Joint Probability Distribution of Prediction Errors of ARIMA

Xin Qin, Jyotirmoy V. Deshmukh
Email: {xinqin, jyotirmoy.deshmukh}@usc.edu
University of Southern California

Abstract. Producing probabilistic guarantee for several steps of a predicted signal follow a temporal logic defined behavior has its rising importance in monitoring. In this paper, we derive a method to compute the joint probability distribution of prediction errors of multiple steps based on Autoregressive Integrated Moving Average (ARIMA) model. We cover scenarios in stationary process and intrinsically stationary process for univariate and multivariate.

Keywords: Probabilistic Reasoning, ARIMA

1 Introduction

Signal Temporal Logic (STL) specify the simultaneous behavior of a signal across different time point, while time series prediction based on Autoregressive Integrated Moving Average (ARIMA) model only give guarantee on each single time point. In this paper, we give the proof of calculating joint probability distribution of prediction errors of multiple steps.

2 Error of Prediction in Univariate Process

The goal is to obtain joint probability distribution of

\[ \{\text{Err}(X_{T+1}), \text{Err}(X_{T+2}), \ldots, \text{Err}(X_{T+h})\} \]

2.1 Error in Stationary Process

Previous time series work [2] [1] concluded that for stationary process

\[
E[\text{Err}(X_{T+h})] = 0 \\
E[\text{Err}(X_{T+h})]^2 = \gamma_X(0) - (a_n^h)^T \gamma_n(h)
\]

Assuming \( \text{Err}(X_{T+h}) \) is a normal distribution, then \( \text{Err}(X_{T+h}) \sim N(0, \gamma_X(0) - (a_n^h)^T \gamma_n(h)) \).
Definition 1.  
\[
\text{Err}(X_{T+h}) = P_nX_{T+h} - X_{T+h} \tag{1}
\]

Definition 2. For best linear predictor [2]
\[
P_n(\alpha_1U + \alpha_2V + \beta) = \alpha_1P_n(U) + \alpha_2P_n(V) + \beta
\]

Corollary 1. The value of best linear predictor at observed datapoint \(\{X_1, X_2, \ldots, X_T\}\) follows
\[
P_n(\sum_{i=1}^{n} \alpha_iX_i + \beta) = \sum_{i=1}^{n} \alpha_iX_i + \beta
\]

Definition 3. If the distribution of \(X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} \) is a multivariate normal distribution, then the random variable vector \(Y = CX + b\), where \(C\) is a \(r \times n\) matrix, is also a multivariate normal distribution. And if \(\{X_t\} \sim N_n(\mu, \Sigma)\), then \(\{Y_t\} \sim N_r(C\mu + b, C\Sigma C^T)\)

Definition 4. If \(\{Z_t\} \sim WN(0, \sigma^2)\) are independent white noise, the joint probability distribution of
\[
\begin{pmatrix} Z_{T-q+1} \\ Z_{T-q+2} \\ \vdots \\ Z_{T+h} \end{pmatrix}
\]
is a multivariate normal distribution with
\[
\mu_Z = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \Sigma_Z = \begin{pmatrix} \sigma^2 & 0 & \ldots & 0 \\ 0 & \sigma^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma^2 \end{pmatrix} \tag{2}
\]

Lemma 1. Given \(Y_t = \theta(B)Z_t\ \ \ t \geq p + 1\), let \(Y_h\) denote \(\{Y_{T+i} \}_{i=1}^{h}\), then
\[
Y_h = C_1 \begin{pmatrix} Z_{T-q+1} \\ Z_{T-q+2} \\ \vdots \\ Z_{T+h} \end{pmatrix} \tag{3}
\]
where $C_1$ is a $[h \times (q + h)]$ transform matrix from

$$
\begin{pmatrix}
(Z_{-q+1}) & \\
(Z_{-q+2}) & \\
\vdots & \\
(Z_{T+h}) & 
\end{pmatrix}
$$

to $Y_h$.

\[ C_1 = \begin{pmatrix}
\theta_q & \theta_{q-1} & \cdots & \theta_1 & 1 & 0 & 0 & \cdots & 0 \\
0 & \theta_q & \theta_{q-1} & \cdots & \theta_1 & 1 & 0 & \cdots & 0 \\
0 & 0 & \theta_q & \theta_{q-1} & \cdots & \theta_1 & 1 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\
0 & 0 & 0 & \cdots & \theta_q & \theta_{q-1} & \cdots & \theta_1 & 1 \\
\end{pmatrix} \] (4)

**Proof.** From the definition of ARMA($p,q$) process,

\[ Y_t = \theta(B)Z_t \quad t \geq p + 1 = Z_t + \theta_1Z_{t-1} + \theta_2Z_{t-2} + \cdots + \theta_qZ_{t-q} \] (5)

where $Z_t \sim WN(0,\sigma^2)$.

