STA 542: Introduction to Time Series Analysis

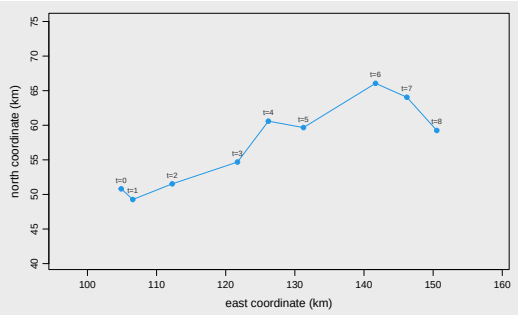
Lecture 3

Lasse Vuursteen

Duke
UNIVERSITY

Mon Aug 31, 2026

Where was the balloon before radar contact?

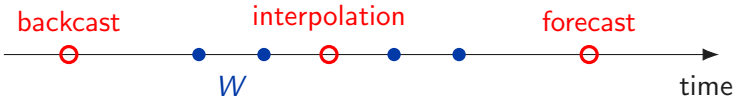


- Radar reports a location every ten minutes from first contact at $t = 0$ through $t = 8$.
- Suppose the route model and all its parameters are known. What can these readings tell us about the hidden position at $t = -6$?

Simulated radar readings; seed 542.

The target need not lie in the future

Let $W = (X_{t_1}, \ldots, X_{t_m})^\top$ be the observed window and let $Y = X_s$ be the unobserved target.



- A predictor $g(W)$ uses the observed window to describe Y.
- For now, the joint process law is known. Fitting comes later.

The loss decides what “best” means

Suppose that, given the observed window $W = w$,

$$\mathbb{P}(Y = 0 \mid W = w) = 0.9, \qquad \mathbb{P}(Y = 10 \mid W = w) = 0.1.$$

Which point a should we report?

The loss decides what “best” means

Suppose that, given the observed window $W = w$,

$$\mathbb{P}(Y = 0 \mid W = w) = 0.9, \qquad \mathbb{P}(Y = 10 \mid W = w) = 0.1.$$

Which point a should we report?

squared loss:	$0.9a^2 + 0.1(10 - a)^2 = (a - 1)^2 + 9,$	minimum at $a = 1,$
absolute loss:	$0.9\lvert a\rvert + 0.1\lvert 10 - a\rvert,$	minimum at $a = 0.$

The conditional mean and a conditional median solve different prediction problems.

A predictor is a function of the observed window

A loss function $L(a, y)$ assigns a cost when prediction a meets target value y. The population risk of g is

$$R(g) = \mathbb{E}\big[L\{g(W), Y\}\big].$$

- The expectation averages over possible observed windows and their possible targets under the known process law.
- This course uses squared-error loss by default:

$$L(a, y) = (a - y)^2.$$

- Under another loss, another feature of the conditional law may be optimal.

Which function minimizes squared error?

Suppose Y is scalar and $\mathbb{E}[Y^2] < \infty$. Write

$$m(W) := \mathbb{E}[Y \mid W].$$

Conditional on W, compare an arbitrary predictor $g(W)$ with $m(W)$:

$$Y - g(W) = \{Y - m(W)\} + \{m(W) - g(W)\}.$$

Which function minimizes squared error?

Suppose Y is scalar and $\mathbb{E}[Y^2] < \infty$. Write

$$m(W) := \mathbb{E}[Y \mid W].$$

Conditional on W, compare an arbitrary predictor $g(W)$ with $m(W)$:

$$Y - g(W) = \{Y - m(W)\} + \{m(W) - g(W)\}.$$

Which term can the choice of g change?

The risk splits into noise and avoidable error

Squaring the preceding decomposition and conditioning on W gives

$$\mathbb{E}\big[(Y - g(W))^2 \mid W\big] = \underbrace{\mathrm{Var}(Y \mid W)}_{\text{irreducible at } W} + \underbrace{\{m(W) - g(W)\}^2}_{\text{choice of predictor}}.$$

The second term is minimized by $g(W) = m(W)$. Hence

$$\widehat{Y}^{\mathrm{sq}} = \mathbb{E}[Y \mid W]$$

is the population-optimal predictor under squared-error loss.

