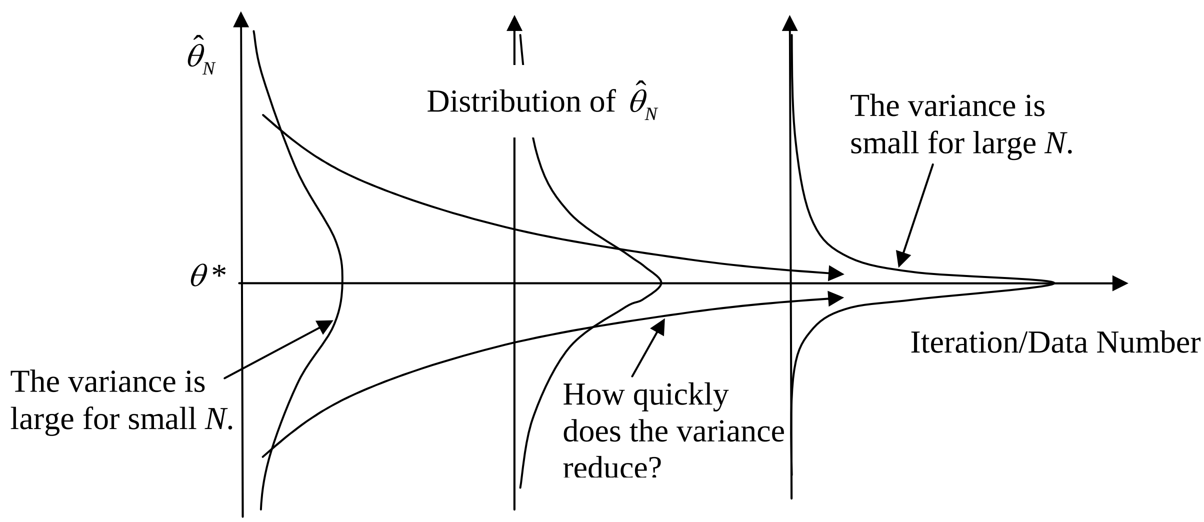
2.160 System Identification, Estimation, and Learning
Lecture Notes No. 18
April 26, 2006

13 Asymptotic Distribution of Parameter Estimates

13.1 Overview

If convergence is guaranteed, then $\hat{\theta}_N \to \theta^*$.

But, how quickly does the estimate $\hat{\theta}_N$ approach the limit θ^*? How many data points are needed? → Asymptotic Variance Analysis



The main points to be obtained in this chapter

The variance analysis of this chapter will reveal

a) The estimate converges to θ^* at a rate proportional to $\frac{1}{\sqrt{N}}$

b) Distribution converges to a Gaussian distribution: $\boldsymbol{N}(0, Q)$.

c) Cov $\hat{\theta}_N$ depends on the parameters sensitivity of the predictor: $\frac{\partial \hat{y}}{\partial \theta}$

Identified model parameter $\hat{\theta}_N$ with $\operatorname{cov}\hat{\theta}_N$: a "quality tag" confidence interval

12.2 Central Limit Theorems.

The mathematical tool needed for asymptotic variance analysis is "Central Limit" theorems. The following is a quick review of the theory.

Consider two independent random variable, X and Y, with PDF, $f_X(x)$ and $f_Y(y)$. Define another random variable Z as the sum of X and Y:

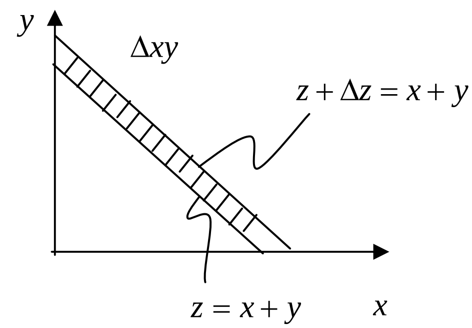
$$Z = X + Y$$

Let us obtain the PDF of Z.

