PART III. Optimal design theory
(LECTURE 1)

Shpilev Petr Valerievich
Faculty of Mathematics and Mechanics, SPbU

September, 2025

Санкт-Петербургский государственный университет

Comments

The third part of our course introduces the principles of regression modeling and optimal experimental design. Many optimization problems rely on expensive function evaluations. To address this, we use surrogate models: smooth, low-cost approximations of the true objective function. These models are fast to evaluate and optimize, guiding us toward the true optimum efficiently. By some sampling of the real function, we refine the surrogate, improving its accuracy over time. Beyond optimization, this approach is powerful for descriptive regression, where we study how variables interact. It helps uncover relationships between inputs (e.g., material properties or hyperparameters) and outputs (e.g., performance or efficiency)—without strict assumptions about data distributions. So the third part of our course dives into regression analysis basics and optimal experiment design – core techniques for building and refining models.

In today's lecture, we introduce the foundational principles of descriptive regression and its application in optimal design theory. We begin by exploring the core idea of descriptive regression, with a focus on empirical data and parametric regression models. A key example is the modeling of height as a function of weight, including a visual representation and the process of estimating the model parameters. We examine the criteria for optimal parameter estimation and the importance of the least squares estimator (LSE), which leads us into the discussion of normal equations and classical linear regression models.

The lecture progresses with a detailed explanation of the Gauss-Markov theorem and the conditions for the Best Linear Unbiased Estimator (BLUE), illustrating the optimality of the LSE in terms of variance minimization. The proof of important lemmas, such as those related to the properties of the OLS estimator and variance of linear unbiased estimators, is presented in-depth. Through specific examples, we further explore the practical application of least squares estimation, both with exact data and in the context of quadratic approximation.

We conclude the lecture by introducing the concept of estimability of linear parametric functions, explaining the conditions under which these estimators are feasible and optimal. The lecture sets the stage for the more advanced topics in optimal design theory that follow.

Descriptive Regression: The Core Idea

Regression analysis is a statistical method for estimating the relationships among variables. It describes how a dependent variable changes as one or more independent variables change.

- **Purpose:** To represent and understand the relationship between a response (output) and several influencing factors (inputs).
- **Focus of Descriptive Regression:** To build a model based on observed data without making strong assumptions about the underlying statistical distribution.

Key Concepts

- **Dependent Variable (y):** The outcome or response variable we are trying to predict or explain.
- **Independent Variables ($x_1, \ldots, x_m$):** The predictor variables or factors that influence the dependent variable. Also called regressors.
- **Model Function ($y = \eta(x_1, \ldots, x_m)$):** A mathematical representation of the relationship.

Comments

Many optimization problems involve expensive function evaluations. For instance, testing a hardware design might require hours of fabrication, an aircraft design needs costly wind tunnel tests, and tuning deep learning hyperparameters could take a week of GPU training. To tackle this, we use surrogate models—smooth, inexpensive approximations of the true objective function. These models are quick to evaluate and optimize, guiding us toward the true optimum without excessive cost. By occasionally evaluating the real function, we can refine the surrogate model, making it more accurate over time. Beyond optimization, this approach is powerful for descriptive regression, which studies how variables interact. It helps us understand relationships between inputs, like material properties or hyperparameters, and outputs, like performance or efficiency, without assuming specific data distributions.

Let's give a formal description of this approach. Imagine we have a system that works like a black box. We input a set of control signals, say x_1 through x_m, which in practice are typically numerical quantities. The output is a single scalar value, y, which depends on x_1 through x_m. Our goal is to determine this relationship, that is, to find a function where y equals η of x_1 through x_m. To solve this task, the first step is to collect experimental data, meaning the results of simultaneous measurements of y and x_1 through x_m.

Empirical Data in Descriptive Regression

Setup: After conducting N experiments, we collect data to model the relationship between inputs and outputs.

- **Empirical Data Matrix:**

$$(Y, X) = \begin{bmatrix} y_1 & x_{11} & \cdots & x_{1m} \\ y_2 & x_{21} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ y_N & x_{N1} & \cdots & x_{Nm} \end{bmatrix}$$

Output Vector (Y): $\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$ Design Matrix (X): $\begin{bmatrix} x_{11} & \cdots & x_{1m} \\ x_{21} & \cdots & x_{2m} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{Nm} \end{bmatrix}$

Key Terms:

Vector of Results (Y): Column of observed output values y_j for $j = 1, \ldots, N$.
Design Matrix (X): Also called plan matrix, stores inputs x_{ij}, often predefined.
Plan (design): Inputs x_{ij} can be set or measured before experiments.

Comments

This slide formalizes how we organize experimental data for regression analysis. After conducting N experiments, we structure our observations into two key components: the output vector Y and design matrix X. The output vector contains all observed values of our dependent variable y, while the design matrix systematically records all corresponding input variable values.

The design matrix plays a crucial role - its structure determines what relationships we can detect between inputs and outputs. Each row represents a complete experimental observation, while columns correspond to different input variables. This arrangement allows us to efficiently analyze multivariate relationships.

The terminology is important here: we distinguish between the "vector of results" (our measured outcomes) and the "design matrix" (our controlled or observed inputs). In many experimental setups, the x-values in the design matrix are carefully chosen in advance (the experimental "design"), while the y-values are measured outcomes that may contain noise. This structured approach enables us to apply mathematical and statistical tools to uncover the underlying relationship η between inputs and outputs.

Parametric Regression Models

Core Challenge: Exact functional relationships are often too complex to determine, so we use simplified parametric models that approximate the statistical dependence with satisfactory accuracy.

Parametric Regression Approach

- Choose a model function $\eta(\mathrm{x}, \theta)$ with parameters θ
- Find θ that best fits the observed data
- Linear vs nonlinear in parameters
- Balance between simplicity and accuracy

Model Types

Linear Models:

$$\eta(\mathrm{x}, \theta) = \theta_1 \mathrm{x}_1 + \cdots + \theta_\mathrm{m} \mathrm{x}_\mathrm{m}$$

With Intercept:

$$\eta(\mathrm{x}, \theta) = \theta_0 + \theta_1 \mathrm{x}_1 + \cdots + \theta_{\mathrm{m}-1} \mathrm{x}_{\mathrm{m}-1}$$

Nonlinear Models:

$\eta(\mathrm{x}, \theta) =$ Any non-linear function

Why Approximate?

- Real-world relationships are complex
- Measurement noise and variability
- Computational tractability
- Interpretability trade-off

Comments

Parametric regression provides a practical framework for modeling complex relationships when exact functional forms are unknown. The key idea is to select a family of functions defined by parameters θ that can approximate the true relationship with sufficient accuracy. Linear models are particularly important due to their simplicity and interpretability - they assume the output is a weighted sum of inputs. The inclusion of an intercept term (θ_0) accounts for baseline effects. More complex nonlinear models can capture intricate patterns but require careful handling to avoid overfitting. The choice between model complexity and simplicity involves trade-offs: simpler models are more robust and interpretable but may miss important relationships, while complex models can fit data better but may capture noise rather than signal. Ultimately, the goal is to find the sweet spot where the model is complex enough to be useful but simple enough to be reliable.

Modeling Height as a Function of Weight

Objective: Investigate the relationship between height (y) and weight (x_1) in a sample of adult men.

- The dataset consists of 35 points (x_{1j}, y_j), where x_1 is weight (kg) and y is height (cm).
- There is no strict functional dependence between x_1 and y: multiple weights can correspond to the same height and vice versa.
- However, the average trend of height as a function of weight can be modeled approximately.

Linear Approximation Model

$$y \approx \theta_0 + \theta_1 x_1$$

This simple linear form is widely used in practice to estimate statistical relationships.

Visualization: Bivariate plots (like the one shown next) help identify trends and statistical dependencies between variables.

Comments

Let's consider a simple but revealing example — modeling height as a function of weight.

Now, intuitively, we expect that taller people tend to weigh more. But does that mean there's a strict mathematical formula linking the two? Not quite. If you've ever compared athletes of the same weight — say, a basketball player and a swimmer — you'll know that height can vary a lot. So what we're really after isn't a precise rule, but an average trend.

In this example, we're working with a dataset of 35 adult males. Each data point is a pair: a person's weight in kilograms and their height in centimeters. If we plot these on a 2D graph, the points won't fall neatly on a line — and that's okay for real data.

Still, we might want to summarize the overall relationship with a simple mathematical model. And one of the most widely used tools for that is linear approximation. Here, we propose a model of the form y approximately equal to $\theta_0 + \theta_1 x_1$, where y is height and x_1 is weight. The idea is to find a straight line that "best fits" the cloud of points — not perfectly, but in terms of average behavior.

This approach has two huge advantages. First, it's simple and interpretable: we can explain what each parameter means. Second, it gives us a baseline. Once we understand the linear trend, we can check if a more complex model is even necessary.

And to really get a feel for the relationship, we'll visualize the data next. Seeing the points and the fitted line side by side gives us intuition about how well the model captures the data — and where it might fall short.

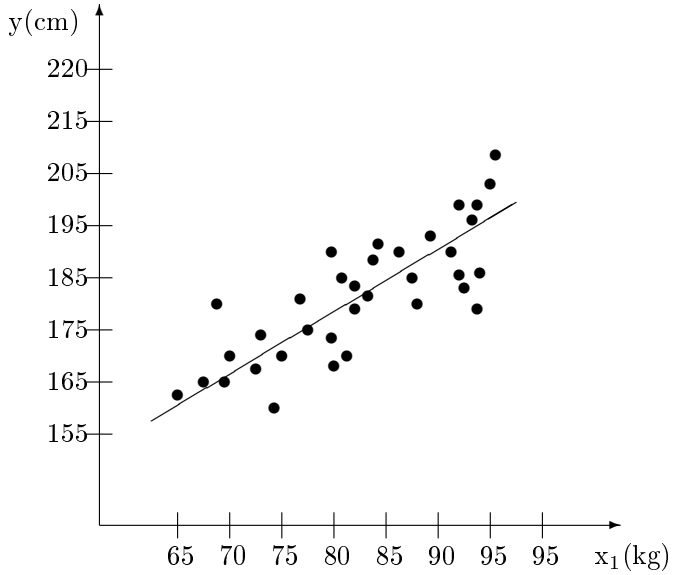

Figure: Data on the ratio of weight and height for 35 men.

Comments

Let's take a look at the actual data.

Here, each dot represents one individual — a pair of weight and height measurements. As you can see, the points are scattered: there's a lot of variation. Some individuals with the same weight have very different heights, and vice versa. This is exactly what we expect from real-world biological data.

But despite the noise, we can still see a clear trend. There's an upward tilt to the cloud of points — as weight increases, height tends to increase as well. This pattern is captured by the straight line we've drawn across the plot. That line is our proposed linear model: it doesn't pass through every point, but it does its best to reflect the average relationship between weight and height.

What's important to notice here is that the model smooths over the individual differences. It gives us a clean, interpretable summary: on average, heavier individuals are taller. Of course, it doesn't explain everything — but that's okay. Our goal at this stage is to capture the overall direction of the relationship, not every detail.

Later, we'll talk about how to actually compute that best-fitting line — and how to measure how well it performs.

Estimating Parameters of the Regression Model

Goal: Determine the parameter values $\theta_1, \ldots, \theta_m$ that make the regression model best describe the observed data.

- After choosing the model form $y \approx \eta(x_1, \ldots, x_m)$, the next task is to find the unknown parameters θ_i.
- A criterion of fit must be introduced to evaluate how well the model matches the experimental data.
- We compare measured values y_j with predicted values $\widetilde{y}_j$.

Deviation of the Linear Model

$$\epsilon_j = y_j - \widetilde{y}_j, \quad j = 1, \ldots, N, \quad \text{where} \quad \widetilde{y}_j = \sum_{i=1}^{m} x_{ji}\theta_i.$$

$$\text{Or equivalently:} \quad y_j = \sum_{i=1}^{m} x_{ji}\theta_i + \epsilon_j$$

$$\text{Matrix form:} \quad y = X\theta + \epsilon,$$

$\theta = (\theta_1, \ldots, \theta_m)^T$ is the vector of unknown parameters,
$\epsilon = (\epsilon_1, \ldots, \epsilon_N)^T$, is the vector of deviations.

Comments

After deciding the model structure, we need to find parameter values that make the model describe the experimental data as well as possible.

To do this, we introduce the concept of error — the difference between the observed value and the predicted value by the model. This error shows how well the model works for each individual observation.

For linear model the predicted value for each data point is calculated as a weighted sum of the input variables, where the weights are the model parameters we want to find.

We can rewrite the relationship by saying that the observed value equals the predicted part plus the error. This highlights that real data usually do not fit perfectly to the model because deviations always exist.

Finally, we represent the whole model compactly using matrix notation: the vector of observed values equals the product of the input data matrix and the parameter vector, plus the vector of errors. This form is very useful for calculations and theoretical analysis.

Criteria for Optimal Parameter Estimation

Error vector and parameter dependence

- The error vector $\epsilon = y - X\theta$ depends on the parameter vector θ, assuming X is fixed.
- The quality of the regression model is thus determined by the choice of θ.

Common optimization criteria

To find the "best" parameters, different criteria can be used:

- Minimize the maximum absolute error: $\max|\epsilon_j|$
- Minimize the sum of absolute errors: $\sum|\epsilon_j|$
- **Minimize the sum of squared errors:** $\epsilon^T\epsilon$

Why least squares is preferred

The last criterion, minimizing the sum of squared errors, is the most widely used due to its computational simplicity and optimality properties.

Comments

With the error vector defined, we now need a meaningful way to choose the best parameter values — that is, those that make the model fit the data as closely as possible.

There are several criteria for this. One option is to minimize the largest absolute error among all predictions. This ensures that no individual prediction is too far off, but it ignores the overall pattern. Another option is to minimize the total sum of absolute deviations. This is more balanced, but the resulting optimization problem is harder to solve analytically.