Equal marginals can imply opposite forecasts

Both processes have $X_t \sim N(0,1)$ at every time:

$$\text{A:}\quad X_t \overset{\text{iid}}{\sim} N(0,1), \qquad \mathrm{Cov}(X_s, X_t) = 0 \quad (s \neq t),$$

$$\text{B:}\quad X_t = Y \text{ for every } t, \quad Y \sim N(0,1), \quad \mathrm{Cov}(X_s, X_t) = 1.$$

Given $W = (X_1, \ldots, X_T)^\top$, what is the squared-loss forecast of X_{T+1}?

Equal marginals can imply opposite forecasts

Both processes have $X_t \sim N(0,1)$ at every time:

$$\text{A:}\quad X_t \overset{\text{iid}}{\sim} N(0,1), \qquad \mathrm{Cov}(X_s, X_t) = 0 \quad (s \neq t),$$
$$\text{B:}\quad X_t = Y \text{ for every } t, \quad Y \sim N(0,1), \quad \mathrm{Cov}(X_s, X_t) = 1.$$

Given $W = (X_1, \ldots, X_T)^\top$, what is the squared-loss forecast of X_{T+1}?

	$\mathbb{E}[X_{T+1} \mid W]$	$\mathrm{Var}(X_{T+1} \mid W)$
A	0	1
B	X_T	0

One-dimensional marginals do not determine a forecast.

A Gaussian law gives the entire predictive distribution

Suppose the target vector Y and observed vector W have joint law

$$\begin{pmatrix} Y \\ W \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_Y \\ \mu_W \end{pmatrix}, \begin{pmatrix} \Sigma_{YY} & \Sigma_{YW} \\ \Sigma_{WY} & \Sigma_{WW} \end{pmatrix} \right],$$

where Σ_{WW} is nonsingular.

The question is no longer only "what is the best point?" We can determine the complete conditional law of Y given W.

Subtract the part of the target carried by the window

Set

$$A := \Sigma_{YW}\Sigma_{WW}^{-1}, \qquad V := (Y - \mu_Y) - A(W - \mu_W).$$

Then

$$\mathrm{Cov}(V, W) = \Sigma_{YW} - A\Sigma_{WW} = 0.$$

- The matrix A chooses the linear combination of the window that removes all covariance with the residual V.
- The pair (V, W) remains jointly Gaussian because it is an affine transformation of (Y, W).

Gaussianity turns orthogonality into independence

Joint Gaussianity and $\text{Cov}(V, W) = 0$ imply that V and W are independent. Moreover,

$$\text{Var}(V) = \Sigma_{YY} - \Sigma_{YW}\Sigma_{WW}^{-1}\Sigma_{WY}.$$

Gaussianity turns orthogonality into independence

Joint Gaussianity and $\mathrm{Cov}(V, W) = 0$ imply that V and W are independent. Moreover,

$$\mathrm{Var}(V) = \Sigma_{YY} - \Sigma_{YW}\Sigma_{WW}^{-1}\Sigma_{WY}.$$

Since $Y = \mu_Y + A(W - \mu_W) + V$,

$$Y \mid W = w \sim N\big(\mu_Y + \Sigma_{YW}\Sigma_{WW}^{-1}(w - \mu_W),\ \Sigma_{YY} - \Sigma_{YW}\Sigma_{WW}^{-1}\Sigma_{WY}\big)\,.$$

The same Gaussian calculation works at any set of times

For $(X_t) \sim \mathrm{GP}(\mu, \Sigma)$, let $W = (X_{t_1}, \ldots, X_{t_m})^\top$, let $\mu_W = (\mu(t_1), \ldots, \mu(t_m))^\top$, and take target X_s. Define

$$\Sigma_W = \big(\Sigma(t_i, t_j)\big)_{i,j=1}^{m}, \qquad \sigma_{sW} = \big(\Sigma(s, t_1), \ldots, \Sigma(s, t_m)\big)^\top.$$

Then

$$\mathbb{E}[X_s \mid W] = \mu(s) + \sigma_{sW}^\top \Sigma_W^{-1}(W - \mu_W),$$
$$\mathrm{Var}(X_s \mid W) = \Sigma(s, s) - \sigma_{sW}^\top \Sigma_W^{-1} \sigma_{sW}.$$

These formulas require Σ_W to be nonsingular. The times need not be ordered, and stationarity is not required.