**Lemma 2.**

\[ P_nY_{T+h} = C_3C_2 \begin{pmatrix}
(Z_{p-q+1}) & \\
(Z_{p-q+2}) & \\
\vdots & \\
(Z_T) & 
\end{pmatrix} \] (6)

where $C_2$ is a $[(T-p) \times (T-p+q)]$ matrix, $C_3$ is a $[h \times h]$ matrix.

\[ C_2 = \begin{pmatrix}
\theta_q & \theta_{q-1} & \cdots & \theta_1 & 1 & 0 & 0 & \cdots & 0 \\
0 & \theta_q & \theta_{q-1} & \cdots & \theta_1 & 1 & 0 & \cdots & 0 \\
0 & 0 & \theta_q & \theta_{q-1} & \cdots & \theta_1 & 1 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\
0 & 0 & 0 & \cdots & \theta_q & \theta_{q-1} & \cdots & \theta_1 & 1 \\
\end{pmatrix} \] (7)

\[ C_3 = \begin{pmatrix}
a_1^{T-p} & a_1^{T-p-1} & \cdots & a_1^1 \\
a_2^{T-p} & a_2^{T-p-1} & \cdots & a_2^1 \\
\vdots & \vdots & \ddots & \vdots \\
a_h^{T-p} & a_h^{T-p-1} & \cdots & a_h^1 \\
\end{pmatrix} \] (8)

**Proof.** \{$Y_t$\} is a stationary process starting from $p + 1$ with zero mean, the prediction of \{\$Y_t\\} follows

\[ P_nY_{T+h} = a_0^h + \sum_{i=1}^{T-p} a_i^hY_{T+1-i} \] (9)

where $a_0^h = 0$ is the coefficient matrix of \{$Y_{p+1},Y_{p+2},\ldots,Y_T$\} when calculating $P_nY_h$, the best linear predictor of \{$Y_{T+1},Y_{T+2},\ldots,Y_{T+h}$\}, a stationary MA($q$) process with zero mean.
Theorem 1. The covariance matrix of prediction error of a moving average process \( Y \) is

\[
\Sigma_{\text{Err}Y} = C_{Z\text{toErr}Y} \Sigma_Z C_{Z\text{toErr}Y}^T
\]  

(10)

Proof. Let \( X_h \) denote \( [X_{T+i}^h]_{i=1}^h \), \( P_n X_h \) denote \( [P_n X_{T+i}]_{i=1}^h \), \( \text{Err}(X_h) \) denote \( [\text{Err}(X_{T+i})]_{i=1}^h \), \( Y_h \) denote \( [Y_{T+i}]_{i=1}^h \), \( \text{Err}(Y_h) \) denote \( [\text{Err}(Y_{T+i})]_{i=1}^h \)

From definition \( \text{Err}(Y_{T+h}) = P_n(Y_{T+h}) - Y_{T+h} \)

(11)

From lemma \( \text{1} \) and \( \text{2} \)

\[
\text{Err}(Y_h) = C_3 C_2 \begin{pmatrix} Z_{p-q+1} \\ Z_{p-q+2} \\ \vdots \\ Z_T \end{pmatrix} - C_1 \begin{pmatrix} Z_{T-q+1} \\ Z_{T-q+2} \\ \vdots \\ Z_{T+h} \end{pmatrix}
\]

(12)

After augmenting the coefficient matrix \( C_1 \) and \( C_2 \) to \( C_1^* \) and \( C_2^* \) by

\[
C_1^* = (O_{h \times (T-p)} \quad C_1) \\
C_2^* = (C_2 \quad O_{(T-p) \times h})
\]

(13)

equation \( \text{12} \) is transformed into

\[
\text{Err}(Y_h) = C_3 C_2^* \begin{pmatrix} Z_{p-q+1} \\ Z_{p-q+2} \\ \vdots \\ Z_{T+h} \end{pmatrix} - C_1^* \begin{pmatrix} Z_{p-q+1} \\ Z_{p-q+2} \\ \vdots \\ Z_{T+h} \end{pmatrix}
\]

(14)

\[
= (C_3 C_2^* - C_1^*) \begin{pmatrix} Z_{p-q+1} \\ Z_{p-q+2} \\ \vdots \\ Z_{T+h} \end{pmatrix}
\]

Let \( C_{Z\text{toErr}Y} \) denote \( C_3 C_2^* - C_1^* \). Follow Lemma \( 3 \) \( \text{Err}(Y_h) \sim N_h(0, \Sigma_{\text{Err}Y}) \),

where

\[
\Sigma_{\text{Err}Y} = C_{Z\text{toErr}Y} \Sigma_Z C_{Z\text{toErr}Y}^T
\]

(15)

the size of \( \Sigma_Z \) here is \( (T-p+q) \times (T-p+q) \).

