0.0.1 Encoding

Suppose the encoder observes data z_t from the RTU at time t. Assume the encoder has developed and estimated a state-space model of the form

$$x_t = Fx_{t-1} + w_t \tag{1a}$$
$$z_t = Hx_t + v_t, \tag{1b}$$

where $w_t \sim N(0, Q)$, $v_t \sim N(0, R)$, and F, Q, H, and R are parameters of the model. States of the Kalman filter are estimated via the following equations:

$$\hat{x}_{t|t-1} = F\hat{x}_{t-1|t-1} \tag{2a}$$
$$P_{t|t-1} = FP_{t-1|t-1}F' + Q \tag{2b}$$
$$e_t = z_t - H\hat{x}_{t|t-1} \tag{2c}$$
$$S_t = HP_{t|t-1}H' + R \tag{2d}$$
$$K_t = P_{t|t-1}H'S_t^{-1} \tag{2e}$$
$$\hat{x}_{t|t} = \hat{x}_{t|t-1} + K_t e_t \tag{2f}$$
$$P_{t|t} = (I - K_t H)P_{t|t-1} \tag{2g}$$
$$e_{t|t} = z_t - H\hat{x}_{t|t}. \tag{2h}$$

Suppose that the system has been observed long enough that asymptotic results can be used. Specifically, Kumar and Varaiya (1986) show that

$$\lim_{t\to\infty} P_{t|t} = \Sigma := (I - KH)S \tag{3a}$$
$$\lim_{t\to\infty} K_t = K := SH'(HSH' + R)^{-1} \tag{3b}$$
$$\lim_{t\to\infty} P_{t|t-1} = S := F\Sigma F' + Q. \tag{3c}$$

Then, state and observation estimates are formed with

$$\hat{x}_t = F\hat{x}_{t-1} + K(z_t - HF\hat{x}_{t-1}) \tag{4a}$$
$$\hat{z}_t = H\hat{x}_t, \tag{4b}$$

where variables with hats represent estimated values. Both of these estimators are unbiased with distributions

$$\hat{x}_t \sim N(x_t, \Sigma) \tag{5a}$$
$$\hat{z}_t = H\hat{x}_t \sim N(Hx_t, H\Sigma H'). \tag{5b}$$

The message is encoded in noise u_t (assumed to be iid), so that the encoder sends y_t, defined as

$$y_t = \hat{z}_t + u_t, \tag{6}$$

where $u_t \sim N(0, T)$. Note that anomaly detectors are expecting to see z_t be sent, which has distribution $z_t \sim N(Hx_t, R)$. Also note that y_t has the distribution

$$y_t \sim N(H\hat{x}_t, T) \equiv N(Hx_t, H\Sigma H' + T). \tag{7}$$

By choosing $T = R - H\Sigma H'$, the encoder ensures that y_t has the same distribution as z_t, which implies the steganography is undetectable. However, the encoder could choose a different T to reduce error rates in the decoding step, at the cost of being more detectable.

0.0.2 Decoding

The decoder receives y_t and attempts to extract noise and infer the message accordingly. The state-space model the decoder uses is already known from the setup of the encoder. Specifically, using Equation (4a), states evolve according to the relationship

$$\begin{aligned}
\hat{x}_t &= F\hat{x}_{t-1} + K(z_t - HF\hat{x}_{t-1}) && \text{(8a)}\\
&= (F - KHF)\hat{x}_{t-1} + Kz_t && \text{(8b)}\\
&= (F - KHF)\hat{x}_{t-1} + K(Hx_t + v_t) && \text{(8c)}\\
&= (F - KHF)\hat{x}_{t-1} + K(HFx_{t-1} + Hw_t + v_t). && \text{(8d)}
\end{aligned}$$

Thus, the evolution of states can be described by

$$\begin{bmatrix} \hat{x}_t \\ x_t \end{bmatrix} = \begin{bmatrix} (F - KHF)\hat{x}_{t-1} + KHFx_{t-1} + KHw_t + Kv_t \\ Fx_{t-1} + w_t \end{bmatrix}. \tag{9}$$

Define $\chi_t = (\hat{x}_t, x_t)'$, $\hat{B} = (I, 0)$, and $B = (0, I)$, where I is the identity matrix and 0 is the square matrix of zeros, each with number of rows and columns equal to the number of states. Notice that $\hat{x}_t = \hat{B}\chi_t$, $x_t = B\chi_t$, and $\hat{B}\hat{B}' = BB' = I$. Then, the state evolution can be written as

$$\begin{aligned}
\chi_t &= \hat{B}'((F - KHF)\hat{B}\chi_{t-1} + KHFB\chi_{t-1} + KHw_t + Kv_t) + B'(FB\chi_{t-1} + w_t) && \text{(10a)}\\
&= (\hat{B}'(F\hat{B} - KHF\hat{B} + KHFB) + B'FB)\chi_{t-1} + \hat{B}'(KHw_t + Kv_t) + B'w_t && \text{(10b)}\\
&= A\chi_{t-1} + c_t, && \text{(10c)}
\end{aligned}$$

where $A = \hat{B}'(F\hat{B} - KHF\hat{B} + KHFB) + B'FB$ and $c_t = \hat{B}'(KHw_t + Kv_t) + B'w_t$. Notice that since w_t and v_t are iid normal, c_t, which is a linear combination of w_t and v_t, is also iid normal. Specifically, the distribution of c_t is given by $c_t \sim N(0, C)$, where

$$C = \hat{B}'K(HQH' + R)K'\hat{B} + B'QB. \tag{11}$$

Further, since $y_t = \hat{z}_t + u_t = H\hat{x}_t + u_t = H\hat{B}\chi_t + u_t$, the state-space model can be written as

$$\begin{aligned}
\chi_t &= A\chi_{t-1} + c_t && \text{(12a)}\\
y_t &= H\hat{B}\chi_t + u_t, && \text{(12b)}
\end{aligned}$$

where $u_t \sim N(0, T)$.

Now, similar to the encoding Kalman, filter, the optimal state and observation estimates are given by

$$\begin{aligned}
\hat{\chi}_t &= A\hat{\chi}_{t-1} + \tilde{K}(y_t - H\hat{B}A\hat{\chi}_{t-1}) && \text{(13a)}\\
y_t &= H\hat{B}\hat{\chi}_t. && \text{(13b)}
\end{aligned}$$

The distributions of these estimates are

$$\begin{aligned}
\hat{\chi}_t &\sim N(\chi_t, \tilde{\Sigma}) && \text{(14a)}\\
\hat{y}_t = H\hat{B}\hat{\chi}_t &\sim N(H\hat{B}\chi_t, H\hat{B}\tilde{\Sigma}\hat{B}'H') \equiv N(H\hat{x}_t, H\hat{B}\tilde{\Sigma}\hat{B}'H'). && \text{(14b)}
\end{aligned}$$

In the equations above, $\tilde{K}$ and $\tilde{\Sigma}$ are defined by the system of equations

$$\begin{aligned}
\tilde{\Sigma} &= (I - \tilde{K}H\hat{B})\tilde{S} && \text{(15a)}\\
\tilde{K} &= \tilde{S}\hat{B}'H'(H\hat{B}\tilde{S}\hat{B}'H' + T)^{-1} && \text{(15b)}\\
\tilde{S} &= A\tilde{\Sigma}A' + C. && \text{(15c)}
\end{aligned}$$

The decoder estimates the residual $\hat{u}_t$ with

$$\begin{aligned} \hat{u}_t &= y_t - \hat{y}_t && (16a) \\ &= (H\hat{x}_t + u_t) - H\hat{B}\hat{\chi}_t && (16b) \\ &= H(\hat{x}_t - \hat{B}\hat{\chi}_t) + u_t. && (16c) \end{aligned}$$

Notice that since $\hat{x}_t \sim N(x_t, \Sigma)$ and $\hat{\chi}_t \sim N(\chi_t, \tilde{\Sigma})$,

$$\begin{aligned} \hat{x}_t - \hat{B}\hat{\chi}_t &\sim N(x_t - \hat{B}\chi_t, \Sigma + \hat{B}\tilde{\Sigma}\hat{B}') && (17a) \\ &\equiv N(x_t - \hat{x}_t, \Sigma + \hat{B}\tilde{\Sigma}\hat{B}') && (17b) \\ &\equiv N(x_t - x_t, 2\Sigma + \hat{B}\tilde{\Sigma}\hat{B}') && (17c) \\ &\equiv N(0, 2\Sigma + \hat{B}\tilde{\Sigma}\hat{B}'). && (17d) \end{aligned}$$

Thus, the estimated residual conditional on the value for the actual residual encoding the message is

$$\hat{u}_t | u_t \sim N(u_t, H(2\Sigma + \hat{B}\tilde{\Sigma}\hat{B}')H'). \tag{18}$$

A central question in determining decoding error rates is whether the decoder will interpret a residual in the same way the encoder intended. Specifically, assume the encoder is sending a message associated with a region R, and the error $u_t \in R$ encodes the message. The error rate for that region is given by the probability that the inferred residual $\hat{u}_t$ will not be in R given that the true error u_t is in R, which can be expressed as

$$\Pr(\hat{u}_t \notin R | u_t \in R) = \int_{u_t \in R} \phi(u_t | 0, T) \int_{\hat{u}_t \notin R} \phi(\hat{u}_t | u_t, H(2\Sigma + \hat{B}\tilde{\Sigma}\hat{B}')H') d\hat{u}_t du_t, \tag{19}$$

where $\phi(x|m, V)$ is the value of the normal density at x given mean m and variance V.

References

Kumar, P. R. and P. Varaiya (1986). *Stochastic Systems: Estimation, Identification, and Adaptive Control.* Prentice-Hall.