Parameter Estimation

Method of Moments & Maximum Likelihood Estimation

CMSC 173: Machine Learning

October 1, 2025

Course Outline

- **Introduction** - What is parameter estimation?
- **Statistical Foundations** - Key concepts and notation
- **Method of Moments** - Classical parameter estimation
- **Maximum Likelihood Estimation** - Optimal parameter estimation
- **Comparison** - When to use which method
- **Applications** - Real-world examples
- **Advanced Topics** - Extensions and modern approaches
- **Best Practices** - Common pitfalls and guidelines

Introduction

What is Parameter Estimation?

Definition

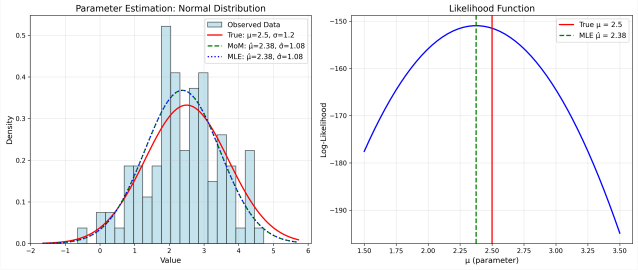
Parameter estimation is the process of inferring the values of unknown parameters that characterize a probability distribution from observed data.

The Problem:

- We have data samples: $\{x_1, x_2, \ldots, x_n\}$
- We assume a distribution family: $f(x|\theta)$
- We need to find: $\hat{\theta}$

Examples:

- Normal distribution: μ, σ^2
- Poisson distribution: λ
- Linear regression: β_0, β_1



Why Parameter Estimation Matters

Machine Learning Applications:

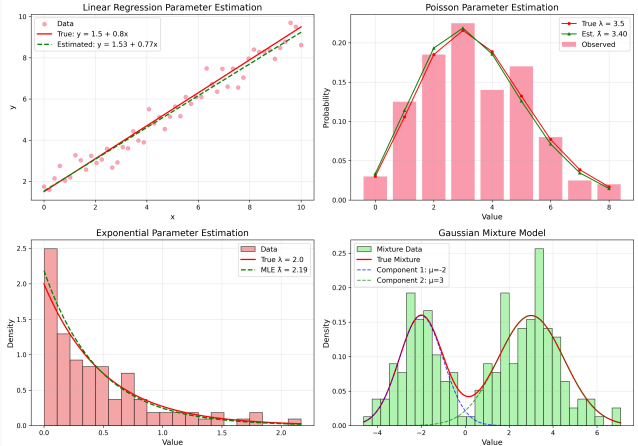
- **Supervised Learning:** Estimating model weights
- **Unsupervised Learning:** Finding cluster parameters
- **Probabilistic Models:** Bayesian inference
- **Time Series:** ARIMA parameters
- **Deep Learning:** Neural network weights

Example

Linear Regression: Given data (x_i, y_i), estimate:

$$y = \beta_0 + \beta_1 x + \epsilon$$

Find $\hat{\beta}_0, \hat{\beta}_1$ that best fit the data.



Estimation Quality Criteria

Desirable Properties of Estimators

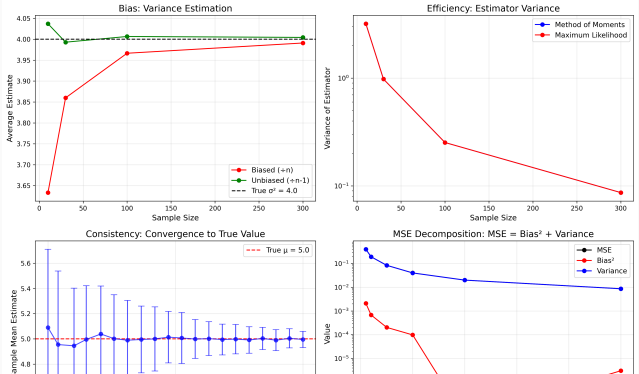
- **Unbiased:** $E[\hat{\theta}] = \theta$
- **Consistent:** $\hat{\theta} \xrightarrow{P} \theta$ as $n \to \infty$
- **Efficient:** Minimum variance among unbiased estimators
- **Sufficient:** Uses all information in the data

Bias-Variance Tradeoff:

$$MSE = Bias^2 + Variance + Noise$$

Cramér-Rao Lower Bound:

$$Var(\hat{\theta}) \geq \frac{1}{I(\theta)}$$



Statistical Foundations

Key Concepts and Notation

Random Variables:

- X: Random variable
- x: Observed value
- θ: True parameter
- $\hat{\theta}$: Estimated parameter

Distributions:

- $f(x|\theta)$: PDF/PMF
- $F(x|\theta)$: CDF
- $L(\theta|x)$: Likelihood

Sample vs Population

- **Population:** $\mu = E[X]$, $\sigma^2 = Var(X)$
- **Sample:** $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Example

For normal distribution $N(\mu, \sigma^2)$:

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

Moments and Central Moments

Definition of Moments

The k-th moment of a random variable X:

$$m_k = E[X^k] = \int_{-\infty}^{\infty} x^k f(x) dx$$

Raw Moments:

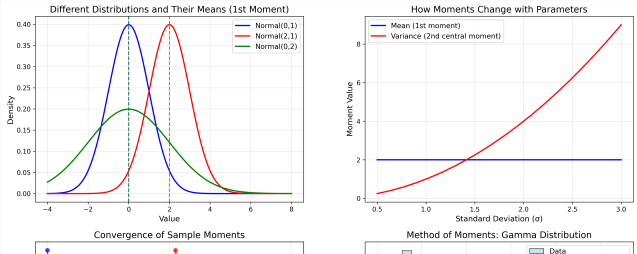
- $m_1 = E[X] = \mu$ (mean)
- $m_2 = E[X^2]$
- $m_3 = E[X^3]$
- $m_4 = E[X^4]$

Central Moments:

- $\mu_1 = 0$
- $\mu_2 = E[(X - \mu)^2] = \sigma^2$ (variance)
- $\mu_3 = E[(X - \mu)^3]$ (skewness)
- $\mu_4 = E[(X - \mu)^4]$ (kurtosis)

Example

For normal distribution $N(\mu, \sigma^2)$: $m_1 = \mu$, $m_2 = \mu^2 + \sigma^2$, $\mu_2 = \sigma^2$



Method of Moments

Method of Moments: Basic Idea

Core Principle

Match sample moments to theoretical moments to estimate parameters.

Algorithm:

1. Express theoretical moments in terms of parameters: $m_k(\theta)$
2. Calculate sample moments: $\hat{m}_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$
3. Set theoretical = sample: $m_k(\theta) = \hat{m}_k$
4. Solve system of equations for $\hat{\theta}$

For p parameters: Use first p moments

$$m_1(\theta) = \hat{m}_1$$

$$m_2(\theta) = \hat{m}_2$$

$$\vdots$$

$$m_p(\theta) = \hat{m}_p$$

Key Insight

If we can express moments as functions of parameters, we can invert to find parameters from moments.

MoM Example: Normal Distribution

Problem: Estimate μ and σ^2 for $N(\mu, \sigma^2)$

Step 1: Theoretical moments

$$m_1 = E[X] = \mu \tag{1}$$

$$m_2 = E[X^2] = \mu^2 + \sigma^2 \tag{2}$$

Step 2: Sample moments

$$\hat{m}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x} \tag{3}$$

$$\hat{m}_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \tag{4}$$

Step 3: Set equal and solve

$$\mu = \bar{x} \tag{5}$$

$$\mu^2 + \sigma^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \tag{6}$$

Step 4: Solution

$$\hat{\mu}_{MoM} = \bar{x} \tag{7}$$

$$\hat{\sigma}^2_{MoM} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{8}$$

MoM Example: Poisson Distribution

Problem: Estimate λ for Poisson(λ)

Theoretical moment: For Poisson distribution:

$$E[X] = \lambda$$

Sample moment:

$$\hat{m}_1 = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

MoM Estimate:

$$\hat{\lambda}_{MoM} = \bar{x}$$

Example

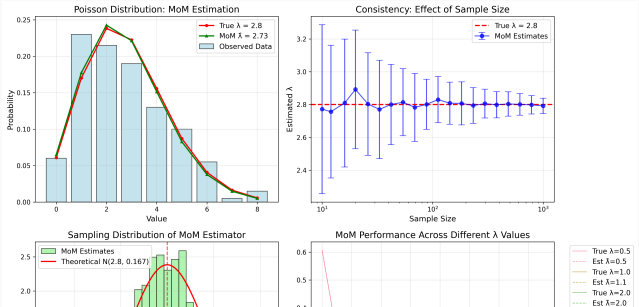
Data: [2, 1, 3, 0, 2, 1, 4, 1]

Sample mean:

$$\bar{x} = \frac{14}{8} = 1.75$$

MoM estimate:

$$\hat{\lambda} = 1.75$$



MoM Example: Gamma Distribution

Problem: Estimate α and β for Gamma(α, β)

Theoretical moments:

$$E[X] = \alpha\beta \tag{9}$$

$$Var(X) = \alpha\beta^2 \tag{10}$$

Also: $E[X^2] = Var(X) + (E[X])^2 = \alpha\beta^2 + \alpha^2\beta^2$

Sample moments:

$$\hat{m}_1 = \bar{x} \tag{11}$$

$$\hat{m}_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \tag{12}$$

Sample variance:

$$\hat{\sigma}^2 = \hat{m}_2 - \hat{m}_1^2$$

Setting moments equal:

$$\alpha\beta = \bar{x} \tag{13}$$

$$\alpha\beta^2 = \hat{\sigma}^2 \tag{14}$$

MoM Estimates:

$$\hat{\beta}_{MoM} = \frac{\hat{\sigma}^2}{\bar{x}}, \quad \hat{\alpha}_{MoM} = \frac{\bar{x}^2}{\hat{\sigma}^2}$$

Properties of Method of Moments

Advantages

- **Simple:** Easy to compute
- **Consistent:** $\hat{\theta} \to \theta$ as $n \to \infty$
- **General:** Works for any distribution
- **Intuitive:** Matches sample to theory

Disadvantages

- **Not optimal:** Higher variance than MLE
- **Existence:** Solutions may not exist
- **Uniqueness:** Multiple solutions possible
- **Boundary:** May give invalid estimates

Asymptotic Distribution

Under regularity conditions:

$$\sqrt{n}(\hat{\theta}_{MoM} - \theta) \xrightarrow{d} N(0, \Sigma)$$

where Σ depends on the moments and their derivatives.

Placeholder for
mom_properties.png

Maximum Likelihood Estimation

Maximum Likelihood: Basic Idea

Core Principle

Find parameter values that make the observed data most likely.

Likelihood Function: For independent observations $x_1, x_2, \ldots, x_n$:

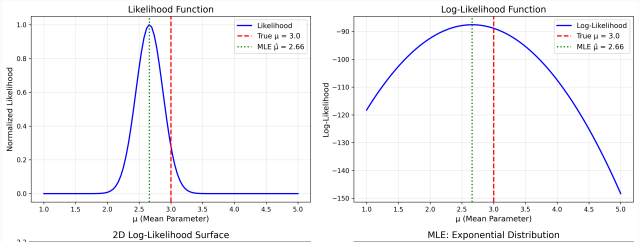
$$L(\theta) = L(\theta|x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\theta)$$

Log-Likelihood:

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(x_i|\theta)$$

Maximum Likelihood Estimator (MLE)

$$\hat{\theta}_{MLE} = \arg\max_{\theta} L(\theta) = \arg\max_{\theta} \ell(\theta)$$



Finding the MLE: Calculus Approach

Method 1: Differentiation

For continuous parameter space, solve:

$$\frac{d\ell(\theta)}{d\theta} = 0$$

Score Function:

$$S(\theta) = \frac{d\ell(\theta)}{d\theta} = \sum_{i=1}^{n} \frac{d \log f(x_i|\theta)}{d\theta}$$

For vector parameters θ:

$$\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) = \mathbf{0}$$

This gives a system of equations to solve.

Second-order condition:

$$\frac{d^2\ell(\theta)}{d\theta^2} < 0$$

Ensures we have a maximum, not minimum.

Example

For normal distribution with known σ^2:

$$\frac{d\ell(\mu)}{d\mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = 0$$

$$\Rightarrow \hat{\mu}_{MLE} = \bar{x}$$

MLE Example: Normal Distribution

Problem: Estimate μ and σ^2 for $N(\mu, \sigma^2)$

Log-likelihood:

$$\ell(\mu, \sigma^2) = \sum_{i=1}^{n} \log f(x_i|\mu, \sigma^2) \tag{15}$$

$$= \sum_{i=1}^{n} \left[-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{(x_i - \mu)^2}{2\sigma^2} \right] \tag{16}$$

$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{17}$$

Taking derivatives:

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = 0 \tag{18}$$

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i - \mu)^2 = 0 \tag{19}$$

MLE Solutions:

$$\hat{\mu}_{MLE} = \bar{x}, \quad \hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

MLE Example: Poisson Distribution

Problem: Estimate λ for Poisson(λ)

PMF: $P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$

Log-likelihood:

$$\ell(\lambda) = \sum_{i=1}^{n} \log P(X_i = x_i|\lambda) \tag{20}$$

$$= \sum_{i=1}^{n} [x_i \log \lambda - \lambda - \log(x_i!)] \tag{21}$$

$$= \log \lambda \sum_{i=1}^{n} x_i - n\lambda - \sum_{i=1}^{n} \log(x_i!) \tag{22}$$

Score function:

$$\frac{d\ell(\lambda)}{d\lambda} = \frac{1}{\lambda} \sum_{i=1}^{n} x_i - n = 0$$

MLE Solution:

$$\hat{\lambda}_{MLE} = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}$$

Note

For Poisson distribution, MLE and MoM give the same result!

MLE Example: Exponential Distribution

Problem: Estimate λ for Exponential(λ)

PDF: $f(x|\lambda) = \lambda e^{-\lambda x}$ for $x \geq 0$

Log-likelihood:

$$\ell(\lambda) = \sum_{i=1}^{n} \log(\lambda e^{-\lambda x_i}) \tag{23}$$

$$= \sum_{i=1}^{n} [\log \lambda - \lambda x_i] \tag{24}$$

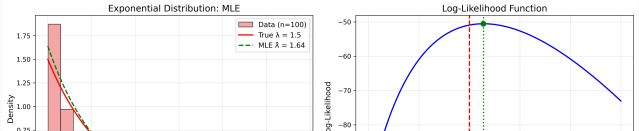
$$= n \log \lambda - \lambda \sum_{i=1}^{n} x_i \tag{25}$$

Score function:

$$\frac{d\ell(\lambda)}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0$$

MLE Solution:

$$\hat{\lambda}_{MLE} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{x}}$$



Properties of Maximum Likelihood Estimators

Asymptotic Properties (Large Sample)

Under regularity conditions:

- **Consistency:** $\hat{\theta}_{MLE} \xrightarrow{p} \theta$
- **Asymptotic Normality:** $\sqrt{n}(\hat{\theta}_{MLE} - \theta) \xrightarrow{d} N(0, I(\theta)^{-1})$
- **Efficiency:** Achieves Cramér-Rao lower bound
- **Invariance:** If $\hat{\theta}$ is MLE of θ, then $g(\hat{\theta})$ is MLE of $g(\theta)$

Fisher Information:

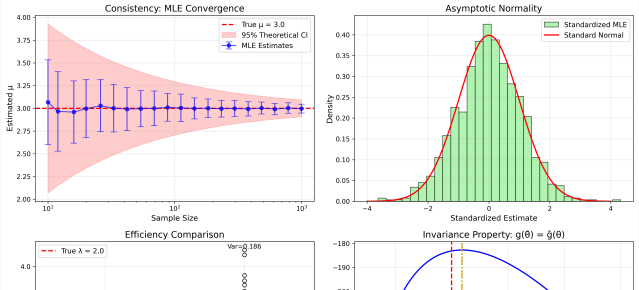
$$I(\theta) = -E\left[\frac{d^2\ell(\theta)}{d\theta^2}\right]$$

Higher information $\Rightarrow$ lower variance

Observed Information:

$$J(\hat{\theta}) = -\frac{d^2\ell(\theta)}{d\theta^2}\bigg|_{\theta=\hat{\theta}}$$

Used for confidence intervals



Numerical Methods for MLE

When Closed-Form Solution Doesn't Exist

Many distributions require numerical optimization to find MLE.

Newton-Raphson Method:

$$\theta^{(t+1)} = \theta^{(t)} - \frac{S(\theta^{(t)})}{J(\theta^{(t)})}$$

where $S(\theta)$ is score and $J(\theta)$ is observed information.

Other Methods:

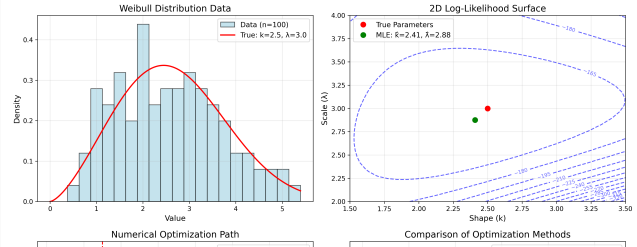
- Gradient ascent
- BFGS optimization
- EM algorithm (for latent variables)
- Grid search (for low dimensions)

Example

For mixture of Gaussians:

$$f(x|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \sigma_k^2)$$

No closed-form MLE $\Rightarrow$ Use EM algorithm



Comparison of Methods

Method of Moments vs Maximum Likelihood

Method of Moments

Pros:

- Simple computation
- Always exists (if moments exist)
- Distribution-free approach
- Good starting values for MLE

Cons:

- Not optimal (higher variance)
- May give invalid estimates
- Doesn't use full data information

Maximum Likelihood

Pros:

- Optimal (minimum variance)
- Uses full data information
- Good theoretical properties
- Invariance property

Cons:

- May require numerical methods
- Can be computationally intensive
- Requires specification of full distribution

Placeholder for
mom_vs_mle_comparison.png

Efficiency Comparison

Relative Efficiency

$$ARE = \frac{Var(\hat{\theta}_{MLE})}{Var(\hat{\theta}_{MoM})}$$

MLE is asymptotically more efficient when $ARE < 1$.

Normal Distribution:

- For μ: $ARE = 1$ (same efficiency)
- For σ^2: $ARE = 0.5$ (MLE better)

Exponential Distribution:

- For λ: $ARE = 1$ (same efficiency)

Gamma Distribution:

- MLE significantly more efficient
- MoM can be quite inefficient

General Rule:

- MLE $\geq$ MoM in efficiency
- Difference larger for complex distributions

Placeholder for
efficiency_comparison.png

When to Use Which Method?

Use Method of Moments When:

- Quick estimates needed
- Computational resources limited
- Distribution family uncertain
- Starting values for optimization
- Robust estimates desired
- Teaching/illustration purposes

Use Maximum Likelihood When:

- Optimal estimates needed
- Distribution well-specified
- Large sample sizes
- Inference required (confidence intervals)
- Model comparison needed
- Production/research applications

Practical Strategy

1. Start with Method of Moments for initial estimates
2. Use MoM estimates as starting values for MLE optimization
3. Compare results and choose based on application needs
4. Consider computational cost vs. statistical efficiency trade-off

Applications

Linear Regression Parameter Estimation

Model: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$

Method of Moments:

$$E[Y] = \beta_0 + \beta_1 E[X] \tag{26}$$

$$E[XY] = \beta_0 E[X] + \beta_1 E[X^2] \tag{27}$$

Solving:

$$\hat{\beta}_1 = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} \tag{28}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{29}$$

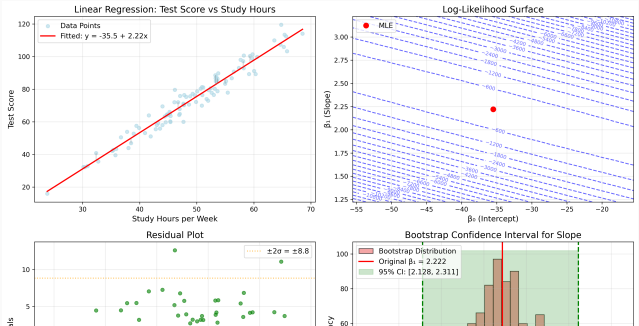
Maximum Likelihood:

$$\ell(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 \tag{30}$$

MLE gives same result:

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \tag{31}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{32}$$



Logistic Regression Parameter Estimation

Model: $P(Y = 1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}}$

No Closed-Form Solution

Logistic regression requires numerical optimization for MLE.

Log-likelihood:

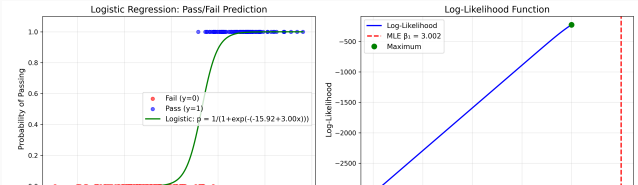
$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i(\beta_0 + \beta_1 x_i) - \log(1 + e^{\beta_0+\beta_1 x_i}) \right]$$

Score equations:

$$\frac{\partial \ell}{\partial \beta_0} = \sum_{i=1}^{n} (y_i - p_i) = 0 \tag{33}$$

$$\frac{\partial \ell}{\partial \beta_1} = \sum_{i=1}^{n} x_i(y_i - p_i) = 0 \tag{34}$$

where $p_i = \frac{1}{1+e^{-(\beta_0+\beta_1 x_i)}}$



Clustering: Gaussian Mixture Models

Model: $f(x|\theta) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \sigma_k^2)$

Parameters to estimate:

- Mixing weights: $\pi_1, \ldots, \pi_K$
- Means: $\mu_1, \ldots, \mu_K$
- Variances: $\sigma_1^2, \ldots, \sigma_K^2$

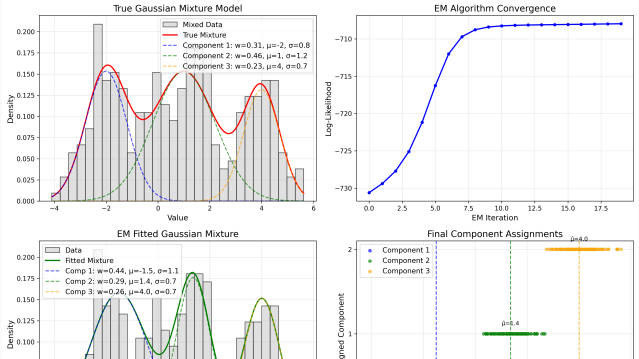
Challenges:

- Latent variables (cluster assignments)
- Complex likelihood surface
- Local optima
- Model selection (choosing K)

EM Algorithm

E-step: Compute posterior probabilities of cluster assignments

M-step: Update parameters using weighted MLE



Time Series: ARIMA Parameters

ARIMA(p,d,q) Model:

$$(1 - \phi_1 B - \cdots - \phi_p B^p)(1 - B)^d X_t = (1 + \theta_1 B + \cdots + \theta_q B^q)\epsilon_t$$

Method of Moments:

- Use sample autocorrelations
- Yule-Walker equations for AR parts
- Moment conditions for MA parts
- Simple but not optimal

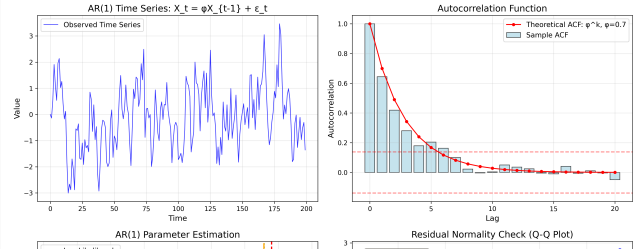
Maximum Likelihood:

- Kalman filter for likelihood
- Numerical optimization required
- More efficient estimates
- Standard errors available

Example

AR(1): $X_t = \phi X_{t-1} + \epsilon_t$

- MoM: $\hat{\phi} = \hat{\rho}_1$ (sample autocorrelation)
- MLE: Optimize $\ell(\phi, \sigma^2)$ numerically



Advanced Topics

Bayesian Parameter Estimation

Bayesian Approach

Treat parameters as random variables with prior distributions.

Bayes' Theorem:

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \propto p(x|\theta)p(\theta)$$

Components:

- $p(\theta)$: Prior distribution
- $p(x|\theta)$: Likelihood
- $p(\theta|x)$: Posterior distribution
- $p(x)$: Marginal likelihood

Estimation:

- MAP: $\hat{\theta}_{MAP} = \arg\max p(\theta|x)$
- Posterior mean: $\hat{\theta}_{PM} = E[\theta|x]$
- Credible intervals available

Example

Normal with normal prior: Prior: $\mu \sim N(\mu_0, \tau^2)$, Likelihood: $X|\mu \sim N(\mu, \sigma^2)$ Posterior: $\mu|x \sim N\left(\frac{\tau^2\bar{x}+\sigma^2\mu_0/n}{\tau^2+\sigma^2/n}, \frac{\tau^2\sigma^2/n}{\tau^2+\sigma^2/n}\right)$

Robust Parameter Estimation

Problem with MLE

MLE can be sensitive to outliers and model misspecification.

Robust Alternatives:

- **M-estimators:** Generalize MLE
- **Huber estimator:** Robust to outliers
- **Trimmed means:** Remove extreme values
- **Median-based:** Use median instead of mean

Example - Huber Loss:

$$\rho(x) = \begin{cases} \frac{1}{2}x^2 & |x| \leq k \\ k|x| - \frac{1}{2}k^2 & |x| > k \end{cases}$$

Quadratic for small errors, linear for large errors.

Trade-offs

Robust methods sacrifice some efficiency for stability against outliers.

Placeholder for
robust_estimation.png

Bootstrap and Resampling

Bootstrap Principle

Estimate sampling distribution by resampling from the observed data.

Algorithm:

1. Draw B bootstrap samples: $\{x_1^*, \ldots, x_n^*\}$ from original data
2. Compute estimate for each sample: $\hat{\theta}_b^*$
3. Use distribution of $\{\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*\}$ for inference

Applications:

- Confidence intervals
- Bias correction
- Variance estimation
- Hypothesis testing

Bootstrap Confidence Interval: For 95

Bias Correction:

$$\hat{\theta}_{corrected} = 2\hat{\theta} - \bar{\theta}^*$$

Placeholder for
bootstrap_estimation.png

Model Selection and Information Criteria

Model Selection Problem

How do we choose between different models or number of parameters?

Information Criteria:

$$AIC = -2\ell(\hat{\theta}) + 2k \tag{35}$$

$$BIC = -2\ell(\hat{\theta}) + k \log n \tag{36}$$

$$AICc = AIC + \frac{2k(k+1)}{n-k-1} \tag{37}$$

where k = number of parameters, n = sample size.

Interpretation:

- Lower values = better models
- Trade-off: fit vs complexity
- AIC: prediction focus
- BIC: true model focus

Cross-Validation:

- Split data into train/validation
- Estimate on training set
- Evaluate on validation set
- Choose model with best CV score

Placeholder for model_selection.png

Best Practices

Common Pitfalls and How to Avoid Them

Pitfall 1: Wrong Distribution

Assuming incorrect distributional family leads to biased estimates.

Solution:

- Exploratory data analysis
- Goodness-of-fit tests
- Residual analysis
- Model comparison

Pitfall 2: Insufficient Data

Small samples lead to unreliable estimates.

Solution:

- Check sample size requirements
- Use bootstrap for uncertainty
- Consider Bayesian methods
- Regularization techniques

Pitfall 3: Outliers

Extreme values can severely affect estimates.

Solution:

- Data visualization
- Robust estimation methods
- Outlier detection and treatment
- Sensitivity analysis

Pitfall 4: Overfitting

Too many parameters relative to data.

Solution:

- Information criteria (AIC, BIC)
- Cross-validation
- Regularization (Ridge, Lasso)
- Domain knowledge constraints

Diagnostic Tools and Validation

Model Validation Checklist

Always validate your parameter estimates and model assumptions.

Residual Analysis:

- Plot residuals vs fitted values
- Check for patterns or heteroscedasticity
- Normal probability plots
- Autocorrelation in residuals

Goodness-of-Fit Tests:

- Kolmogorov-Smirnov test
- Anderson-Darling test
- Chi-square test
- Likelihood ratio tests

Confidence Intervals:

- Asymptotic (Fisher Information)
- Profile likelihood
- Bootstrap intervals
- Bayesian credible intervals

Sensitivity Analysis:

- Remove potential outliers
- Subsample analysis
- Perturbation studies
- Cross-validation

Placeholder for
diagnostic_tools.png

Computational Considerations

Optimization Tips

- **Starting values:** Use MoM for MLE initialization
- **Scaling:** Normalize variables for numerical stability
- **Constraints:** Handle parameter bounds properly
- **Convergence:** Check multiple starting points

Implementation

- **Vectorization:** Use efficient matrix operations
- **Automatic differentiation:** For complex models
- **Parallel computing:** Bootstrap and cross-validation
- **Memory management:** For large datasets

Software Tools

- **Python:** scipy.optimize, statsmodels, scikit-learn
- **R:** optim(), nlm(), maxLik package
- **Specialized:** Stan, PyMC for Bayesian methods
- **Deep Learning:** TensorFlow, PyTorch for gradient-based optimization

Placeholder for
computational_tools.png

Summary and Key Takeaways

Method of Moments

When to use:

- Quick estimates needed
- Simple distributions
- Starting values for MLE
- Computational constraints

Key insight: Match theoretical and sample moments

Maximum Likelihood

When to use:

- Optimal estimates desired
- Large sample sizes
- Inference required
- Model comparison

Key insight: Find parameters that maximize data likelihood

General Principles

- **Start simple:** Use MoM, then refine with MLE if needed
- **Validate assumptions:** Check distributional assumptions
- **Assess uncertainty:** Always provide confidence intervals
- **Consider alternatives:** Robust methods for outliers
- **Use diagnostics:** Residual analysis and goodness-of-fit

Parameter estimation is fundamental to statistical modeling and machine learning!

Next Steps and Further Reading

Advanced Topics to Explore:

- Generalized Method of Moments (GMM)
- Quasi-Maximum Likelihood
- Empirical likelihood methods
- Regularized estimation (Ridge, Lasso)
- Bayesian computation (MCMC)

Applications in ML:

- Neural network training
- Variational autoencoders
- Gaussian processes
- Hidden Markov models
- Reinforcement learning

Recommended Resources

- **Books:** Casella & Berger "Statistical Inference", Lehmann & Casella "Theory of Point Estimation"
- **Software:** Practice with scipy.optimize, statsmodels, Stan
- **Datasets:** UCI ML Repository, Kaggle competitions

Thank you! Questions?