Second moments determine a different prediction target

A full process law determines $\mathbb{E}[Y \mid W]$. Means and covariances alone generally do not.

Second moments determine a different prediction target

A full process law determines $\mathbb{E}[Y \mid W]$. Means and covariances alone generally do not.

If only second-order structure is specified, restrict attention to affine predictors

$$g(W) = b + a^\top W.$$

The *best linear predictor* minimizes mean squared error within this class. It is a projection target, not automatically a conditional mean.

Weak stationarity makes one linear rule reusable

For a weakly stationary process with mean μ and ACVF γ, write

$$W_{T,p} = \begin{pmatrix} X_T - \mu \\ X_{T-1} - \mu \\ \vdots \\ X_{T-p+1} - \mu \end{pmatrix}, \qquad \Gamma_p = \big(\gamma(i-j)\big)_{i,j=1}^{p},$$

$$\gamma_{p,h} = \big(\gamma(h), \gamma(h+1), \ldots, \gamma(h+p-1)\big)^\top.$$

These are the covariance matrix of $W_{T,p}$ and its covariance vector with X_{T+h}. They depend on lags, not on the forecast origin T.

The linear risk is a quadratic

Insert μ into the error:

$$X_{T+h} - b - a^\top W_{T,p} = (X_{T+h} - \mu) + (\mu - b) - a^\top W_{T,p}.$$

The linear risk is a quadratic

Insert μ into the error:

$$X_{T+h} - b - a^\top W_{T,p} = (X_{T+h} - \mu) + (\mu - b) - a^\top W_{T,p}.$$

Square and take expectation. The cross terms with $\mu - b$ vanish, because $\mathbb{E}[X_{T+h}] = \mu$ and $\mathbb{E}[W_{T,p}] = 0$:

$$\begin{aligned}
&\mathbb{E}\big[(X_{T+h} - b - a^\top W_{T,p})^2\big] \\
&\quad = (\mu - b)^2 + \mathbb{E}[(X_{T+h} - \mu)^2] \\
&\qquad - 2a^\top \mathbb{E}[(X_{T+h} - \mu) W_{T,p}] + a^\top \mathbb{E}[W_{T,p} W_{T,p}^\top] a.
\end{aligned}$$

The linear risk is a quadratic

Insert μ into the error:

$$X_{T+h} - b - a^\top W_{T,p} = (X_{T+h} - \mu) + (\mu - b) - a^\top W_{T,p}.$$

Square and take expectation. The cross terms with $\mu - b$ vanish, because $\mathbb{E}[X_{T+h}] = \mu$ and $\mathbb{E}[W_{T,p}] = 0$:

$$\begin{aligned}
&\mathbb{E}\big[(X_{T+h} - b - a^\top W_{T,p})^2\big] \\
&\quad = (\mu - b)^2 + \mathbb{E}[(X_{T+h} - \mu)^2] \\
&\qquad - 2a^\top \mathbb{E}[(X_{T+h} - \mu) W_{T,p}] + a^\top \mathbb{E}[W_{T,p} W_{T,p}^\top] a.
\end{aligned}$$

Those three second-moment factors are $\gamma(0)$, $\gamma_{p,h}$, and Γ_p, so the risk equals

$$(\mu - b)^2 + \gamma(0) - 2a^\top \gamma_{p,h} + a^\top \Gamma_p a.$$

The normal equations determine the best coefficients

The intercept is minimized at $b = \mu$. Differentiating with respect to a gives the *normal equations* $\Gamma_p a = \gamma_{p,h}$. If Γ_p is nonsingular,

$$a_h = \Gamma_p^{-1}\gamma_{p,h}, \qquad \widehat{X}^{\text{lin}}_{T+h|T} = \mu + a_h^\top W_{T,p}.$$

Its mean squared forecast error is

$$\gamma(0) - \gamma_{p,h}^\top \Gamma_p^{-1}\gamma_{p,h}.$$

Weak stationarity makes this population rule the same at every origin.

For Gaussian processes, the two predictors coincide

For a scalar target Y, when (Y, W) is jointly Gaussian,

$$\mathbb{E}[Y \mid W] = \mu_Y + \Sigma_{YW}\Sigma_{WW}^{-1}(W - \mu_W),$$

which is affine in W.

- The conditional mean is therefore already inside the class of linear predictors.
- The best linear predictor equals the squared-loss optimal predictor.
- Its mean squared error equals the conditional variance, which does not depend on the realized value of W in the Gaussian case.

Only X_T enters the AR(1) linear forecast

For the stationary causal solution of

$$X_t - \mu = \phi(X_{t-1} - \mu) + Z_t, \qquad (Z_t) \sim \mathrm{WN}(0, \sigma^2), \qquad |\phi| < 1.$$

The calculation uses two facts:

- $\gamma(k) = \phi^{|k|}\gamma(0)$;
- $\gamma_{p,h}$ is ϕ^h times the first column of Γ_p, so $a_h = (\phi^h, 0, \ldots, 0)^\top$.

Plugging these into the best-linear-prediction formula gives

$$\widehat{X}^{\mathrm{lin}}_{T+h|T} = \mu + \phi^h(X_T - \mu).$$

Once X_T is included, the coefficients on older observations are zero.

Workbook 3 carries out the calculation; Homework 1 applies it to Gaussian and binary innovation laws.

Future innovations determine the remaining uncertainty

Iterating the AR equation gives

$$X_{T+h} = \underbrace{\mu + \phi^h (X_T - \mu)}_{\text{best linear forecast}} + \underbrace{\sum_{j=0}^{h-1} \phi^j Z_{T+h-j}}_{e_{T,h}\text{: forecast error}} .$$

Under white noise, the terms in $e_{T,h}$ have mean zero and are mutually uncorrelated. Hence

$$s_h^2 = \text{Var}(e_{T,h}) = \sigma^2 \sum_{j=0}^{h-1} \phi^{2j} = \sigma^2 \frac{1 - \phi^{2h}}{1 - \phi^2}.$$

White noise still does not identify a conditional mean

Linearity of conditional expectation gives

$$\mathbb{E}[X_{T+h} \mid X_1, \ldots, X_T] = \mu + \phi^h (X_T - \mu) + \mathbb{E}[e_{T,h} \mid X_1, \ldots, X_T].$$

Does the white-noise assumption force the last term to vanish?

White noise still does not identify a conditional mean

Linearity of conditional expectation gives

$$\mathbb{E}[X_{T+h} \mid X_1, \ldots, X_T] = \mu + \phi^h(X_T - \mu) + \mathbb{E}[e_{T,h} \mid X_1, \ldots, X_T].$$

Does the white-noise assumption force the last term to vanish?

- No. White noise is uncorrelated across time, but it may still be dependent across time.
- If the driving sequence is i.i.d., future innovations are independent of the observed past. Then the conditional mean equals the linear forecast and the conditional variance equals s_h^2.

A Gaussian noise law completes the forecast

If

$$Z_t \overset{\text{iid}}{\sim} N(0, \sigma^2),$$

then the future innovation sum is Gaussian and independent of the observed window. Therefore

$$X_{T+h} \mid X_1, \ldots, X_T \sim N\left(\mu + \phi^h (X_T - \mu),\ \sigma^2 \frac{1 - \phi^{2h}}{1 - \phi^2}\right).$$

As the horizon h grows, the conditional mean approaches μ and the conditional variance approaches the stationary marginal variance $\sigma^2/(1 - \phi^2)$.

What does a 95% prediction interval promise?

Let $I(W)$ be an interval computed from the observed window.

$$\text{conditional coverage:} \quad \mathbb{P}\{Y \in I(W) \mid W\} \geq 1 - \alpha,$$
$$\text{marginal coverage:} \quad \mathbb{P}\{Y \in I(W)\} \geq 1 - \alpha.$$

Conditional coverage is a statement given the observed window. Averaging it over possible windows gives marginal coverage:

$$\mathbb{P}\{Y \in I(W)\} = \mathbb{E}\big[\mathbb{P}\{Y \in I(W) \mid W\}\big].$$

The converse need not hold.

Two moments buy a valid but wide interval

Suppose the conditional mean $m(W)$ and conditional variance $v(W)$ are known. Conditional Chebyshev gives

$$\mathbb{P}(|Y - m(W)| \geq c \mid W) \leq \frac{v(W)}{c^2}.$$

Setting $c = \sqrt{v(W)/\alpha}$ gives conditional coverage at least $1 - \alpha$ for

$$m(W) \pm \sqrt{\frac{v(W)}{\alpha}}.$$

Two moments buy a valid but wide interval

Suppose the conditional mean $m(W)$ and conditional variance $v(W)$ are known. Conditional Chebyshev gives

$$\mathbb{P}(|Y - m(W)| \geq c \mid W) \leq \frac{v(W)}{c^2}.$$

Setting $c = \sqrt{v(W)/\alpha}$ gives conditional coverage at least $1 - \alpha$ for

$$m(W) \pm \sqrt{\frac{v(W)}{\alpha}}.$$

At 95%, the half-width is $\sqrt{20} \approx 4.47$ conditional standard deviations.

The available assumptions determine the interval claim

For a best linear predictor with only marginal mean squared error s_h^2, ordinary Chebyshev gives only the marginal guarantee

$$\widehat{X}^{\mathrm{lin}}_{T+h|T} \ \pm \ \frac{s_h}{\sqrt{\alpha}}.$$

If instead the full conditional law is Gaussian,

$$X_{T+h} \mid W \sim N\{m_h(W), v_h(W)\},$$

then the exact equal-tail conditional interval is

$$m_h(W) \ \pm \ z_{1-\alpha/2}\sqrt{v_h(W)}.$$

At 95%, Gaussian shape replaces 4.47 by 1.96.

A random-walk forecast spreads at rate $\sqrt{h}$

Starting from a fixed X_0, suppose

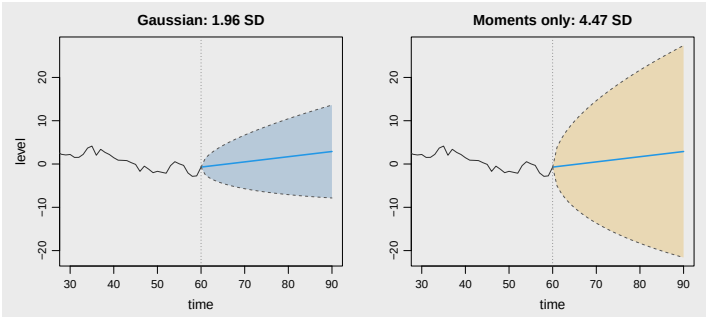
$$X_t = X_{t-1} + \delta + Z_t, \qquad (Z_t) \sim \mathrm{WN}(0, \sigma^2).$$

Then

$$X_{T+h} = X_T + h\delta + \sum_{j=1}^{h} Z_{T+j}.$$

- White noise gives the best linear forecast $X_T + h\delta$ and mean squared error $h\sigma^2$.
- I.i.d. Gaussian noise gives the exact conditional law $N(X_T + h\delta, h\sigma^2)$.
- The level process is nonstationary; stationarity is not needed for this forecast calculation.

The same center can support very different intervals



- Both fans use the same simulated history and forecast center.
- Gaussian shape gives pointwise 95% conditional intervals; white-noise moments alone give the wider pointwise marginal Chebyshev guarantee.

Simulated random walk; $\delta = 0.12$, $\sigma = 1$, seed 542.

A pointwise fan is not a region for the whole path

Pointwise coverage means, for every fixed horizon h,

$$\mathbb{P}\{X_{T+h} \in I_h(W) \mid W\} \geq 1 - \alpha.$$

Does this imply that the entire future path lies inside the fan with probability at least $1 - \alpha$?

A pointwise fan is not a region for the whole path

Pointwise coverage means, for every fixed horizon h,

$$\mathbb{P}\{X_{T+h} \in I_h(W) \mid W\} \geq 1 - \alpha.$$

Does this imply that the entire future path lies inside the fan with probability at least $1 - \alpha$?

$$\mathbb{P}\Big(X_{T+1} \in I_1(W), \ldots, X_{T+H} \in I_H(W) \;\Big|\; W\Big)$$

is a different probability. A plotted collection of pointwise intervals is not a simultaneous path region unless it was constructed as one.

The balloon model separates route and measurement

Return to the radar readings. Let $R_t \in \mathbb{R}^2$ be the hidden position and $O_t \in \mathbb{R}^2$ the recorded position, in kilometers east and north:

$$\underbrace{R_t = b + dt + U_t}_{\text{mean route plus deviation}}, \qquad \underbrace{O_t = R_t + \varepsilon_t}_{\text{radar reading}}.$$

Take $(U_t)_{t\in\mathbb{Z}}$ to be the stationary solution of

$$U_t = \phi U_{t-1} + Z_t, \qquad Z_t \overset{\text{iid}}{\sim} N_2(0, qI_2), \qquad |\phi| < 1,$$

while $\varepsilon_t \overset{\text{iid}}{\sim} N_2(0, rI_2)$. The sequences (Z_t) and (ε_t) are independent, and all parameters are known. These assumptions make the hidden target and radar window jointly Gaussian.

Covariance carries information backward

The stationary deviation variance in either coordinate is $\tau^2 = q/(1-\phi^2)$. For observed times $t_1, \ldots, t_m$, define

$$C_{ij} = \underbrace{\tau^2 \phi^{|t_i - t_j|}}_{\text{route covariance}} + \underbrace{r\mathbf{1}\{i = j\}}_{\text{radar noise}},$$

$$c_s = \tau^2 \begin{pmatrix} \phi^{|s-t_1|} \\ \vdots \\ \phi^{|s-t_m|} \end{pmatrix}.$$

The matrix C describes the observed radar window; c_s connects the hidden position at time s to that window.

Gaussian conditioning gives the backcast

Let row i of O be $O_{t_i}^\top$ and row i of M be $(b + dt_i)^\top$. Then

$$R_s \mid O_{t_1}, \ldots, O_{t_m} \sim N_2\big(m_s, s_s^2 I_2\big),$$

where

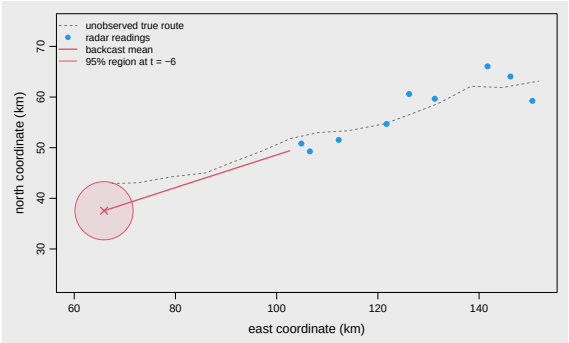
$$m_s = b + ds + (O - M)^\top C^{-1} c_s, \qquad s_s^2 = \tau^2 - c_s^\top C^{-1} c_s.$$

At $s = -6$,

$$m_{-6} = (65.9, 37.6)^\top \text{ km}, \qquad s_{-6}^2 = 5.51 \text{ km}^2.$$

A 95% position region is $\|R_s - m_s\|^2 \leq \chi^2_{2,0.95} s_s^2$.

The readings localize the earlier position



- The pointwise 95% disk at $t = -6$ has radius 5.7 km around the conditional mean.
- The hidden simulated position was $(67.4, 42.9)^\top$ km; it was not used in the backcast.

Simulated route and radar readings; seed 542.

Later readings help because first contact is noisy

Using only O_0 gives conditional variance $6.27 I_2$ km^2. Using all nine readings gives $5.51 I_2$ km^2, a 12.1% reduction.

What happens if the radar error variance is $r = 0$?

Later readings help because first contact is noisy

Using only O_0 gives conditional variance $6.27 I_2$ km^2. Using all nine readings gives $5.51 I_2$ km^2, a 12.1% reduction.

What happens if the radar error variance is $r = 0$?

- Then $O_0 = R_0$: the boundary position is observed exactly.
- Under the Gaussian AR(1) deviation law, readings after $t = 0$ add no information about R_s for $s < 0$ once R_0 is known.
- With noisy first contact, later readings help infer the hidden boundary position and thereby sharpen the backcast.

