# Bayesian State Space Model


## Abstract

This methodological framework examines the Bayesian State Space Model (BSSM), which is particularly suitable for estimating model parameters with a high degree of confidence when working with short time series. The BSSM is implemented through several stages, including the Variational Bayesian M-step (maximisation), Variational Bayesian E-step (expectation), calculation of the lower bound of the logarithmic marginal likelihood, updating of hyperparameters, and comparison of the lower bound of the logarithmic marginal likelihood. The objective of this article is to expand the range of modelling approaches used to improve the accuracy of estimates of the fundamental value of housing prices in Uzbekistan's real estate market.


**Keywords:** covariance matrix, marginal probability, error, prior distribution, posterior distribution, time series.

Prepared by the Financial Stability Department of the Central Bank of Uzbekistan.

If you have any comments, please send them to udjumanazarov@cbu.uz.

# Bayesian State Space Model

Bayesian state space model (BSSM) is considered more efficient than the state space model (SSM)[1] for estimating parameters with a high degree of confidence in short time series. In particular, BSSM provides opportunities for managing uncertainty and adapting parameters to observed data. It allows for rapid updating of posterior distributions[2], enabling better representation of dynamic changes in data over short time series.

BSSM treats all uncertain quantities as random variables and uses the laws of probability. These models are widely used in the fields of signal filtering and prediction. To quantify uncertainties in the model, posterior distributions over all the parameters are required. In particular, the model consists of an observable or measurement equation representing the level of interdependence and a transition equation[3] representing dynamic state variables:

$$y_t = Cx_t + Du_t + v_t$$
$$x_t = Ax_{t-1} + Bu_t + w_t .$$

Where,
$y_t$ – vector of observable endogenous variables;
$x_t$ – vector of hidden state variables
$u_t$ – vector of observable exogenous variables;
$A$ – $kxk$ state dynamics matrix;
$C$ – $pxk$ observation matrix;
$B$ – $kxd$ input-to-state matrix;
$D$ – $pxd$ input-to-observation matrix;
$v_t \sim N(0, R)$ – vector of measurement errors for the observation equation;
$w_t \sim N(0, Q)$ – vector of measurement errors for the state equation;
$A$, $B$, $C$ and $D$ matrices – matrices of estimated coefficients;
$Q$ and $R$ – covariance matrices for errors.

The displacements in the observation space can be learnt as parameters of the input-to-state and input-to-observation matrices. Since the hidden state is unobserved, state

[1] State space model (SSM) methodology is presented in the Methodological Framework of Central Bank.

[2] Posterior distribution is the updated distribution of a parameter after observing the data.

[3] Beal, M. (2003). Variational Algorithms for Approximate Bayesian Inference. University of Cambridge.

noise covariance matrix $Q$ can be incorporated into the state dynamics matrix $A$. As a result, hidden state rescaled according to this change. Covariance matrix, $R$, of output noise can not be incorporated to observation matrix $C$ since the data is observed. In particular, the output noise covariance matrix $R$, which is assumed diagonal, is defined through the precision vector $\rho$ as follows:

$$R^{-1} = diag(\rho)$$

Here, $R$ denotes the covariance matrix of output noise, $diag(p)$ denotes a matrix with all offdiagonal elements equal to zero.

For conjugacy[4], each dimension of $\rho$ is assumed to be gamma distributed with hyperparameters $a$ and $b$. Prior distribution of $\rho$ is defined as follows:

$$p(\rho|a,b) = \prod_{s=1}^{p} \frac{b^a}{\Gamma(a)} \rho_s^{a-1} \exp\{-b\rho_s\}.$$

The parameters and hyperparameters are linked through following probabilities:

$$p(a_{(j)}|\alpha) = N(a_{(j)}|0, diag(\alpha)^{-1})$$
$$p(b_{(j)}|\beta) = N(b_{(j)}|0, diag(\beta)^{-1})$$
$$j = 1,2,\dots,k$$
$$p(c_{(s)}|\rho_s,\gamma) = N(c_{(s)}|0, {\rho_s}^{-1} diag(\gamma)^{-1})$$
$$p(d_{(s)}|\rho_s,\delta) = N(d_{(s)}|0, {\rho_s}^{-1} diag(\delta)^{-1})$$
$$p(\rho_s|a,b) = Ga(\rho_s|a,b)$$
$$s = 1,2,\dots,p.$$

Where,

$a_{(j)}$ – $j$th column of matrix $A$;

