The Gaussian Distribution: Universality and Applications

March 27, 2025

Outline

The Univariate Gaussian Distribution

Definition

A random variable X follows a Gaussian (or normal) distribution with mean μ and variance σ^2, denoted $X \sim \mathcal{N}(\mu, \sigma^2)$, if its probability density function (PDF) is:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}$$

- The standard normal distribution: $Z \sim \mathcal{N}(0, 1)$
- Bell-shaped, symmetric curve with inflection points at $\mu \pm \sigma$
- 68%-95%-99.7% rule: Probability mass within 1, 2, and 3 standard deviations

Key Properties

- **Mean equals mode equals median** $= \mu$
- **Variance** $= \sigma^2$, **Standard deviation** $= \sigma$
- **Moment generating function:** $M_X(t) = \exp(\mu t + \frac{\sigma^2 t^2}{2})$
- **Standardization:** If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$
- **Linear transformations:** If $Y = aX + b$, then $Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$

The Central Limit Theorem

Statement of the Theorem

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables with mean μ and finite variance σ^2. Define the sample mean:

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Then, as $n \to \infty$:

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1)$$

The $\sqrt{n}$ Denominator in the CLT: Variance Calculation

Variance of the Sample Mean

$$\begin{aligned}
\mathrm{Var}(\bar{X}_n) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) \\
&= \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}
\end{aligned}$$

- The variance of the sample mean decreases as sample size increases
- This means the standard deviation is $\sigma/\sqrt{n}$
- This $1/\sqrt{n}$ rate of decrease is crucial for the CLT statement

The $\sqrt{n}$ Denominator in the CLT: Significance

Why the $\sqrt{n}$ Scaling Matters

- Standard deviation of $\bar{X}_n$ is $\sigma/\sqrt{n}$ (shrinks with sample size)
- Dividing by $\sigma/\sqrt{n}$ in the CLT normalizes this decreasing variance
- $\sqrt{n}$ scaling "magnifies" deviations at exactly the right rate

Mathematical Significance

- Without scaling: $\bar{X}_n - \mu \xrightarrow{p} 0$ (Law of Large Numbers)
- With scaling: reveals the Gaussian limiting behavior
- Enables construction of confidence intervals using normal approximation
- For large n: $\bar{X}_n \approx \mathcal{N}(\mu, \sigma^2/n)$

Importance of the Central Limit Theorem

- Explains why many natural phenomena follow a Gaussian distribution
- Provides justification for statistical methods that assume normality
- Applies regardless of the original distribution (with finite variance)
- Convergence rate depends on the original distribution

Example

The distribution of heights in a population results from many small, independent genetic and environmental factors.

The Maximum Entropy Principle

Entropy

For a continuous random variable with PDF $f(x)$, the differential entropy is:

$$H[f] = -\int f(x) \log f(x)\, dx$$

Maximum Entropy Theorem

Among all continuous probability distributions on $\mathbb{R}$ with a specified mean μ and variance σ^2, the Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ has the maximum entropy.

- The Gaussian is the "least informative" distribution consistent with the given constraints
- It makes the minimum assumptions beyond the specified mean and variance

The Multivariate Gaussian Distribution

Definition

A random vector $\mathbf{X} = (X_1, X_2, \ldots, X_d)^T$ follows a d-dimensional multivariate Gaussian distribution with mean vector $\boldsymbol{\mu} \in \mathbb{R}^d$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$, denoted $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, if its PDF is:

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)$$

- $\boldsymbol{\Sigma}$ must be symmetric and positive semi-definite
- The level sets are ellipsoids in $\mathbb{R}^d$
- Linear transformations preserve Gaussian structure

Geometric Interpretation

The covariance matrix $\boldsymbol{\Sigma}$ determines the shape and orientation of the PDF:

- **Eigenvectors** of $\boldsymbol{\Sigma}$ are the principal directions of the ellipsoid
- **Eigenvalues** determine the lengths of the semi-axes
- When $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$, level sets are spheres (isotropic variation)
- The density decreases exponentially with the Mahalanobis distance from the mean: $(\mathbf{x} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})$

Marginal Distributions

Proposition

If $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then any subset of components follows a multivariate Gaussian distribution with the corresponding subset of means and submatrix of $\boldsymbol{\Sigma}$.

Bivariate Example

For $\mathbf{X} = (X_1, X_2)^T \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with:

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$$

The marginal distributions are $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$.

Derivation of Marginal Distributions: Bivariate Case

Marginalizing X_2

To derive the marginal distribution of X_1, we integrate over X_2:

$$
\begin{aligned}
f_{X_1}(x_1) &= \int_{-\infty}^{\infty} f_{X_1,X_2}(x_1, x_2)\, dx_2 \\
&= \int_{-\infty}^{\infty} \frac{1}{2\pi\sqrt{|\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)\, dx_2
\end{aligned}
$$

Derivation of Marginal Distributions: Key Steps

Quadratic Form in the Exponent

For a bivariate Gaussian, the exponent can be written as:

$$-\frac{1}{2}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2(1-\rho^2)}-\frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2(1-\rho^2)}+\frac{(x_2-\mu_2)^2}{\sigma_2^2(1-\rho^2)}\right]$$

Where $\rho = \frac{\sigma_{12}}{\sigma_1\sigma_2}$ is the correlation coefficient.

Completing the Square in x_2

Group terms with x_2, complete the square, and integrate the resulting Gaussian form:

$$\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2a}(x_2-b)^2\right)\,dx_2 = \sqrt{2\pi a}$$

The result is a Gaussian distribution: $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$

Conditional Distributions

Proposition

Let $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)^T \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with:

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$$

Then $\mathbf{X}_1 | (\mathbf{X}_2 = \mathbf{x}_2) \sim \mathcal{N}(\boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2})$ where:

$$\boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)$$
$$\boldsymbol{\Sigma}_{1|2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$$

Independence and Correlation

Proposition

If $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then components X_i and X_j are independent if and only if $\Sigma_{ij} = 0$.

- For multivariate Gaussian distributions, uncorrelated variables are independent
- This is a special property that does not hold for most other distributions
- Independence corresponds to a diagonal covariance matrix

Gaussian Distributions in Physical Systems

- **Brownian motion:** Position of particles due to molecular collisions
- **Thermal noise:** Johnson-Nyquist noise in electronic circuits
- **Measurement errors:** Multiple independent sources of error
- **Diffusion processes:** Spread of heat, particles, or information

Mathematical Reasons for Universality

- Central Limit Theorem
- Maximum Entropy Principle
- Stability under convolution
- Self-similarity under scaling and translation

Biological and Social Systems

- **Human heights:** Multiple genetic and environmental factors
- **IQ scores:** Multiple cognitive abilities and environmental influences
- **Measurement errors:** In scientific experiments and observations
- **Financial markets:** Small price movements (although larger moves follow heavier-tailed distributions)

Robustness

The Gaussian appears even when underlying mechanisms are complex or unknown, provided they involve many small, independent contributions.

Linear Regression and MSE

Model Assumption

$$y = \mathbf{x}^T \boldsymbol{\beta} + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

- Maximum likelihood estimation with Gaussian errors leads to least squares solution
- Minimizing MSE is equivalent to maximum likelihood under Gaussian noise
- OLS estimator: $\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T y$

Gaussian Processes

Definition

A Gaussian process is a collection of random variables, any finite subset of which follows a multivariate Gaussian distribution.

- Powerful framework for Bayesian nonparametric regression
- Defined by a mean function $m(x)$ and covariance (kernel) function $k(x, x')$
- Provides uncertainty estimates for predictions
- Applications: time series forecasting, spatial statistics, Bayesian optimization

Probabilistic Models with Gaussian Components

- **Gaussian Mixture Models (GMMs):**

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$$

- **Principal Component Analysis (PCA):** Linear transformation to uncorrelated Gaussian variables
- **Variational Autoencoders:** Assume latent variables follow a Gaussian prior
- **Kalman Filter:** State estimation with Gaussian process and observation models

Heavy-Tailed Phenomena

- Gaussian distribution has light tails: extreme values are very rare
- Many real-world phenomena exhibit heavier tails
- Examples: financial returns, internet traffic, earthquake magnitudes

Alternatives for Heavy Tails

- **Student's t-distribution:** Heavier tails, controlled by degrees of freedom
- **Laplace distribution:** Leads to L1 regularization (lasso)
- **Stable distributions:** Generalization of Gaussian with heavy-tail properties

Mixture Models and Beyond

- **Gaussian Mixture Models:** For multimodal or heterogeneous data
- **Skewed distributions:** When symmetry assumption is violated
- **Robust alternatives:** Huber loss for robust regression
- **Copulas:** Flexible modeling of dependencies with arbitrary marginals

When to Move Beyond Gaussian

- When data exhibits multimodality, skewness, or heavy tails
- When outliers are frequent or expected
- When complex dependency structures exist between variables

Bias and Variance: Fundamental Concepts

Definition (Bias)

The bias of an estimator $\hat{\theta}$ for a parameter θ is:

$$\text{Bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$$

Definition (Variance)

The variance of an estimator $\hat{\theta}$ is:

$$\text{Var}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2]$$

Interpretation

- **Bias:** Systematic error; how far predictions are from true values on average
- **Variance:** Statistical dispersion; how much predictions fluctuate