The most commonly used and most practical criterion is to minimize the sum of squared deviations. That is, we square each residual, then sum them all up. This criterion — known as the least squares method — is not only computationally convenient but also leads to estimators with strong theoretical guarantees, such as unbiasedness and efficiency under classical assumptions.

Its popularity comes precisely from this combination of simplicity and optimality. In most applications, minimizing the squared error is the default strategy, unless there's a specific reason to choose a different loss function.

Definition: Least Squares Estimator

The vector

$$\widehat{\theta} = \arg\min_{\theta \in \mathbb{R}^m} (\mathrm{Y} - \mathrm{X}\theta)^{\mathrm{T}} (\mathrm{Y} - \mathrm{X}\theta) = \arg\min_{\theta \in \mathbb{R}^m} \epsilon^{\mathrm{T}} \epsilon$$

is called the empirical least squares estimator (LSE) .

Lemma 1 (Normal Equations)

For any matrix X and vector Y of compatible dimensions, the system

$$\mathrm{X}^{\mathrm{T}}\mathrm{X}\theta = \mathrm{X}^{\mathrm{T}}\mathrm{Y}$$

— called the *system of normal equations* — always has at least one solution. Any vector θ^* satisfying this system is a least squares estimator.

Features of the system of normal equations

- The least squares criterion is quadratic in θ, leading to a convex optimization problem.
- The normal equations provide a closed-form condition for optimality.

Comments

So, as a rule to determine the best-fit parameters for a linear model, we apply the least squares method, minimizing the sum of squared residuals. The parameter vector that achieves this minimum is called the empirical least squares estimator.

The following lemma holds: for any matrix X and any compatible vector Y, the equation system $\mathrm{X}^{\mathrm{T}}\mathrm{X}\theta = \mathrm{X}^{\mathrm{T}}\mathrm{Y}$, known as the system of normal equations, always has at least one solution. Any vector that satisfies these equations is a least squares estimator.

This formulation is critical in both theoretical and computational aspects of regression analysis. It allows us to find an optimal estimate analytically — at least when the normal matrix $\mathrm{X}^{\mathrm{T}}\mathrm{X}$ is invertible. Even if it's not, a solution still exists, possibly non-unique.

The optimization problem itself is convex, since the objective function is quadratic in θ. This means that any solution to the normal equations corresponds to a global minimum. The name "normal equations" comes from classical least squares theory and is standard terminology in linear regression.

Lemma 1: Proof (part 1)

Proof: To begin with, we show that there is always a solution to a system of normal equations. For any matrix A, denote by $\mathcal{L}(A)$ the linear span of its columns.
We claim that:

$$\mathcal{L}(A^T) = \mathcal{L}(A^T A).$$

- Let b be a vector orthogonal to $\mathcal{L}(A^T)$. Then

$$b^T A^T = 0 \quad \Rightarrow \quad b^T A^T A = 0,$$

so $b \perp \mathcal{L}(A^T A)$.

- Conversely, if $b \perp \mathcal{L}(A^T A)$, then

$$b^T A^T A = 0 \quad \Rightarrow \quad (Ab)^T Ab = 0 \quad \Rightarrow \quad Ab = 0 \quad \Rightarrow \quad b^T A^T = 0.$$

So $b \perp \mathcal{L}(A^T)$.

Hence, the spaces coincide:

$$\mathcal{L}(A^T) = \mathcal{L}(A^T A).$$

Thus, for any $X^T Y \in \mathcal{L}(X^T)$, there exists θ^* such that

$$X^T X\theta^* = X^T Y.$$

Comments

We begin the proof by showing that the system of normal equations always has a solution.

To do this, we prove that for any matrix A, the space spanned by the rows of A, which is the same as the column space of A transposed, coincides with the column space of the matrix $A^T A$.

To show this, take any vector b orthogonal to the column space of A transposed. Then $b^T A^T = 0$, and multiplying by A again yields zero. This implies b is orthogonal to the column space of $A^T A$.

Conversely, if b is orthogonal to the column space of A transposed A, then the product $(Ab)^T Ab = 0$, so Ab is zero, and hence $b^T A^T$ is also zero.This proves the spaces coincide. It follows that the vector X transposed Y lies in the column space of $X^T X$, so the equation $X^T X\theta = X^T Y$ has a solution θ^*.

Lemma 1: Proof (part 2)

Let $\widehat{\theta}$ be any solution to the normal equations:

$$\mathrm{X}^{\mathrm{T}}\mathrm{X}\widehat{\theta} = \mathrm{X}^{\mathrm{T}}\mathrm{Y}.$$

Then for arbitrary θ:

$$\begin{aligned}&(\mathrm{Y} - \mathrm{X}\theta)^{\mathrm{T}}(\mathrm{Y} - \mathrm{X}\theta) = \\ &= (\mathrm{Y} - \mathrm{X}\widehat{\theta} + \mathrm{X}(\widehat{\theta} - \theta))^{\mathrm{T}}(\mathrm{Y} - \mathrm{X}\widehat{\theta} + \mathrm{X}(\widehat{\theta} - \theta)) = \\ &= (\mathrm{Y} - \mathrm{X}\widehat{\theta})^{\mathrm{T}}(\mathrm{Y} - \mathrm{X}\widehat{\theta}) + (\widehat{\theta} - \theta)^{\mathrm{T}}\mathrm{X}^{\mathrm{T}}\mathrm{X}(\widehat{\theta} - \theta) \geq \\ &\geq (\mathrm{Y} - \mathrm{X}\widehat{\theta})^{\mathrm{T}}(\mathrm{Y} - \mathrm{X}\widehat{\theta}).\end{aligned}$$

The cross term vanishes due to:

$$(\widehat{\theta} - \theta)^{\mathrm{T}}\mathrm{X}^{\mathrm{T}}(\mathrm{Y} - \mathrm{X}\widehat{\theta}) = 0.$$

Therefore, $\widehat{\theta}$ minimizes the squared error. The Lemma is proved. □

Remark

If $\mathrm{X}^{\mathrm{T}}\mathrm{X}$ is nonsingular, then the normal system has a unique solution:

$$\widehat{\theta} = (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{Y}.$$

Comments

Now we show that any solution to a system of normal equations is least squares estimate. To be precise let θ hat be any solution to the system of normal equations. Then we compute the squared error norm of any arbitrary θ by rewriting the expression in terms of θ hat. This is done using a standard identity from linear algebra.

The key point is that the cross-term vanishes due to the fact that θ hat solves the normal equations. As a result, the squared error of any θ equals the squared error of θ hat plus a non-negative term. This implies that the minimum is achieved exactly when $\theta = \widehat{\theta}$. Therefore, any solution to the normal equations minimizes the residual norm and is indeed least squares estimate. The Lemma is proved.

We also note that if the matrix X transposed times X is nonsingular, then the normal equations have a unique solution. This unique estimator is then given explicitly as $(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{Y}$.

Mathematical formulation

The classical linear regression model is written as:

$$y_j = \sum_{i=1}^{m} x_{ji}\theta_i + \epsilon_j, \quad j = 1, \dots, N, \tag{1}$$

or in matrix form:

$$Y = X\theta + \epsilon. \tag{2}$$

- $\theta = (\theta_1, \dots, \theta_m)^T$ — vector of unknown coefficients.
- $Y = (y_1, \dots, y_N)^T$ — vector of observed responses.
- $X = (x_{ij})_{i=1,j=1}^{m\ N}$ — design matrix:

$$X = \begin{pmatrix} x_{11} & \dots & x_{1m} \\ x_{21} & \dots & x_{2m} \\ \vdots & \ddots & \vdots \\ x_{N1} & \dots & x_{Nm} \end{pmatrix}$$

- $\epsilon = (\epsilon_1, \dots, \epsilon_N)^T$ — vector of random errors.

Comments

In the regression setting, the observed output is modeled as a deterministic part plus an error term. The deterministic part is a linear combination of the inputs, with unknown coefficients that describe how each input influences the output. The error term accounts for noise and other unmodeled effects.

The classical linear regression model assumes that each observation y_j is equal to the sum of the input values x_{ji} multiplied by unknown coefficients θ_i, plus a random error ϵ_j. This can be written as: y_j equals the sum over i from 1 to m of x_{ji} times θ_i, plus ϵ_j, or j from 1 to N.

To make the notation more compact and the analysis more convenient, the model is expressed in matrix form as: Y equals X times θ plus ϵ, where Y is the column vector of all observed outputs, X is the design matrix that stores all input values, θ is the vector of unknown coefficients, and ϵ is the vector of random errors.

Here, the vector ϵ is modeled as a random variable — typically assumed to follow a normal distribution centered at zero. This stochastic interpretation reflects the fact that measurements are never perfectly accurate and that real-world systems include inherent randomness.

By explicitly including this randomness in the model, statistical inference becomes possible: one can estimate the coefficients, assess the quality of the model, and make probabilistic predictions.

Classical Linear Regression: Notation and Assumptions

Model Notation: We denote the classical linear regression model by the triplet

$$(\mathrm{Y}, \mathrm{X}\theta, \Sigma),$$

where Σ is a fixed $\mathrm{N} \times \mathrm{N}$ covariance matrix of the random errors:

$$\Sigma = \mathrm{E}[\epsilon\epsilon^{\mathrm{T}}] = \|\mathrm{E}[\epsilon_{\mathrm{i}}\epsilon_{\mathrm{j}}]\|_{\mathrm{i,j}=1}^{\mathrm{N}}.$$

(1) **Unbiasedness:** $\mathrm{E}[\epsilon_{\mathrm{i}}] = 0$.
(2) **Homoscedasticity:** $\mathrm{E}[\epsilon_{\mathrm{i}}^2] = \sigma^2$.
(3) **Uncorrelated Errors:** $\mathrm{E}[\epsilon_{\mathrm{i}}\epsilon_{\mathrm{j}}] = 0$ for $\mathrm{i} \neq \mathrm{j}$.

(a) **Estimator Unbiasedness:** $\mathrm{E}[\hat{\theta}] = \theta$.
(b) **Minimum Variance:** For any vector z of appropriate dimension and any unbiased estimator $\tilde{\theta}$, the inequality

$$\mathrm{D}\left(\mathrm{z}^{\mathrm{T}}(\hat{\theta} - \theta)\right) \leq \mathrm{D}\left(\mathrm{z}^{\mathrm{T}}(\tilde{\theta} - \theta)\right)$$

holds, where D denotes the covariance matrix.
(c) **Linearity:** $\hat{\theta} = \mathrm{SY}$, where S is a matrix independent of Y.

Comments

From now on, to represent the classical regression model, we will use the following triplet: the observed response vector, the deterministic part involving the design matrix and parameters, and the covariance matrix of the error terms. This covariance matrix is fixed and encodes how errors vary and how they might be correlated.

As a rule, the following standard assumptions about errors and parameter estimates are accepted.

First, we assume errors have zero mean, which means that on average the errors do not systematically bias the observations. This ensures that our model predictions are centered around the true values rather than consistently over- or underestimating them.

Second, the errors are assumed to have equal variance, called homoscedasticity. This reflects the idea that the variability of errors is uniform across all observations, so no particular measurement is inherently more uncertain than ‘another. This assumption simplifies both estimation and inference.

Third, we assume errors are uncorrelated, meaning the error in one observation does not influence the error in another. This is crucial because correlations would imply some hidden structure or dependence in the noise, which requires more complex modeling.

Regarding parameter estimates, the goal is to find estimators that, on average, correctly recover the true parameters (unbiasedness). Among all unbiased estimators, we prefer those with minimal variance for any linear combination of parameters, ensuring the estimates are as precise as possible.

Finally, insisting that the estimator be a linear function of the observed data allows for elegant mathematical treatment and computational efficiency. This linearity condition means the estimator can be expressed as a fixed matrix multiplying the data vector, independent of the data realization itself.

Variance Minimization Condition and BLUE

Let's comment on condition (b).

- Let $\mathrm{D}_{\tilde{\theta}} = \mathrm{E}[(\tilde{\theta} - \theta)(\tilde{\theta} - \theta)^{\mathrm{T}}]$ denote the covariance matrix of an unbiased estimator $\tilde{\theta}$.
- For any vector z of compatible dimension,

$$\mathrm{D}\big(\mathrm{z}^{\mathrm{T}}(\tilde{\theta} - \theta)\big) = \mathrm{E}\big(\mathrm{z}^{\mathrm{T}}(\tilde{\theta} - \theta)\big)^2 = \mathrm{z}^{\mathrm{T}}\mathrm{D}_{\tilde{\theta}}\,\mathrm{z},$$

  where the last equality holds due to the unbiasedness of $\tilde{\theta}$.
- The variance minimization condition (b) implies:

$$\mathrm{z}^{\mathrm{T}}\mathrm{D}_{\hat{\theta}}\,\mathrm{z} \le \mathrm{z}^{\mathrm{T}}\mathrm{D}_{\tilde{\theta}}\,\mathrm{z},$$

  i.e., the matrix $\mathrm{D}_{\tilde{\theta}} - \mathrm{D}_{\hat{\theta}}$ is positive semi-definite.
- Note: An estimator satisfying this condition does not always exist.

Definition: Best Linear Unbiased Estimator (BLUE)

An estimator that satisfies conditions (a) unbiasedness, (b) minimal variance, and (c) linearity is called the Best Linear Unbiased Estimator (BLUE).

Comments

Let's now take a closer look at condition (b), the variance minimization requirement. At first glance, it may seem unclear why we're interested not simply in minimizing the variances of the individual components of the estimator vector $\hat{\theta}$, but instead in minimizing the variance of all possible linear combinations $\mathrm{z}^{\mathrm{T}}\hat{\theta}$, where z is any fixed vector.

Here's why. In many practical problems, we're not always interested in all components of the parameter vector individually. Sometimes we care about a specific function of the parameters — say, a contrast like θ_1 minus θ_2, or a predicted value at some point, which also ends up being a linear combination of parameters. Therefore, it is natural to assess the precision of $\hat{\theta}$ not just globally, but in every possible direction in parameter space — i.e., for all vectors z.

Condition (b) says: if we take any linear combination $\mathrm{z}^{\mathrm{T}}\hat{\theta}$, then its variance should be no greater than for any other unbiased estimator $\tilde{\theta}$. This guarantees that $\hat{\theta}$ is not just good on average, but it delivers the tightest possible confidence intervals for any estimable quantity derived from θ.

In matrix terms, this condition means that the difference between the covariance matrices of $\tilde{\theta}$ and $\hat{\theta}$ is a positive semi-definite matrix. That is: for all z, the quadratic form with this difference is non-negative.

An estimator satisfying conditions (a), (b), and (c) is called a BLUE — Best Linear Unbiased Estimator. It is the benchmark against which all other linear estimators are judged.

Example: Linear Regression Model

Consider a linear regression model of the form:

$$\eta(t) = a + bt.$$

This model describes, for instance, the change in length of a metal rod as a function of temperature.
Measurements are taken at points $t_1, \ldots, t_N$:

$$y_j = a + bt_j + \epsilon_j, \quad j = 1, \ldots, N.$$

We represent the model in matrix form:

$$Y = X\theta + \epsilon, \quad \theta = \begin{pmatrix} a \\ b \end{pmatrix}, \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_N \end{pmatrix},$$

$$X = \begin{pmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_N \end{pmatrix}, \quad X\theta = \begin{pmatrix} a + bt_1 \\ \vdots \\ a + bt_N \end{pmatrix}.$$

Comments

Let's consider a simple example. Here we're modeling how a metal rod's length changes with temperature - a classic physics experiment. We assume this relationship is linear and can be described by an equation where the outcome depends on the sum of two components: a constant parameter plus another parameter multiplied by temperature. Both parameters are unknown and need to be estimated.

We take measurements at different temperature points - from the first to the nth observation - obtaining corresponding measured values. Each measurement differs from the true value by some random error. Thus, each observed value equals the sum of: (1) the constant parameter, (2) temperature multiplied by the second parameter, and (3) random measurement error.

This model can be conveniently expressed in matrix form. Here, the parameter vector contains two elements: the intercept (constant term) and the temperature coefficient. The design matrix has two columns: one column of ones and another column of temperature values at each measurement. The observation vector contains all measured lengths, while the error vector collects all random deviations.

The matrix formulation offers key advantages: it lets us use standard linear algebra tools like pseudoinverse for parameter estimation, analyze estimator properties, construct confidence intervals, and test hypotheses. Moreover, it naturally extends to more complex models and simplifies computations with large datasets.

Definition

For a classical linear regression model $(Y, X\theta, \Sigma)$ with the nonsingular matrix X^TX, the vector

$$\widehat{\theta} = (X^TX)^{-1}X^TY$$

is called the *Ordinary Least Squares estimator (OLS estimator)*.

Example: Linear Model

Consider again the model describing the dependence of a metal rod's length on temperature:

$$y_j = a + bt_j + \epsilon_j, \quad j = 1, \ldots, N,$$

where $t_1, \ldots, t_N$ are given real numbers.
The matrix X^TX takes the form:

$$X^TX = \begin{pmatrix} 1 & \ldots & 1 \\ t_1 1 & \ldots & t_N \end{pmatrix} \begin{pmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_N \end{pmatrix} = \begin{pmatrix} N & \sum t_j \\ \sum t_j & \sum t_j^2 \end{pmatrix}.$$

Comments

Earlier, we introduced the empirical least squares estimate. For the mathematical model of linear regression represented by the triple $(Y, X\theta, \Sigma)$, we now define the standard Ordinary Least Squares estimator as: the inverse of X^TX, multiplied by X^T, and then by the vector Y. Note that in this case, the estimator is obtained as the solution to the system of normal equations. It is assumed here that the matrix X^TX is nonsingular. Later on, we will generalize this definition to the case where this matrix is singular.

Let us return to the example we considered earlier — the model describing the dependence of a metal rod's length on temperature. In this case, the matrix X^TX is nonsingular and has the following form: the number of observations N, the sum of t_j, the sum of t_j again, and the sum of squared t_j.

Example (continued)

Using the standard formula for the inverse of a 2×2 matrix, we obtain:

$$(X^T X)^{-1} = \begin{pmatrix} \frac{\sum t_j^2}{\Delta} & -\frac{\sum t_j}{\Delta} \\ -\frac{\sum t_j}{\Delta} & \frac{N}{\Delta} \end{pmatrix}, \quad \Delta = N \sum t_j^2 - \left(\sum t_j\right)^2.$$

$$X^T Y = \begin{pmatrix} \sum y_j \\ \sum y_j t_j \end{pmatrix}.$$

Hence, the OLS estimator is:

$$\begin{pmatrix} \widehat{a} \\ \widehat{b} \end{pmatrix} = (X^T X)^{-1} X^T Y = \begin{pmatrix} \frac{\sum t_j^2 \sum y_j - \sum t_j \sum y_j t_j}{\Delta} \\ \frac{N \sum y_j t_j - \sum t_j \sum y_j}{\Delta} \end{pmatrix}.$$

- This is the explicit form of the OLS estimates for the intercept and slope in the simple linear model.
- These formulas depend only on the sample sums and can be computed directly from the data.

Comments

Using the standard formula for the inverse of a two-by-two matrix, we obtain an explicit expression for the inverse of the matrix $X^T X$.

Multiplying the inverse of $X^T X$ with $X^T Y$, we obtain explicit formulas for the estimated intercept and slope — that is, the least squares estimates of the coefficients a and b in our linear model. These estimates depend only on basic sample statistics and are easy to compute in practice. They form the foundation of many methods used in applied regression analysis.

Theorem 1 (Gauss–Markov)

Consider the classical linear regression model $(Y, X\theta, \sigma^2 I_N)$, where:

- σ^2 is the common variance of the errors.
- I_N is the identity matrix of size N,
- the error vector ϵ satisfies the assumptions (1)–(3):
  (1) $E[\epsilon] = 0$,
  (2) $\mathrm{Cov}(\epsilon) = \sigma^2 I_N$,
  (3) the components of ϵ are uncorrelated.
- and the matrix $X^T X$ is nonsingular.

Then the vector

$$\widehat{\theta} = (X^T X)^{-1} X^T Y$$

is the *Best Linear Unbiased Estimator (BLUE)* of θ. In other words, it has the *minimum variance* among all linear unbiased estimators.

Its covariance matrix is given by:

$$D_{\widehat{\theta}} = \sigma^2 (X^T X)^{-1}.$$

Interpretation

Comments

On this slide, we present the Gauss–Markov theorem, which is a key result in the theory of linear regression. We consider the classical model, in which the error vector has zero mean, uncorrelated components, and equal variance — in other words, the covariance matrix of the errors is σ^2 times the identity matrix. In addition, we assume that the matrix $X^T X$ is invertible.

Under these conditions, the Ordinary Least Squares estimator, which is $(X^T X)^{-1} X^T Y$, turns out to be the best possible — that is, it is the most efficient among all linear and unbiased estimators. This is what we mean when we say that it is the best linear unbiased estimator, or BLUE.

The theorem also gives us the exact formula for the covariance matrix of the estimator: it equals σ^2 times the inverse of $X^T X$. This matrix reflects the precision of our estimates and plays a central role in evaluating statistical reliability.

In summary, this theorem not only guarantees the optimality of the estimator under the given assumptions, but also allows us to quantify the variability of the estimated parameters.

Lemma 2

The OLS estimator is a linear and unbiased estimator, i.e., it satisfies conditions (a) and (c):

(a) Estimator Unbiasedness: $\mathrm{E}[\hat{\theta}] = \theta$.

(c) Linearity: $\hat{\theta} = \mathrm{SY}$, where S is a matrix independent of Y.

Proof: Linearity of the OLS estimator follows directly from its form: the matrix $\mathrm{S} = (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}$.
To verify unbiasedness, consider the expectation: since $\mathrm{E}(\mathrm{Y}) = \mathrm{X}\theta$, we get $\mathrm{E}(\widehat{\theta}) = \mathrm{E}(\mathrm{SY}) = \mathrm{S} \cdot \mathrm{E}(\mathrm{Y}) = \mathrm{SX}\theta = (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{X}\theta = \theta$. Lemma is proved. □

Lemma 3

A linear estimator $\widetilde{\theta} = \mathrm{AY}$ is unbiased if and only if $\mathrm{AX} = \mathrm{I}$.

Comment

Lemma 3 is the key criterion for verifying the unbiasedness of any linear estimator. It will be used in the proof of the Gauss–Markov theorem.

Comments

This slide presents two supporting lemmas that we need in order to prove the Gauss–Markov theorem.

The first lemma states that the OLS estimator is both linear and unbiased. Linearity means that the estimator is equal to a fixed matrix times the observation vector Y. In this case, the matrix is defined as the inverse of $\mathrm{X}^{\mathrm{T}}\mathrm{X}$, multiplied by X^{T}. This property is important because it means the estimator reacts to data in a predictable, linear way.

To verify that the estimator is unbiased, we consider the expected value of the estimator. Due to the fact that the expected value of error is zero, we have that the expected value of Y is equal to X times θ. When we substitute this into the formula for the estimator, all the matrices cancel out, and we obtain exactly θ. This confirms that the estimator is unbiased.

The second lemma gives us a general condition for checking whether a linear estimator is unbiased. Suppose we define an estimator as a matrix A times the vector Y. Then this estimator is unbiased if and only if the product of A and the matrix X is equal to the identity matrix.

This second lemma will serve as the main technical tool in the proof of the Gauss–Markov theorem, which we will see shortly.

Proof of Lemma 3

Proof:

Sufficiency: Assume $AX = I$. Then:

- The estimator has the form: $\widetilde{\theta} = AY$.
- Since $E(Y) = X\theta$, we have:

$$E[\widetilde{\theta}] = AE[Y] = AX\theta = \theta.$$

- Hence, $\widetilde{\theta}$ is unbiased.

Necessity: Suppose $\widetilde{\theta} = AY$ is unbiased.

- Then $E[\widetilde{\theta}] = AX\theta = \theta$ for all $\theta \in \mathbb{R}^m$.
- This implies that:

$$AX\theta = \theta \quad \text{for all } \theta.$$

- Let us test this identity on the standard basis vectors:

$$\theta = e_i = (0, \dots, 0, 1, 0, \dots, 0)^T \text{ with one in the i-th place..}$$

- Then $AXe_i = e_i$, so the i-th column of AX equals the i-th column of the identity matrix.
- Repeating this for all $i = 1, \dots, m$, we conclude:

$$AX = I.$$

The Lemma is proved. □

Comments

This slide presents a full proof of Lemma 3, which defines the necessary and sufficient conditions for the estimator to be unbiased.

We begin with sufficiency. Suppose that the matrix A multiplied by X equals the identity. Then the estimator AY has expectation equal to A times the expected value of Y. Since the expectation of Y is X times θ, we get $AX\theta$, which is simply θ. Therefore, the estimator is unbiased.

For necessity, we start from the assumption that AY is an unbiased estimator. This means that the expectation of AY must be equal to θ for all possible values of θ. In other words, $AX\theta$ equals θ for any vector θ.

To show that this leads to AX being the identity matrix, we substitute standard basis vectors one by one into the expression. Each substitution tells us that one column of AX must match the identity matrix. After going through all components, we conclude that AX equals the identity.

This completes the proof.

Lemma 4

Under the assumptions of the Gauss–Markov Theorem, the covariance matrix of any linear unbiased estimator $\widetilde{\theta} = \mathrm{AY}$ has the form:

$$\mathrm{D}_{\widetilde{\theta}} = \sigma^2 \mathrm{A}\mathrm{A}^{\mathrm{T}}.$$

In particular, for the OLS estimator:

$$\mathrm{D}_{\widehat{\theta}} = \sigma^2 (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}.$$

Proof:

Consider the definition of the covariance matrix:

$$\mathrm{D}_{\widetilde{\theta}} = \mathrm{E}[(\widetilde{\theta} - \mathrm{E}[\widetilde{\theta}])(\widetilde{\theta} - \mathrm{E}[\widetilde{\theta}])^{\mathrm{T}}].$$

Since $\widetilde{\theta} = \mathrm{AY}$ and $\mathrm{E}[\widetilde{\theta}] = \theta$, we have:

$$\mathrm{D}_{\widetilde{\theta}} = \mathrm{E}[(\mathrm{AY} - \theta)(\mathrm{AY} - \theta)^{\mathrm{T}}].$$

Comments

This lemma establishes the form of the covariance matrix of a linear, unbiased estimator under the same assumptions as in the Gauss–Markov theorem.

Since the estimator is linear, it can be written as a matrix A multiplied by the vector of observations Y. Furthermore, the estimator is unbiased, which means that the expected value of A times Y equals the true parameter vector θ.

To find the covariance matrix of this estimator, we recall the general formula: it's the expected value of the deviation from the mean, multiplied by its own transpose.

In our case, the deviation is A times Y minus θ. So the covariance matrix becomes the expected value of that expression times its transpose.

Variance of Linear Unbiased Estimators (continued)

Since $\mathrm{E}[\mathrm{AY}] = \mathrm{E}[\mathrm{Y}^{\mathrm{T}}\mathrm{A}^{\mathrm{T}}]^{\mathrm{T}} = \theta$ we have:

$$\mathrm{E}[(\mathrm{AY}-\theta)(\mathrm{AY}-\theta)^{\mathrm{T}}] = \mathrm{E}[\mathrm{AYY}^{\mathrm{T}}\mathrm{A}^{\mathrm{T}}] - \theta\theta^{\mathrm{T}}.$$

To compute $\mathrm{E}[\mathrm{AYY}^{\mathrm{T}}\mathrm{A}^{\mathrm{T}}]$, substitute the expression for Y:

$$\mathrm{Y} = \mathrm{X}\theta + \varepsilon \quad \Rightarrow \quad \mathrm{YY}^{\mathrm{T}} = (\mathrm{X}\theta+\varepsilon)(\mathrm{X}\theta+\varepsilon)^{\mathrm{T}}.$$

Then:

$$\mathrm{E}[\mathrm{AYY}^{\mathrm{T}}\mathrm{A}^{\mathrm{T}}] = \mathrm{AE}[(\mathrm{X}\theta+\varepsilon)(\mathrm{X}\theta+\varepsilon)^{\mathrm{T}}]\mathrm{A}^{\mathrm{T}}.$$

Using independence and zero mean of ε, we get:

$$\mathrm{E}[\mathrm{YY}^{\mathrm{T}}] = \mathrm{X}\theta\theta^{\mathrm{T}}\mathrm{X}^{\mathrm{T}} + \sigma^2\mathrm{I} \quad \Rightarrow \quad \mathrm{E}[\mathrm{AYY}^{\mathrm{T}}\mathrm{A}^{\mathrm{T}}] = \mathrm{AX}\theta\theta^{\mathrm{T}}\mathrm{X}^{\mathrm{T}}\mathrm{A}^{\mathrm{T}} + \sigma^2\mathrm{AA}^{\mathrm{T}}.$$

Since $\mathrm{AX} = \mathrm{I}$, this simplifies to:

$$\mathrm{E}[\mathrm{AYY}^{\mathrm{T}}\mathrm{A}^{\mathrm{T}}] = \theta\theta^{\mathrm{T}} + \sigma^2\mathrm{AA}^{\mathrm{T}}.$$

Subtracting $\theta\theta^{\mathrm{T}}$, we conclude:

$$\mathrm{D}_{\widetilde{\theta}} = \sigma^2\mathrm{AA}^{\mathrm{T}}.$$

For the OLS estimator, where $\mathrm{A} = (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}$, we have:

$$\mathrm{D}_{\widehat{\theta}} = \sigma^2(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{X}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1} = \sigma^2(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}. \qquad \square$$

Comments

To complete the proof, note that since the expected value of AY equals θ we can rewrite the covariance matrix of the estimator as expectation of $\mathrm{AYY}^{\mathrm{T}}\mathrm{A}^{\mathrm{T}}$ minus $\theta\theta^{\mathrm{T}}$.

To compute this, we use the fact that the vector of observations Y can be written as X times θ plus ϵ, where ϵ is the vector of random errors. Then the matrix Y times Y^{T} becomes the product of X times θ plus ϵ with its transpose. Taking expectation of the resulting expression, and using linearity along with the assumption that the error vector ϵ has zero mean and is uncorrelated with the regressors, we obtain: expectation of Y times Y^{T} equals X times θ times θ^{T} times X^{T} plus σ^2 times the identity matrix.

Multiplying by A from the left and by A^{T} from the right, we get that the desired expectation is equal to A times X times θ times θ^{T} times X^{T} times A^{T} plus σ^2 times A times A^{T}.

Now, since we assumed that A times X equals the identity matrix, this simplifies to θ times θ^{T} plus σ^2 times A times A^{T}.

Subtracting the outer product θ times θ^{T}, we arrive at the final result: the covariance matrix of the estimator equals σ^2 times A times A^{T}.

Finally, in the case of the OLS estimator, the matrix A is given by the inverse of $\mathrm{X}^{\mathrm{T}}\mathrm{X}$, all multiplied by X^{T}. Substituting this into the general formula and simplifying, we get that the covariance matrix of the OLS estimator is equal to σ^2 times the inverse of $\mathrm{X}^{\mathrm{T}}\mathrm{X}$, which completes the proof.

Proof of Theorem 1:
Let us verify the matrix inequality:

$$\mathrm{D}_{\widehat{\theta}} \leq \mathrm{D}_{\widetilde{\theta}}.$$

Denote $\mathrm{S} = (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}$, so that $\widehat{\theta} = \mathrm{SY}$. For any linear unbiased estimator $\widetilde{\theta} = \mathrm{AY}$, we compute:

$$\mathrm{D}_{\widetilde{\theta}} = \mathrm{D}(\mathrm{AY}) = \mathrm{D}((\mathrm{A}-\mathrm{S})\mathrm{Y} + \mathrm{SY}).$$

Using the identity $\mathrm{D}(\mathrm{U}+\mathrm{V}) = \mathrm{D}(\mathrm{U}) + \mathrm{D}(\mathrm{V})$ when U and V are uncorrelated, we obtain:

$$\mathrm{D}_{\widetilde{\theta}} = \mathrm{D}((\mathrm{A}-\mathrm{S})\mathrm{Y}) + \mathrm{D}(\mathrm{SY}) = \mathrm{D}((\mathrm{A}-\mathrm{S})\mathrm{Y}) + \mathrm{D}_{\widehat{\theta}} \geq \mathrm{D}_{\widehat{\theta}}.$$

The cross term vanishes because $(\mathrm{A}-\mathrm{S})\mathrm{X} = \mathrm{I} - \mathrm{I} = 0$, and thus:

$$\mathrm{E}[(\mathrm{A}-\mathrm{S})\mathrm{YY}^{\mathrm{T}}\mathrm{S}^{\mathrm{T}}] = (\mathrm{A}-\mathrm{S})\mathrm{E}[\mathrm{YY}^{\mathrm{T}}]\mathrm{S}^{\mathrm{T}} = (\mathrm{A}-\mathrm{S})(\mathrm{X}\theta\theta^{\mathrm{T}}\mathrm{X}^{\mathrm{T}} + \sigma^2\mathrm{I})\mathrm{S}^{\mathrm{T}} = 0.$$

This concludes the proof. □

Comments

Now we return to the proof of the Gauss–Markov Theorem.Our goal is to show that the covariance matrix of the ordinary least squares estimator is less than or equal to the covariance matrix of any other linear unbiased estimator — in the sense of matrix inequality. This means that the OLS estimator is the best, in the sense of having minimal variance, within the class of all linear unbiased estimators.

We denote the OLS matrix by capital S, where S equals the inverse of $\mathrm{X}^{\mathrm{T}}\mathrm{X}$, multiplied by X^{T}. Then the OLS estimator, $\hat{\theta}$, is equal to S times Y.

Now consider any linear unbiased estimator, denoted by $\tilde{\theta}$, which has the form A times Y. We rewrite this as: A times Y equals the sum of two parts — namely, A minus S times Y, plus S times Y.

Then, the covariance matrix of $\tilde{\theta}$ equals the covariance of the first term, A minus S times Y, plus the covariance of the second term, S times Y.

We are allowed to add the covariances because the two components are uncorrelated. This is due to the fact that the matrix A minus S multiplied by X gives zero. Therefore, A minus S times Y is uncorrelated with S times Y, and the mixed covariance term vanishes.

As a result, the covariance matrix of any linear unbiased estimator is always greater than or equal to that of the OLS estimator. This proves that the OLS estimator has the smallest possible variance among all linear unbiased estimators. The theorem is thus proven.

Example: Exact Data and Quadratic Approximation

Example

$$y = \frac{2}{4 - 3x}$$

This true function is known, but used only for analyzing model performance — not for building the model.

Measurement Points:

- $x_1 = 0, \quad y_1 = 1/2$
- $x_2 = \frac{2}{3}, \quad y_2 = 1$
- $x_3 = 1, \quad y_3 = 2$

Model: Quadratic Regression (No Measurement Error)

$$\eta(x, \theta) = \theta_0 + \theta_1 x + \theta_2 x^2$$

Vectors and Matrix:

$$Y = \begin{pmatrix} 1/2 \\ 1 \\ 2 \end{pmatrix}, \quad X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & \frac{2}{3} & \frac{4}{9} \\ 1 & 1 & 1 \end{pmatrix}$$

Comments

In this example, we consider a situation where the true dependence between input and output is exactly known and given by the function: y equals two divided by the quantity four minus three x. This information is not used to construct the model — it is included solely to evaluate its accuracy.

We conduct measurements at three input points: x equals zero, x equals two-thirds, and x equals one. Evaluating the true function at these points, we obtain the corresponding outputs: one-half, one, and two. These values form the observation vector Y.

The model we use for approximation is a quadratic regression of the form: θ_0 plus θ_1 times x plus θ_2 times x squared. This means we assume that the data can be well-approximated by a second-degree polynomial.

Additionally, we assume that there are no measurement errors. Of course, this is an artificial assumption made only to simplify the analysis.

The design matrix X is constructed by applying the basis functions — constant, linear, and quadratic — to each of the three input values. At x equals zero, the row is: one, zero, zero. At x equals two-thirds, the row is: one, two-thirds, and four-ninths. At x equals one, the row is: one, one, one. These will be used to compute the least squares estimate and analyze how well the model fits the data.

Example (continued)

To compute the OLS estimator, we start by finding the matrix $\mathrm{X}^{\mathrm{T}}\mathrm{X}$:

$$\mathrm{X}^{\mathrm{T}}\mathrm{X} = \begin{pmatrix} 3 & \frac{5}{3} & \frac{13}{9} \\ \frac{5}{3} & \frac{13}{9} & \frac{35}{27} \\ \frac{13}{9} & \frac{35}{27} & \frac{97}{81} \end{pmatrix}, \quad (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1} = \begin{pmatrix} 1 & -\frac{5}{2} & \frac{3}{2} \\ -\frac{5}{2} & \frac{61}{2} & -30 \\ \frac{3}{2} & -30 & \frac{63}{2} \end{pmatrix}.$$

Then, the least squares estimator is:

$$\widehat{\theta} = (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{Y} = \begin{pmatrix} 1/2 \\ -3/4 \\ 9/4 \end{pmatrix}.$$

Thus, the original function $\frac{2}{4-3x}$ is approximated by a quadratic function obtained via the method of least squares.

Note: This quadratic approximation matches the Lagrange interpolating polynomial for the same nodes.

Comments

Now let's compute the least squares estimates step by step. First, we multiply the design matrix X by its transpose - this gives us the matrix $\mathrm{X}^{\mathrm{T}}\mathrm{X}$ shown here. Then we calculate the inverse matrix.

Finally we multiply this inverse by $\mathrm{X}^{\mathrm{T}}\mathrm{Y}$ (which combines our observations with the basis functions), we get our parameter estimates: θ_0 equals one-half, θ_1 equals minus three-fourths, and θ_2 equals nine-fourths.

These coefficients define our quadratic approximation. Interestingly, this exact same polynomial would emerge if we used Lagrange interpolation - but here we derived it through least squares, which is more general. The perfect fit at our three points was guaranteed because we have three parameters and three exact observations - but remember, this is a special case with no measurement errors.

Example: Least Squares Estimation (Figure)

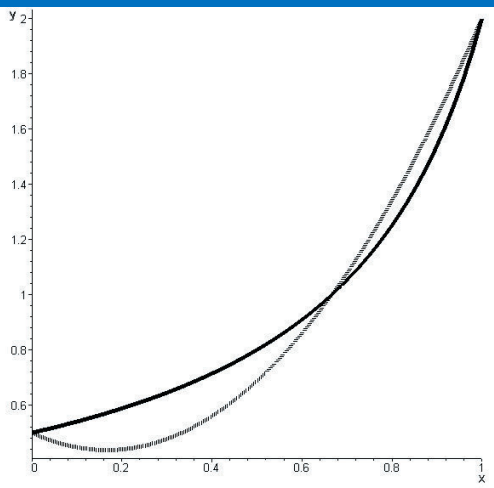

Figure: Approximation of the function $y = \frac{2}{4-3x}$ (solid line) by the function $\bar{y} = 1/2 - 3/4x + 9/4x^2$ (dashed line) constructed using the least squares method.

Comments

This figure shows the result of approximating the function y equals two divided by the quantity four minus three x using the least squares method. The original nonlinear function is drawn with a solid curve, while the dashed curve represents the quadratic approximation defined by the formula: $\bar{y}$ equals one half minus three fourths x plus nine fourths x squared.

As we have already noted, in this particular case the least squares approximation coincides with the Lagrange interpolating polynomial. However, this is not a general feature of least squares methods. It occurs here only because the number of parameters in the model is equal to the number of observations and there are no observation errors in our measurements - we're using the exact function values. Let me emphasize that in the general case, the least squares method does not interpolate the data. Instead, it finds the function that minimizes the total squared error between the predicted values and the observed ones. That is the essence of least squares approximation — it aims not to match the data exactly, but to provide the best possible overall fit in terms of squared deviations.

A key point to emphasize is that the result of a least squares approximation depends not only on the form of the model but also on the locations of the data points. If we had chosen different x values, even with the same model, we would have obtained different coefficients. Thus, the design of the experiment — that is, the selection of input values at which the system is observed — plays a crucial role in the quality of the approximation.

This leads us naturally to the topic of optimal experimental design, which studies how to choose data points to achieve the most informative or precise estimates. In practical applications, especially when experiments are costly or time-consuming, choosing the right points can significantly improve the accuracy of estimation. We will explore this topic in more detail when we study optimality criteria and planning strategies in regression analysis.

Estimability of Linear Parametric Functions

Definition: OLS

A linear parametric function $\tau = \mathrm{T}\theta$ $(\mathrm{T} \in \mathbb{R}^{k\times m},\ k \in \{1,\dots,m\})$ is called estimable if there exists an unbiased estimator of the form $\widehat{\tau} = \mathrm{A}\mathrm{Y}$ with $\mathrm{A} \in \mathbb{R}^{k\times N}$.

Lemma 5: Equivalent Conditions

Consider the classical linear regression model $\mathrm{Y} = \mathrm{X}\theta + \varepsilon$ under the assumption that $\mathbb{E}[\varepsilon] = 0$. Then the following statements are equivalent:

(a) Each row of T lies in the row space of X, i.e., $\mathcal{L}(\mathrm{T}^{\mathrm{T}}) \subseteq \mathcal{L}(\mathrm{X}^{\mathrm{T}}) = \mathcal{L}(\mathrm{X}^{\mathrm{T}}\mathrm{X})$;

(b) The value of $\mathrm{T}\widehat{\theta}$ is the same for all solutions of the system of normal equations $\mathrm{X}^{\mathrm{T}}\mathrm{X}\theta = \mathrm{X}^{\mathrm{T}}\mathrm{Y}$;

(c) The parametric function $\tau = \mathrm{T}\theta$ is estimable.

Note: Estimability ensures that a parametric function $\tau = \mathrm{T}\theta$ can be recovered from data without bias, using only the information contained in the design matrix X.

Comments

We now generalize the classical estimation problem. Previously, we estimated the full parameter vector θ, but in practice, we are often interested in specific functions of parameters—for example, differences between them or averages. In this context, we define a parametric function τ as T times θ, where T is a known matrix of size k by m. The number k indicates how many different linear combinations of parameters we are consider.

Let us now state the formal definition. We say that a parametric function $\tau = \mathrm{T}\theta$ is estimable if there exists an unbiased linear estimator of the form $\hat{\tau} = \mathrm{A}\mathrm{Y}$, where A is a matrix of size k by N. That is, $\hat{\tau}$ is a linear function of the data, and its expected value equals τ.

This is not just a technical detail. In many statistical applications, we cannot recover the full vector θ, but we can still estimate certain combinations of its components—and this definition tells us what it means to do so without bias.

The lemma on the slide provides necessary and sufficient conditions for estimability in the classical linear model $\mathrm{Y} = \mathrm{X}\theta + \epsilon$, assuming the error vector has zero expectation. It states that the following three statements are equivalent:

The rows of T lie in the row space of X.

The value of $\mathrm{T}\hat{\theta}$ does not depend on the particular solution $\hat{\theta}$ of the system of normal equations.

The function $\tau = \mathrm{T}\theta$ is estimable.

This lemma gives us a complete characterization of which parametric functions can be estimated without bias in the linear model.

Remark

We make no assumptions about the rank of the matrix X. Therefore, Lemma 5 holds even if the matrix $X^T X$ is singular.

Proof Strategy:

- We demonstrate the proof for the case $k = 1$; the general case follows similarly.
- The proof proceeds by showing a cyclic implication:

$$\textbf{(a)} \Rightarrow \textbf{(b)} \Rightarrow \textbf{(c)} \Rightarrow \textbf{(a)}.$$

Step 1: Proving (a) ⇒ (b)

- Assume **(a)**: $T^T \in \mathcal{L}(X^T X)$. This implies there exists a vector p such that $T^T = X^T X p$.
- For any solution $\widehat{\theta}$ of the system normal equations $X^T X\widehat{\theta} = X^T Y$, we have:

$$T\widehat{\theta} = (X^T X p)^T \widehat{\theta} = p^T X^T X\widehat{\theta} = p^T X^T Y.$$

- Since $p^T X^T Y$ does not depend on $\widehat{\theta}$, the value $T\widehat{\theta}$ is unique, satisfying condition **(b)**.

Comments

Let us now prove the lemma that establishes equivalence between three conditions of estimability for a linear parametric function of the form τ equals $T\theta$. We begin with an important observation: we do not assume anything about the rank of the matrix X. Thus, the lemma remains valid even when the matrix $X^T X$ is singular, that is, when the design matrix has linearly dependent columns.

We present the proof for the case when the number of components in τ equals one, that is, k equals one. The general case is proved analogously. The proof proceeds along a cycle of implications: condition (a) implies (b), which implies (c), which in turn implies (a).

First, suppose that the transpose of T lies in the column space of the matrix $X^T X$. This means there exists a vector p such that T^T equals $X^T X$ times p. Now, let $\hat{\theta}$ be any solution of the system of normal equations.

Then, T times $\hat{\theta}$ equals p^T times $X^T X$ times $\hat{\theta}$. But this simplifies to p^T times X^T times Y. Thus, the value of T times $\hat{\theta}$ does not depend on the particular choice of $\hat{\theta}$, which proves that condition (b) holds.

Proof of Lemma: Estimability Conditions (continued)

Step 2: Proving (b) ⇒ (c)

- For any λ such that $X^TX\lambda = 0$ we have that if $\widehat{\theta}$ is a solution of the system normal equations ($X^TX\widehat{\theta} = X^TY$) then $\bar{\theta} = \widehat{\theta} - \lambda$ is also the solution.
- Condition (b) implies $T\widehat{\theta} = T\bar{\theta}$, so we have that $0 = T\widehat{\theta} - T\bar{\theta} = T\lambda$
- since this equality holds for any λ which satisfies $X^TX\lambda = 0$, we have that $T \in \mathcal{L}(X^TX)$
- this means that there exist a vector p such that $T^T = X^TXp$ and

$$ET\widehat{\theta} = p^TX^TEY = p^TX^TX\theta = T\theta,$$

- Thus, $\tau = T\theta$ is an estimable function, satisfying condition (c).

Step 3: Proving (c) ⇒ (a)

- Suppose the function $\tau = T\theta$ is estimable.
- By definition, there exists a matrix A such that $\mathbb{E}[AY] = T\theta$.
- Thus we have $AX\theta = T\theta$, which implies $AX = T$ (see the proof of Lemma 3).
- Consequently, $T^T \in \mathcal{L}(X^T)$, which is equivalent to condition **(a)**.

The Lemma is proved. □

Comments

We will now show that condition (b) implies condition (c), and then we will complete the logical cycle by proving that condition (c) implies condition (a).

Assume condition (b) holds. This means that the value of T times $\hat{\theta}$ is the same for all solutions of the normal equation. Recall that the normal equation is $X^TX\hat{\theta} = X^TY$. If this system has more than one solution, then any other solution has the form $\hat{\theta}$ minus λ, where λ is a vector that satisfies $X^TX\lambda = 0$.

Now, since both $\hat{\theta}$ and $\hat{\theta}$ minus λ are valid solutions, and the value of T times $\hat{\theta}$ must be the same for both, we subtract the two expressions and get T times λ equals zero. And this must hold for any vector λ such that $X^TX\lambda$ equals zero.

This condition tells us that there exists a vector p such that T^T equals X^TX times p. In other words, T^T can be expressed as a linear combination of the columns of X^TX.

Now consider the expectation of T times $\hat{\theta}$. Since $\hat{\theta}$ is a solution to the normal equation, we have: the expectation of T times $\hat{\theta}$ equals p^T times X^T times the expectation of Y. Because the expectation of Y is equal to X times θ, this becomes: p^T times X^T times X times θ, which simplifies to T times θ.

This shows that T times θ is an unbiased linear estimator, so the function is estimable. Hence, condition (c) is satisfied.

To complete the proof, we now assume that τ equals $T\theta$ is estimable. Then by definition, there exists a matrix A such that the expected value of A times Y equals $T\theta$. This implies that A times X equals T. Taking transposes, we obtain that T^T belongs to the column space of X^T, which implies condition (a).

Thus, the equivalence of all three conditions is proven.

Theorem 2 (BLUE for τ)

If the matrix $\mathrm{X}^{\mathrm{T}}\mathrm{X}$ is non-singular and the standard assumptions of the linear regression model are satisfied, then for any matrix T of compatible dimension:

- The function $\tau = \mathrm{T}\theta$ is estimable.
- Its best linear unbiased estimator (BLUE) is

$$\widehat{\tau} = \mathrm{T}\widehat{\theta} = \mathrm{T}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{Y}.$$

- The covariance matrix of the estimator $\widehat{\tau}$ is

$$\mathrm{D}_{\widehat{\tau}} = \sigma^2\mathrm{T}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{T}^{\mathrm{T}}.$$

Model assumptions:

- The model is $\mathrm{Y} = \mathrm{X}\theta + \varepsilon$, where ε is a random error vector;
- $\mathbb{E}[\varepsilon] = 0, \quad \mathrm{Var}(\varepsilon) = \sigma^2\mathrm{I}$;
- The matrix X has full column rank: $\mathrm{rank}(\mathrm{X}) = \mathrm{k}$.

Comments

From the previously proven lemma and the Gauss–Markov theorem, we now obtain an important result concerning the estimation of linear functions of the parameter vector θ when the design matrix X has full column rank.

The theorem states the following: if the matrix $\mathrm{X}^{\mathrm{T}}\mathrm{X}$ is invertible and all classical assumptions of the linear regression model are satisfied, then for any matrix T of compatible dimensions, the linear function τ equals $\mathrm{T}\theta$ is estimable. Moreover, its best linear unbiased estimator — also known as the BLUE — is given by T times $\hat{\theta}$, where $\hat{\theta}$ equals $(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{Y}$ is the usual least squares estimator.

The covariance matrix of this estimator is given by σ^2 times T times $(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}$ times T^{T}, which shows that the uncertainty in estimating τ depends both on the design matrix X and the matrix T.

It is important to note that this result relies on the invertibility of the matrix $\mathrm{X}^{\mathrm{T}}\mathrm{X}$, which corresponds to the assumption that the columns of X are linearly independent. This condition is necessary for the uniqueness of the least squares estimator. Otherwise, not all model parameters can be unbiased estimated.

In the next part, we will generalize this result to the case when the matrix $\mathrm{X}^{\mathrm{T}}\mathrm{X}$ is singular, that is, when the model is not of full rank.

PART III. Optimal design theory
(LECTURE 2)

Shpilev Petr Valerievich

Faculty of Mathematics and Mechanics, SPbU

September, 2025

Санкт-Петербургский государственный университет

Comments

In this lecture, we extend the theory of regression estimation to more general and challenging settings. We begin with the ordinary least squares (OLS) estimator in the singular case, introducing the concept of the generalized inverse, its definition, existence, and key properties, with special attention to the Moore–Penrose inverse. This framework allows us to address estimation when the design matrix is singular and to revisit the estimability of linear parametric functions, illustrated through a detailed example from weighing experiments.

We then turn to maximum likelihood estimation (MLE) in linear regression, proving its equivalence with OLS under normal assumptions and highlighting classical results such as the IID normal sample as a special case. The lecture continues with the generalized linear regression model and the derivation of the generalized least squares (GLS) estimator, which accommodates correlated and heteroscedastic errors.

Finally, we consider regression under additional structural information, including prior knowledge of parameters, linear constraints, and estimation within restricted parameter sets. This leads to the study of minimax estimators under quadratic constraints, their special cases, and their optimality in the class of unbiased estimators. The lecture concludes with a Bayesian perspective, introducing estimators based on a known prior distribution.

OLS Estimation in the Singular Case

Motivation

In many applications, including analysis of variance (ANOVA), the matrix $X^T X$ is singular. Therefore, generalizing OLS estimation to this case is essential.

- So far, we have assumed that the matrix $X^T X$ is non-singular.
- Now we address the case where the matrix $X^T X$ is singular, i.e., $\operatorname{rank}(X^T X) < m$.
- In this case, the normal equation

$$X^T X\theta = X^T Y$$

still has solutions, but not a unique one.

- Any vector $\widehat{\theta}$ that satisfies the normal equation minimizes the residual sum of squares $\|Y - X\theta\|^2$, and is called an OLS estimator.

Key Point

In the singular case, the OLS estimator exists but is not unique: the solution set of the normal equations is infinite. These solutions can be represented using generalized inverse matrices.

Comments

Let us now transition to the general case where the matrix $X^T X$ is singular. This situation arises frequently in applied settings, especially in analysis of variance, where the design matrix includes linearly dependent columns. Such linear dependencies make the matrix $X^T X$ non-invertible.

Despite this complication, we still define the ordinary least squares estimator as any vector $\widehat{\theta}$ that minimizes the squared norm of $Y - X\theta$. According to a previously proven Lemma 1, this vector $\widehat{\theta}$ satisfies the normal equation: $X^T X\widehat{\theta} = X^T Y$. However, because the matrix on the left is singular, the normal equation no longer has a unique solution. Instead, it has infinitely many solutions.

Each of these solutions provides the same minimum value of the residual sum of squares, but the vector $\widehat{\theta}$ itself is not uniquely defined. This distinguishes the singular case from the non-singular one, where the least squares estimator is given explicitly by the inverse of $X^T X$ times $X^T Y$.

This transition marks an important shift in focus. We no longer seek a unique estimator, but instead investigate which functions of θ remain estimable in the presence of singularity. This perspective will guide the development of generalized estimation theory for linear models.

Definition

Let $A \in \mathbb{R}^{n\times m}$. A matrix $A^- \in \mathbb{R}^{m\times n}$ is called a generalized inverse of A if, for every vector $y \in \mathbb{R}^n$ such that the system $Ax = y$ is consistent, the vector $x = A^-y$ is a solution.

Remarks

- Generalized inverse matrices are not unique.
- Any matrix admits at least one generalized inverse.
- Generalized inverses allow representation of all solutions of consistent linear systems.

Moore–Penrose Pseudoinverse

A matrix $A^+ \in \mathbb{R}^{m\times n}$ is called the Moore–Penrose pseudoinverse of A if it satisfies all four Penrose conditions:

$$(1)\quad AA^+A = A, \qquad (3)\quad (AA^+)^\top = AA^+,$$
$$(2)\quad A^+AA^+ = A^+, \qquad (4)\quad (A^+A)^\top = A^+A.$$

The pseudoinverse always exists and is unique.

Comments

We now introduce the notion of a generalized inverse. Let A be a real matrix of size $n \times m$. A matrix A^- of size $m \times n$ is called a generalized inverse of A if it maps every vector y in $\mathbb{R}^n$, for which the equation $Ax = y$ is consistent, to a solution $x = A^-y$. This concept is used to represent all possible solutions of linear systems in the case when A is not invertible or not square. It is important to note that generalized inverses are not uniquely defined — a matrix may admit infinitely many generalized inverses.

Among generalized inverses, a particularly important role is played by the Moore–Penrose pseudoinverse, denoted by A^+. This matrix is uniquely defined for any real matrix A and satisfies four algebraic conditions, known as the Penrose equations. These conditions ensure symmetry and minimality properties that make the pseudoinverse especially valuable in linear estimation theory. In particular, the pseudoinverse yields the solution of minimal Euclidean norm to a consistent linear system.

Every Moore–Penrose pseudoinverse is a generalized inverse, but not every generalized inverse is a pseudoinverse. This distinction is essential in the theory of linear models, where various estimation criteria may select different generalized inverses.

Lemma 6: Condition for Generalized Inverse

For a matrix B to be a *generalized inverse* of matrix A, it is necessary and sufficient that B satisfies the equality:

$$ABA = A.$$

Proof:

Sufficiency:

- Assume a generalized inverse $B = A^-$ exists for matrix A.
- Consider the i-th column of matrix A, denoted as a_i.
- The system $Ax = a_i$ is obviously consistent, as a solution exists (e.g., $x = e_i$).
- By definition, the vector $x = A^-a_i$ is a particular solution to this system.
- Therefore, substituting this solution into the system gives:

$$AA^-a_i = a_i, \quad \text{for all } i.$$

- This equality, holding for every column a_i of A, directly implies that $AA^-A = A$.

Comments

This lemma provides a necessary and sufficient condition for a matrix to be a generalized inverse. Specifically, a matrix B is a generalized inverse of a matrix A if and only if the product $ABA = A$.

To prove sufficiency, assume that a generalized inverse exists. Denote this inverse by A^-. We examine each column of the matrix A — let a_i be the i-th column. The system $Ax = a_i$ is clearly consistent; for example, it is solved by the i-th standard basis vector.

According to the definition, the product A^-a_i must also be a solution. Substituting this into the original system yields: $AA^-a_i = a_i$. Since this holds for every column of A, we conclude that $AA^-A = A$.

This condition captures the essential role of the generalized inverse: it reproduces any vector in the image of A via this triple product. It also demonstrates that the generalized inverse does not require the matrix to be square or invertible.

Generalized Inverse: Proof (continued)

Necessary ($ABA = A \implies \exists\, B = A^-$):

- Now, suppose the system $\mathbf{Ax = y}$ is consistent, and matrix B satisfies the equality $\mathbf{ABA = A}$.
- Multiply both sides of this equality by x from the right:

$$\mathbf{AB}\underbrace{\mathbf{Ax}}_{\mathbf{y}} = \underbrace{\mathbf{Ax}}_{\mathbf{y}} \quad \Rightarrow \quad \mathbf{ABy = y}.$$

- This result means that the vector $\mathbf{x = By}$ is a solution to the consistent system $\mathbf{Ax = y}$.
- By definition, if for every consistent system Ax = y, By is a solution, then matrix B is a *generalized inverse* for A.
- Thus, $B = A^-$.

□

Remark

The condition ABA = A is fundamental. Unlike the regular inverse, generalized inverses exist for any matrix, regardless of its shape or rank. This ensures that if a linear system Ax = y has a solution, By will provide one.

Comments

Now we prove the necessity. Suppose that the matrix B satisfies the condition ABA = A. Consider a consistent system Ax = y. That means there exists a vector x for which Ax = y.

Multiplying both sides of the identity ABA = A on the right by this vector x, we obtain ABAx = Ax. Since Ax = y, this implies that ABy = y. Hence, the vector x = By is a solution of the system Ax = y.

Therefore, the matrix B provides a solution to every consistent system. By definition, this means that B is a generalized inverse of A.

In conclusion, the condition ABA = A is both necessary and sufficient for B to be a generalized inverse.

The remark at the end emphasizes the generality of this concept. Unlike the ordinary inverse, which only exists for square and nonsingular matrices, a generalized inverse exists for any matrix, regardless of its dimensions or rank. It guarantees that for every consistent system Ax = y, the product By gives a solution.

Lemma 7: Existence of Generalized Inverse

For any matrix A, there exists a generalized inverse A^-.

Proof:

Part 1: Symmetric Square Matrices

- Consider the case where A is a *symmetric square matrix*.
- By the *spectral decomposition theorem*, A can be expressed as:

$$A = P\Lambda P^T, \quad \text{where} \quad P^TP = PP^T = I.$$

- Here, Λ is a diagonal matrix containing the eigenvalues of A.
- Without loss of generality, assume Λ has the form:

$$\Lambda = \begin{pmatrix} \Lambda_1 & 0 \\ 0 & 0 \end{pmatrix},$$

where Λ_1 is a diagonal matrix with **nonzero** entries.

- Now, we define a candidate for the generalized inverse:

$$A^- = P\Lambda^- P^T, \quad \text{where} \quad \Lambda^- = \begin{pmatrix} \Lambda_1^{-1} & 0 \\ 0 & 0 \end{pmatrix}.$$

Comments

This slide addresses the existence of a generalized inverse for any matrix. We begin by considering the special case when the matrix A is square and symmetric. According to the spectral decomposition theorem, any symmetric matrix A can be represented as the product of three matrices: $A = P\Lambda P^T$, where P is an orthogonal matrix such that the product of its transpose and itself equals the identity matrix, that is, $P^TP = I$, and Λ is a diagonal matrix containing the eigenvalues of A.

Without loss of generality, we assume that the matrix Λ is partitioned into a block form where the top-left block, denoted by Λ_1, contains all the nonzero eigenvalues on its diagonal, while the bottom-right block consists of zeros.

We then construct a matrix Λ^- by taking the reciprocal of each nonzero diagonal entry of Λ_1 and leaving all other entries as zero. Using this, we define a candidate for the generalized inverse of A as follows: $A^- = P\Lambda^- P^T$.

This construction uses the same eigenbasis as A, and the diagonal inversion is only applied to the nonzero spectrum, which ensures stability in the presence of zero eigenvalues.

Verification for Symmetric Case:

- Let's verify that the defined A^- satisfies the condition $AA^-A = A$:

$$AA^-A = P\Lambda P^T P\Lambda^- P^T P\Lambda P^T = P\Lambda\Lambda^-\Lambda P^T = P\Lambda P^T = A.$$

- Therefore, by Lemma 6, A^- is indeed a generalized inverse.

Part 2: General Case (Arbitrary Matrices)

- For any arbitrary matrix A, there exists a representation (e.g., Singular Value Decomposition) in the form **A=BΛC**.
- Here, **B** and **C** are *non-singular* matrices, and Λ is a diagonal matrix.
- Similar to the symmetric case, we can write Λ and define Λ^- as:

$$\Lambda = \begin{pmatrix} \Lambda_1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \Lambda^- = \begin{pmatrix} \Lambda_1^{-1} & 0 \\ 0 & 0 \end{pmatrix}.$$

- Now, we define the generalized inverse for A as: $A^- = C^{-1}\Lambda^- B^{-1}$.
- Let's verify that for this A^-, the condition $AA^-A = A$ holds:

$$AA^-A = B\Lambda CC^{-1}\Lambda^- B^{-1} B\Lambda C = B\Lambda\Lambda^-\Lambda C = B\Lambda C = A.$$

- Therefore, by Lemma 6, this A^- is also a generalized inverse.

The Lemma is proved. □

Comments

We now verify that the matrix constructed in the symmetric case truly satisfies the condition for being a generalized inverse. We compute the product AA^-A, which becomes $P\Lambda P^T$ multiplied by $P\Lambda^- P^T$ multiplied by $P\Lambda P^T$. Using the fact that the transpose of P times P equals the identity matrix, this expression simplifies to $P\Lambda\Lambda^-\Lambda P^T$. Since $\Lambda\Lambda^-\Lambda = \Lambda$, the final result is $P\Lambda P^T$, which is equal to A. Thus, the required condition is satisfied, and the constructed matrix is indeed a generalized inverse.

To extend this result to arbitrary matrices, we use a general factorization, such as the singular value decomposition. In this approach, any matrix A can be written as the product of three matrices: $A = B\Lambda C$, where B and C are nonsingular matrices and Λ is a diagonal matrix.

Again, we assume that Λ has a block structure where the top-left block Λ_1 contains nonzero diagonal entries, and the rest are zero. Then, we define Λ^- by inverting the nonzero entries of Λ_1 and leaving the rest as zero. Using this, we define the generalized inverse of A as $A^- = C^{-1}\Lambda^- B^{-1}$.

Finally, we verify the condition $AA^-A = A$. Multiplying out, we get $B\Lambda C$ multiplied by $C^{-1}\Lambda^- B^{-1}$ multiplied by $B\Lambda C$. This simplifies to $B\Lambda\Lambda^-\Lambda C$, which again equals $B\Lambda C$, and this is precisely A. Thus, this construction works for any matrix.

Theorem 3

Let A^- be a generalized inverse of a matrix A, and define $H = A^-A$. Then:

(a) $H^2 = H$; that is, H is idempotent.

(b) $AH = A$ and $\text{rank}(A) = \text{rank}(H) = \text{tr}(H)$.

(c) The general solution to $Ax = 0$ is given by $x = (H - I)z$, where z is arbitrary.

(d) The general solution to the consistent system $Ax = y$ is given by $x = A^-y + (H - I)z$, where z is arbitrary.

(e) The product Tx is uniquely defined for all x satisfying $Ax = y$ if and only if $TH = T$.

Proof: Items (a)–(e) follow from Lemma 6. Let us comment on the second part of (b). Assume $\text{rank}(A) = k$. Then:

$$\text{tr}(A^-A) = \text{tr}(C^{-1}\Lambda^-B^{-1}B\Lambda C) = \text{tr}(C^{-1}\Lambda^-\Lambda C)$$
$$= \text{tr}(\Lambda^-\Lambda CC^{-1}) = \text{tr}(\Lambda^-\Lambda) = \text{tr}(I_k) = k = \text{rank}(A).$$

where I_k is diagonal with k ones ($k = \text{rank}(A)$) and zeros elsewhere. □

Comments

This slide presents a fundamental theorem about the structure of generalized inverses. Let A be an arbitrary matrix, and let A^- be any of its generalized inverses. We define the matrix H as A^-A.

The theorem consists of five key statements.

First, item (a): the matrix H is idempotent. That means, if we multiply H by itself, we obtain H again.

Item (b): multiplying A by H on the right gives A. Moreover, the rank of A equals the rank of H, and this also equals the trace of H. The trace, being the sum of the diagonal elements, gives a numerical measure of the rank in this context.

Item (c): the general solution of the homogeneous system $Ax = 0$ is given by the formula $x = (H - I)z$, where z is an arbitrary vector.

Item (d): the general solution to the consistent system $Ax = y$ is given by $A^-y + (H - I)z$, where z is an arbitrary vector.

Finally, item (e): the linear expression Tx takes a unique value for all x satisfying $Ax = y$, if and only if $TH = T$.

The proof follows directly from the characterization of generalized inverses. Let us only comment on the trace identity in item (b). Suppose the rank of A is equal to k. Then, using a canonical factorization of A, we compute the trace of A^-A. This is equal to the trace of a matrix product involving the inverse of C, a diagonal matrix Λ^-, and the matrices B and C. After simplification using cyclicity of trace and orthogonality properties, we obtain the trace of $\Lambda^-\Lambda$, which is simply the trace of the identity matrix of order k. Therefore, the result equals k, which is the rank of A.

Background and Motivation

- In the proof, we used the well-known formula: $\mathrm{tr}(AB) = \mathrm{tr}(BA)$ (provided the matrices are conformable).
- If design matrix X is ***singular*** ($\mathrm{rank}(X) < m$):
  - Unbiased estimates for all parameters θ are ***impossible***.
  - However, ***certain parametric functions*** $\tau = T\theta$ ***can still be estimated***.

Why Generalized Inverses?

- The Moore-Penrose pseudoinverse might not preserve properties like unbiasedness.
- ***Generalized inverses*** allow constructing estimators that meet minimal conditions for ***unbiasedness*** and ***consistency***.
- They provide the necessary flexibility while maintaining $ABA = A$, which is crucial in deriving estimator formulas.

Comments

In the proof, we used the well-known equality: the trace of a matrix product AB equals the trace of BA, provided the matrices are conformable.

In regression settings, when the design matrix X has rank strictly less than the number of columns m, the full parameter vector θ is no longer estimable in an unbiased way. However, some linear combinations of θ, written as $\tau = T\theta$, may still be estimable. Determining which combinations are estimable becomes a central concern when working with singular models.

At this point, let us clarify why generalized inverses, rather than the Moore–Penrose pseudoinverse, are often used in theoretical estimation. The Moore–Penrose pseudoinverse provides the unique solution of minimal Euclidean norm, which is suitable for numerical computations. However, this solution is not always unbiased and may not minimize variance under statistical constraints. Generalized inverses offer the flexibility to construct estimators that meet specific conditions of unbiasedness or optimality. This is why in linear model theory, generalized inverses are preferred over pseudoinverses when analyzing estimability.

Theorem 4 (on Estimability of Linear Functions)

Consider the classical linear regression model (1) with a vector of errors satisfying $\mathrm{E}\epsilon = 0$.

(1) A parametric function $\tau = \mathrm{T}\theta$, where $\mathrm{T} \in \mathbb{R}^{\mathrm{k}\times\mathrm{m}}$ and $\mathrm{k} \in \{1, \ldots, \mathrm{m}\}$, is estimable if and only if

$$\mathrm{T}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-}\mathrm{X}^{\mathrm{T}}\mathrm{X} = \mathrm{T}.$$

(2) If condition (1) is satisfied, then the OLS-estimator for τ is

$$\widehat{\tau} = \mathrm{T}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-}\mathrm{X}^{\mathrm{T}}\mathrm{Y},$$

which is uniquely defined and represents the best linear unbiased estimator. Its covariance matrix is given by

$$\mathrm{D}_{\widehat{\tau}} = \sigma^2\mathrm{D}, \quad \mathrm{D} = \mathrm{T}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-}\mathrm{T}^{\mathrm{T}}.$$

Proof:
This theorem follows directly from Lemma 5, Theorem 2, and Theorem 3. □

Comments

This theorem addresses the problem of estimating a linear function of the parameter vector in the classical regression model. Assume that the vector of errors has expected value equal to zero. We consider a parametric function τ defined as $\mathrm{T}\theta$, where T is a matrix of size $\mathrm{k} \times \mathrm{m}$.