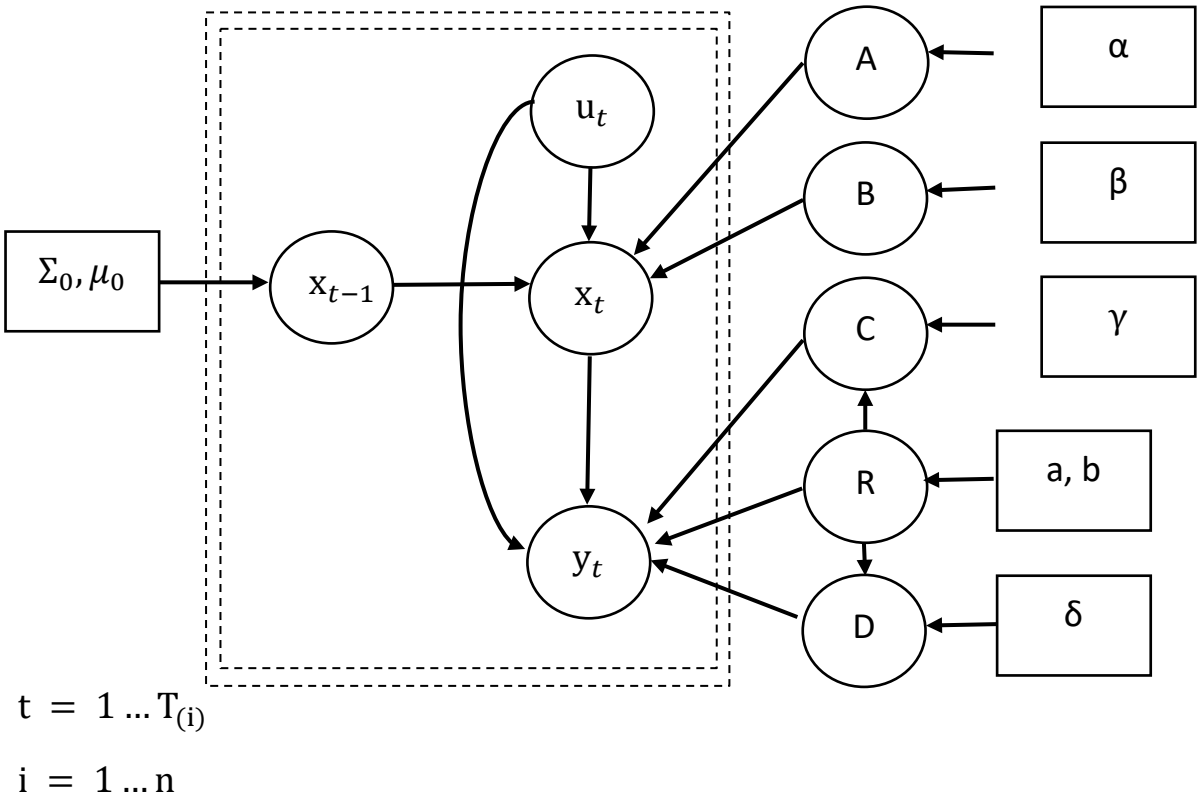
$b_{(j)}$ – $j$th column of matrix $B$;

$c_{(s)}$ – $s$th column of matrix $C$;

$d_{(s)}$ – $s$th column of matrix $D$.

[4] Conjugation is the method of expressing the prior and posterior distributions in the same form to simplify calculations.

**Figure 1. BSSM Overview**




Source: Beal, M. (2003). Variational Algorithms for Approximate Bayesian Inference. University of Cambridge.

In addition, for each row of matrix $A$, zero mean Gaussian prior has precision[5] equal to $diag(\alpha)$. Each row of $C$ matrix is given a zero mean Gaussian prior with precision equal to $diag(\rho s \gamma)$. The dependence of the precision of $c(s)$ on the noise output precision $\rho s$ is motivated by conjugacy. Two more hyperparameter vectors $\beta$ and $\delta$ are placed on the rows of the inputrelated matrices $B$ and $D$. The covariances across the rows of $A$ and $B$ matrices remain unchanged. However, the covariances within the rows of matrices $C$ and $D$ may change as a result of $\rho s$ . If the $j$th column of $A$ become zero, the $j$th column at time $t - 1$ is not involved in generating the hidden state at time $t$. However, $j$th hidden dimension is used in producing covariance structure in the data at each time step.

BSSM is implemented in six steps. The first step is initialization, in which the auxiliary hidden state is set, prior distribution[6] of parameters is specified, and the log

[5] In the distributions, the precision is equal to the inverse of the variance, meaning that the product of the precision and the variance is equal to one.

[6] Prior distribution is the distribution of the parameters before any data are observed.

marginal probability[7] is evaluated using Jensen's inequality. The second step is Variational Bayesian M-step (Maximisation). At this step, the optimal forms of the posterior distributions of the parameters are determined and the natural parameter vector is calculated. Next, in the Variational Bayesian E-step (Expectation), the distributions of hidden state variables vector are determined. In the fourth step, the lower bound on the marginal probability, the value of $F$ is computed. The fifth is the hyperparameter updating step, in which hyperparameters are adjusted to maximize lower bound of the marginal probability. In the final step, the lower bound estimate of the log marginal probability is compared with its previous value, and if the lower bound has increased, the Variational Bayesian M-step is repeated. These computational steps are continued iteratively until the lower bound estimate no longer exceeds its previous value.

**Initialization**

At first, for time $t = 0$, auxiliary hidden state $x_0$ is initialized. In this context, the auxiliary hidden state $x_0$ follows Gaussian distribution with mean $\mu_0$, covariance $\Sigma_0$, and prior distribution is defined as follows:

$$\begin{aligned} p(x_0 | \mu_0, \Sigma_0) &= N(x_0 | \mu_0, \Sigma_0) \\ \mu_0 &\sim N(\mu_0 | 0, b_{\mu_0} \mathrm{I}) \\ \Sigma_0 &\sim \prod_{j=1}^{k} Ga\,(\Sigma_{0jj}^{-1} | a_{\Sigma_0}, b_{\Sigma_0}) \end{aligned}$$

Where,
$x_0$ – auxiliary hidden state;
$\mu_0$ – mean of auxiliary hidden state;
$\Sigma_0$ – covariance of auxiliary hidden state.

In this case, the prior of $x_1$ via the state dynamics process is denoted by the following:

$$\begin{aligned} p(x_1 | \mu_0, \Sigma_0, \theta) &= \int dx_0\, p(x_0 | \mu_0, \Sigma_0)\, p(x_1 | x_0, \theta) \\ &= N(x_1 | A\mu_0 + Bu_1,\ A^T \Sigma_0 A + Q)\,. \end{aligned}$$

[7] Marginal probability is the likelihood of an event occurring independently of other events.

For simplicity, the parameters are integrated into a single parameter vector $\theta = (A, B, C, D, R)$. In addition, the dependence of $A$, $B$, $C$ and $D$ matrices with hyperparameters is given by the following hyperprior distribution:

$$\alpha \sim \prod_{j=1}^{k} Ga\,(\alpha_{(j)}|a_{(\alpha)}, b_{(\alpha)})$$
$$\beta \sim \prod_{c=1}^{d} Ga\,(\beta_{(c)}|a_{(\beta)}, b_{(\beta)})$$
$$\gamma \sim \prod_{j=1}^{k} Ga\,(\gamma_{(j)}|a_{(\gamma)}, b_{(\gamma)})$$
$$\delta \sim \prod_{c=1}^{d} Ga\,(\delta_{(c)}|a_{(\delta)}, b_{(\delta)})$$

The marginal probability takes the following form:

$$p(y_{1:T}) = \int dA\, dB\, dC\, dD\, d\rho\, dx_{0:T}\, p(A, B, C, D, \rho, x_{0:T}, y_{1:T})$$

Due to some conditional independencies between the parameters of the model, it is possible to see how posterior distributions for the dynamics and output processes decompose. The full joint probability for parameters, hidden variables and observed data, taking into account the inputs, is given as follows:

$$p(A, B, C, D, \rho, x_{0:T}, y_{1:T}|u_{1:T}) =$$
$$p(A|\alpha)p(B|\beta)p(\rho|a, b)p(C|\rho, \gamma)p(D|\rho, \delta)p(x_0|\,\mu_0, \Sigma_0) \cdot$$
$$\prod_{t=1}^{T} p(x_t|\,x_{t-1}, A, B, u_t)\, p(y_t|\,x_t, C, D, \rho, u_t)$$

Next, the dependence on the input sequence $u_{1:T}$ is dropped and inputs is left as implicit. A distribution $q(\theta, x)$ over the parameters and hidden variables is introduced. By applying Jensen's inequality lower bound the marginal probability is estimated:

$$\ln p(y_{1:T}) = \ln\left(\int dA\, dB\, dC\, dD\, d\rho\, dx_{0:T}\, p(A, B, C, D, \rho, x_{0:T}, y_{1:T})\right) \geq$$
$$\int dA\, dB\, dC\, dD\, d\rho\, dx_{0:T}\, q(A, B, C, D, \rho, x_{0:T}) \ln \frac{p(A,B,C,D,\rho,x_{0:T},y_{1:T})}{q(A,B,C,D,\rho,x_{0:T})} = F$$

The next step in the variational approximation is to assume some approximate form for the distribution $q(\cdot)$. First, hidden variables and parameters are separated based on conditional independency:

$$q(A, B, C, D, \rho, x_{0:T}) = q_\theta(A, B, C, D, \rho) q_x(x_{0:T})$$
$$q(A, B, C, D, \rho, x_{0:T}) = q_{AB}(A, B)\, q_{CD\rho}(C, D, \rho)\, q_x(x_{0:T})\,.$$

To simplify the estimation of the lower bound of the log marginal probability, the following terms of conditional probability are used:

$$q_{AB}(A,B) = q_B(B)\, q_A(A|B)$$
$$q_{CD\rho}(C,D,\rho) = q_\rho(\rho)\, q_D(D|\rho)\, q_C(C|D,\rho)$$