The Bias-Variance Decomposition

Theorem

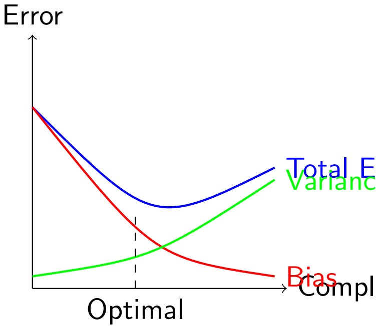
For a given point x, the expected mean squared error can be decomposed as:

$$\mathbb{E}[(y - \hat{f}(x))^2] = \underbrace{(f(x) - \mathbb{E}[\hat{f}(x)])^2}_{\text{Bias}^2} + \underbrace{\mathbb{E}[(\hat{f}(x) - \mathbb{E}[\hat{f}(x)])^2]}_{\text{Variance}} + \underbrace{\sigma_\varepsilon^2}_{\text{Irreducible Error}}$$

- $f(x)$ is the true function
- $\hat{f}(x)$ is our model's prediction
- σ_ε^2 is the noise variance (can't be reduced)

The Tradeoff

- Simple models: **High bias**, **low variance**
- Complex models: **Low bias**, **high variance**
- The goal: Find optimal complexity that minimizes total error



Estimating Parameters of a Gaussian Distribution

Given a sample $X_1, X_2, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$:

Maximum Likelihood Estimators

$$\hat{\mu}_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

$$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{\mu}_{\text{MLE}})^2$$

Properties

- $\hat{\mu}_{\text{MLE}}$ is unbiased: $\mathbb{E}[\hat{\mu}_{\text{MLE}}] = \mu$
- $\hat{\sigma}^2_{\text{MLE}}$ is biased: $\mathbb{E}[\hat{\sigma}^2_{\text{MLE}}] = \frac{n-1}{n}\sigma^2$
- Unbiased variance estimator: $\hat{\sigma}^2_{\text{unbiased}} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \hat{\mu})^2$

Temperature Measurements: A Case Study

Suppose daily temperatures follow $\mathcal{N}(20^\circ\mathrm{C}, 25\ {}^\circ\mathrm{C}^2)$:

Small Sample ($n = 10$)

$$\hat{\mu} = 18.5^\circ\mathrm{C}$$
$$\hat{\sigma}^2 = 19.8\ {}^\circ\mathrm{C}^2$$

High bias, high variance

Large Sample ($n = 1000$)

$$\hat{\mu} = 20.1^\circ\mathrm{C}$$
$$\hat{\sigma}^2 = 24.9\ {}^\circ\mathrm{C}^2$$

Low bias, low variance

Sampling Distribution

$$\hat{\mu} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) = \mathcal{N}\left(20, \frac{25}{n}\right)$$

Model Complexity: Underfitting vs. Overfitting

Underfitting (High Bias)

- Model is too simple to capture underlying structure
- Example: Linear model for seasonal temperature variation
- $f(x) = \beta_0 + \beta_1 x$

Good Fit (Balanced)

- Model captures the main patterns without fitting noise
- Example: Cubic model for seasonal temperature
- $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$

Overfitting (High Variance)

- Model captures noise, not just underlying structure
- Example: 30-degree polynomial for temperature data
- Performs well on training data, poorly on new data

Practical Techniques for Bias-Variance Management

Regularization

- **Ridge Regression**: L2 penalty, shrinks coefficients toward zero
- **Lasso Regression**: L1 penalty, performs feature selection
- **Elastic Net**: Combines L1 and L2 penalties

Cross-Validation

1. Split data into k folds
2. For each model complexity:
   - Train on $k - 1$ folds, test on remaining fold
   - Repeat for all folds and average results
3. Choose complexity with best validation performance

Practical Guidelines for Machine Learning

- **Small datasets**: Use simpler models (prioritize variance reduction)
- **Large datasets**: Can use more complex models (focus on bias reduction)
- **Start simple**: Begin with simpler models and gradually increase complexity
- **Ensemble methods**:
  - **Bagging** (Random Forests): Reduces variance
  - **Boosting**: Reduces bias
  - **Stacking**: Combines multiple models
- **Feature selection**: Remove irrelevant features to reduce variance
- **Regularization**: Add constraints to control overfitting

Key Takeaways

- The Gaussian distribution is mathematically elegant and computationally tractable
- The Central Limit Theorem explains its pervasiveness in natural phenomena
- Maximum Entropy Principle establishes its information-theoretic optimality
- Multivariate Gaussians have remarkable properties for marginal and conditional distributions
- The bias-variance decomposition helps understand prediction error
- The tradeoff between model complexity and performance is fundamental
- Regularization and cross-validation help manage the bias-variance tradeoff

Further Reading

- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
- Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press.
- Cover, T. M., & Thomas, J. A. (2006). *Elements of Information Theory*. Wiley-Interscience.
- Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer.