

APPENDIX: BAYESIAN REGRESSION METHODS

Introduction

Many of the regressions presented Chapters III to VI derive from Bayesian statistics. Bayesian methods are less familiar than some other statistical approaches in ecology, and for this reason it is appropriate to provide some methodological background. This Appendix provides a brief summary of the Bayesian regression methods that are employed for many analyses presented in the book.

Methods presented below come from the Bayesian theory for normally-distributed random variables (Gelman et al. 1995). Bayesian methods can be used for any probability distribution. The methods shown here are but a narrow subset of the possibilities.

The explanations in this Appendix assume that the reader is familiar with matrix algebra. Matrix algebra is explained in many elementary mathematics texts. The matrix algebra needed to follow most statistical calculations is explained in many textbooks, for example Draper and Smith (1981).

Prior Probabilities and Bayes' Theorem

The task of each Bayesian analysis is to build a model for the relationship between parameters (θ) and observables (y), and then calculate the probability distribution of parameters conditional on the data, $p(\theta|y)$. In addition, the Bayesian analysis may calculate the predicted distribution of unobserved data.

Bayesian statistics begins with a model for the joint probability distribution of θ and y , $p(\theta, y)$. θ may be a single parameter or a vector of many parameters, and y may be a vector of observations of a single variable or a matrix with multiple observations of many variables. The function p is a probability distribution. An example of a model is the familiar one for estimating the mean and variance of a normally distributed population, in which $p(\theta, y)$ is a normal distribution with mean and variance given by the parameter vector θ , and y is a sample of independent measurements.

Using the definition of conditional probability (Mangel and Clark 1988, Howson and Urbach 1989), $p(\theta, y)$ can be decomposed into two components:

$$p(\theta, y) = p(\theta) p(y | \theta) \tag{A.1}$$

By convention, $p(\theta)$ is called the prior distribution of θ (i.e. the distribution prior to observing the data y) and $p(y | \theta)$ is called the likelihood function (i.e. the likelihood of observing the data given a particular parameter value θ). Bayes' theorem provides the posterior probability distribution $p(\theta | y)$ (i.e. the distribution of θ obtained after observing y and combining the information in the data with the information in the prior distribution):

$$p(\theta | y) = p(\theta) p(y | \theta) / p(y) \tag{A.2}$$

Equation A.2 provides a probability distribution of θ given observations of the data y .

In this equation, $p(y)$ is the sum (or integral) of $p(\theta) p(y|\theta)$ over all possible values of θ . For discussions of these basic points of Bayesian analysis, see Mangel and Clark (1988) or Howson and Urbach (1989).

Subjectivity

Bayesian probabilities are sometimes called subjective probabilities. It is important to understand exactly what is meant by “subjective” in this context.

Decision analyses are often unique. The situation in which one is making the decision may occur only once. It cannot be replicated, so there is no possibility for measuring probabilities by repeated sampling. Nevertheless, Bayesian analysis may be used to compute the probabilities needed to make decisions. Because these probabilities cannot be measured by repeated sampling, they are called “subjective” and they represent a degree of belief in a particular outcome. For thorough discussions of this notion of subjectivity in the context of decision making, see Lindley (1985), Howson and Urbach (1989) and Pratt et al. (1995).

Also, if there is no basis in observed data for estimating the prior probability distribution, then the analyst may simply assume a particular prior distribution. The consequences of this assumption can be tested by sensitivity analyses that compare the response of the posterior distribution to different assumptions about the prior distribution. Most commonly, a *noninformative* prior distribution is assumed. A noninformative prior distribution assigns the same probability to each possible value of the parameters. If the number of observations is at least moderately large, a noninformative prior distribution will have negligible impact on the

posterior distribution. If the data y are limited, however, the choice of prior distribution may have a substantial impact on the posterior distribution. In this case, sensitivity analysis is needed to evaluate the consequences of different assumptions about the prior distribution.

Further reading

This Appendix is not intended to be a thorough review of Bayesian methods. Readers who are interested in Bayesian methods should consult one of the many excellent texts on the topic. Howson and Urbach (1989) describe the Bayesian approach to scientific reasoning from a philosophical perspective. The book clarifies the similarities and differences between Bayesian inference and other types of statistical inference. Bernardo and Smith (2000) give a thorough summary of Bayesian theory. On a more practical level, Gelman et al. (1995) provide a user's guide to Bayesian calculations. Their book covers several of the statistical models most familiar to ecologists from a Bayesian perspective. It is oriented toward application of statistics more than toward theory. Many of the methods presented by Gelman et al. (1995) can be computed using the software package WINBUGS which is available from the Internet (URL: <http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml>).

Bayesian time-series methods are also called dynamic regressions, because the parameter estimates are updated over time in a dynamic way. Pole et al. (1994) have written a practical guide to Bayesian time series analysis using models that are linear in the parameters. They also provide software for calculating Bayesian dynamic regressions. Detailed descriptions of their algorithm as implemented for ecological analyses are presented in appendices to Cottingham and Carpenter (1998) and Lamon et al. (1998). West and Harrison (1989) present the nonlinear

method used in this book, which is a straightforward extension of the linear method presented by Pole et al. (1994).

Bayesian statistics are used extensively in decision analysis. A brief introduction to decision theory was written by Lindley (1985). His book also covers many basic aspects of Bayesian statistics. The book by Pratt et al. (1995) is as much a textbook of Bayesian statistics as it is a book about decision theory. Decision theory is most familiar to ecologists as a method in behavioral ecology. Mangel and Clark (1988) provide an excellent introduction while Clark and Mangel (2000) have written an updated and wider survey of the uses of dynamic optimization models in ecology.

Bayesian Linear Regression

Linear Regression with Noninformative Prior

In linear regression, the observations consist of a response variable in a vector y and one or more predictor variables in a matrix X . The vector y has n elements, corresponding to n observations. The matrix X has n rows, corresponding to the observations, and k columns corresponding to the number of predictors. If the regression includes an intercept, one of the columns of X is a column of ones. The parameters are the regression coefficients β and the error variance of the fitted model, σ^2 . The model that relates observations and parameters is written:

$$(y \mid \beta, \sigma^2, X) \sim \text{Normal}(X\beta, \sigma^2 I) \tag{A.3}$$

In words, this model states that the distribution of y given parameters β and σ^2 and predictors X is a normal distribution with mean $X\beta$ and variance σ^2 . The identity matrix is I . A normal distribution is completely specified by its mean and variance.

Once the model is specified, the Bayesian analysis seeks the posterior distribution for the parameters and a predictive distribution for the model's predictions. The analysis begins with a prior distribution. A noninformative prior distribution that is commonly used for linear regression is

$$p(\beta, \sigma^2) \propto 1/\sigma^2 \quad (\text{A.4})$$

In words, this expression means that the joint probability distribution of β and σ^2 given X is a flat surface with a constant level proportional to $1/\sigma^2$. In many situations encountered in ecology, the outcome is not overly sensitive to this choice of prior distributions. For example, if the prior value of σ^2 is sufficiently large and the observations are many and informative, the effect of the likelihood will overshadow that of the prior distribution. It may be desirable to explore the sensitivity of the posterior distribution to the choice of prior distribution.

The posterior distribution of β given σ^2 is

$$\beta \mid \sigma^2, y \sim \text{Normal}(\beta_E, V_\beta \sigma^2) \quad (\text{A.5})$$

Expression A.5 states that the probability distribution of β given σ^2 and y is normal with mean β_E and variance $V_\beta \sigma^2$. The parameters of this normal distribution are computed from

$$\beta_E = (X' X)^{-1} X' y \quad (\text{A.6})$$

$$V_\beta = (X' X)^{-1} \quad (\text{A.7})$$

The apostrophe (') denotes matrix transposition.

The marginal posterior distribution of σ^2 (i.e. the integral over all possible values of β of the joint distribution of β and σ^2) is

$$\sigma^2 | y \sim \text{Inverse } \chi^2 (n - k, s^2) \quad (\text{A.8})$$

Expression A.8 says that the probability distribution of σ^2 given y follows an inverse χ^2 distribution. The inverse χ^2 distribution, presented by Gelman et al. (1995), is fully defined by two parameters, the degrees of freedom and the scale factor. In this case there are $n - k$ degrees of freedom (where n is the number of observations of y and k is the number of parameters to be estimated, i.e. the number of columns of X). The scale factor s^2 is computed by

$$s^2 = (y - X \beta_E)' (y - X \beta_E) / (n - k) \quad (\text{A.10})$$

Note that $y - X \beta_E$ is the vector of residuals, or deviations of observations from predictions.

The marginal posterior distribution of β given y is written

$$\beta | y \sim \text{Multivariate Student } t(n - k, \beta_E, s^2) \quad (\text{A.11})$$

The multivariate Student t distribution (presented by Gelman et al. 1995) has three parameters, the degrees of freedom $n - k$, the mean β_E , and the scale factor s^2 . This distribution is derived by integrating the posterior distribution of β given σ^2 (A.5) over all possible values of σ^2 (A.8).

Regressions are often fitted in order to make predictions. The predictive distribution, y_p , given a new set of predictors X_p has mean

$$E(y_p | y) = X_p \beta_E \quad (\text{A.12})$$

The marginal posterior distribution of the variance of this prediction is

$$\text{var}(y_p | \sigma^2, y) = (I + X_p V_\beta X_p') \sigma^2 \quad (\text{A.13})$$

where I is the identity matrix. This variance formula has two components, $I \sigma^2$ for sampling variance of the new observations and $X_p V_\beta X_p' \sigma^2$ for uncertainty about β . The marginal posterior distribution of y_p given y is

$$p(y_p | y) \sim \text{Multivariate Student-t}[n - k, X_p \beta_E, (I + X_p V_\beta X_p') s^2] \quad (\text{A.14})$$

Ecologists usually use Bayesian regression analysis to calculate the probability distribution of β or the probability distribution of predictions for new observations X_p . The probability distribution of β is given by A.11 with β_E computed by equation A.6 and s^2 computed by equation A.10. The probability distribution for new observations is given by A.14 with β_E computed by equation A.6, V_β computed by equation A.7 and s^2 computed by equation A.10.

The multivariate Student t distribution commonly arises in Bayesian regression. A number of useful properties of the univariate and multivariate Student t distributions are presented by Gelman et al. (1995), including the equation for ordinates of the distributions. These authors also explain how to compute a random sample from a Student t distribution. Probabilities are computed by integrating the Student t distribution using the incomplete beta function, as explained by Press et al. (1989). Many scientific software packages include subroutines for computing random samples and probabilities from the univariate and multivariate Student-t distributions.

The rationale for Bayesian regression analysis is different from the rationale for classical regression (e.g. Draper and Smith 1981). The Bayesian analysis is focused toward a conditional posterior distribution for the parameters and a predictive distribution for the model. The classical analysis focuses on point estimates of parameters or predictions, as well as the variances of these point estimates. The classical analysis is derived by a different sequence of steps which involve maximization of a likelihood function for model errors (Burnham and Anderson 1998). Despite these differences, the computations turn out to be similar to the

Bayesian analysis using the noninformative prior of equation A.4. The classical estimates of β and σ^2 are β_E and s^2 , respectively. The classical standard error estimate for β is $V_\beta s^2$. The classical prediction for new data is $y_p = X_p \beta_E$ with variance $(I + X_p V_\beta X_p') s^2$.

Linear Regression with Informative Prior

Bayesian analysis can be used to combine two different sources of information in a single model to estimate parameters or make predictions. The results can then be combined with a third source of information to improve the parameter estimates or predictions. This process can be repeated over and over again to combine information from any number of sources. Combining multiple sources of information is one of the most important uses of Bayesian statistics in ecology (Hilborn and Mangel 1997). Ecologists frequently have information from diverse experiments or field observations that could be combined to make predictions under a single model. Bayesian statistics offer a family of methods for such mergers of information. One of these methods is linear regression with an informative prior distribution.

In linear regression with an informative prior distribution, there are two statistically-independent data sets that provide information about the model to be analyzed. We assign one data set to be the prior, and use the other data set for the likelihood.

Usually it is convenient to assume that the prior distribution of the k regression parameters is multivariate Student-t. This distribution has three parameters, a vector of mean regression parameters, a matrix with variances along the main diagonal and covariances elsewhere, and degrees of freedom. In this case, the mean vector contains the k prior estimates

of the mean regression parameters (B_0) and the $k \times k$ parameter covariance matrix S_0 , model variance s_0^2 and degrees of freedom n_0 . For the second data set, we have a $n_1 \times 1$ response vector y_1 and a $n_1 \times k$ matrix of predictors X_1 .

The posterior can be computed by treating the prior as additional data points, and then weighting their contribution to the posterior (Gelman et al. 1995). The weights are inversely proportional to the variance of the prior parameter estimates. To perform the computations, construct a new vector of observations y , predictor matrix X , and weight matrix Σ as follows:

$$y = [y_1' \ B_0']' \quad (\text{A.15})$$

$$X = [X_1' \ I_k']' \quad (\text{A.16})$$

$$\Sigma = [I_n \ 0 ; 0 \ S_0] \quad (\text{A.17})$$

Here I_k and I_n stand for the identity matrices of dimension k and n_1 , respectively. 0 denotes the zero elements necessary to fill out the square matrix Σ , which has I_n as the upper left elements and S_0 as the lower right elements.

The posterior estimate of the parameters follows a multivariate Student-t distribution (Gelman et al. 1995). The mean vector β_E is

$$\beta_E = (X' \Sigma^{-1} X)^{-1} X' \Sigma^{-1} y \quad (\text{A.18})$$

The scale matrix is $s^2 V$ with $n-k$ degrees of freedom, where n is the number of observations, k is the number of parameters estimated from the data, and

$$V = (X' \Sigma^{-1} X)^{-1} \quad (\text{A.19})$$

The variance s^2 with $n_0 + n_1 - k$ degrees of freedom is calculated as

$$s^2 = (n_0 s_0^2 + n_1 s_1^2) / (n_0 + n_1) \quad (\text{A.20})$$

where the variance s_1^2 is calculated as

$$s_1^2 = (y - X \beta_E)'(y - X \beta_E) / (n_1 - k) \quad (\text{A.21})$$

Predictions for new data X_p using the regression for both existing data sets also follow a multivariate Student t distribution:

$$p(y_p | y) \sim \text{Multivariate Student-t} [n_0 + n_1 - k, X_p \beta_E, (I + X_p V X_p') s^2] \quad (\text{A.22})$$

In this distribution X_n is the new set of predictors, β_E is computed with A.18, V is computed with A.19 and s^2 is computed with A.20 and A.21.

In summary, the linear regression with two independent data sets is computed as follows. First, express the results from the first data set as a multivariate Student t distribution for the

parameters. This distribution follows from a Bayesian regression analysis of the first set of data using noninformative priors. Then, form the new response vector, predictor matrix, and weight matrix using A.15 – A.17. The posterior distribution of parameters from both sets of data is multivariate Student t with mean given by A.18, covariance matrix $s^2 V$ computed using A.19 and A.20, and $n_0 + n_1 - k$ degrees of freedom. Predictions for new data follow the multivariate Student t distribution given by A.22.

Bayesian Inverse Modeling

Methods covered so far in this appendix are appropriate for models in which the response variable is a linear function of the predictors. Some models in this book involve nonlinear functions of the predictors. Parameters of these models can be estimated by various methods. The method used here estimates parameters from multiple data sets using Bayes' theorem. Earth scientists call this method Bayesian inverse modeling (Jackson and Matsu'ura 1985). "Inverse modeling" is a term used by physical scientists to describe parameter estimation for complex models, often in situations where the models are poorly constrained by data.

The general form of the nonlinear models used in this book is

$$y = f(\beta, X) + E, E \sim N(0, \sigma^2) \tag{A.23}$$

where f is some nonlinear function, y is a vector of observations of the response variable, X is a matrix of predictors, β is a vector of parameters, and E is a vector of residuals (deviations of y

from model predictions). Usually the errors are assumed to be normally distributed as shown in A.23. In this situation, E can be written as a function of the unknown parameters, $E = y - f(\beta, X)$.

In this book, equations like A.23 were fit to multiple sources of data using Bayes' theorem, as described below. Before explaining that method, however, it is useful to mention the maximum likelihood method for estimating the parameters of A.23. The unknown parameters β and σ^2 (the variance of the residuals) can be found by minimizing

$$-\log(L) = 0.5 [(E' E / \sigma^2) + n \log (2 \pi \sigma^2)] \quad (\text{A.24})$$

This function is called the negative log likelihood function (Hilborn and Mangel 1997). The values of β and σ^2 that minimize A.24 will maximize the likelihood L . The likelihood of a single residual is its y-axis value from the normal distribution of residuals. The likelihood of all the residuals taken together is the product of the likelihoods of the individual residuals. The log likelihood, then, is the sum of the logs of the likelihoods of the individual residuals. Equation A.24 is the negative log of that collective likelihood. Computationally, it is more convenient to minimize the negative log likelihood than it is to maximize the likelihood.