Using the above terms, the lower bound of the log marginal probability, $F$ is expressed as the following sum:

$$F = \int dB\; q_B(B) \ln\frac{p(B|\beta)}{q_B(B)} + \int dB\; q_B(B) \int dA\; q_A(A|B)\; \ln\frac{p(A|\alpha)}{q_A(A|B)} +$$
$$\int d\rho\; q_\rho(\rho)\; \ln\frac{p(\rho|a,b)}{q_\rho(\rho)} + \int d\rho\; q_\rho(\rho) \int dD q_D(D|\rho)\; \ln\frac{p(D|\rho,\delta)}{q_D(D|\rho)} +$$
$$\int d\rho\; q_\rho(\rho) \int dD\; q_D(D|\rho) \int dC\; q_C(C|\rho,D)\; \ln\frac{p(C|\rho,\gamma)}{q_C(C|\rho,D)} -$$
$$\int dx_{0:T}\; q_x(x_{0:T}) \ln q_x(x_{0:T}) + \int dB\; q_B(B) \int dA\; q_A(A|B) \int d\rho\; q_\rho(\rho) \cdot$$
$$\int dD\; q_D(D|\rho) \int dC\; q_C(C|\rho,D) \int dx_{0:T}\; q_x(x_{0:T}) \ln p(x_{0:T}, y_{1:T}|A,B,C,D,\rho) =$$
$$F(q_x(x_{0:T}), q_B(B), q_A(A|B), q_\rho(\rho), q_D(D|\rho), q_C(C|\rho,D))$$
$$H\big(q_x(x_{0:T})\big) = -\int dx_{0:T} q_x(x_{0:T}) \ln q_x(x_{0:T})$$

The lower bound of the log marginal probability depends on the hyperparameters. The optimum forms of these approximate posteriors can be found by setting functional derivatives of $F$ equal to zero with respect to each distribution over parameters and hidden variable sequences.

**Variational Bayesian M-step (Maximisation), VBM**

In this step, the variational posterior distributions over the parameters are determined by applying Bayesian theorem and from these the expected natural parameter vector is computed. The posterior distribution is factorized in the following form:

$$q_\theta(A,B,C,D,\rho) = \prod_{j=1}^{k} q\big(b_{(j)}\big)\, q\big(a_{(j)}\big|b_{(j)}\big)\; \prod_{s=1}^{p} q(\rho_s) \cdot$$
$$q(d_{(s)}|\rho_s) q(c_{(s)}|\rho_s, d_{(s)})$$

Where,
$b_{(j)}$ – $j$th row of $B$ matrix;

$a_{(j)}$ – $j$th row of $A$ matrix;
$d_{(s)}$ – $s$th row of $D$ matrix;
$c_{(s)}$ – $s$th row of $C$ matrix;

Some statistics of the input and output variable vectors, required for the calculations, is defined as follows:

$$\ddot{U} = \sum_{t=1}^{T} u_t\, u_t^T$$
$$U_y = \sum_{t=1}^{T} u_t\, y_t^T$$
$$\ddot{Y} = \sum_{t=1}^{T} y_t\, y_t^T$$

The sufficient statistics used to compute the posterior distributions of the parameters and the natural parameter vector are obtained using the following equations:

$$W_A = \sum_{t=1}^{T} < x_{t-1} x_{t-1}{}^T > = \sum_{t=1}^{T} \Upsilon_{t-1,t-1} + \omega_{t-1}\omega_{t-1}{}^T$$
$$G_A = \sum_{t=1}^{T} < x_{t-1} > u_t{}^T = \sum_{t=1}^{T} \omega_{t-1}\, u_t{}^T$$
$$\widetilde{M} = \sum_{t=1}^{T} u_t < x_t{}^T > = \sum_{t=1}^{T} u_t\, \omega_t{}^T$$
$$S_A = \sum_{t=1}^{T} < x_{t-1} x_t{}^T > = \sum_{t=1}^{T} \Upsilon_{t-1,t} + \omega_{t-1}\omega_t{}^T$$
$$W_C = \sum_{t=1}^{T} < x_t x_t{}^T > = \sum_{t=1}^{T} \Upsilon_{t,t} + \omega_t\omega_t{}^T$$
$$G_c = \sum_{t=1}^{T} < x_t > u_t{}^T = \sum_{t=1}^{T} \omega_t\, u_t{}^T$$
$$S_c = \sum_{t=1}^{T} < x_t > y_t{}^T = \sum_{t=1}^{T} \omega_t\, y_t{}^T\ .$$

The posterior distributions for $A$ and $B$ are defined by the following equations:

$$q_B(B) = \prod_{j=1}^{k} N(b_{(j)} |\, \Sigma_B \bar{b}_{(j)},\, \Sigma_B)$$
$$q_A(A|B) = \prod_{j=1}^{k} N(a_{(j)} |\, \Sigma_A \left[S_{A(j)} - G_A b_{(j)}\right], \Sigma_A)$$
$$\Sigma_A^{-1} = diag(\alpha) + W_A$$
$$\Sigma_B^{-1} = diag(\beta) + \ddot{U} - G_A^T \Sigma_A G_A$$
$$\bar{B} = \widehat{M}^T - S_A^T\, \Sigma_A G_A$$
$$q_A(A) = \prod_{j=1}^{k} N(a_{(j)} |\, \Sigma_A \left[S_{A(j)} - G_A \Sigma_B \bar{b}_{(j)}\right],\, \tilde{\Sigma}_A)$$
$$\tilde{\Sigma}_A = \Sigma_A + \Sigma_A G_A \Sigma_B G_A^T \Sigma_A$$

Here, $\bar{b}_{(j)}$ denotes $j$th row of $\bar{B}$ matrix, while $S_{A(j)}$ defines $j$th column of $S_A$ matrix.

