### 2.6.2 Optimal Prediction

Prediction error is defined by

$$\varepsilon(t,\theta) = y_t - \hat{y}(t|t-1)$$

where $\hat{y}(t|t-1)$ now denotes a prediction of $y_t$ given *all* data up to and including time $t-1$ (i.e., $y_{t-1}, u_{t-1}, y_{t-2}, u_{t-2}, \ldots$).

Consider the general model structure

$$y_t = G_p(z^{-1};\theta)u_t + G_l(z^{-1};\theta)e_t$$

with the assumption that $G_p(0;\theta) = 0$. A general linear one-step ahead predictor is described as [10]:

$$\hat{y}(t|t-1) = L_1(z^{-1};\theta)y_t + L_2(z^{-1};\theta)u_t \tag{2.15}$$

which is a function of past data if the filters $L_1(z^{-1};\theta)$ and $L_2(z^{-1};\theta)$ are constrained by

$$L_1(0;\theta) = 0 \tag{2.16}$$
$$L_2(0;\theta) = 0 \tag{2.17}$$

Thus, the prediction error can be further written as

$$\begin{aligned}
\varepsilon(t,\theta) &= G_p(z^{-1};\theta)u_t + G_l(z^{-1};\theta)e_t - L_1(z^{-1};\theta)y_t - L_2(z^{-1};\theta)u_t \\
&= G_p(z^{-1};\theta)u_t + (G_l(z^{-1};\theta) - I)e_t + e_t - L_1(z^{-1};\theta)y_t - L_2(z^{-1};\theta)u_t \\
&= (G_p(z^{-1};\theta) - L_2(z^{-1};\theta))u_t + (G_l(z^{-1};\theta) - I)e_t - L_1(z^{-1};\theta)y_t + e_t
\end{aligned}$$

According to the model,

$$y_t = G_p(z^{-1};\theta)u_t + G_l(z^{-1};\theta)e_t$$

$e_t$ can be derived as

$$e_t = G_l^{-1}(z^{-1};\theta)(y_t - G_p(z^{-1};\theta)u_t)$$

Using this relation, we can further write the expression of the prediction error as

$$\begin{aligned}
\varepsilon(t,\theta) &= (G_p(z^{-1};\theta) - L_2(z^{-1};\theta))u_t + (G_l(z^{-1};\theta) - I) \\
&\quad \times G_l^{-1}(z^{-1};\theta)(y_t - G_p(z^{-1};\theta)u_t) - L_1(z^{-1};\theta)y_t + e_t \\
&= (G_p(z^{-1};\theta) - L_2(z^{-1};\theta))u_t + (I - G_l^{-1}(z^{-1};\theta))(y_t - G_p(z^{-1};\theta)u_t) \\
&\quad - L_1(z^{-1};\theta)y_t + e_t \\
&= (G_l^{-1}(z^{-1};\theta)G_p(z^{-1};\theta) - L_2(z^{-1};\theta))u_t \\
&\quad + (I - G_l^{-1}(z^{-1};\theta) - L_1(z^{-1};\theta))y_t + e_t \\
&\triangleq \Psi_u(z^{-1};\theta)u_t + \Psi_y(z^{-1};\theta)y_t + e_t
\end{aligned}$$

Given the conditions $G_p(0;\theta) = 0$, $G_l(0;\theta) = I$, $L_1(0;\theta) = 0$, and $L_2(0;\theta) = 0$, it can be verified that

$$\Psi_u(0;\theta) = 0$$
$$\Psi_y(0;\theta) = 0$$

Namely, both $\Psi_u(z^{-1};\theta)$ and $\Psi_y(z^{-1};\theta)$ have at least one sample time delay. Thus, by expanding transfer functions into impulse response functions, we have

$$\Psi_u(z^{-1};\theta)u_t = \psi_{u1}u_{t-1} + \psi_{u2}u_{t-2} + \dots$$
$$\Psi_y(z^{-1};\theta)y_t = \psi_{y1}y_{t-1} + \psi_{y2}y_{t-2} + \dots$$

Being a future white-noise disturbance relative to $\Psi_u(z^{-1};\theta)u_t$ and $\Psi_y(z^{-1};\theta)y_t$, $e_t$ is independent of both $\Psi_u(z^{-1};\theta)u_t$ and $\Psi_y(z^{-1};\theta)y_t$. As a result

$$Cov(\varepsilon(t,\theta)) = Cov[\Psi_u(z^{-1};\theta)u_t + \Psi_y(z^{-1};\theta)y_t] + Cov[e_t] \succcurlyeq Cov(e_t)$$

or as a norm expression

$$trace[Cov(\varepsilon(t,\theta))] \geq trace[Cov(e_t)]$$

Therefore, the minimum is given by $Cov(e_t)$, i.e., the covariance of white noise $e_t$, which is $\Sigma_e$. Consequently, an optimal one-step predictor should give this lower bound as its prediction error. This lower bound is achieved if

$$\Psi_u(z^{-1};\theta) = 0$$
$$\Psi_y(z^{-1};\theta) = 0$$

Solving these two equations gives, respectively,

$$L_2(z^{-1};\theta) = G_l^{-1}(z^{-1};\theta)G_p(z^{-1};\theta)$$
$$L_1(z^{-1};\theta) = I - G_l^{-1}(z^{-1};\theta)$$

As the result, the optimal predictor can be derived.

Let's start from a simple example to demonstrate how the optimal prediction can be derived. Consider the following ARMAX model:

$$y_t = \frac{bz^{-1}}{1+az^{-1}}u_t + \frac{1+cz^{-1}}{1+az^{-1}}e_t$$

The white noise term $e_t$ can be derived from this equation as

$$e_t = \frac{1+az^{-1}}{1+cz^{-1}}(y_t - \frac{bz^{-1}}{1+az^{-1}}u_t) \tag{2.18}$$

The following derivation yields optimal one-step prediction: