ON MODEL SELECTION FOR STATE ESTIMATION FOR NONLINEAR SYSTEMS

Robert Bos [*,1] **Xavier Bombois** [*]
Paul M. J. Van den Hof [*]

[*] *Delft Center for Systems and Control, Delft University of Technology, Mekelweg 2, 2628 CD Delft, The Netherlands, email: r.bos@dcsc.tudelft.nl*

Abstract: State filters can be used to produce online estimates of the state of a process. If an exact model for the true process is not known, but multiple candidate models are available to describe the current behavior of the true system, it is necessary to select that model that leads to the optimal state estimates. This paper describes a novel approach for model selection for state estimation by comparing the expected weighted prediction error using estimated states of different candidate models. The expected prediction error can not be computed exactly, but can be estimated using a newly derived generalized version of the FPE selection criterion. A simulation example of a time varying system is used to illustrate the performance of the selection method.

Keywords: State estimation, FPE, model selection, fault detection

1. INTRODUCTION

The goal of monitoring can be to provide an operator with accurate online information on critical process variables. In this paper these physical variables will be referred to as the state of the system. Since the complete state of a complex process often cannot be measured directly, it has to be estimated using measurements related to the state. To obtain these estimates a model of the system is required that describes the dynamic behavior of the state and relates the states to the available measurements.

Assume that the model can be written in the following state-space form:

$$x_{k+1} = f(x_k, u_k, \theta) + w_k \quad (1)$$

$$y_k = h(x_k, \theta) + v_k, \quad (2)$$

in which x_k is the system state at time index k, u_k are the known inputs to the system, y_k are the measured system outputs, θ is a known parameter vector and w_k and v_k are zero mean independent gaussian noises that operate on the states and measurements respectively. In process industry,$f(\cdot)$ and $h(\cdot)$ will generally be complex nonlinear functions. If a model in this form is given, it is usually possible to construct a state filter, that produces an online estimate of a true state x_k using all measurements $y_0, \ldots, y_k$. We will denote this estimate as $\hat{x}_{k|k}$. Examples of such filters are the Extended Kalman Filter (EKF) (Anderson and Moore, 1979), Unscented Kalman Filter (UKF) (Julier *et al.*, 2000), Moving Horizon estimators (MHE) (Robertson and Lee, 1995) and Particle Filters (Doucet *et al.*, 2001). Since a state filter estimates states based on measurements y_k and the model (1)-(2), the accuracy of the estimated states not only depends on the accuracy of the measurements, but also on the quality of the

[1] The work of Robert Bos is supported by Dutch research organization TNO-TPD.

model. If, for instance, the model is incorrect, the state estimates could be severely biased.

In some situations multiple candidate models are available for describing the current behavior of the true system. An example of such a situation is in model based fault detection and isolation (FDI). In model based FDI, at least one model is given to describe the nominal process behavior, while a second set of models describe the behavior of the process after corresponding faults have occurred. These fault models could be the same as the nominal models, but with time varying parameters. In this paper we will focus on the problem of selecting the best possible model for state estimation.

Model selection for filtering problems is often described using Bayes conditional probability theory (Gustafsson, 2000). Suppose n models of the form (1)-(2) are available, denote these models as $\mathcal{M}_i$, with $i = 1, 2, \ldots, n$. Then the filtering procedure is carried out for each model, on data y. Afterwards, using Bayes conditional probability theory, the conditional probability $p(\mathcal{M}_i|y)$ is computed. The model with the highest conditional probability is then selected, and the state estimates based on this model are used. The conditional probability $p(\mathcal{M}_i|y)$ can be computed via: $p(\mathcal{M}_i|y) = p(y|\mathcal{M}_i)p(\mathcal{M}_i)/p(y)$. Using this equation in practice for complex process models is generally difficult, because the term $p(y|\mathcal{M}_i)$ is not trivial to compute for non-linear systems and knowledge of the a-priori probability of each model $p(\mathcal{M}_i)$ is rarely available.