According to the theorem, this function τ is estimable — that is, there exists an unbiased linear estimator — if and only if the matrix identity $\mathrm{T}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-}\mathrm{X}^{\mathrm{T}}\mathrm{X} = \mathrm{T}$ holds. In words: τ is estimable if and only if the product $\mathrm{T}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-}\mathrm{X}^{\mathrm{T}}\mathrm{X}$ equals T.

If this condition is satisfied, then the function τ admits a unique best linear unbiased estimator. It is given by the formula: $\widehat{\tau} = \mathrm{T}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-}\mathrm{X}^{\mathrm{T}}\mathrm{Y}$.

Moreover, the covariance matrix of this estimator is equal to $\sigma^2\mathrm{D}$, where D is defined as $\mathrm{T}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-}\mathrm{T}^{\mathrm{T}}$.

This result generalizes the classical Gauss–Markov theorem to the case where the matrix $\mathrm{X}^{\mathrm{T}}\mathrm{X}$ is singular. In this setting, not all components of θ may be estimable, but linear functions satisfying the stated matrix identity remain estimable with minimal variance.

Example: Model setup

Consider a regression model corresponding to the weighing of three objects using a two-pan balance. The result of each weighing is the difference in weight between the left and right pans plus random noise.

- Let $Y \in \mathbb{R}^N$ be the vector of observed differences.
- Let $\theta = (\theta_1, \theta_2, \theta_3)^T$ be the unknown weights.
- The design matrix X has entries

$$x_{ij} = \begin{cases} 1, & \text{object j on the left pan in weighing i,} \\ -1, & \text{object j on the right pan in weighing i,} \\ 0, & \text{otherwise.} \end{cases}$$

- Regression model: $Y = X\theta + \epsilon$, where $E\epsilon = 0$.

Comments

This example considers a classical weighing problem with three objects and a two-pan scale. The outcome of each weighing is the difference in total mass between the left and the right pan, plus a random error. We assume that the error vector satisfies the standard regression assumptions: first, the expected value of the error is zero; second, the errors have constant variance; and third, they are uncorrelated. The regression equation can be written in matrix form as $Y = X\theta + \epsilon$, where Y is the vector of observed differences, X is the design matrix, θ is the vector of unknown object weights, and ϵ is the vector of random errors.

Each row of the design matrix X represents a weighing configuration. For element x_{ij}, we assign the value one if the j-th object is on the left pan during the i-th weighing, the value minus one if it is on the right pan, and zero if it is excluded from the weighing.

Example: measurement scheme

Let $N = 3$ and the weighings are: $(0, 1, -1)$, $(-1, 1, 0)$, $(1, 0, -1)$. Then

$$X = \begin{pmatrix} 0 & 1 & -1 \\ -1 & 1 & 0 \\ 1 & 0 & -1 \end{pmatrix}, \quad X^T X = \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix}$$

$$X^T Y = \begin{pmatrix} y_3 - y_2 \\ y_1 + y_2 \\ -y_1 - y_3 \end{pmatrix}$$

- The normal equations $X^T X\theta = X^T Y$ become:

$$\begin{aligned} 2\theta_1 - \theta_2 - \theta_3 &= y_3 - y_2 \\ -\theta_1 + 2\theta_2 - \theta_3 &= y_1 + y_2 \\ -\theta_1 - \theta_2 + 2\theta_3 &= -y_1 - y_3 \end{aligned}$$

Comments

In our case, there are three weighings. The first configuration places object two on the left and object three on the right, corresponding to the row $(0, 1, -1)$. The second weighing puts object one on the right and object two on the left, giving $(-1, 1, 0)$. The third weighing puts object one on the right and object three on the left, resulting in the row $(-1, 0, 1)$.

Given the design matrix X from the weighing scheme, we compute the matrix product $X^T X$, which yields a symmetric three-by-three matrix with diagonal elements equal to two and off-diagonal elements equal to minus one. We also compute $X^T Y$, which results in a three-dimensional column vector: the first element is $y_3 - y_2$, the second is $y_1 + y_2$, and the third is $-y_1 - y_3$.

Using these, we write the normal equations of the regression model: $2\theta_1 - \theta_2 - \theta_3 = y_3 - y_2$; $-\theta_1 + 2\theta_2 - \theta_3 = y_1 + y_2$; and $-\theta_1 - \theta_2 + 2\theta_3 = -y_1 - y_3$.

Example (continued): General solution of the normal system

Solving the normal equations yields:

$$\widehat{\theta}_1 = \frac{2y_3 - y_2 + y_1}{3} + \theta_3, \quad \widehat{\theta}_2 = \frac{y_2 + 2y_1 + y_3}{3} + \theta_3$$

- The parameter θ_3 remains arbitrary.
- The system is not of full rank, so X^TX is singular.
- Use of a generalized inverse is necessary for estimation.

Why generalized inverse?

In underdetermined models, the Moore–Penrose pseudoinverse may not yield estimators for identifiable parameter functions. Generalized inverses allow for flexible estimation of identifiable combinations such as $T\theta$.

Comments

Solving the system of normal equations gives the following expressions for the least squares estimators: $\widehat{\theta}_1 = \frac{2y_3 - y_2 + y_1}{3} + \theta_3$; and $\widehat{\theta}_2 = \frac{y_2 + 2y_1 + y_3}{3} + \theta_3$. These formulas show that both estimators depend on θ_3, which remains arbitrary. This arbitrariness reflects the fact that the design matrix X does not have full rank, so the matrix X^TX is singular. Therefore, the system has infinitely many solutions.

In such cases, we estimate only those linear functions of the parameter vector θ that are estimable. This means we focus on combinations like $\theta_1 + \theta_3$, or $\theta_2 + \theta_3$, for which unbiased linear estimators exist. To compute these, we need to choose a generalized inverse of the matrix X^TX.

One might ask why we do not simply use the Moore–Penrose pseudoinverse. The reason is practical: the Moore–Penrose inverse yields one specific solution — the one with minimal Euclidean norm — but it is not required in our setup. Any generalized inverse satisfying the standard conditions can be used to compute the best linear unbiased estimator of an estimable function. In our example, it is convenient to choose a generalized inverse that yields simple expressions for the estimators of the combinations we care about. The Moore–Penrose inverse, though perfectly valid, would result in more complicated formulas without improving the statistical properties of the estimators.

Example (continued): Application of Gauss–Markov-type theorem

To estimate $\tau = \mathrm{T}\theta$ using Theorem 4, choose:

$$(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-} = \begin{pmatrix} \frac{2}{3} & \frac{1}{3} & 0 \\ \frac{1}{3} & \frac{2}{3} & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \mathrm{T} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

- These choices satisfy the conditions of Theorem 4.
- Hence,

$$\widehat{\tau} = \mathrm{T}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-}\mathrm{X}^{\mathrm{T}}\mathrm{Y} = \begin{pmatrix} \frac{2\mathrm{y}_3 - \mathrm{y}_2 + \mathrm{y}_1}{3} \\ \frac{\mathrm{y}_2 + 2\mathrm{y}_1 + \mathrm{y}_3}{3} \\ 0 \end{pmatrix}$$

- Therefore,

$$\widehat{\theta}_1 = \frac{2\mathrm{y}_3 - \mathrm{y}_2 + \mathrm{y}_1}{3}, \quad \widehat{\theta}_2 = \frac{\mathrm{y}_2 + 2\mathrm{y}_1 + \mathrm{y}_3}{3}$$

Comments

To resolve the identifiability issue in our underdetermined regression model, we now apply Theorem 4, which provides the best linear unbiased estimator for a linear transformation of the parameter vector. In our case, we wish to estimate the vector τ, defined as $\mathrm{T}\theta$, where T is a three-by-three matrix designed to extract identifiable combinations. Specifically, the first row of T corresponds to $\theta_1 + \theta_3$, and the second row to $\theta_2 + \theta_3$. The third row is all zeros, since θ_3 is not estimable on its own.

To apply the theorem, we must select a generalized inverse of the matrix $\mathrm{X}^{\mathrm{T}}\mathrm{X}$. We choose a matrix whose first two rows contain two-thirds and one-third in symmetric positions, and whose third row is entirely zero. This choice ensures that θ_3 remains arbitrary, while θ_1 and θ_2 are estimable through appropriate linear combinations.

Substituting these matrices into the formula from Theorem 4, we compute the best linear unbiased estimator for τ as $\mathrm{T}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-}\mathrm{X}^{\mathrm{T}}\mathrm{Y}$. The result is a three-dimensional vector, where the first element is $\frac{2\mathrm{y}_3 - \mathrm{y}_2 + \mathrm{y}_1}{3}$, the second element is $\frac{\mathrm{y}_2 + 2\mathrm{y}_1 + \mathrm{y}_3}{3}$, and the third element is zero.

Hence, the corresponding estimates for θ_1 and θ_2 are uniquely determined, while θ_3 remains unidentifiable. This shows how the use of a generalized inverse, together with the transformation matrix T, allows us to extract the estimable parts of the parameter vector, even when the model is not of full rank.

The Likelihood Function in Linear Regression

- Let's consider the classic linear regression model (1):

$$Y = X\boldsymbol{\theta} + \boldsymbol{\epsilon}$$

- In this model, the observed data vector Y is a sample drawn from a distribution.
- The ***probability density function*** (or probability mass function for discrete data) of Y, parameterized by the unknown parameters $\boldsymbol{\theta}$, is denoted as $L(Y, \boldsymbol{\theta})$.
- This function, $L(Y, \boldsymbol{\theta})$, is known in mathematical statistics as the ***Likelihood Function***.

Definition: Maximum Likelihood Estimator (MLE)

The value $\widehat{\boldsymbol{\theta}}_{\text{MLE}}$ that ***maximizes*** the likelihood function $L(Y, \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ is called the *Maximum Likelihood Estimator*.

Comments

On this slide, we begin our study of the maximum likelihood method – a powerful tool for parameter estimation in statistical models.

We already know that the quality of regression analysis largely depends on the accuracy of parameter estimation. The least squares method (LSM) is a common and reliable approach, but it's not the only one. When we know the error distribution law (for example, normal distribution), we can use this additional information to find the most probable values of unknown parameters (θ) given the observed data.

The function $L(Y, \theta)$, written as $L(Y, \theta)$, denotes the likelihood function. This is the probability density function of the observed data Y, treated as a function of the unknown parameters θ. In essence, the likelihood function measures how "likely" a given value of θ is, in light of the data we observed.

The maximum likelihood estimator, denoted $\widehat{\theta}_{\text{MLE}}$, is defined as the value of θ that maximizes this likelihood function. That is, it is the value of the parameter vector under which the observed data are most probable.

However, the maximum likelihood framework is more general. It allows for flexible modeling assumptions, such as non-normal error distributions, heteroskedasticity, or even discrete outcomes. On the other hand, maximum likelihood estimation requires stronger distributional assumptions and may be computationally more complex than least squares, especially in high-dimensional or non-linear models.

Thus, while both methods often yield similar estimates in linear models, maximum likelihood provides a richer and more principled framework when full probabilistic modeling is desired.

Theorem 5: Properties of Maximum Likelihood Estimators

Let the classical linear regression model $\mathrm{Y} = \mathrm{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \mathrm{I}_\mathrm{N})$ be given. Then the maximum likelihood estimator of $\boldsymbol{\theta}$ coincides with the least squares estimator, and

$$\widehat{\mathrm{s}}^2_{\mathrm{MLE}} = \frac{1}{\mathrm{N}} \left(\mathrm{Y} - \mathrm{X}\widehat{\boldsymbol{\theta}}\right)^{\mathrm{T}} \left(\mathrm{Y} - \mathrm{X}\widehat{\boldsymbol{\theta}}\right)$$

is the maximum likelihood estimator of σ^2.

Proof:

- For a normally distributed observation vector Y (which follows from $\boldsymbol{\epsilon} \sim \mathrm{N}(0, \sigma^2 \mathrm{I}_\mathrm{N})$), the ***likelihood function*** $\mathrm{L}(\mathrm{Y}, \boldsymbol{\theta}, \sigma^2)$ is defined as:

$$\mathrm{L}(\mathrm{Y}, \boldsymbol{\theta}, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{\mathrm{N}/2} \exp\left\{-\frac{1}{2\sigma^2} (\mathrm{Y} - \mathrm{X}\boldsymbol{\theta})^{\mathrm{T}} (\mathrm{Y} - \mathrm{X}\boldsymbol{\theta})\right\}.$$

- To simplify the optimization process, it is significantly easier to work with the ***logarithm of the likelihood function*** (log-likelihood), denoted as $\ln \mathrm{L}$.
- Taking the natural logarithm of L, we obtain the log-likelihood function:

$$\ln \mathrm{L}(\mathrm{Y}, \boldsymbol{\theta}, \sigma^2) = -\frac{\mathrm{N}}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} (\mathrm{Y} - \mathrm{X}\boldsymbol{\theta})^{\mathrm{T}} (\mathrm{Y} - \mathrm{X}\boldsymbol{\theta}).$$

Comments

Let us consider the classical linear regression model under the standard assumptions on the error term: zero mean (unbiasedness), uncorrelated components, and homoskedasticity, meaning that all errors have the same variance. In addition to these standard conditions, we now impose a stronger assumption: we assume that the error vector follows a multivariate normal distribution.

Under these conditions, the observation vector Y has a normal distribution with mean $\mathrm{X}\theta$ and covariance matrix $\sigma^2\mathrm{I}$. In this setting, the likelihood function can be written down explicitly, and we can formulate the following result: the maximum likelihood estimator for the parameter vector θ coincides with the usual least squares estimator. The estimator for the variance parameter σ^2 is given by $\frac{1}{\mathrm{N}}$ times the squared norm of the residual vector $\mathrm{Y} - \mathrm{X}\widehat{\theta}$, evaluated at the estimated value of θ. This quantity is typically denoted by $\widehat{\mathrm{s}}^2_{\mathrm{MLE}}$.