This book uses a conceptually different, though computationally similar, method for estimating parameters of nonlinear models. This approach combines multiple sources of information using Bayes' theorem. Bayesian inverse modeling is the name given by geophysicists to the process of pooling two types of measurements for the same parameter (Jackson and Matsu'ura 1985). In Bayesian inverse modeling, a prior distribution for certain

parameters is combined with data from a second, independent source to calculate a posterior distribution for the parameters. By an argument analogous to the one used for pooling data to estimate parameters for linear models, Jackson and Matsu'ura (1985) show that parameters of a nonlinear model can be estimated by minimizing the sum of a negative log likelihood plus the negative log prior probability density:

$$-\log(L_P) = -\log(L_0) - \log(L_1) \quad (\text{A.25})$$

L_P stands for the negative log posterior density that results from pooling a prior (L_0) probability distribution from the first data set and a likelihood (L_1) from the second data set. The prior distribution for the parameters has mean B_0 and covariance matrix S_0 . As above, $E = y - f(\beta, X)$. The negative log of the prior distribution is

$$-\log(L_0) = 0.5 [(\beta - B_0)' S_0^{-1} (\beta - B_0) + k \log(2 \pi) + \log (|S_0|)] \quad (\text{A.26})$$

where β is the vector of parameters to be estimated. The negative log likelihood function from the second set of data is

$$-\log(L_1) = 0.5 [(E' E / \sigma^2) + n \log (2 \pi \sigma^2)] \quad (\text{A.27})$$

In summary, to estimate parameters of a nonlinear model from two data sets, first obtain parameter estimates B_0 and an approximate covariance matrix S_0 for the data set which will provide the prior distribution. Often, the covariance matrix contains only the prior variances,

with zeros elsewhere. The prior parameters B_0 and S_0 can be estimated by minimizing a negative log likelihood such as A.24 and then bootstrapping the covariance matrix of parameters, for example (Efron and Tibshirani 1993). Then, compute the pooled parameter estimates by minimizing the negative log probability density A.25 over the parameters β and σ^2 . Several computational methods are available for minimizing the negative log probability density (Press et al. 1989). Results reported in this book used the simplex method of Nelder and Mead (1964).

This method is an approximation, and there is no guarantee that it will work. It is a good idea to examine plots of the likelihood against each parameter, as well as the usual plots of residuals recommended in statistics textbooks (e.g. Draper and Smith 1981). Often the approximation seems reasonable.

The posterior probability distribution of the parameters will not resemble a standard distribution, such as the multivariate normal or multivariate Student t distribution. It is possible to calculate an approximate covariance matrix for the parameters by various methods, such as linearizing near the optimal parameter estimates or bootstrapping (Draper and Smith 1981, Efron and Tibshirani 1993). Full Bayesian analysis of the posterior distribution may be difficult. Unlike the linear case in the preceding section, it is not possible to write a general summary of Bayesian results for nonlinear models.

Bayesian Dynamic Regression

Previous sections showed how Bayesian methods can be used to estimate parameters from two or more data sets. This approach can be used to update parameter estimates using

sequential monitoring data in a process called Bayesian Dynamic Regression (Pole et al. 1994, West and Harrison 1989). Suppose that an ecosystem is observed at regular intervals in time, and that some error is associated with each observation. Prior to making a new observation, one can predict the next state of the system using a model for the observations so far. This predictive distribution is a prior distribution conditional on the time series data observed so far. Then new observations provide a second estimate of the state of the system, which may be summarized in a likelihood function. The prior distribution and the likelihood are combined by Bayesian dynamic regression into new posterior distributions for the parameters, the current state of the system, and the predictive distribution for the next time step. Then the process is repeated for successive time steps. The result is a regularly-updated calculation of the conditional posterior density of the parameters and the predictive probability distribution of future ecosystem states.

Chapters IV and V use nonlinear Bayesian dynamic regressions, meaning that the predictions of the regression model are a nonlinear function of the parameters. Theory is well developed for linear Bayesian dynamic regression models, in which the predictions are a linear function of the parameters (Pole et al. 1994, West and Harrison 1989). Nonlinear Bayesian dynamic regressions are still an active area of research. I used the algorithm of West and Harrison (1989). It adapts the well-understood linear case to nonlinear models by using the first-order terms of a Taylor expansion of the model in the neighborhood of the most recent observation.

In dynamic regression, the relationship between observations and unknown parameters is described by an *observation equation*

$$Y_t = F_t(\theta_t, X_t) + w_t \quad (\text{A.28})$$

The observation equation relates system state variables, Y_t , to parameters θ , and observed regressors X_t . The parameters θ are not observed directly, but can be estimated from the observations of system state Y_t and predictors X_t . Changes in the parameters over time are described by a *system equation*:

$$\theta_t = g_t(\theta_{t-1}) + v_t \quad (\text{A.29})$$

F and g are functions (possibly nonlinear) of the parameter vector θ . F and g must be differentiable. We assume that observation and process errors w and v are independent and normally distributed with means of zero and standard deviations s_w and s_v , respectively. Note that θ receives a random shock v_t at each time step. We assume that this random drift in θ can be represented by inflating the variance of our estimate at each time step (or, equivalently, discounting the precision of our estimate at each time step). We will specify a parameter δ ($0 < \delta < 1$) to discount the precision (i.e. increment the variance) of θ at each time step. Maximum-likelihood estimates of δ typically lie between 0.95 and 0.99. In practice, it turns out that results are not very sensitive to δ when δ is near 1 (Pole et al. 1994, West and Harrison 1989).

At the beginning of any time step t , we have posterior information which consists of the current estimate of θ , m_{t-1} , the covariance of this estimate C_{t-1} , and the current degrees of freedom n_{t-1} . These are the parameters of a multivariate Student t distribution for θ_{t-1} given data

available through time $t-1$. For the first time step, before any observations are made, the values of m , C and n will depend on the amount of prior information. If there is no prior information, the diagonal elements of C (which correspond to the parameter variances) can be set to a large number. The computations can be repeated with different values for the initial parameter variance to assess the sensitivity of the outcome to assumptions about the prior. Usually the effects of the prior are negligible after the first few observations have been made.

The sequence of calculations for any time step t is as follows. First, calculate the prior estimate for θ , a_t , which is

$$a_t = g_t(m_{t-1}) \tag{A.30}$$

with covariance matrix R_t calculated as

$$R_t = G_t C_{t-1} G_t' + ((1/\delta) - 1)C_{t-1} \tag{A.31}$$

G_t is the matrix of partial derivatives of g_t with respect to each parameter (element of m_{t-1}), evaluated at the posterior values for time $t-1$. In the linear case, G is simply the matrix of predictors at the previous time step.

Next, calculate the one-step ahead forecast of the system state Y_t , which will be labeled f_t :

$$f_t = F_t(a_t) \quad (\text{A.32})$$

The covariance matrix for the one-step ahead forecast is calculated as

$$Q_t = H_t' R_t H_t + S_{t-1} \quad (\text{A.33})$$

H_t is the matrix of partial derivatives of F_t with respect to each element of a_t , evaluated using the prior estimates for time t . The marginal distribution of the predictions is multivariate Student t , with n_{t-1} degrees of freedom and scale matrix $Q_t S_{t-1}$. In the linear case, H is simply the vector of parameter estimates at time t , a_t .

At this point, the new observation is made and the parameters are updated as follows.

The prediction errors are $E_t = Y_t - f_t$. It is convenient to calculate a matrix $A_t = R_t H_t Q_t^{-1}$. Then the updated parameter vector is

$$m_t = a_t + A_t E_t \quad (\text{A.34})$$

with degrees of freedom

$$n_t = \delta n_{t-1} + 1 \quad (\text{A.35})$$

and covariance matrix

$$C_t = R_t - A_t Q_t A_t' \quad (\text{A.36})$$

The updated covariance matrix for the estimate of current system state is

$$S_t = [n_{t-1} S_{t-1} + E_t Q_t^{-1} E_t'] / n_t \quad (\text{A.37})$$

The calculations A.30 to A.37 are repeated for each time step.

In summary, the Bayesian dynamic regression is computed as follows. Before any data are observed, a prior multivariate Student t distribution is chosen for the parameters. This distribution may be based on data, or it may be a prior with low information content (large diagonal elements in the covariance matrix). At the beginning of each time step, prior estimates of the parameters are calculated by A.30 and A.31. Prior forecasts of the system state variables are computed using A.32 and A.33. Then, the new data are observed. One then calculates the updated parameter distribution (A. 34 – A.36) and the new covariance matrix for the estimate of current system state (A.37). The sequence A.30 – A.37 is repeated at each time step.