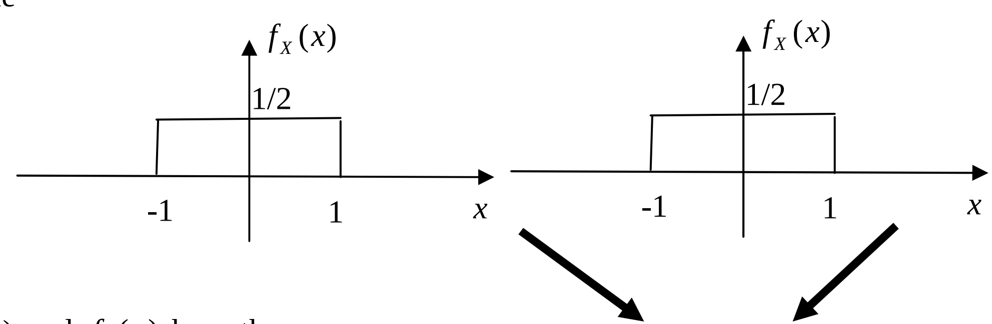
$$\Pr ob\left(z \le Z \le z + \Delta z\right)$$

$$= \int_{\Delta XY}\int f_X(x) f_Y(y) dxdy$$

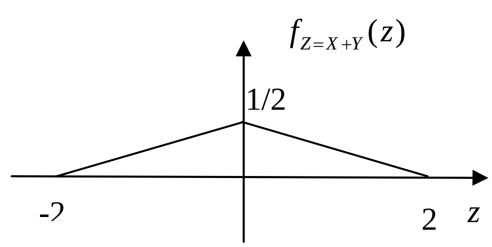
$$= \left[\int_{-\infty}^{\infty} f_X(x) f_Y(z-x) dx\right] \Delta z = f_Z(z) \Delta z$$



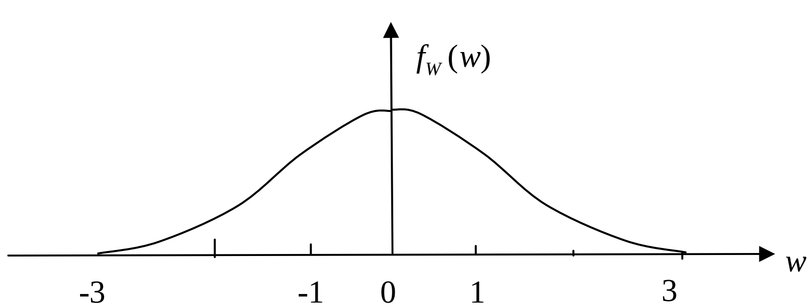
Example



$f_X(x)$ and $f_Y(y)$ have the same uniform distribution. Combining the two distributions, we can obtain the distribution of Z.



Further, consider $W = X + Y + V$, $f_V(v)$ has the same rectangular PDF as X and Y.



The resultant PDF is getting close to a Gaussian distribution.

In general, the PDF of a random variable $\sum_{i=1}^{N} X_i$ approaches a Gaussian distribution, regardless of the PDF of each X_i, as N gets larger. More rigorously, the following central limit theorem has been proven.

A Central Limit Theorem of Independent Random Variables

Let $X_t, t = 0,1,\cdots$ be a *d*-dimensional random variable with

$$\begin{aligned} &\text{Mean} && m = E(X_t) \\ &\text{Co-variance} && Q = E\left[(X_t - m)(X_t - m)^T\right] \end{aligned} \tag{1}$$

Consider the sum of $X_t - m$ given by

$$Y_N = \frac{1}{\sqrt{N}} \sum_{t=1}^{N} (X_t - m) \tag{2}$$

Then, as N tends to infinity, the distribution of Y_N converges to the Gaussian distribution given by PDF:

$$f_Y(y) = \frac{1}{(2\pi)^{d/2}\sqrt{\det Q}} \exp\left\{-\frac{1}{2} y^T Q^{-1} y\right\} \tag{3}$$

where

$$y = \lim_{N\to\infty} \frac{1}{\sqrt{N}} \sum_{t=1}^{N} (X_t - m).$$

13.3 Distribution of Estimate $\hat{\theta}_N$

Applying the Central Limit Theorem, we can obtain the distribution of estimate $\hat{\theta}_N$ as N tends to infinity.
Let $\hat{\theta}_N$ be an estimate based on the prediction error method (PEM);

$$\hat{\theta}_N = \arg\min_{\theta \in D_M} V_N(\theta, Z^N) \tag{4}$$

$$V_N(\theta, Z^N) = \frac{1}{N} \sum_{t=1}^{N} \frac{1}{2} \varepsilon^2(t,\theta) \tag{5}$$