An alternative approach to the model selection problem is given in (Tyler *et al.*, 2000). If a moving horizon state estimator (MHE) is used for state estimation, the state estimation problem is written as a weighted and regularized least squares problem. The problem of model selection is therefore approximately similar to model selection in system identification theory. Given this similarity, the model selection is done using the Akaike Information Criterion (AIC). Advantages of this approach are that exact probability distributions are no longer required, and the technique can also be easily adapted for non-linear models. Drawbacks of this approach are that the technique can only be used in conjunction with moving horizon estimators. Another drawback is that the AIC criterion may not be the best criterion, since it was derived only for least squares problems without weighting and regularization, while states are estimated with weighting and regularization.

In this paper a model selection procedure closely related to (Tyler *et al.*, 2000) will be considered. Instead of using AIC, a specialized criterion for weighted and regularized least squares problems is derived. Secondly we will show that the model selection can also be used with estimators different from the MHE. Finally a simulation example is presented in which model selection is used to accurately estimate the state of a time varying system.

2. MODEL SELECTION FOR STATE ESTIMATION

As mentioned in the introduction, our objective is to estimate the state vector of a true process using the output vector y_k, the input signal u_k and a model of the true process. Assume that this model of the true system is given by the linear state space equations:

$$x_{k+1} = Ax_k + Bu_k + w_k \tag{3}$$

$$y_k = Cx_k + v_k. \tag{4}$$

where w_k and v_k are gaussian noises with

$$\mathbb{E}\begin{bmatrix} w_k \\ v_k \end{bmatrix} = 0 \quad \mathbb{E}\begin{bmatrix} w_k \\ v_k \end{bmatrix} [w_l^T \;\; v_l^T] = \begin{bmatrix} Q & 0 \\ 0 & R \end{bmatrix} \delta_{k,l} \tag{5}$$

If the model (3)-(4) perfectly describes the true system, an optimal estimate of the state vector x_k at each time k can be obtained using a Kalman filtering procedure implemented using (3)-(4). Such a filtering procedure consists of two steps. In the prediction step, a prediction $\hat{x}_{k+1|k}$ of the state vector is given, along with its covariance matrix $\hat{P}_{k+1|k}$. In the measurement update, the measurement y_{k+1} is used to refine the prediction into estimate $\hat{x}_{k+1|k+1}$ and to produce its covariance matrix $\hat{P}_{k+1|k+1}$. An important result for the sequel is that the measurement update in a Kalman filtering procedure can be seen as the solution of a Weighted and Regularized Least Squares (WRLS) problem (Kailath *et al.*, 1999):

$$\hat{x}_{k|k} = \arg\min_x \|y_k - Cx\|^2_{R^{-1}} + \|x - \hat{x}_{k|k-1}\|^2_{\hat{P}^{-1}_{k|k-1}}, \tag{6}$$

with $\|z\|^2_W = z^T W z$ for $z \in \mathbb{R}^n$.

If the model of the true system that is assumed to perfectly describe the true system is non-linear, such as e.g. in (1)-(2), an estimate of the state vector x_k can again be obtained using very similar procedures (e.g. using the EKF or UKF instead of the normal Kalman Filter) which also consist of two steps: a prediction step and a measurement update step. Moreover, it can be proven that, if the output function (2) can be well approximated by a linear function around $\hat{x}_{k|k-1}$, the measurement update in this procedure is still approximately a WRLS problem:

$$\hat{x}_{k|k} = \arg\min_x \|y_k - \hat{y}(x)\|^2_{R^{-1}} + \|x - \hat{x}_{k|k-1}\|^2_{\hat{P}^{-1}_{k|k-1}}, \quad (7)$$

where $\hat{y}(x)$ denotes the predictor of the output vector y_k using the state vector x and the available model (1)-(2) i.e. $\hat{y}(x) = h(x)$. By comparing (6) and (7), we see that (7) is also valid in the linear case with $\hat{y}(x) = Cx$.

The state estimation procedure delivers an estimate $\hat{x}_k = \hat{x}_{k|k}$ for x_k using the assumption that the available model perfectly describes the true system. In practice, the model is only an approximation of the true system and the quality of the estimate of x_k will depend on the quality of the chosen model. Consequently, the model will have to be chosen in such a way that the state estimation procedure based on this model delivers a good estimate of x_k. In this paper, we will consider the particular situation where we have several candidate-models and we have to select, among these models, the model which will deliver the best estimate of the state vector. To make this selection, only known inputs and measured outputs are available.

Our selection procedure will be based on a measure of the quality of the model for the estimation of the state vector. Different measures could be considered for this purpose. In this paper, we will define a measure of quality which is very similar to the quality measures used in system identification. Given a model of the type (1)-(2) and given the estimate $\hat{x}_k$ of the state vector at time k obtained using the WRLS problem (7), the measure of quality $\mathcal{V}(k)$ at time k is defined as follows:

$$\mathcal{V}(k) = \mathbb{E}_{\hat{x}_k} \overline{V}(\hat{x}_k, k). \quad (8)$$

$$\text{with } \overline{V}(x,k) = \mathbb{E}_{y_k} \|y_k - \hat{y}(x)\|^2_{R^{-1}}. \quad (9)$$

Using (9), we see that $\overline{V}(\hat{x}_k, k)$ represents the ability of the model and the estimate $\hat{x}_k$ to predict not only the particular realization of the output vector y_k that we used to estimate $\hat{x}_k$, but also all other possible realizations of y_k. In (9) the prediction error $y_k - \hat{y}(x)$ is weighted with R^{-1}, to take into account the variability of the measurements y_k. The quantity $\overline{V}(\hat{x}_k, k)$ is still a random variable since $\hat{x}_k$ is determined using noisy data. Therefore, it is safer to consider its mean as measure of quality for the model such as we have done in (8). From the definition of $\mathcal{V}(k)$, we see that the smaller the time function $\mathcal{V}(k)$ is, the better is the model.

We have thus defined a measure $\mathcal{V}(k)$ of the quality of a model. This quantity can only be used in a quality assessment procedure if it is possible to compute (or to approximate) $\mathcal{V}(k)$ using the available data. In order to find a method for computing (or approximating) $\mathcal{V}(k)$, we first notice the strong analogy between $\mathcal{V}(k)$ and the Final Prediction Error (FPE) that is used to assess the quality of a model in system identification theory (see (Ljung, 1999)). We can indeed see the FPE as a special case of $\mathcal{V}(k)$ where the matrix R in (9) has been replaced by the identity matrix and where $\hat{x}_k$ has been estimated using the criterion (7) with $R = I$ and with $\hat{P}^{-1}_{k+1|k} = 0$. In (Ljung, 1999) an approximation for FPE is also given for the case where $\hat{x}_k$ has been estimated using the problem (7) with $R = I$ and with $\hat{P}^{-1}_{k+1|k} = \delta I$ with δ a positive real constant. However, in practice $\hat{P}^{-1}_{k+1|k}$ cannot be written as δI, which implies that the approximations in (Ljung, 1999) can not be used directly in order to approximate $\mathcal{V}(k)$ if $\hat{x}_k$ has been estimated using the general WRLS problem (7). We have therefore extended the results of (Ljung, 1999) in order to be able to approximate $\mathcal{V}(k)$.

Proposition 1. Let us consider the time instant k and the output vector y_k collected from the true system at that instant. Let us also consider the measure of quality $\mathcal{V}(k)$ defined in (8)-(9). Furthermore assume that the estimate $\hat{x}_k$ of the state vector in (8)-(9) is obtained via the following weighted and regularized least squares problem which is equivalent to (7):

$$\hat{x}_k = \arg\min_x \left(V(x,k) + (x - x_k^{\#})^T P^{-1} (x - x_k^{\#})\right) \quad (10)$$

in which $x_k^{\#}$ is a given state vector, P^{-1} is a positive semi-definite regularization matrix, and $V(x,k)$ is a weighted least squares criterium:

$$V(x,k) = \|y_k - \hat{y}(x)\|^2_{R^{-1}}. \quad (11)$$

Then, if we can further assume that $x_k^{\#} \approx x_k^* = \arg\min_x \overline{V}(x,k)$, that $y_k - \hat{y}(x_k^*)$ is approximately a white noise and that the dimension of the vector y_k is sufficiently large, we have that

$$\mathcal{V}(k) \approx V(\hat{x}_k, k) + 4\text{tr}\left\{[(\psi^T R^{-1} L R^{-T} \psi)] \times \left[\overline{V}''(x_k^*, k) + 2P^{-1}\right]^{-1}\right\}. \quad (12)$$

where $L = \mathbb{E}(ee^T)$ with $e = y_k - \hat{y}(x_k^*)$ and $\psi = \frac{\partial}{\partial x}\hat{y}(x)$ evaluated at $x = x_k^*$.

Proof. See Appendix A. □

As mentioned in the statement of Proposition 1, the derivation of the approximation (12) requires some additional assumptions. These assumptions are also present in the less general version of (Ljung, 1999) and in the same reference the consequences of these assumptions are discussed.

The estimated asymptotic fit $\mathcal{V}(k)$ in (12) is the sum of the achieved fit $V(\hat{x}_k, k)$ on the measurement data and a complex term. This term contains the second derivative of $\overline{V}(x_k^*, k)$. For general models this quantity can be approximated by the second derivative of $V(\hat{x}_k, k)$. Similarly, ψ can be approximated using the derivative of $\hat{y}(x)$ evaluated in $\hat{x}_k$ instead of x_k^*. If the output model is linear, for instance as is the case in (4), such approximations are not necessary, because in this case it is easy to show that $\overline{V}''(x_k^*, k) = 2CR^{-1}C^T$ and $\psi = C$. L represents the covariance of the minimal asymptotic prediction error. For additive measurement noise, $L = R$.

Let us now summarize and define our model selection procedure. We wanted to select, among a set of candidate-models, the model which delivers the best estimate of the state vector. We have defined for this purpose a measure $\mathcal{V}(k)$ of the quality of a model. This measure is a time function and can be approximated using (12). In order to make the selection, let us compute, with each model, the estimate of the state vector at time instants $k = 1...M$. Let us then approximate $\mathcal{V}_i(k)$ for each model and for $k = 1...M$ ($\mathcal{V}_i(k)$ denotes the quality measure of the i^{th} model). Then, let us determine for each model i, the average quality measure over $k = 1...M$ i.e. $\mathcal{V}_{av,i} = (1/M)\sum_{k=1}^{M} \mathcal{V}_i(k)$. The best model is then defined as the model for which $\mathcal{V}_{av,i}$ is minimal.

Using the average $\mathcal{V}_{av,i}$ of the quality measure over $k = 1...M$ as selection criterion has two advantages: we replace a time function $\mathcal{V}_i(k)$ by a single number and we generally decrease the influence of the approximation errors in $\mathcal{V}_i(k)$.

3. MODEL SELECTION AND TIME VARYING SYSTEMS

State estimation for time varying systems can be improved by treating the estimation problem as a model selection problem which can be solved using the model selection procedure of section 2. This particular estimation setup is explained in the sequel of this paper.

Time varying systems are very common in the industrial world: the true behavior of a complex process will generally change over time. Changes can occur slowly, for example due to wear, but can also occur abruptly, e.g. due to a sudden fault in the system. If the system behavior changes, but the model used for state estimation remains the same, state estimates will most likely become biased.

If we assume that the dynamics of the true system can be described by (1)-(2) and that any change in the dynamics of this true system can be modelled as changes in the model parameters θ, then an extra equation can be added to the model, such that the model parameters are allowed to change:

$$\theta_{k+1} = f_\theta(x_k, \theta_k, u_k) + w_k^\theta, \tag{13}$$

where w_k^θ is a independent gaussian noise. If we define a new vector $\tilde{x} = [x_k^T \ \theta_k^T]^T$, it is always possible to combine (1)-(2) with (13) into a new model:

$$\tilde{x}_{k+1} = \tilde{f}(\tilde{x}_k, u_k) + \tilde{w}_k \tag{14}$$

$$y_k = \tilde{h}(\tilde{x}_k, u_k) + v_k, \tag{15}$$

with $\tilde{w}_k = [w_k^T \ w_k^{\theta T}]^T$. Since the form of (14)-(15) is the same as that of the original model (1)-(2), a state filter can usually be designed to simultaneously estimate states x_k and to update parameters θ_k. The price for simultaneously estimating states and parameters is that the variance of the estimated states will increase, even if the true model parameters are constant. If we assume that parameters only change very slowly over time, this may be wasteful, because the estimation bias caused by slightly incorrect parameters may be less than the extra variance that is caused by simultaneously estimating states and parameters.

