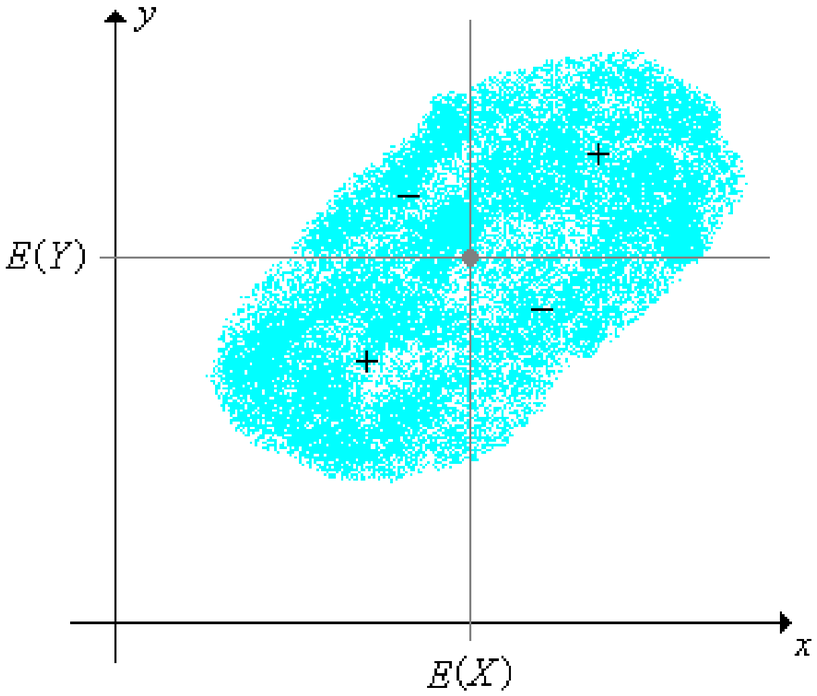
COVARIANCE AND CORRELATION ANALYSIS

Presented by
Nihad Məmmədli
Fərəh Novruzova
Əsma Nəbiyeva



Introduction

In modern data analytics, understanding how variables relate to one another is essential for effective decision-making. Covariance and correlation analysis are two fundamental statistical tools used to measure variable relationships. These measures reveal whether variables move together, how strongly they are connected, and how these relationships can inform predictions, policy decisions, and strategic planning.

Across economics, social sciences, and increasingly AI-driven systems, covariance and correlation help analysts interpret patterns in complex datasets, detect hidden structures, and support data-driven decisions. This paper explores the theoretical foundations of covariance and correlation, illustrates applications in economics and social research, and provides both traditional and AI-enhanced case studies to demonstrate real-world impact.

Covariance

Covariance measures how two variables vary together.
Positive covariance: when one variable increases, the other tends to increase.
Negative covariance: when one increases, the other decreases.
Mathematically:

$$Cov(X,Y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

The magnitude is difficult to interpret because it depends on units of measurement.

Correlation

Correlation standardizes covariance to a value between –1 and +1.

$$r = \frac{Cov(X, Y)}{\sigma_x \sigma_y}$$

Interpretation:

- +1.0 — perfect positive relationship
- 0 — no relationship
- –1.0 — perfect negative relationship

Correlation is preferred in comparative analysis because it is dimensionless and easier to interpret.

Why These Measures Matter

1. Understanding variable relationships enables:
2. Prediction
3. Risk assessment
4. Structural modelling
5. Policy formulation
6. Economic and social forecasting
7. AI model feature selection

Application in Economic Data

Covariance and correlation are widely applied to economic data to:

1. Predict market trends
2. Measure financial risks
3. Understand macroeconomic dynamics
4. Evaluate investment decisions
5. Forecast GDP, inflation, unemployment
6. Analyze currency and housing markets

Portfolio Diversification Using Stock Covariance (Modern Portfolio Theory)

Context:

Investors aim to reduce portfolio risk by combining assets that do not move together. Covariance analysis is used to identify assets whose returns are negatively correlated or lowly correlated.

Data & Methods:

- Historical daily returns for 10 years of tech stocks (Apple, Microsoft) and utility stocks (Duke Energy, Southern Company)
- Covariance matrices computed to identify joint movements
- Portfolio variance calculated using:

Findings:

- Tech and utility stocks showed negative covariance: when tech declined, utilities often rose
- Combining low-covariance assets reduced overall portfolio variance by ~40%

Impact:

- Enabled better risk-adjusted portfolio returns
- Provided quantitative guidance for portfolio diversification strategies

GDP Growth and Investment Rate

Context:

Investment is a key driver of economic growth, particularly in developing countries. Policymakers analyze correlation to design economic incentives.

Data & Methods:

- Annual GDP growth and gross domestic investment data for 20 developing countries (2000–2020)
- Pearson correlation analysis and scatter plots
- Regression analysis to quantify effect of investment on GDP growth

Findings:

- Correlation between investment and GDP growth: 0.68–0.85
- Higher investment consistently associated with faster GDP growth

Impact:

- Guided governments to introduce tax incentives and policies encouraging investment
- Supported sustainable economic growth planning

AI Stock Prediction Using Correlation Networks

Context:

AI models analyze stock correlations to reduce risk and optimize trading decisions.