The normal equations determine the projection

The risk is minimized at $b = \mu$. If Γ_p is nonsingular, differentiating the quadratic risk and substituting the solution gives

$$
\begin{aligned}
\Gamma_p a &= \gamma_{p,h}, && \text{normal equations,} \\
a_h &= \Gamma_p^{-1}\gamma_{p,h}, && \text{best coefficients,} \\
\widehat{X}^{\mathrm{lin}}_{T+h|T} &= \mu + a_h^\top W_{T,p}, && \text{best linear predictor,} \\
\mathrm{MSE} &= \gamma(0) - \gamma_{p,h}^\top \Gamma_p^{-1}\gamma_{p,h}.
\end{aligned}
$$

Weak stationarity makes this population rule the same at every forecast origin.

The normal equations are an orthogonality condition

By the definitions of Γ_p and $\gamma_{p,h}$,

$$\text{Cov}\Big(W_{T,p}, X_{T+h} - \mu - a^\top W_{T,p}\Big) = \gamma_{p,h} - \Gamma_p a.$$

The normal equations are an orthogonality condition

By the definitions of Γ_p and $\gamma_{p,h}$,

$$\mathrm{Cov}\Big(W_{T,p}, X_{T+h} - \mu - a^\top W_{T,p}\Big) = \gamma_{p,h} - \Gamma_p a.$$

Therefore

$$\Gamma_p a = \gamma_{p,h} \quad \Longleftrightarrow \quad \text{the prediction error is uncorrelated with every entry of } W_{T,p}.$$

This is the population counterpart of the familiar least-squares fact that the residual is orthogonal to the predictors.

The AR equation supplies a candidate

Consider the stationary causal solution of

$$X_t - \mu = \phi(X_{t-1} - \mu) + Z_t, \qquad |\phi| < 1.$$

Iterating h steps forward gives

$$X_{T+h} = \underbrace{\mu + \phi^h(X_T - \mu)}_{\text{candidate predictor}} + \underbrace{\sum_{j=0}^{h-1} \phi^j Z_{T+h-j}}_{e_{T,h}\text{: prediction error}} .$$

The AR equation supplies a candidate

Consider the stationary causal solution of

$$X_t - \mu = \phi(X_{t-1} - \mu) + Z_t, \qquad |\phi| < 1.$$

Iterating h steps forward gives

$$X_{T+h} = \underbrace{\mu + \phi^h(X_T - \mu)}_{\text{candidate predictor}} + \underbrace{\sum_{j=0}^{h-1} \phi^j Z_{T+h-j}}_{e_{T,h}\text{: prediction error}} .$$

This identity suggests a predictor; it does not yet prove optimality. The normal-equation criterion asks whether $e_{T,h}$ has mean zero and is uncorrelated with every observed X_s, $s = 1, \ldots, T$.

The observations and the error use different innovations

For every $s \leq T$, causality gives

$$X_s - \mu = \underbrace{\sum_{k=0}^{\infty} \phi^k Z_{s-k}}_{\text{innovations dated at or before } T},$$

$$e_{T,h} = \underbrace{\sum_{j=0}^{h-1} \phi^j Z_{T+h-j}}_{\text{innovations dated } T+1,\ldots,T+h}.$$

The two sets of dates do not overlap. If $(Z_t) \sim \mathrm{WN}(0, \sigma^2)$, innovations at distinct dates are uncorrelated. Hence

$$\mathbb{E}[e_{T,h}] = 0, \qquad \mathrm{Cov}(e_{T,h}, X_s) = 0, \qquad s = 1, \ldots, T.$$

The candidate therefore satisfies the normal equations.

White noise determines the best linear forecast

The orthogonality criterion now gives

$$\widehat{X}^{\text{lin}}_{T+h|T} = \mu + \phi^h(X_T - \mu),$$

$$s_h^2 := \text{Var}(e_{T,h}) = \sigma^2 \sum_{j=0}^{h-1} \phi^{2j} = \sigma^2 \frac{1-\phi^{2h}}{1-\phi^2}.$$

Only means and covariances have been used. White noise does not yet tell us whether this linear forecast is also $\mathbb{E}[X_{T+h} \mid X_1, \ldots, X_T]$.

Next class: parameter uncertainty becomes part of prediction

- Put a prior distribution on an unknown process parameter.
- Update that distribution using the observed path.
- Average the known-parameter forecasts over the posterior to obtain the posterior predictive law.