J. Fish. Acquat. Sci.,* Vol. 46:138-144.
7. Ludwig D. and Walters C.J. 1988. Comparison of two Models and two Estimation Methods for Catch and Effort Data. *Natural Resource Modeling.* Vol. 2, No. 2: 447-499.
8. Polacheck, T., Hilborn, R. and Punt, A.E. 1993. Fitting Surplus Production Models: Comparing Methods and Measuring Uncertainty. *Can. J. Fish. Acquat. Sci.* Vol. 50: 2579-2607.

9. Russell, S.U. 1980. Least Squares Regression Estimates of the Schaefer Production Model: Some Monte Carlo Simulation Results. *Can. J. Fish. Aquat. Sci.* Vol. 37: 1284-1294.
10. Sandal, K.L and Steinshamn, S.I. 1997c. Optimal Steady State and the Effects of Discounting. *Marine Resource Economics*, Vol. 12, 95-105.
11. Schaefer, M.B. 1957. A Study of the Dynamics of the Fishery for Yellowfin Tuna in the Eastern Pacific Ocean. *Inter-Am. Trop. Tuna Comm. Bull.* 11:247-268.
12. Spitz, H.Y., Moisan, J.R., Abbott, M.R. and Richman, J.G. 1998. Data Assimilation and A Pelagic Ecosystem Model. Parameterization Using Time Series Observations. *J. Mar. Syst.*, in Press.

TABLE 1: Results of the VAM Using Simulated Measurements.

Parameter Estimates and Biases: Model (1-2a)

Actual Parameters $r=0.40$, $K = 1.0$, $\rho = -0.2$, $\alpha =0.01$

% Noise	0	5	10	15	20
$\bar{\hat{r}}$	0.4000	0.4007	0.4014	0.4019	0.4025
$(\bar{\hat{r}} - r)/r$	0.0000	0.0017	0.0035	0.0047	0.0063
$\bar{\hat{K}}$	1.0000	1.0097	1.0095	1.0118	1.0152
$(\bar{\hat{K}} - K)/K$	0.0000	0.0097	0.0095	0.0118	0.0152
$\bar{\hat{\rho}}$	-0.2000	-0.1965	-0.2004	-0.2056	-0.2155
$(\bar{\hat{\rho}} - \rho)/\rho$	0.0000	-0.0175	0.0020	0.0280	0.0775
$\bar{\hat{\alpha}}$	0.0100	0.0109	0.0113	0.0112	0.0119
$(\bar{\hat{\alpha}} - \alpha)/\alpha$	0.0000	0.0900	0.1300	0.1120	0.1190

Sample size=95

Time step =0.1, Convergence criterion=1.0E-08

Number of simulations = 500

TABLE 2: Results of the VAM using simulated measurements.

Parameter Estimates and Biases: Model 2

Actual Parameters $r=0.40$, $K = 1.0$, $\rho_0 = -0.2$, $\rho_1 = 0.2$, $\alpha = 0.01$

% Noise	0	5	10	15	20
$\bar{\hat{r}}$	0.4000	0.3994	0.3989	0.3999	0.3983
$(\bar{\hat{r}} - r)/r$	0.0000	-0.0015	-0.0028	-0.0003	-0.0043
$\bar{\hat{K}}$	1.0000	1.0052	1.0143	1.0186	1.0402
$(\bar{\hat{K}} - K)/K$	0.0000	0.0052	0.0143	0.0186	0.0402
$\bar{\hat{\rho}}_0$	-0.2000	-0.1969	-0.1999	-0.2072	-0.1833
$(\bar{\hat{\rho}}_0 - \rho_0)/\rho_0$	0.0000	-0.0155	-0.0005	0.0360	-0.0835
$\bar{\hat{\rho}}_1$	0.0200	0.1924	0.1910	0.1916	0.1319
$(\bar{\hat{\rho}}_1 - \rho_1)/\rho_1$	0.0000	-0.0380	-0.0450	-0.0420	-0.3405
$\bar{\hat{\alpha}}$	0.0100	0.0100	0.0101	0.0103	0.0102
$(\bar{\hat{\alpha}} - \alpha)/\alpha$	0.0000	0.0000	0.0101	0.0103	0.0102

Sample size=95

Time step =0.1, Convergence criterion=1.0E-08

Number of simulations = 500

TABLE 3: Results of the VAM using simulated measurements.

