Wold Decomposition Theorem

Pranav Tikkawar

May 6, 2025

Outline

Introduction

My goal is to frame the Wold Decomposition Theorem.

- What is the Wold Decomposition Theorem and where does it apply?
- What content is required to understand the Wold Decomposition Theorem?
- Why we care about this theorem?

Definition

Simply put, any weakly stationary stochastic process can be expressed as a sum of a deterministic component and a stochastic component which is a Infinite Order Moving Average model.

Time Series Data VS IID Data

- Time Series Data
  - A sequence of data points indexed in time order by Integers
  - Each data point is dependent on the previous one.
- IID Data
  - Independent and identically distributed data.
  - Each data point is independent of the others.

Theorem (Glivenko-Cantelli Theorem)

Suppose $X_1, X_2, \ldots, X_n$ are IID random variables with a common distribution function F. Then the empirical distribution function $F_n(x)$ converges uniformly to $F(x)$ as $n \to \infty$.

However, unlike IID data, no uniform convergence is guaranteed for time series data. due to the fact that the data points are dependent on each other.

Ergodicity

Definition (Almost Sure Ergodic Property with a Constant Limit (EPCL))

$X_t \in \mathbb{R}$ $(t \in \mathbb{Z})$ has this property if $\exists \mu \in \mathbb{R}$ such that $\mathbb{P}(\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} X_t = \mu) = 1$.
In other words the time average converges to a constant almost surely.

Similarly, L^2-EPCL requires convergence in the $L-2$ sense i.e. $\lim_{n\to\infty} E[(\frac{1}{n} \sum_{t=1}^{n} X_t - \mu)^2] = 0$.
These properties are nice, but there are some circumstances where they do not hold like

- Lack of Stability
- High Variance
- Absorbing States

Stationarity

We can use stationarity to help solves these issues that might come with a lack of ergodicity.

Definition (Weak Stationarity)

$X_t \in \mathbb{R}$ $(t \in \mathbb{Z})$ is weakly stationary if:

- $E[X_t^2] < \infty$
- $\exists \mu \in \mathbb{R}$ such that $\forall t \in \mathbb{Z}, E[X_t] = \mu$ for all t.
- $\exists \gamma : \mathbb{Z} \to \mathbb{R}$ such that $\forall s, t \in \mathbb{Z}$, $Cov(X_s, X_t) = \gamma(t - s)$.

Simply put, a weakly stationary process has a constant mean and its covariance only depends on the time difference (lag) between two points, not on the actual time points themselves.

Theorem (Birkhoff's Ergodic Theorem)

If X_t is a L^2 (weakly stationary) process and μ is the mean of the process, then $\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} X_t = \mu$ almost surely.

Hilbert Spaces

Consider $\mathscr{R}$ which is the space of real-valued random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.
Define $L^2(\Omega) = \{X : X \in \mathscr{R} \text{ s.t. } E[X^2] < \infty\}$, which is the space of square integrable random variables.
This space is a Hilbert space with the inner product defined as:

$$\langle X, Y \rangle = E[XY] \quad \text{for } X, Y \in L^2(\Omega)$$

Definition (Projection on subspace)

We can also define the projection of a random variable X onto a subspace $V \subseteq L^2(\Omega)$ as $P_V(X)$ such that $P_V(X)$ is the unique element in V that minimizes the distance to X, i.e.
$\|X - P_V(X)\|^2 = \inf_{Y \in V} \|X - Y\|^2$.

Sigma Algebra and Filtration

Definition (Sigma Algebra)

A σ-algebra $\mathcal{F}$ on a set Ω is a collection of subsets of Ω that is closed under complementation and countable unions.

Definition (Filtration)

A filtration $\{\mathcal{F}_t\}_{t\in\mathbb{Z}}$ is a family of σ-algebras such that $\mathcal{F}_s \subseteq \mathcal{F}_t$ for all $s < t$.

This allows us to define the information available at time t.

Optimal Forecast

Definition (K-step prediction Mean Square Error)

Let $\mathcal{F}_t$ be the σ-algebra generated by the process up to time t. And $\mathscr{X} = \{Y | Y \in L^2(\Omega), \mathcal{F}_t - \text{measureable}\}$ Then for $k \geq 1$ and $Y \in \mathscr{X}$, the k-step prediction mean square error of Y is defined as: $\sigma_k^2(Y) = E[(X_{t+k} - Y)^2]$

The prediction of X_{t+k} given $\mathcal{F}_t$ is optimal if it minimizes the k-step prediction mean square error.

Theorem (Optimal Forecast)

The optimal forecast $(\hat{X}_{t+k})$ *of* X_{t+k} *given* $\mathcal{F}_t$ *is*
$\sigma_k^2(X_{t+k}) = \inf_{Y \in \mathscr{X}} \sigma_k^2(Y)$
Or equivalently $\hat{X}_{t+k} = E[X_{t+k} | \mathcal{F}_t]$.
In other words, the best prediction of X_{t+k} *based on the information available up to time* t *is the conditional expectation of* X_{t+k} *given* $\mathcal{F}_t$.

Theorem

If $E(X)$ *exists then* $E(X|\mathcal{F}_t)$ *exists and is unique.*

Optimal Linear Forecast

Similarly we can consider the optimal linear forecast of X_{t+k} given $\mathcal{F}_t$.

Definition (Linear Past)

The linear past of a process X_t is the closure of the set of all linear combinations of past values of the process, i.e.
$L_t = \{Y | Y \in L^2(\Omega), Y = \sum_{i=0}^{n} a_i X_{t-i}, a_i \in \mathbb{R}, n \in \mathbb{N}\}$.
The infinite linear past is defined as $L_{-\infty} = \bigcap_{t=-\infty}^{\infty} L_t$.

Definition (Optimal Linear Forecast)

The optimal linear forecast of X_{t+k} given $\mathcal{F}_t$ when the follow holds: $\sigma_k^2(X_{t+k}) = \inf_{Y \in L_t} \sigma_k^2(Y)$,
Then we can write $\sigma_{k,\text{opt}}^2 = \sigma_k^2(X_{t+k})$
Note that the optimal linear forecast is essentaly the orthogonal projection of X_{t+k} onto the linear past L_t.

Deterministic Process

Definition (Deterministic)

We say a process X_t (weakly stationary) is deterministic if $\sigma^2_{1,\text{opt}} = 0$

More generally, we say a process Z_t is deterministic with respect to X_t if $\forall t \in \mathbb{Z}$, $\inf_{Y \in L_t} E[(Z_t - Y)^2] = 0$.

This implies that $\sigma^2_{k,\text{opt}} = 0$ for all $k \geq 1$, meaning that the process can be perfectly predicted from its past values.

Note that if Z_t is a deterministic process with respect to X_t then Z_t is in the infinite linear past of X_t.

Now we are ready to define the Wold Decomposition Theorem.

Wold Decomposition Theorem

Theorem (Wold Decomposition Theorem)

Plainly put, any L^2/weakly stationary process $X_t \in \mathbb{R}(t \in \mathbb{Z})$ can be decomposed into a deterministic component and a stochastic component where the stochastic component is an Infinite Order Moving Average (IOMA) process.
More formally:

$$X_t \in \mathbb{R}(t \in \mathbb{Z}) \textit{weakly stationary} \implies$$

$$\exists a_n, \in \mathbb{R}, n \in \mathbb{Z}, a_0 = 1, \sum_{n=-\infty}^{\infty} a_n^2 < \infty$$

$$\exists \epsilon_t, \mu_t (t \in \mathbb{Z}) \forall s, t \in \mathbb{Z} \textit{ such that}$$

$$\epsilon_t \in L_t, \mu_t \in L_{-infty},$$

$$\mathbb{E}(\epsilon_t) = 0, Cov(\epsilon_s, \epsilon_t) = \delta_{s,t}\sigma^2 < \infty, Cov(\epsilon_s, \mu_t) = 0$$

$$X_t = \mu_t + \sum_{j=0}^{\infty} a_j \epsilon_{t-j}$$

Wold Decomposition Theorem Proof

Since X_t is weakly stationary, can write is as the sum of its projection onto the linear past upt to time $t-1$ and an element of the orthogonal complement of the linear past, i.e.

$$X_t = P_{L_{t-1}}(X_t) + \epsilon_t, \text{ where } \epsilon_t \in L_{t-1}^{\perp} \cap L_t$$

$$\sigma_\epsilon^2 = \text{Var}(\epsilon_t) = \sigma_{1,\text{opt}}^2, \text{Cov}(X_s, \epsilon_t) = 0 (s \leq t-1)$$

Next we can define a_j as the coefficients of the linear combination of given by $\frac{\langle X_t, \epsilon_{t-j} \rangle}{\sigma_\epsilon^2}$ and since X_t is weakly stationary, a_j is independent of t.
Now we can consider
$E_t^0 = \{Y \in L_t | Y = \sum_{j=1}^k a_j \epsilon_{t_j}, k \in \mathbb{N}, a_j \in \mathbb{R}, t_j \in \mathbb{Z}, t_j \leq t\}$ and its closure E_t Then we can $\sum_{j=0}^\infty a_j \epsilon_{t-j} \in E_t \subseteq L_t$ and then we can define $\mu_t := X_t - \sum_{j=0}^\infty a_j \epsilon_{t-j} \in L_t$,

Wold Decomposition Theorem Proof (cont.)

Clearly with this we have the the inner product $\langle \mu_t, \epsilon_{t-j} \rangle = 0$ for all j

Then $\forall l \geq 1$: $P_{L_{t-l}}(X_t) = P_{L_{t-l}}(\mu_t) + \sum_{j=0}^{\infty} a_j P_{L_{t-l}}(\epsilon_{t-j})$

$= \mu_t + \sum_{j=l}^{\infty} a_j \epsilon_{t-j} \in L_{t-1}$

Since $\sum_{j=l}^{\infty} a_j \epsilon_{t-j} \in L_{t-1}$

We can write $\forall l \geq 1$: $\mu_t = P_{L_{t-1}}(X_t) - \sum_{j=l}^{\infty} a_j \epsilon_{t-j}$

And thus $\mu_t \in L_{-\infty}$, and deterministic with respect to X_t

Remark: The ϵ_t are only uncorrelated not necessarily independent.

Applications of Wold Decomposition Theorem

The largest implication of the Wold Decomposition Theorem is that it allows us to decompose any weakly stationary process into something that is reliably predictable (the deterministic component) and something that is just pure noise (the stochastic component).
This is useful in many applications such as:

- Economic Forecasting
- Signal Processing
- Weather Prediction

Thanks for Listening!

I would like to thank my graduate mentor Forrest Thurman for his guidance and the DRP program for facilitating these amazing presentations! The book I used for this presentation is *Mathematical Foundations of Time Series Analysis: A Concise Introduction* by Jan Beran.