Additionally, the posterior distributions over $\rho$, $C$ and $D$ are determined by the following:

$$q_\rho(\rho) = \prod_{s=1}^{p} Ga(\rho_s|\, a + \frac{T}{2}, b + \frac{1}{2}G_{ss})$$
$$q_D(D|\rho) = \prod_{s=1}^{p} N(d_{(s)}|\, \Sigma_D \bar{d}_{(s)},\, \rho_s^{-1}\Sigma_D)$$
$$q_C(C|D,\rho) = \prod_{s=1}^{p} N(c_{(s)}|\, \Sigma_C \left[S_{C(s)} - G_C d_{(s)}\right], \rho_s^{-1}\Sigma_C)$$
$$\Sigma_C^{-1} = diag(\gamma) + W_C$$
$$\Sigma_D^{-1} = diag(\delta) + \ddot{U} - G_C^T \Sigma_C G_C$$
$$G = \ddot{Y} - S_C^T \Sigma_C S_C - \bar{D}\Sigma_D \bar{D}^T$$
$$\bar{D} = {U_Y}^T - S_C^T \Sigma_C G_C$$
$$q_C(C|\rho) = \prod_{s=1}^{p} N\left(c_{(s)}\middle|\, \Sigma_C \left[S_{C(s)} - G_C \Sigma_D \bar{d}_{(s)}\right], \rho_s^{-1}\hat{\Sigma}_C\right)$$
$$\hat{\Sigma}_C = \Sigma_C + \Sigma_C G_C \Sigma_D G_C^T \Sigma_C$$

Here, $\bar{d}_{(s)}$ and $S_{C(s)}$ denote $s$th row of $\bar{D}$ and $s$th column of $S_C$ matrices, respectively.

In the Variational Bayesian M-step (Maximisation), $\varphi(\theta)$ is considered the vector of expected natrural parameters. These will then be used in the Variational Bayesian E-step (Expectation) which infers the distribution $q_x(x_{0:T})$ over hidden states in the system. The vector of relevant natural parameters is given by the following form:

$$\varphi(\theta) = \varphi(A,B,C,D,R) = [A, A^T A, B, A^T B, C^T R^{-1} C,$$
$$R^{-1}C,\, C^T R^{-1} D, B^T B, R^{-1}, \ln|R^{-1}|,\, D^T R^{-1} D,\, R^{-1} D]\,.$$
$$< A > = [S_A - G_A \Sigma_B \bar{B}^T]^T \Sigma_A$$
$$< A^T A > = < A >^T < A > + k\,[\Sigma_A + \Sigma_A G_A \Sigma_B G_A^T \Sigma_A]$$
$$< B > = \bar{B}\Sigma_B$$
$$< A^T B > = \Sigma_A\,[S_A < B > - G_A\,\{< B >^T < B > + k\Sigma_B\}]$$
$$< B^T B > = < B >^T < B > + k\Sigma_B$$
$$< \rho_s > = \bar{\rho_s} = \frac{a_\rho + {}^{T}/_{2}}{b_\rho + {}^{G_{ss}}/_{2}}$$
$$< ln\rho_s > = \overline{ln\rho_s} = \Psi\left(a_\rho + \frac{T}{2}\right) - \ln(b_\rho + \frac{G_{ss}}{2})$$
$$< R^{-1} > = diag(\bar{\rho})$$
$$< C > = [S_C - G_C \Sigma_D \bar{D}^T]^T \Sigma_C$$
$$< D > = \bar{D}\Sigma_D$$
$$< C^T R^{-1} C > = < C >^T\, diag(\bar{\rho}) < C > + p\,[\Sigma_C + \Sigma_C G_C \Sigma_D G_C^T \Sigma_C]$$
$$< R^{-1} C > = diag(\bar{\rho}) < C >$$

$$< C^T R^{-1} D \ > = \Sigma_C \ [S_C diag(\bar{\rho}) < D > - G_C < D >^T \ diag(\bar{\rho}) < D > - pG_C \Sigma_D]$$
$$< R^{-1} D > = \ diag(\bar{\rho}) < D >$$
$$< D^T R^{-1} D \ > = < D >^T \ diag(\bar{\rho}) < D > \ + p\Sigma_D \ .$$

Here, $<\cdot>$ denotes expectation with respect to the variational posterior.

**Variational Bayesian E-step (Expectation), VBE**

The posterior distribution of the hidden state variable vector $x_{0:T}$ is jointly Gaussian over the time steps, and the logarithm of the posterior distribution of the hidden state vector is expressed as follows:

$$\ln q_x(x_{0:T}) = \ - \ln Z + < \ln p(A, B, C, D, \rho, x_{0:T}, y_{1:T}) >$$
$$= - \ln Z' + < \ln p(x_{0:T}, y_{1:T} | \ A, B, C, D, \rho) >$$
$$Z' \ = \ \int dx_{0:T} \ exp < \ln p(x_{0:T}, y_{1:T} | \ A, B, C, D, \rho) > .$$
$$\ln Z' = \ \sum_{t=1}^{T} \ln \zeta_t{}'(y_t)$$
$$\ln \zeta_t{}'(y_t) = \ -\frac{1}{2}[< \ln |2\pi R| > \ - \ln |\Sigma_{t-1}^{-1} \Sigma_{t-1}^{*} \Sigma_t| + \mu_{t-1}^T \Sigma_{t-1}^{-1} \mu_{t-1} - \ \mu_t^T \Sigma_t^{-1} \mu_t$$
$$+ \ y_t^T < R^{-1} > y_t - 2 y_t^T < R^{-1} D > u_t + \ u_t^T < D^T R^{-1} D > u_t$$
$$- \ (\Sigma_{t-1}^{-1} \mu_{t-1} - < A^T B > u_t)^T \Sigma_{t-1}^{*} (\Sigma_{t-1}^{-1} \mu_{t-1} - < A^T B > u_t) \ .$$

$\alpha_t(x_t)$ is posterior over the hidden state at time $t$ given observed data up to time $t$ ($t = \{1, 2, \dots, T\}$). At the same time, filtered mean $\mu_t$ and covariance $\Sigma_t$ of hidden state $\alpha_t(x_t)$ are estimated:

$$\alpha_t(x_t) \ \equiv \ p(x_t | \ y_{1:t}) \ .$$

In addition, recursion of $\alpha_t(x_t)$ with previous $\alpha_{t-1}(x_{t-1})$ is formed as following form:

$$\alpha_t(x_t) \ = \ \int dx_{t-1} \ \frac{p(x_{t-1} | \ y_{1:t-1}) \ p(x_t | \ x_{t-1}) p(y_t | \ x_t)}{p(y_t | \ y_{1:t-1})}$$
$$= \frac{1}{\zeta_t(y_t)} \int dx_{t-1} \alpha_{t-1}(x_{t-1}) \ p \ (x_t | \ x_{t-1}) \ p(y_t | \ x_t)$$
$$= \frac{1}{\zeta_t(y_t)} \int dx_{t-1} N(x_{t-1} | \ \mu_{t-1}, \Sigma_{t-1}) \ N(x_t | \ A x_{t-1}, \mathrm{I}) \ N(y_t | \ C x_t, R)$$
$$= N(x_t | \ \mu_t, \Sigma_t)$$

Here, $\zeta_t(y_t) \equiv\ p(y_t|\ y_{1:t-1})$ is the filtered output probability.

$$\Sigma^*_{t-1}\ =\ (\Sigma^{-1}_{t-1} + A^T A)^{-1}$$
$$x^*_{t-1} = \Sigma^*_{t-1}[\Sigma^{-1}_{t-1}\mu_{t-1}\ +\ A^T x_t]$$
$$\Sigma_t\ = [\mathrm{I} + C^T R^{-1} C - A\Sigma^*_{t-1}A^T]^{-1}$$
$$\mu_t\ =\ \Sigma_t\ [C^T R^{-1} y_t\ +\ A\Sigma^*_{t-1}\Sigma^{-1}_{t-1}\mu_{t-1}]\ .$$

At each step the normalising constant $\zeta_t(y_t)$ obtained as the denominator of $\alpha_t(x_t)$, contributes to the calculation of the probability of the data. In this context, the probability of the data can be written in the following form:

$$p(y_{1:T})\ =\ p(y_1)p(y_2|\ y_1)\ \ldots\ p(y_t|\ y_{1:t-1})\ \ldots\ p(y_T|\ y_{1:T-1})\ =$$
$$p(y_1)\prod_{t=2}^{T} p(y_t|\ y_{1:t-1}) = \prod_{t=1}^{T}\zeta_t(y_t)$$
$$\zeta_t(y_t) =\ N(y_t|\ \varpi_t, \xi_t)$$
$$\xi_t\ =\ (R^{-1} - R^{-1}C\Sigma_t C^T R^{-1})^{-1}$$
$$\varpi_t\ = \xi_t R^{-1} C\Sigma_t A\Sigma^*_{t-1}\Sigma^{-1}_{t-1}\mu_{t-1}\ .$$

With these distributions, the probability of each observation $yt$ given the previous observations in the sequence is computed, and predictive mean and variance are assigned to the data at each time step as it arrives. However, this predictive distribution will change once the hidden state sequence has been smoothed. To obtain the posterior over the hidden state given all the data, the smoothed estimate $\gamma$ of the hidden state is determined. In particular, $\gamma_t(x_t) = p(x_t|y_{1:T})$, the smoothed estimate of the hidden state, is expressed as follows:

$$\gamma_t(x_t) = p(x_t|\ y_{1:T}) =\ \int dx_{t+1}\ p(x_t, x_{t+1}|\ y_{1:T})$$
$$= \int dx_{t+1}\ p(x_t|\ x_{t+1}, y_{1:T})p(x_{t+1}|\ y_{1:T})$$
$$= \int dx_{t+1}\ p(x_t|\ x_{t+1}, y_{1:t})p(x_{t+1}|\ y_{1:T})$$
$$= \int dx_{t+1}\ \frac{p(x_t|\ y_{1:t})p(x_{t+1}|\ x_t)}{\int dx'_t\ p(x'_t|\ y_{1:t})p(x_{t+1}|\ x'_t)}\ p(x_{t+1}|\ y_{1:T})$$
$$= \int dx_{t+1}\left[\frac{\alpha_t(x_t)\ p(x_{t+1}|x_t)}{\int dx'_t\ \alpha_t(x'_t)\ p(x_{t+1}|x'_t)}\right]\gamma_{t+1}(x_{t+1})$$
$$\gamma_t(x_t)\ =\ N(x_t|\ \omega_t, \Upsilon_{tt})$$
$$K_t\ =\ (\Upsilon^{-1}_{t+1,t+1}\ +\ A\Sigma^*_t A^T)^{-1}$$

$$\Upsilon_{tt} = [\Sigma_t^{*-1} - A^T K_t A]^{-1}$$
$$\omega_t = \Upsilon_{tt}\,[\Sigma_t^{-1}\mu_t + A^T K_t(\Upsilon_{t+1,t+1}^{-1}\omega_{t+1} - A\Sigma_t^{*}\Sigma_t^{-1}\mu_t)]\,.$$

In this case, once the data at time $t$, $y_t$ has been filtered, $\alpha_t(x_t)$ is used instead of hidden variables.

To estimate the states at different time steps, parallel recursion $\beta_t(x_t) = p(y_{t+1:T}|x_t)$ is calculated:

$$\beta_{t-1}(x_{t-1}) = \int dx_t\, p(x_t|\, x_{t-1})\, p(y_t|\, x_t) p(y_{t+1:T}|\, x_t) =$$
$$\int dx_t\, p(x_t|\, x_{t-1})\; p(y_t|\, x_t)\, \beta_t(x_t) \propto N(x_{t-1}|\, \eta_{t-1}, \Psi_{t-1})\,.$$

$\beta_T(x_T) = 1$ is defined in ${\Psi_T}^{-1} = 0$ form to satisfy the end condition. The recursion proceeds from $t = T$ to $t = 1$. At the last step of this recursion, the probability of all data on the auxiliary hidden state variable $x_0$ is determined:

$${\Psi_t}^{*} = (\mathrm{I} + C^T R^{-1} C + {\Psi_t}^{-1})^{-1}$$
$$\Psi_{t-1} = [A^T A - A^T\, {\Psi_t}^{*} A]^{-1}$$
$$\eta_{t-1} = \Psi_{t-1}\, A^T {\Psi_t}^{*} [C^T R^{-1} y_t + {\Psi_t}^{-1}\eta_t]\,.$$

Using the formulas presented above, $\{\eta_t, \Psi_t\}_{t=0}^{T}$ is computed over the interval from $t = 0$ to $t = T$.

In the Variational Bayesian scenario the marginals cannot be obtained easily with backward sequential pass and instead they are computed by combining $\alpha$ and $\beta$ messages described as follows. In particular, $\omega_t$ and $\Upsilon_{t,t}$ are computed for $t = \{0, 1, 2, \ldots, T-1\}$ in the following form:

$$p(x_t|\, y_{1:T}) \propto p(x_t|\, y_{1:t}) p(y_{t+1:T}|\, x_t) =$$
$$\alpha_t(x_t)\, \beta_t(x_t) = N(x_t|\, \omega_t, \Upsilon_{tt})$$
$$\Upsilon_{t,t} = [\Sigma_t^{-1} + {\Psi_t}^{-1}]^{-1}$$
$$\omega_t = \Upsilon_{t,t}\,[\Sigma_t^{-1}\mu_t + {\Psi_t}^{-1}\eta_t]\,.$$

For $t = \{0, 1, 2, \ldots, T-1\}$ taking into account all observations, the cross-covariance of the hidden variable vector, $\Upsilon_{t,t+1}$, is defined as follows:

$$\Upsilon_{t,t+1} = \Sigma_t^{*} A^T \Upsilon_{t+1,t+1}.$$

**Calculation of the lower bound on the log marginal probability**

Due to the complexity of directly computing the log marginal probability, its lower bound is evaluated and denoted by $F$ as the estimate of the lower bound of the log marginal probability. As this estimate of the lower bound of the log marginal probability is itself a complex function, it is derived through several steps and ultimately expressed as a simplified summ.