Let us summarize. We cannot use a model with a fixed parameter vector to estimate the state vector of time varying systems, because the estimate will become biased if the true parameter vector has changed. An estimator based on a flexible parameter model such as in (14)-(15) is also a suboptimal solution in this case since the estimate will have a large variance in the periods where the true parameter vector is constant.

Based on this, it is obvious that a better estimate of the state vector of a time varying true system would be obtained if the fixed parameter model (with the correct parameter values) is used at times where the true parameter remains constant and if the flexible parameter model is used when the true parameter vector is actually changing. In order to decide at a given time instant k, whether we have to use the fixed or the flexible parameter vector model, we apply our model selection procedure of section 2. The details of this model selection procedure are illustrated in an example in the next section.

4. SIMULATION EXAMPLE

In the simulation example we will use model selection to estimate states of a time varying system. Consider the following simple system:

$$x_{k+1} = \begin{bmatrix} \theta_k & 0.7 \\ 0 & 0.9 \end{bmatrix} x_k + \begin{bmatrix} 1 \\ 1.5 \end{bmatrix} u_k + w_k \tag{16}$$

$$y_k = [1 \ 1] x_k + v_k. \tag{17}$$

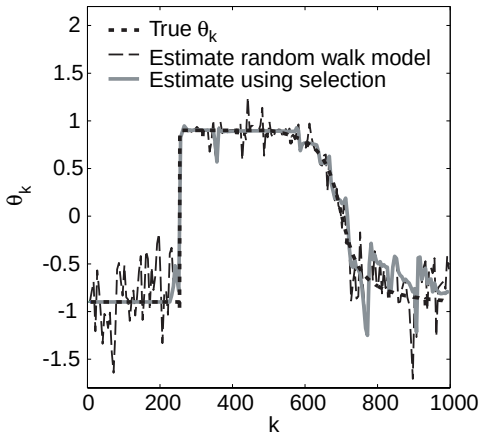

Fig. 1. True and estimated parameter θ_k in the simulation example.

The input signal u_k was chosen as a random binary signal with a switching probability of 0.05, the noises w_k and v_k were chosen as gaussian with:

$$\mathbb{E}\begin{bmatrix} w_k \\ v_k \end{bmatrix} = 0 \quad \mathbb{E}\begin{bmatrix} w_k \\ v_k \end{bmatrix}[w_k^T v_k^T] = \begin{bmatrix} Q & 0 \\ 0 & R \end{bmatrix} \quad (18)$$

$$Q = I \quad R = 10. \quad (19)$$

If the parameter θ_k is constant, these system equations describe a linear system for which the standard Kalman filter is the optimal state estimator. In this example however the parameter θ_k is chosen to be time varying. The parameter θ_k jumps from -0.90 to 0.90 at $k = 255$ and starting from $k = 500$, the parameter gradually returns to its original value, see Figure 1. This way, both a sudden jump (fault) and slowly changing behavior are modelled.

Using this system $N = 1000$ measurements y_k are generated. The true states x_k and parameters θ_k are also stored. Afterwards, using only the recorded measurements and inputs, the state of the system is estimated using the procedure based upon model selection as presented in the previous section. The estimation error is expressed in the mean squared error: $\frac{1}{N}\sum_{k=0}^{N-1} \|x_k - \hat{x}_{k|k}\|^2$.

For the model selection procedure two candidate models of the system are available. The first model is equal to the true model (16)-(17), but uses constant parameters, i.e.:

$$\theta_{k+1} = \theta_k. \quad (20)$$

The initial value for the constant parameter is -0.90, which is equal to the true starting value of θ_k. As proposed in section 3, the second candidate model is a flexible parameter model, i.e. the model (16)-(17) augmented by:

$$\theta_{k+1} = \theta_k + w_k^\theta \quad (21)$$

with w_k^θ chosen as a gaussian white noise with variance 0.1.

Table 1. Mean squares errors of estimated states for a time varying system. States were estimated using a fixed parameter model, a random walk parameter model, and using model selection. Results are averaged over 200 simulation runs.