Parameter Estimates and Biases: Model 3

Actual Parameters: $r=0.40$, $\theta=1.0$, $K = 1.0$, $\rho = -0.2$, $\alpha =0.01$

% Noise	0	5
$\bar{\hat{r}}$	0.4000	0.4080
$ (\bar{\hat{r}} - r) / r $	0.0000	0.0200
$\bar{\hat{\theta}}$	1.0000	1.0093
$(\bar{\hat{\theta}} - \theta) / \theta$	0.0000	0.0093
$\bar{\hat{K}}$	1.0000	0.9999
$(\bar{\hat{K}} - K) / K$	0.0000	-0.0001
$\bar{\hat{\rho}}$	-0.2000	-0.1995
$(\bar{\hat{\rho}} - \rho) / \rho$	0.0000	-0.0025
$\bar{\hat{\alpha}}$	0.0100	0.0100
$(\bar{\hat{\alpha}} - \alpha) / \alpha$	0.0000	0.0000

Sample size=95

Time step =0.1, Convergence criterion=1.0E-08

Number of simulations = 500

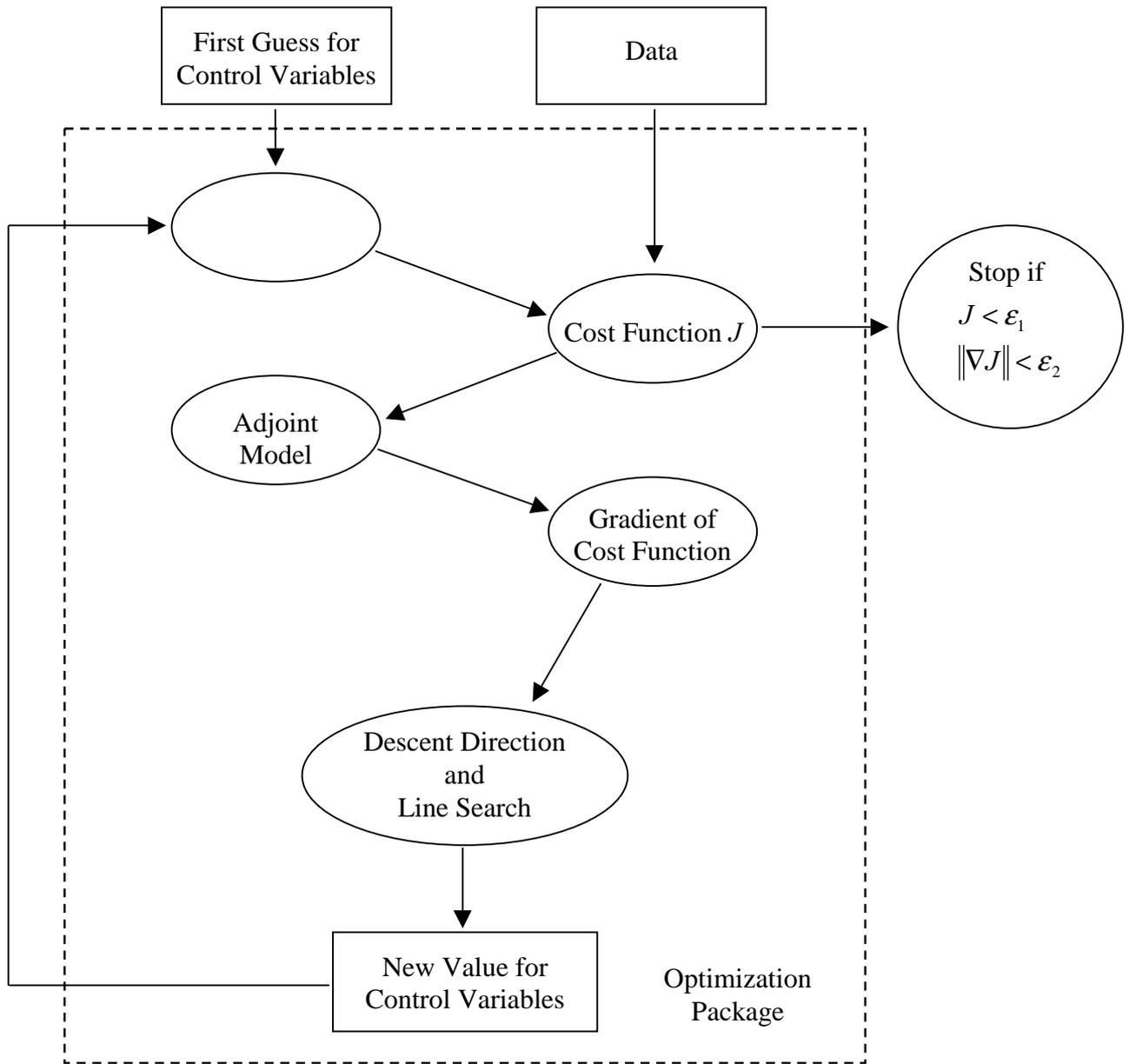


Figure 1. Steps involved in the data assimilation scheme. Figure adapted from Spitz et al. 1998. The convergence criterion is $J < \varepsilon_1$, $\|\nabla J\| < \varepsilon_2$, ε denotes a small number and $\|\cdot\|$ denotes the norm of the gradient of the cost function.