KL (Kullback-Leibler) is divergence between two normal or two gamma distributions over the same variables and computed using the following equations for a pair of variables $J$ and $K$:

$$KL(J) = \int dJ\, q(J)\, \ln\frac{q(J)}{p(J)}$$

$$KL(J|K) = \int dJ\, q(J|K)\, \ln\frac{q(J|K)}{p(J|K)}$$

$$< KL(J|K) >_{q(K)} = \int dK\, q(K)\; KL(J|K)$$

Where,
$KL(J)$ – $KL$ divergence,
$KL(J|K)$ – conditional $KL$,
$< KL(J|K) >_{q(K)}$ – expected conditional $KL$.

By applying these equations, $F$ can be redefined in the following form:

$$F = -KL(B) - < KL(A|\, B) >_{q(B)} - \; KL(\rho) - < KL(D|\, \rho) >_{q(\rho)} - < KL(C|\, \rho, D) >_{q(\rho,D)} + \; H(q_x(x_{0:T})) \; + < \ln p(x_{1:T}, y_{1:T}|\, A, B, C, D, \rho) >_{q(A,B,C,D,\rho)q(x_{1:T})} .$$

$$H\big(q_x(x_{0:T})\big) = -\int dx_{0:T}\; q_x(x_{0:T}) \ln q_x(x_{0:T}) =$$

$$-\int dx_{0:T}\; q_x(x_{0:T})[-\ln Z' + \; < \ln p(x_{0:T}, y_{1:T}|\, A, B, C, D, \rho, \mu_0, \Sigma_0) >_{q_\theta(A,B,C,D,\rho)}]$$

$$= \ln Z' \; - \; < \ln p(x_{0:T}, y_{1:T}|\, A, B, C, D, \rho, \mu_0, \Sigma_0) >_{q_\theta(A,B,C,D,\rho)q_x(x_{0:T})}$$

Using $H(q_x(x_{0:T}))$, the equation for $F$ is transformed into the following simplified form, which facilitates its computation:

$$F = -KL(B) - < KL(A|\, B) >_{q(B)} - KL(\rho) - < KL(D|\, \rho) >_{q(\rho)} - < KL(C|\, \rho, D) >_{q(\rho,D)} + ln\, Z'.$$

**Updating hyperparameters**

The hyperparameters $\alpha$, $\beta$, $\gamma$, $\delta$, a, b, and the prior parameters $\Sigma_0$ and $\mu_0$, are updated to maximise the lower bound on the marginal probability. In this context, hyperparameters is updated as follows:

$$\alpha_j^{-1} \leftarrow \frac{1}{k}\big[k\Sigma_A + \Sigma_A[S_A {S_A}^T - 2G_A < B >^T {S_A}^T + G_A\{k\Sigma_B + < B >^T < B >\}{G_A}^T]\,\Sigma_A\big]_{jj}$$

$$\beta_j^{-1} \leftarrow \frac{1}{k}[k\Sigma_B + < B >^T < B >]_{jj}$$

$$\gamma_j^{-1} \leftarrow \frac{1}{p}[p\Sigma_C + \Sigma_C[S_C diag(\bar{\rho}) {S_C}^T - 2S_C diag(\bar{\rho}) < D > {G_C}^T + pG_C\Sigma_D {G_C}' + G_C < D >^T diag(\bar{\rho}) < D > {G_C}^T]\,\Sigma_C]_{jj}$$

$$\delta_j^{-1} \leftarrow \frac{1}{p}[p\Sigma_D + < D >^T diag(\bar{\rho}) < D >]_{jj}$$

Here, $[\cdot]_{jj}$ denotes $(j, j)$th element of matrix $[\cdot]$.

In order to maximise the probability of the hidden state sequence under the prior, the hyperparameters of the prior over the auxiliary hidden state are set according to the distribution of the smoothed estimate of $x_0$:

$$\Sigma_0 \leftarrow \Upsilon_{0,0}$$
$$\mu_0 \leftarrow \omega_0$$

The hyperparameters $a$ and $b$ governing the prior distribution over the output noise, $R = diag(\rho)$, are set to the fixed point of the equations:

$$\Psi(a) = \ln b + \frac{1}{p}\sum_{s=1}^{p} \overline{ln\ \rho_s}$$
$$\frac{1}{b} = \frac{1}{pa}\sum_{s=1}^{p} \overline{\rho_s}$$

**Comparison of the lower bound estimates on the log marginal probability**

As the estimate of the lower bound of the log marginal probability is determined, in the final step, this estimate is compared with its value from the previous iteration. If the lower bound has increased relative to its value in the preceding iteration, the computations are repeated starting from the Variational Bayesian M-step. This iterative process continues until the lower bound no longer increases. When the estimate of the lower bound on the log marginal probability remains unchanged from its previous value, it is considered to have reached its optimal value. At this stage, the parameters obtained through the hyperparameters also achieve their optimal values. Based on the identified parameters and the hidden variables, the observable endogenous variable $Y$ is estimated.