For simplicity, we first assume that the predictor $\hat{y}(t|\theta)$ is given by a linear regression:

$$\hat{y}(t|\theta) = \varphi^T \theta \tag{6}$$

and the parameter vector of the true system, θ_0, is involved in the model set, $\theta_0 \in D_M$.

The actual data is generated by

$$\hat{y}(t|\theta) = \varphi^T \theta_0 + e_0(t) \tag{7}$$

where

$$E[e_0(t)e_0(s)] = \begin{cases} \lambda_0 & t = s \\ 0 & t \neq s \end{cases}.$$

Since $\hat{\theta}_N$ minimizes $V_N(\theta, Z^N)$

$$V'_N(\hat{\theta}_N, Z^N) = \frac{d}{d\theta} V_N(\theta_N, Z^N)\Big|_{\theta=\theta_N} = 0, \qquad V'_N \in R^{d\times 1} \tag{8}$$

Using the Mean Value Theorem, V'_N can be expressed as

$$V'_N(\hat{\theta}_N, Z^N) = V'_N(\theta_0, Z^N) + V''_N(\xi_N, Z^N)(\hat{\theta}_N - \theta_0) \qquad \hat{\theta}_N \leq \xi_N \leq \theta_0 \; or \quad \theta_0 \leq \xi_N \leq \hat{\theta}_N \tag{9}$$

where ξ_N is a parameter vector somewhere between θ_0 and $\hat{\theta}_N$.

Assuming that $V''_N(\xi_N, Z^N) = \frac{d}{d\theta} V'_N$ is non-singular and using (8) for (9),

$$\hat{\theta}_N - \theta_0 = -\left[V''_N(\xi_N, Z^N)\right]^{-1} V'_N(\theta_0, Z^N) \tag{10}$$

To obtain the distribution of $\hat{\theta}_N - \theta_0$, let us first examine $V'_N(\theta_0, Z^N)$ as N tends to infinity.

$$V'_N(\theta_0, Z^N) = \frac{1}{N}\sum_{t=1}^{N} \varepsilon(t, \theta_0) \frac{d\varepsilon}{d\theta}\Big|_{\theta=\theta_0}, \tag{11}$$

Recall $\varepsilon(t,\theta) = y(t) - \hat{y}(t|\theta)$ and (6)

$$\frac{d\varepsilon}{d\theta}\Big|_{\theta_0} = -\frac{d}{d\theta}\hat{y}(t|\theta)\Big|_{\theta_0} = -\varphi^T(t), \tag{12}$$

and

$$\varepsilon(t, \theta_0) = \varphi^T(t)\theta_0 + e_0(t) - \varphi^T(t)\theta_0 = e_0(t)$$

Therefore, (11) reduces to

$$-V'_N(\theta_0, Z^N) = \frac{1}{N}\sum_{t=1}^{N}\varphi(t)e_0(t) \tag{13}$$

Let us treat $\varphi(t)e_0(t) \equiv X_t$ as a random variable. Its mean is zero, since

$$m = \bar{E}[\varphi(t)e_0(t)] = \bar{E}[\varphi(t)]\bar{E}[e_0(t)] = 0 \tag{14}$$

The covariance is

$$\mathrm{cov}(X_t X_s^T) = \bar{E}\left[\left(X_t - m\right)\left(X_s - m\right)^T\right] = \bar{E}\left[\varphi(t)e_0(t)e_0(s)\varphi(s)^T\right] = 0, \quad \text{for } t \neq s \tag{15}$$

$$\mathrm{cov}(X_t X_t^T) = \bar{E}\left[e_0^2(t)\right]\bar{E}\left[\varphi(t)\varphi(t)^T\right] = \lambda_0 \bar{R} \tag{16}$$

Note that $X_1, X_2, \ldots, X_N$ are independent, since $e_0(t)$ is independent.

Consider

$$Y_N = \frac{1}{\sqrt{N}}\sum_{t=1}^{N}\left(X_t - m\right) = \frac{1}{\sqrt{N}}\sum_{t=1}^{N}\varphi(t)e_0(t)$$

and apply the Central Limit Theorem. The distribution of Y_N, i.e. $-\sqrt{N}V'_N(\theta_0, Z^N)$, converges to a Gaussian distribution as N tends to infinity.

$$Y_N = -\sqrt{N}V'_N(\theta_0, Z^N) \sim N(0, \lambda_0 \bar{R}) \tag{17}$$

Next, compute $V''_N(\xi_N, Z^N)$

$$\begin{aligned}
V''_N(\xi_N, Z^N) &= \frac{d}{d\theta}V'_N\left(\theta, Z^N\right)\Big|_{\theta=\xi_N} \\
&= \frac{d}{d\theta}\frac{1}{N}\sum_{t=1}^{N}\varepsilon(t,\theta)\frac{d\varepsilon}{d\theta}\Big|_{\theta=\xi_N} \\
&= \frac{1}{N}\left\{\frac{d\varepsilon}{d\theta}\left(\frac{d\varepsilon}{d\theta}\right)^T + \varepsilon(t,\theta)\frac{d^2\varepsilon}{d\theta^2}\right\}\Big|_{\theta=\xi_N} \\
&= \frac{1}{N}\sum_{t=1}^{N}\left(\varphi(t)\varphi^T(t)\right)
\end{aligned} \tag{18}$$

Therefore, under the ergodicity assumption,

$$\bar{V}''_N(\xi_N, Z^N) = \lim_{N\to\infty}\frac{1}{N}\sum_{t=1}^{N}\left(\varphi(t)\varphi^T(t)\right) = \bar{R} \tag{19}$$

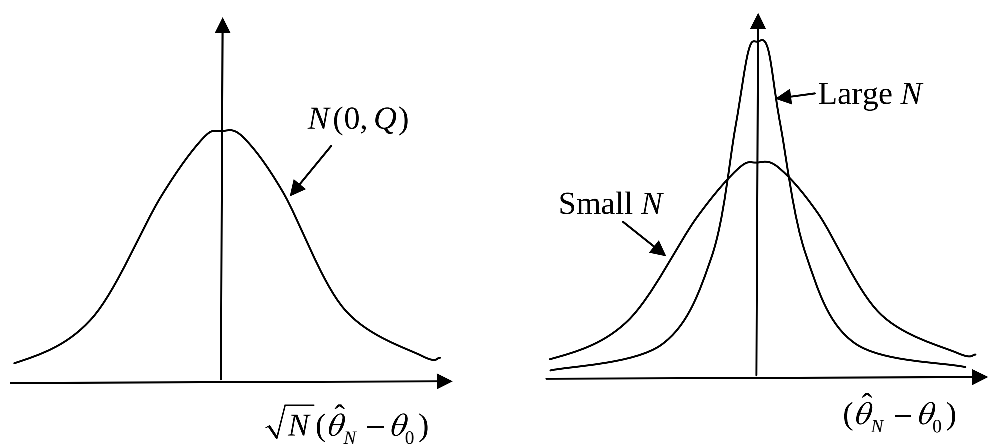
From (10), (17) and (19), the distribution of $\sqrt{N}(\hat{\theta}_N - \theta_0)$ converges to the Gaussian distribution given by

$$\sqrt{N}(\hat{\theta}_N - \theta_0) \sim N(0,Q)\, as \quad N \to \infty \tag{20}$$

where

$$Q = \overline{R}^{-1}(\lambda_0 \overline{R})\overline{R}^{-1} = \lambda_0 \overline{R}^{-1} \tag{21}$$

Note that, as coordinate transformation $y=Ax$ is performed, the covariant matrix C associated with a multivariate Gaussian distribution is transformed to ACA^T. This is used in (21).



<u>Remarks</u>

1) Eq.(20) manifests that the standard deviation of $\hat{\theta}_N - \theta_0$ decreases at the rate of $\frac{1}{\sqrt{N}}$ for large N. See the figure above. Note that $\operatorname{cov}\hat{\theta}_N = \frac{1}{N}Q$.

2) The above result is for a very restrictive case. A similar result can be obtained for general cases with mild assumptions.
   - The true system (7) does not have to be assumed. Instead, $\theta^* = \arg\min \overline{V}(\theta)$ must be involved in D_M.
   - The linear regression (6) can be extended to a general predictor where the model parameter θ is determined based on the prediction error method (4), (5).

The extended result of estimate distribution is summarized in the following theorem, i.e. Ljun'g Textbook Theorem 9-1.

Theorem 1 Consider the estimate $\hat{\theta}_N$ determined by (4) and (5). Assume that the model structure is linear and uniformly stable and that the data set Z^∞ satisfies the quasi stationary and ergodicity requirements. Assume also that $\hat{\theta}_N$ converges with probability 1 to a unique parameter vector θ^* involved in D_M:

$$\hat{\theta}_N \to \theta^* \in D_M \qquad w.p.1 \quad as\, N \to \infty \tag{22}$$

and that

$$\overline{V}''_N(\theta^*) > 0 \;\; ; \text{ positive definite} \tag{23}$$

and that

$$V'(\theta^*) = \lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} \left(\frac{d}{d\theta} \hat{y}(t|\theta) \right) \varepsilon(t,\theta) \Big|_{\theta^*} \text{ converges to } m_t \text{ with probability 1} \tag{24}$$

where m_t is the ensemble mean given by

$$m_t = \overline{E}\left[\sum_{t=1}^{N} \left(\frac{d}{d\theta} \hat{y}(t|\theta) \right) \varepsilon(t,\theta) \Big|_{\theta^*} \right] \tag{25}$$

Then, the distribution of $\sqrt{N}(\hat{\theta}_N - \theta_0)$ converges to the Gaussian distribution given by

$$\sqrt{N}(\hat{\theta}_N - \theta_0) \sim N(0, P_\theta) \tag{26}$$

where P_θ is given by

$$P_\theta = \left[\overline{V}''_N(\theta^*)\right]^{-1} Q \left[\overline{V}''_N(\theta^*)\right]^{-1} \tag{27}$$

$$Q = \lim_{N\to\infty} N \cdot E[\left(\overline{V}'_N(\theta^*)\right)\left(\overline{V}'_N(\theta^*)\right)^T] \tag{28}$$

The proof is quite complicated, since the random variable $\left(\frac{d}{d\theta} \hat{y}(t|\theta) \right) \varepsilon(t,\theta) \Big|_{\theta^*}$ is not independent. Therefore, the standard central limit theorem is not applicable.

Appendix 9A, at p.309 of Ljung's textbook, shows the outline of proof. Since the model structure is assumed to be stable uniformly in θ, X_t and X_s are independent as t and s are distal. Because of this property, the sum, $\frac{1}{\sqrt{N}} \sum_{t=1}^{N} \left(X_t - m_t \right)$, converges to the Gaussian distribution.

13.4 Expression for the Asymptotic Variance.

As stated formally in Theorem 1, the distribution of $\sqrt{N}\left(\hat{\theta}_N - \theta^*\right)$ converges to a Gaussian distribution for the broad class of system identification problems. This implies that the covariance of $\hat{\theta}_N$ asymptotically converges to:

$$Cov\hat{\theta}_N \sim \frac{1}{N} P_\theta \tag{29}$$

This is called the asymptotic covariance matrix.

The asymptotic variance depends not only on

(a) the member of samples/data set size: N, but also on

(b) the parameter sensitivity of the predictor:

$$\psi\left(t,\theta^*\right) = \frac{d}{d\theta}\hat{y}(t|\theta)\Big|_{\theta^*} = -\frac{d}{d\theta}\varepsilon(t,\theta)\Big|_{\theta^*} \text{ and} \tag{30}$$

(c) Noise variance λ_0.

Let us compute the covariance once again for the general case. Form (5) and (30),

$$\frac{d}{d\theta}V_N\left(\theta, Z^N\right) = \frac{1}{N}\sum_{t=1}^{N}\varepsilon(t,\theta)\frac{d\varepsilon}{d\theta} = -\frac{1}{N}\sum_{t=1}^{N}\varepsilon(t,\theta)\psi\left(t,\theta\right) \tag{31}$$

Unlike the linear regression, the sensitivity $\psi\left(t,\theta\right)$is a function of θ,

$$\begin{aligned}\frac{d^2}{d\theta^2}V_N\left(\theta, Z^N\right) &= -\frac{1}{N}\sum\left(\frac{d\varepsilon}{d\theta}\psi + \varepsilon\frac{d\psi}{d\theta}\right)\\ &= \frac{1}{N}\sum_{t=1}^{N}\left(\psi(t,\theta)\psi^T(t,\theta) - \varepsilon(t,\theta)\frac{d^2}{d\theta^2}\hat{y}(t|\theta)\right)\end{aligned} \tag{32}$$

When the true system is contained in the model structure, $\theta_0 \in D_M$, and that is unique,

$$\varepsilon(t,\theta_0) = e_0(t) \tag{33}$$

from (28), (31), and (33)

$$Q = \lim_{N\to\infty} \frac{N}{N^2} \sum_{t=1}^{N} \sum_{s=1}^{N} E\left[e_0(t)\psi(t,\theta_0)\psi^T(s,\theta_0)e_0(s)\right]$$
$$= \lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} \lambda_0 E\left[\psi(t,\theta_0)\psi^T(t,\theta_0)\right] = \lambda_0 \overline{E}\left[\psi(t,\theta_0)\psi^T(t,\theta_0)\right] \tag{34}$$

Also from (32)

$$\overline{V}''(\theta_0) = \lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} \left[E\left[\psi(t,\theta_0)\psi^T(t,\theta_0)\right] - \varepsilon(t,\theta_0) \frac{d^2}{d\theta^2} \hat{y}(t|\theta)\Big|_{\theta_0} \right]$$
$$= \overline{E}\left[\psi(t,\theta_0)\psi^T(t,\theta_0)\right] - \cancelto{0}{\overline{E}\left[e_0(t) \frac{d^2}{d\theta^2} \hat{y}\Big|_{\theta_0} \right]} \tag{35}$$

This depends on Z^{t-1} not on Z^t

Since $e_0(t)$ and $\frac{d^2}{d\theta^2}\hat{y}$ are independent, the second term varnishes. Substituting (34) and (35) into (29),

$$Cov\hat{\theta}_N \sim \frac{1}{N} P_\theta = \frac{\lambda_0}{N} \left[\overline{E}\left(\psi(t,\theta_0)\psi^T(t,\theta_0)\right)\right]^{-1} \tag{36}$$

The asymptotic variance is therefore a) inversely proportional to the number of samples, b) proportional to the noise variance, and c) inversely related to the parameter sensitivity. <u>The more a parameter affects the prediction, the smaller the variance becomes.</u>

Since θ_0 is not known, the asymptotic variance cannot be determined. In practice, however, an empirical estimate, like the following formula, works well for large N.

$$\widehat{P}_N = \hat{\lambda}_N \left[\frac{1}{N} \sum_{t=1}^{N} \psi(t,\hat{\theta}_N)\psi^T(t,\hat{\theta}_N) \right]^{-1} \tag{37}$$

$$\hat{\lambda}_N = \frac{1}{N} \sum_{t=1}^{N} \varepsilon^2(t,\hat{\theta}_N) \tag{38}$$

If one computes $\widehat{P}_N$ during experiments, sufficient data samples needed for assuming the model accuracy may be obtained.

13.5 Frequency-Domain Expressions for the Asymptotic Variance.

The asymptotic variance has different expression in the frequency domain, which we will find useful for variance analysis and experiment design.

Let transfer function $G(q,\theta)$ and noise model $H(q,\theta)$ be consolidated into as 1X2 matrix:

$$T(q,\theta) = [G(q,\theta), H(q,\theta)] \tag{39}$$

The gradient of T, that is, the sensitivity of T to θ, is

$$T'(q,\theta) = \frac{d}{d\theta}T(q,\theta) = [G'(q,\theta), H'(q,\theta)] \tag{40}$$

For a predictor, we have already defined $W(q,\theta)$ and z(t), s.t.

$$\hat{y}(t|\theta) = W_u(q)u(t) + W_y(q)y(t) = \begin{bmatrix} W_u & W_y \end{bmatrix} \begin{bmatrix} u \\ y \end{bmatrix} = WZ(t)$$

Therefore the predictor sensitivity $\psi(t,\theta)$ is given by

$$\psi(t,\theta) = \frac{d}{d\theta}\hat{y}(t|\theta) = \begin{bmatrix} W'_u & W'_y \end{bmatrix} Z(t) \tag{41}$$

$W_u^{'}$ and $W_y^{'}$ are computed as

$$W_u^{'} \Rightarrow \frac{d}{d\theta}W_u(z,\theta) = \frac{d}{d\theta}H^{-1}(z,\theta)G(z,\theta) = \frac{HG' - H'G}{H^2(z,\theta)} \tag{42}$$

$$W_y^{'} \Rightarrow \frac{d}{d\theta}W_y(z,\theta) = \frac{d}{d\theta}\left[1 - H^{-1}(z,\theta)\right] = \frac{H'(z,\theta)}{H^2(z,\theta)}$$

Substituting these back to $\psi(t,\theta)$

$$\begin{aligned} \psi(t,\theta) &= \frac{1}{H^2(q,\theta)}\begin{bmatrix} HG' - H'G, & H' \end{bmatrix} Z(t) \\ &= \frac{1}{H^2(q,\theta)}\begin{bmatrix} G', & H' \end{bmatrix} \begin{bmatrix} H & 0 \\ -G & 1 \end{bmatrix} \begin{bmatrix} u(t) \\ y(t) \end{bmatrix} \\ &= \frac{1}{H(q,\theta)} T'(q,\theta) \begin{bmatrix} u(t) \\ -H^{-1}Gu + H^{-1}y \end{bmatrix} \end{aligned} \tag{43}$$

At $\theta = \theta_0$ (the true system), note $\varepsilon(t,\theta_0) = e_0(t)$ and
$-H^{-1}(q,\theta_0)G(q,\theta_0)u(t) + H^{-1}(q,\theta_0)y(t) = e_0(t)$

$$\therefore \psi(t,\theta_0) = H^{-1}(q,\theta_0)T'(q,\theta_0)x_0(t) \tag{44}$$

where $x_0(t) = \begin{bmatrix} u(t) & e_0(t) \end{bmatrix}^T$.

Let $\Phi_{x_0}(\omega)$ be the spectrum matrix of $x_0(t)$

$$\Phi_{x_0}(\omega)=\begin{bmatrix}\Phi_u(\omega) & \Phi_{ue_0}(\omega)\\ \Phi_{ue_0}(-\omega) & \Phi_{e_0}(\omega)\end{bmatrix}\quad \begin{array}{l}\Phi_{e_0}(\omega)=\lambda_0\\ \Phi_{ue_0}(\omega)=0\ for\quad open-loop\end{array} \tag{45}$$

Using the familiar formula: $R_s(0)=\frac{1}{2\pi}\int_{-\pi}^{\pi}\Phi_s(\omega)d\omega$

$$\bar{E}\left[\psi(t,\theta_0)\psi^T(t,\theta_0)\right]=\frac{1}{2\pi}\int_{-\pi}^{\pi}\left|H(e^{i\omega},\theta_0)\right|^{-2}T'(e^{i\omega},\theta_0)\Phi_{x_0}(\omega)T'^T(e^{i\omega},\theta_0)d\omega \tag{46}$$

For the noise spectrum,

$$\Phi_v(\omega)=\lambda_0\left|H(e^{i\omega},\theta_0)\right|^2 \tag{47}$$

Using these in (36)

$$Cov\hat{\theta}_N\sim\frac{1}{N}\left[\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{1}{\Phi_v(\omega)}T'(e^{i\omega},\theta_0)\Phi_{x_0}(\omega)T'^T(e^{i\omega},\theta_0)d\omega\right]^{-1}$$

The asymptotic variance in the frequency domain.