Theorem 2. \( \text{Err}(X_h) \) is a multivariate normal distribution \( \text{Err}(X_h) \sim N_h(0, \Sigma_{\text{Err}X}) \) in ARMA\((p,q)\) process \( \{X_t\} \), using the best linear predictor \( P_n X_{T+h} \). The covariance matrix of prediction error is

\[
\Sigma_{\text{Err}X} = C_{\text{Err}Y\to\text{Err}X} C_{Z\text{toErr}Y} \Sigma_Z C_{Z\text{toErr}Y}^T C_{\text{Err}Y\to\text{Err}X}^T
\]

(16)

where

\[
C_{\text{Err}Y\to\text{Err}X} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ c_{21} & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \cdots & \vdots \\ c_{h1} & c_{h2} & c_{h3} & \cdots & 1 \end{pmatrix}
\]

(17)
in which \( c_{ij} \) is the coefficient of \( \sigma_{Err}(Y_{T+i}) \) when calculating \( Err(X_{T+j}) \). \( c_{ij} \) can be recursively given by: For any \( i \geq 2 \),

\[
c_{i,1} = \sum_{k=1}^{\min(p,h-1)} \phi_k c_{i-k,1}
\]

(18)

For any \( j \geq 2 \),

\[
c_{i,j} = c_{i-1,j-1}
\]

(19)

**Proof.** From ARIMA we know

\[
Y_t = \phi(B)X_t \quad t > \max(p,q)
\]

(20)

Then from equation (20)

\[
X_{T+h} = Y_{T+h} + \sum_{i=1}^{p} \phi_i X_{T+h-i}
\]

(21)

From definition \( P \)

\[
P_n(X_{T+h}) = P_n(Y_{T+h} + \sum_{i=1}^{p} \phi_i X_{T+h-i})
\]

\[
= P_n(Y_{T+h}) + \sum_{i=1}^{p} \phi_i P_n(X_{T+h-i})
\]

(22)

From corollary \( I \)

\[
P_n(X_{T+h}) = \begin{cases} 
P_n(Y_{T+h}) + \sum_{i=1}^{h-1} \phi_i P_n(X_{T+h-i}) + \sum_{i=h}^{p} \phi_i X_{T+h-i} & h \leq p \\
P_n(Y_{T+h}) + \sum_{i=1}^{p} \phi_i P_n(X_{T+h-i}) & h > p 
\end{cases}
\]

(23)

Then we can see

\[
Err(X_{T+h}) = P_n(X_{T+h}) - X_{T+h}
\]

\[
= P_n(Y_{T+h}) + \sum_{i=1}^{h-1} \phi_i P_n(X_{T+h-i}) + \sum_{i=h}^{p} \phi_i X_{T+h-i}
\]

\[
- Y_{T+h} - \sum_{i=1}^{p} \phi_i X_{T+h-i}
\]

\[
= Err(Y_{T+h}) + \sum_{i=1}^{\min(p,(h-1))} \phi_i Err(X_{T+h-i})
\]

(24)

which represented in matrix form is

\[
Err(X_h) = C_{ErrYtoErrX} Err(Y_h)
\]

(25)
Let $C_{\text{Err}Y \to \text{Err}X}$ denote
\[
\begin{pmatrix}
1 & 0 & 0 & \ldots & 0 \\
c_{21} & 1 & 0 & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
c_{h_1} & c_{h_2} & c_{h_3} & \ldots & 1
\end{pmatrix}
\] (26)

Using Lemma 3, $\text{Err}(X_h) \sim N_h(0, \Sigma_{\text{Err}X})$ where
\[
\Sigma_{\text{Err}X} = C_{\text{Err}Y \to \text{Err}X} \Sigma_{\text{Err}Y} C_{\text{Err}Y \to \text{Err}X}^T
= C_{\text{Err}Y \to \text{Err}X} Z_{\text{Err}Y} \Sigma_{Z} Z_{\text{Err}Y} C_{\text{Err}Y \to \text{Err}X}^T
\] (27)

Q.E.D.

From definition $\text{Err}(X_{T+h}) \sim N_h(0, \Sigma_{\text{Err}X})$ where $\Sigma_{\text{Err}X} = C_{\text{Err}Y \to \text{Err}X} \Sigma_{\text{Err}Y} C_{\text{Err}Y \to \text{Err}X}^T$ leads us to the following conclusion:

**Theorem 3.**
\[
P(X_{T+h} \in [c_1, c_2]) = P(\text{Err}(X_{T+h}) \in [P_n(X_{T+h}) - c_2, P_n(X_{T+h}) - c_1])
\]

As STL formula produce target prediction interval at each time step $T+h$, denote the event $(X_{T+h} \in [c_{h_1,1}, c_{h_2,2}])$ as event $A_h$, then while $\{X_t\}$ is a stationary process, the joint probability of $(\bigcup_{t=1}^{h} A_t)$ is given by Theorem 2.

### 2.2 Error in Intrinsically Stationary Process

**Definition 5.** A process $\{X_t\}$ is defined as a $d$-ordered intrinsically stationary process when its $d$-ordered differencing is a stationary process while its $(d-1)$-ordered process is still a non-stationary process.

\[
\nabla^d X_t = \sum_{k=0}^{d} \binom{d}{k} (-1)^k X_{t-k}
\] (29)

**Lemma 3.** To a $d$-ordered intrinsically stationary process, the best linear prediction of $X_{T+h}$ based on observation $\{1, X_1, X_2, \ldots, X_T\}$ is:

\[
P_T X_{T+h} = \begin{cases}
P_T \nabla^d X_{T+h} - \sum_{k=h}^{d} \binom{d}{k} (-1)^k P_T X_{T+h-k} & h \leq d + 1 \\
- \sum_{k=1}^{h-1} \binom{d}{k} (-1)^k P_T X_{T+h-k} & h = d + 1
\end{cases}
\] (30)

\[
P_T \nabla^d X_{T+h} - \sum_{k=1}^{d} \binom{d}{k} (-1)^k P_T X_{T+h-k} & h > d + 1
\]
Proof. From the definition of differencing function, equation 29:

\[ X_{T+h} = \nabla^d X_{T+h} - \sum_{k=1}^{d} \binom{d}{k} (-1)^k X_{T+h-k} \]  

(31)

Therefore the prediction is in the form

\[ P_T X_{T+h} = P_T (\nabla^d X_{T+h} - \sum_{k=1}^{d} \binom{d}{k} (-1)^k X_{T+h-k}) \]  

(32)

using Property 2

\[ P_T X_{T+h} = P_T \nabla^d X_{T+h} - \sum_{k=1}^{d} \binom{d}{k} (-1)^k P_T X_{T+h-k} \]  

(33)

Introducing Property 1, In the case \( h \leq d + 1 \) and \( \{X_T, X_{T-1}, \ldots, X_{T+h-d}\} \) is observed, we have

\[
\begin{align*}
P_T X_{T+h} &= P_T \nabla^d X_{T+h} - \sum_{k=1}^{d} \binom{d}{k} (-1)^k P_T X_{T+h-k} \\
&= P_T \nabla^d X_{T+h} - \sum_{k=1}^{h} \binom{d}{k} (-1)^k X_{T+h-k} - \sum_{k=1}^{h-1} \binom{d}{k} (-1)^k P_T X_{T+h-k} \\
&= P_T \nabla^d X_{T+h} - \sum_{k=1}^{d} \binom{d}{k} (-1)^k P_T X_{T+h-k}
\end{align*}
\]  

(34)

In the case \( h > d + 1 \),

\[
\begin{align*}
P_T X_{T+h} &= P_T \nabla^d X_{T+h} - \sum_{k=1}^{d} \binom{d}{k} (-1)^k P_T X_{T+h-k}
\end{align*}
\]  

(35)

Q.E.D.

Lemma 4. Error of \( T + h \) time step \( \text{Err}(X_{T+h}) \) can be represent recursively through error of its differenced process \( \text{Err}(\nabla^d X_h) \)

\[
\text{Err}(X_{T+h}) = \text{Err}(\nabla^d X_{T+h}) - \sum_{k=1}^{\min(h-1,d)} \binom{d}{k} (-1)^k \text{Err}(X_{T+h-k})
\]  

(36)

Proof. When \( h \leq d + 1 \), using the definition of differencing (equation 29):

\[
X_{T+h} = \nabla^d X_{T+h} - \sum_{k=h}^{d} \binom{d}{k} (-1)^k X_{T+h-k} - \sum_{k=1}^{h-1} \binom{d}{k} (-1)^k X_{T+h-k}
\]  

(37)
Then using Lemma 3, the error \( \text{Err}(X_{T+h}) \) is:

\[
\text{Err}(X_{T+h}) = \bar{P}(X_{T+h}) - X_{T+h}
\]

\[
= \bar{P}(\nabla^d X_{T+h}) - \sum_{k=1}^{d} \binom{d}{k} (-1)^k X_{T+h-k}
\]

\[
- \sum_{k=1}^{h-1} \binom{d}{k} (-1)^k \bar{P}(X_{T+h-k}) - X_{T+h}
\]

\[
= \bar{P}(\nabla^d X_{T+h}) - \sum_{k=1}^{d} \binom{d}{k} (-1)^k X_{T+h-k}
\]

\[
- \sum_{k=1}^{h-1} \binom{d}{k} (-1)^k \bar{P}(X_{T+h-k}) - \nabla^d X_{T+h}
\]

\[
+ \sum_{k=h}^{d} \binom{d}{k} (-1)^k X_{T+h-k} + \sum_{k=1}^{h-1} \binom{d}{k} (-1)^k X_{T+h-k}
\]

\[
= \text{Err}(\nabla^d X_{T+h}) - \sum_{k=1}^{h-1} \binom{d}{k} (-1)^k \text{Err}(X_{T+h-k})
\]

For \( h > d + 1 \), similarly, using equation 29:

\[
X_{T+h} = \nabla^d X_{T+h} - \sum_{k=1}^{d} \binom{d}{k} (-1)^k X_{T+h-k}
\]

Then from Theorem 3, the error \( \text{Err}(X_{T+h}) \) is:

\[
\text{Err}(X_{T+h}) = \bar{P}(X_{T+h}) - X_{T+h}
\]

\[
= \bar{P}(\nabla^d X_{T+h}) - \sum_{k=1}^{d} \binom{d}{k} (-1)^k \bar{P}(X_{T+h-k}) - X_{T+h}
\]

\[
= \bar{P}(\nabla^d X_{T+h}) - \sum_{k=1}^{d} \binom{d}{k} (-1)^k \bar{P}(X_{T+h-k})
\]

\[
- \nabla^d X_{T+h} + \sum_{k=1}^{d} \binom{d}{k} (-1)^k X_{T+h-k}
\]

\[
= \text{Err}(\nabla^d X_{T+h}) - \sum_{k=1}^{d} \binom{d}{k} (-1)^k \text{Err}(X_{T+h-k})
\]

Q.E.D.

**Theorem 4.** The joint probability distribution among errors of different time steps is a multivariate normal distribution

\[
\text{Err}(X_h) \sim N_h(0, \Sigma_{\text{Err}X})
\]
where
\[ \Sigma_{Err} X = C_{Err XtoErr X} \Sigma_{Err \nabla^d X} C_{Err XtoErr X}^T \]  
(42)

\[ \Sigma_{Err \nabla^d X} \] is the covariance matrix of \( Err(\nabla^d X_h) \) given by Theorem 2 and

\[ C_{Err XtoErr X} = \begin{pmatrix}
1 & 0 & 0 & \ldots & 0 \\
T_{21} & 1 & 0 & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
T_{h1} & T_{h2} & T_{h3} & \ldots & 1
\end{pmatrix} \]  
(43)

\( T_{ij} \) is the coefficient of \( Err(\nabla^d X_{T+i}) \) when calculating \( Err(X_{T+j}) \).

We give \( T_{ij} \) recursively by:

For any \((i \geq 2)\),
\[ T_{i,1} = \sum_{k=1}^{\min(h-1,d)} \binom{d}{k} (-1)^{k+1} T_{i-k,1} \]  
(44)

For any \((j \geq 2)\),
\[ T_{i,j} = T_{i-1,j-1} \]  
(45)

**Proof.** From Lemma 4

\[ \begin{pmatrix} Err(X_{T+1}) \\ Err(X_{T+2}) \\ \vdots \\ Err(X_{T+h}) \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\
T_{21} & 1 & 0 & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
T_{h1} & T_{h2} & T_{h3} & \ldots & 1 \end{pmatrix} \begin{pmatrix} Err(\nabla^d X_{T+1}) \\ Err(\nabla^d X_{T+2}) \\ \vdots \\ Err(\nabla^d X_{T+h}) \end{pmatrix} \]  
(46)

\( T_{ij} \) is the coefficient of \( Err(\nabla^d X_{T+i}) \) when calculating \( Err(X_{T+j}) \).

We give \( T_{ij} \) recursively by:

For any \((i \geq 2)\),
\[ T_{i,1} = \sum_{k=1}^{\min(h-1,d)} \binom{d}{k} (-1)^{k+1} T_{i-k,1} \]  
(47)

For any \((j \geq 2)\),
\[ T_{i,j} = T_{i-1,j-1} \]  
(48)

Denoting
\[ Err(\nabla^d X_h) := \begin{pmatrix} Err(\nabla^d X_{T+1}) \\ Err(\nabla^d X_{T+2}) \\ \vdots \\ Err(\nabla^d X_{T+h}) \end{pmatrix} \]

\[ C_{Err \nabla^d XtoErr X} := \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\
T_{21} & 1 & 0 & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
T_{h1} & T_{h2} & T_{h3} & \ldots & 1 \end{pmatrix} \]
\[
Err(X_h) := \begin{pmatrix}
Err(X_{T+1}) \\
Err(X_{T+2}) \\
\vdots \\
Err(X_{T+h})
\end{pmatrix}
\]

Following Definition 3, \( Err(X_h) \sim N_h(0, \Sigma_{ErrX}) \) where
\[
\Sigma_{ErrX} = C_{Err\nabla XtoErrX} \Sigma_{ErrX} C_{Err\nabla XtoErrX}^T
\]

Q.E.D.

Till now we obtained the joint guarantee of any prediction interval of a series of time steps \( \{X_{T+1}, X_{T+2}, \ldots, X_{T+h}\} \). Conclusion of Theorem 4 can be used to deal with intrinsically stationary processes.

3 Error of Prediction in Multivariate Process

Previous sections discussed univariate time series, now we want to generalize our conclusion to multivariate cases. Similarly, we look at stationary cases firstly.

3.1 Error in Stationary Multivariate Process

**Definition 6.** The best linear prediction of \( X_{T+h} \) is:
\[
P(X_{T+h} | X_T, X_{T-1}, \ldots, X_1, \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}) = A^h_0 + \sum_{i=1}^{T} A^h_i X_{T+1-i}
\]  

\( \{X_t\} \) is a time series with \( n \) variate, \( A^h_0 \) is a \( n \times 1 \) vector and \( A^h_i \) are \( n \times n \) matrices. The optimized choice of \( A^h_i \) are
\[
\sum_{j=1}^{T} A^h_i \Gamma(i-j) = \Gamma(i+h) \quad i = 1, 2, \ldots, T
\]  

**Definition 7.** Denoting \( Z_i := [Z_{ij}]_{j=1}^{h} \), where \( Z_i \) is the \( [n \times n] \) white noise of time step \( i \).
\[
Z_i \sim N_{nt}(0, \Sigma_Z)
\]  

where
\[
\Sigma_Z = \sigma^2 E_{nt \times nt}
\]

**Lemma 5.** Defining a multivariate MA process \( \{Y_t\} \)
\[
Y_t = \Phi(B)X_t \quad t > \max(p, q)
\]  

where \( \Phi_i \) is a \( n \times n \) matrix.
Denoting $Y_h := [Y_{T+1}]_{i=0}^{h}$, then

\[
Y_h = C_1 \begin{pmatrix} Z_{T-q+1} \\ Z_{T-q+2} \\ \vdots \\ Z_{T+h} \end{pmatrix}
\]

(55)

$C_1$ is a $[nh \times n(q+h)]$ transform matrix from $\begin{pmatrix} Z_{T-q+1} \\ Z_{T-q+2} \\ \vdots \\ Z_{T+h} \end{pmatrix}$ to $Y_h$.

\[
C_1 = \begin{pmatrix} \Theta_q & \Theta_{q-1} & \ldots & \Theta_1 & E_{n \times n} & O_{n \times n} & O_{n \times n} & \ldots & O_{n \times n} \\ O_{n \times n} & \Theta_q & \Theta_{q-1} & \ldots & \Theta_1 & E_{n \times n} & O_{n \times n} & \ldots & O_{n \times n} \\ O_{n \times n} & O_{n \times n} & \Theta_q & \Theta_{q-1} & \ldots & \Theta_1 & E & \ldots & O_{n \times n} \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ O_{n \times n} & O_{n \times n} & O_{n \times n} & \ldots & \Theta_q & \Theta_{q-1} & \ldots & \Theta_1 & E \end{pmatrix}
\]

(56)

Proof. From the definition of multivariate ARMA(p,q) process,

\[
Y_t = \Theta(B)Z_t \quad t \geq p + 1
\]

\[
= Y_t = Z_t + \Theta_1 Z_{t-1} + \Theta_2 Z_{t-2} + \cdots + \Theta_q Z_{t-q}
\]

(57)

Q.E.D.

Lemma 6. The best linear predictor of $Y_h$ is a linear combination of $\begin{pmatrix} Z_{p-q+1} \\ Z_{p-q+2} \\ \vdots \\ Z_T \end{pmatrix}$.

\[
P_T Y_h = C_3 C_2 \begin{pmatrix} Z_{p-q+1} \\ Z_{p-q+2} \\ \vdots \\ Z_T \end{pmatrix}
\]

(58)
Proof.
Denoting \( Y \) where \( Y \) steps of
The joint probability distribution among errors of different time
Theorem 5. 

\[
\begin{pmatrix}
Y_{p+1} \\
Y_{p+2} \\
\vdots \\
Y_T
\end{pmatrix}
\]

\[
C_2 = \begin{pmatrix}
\Theta_q & \Theta_{q-1} & \cdots & \Theta_1 & E_{n \times n} & O_{n \times n} & \cdots & O_{n \times n} \\
O_{n \times n} & \Theta_q & \Theta_{q-1} & \cdots & \Theta_1 & E_{n \times n} & \cdots & O_{n \times n} \\
O_{n \times n} & O_{n \times n} & \Theta_q & \Theta_{q-1} & \cdots & \Theta_1 & \cdots & O_{n \times n} \\
\vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\
O_{n \times n} & O_{n \times n} & \cdots & \Theta_q & \Theta_{q-1} & \cdots & \Theta_1 & E_{n \times n}
\end{pmatrix}
\]

(59)

\( C_3 \) is the coefficient matrix of \( (Y_{p+1}, Y_{p+2}, \ldots, Y_T) \) while calculating \( P_T Y_h \).
\( P_T Y_h \) is the best linear predictor of \( \{Y_{T+1}, Y_{T+2}, \ldots, Y_{T+h}\} \), where \( \{Y_t\} \) is mul-
tivariate stationary MA(q) process with zero mean.

\[
C_3 = \begin{pmatrix}
A_{T-p}^1 & A_{T-p-1}^1 & \cdots & A_1^1 \\
A_{T-p}^2 & A_{T-p-1}^2 & \cdots & A_1^2 \\
\vdots & \vdots & \ddots & \vdots \\
A_{T-p}^h & A_{T-p-1}^h & \cdots & A_1^h
\end{pmatrix}
\]

(60)

\( A_t^i \) will be given by equation \( 51 \).

Theorem 5. The joint probability distribution among errors of different time
steps of \( Y_h \) is a multivariate distribution

\[
\text{Err}(Y_h) \sim N_{nh}(0, \Sigma_{\text{Err}Y})
\]

where

\[
\Sigma_{\text{Err}Y} = C_{ZtoErrY} \Sigma_Z C_{ZtoErrY}^T
\]

\[
= \sigma^2 C_{ZtoErrY} E_{n(T+h-p+q) \times n(T+h-p+q)} C_{ZtoErrY}^T
\]

(61)

Proof. Denoting \( X_h := [X_{T+i}]_{i=1}^h \), \( P_T X_h := [P_T X_{T+i}]_{i=1}^h \), \( \text{Err}(X_h) := [\text{Err}(X_{T+i})]_{i=1}^h \),
\( Y_h := [Y_{T+i}]_{i=1}^h \), \( P_T Y_h := [P_T Y_{T+i}]_{i=1}^h \), \( \text{Err}(Y_h) := [\text{Err}(Y_{T+i})]_{i=1}^h \)

Error of \( Y_{T+h} \) is defined as

\[
\text{Err}(Y_{T+h}) = P_T Y_{T+h} - Y_{T+h}
\]

(62)

From Lemma 5 and Lemma 6

\[
\text{Err}(Y_h) = C_3 C_2 \begin{pmatrix}
Z_{p-q+1} \\
Z_{p-q+2} \\
\vdots \\
Z_T
\end{pmatrix} - C_1 \begin{pmatrix}
Z_{T-q+1} \\
Z_{T-q+2} \\
\vdots \\
Z_{T+h}
\end{pmatrix}
\]

(63)
Augment the coefficient matrices $C_1$ and $C_2$ to $C_1^*$ and $C_2^*$ by

$$C_1^* = \begin{pmatrix} O_{nh \times n(T-p)} & C_1 \\ C_2 & O_{n(T-p) \times nh} \end{pmatrix}$$

Thus equation (63) is transformed to

$$\text{Err}(Y_h) = C_3 C_2^* - C_1^*$$

(65)

Denoting $C_{ZtoErrY} := C_3 C_2^* - C_1^*$. Follow Definition (6)

$$\text{Err}(Y_h) \sim N_{nh}(0, \Sigma_{ErrY})$$

$$\Sigma_{ErrY} = C_{ZtoErrY} \Sigma_Z C_{ZtoErrY}^T$$

(66)

**Q.E.D.**

**Theorem 6.** Given $\{X_t\}$ is a multivariate ARMA(p,q) process with n variables, using the best linear predictor $P_T X_{T+h}$, joint probability distribution among errors of different time steps is a multivariate normal distribution

$$\text{Err}(X_h) \sim N_{nh}(0, \Sigma_{ErrX})$$

$$\Sigma_{ErrX} = C_{ErrYtoErrX} C_{ZtoErrX} \Sigma_Z C_{ZtoErrY}^T C_{ErrYtoErrX}^T$$

$$= \sigma^2 C_{ErrYtoErrX} C_{ZtoErrX} \Sigma_Z C_{ZtoErrY} C_{ErrYtoErrX}^T$$

(67)

where

$$C_{ErrYtoErrX} = \begin{pmatrix} E_{n \times n} & O_{n \times n} & O_{n \times n} & \cdots & O_{n \times n} \\ C_{21} & E_{n \times n} & O_{n \times n} & \cdots & O_{n \times n} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ C_{h1} & C_{h2} & C_{h3} & \cdots & E_{n \times n} \end{pmatrix}$$

(68)

$C_{ij}$ is the coefficient matrix of $\sigma \text{Err}(Y_{T+i})$ when calculating $\text{Err}(X_{T+j})$. We give $C_{ij}$ recursively by: For any $(i \geq 2)$,

$$C_{i,1} = \sum_{k=1}^{\min(p,h-1)} \Phi_k C_{i-k,1}$$

(69)

For any $(j \geq 2)$,

$$C_{i,j} = C_{i-1,j-1}$$

(70)
Proof. From the definition of $Y_t$, equation 54, we have

$$X_{T+h} = Y_{T+h} + \sum_{i=1}^{p} \Phi_i X_{T+h-i}$$  \hspace{1cm} (71)$$

The best linear prediction of $X_{T+h}$

$$P_T X_{T+h} = P_T (Y_{T+h} + \sum_{i=1}^{p} \Phi_i X_{T+h-i})$$  \hspace{1cm} (72)$$

From Property 2 we have:

$$P_T X_{T+h} = P_T Y_{T+h} + \sum_{i=1}^{p} \Phi_i P_T X_{T+h-i}$$  \hspace{1cm} (73)$$

Introducing Property 1:

$$P_T X_{T+h} =
\begin{cases}
  P_T Y_{T+h} + \sum_{i=1}^{h-1} \Phi_i P_T X_{T+h-i} + \sum_{i=h}^{p} \Phi_i X_{T+h-i} & h < p \\
  P_T Y_{T+h} + \sum_{i=1}^{p} \Phi_i P_T X_{T+h-i} & h \leq p
\end{cases}$$  \hspace{1cm} (74)$$

According to equation 71

$$\text{Err}(X_{T+h}) = P_T X_{T+h} - X_{T+h}$$

$$= P_T Y_{T+h} + \sum_{i=1}^{h-1} \Phi_i P_T X_{T+h-i} + \sum_{i=h}^{p} \Phi_i X_{T+h-i}$$

$$- Y_{T+h} - \sum_{i=1}^{p} \Phi_i X_{T+h-i}$$

$$= \text{Err}(Y_{T+h}) + \sum_{i=1}^{\min(p,(h-1))} \Phi_i \text{Err}(X_{T+h-i})$$  \hspace{1cm} (75)$$

represented in matrix form

$$\text{Err}(X_h) =
\begin{pmatrix}
  E_{n\times n} & O_{n\times n} & O_{n\times n} & \cdots & O_{n\times n} \\
  C_{21} & E_{n\times n} & O_{n\times n} & \cdots & O_{n\times n} \\
  \vdots & \vdots & \ddots & \ddots & \vdots \\
  C_{h1} & C_{h2} & C_{h3} & \cdots & E_{n\times n}
\end{pmatrix}
\text{Err}(Y_h)$$  \hspace{1cm} (76)$$

$C_{ij}$ is the coefficient matrix of $\sigma \text{Err}(Y_{T+i})$ when calculating $\text{Err}(X_{T+j})$.

We give $C_{ij}$ recursively by:

For any $(i \geq 2)$,

$$C_{i,1} = \sum_{k=1}^{\min(p,h-1)} \Phi_k C_{i-k,1}$$  \hspace{1cm} (77)$$
For any \((j \geq 2)\),
\[ C_{i,j} = C_{i-1,j-1} \]  
(78)

Denoting
\[ C_{\text{Err}Y \to \text{Err}X} := \begin{pmatrix} E_{n \times n} & O_{n \times n} & \cdots & O_{n \times n} \\ C_{21} & E_{n \times n} & \cdots & O_{n \times n} \\ \vdots & \vdots & \ddots & \vdots \\ C_{h1} & C_{h2} & C_{h3} & \cdots E_{n \times n} \end{pmatrix} \]  
(79)

From Definition 3, \( \text{Err}(X_h) \sim N_{nh}(0, \Sigma_{\text{Err}X}) \), where
\[ \Sigma_{\text{Err}X} = C_{\text{Err}Y \to \text{Err}X} \Sigma_{\text{Err}Y} C_{\text{Err}Y \to \text{Err}X}^T \]
\[ = C_{\text{Err}Y \to \text{Err}X} C_{\text{Err}Y \to \text{Err}X} \Sigma_{\text{Err}Y} C_{\text{Err}Y \to \text{Err}X}^T \]
(80)

Q.E.D.

With Theorem 6, we can calculate the joint probability of any prediction interval of any certain series of time steps \( \{X_{T+1}, X_{T+2}, \ldots, X_{T+h}\} \), where \( \{X_t\} \) is a multivariate stationary process.

### 3.2 Error of Multivariate Non-Stationary Process

**Definition 8.** The definition of a \(d\)-ordered multivariate intrinsically stationary process is exactly the same as univariate one. And the differencing method is also:

\[ \nabla^d X_t = \sum_{k=0}^d \binom{d}{k} (-1)^k X_{t-k} \]  
(81)

**Lemma 7.** For any certain series \( \{X_t^s\} \), we have:

\[ \nabla^d X_t^s = \sum_{k=0}^d \binom{d}{k} (-1)^k X_{t-k}^s \]  
(82)

**Theorem 7.** The joint probability distribution of
\[ \begin{pmatrix} \text{Err}(X_{T+1}) \\ \text{Err}(X_{T+2}) \\ \vdots \\ \text{Err}(X_{T+h}) \end{pmatrix} \]
is a multivariate normal distribution
\[ \text{Err}(X_h) \sim N_{nh}(0, \Sigma_{\text{Err}X}) \]
\[ \Sigma_{\text{Err}X} = C_{\text{Err}Y \to \text{Err}X} \Sigma_{\text{Err}Y} C_{\text{Err}Y \to \text{Err}X}^T \]  
(83)
where $\Sigma_{\nabla h}$ is the covariance matrix of errors of the differenced stationary multivariate process given by Theorem 2 and

$$C_{\text{Err} \nabla X \to \text{Err} X} = \begin{pmatrix} E_{n \times n} & O_{n \times n} & O_{n \times n} & \cdots & O_{n \times n} \\ T_{21}E_{n \times n} & E_{n \times n} & O_{n \times n} & \cdots & O_{n \times n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ T_{h1}E_{n \times n} & T_{h2}E_{n \times n} & T_{h3}E_{n \times n} & \cdots & E_{n \times n} \end{pmatrix}$$

(84)

$T_{ij}$ can be recursively given by:

For any ($i \geq 2$),

$$T_{i,1} = \sum_{k=1}^{h-1} \binom{d}{k} (-1)^{k+1} T_{i-k,1}$$

(85)

For any ($j \geq 2$),

$$T_{i,j} = T_{i-1,j-1}$$

(86)

Now we obtain joint guarantee of any prediction interval of a certain series of time steps $\{X_{T+1}^x, X_{T+2}^x, \ldots, X_{T+h}^x\}$, here $\{X_t\}$ is a multivariate intrinsically stationary process.

4 Conclusion

We use the property of both prediction and modeling of ground truth value are linear combination of white noise, which is the assumption of ARIMA model, to prove that the prediction error: difference between ground truth value and predicted value, are also linear combination of white noise, thus prediction errors among different time steps is eligible to form a multivariate distribution and the according expression for different scenarios is in Theorem 2, 4, 6, 7.

References

1. Peter J. Brockwell and Richard A. Davis. Time series: Theory and methods. Springer-Verlag New York, 1987.
2. Peter J. Brockwell and Richard A. Davis. Introduction to time series and forecasting. Springer, Cham, 2016.