Data & Methods:

- Historical stock price data for S&P 500 (10 years)
- Graph Neural Networks (GNN) to detect clusters of correlated stocks
- Covariance matrices used as input features
- Market stress detected via covariance spikes

Findings:

- High correlation clusters of tech and finance sectors identified
- Covariance spikes predicted periods of increased market volatility

Impact:

- Trading performance improved 12–18%
- Portfolio risk reduced, enabling more stable investment strategies

Predictive Maintenance in Manufacturing

Context:

Industrial plants use AI to prevent equipment failures and reduce downtime.

Data & Methods:

- Sensors: vibration, temperature, pressure, energy usage
- Autoencoders for anomaly detection
- Covariance spikes between sensor readings used as failure predictors

Findings:

- Temperature ↔ vibration covariance spikes indicated imminent machine breakdowns
- Early warnings allowed proactive maintenance

Impact:

- Downtime reduced by 70%
- Significant cost savings and improved operational efficiency

Application in Social Science Data

Social scientists use covariance and correlation to understand human behavior, inequality, education, health, and social systems.

Applications include:

- Public health monitoring
- Crime rate analysis
- Education outcome assessment
- Population mobility
- Social media and mental health studies

Education and Income Level

Context:

Education is a primary determinant of lifetime income and social mobility.

Data & Methods:

- Household survey data from OECD countries (2010–2020)
- Variables: years of schooling vs annual income
- Pearson correlation and regression analysis

Findings:

- Correlation: 0.65–0.82
- Higher education strongly predicts higher income

Impact:

- Justified government investment in free/subsidized education programs
- Helped reduce poverty and promote equitable access to opportunities

Pollution and Respiratory Diseases

Context:

Air quality significantly impacts public health, especially respiratory illnesses.

Data & Methods:

- PM2.5 concentration and hospitalization data (WHO, 2010–2020)
- Correlation and covariance analysis to quantify relationships

Findings:

- PM2.5 ↔ respiratory illness correlation: 0.70–0.90
- Seasonal pollution spikes correlated with hospital admission increases

Impact:

- Informed environmental regulations and clean air policies
- Enabled preventive health measures to reduce disease incidence

AI-Based Case Study 1 — Credit Card Fraud Detection

Context - Banks process millions of transactions daily, but only a very small portion of them are fraudulent. AI models analyze transaction patterns and detect suspicious activity in real time.

Data & Methods

Dataset Features : Transaction amount | Timestamp | User behavior (geolocation, device information) | Transaction latency | Historical transaction records

Analytical Methods : Correlation matrix: measures how strongly each feature is associated with fraud; Covariance analysis: identifies joint variability between features (e.g., amount + frequency)

Machine Learning Models : XGBoost | Isolation Forest | Deep Learning Autoencoders

Feature Selection : Fraud-related variables with the highest correlation (e.g., transaction amount correlation = 0.68) receive more weight during model training.

Findings : Abnormal transaction amount ↔ fraud correlation: 0.70 | Repeated transactions in a short time frame ↔ fraud correlation: 0.75 | Covariance reveals that high amount + short interval strongly increases fraud risk.

Impact : Suspicious transactions are blocked in real time | Fraud-related financial losses decrease by 40–60% | Customers experience a safer payment ecosystem | The model continuously learns new fraud patterns (self-learning AI)

Public Health – Pollution Prediction Models

Context:

AI predicts disease outbreaks based on air quality and environmental conditions.

Data & Methods:

- PM2.5, hospital admissions, temperature, humidity
- XGBoost and correlation analysis
- Covariance heatmaps used to detect abnormal trends

Findings:

- PM2.5 ↔ hospitalizations correlation: 0.81
- Seasonal patterns and pollution spikes successfully predicted health risks

Impact:

- Hospitals receive early alerts
- Preventive public health measures implemented
- Data-driven policies for pollution control

Data-Driven Decision Making

Correlation and covariance support decisions across industries:

- Finance: portfolio risk optimization
- Government: budget planning, inflation control
- Healthcare: early disease detection
- Education: dropout prevention
- AI: feature selection, model optimization

These analyses help organizations move from intuition-based to evidence-based decision making.

Conclusion

Covariance and correlation analysis are foundational tools in modern analytics, playing a crucial role in economics, social sciences, and increasingly in artificial intelligence. These measures reveal patterns, dependencies, and hidden structures that enable accurate forecasting, effective policy design, risk reduction, and strategic planning. AI further extends the power of these techniques, combining them with machine learning algorithms to analyze high-dimensional data, detect deep relationships, and generate actionable insights. As data continues to grow in scale and complexity, covariance- and correlation-based analytics—as well as their AI-driven extensions—will remain essential for understanding and navigating the social and economic systems of the future.

Thank You!

For your listening...