Fixed parameter	Flexible par. filter	Selection filter
$1.1 \cdot 10^4$	15.1	11.8

In this example, the optimal model for state estimation will be selected using the model selection procedure of section 2. Two UKF filters are designed using the given models. At time instant $k = k_0$, both filters are initialized with the same estimate $\tilde{x}_{k_0|k_0}$ and $\tilde{P}_{k_0|k_0}$. Then both filters are run for $M = 15$ time steps and $\mathcal{V}_{av,i}$ of each filter model is computed. The results of the model with the lowest $\mathcal{V}_{av,i}$ is then chosen as the final selected result. At time $k_0 + M$ all filters are again initialized with the selected result and all filters are again run for M time steps, etcetera.

For illustration purposes, we have also estimated the state vector using only the fixed parameter model and using only the flexible parameter model. The results of all the simulations are given in Table 1. As expected, the fixed parameter model alone produces very poor results, due to the bias caused by using an inaccurate model. The results obtained using the flexible parameter model alone are much better. The extra degree of freedom in this model allows the filter to correct the parameter in the measurement update of the filtering procedure. The results using both models and our new model selection procedure are, as expected, much better than the results using the fixed model alone and better than the results of using the flexible parameter model alone. By looking at Figure 1, we indeed observe that the estimate using our new procedure is made using the fixed parameter model in the time periods where the true parameter remains constant, so that there is no extra variance in the estimated state.

5. CONCLUSIONS

This paper discusses a new algorithm for model selection for state estimation problems. The presented algorithm uses the expected weighted prediction error of estimated states, denoted as $\mathcal{V}(k)$, as a selection criterion. The expected weighted prediction error cannot be estimated using the standard FPE criterion, since parameters are estimated using a weighted and regularized least squares problem. To overcome this problem a new estimator for $\mathcal{V}(k)$ has been derived, that is also valid if states are estimated using weighting and

regularization. Using $\mathcal{V}(k)$ as the selection criterion has the advantage that the exact conditional probability of a model structure given the data does not need to be computed. The new algorithm was also demonstrated in a time varying system simulation example.

REFERENCES

Anderson, B. D. O. and J. B. Moore (1979). *Optimal Filtering*. Prentice-Hall. Englewood Cliffs, N.J.

Doucet, A., N. De Freitas and N. Gordon (2001). *Sequential Monte Carlo Methods in Practice*. Springer. New York.

Gustafsson, F. (2000). *Adaptive Filtering and Change Detection*. Wiley & Sons. Chichester, United Kingdom.

Julier, S., J. Uhlmann and H.F. Durrant-Whyte (2000). A new method for the non-linear transformation of means and covariances in filters and estimators. *IEEE Trans. Autom. Control* **45**(3), 477–482.

Kailath, T., A.H. Sayed and B. Hassibi (1999). *Linear Estimation*. Prentice Hall. Upper Saddle River.

Ljung, L. (1999). *System Identification-Theory for the User*. second ed.. Prentice Hall. Upper Saddle River.

Robertson, D. G. and J. H. Lee (1995). A least squares formulation for state estimation. *J. Process Control* **5**(4), 291–299.

Tyler, M. L., K. Asano and M. Morari (2000). Application of moving horizon estimation based fault detection to cold tandem steel mill. *Int. J. Control* **73**(5), 427–438.

Appendix A. PROOF OF PROPOSITION 1

Define $W(x,k)$ as:

$$W(x,k) \;=\; V(x,k) \;+\; \|x \;-\; x_k^{\#}\|^2_{P^{-1}}, \quad \text{(A.1)}$$

then

$$\hat{x}_k = \arg\min_x W(x,k). \quad \text{(A.2)}$$

Define also $\overline{W}(x,k)$ as:

$$\overline{W}(x,k) \;=\; \overline{V}(x,k) \;+\; \|x \;-\; x_k^{\#}\|^2_{P^{-1}}. \quad \text{(A.3)}$$

In order to prove the proposition, let us first expand $\overline{V}(\hat{x}_k,k)$ around $x_k^* = \arg\min_x \overline{V}(x,k)$:

$$\overline{V}(\hat{x}_k,k) = \overline{V}(x_k^*,k) + \frac{1}{2}(\hat{x}_k - x_k^*)^T \overline{V}''(\zeta_k,k)(\hat{x}_k - x_k^*) \quad \text{(A.4)}$$

In the equation above, a $'$ means taking the partial derivative with respect to x. Similarly since $W'(\hat{x}_k,k) = 0$:

$$W(x_k^*,k) = W(\hat{x}_k,k) + \frac{1}{2}(\hat{x}_k - x_k^*)W''(\overline{\zeta}_k,k)(\hat{x}_k - x_k^*), \quad \text{(A.5)}$$

which is easily rewritten into:

$$W(\hat{x}_k,k) = W(x_k^*,k) - \frac{1}{2}(\hat{x}_k - x_k^*)W''(\overline{\zeta}_k,k)(\hat{x}_k - x_k^*). \quad \text{(A.6)}$$

The second derivative of $W(x,k)$ is:

$$W''(x,k) = V''(x,k) + 2P^{-1}. \quad \text{(A.7)}$$

Inserting (A.3) and (A.7) into (A.6) and using the assumption that $x_k^{\#} = x_k^*$, we obtain:

$$V(\hat{x}_k,k) = V(x_k^*,k) - \frac{1}{2}(\hat{x}_k - x_k^*)(V''(\overline{\zeta}_k,k) + 4P^{-1})(\hat{x}_k - x_k^*) \quad \text{(A.8)}$$

Take the expectations of (A.4) and (A.8) and use the following asymptotical relations:

$$\mathbb{E}_{\hat{x}_k}(\hat{x}_k - x_k^*)(\overline{V}''(\zeta_k,k))(\hat{x}_k - x_k^*) = \mathbb{E}_{\hat{x}_k}\ \mathrm{tr}\left\{\overline{V}''(\zeta_k,k)(\hat{x}_k - x_k^*)(\hat{x}_k - x_k^*)^T\right\} \approx \mathrm{tr}\ \overline{V}''(x_k^*,k)P_x \quad \text{(A.9)}$$

in which P_x is the asymptotic covariance matrix of $\hat{x}_k$. Also,

$$\mathbb{E}_{\hat{x}_k}(\hat{x}_k - x_k^*)(V''(\overline{\zeta}_k,k))(\hat{x}_k - x_k^*) \approx \mathrm{tr}\overline{V}''(x_k^*,k)P_x \quad \text{and} \quad V(x^*,k) \approx \overline{V}(x^*,k).$$

Using the last three relations together with (A.4) and (A.8) gives:

$$\mathbb{E}_{\hat{x}_k}\overline{V}(\hat{x}_k,k) \approx \overline{V}(x_k^*,k) + \frac{1}{2}\mathrm{tr}\overline{V}''(x_k^*,k)P_x$$

$$\mathbb{E}_{\hat{x}_k}V(\hat{x}_k,k) \approx \overline{V}(x_k^*,k) - \frac{1}{2}\mathrm{tr}(\overline{V}''(x_k^*,k) + 4P^{-1})P_x$$

Combining the two last expressions yields:

$$\mathbb{E}_{\hat{x}_k}\overline{V}(\hat{x}_k,k) \approx \mathbb{E}_{\hat{x}_k}V(\hat{x}_k,k) + \mathrm{tr}(\overline{V}''(x_k^*,k) + 2P^{-1})P_x \quad \text{(A.10)}$$

Using the theory of chapter 9, pages 281-282 in (Ljung, 1999), it can be shown that the covariance matrix P_x equals:

$$P_x = 4\left[\overline{W}''(x_k^*,k)\right]^{-1}\left[\psi^T RLR^T\psi\right]\left[\overline{W}(x_k^*,k)\right]^{-1} = 4\left[\overline{V}''(x_k^*,k) + 2P^{-1}\right]^{-1}\left[\psi^T RLR^T\psi\right] \times \left[\overline{V}''(x_k^*,k) + 2P^{-1}\right]^{-1} \quad \text{(A.11)}$$

with ψ and L as defined in the statement of the theorem. To derive this last expression, we have used the assumption that $y_k - \hat{y}(x_k^*)$ is approximately a white noise and that $\overline{W}''(x_k^*,k)$ exists and is regular.

If we take result (A.10) (after replacing $\mathbb{E}_{\hat{x}_k}V(\hat{x}_k)$ with $V(\hat{x}_K)$, the only observation we have of it) and combine it with (A.11), we obtain the approximation (12). □