To derive this result, we maximize the likelihood function with respect to the unknown parameters — namely, the vector θ and the scalar σ^2. Since the likelihood depends on $\mathrm{Y} - \mathrm{X}\theta$, the essential part of the optimization reduces to minimizing the corresponding quadratic form.

Rather than maximizing the likelihood directly, we take the natural logarithm. This simplifies the expression while preserving the maximizers. The resulting log-likelihood separates into two terms: the first involves the logarithm of σ^2, and the second involves the squared norm of $\mathrm{Y} - \mathrm{X}\theta$, scaled by σ^2. The structure of the log-likelihood makes clear the equivalence with least squares in θ and yields an explicit formula for the MLE of σ^2, written as $\frac{1}{\mathrm{N}}$ times the squared norm of the residual vector.

MLE in Linear Regression: Completing the Proof

- Differentiating the log-likelihood with respect to the unknown parameters and setting the derivatives to zero, we obtain:

$$\sigma^2 \frac{\partial \ln \mathrm{L}(\mathrm{Y}, \boldsymbol{\theta}, \sigma^2)}{\partial \boldsymbol{\theta}} = \mathrm{X}^{\mathrm{T}}(\mathrm{Y} - \mathrm{X}\boldsymbol{\theta}) = \mathrm{X}^{\mathrm{T}}\mathrm{Y} - \mathrm{X}^{\mathrm{T}}\mathrm{X}\boldsymbol{\theta} = 0,$$

$$2\sigma^4 \frac{\partial \ln \mathrm{L}(\mathrm{Y}, \boldsymbol{\theta}, \sigma^2)}{\partial \sigma^2} = -\mathrm{N}\sigma^2 + (\mathrm{Y} - \mathrm{X}\boldsymbol{\theta})^{\mathrm{T}}(\mathrm{Y} - \mathrm{X}\boldsymbol{\theta}) = 0.$$

- The first equation is the normal equation. By Lemma 1, its solution is the least squares estimator $\widehat{\boldsymbol{\theta}}$.
- Solving the second equation for σ^2, we obtain the MLE:

$$\widehat{\mathrm{s}}^2_{\mathrm{MLE}} = \frac{1}{\mathrm{N}} \|\mathrm{Y} - \mathrm{X}\widehat{\boldsymbol{\theta}}\|^2.$$

□

Remark: Asymptotic Unbiasedness of $\widehat{\mathrm{s}}^2_{\mathrm{MLE}}$

The MLE $\widehat{\mathrm{s}}^2_{\mathrm{MLE}}$ is asymptotically unbiased. That is,

$$\mathbb{E}\widehat{\mathrm{s}}^2_{\mathrm{MLE}} \xrightarrow[\mathrm{N}\to\infty]{} \sigma^2.$$

Comments

To complete the proof, we compute the partial derivatives of the log-likelihood with respect to the unknown parameters — the vector θ and the scalar σ^2 — and set them equal to zero. The derivative with respect to θ yields the matrix equation: σ^2 times the gradient equals the transpose of matrix X multiplied by vector $\mathrm{Y} - \mathrm{X}\theta$. Simplifying, we obtain the normal equations: $\mathrm{X}^{\mathrm{T}}\mathrm{X}\theta = \mathrm{X}^{\mathrm{T}}\mathrm{Y}$. By Lemma 1, the unique solution of this system, provided X has full rank, is the ordinary least squares estimator of θ.

Next, we differentiate with respect to σ^2. After simplification and clearing denominators, we find that $2\sigma^4$ times the derivative equals $-\mathrm{N}\sigma^2$ plus the squared norm of the residual vector. Solving this equation for σ^2 yields the maximum likelihood estimator: $\widehat{\mathrm{s}}^2_{\mathrm{MLE}} = \frac{1}{\mathrm{N}}\|\mathrm{Y} - \mathrm{X}\widehat{\theta}\|^2$.

The theorem is thus fully proven.

Finally, we note that although this estimator for σ^2 is biased in finite samples, it becomes unbiased in the limit. That is, the expectation of $\widehat{\mathrm{s}}^2_{\mathrm{MLE}}$ converges to σ^2 as N tends to infinity. This property, known as asymptotic unbiasedness, justifies using this estimator in large-sample contexts.

MLE in the Normal Case

Let $Y = (y_1, \dots, y_N)^T \sim \mathcal{N}(\theta 1, \sigma^2 I_N)$, with unknown mean θ and variance σ^2. Then the MLEs are:

- Mean: $\widehat{\theta} = \overline{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$;
- Variance: $\widehat{s}^2_{MLE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{y})^2$.

Derivation via linear model:
We write $Y = X\theta + \varepsilon$, with

$$X = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^{N\times 1}, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_N).$$

Then:

$$X^T X = N, \quad X^T Y = \sum y_i \quad \Rightarrow \quad \widehat{\theta} = (X^T X)^{-1} X^T Y = \overline{y},$$

$$\widehat{s}^2_{MLE} = \frac{1}{N}\|Y - X\widehat{\theta}\|^2 = \frac{1}{N}\sum (y_i - \overline{y})^2.$$

Comments

As an immediate corollary of Theorem Five, we obtain a classical result from mathematical statistics. Let us consider a one-dimensional repeated sample from a normal distribution with unknown mean θ and unknown variance σ^2. The maximum likelihood estimator of θ in this case is the sample mean, and the maximum likelihood estimator of σ^2 is the uncorrected sample variance, that is, the sum of squared deviations from the mean divided by N.

To formalize this, we write the sample as a linear regression model. The response vector Y consists of the values y_1 through y_N. The design matrix X consists of a single column of ones. Then the model $Y = X\theta + \epsilon$ corresponds to assuming all observations are identically distributed with common mean θ and independent, homoscedastic normal errors.

In this setting, the product of X^T and X is simply N, and the product of X^T and Y is the sum of all sample values. Therefore, the MLE for θ, which equals the inverse of $X^T X$ times $X^T Y$, reduces to the sample mean.

By applying the general formula for the MLE of σ^2, we substitute the estimated θ and obtain the uncorrected sample variance: $\frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{y})^2$.

This classical example illustrates how the general theory of linear models includes basic parametric estimation problems as special cases.

Definition

Model $(Y,\ X\theta,\ \sigma^2 W)$, where $W \in \mathbb{R}^{N\times N}$ is a known positive definite matrix, and $\sigma^2 > 0$ is an unknown scalar parameter, is called the generalized linear regression model.

Motivation: Repeated Measurements

In practice, generalized regression models often arise in the context of repeated measurements at the same design points. Suppose:

$$y_i = x_{(i)}\theta + \varepsilon_i,$$

where $x_{(i)} = (x_{i1}, \ldots, x_{im}) \in \mathbb{R}^{1\times m}$ are known inputs, and $\varepsilon_i \sim$ i.i.d. $(0, \sigma^2)$ are uncorrelated random errors with equal variance.

Comments

We now extend the classical linear regression model to a more general setting. Consider the model where the response vector Y has expectation equal to $X\theta$ and covariance matrix equal to $\sigma^2 W$, where W is a known symmetric positive definite matrix of size $N \times N$. This model is called the generalized linear regression model.

The key difference from the classical linear model is that the covariance matrix of the errors is no longer proportional to the identity. Instead, it is a general known matrix W, while σ^2 remains an unknown scalar. This structure allows modeling of heteroskedasticity and correlation among observations.

A common practical situation that leads to this model is when repeated measurements are taken at the same set of design points. Suppose that each observation y_i is generated according to the linear relation $y_i = x_i^T\theta + \epsilon_i$, where x_i is a known row vector and ϵ_i is a random error with variance σ^2. If we have multiple measurements at the same point, the errors across observations may be uncorrelated but not identically distributed, leading to a non-scalar covariance structure.

This motivates the use of a general weight matrix W in the model. In such cases, generalized least squares methods are appropriate for estimation.

We have N total observations $x_{(i)}$, but only M are ***distinct*** points $t_{(1)}, \ldots, t_{(M)}$. Let r_k be the ***number of measurements*** at each distinct point $t_{(k)}$ ($\sum_{k=1}^{M} r_k = N$).

- By ***averaging observations*** (y_j) at each $t_{(i)}$, we form a new model for $\widetilde{y}_i$:
$$\widetilde{y}_i = t_{(i)}\boldsymbol{\theta} + \widetilde{\epsilon}_i, \quad i = 1, \ldots, M.$$
- Here, $\widetilde{y}_i$ are mean observations, $\widetilde{\epsilon}_i$ are mean errors.

Averaged Error Properties and Generalized Model Form

- **Error Properties:** $\mathbb{E}[\widetilde{\epsilon}_i] = 0$, $\mathbb{E}[\widetilde{\epsilon}_i \widetilde{\epsilon}_j] = 0$ for $i \neq j$.
- **Scaled Variance:** $\mathbb{E}[\widetilde{\epsilon}_i^2] = \frac{\sigma^2}{r_i}$.
- This leads to a ***Generalized Linear Regression Model***:
$$\left(\widetilde{Y}, \widetilde{X}\boldsymbol{\theta}, \sigma^2 W\right), \quad W = \operatorname{diag}\left(\frac{1}{r_1}, \ldots, \frac{1}{r_M}\right).$$
- ***Key Implication:*** This model structure directly implies the need for ***Weighted Least Squares*** (WLS) for efficient estimation.

Comments

In many experimental settings, repeated measurements are collected at a limited number of design points. Suppose that among the N total measurements, only M design points are distinct. Denote these distinct points by t_1 through t_M. Let r_k denote the number of repeated observations taken at point t_k. Then the total number of measurements equals the sum over k from one to M of r_k, which is equal to N.

By averaging all measurements taken at each design point, we can reduce the original model to a new aggregated model with only M observations. This aggregated model has the form: $\widetilde{y}_i = t_i\theta + \widetilde{\epsilon}_i$, for i from one to M, where $\widetilde{y}_i$ is the average of all measurements at point t_i, and $\widetilde{\epsilon}_i$ is the corresponding average of the error terms.

Because the original errors were uncorrelated and had equal variance σ^2, the aggregated errors also remain uncorrelated, and their variances become σ^2/r_i. Hence, the new model has uncorrelated errors with unequal variances.

The result is a generalized linear regression model with response vector $\widetilde{Y}$, design matrix $\widetilde{X}$, and a diagonal covariance matrix W, where the i-th diagonal entry is $1/r_i$. The covariance structure of the noise becomes $\sigma^2 W$.

Remark: Model Equivalence

A ***generalized linear regression model*** $(\mathrm{Y}, \mathrm{X}\theta, \sigma^2\mathrm{W})$ is equivalent to a ***classical linear regression model*** $(\widetilde{\mathrm{Y}}, \widetilde{\mathrm{X}}\theta, \sigma^2\mathrm{I_N})$.

Justification for the Remark

- Start with generalized model $\left(\mathrm{Y}, \mathrm{X}\boldsymbol{\theta}, \sigma^2\mathrm{W}\right)$.
- Since W is ***positive definite***, $\exists$ non-singular A s.t. $\mathrm{W} = \mathrm{A}\mathrm{A}^{\mathrm{T}}$.
- ***Transform variables*** by A^{-1}: $\widetilde{\mathrm{Y}} = \mathrm{A}^{-1}\mathrm{Y}$, $\widetilde{\mathrm{X}} = \mathrm{A}^{-1}\mathrm{X}$, $\widetilde{\boldsymbol{\epsilon}} = \mathrm{A}^{-1}\boldsymbol{\epsilon}$.
- The ***covariance of transformed errors*** becomes:

$$\begin{aligned}\mathrm{D}(\widetilde{\boldsymbol{\epsilon}}) &= \mathbb{E}[\mathrm{A}^{-1}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathrm{T}}(\mathrm{A}^{-1})^{\mathrm{T}}] \\ &= \mathrm{A}^{-1}(\sigma^2\mathrm{W})(\mathrm{A}^{-1})^{\mathrm{T}} \\ &= \sigma^2\mathrm{A}^{-1}(\mathrm{A}\mathrm{A}^{\mathrm{T}})(\mathrm{A}^{-1})^{\mathrm{T}} = \sigma^2\mathrm{I}.\end{aligned}$$

- This confirms the transformed model $\left(\widetilde{\mathrm{Y}}, \widetilde{\mathrm{X}}\boldsymbol{\theta}, \sigma^2\mathrm{I_N}\right)$ has ***homoscedastic*** and ***uncorrelated errors***, acting as a classical regression model.

Comments

This slide establishes an equivalence between the generalized and classical linear regression models. We consider a generalized model where the error covariance matrix is $\sigma^2\mathrm{W}$, a known positive definite matrix. According to the spectral factorization theorem, such a matrix W can be written as the product $\mathrm{A}\mathrm{A}^{\mathrm{T}}$, where A is an invertible matrix.

We then transform the model by multiplying both sides by the inverse of A. This gives us a new response vector, design matrix, and error vector — denoted by $\widetilde{\mathrm{Y}}$, $\widetilde{\mathrm{X}}$, and $\widetilde{\epsilon}$. The transformed error term becomes $\widetilde{\epsilon} = \mathrm{A}^{-1}\epsilon$. Since ϵ had covariance $\sigma^2\mathrm{W}$, the new error has covariance $\sigma^2\mathrm{I}$. This transforms the model into the classical regression form, where the errors are independent and identically distributed with constant variance.

This result is highly useful in both theory and applications. It shows that every generalized linear model can be reduced to a classical one via an appropriate linear transformation, without changing the underlying structure of the model or the parameter of interest.

If the matrix $\widetilde{\mathrm{X}}^{\mathrm{T}}\widetilde{\mathrm{X}}$ is non-singular, the least squares estimator for the *transformed model* is derived as:

$$\widehat{\boldsymbol{\theta}} = (\widetilde{\mathrm{X}}^{\mathrm{T}}\widetilde{\mathrm{X}})^{-1}\widetilde{\mathrm{X}}^{\mathrm{T}}\mathrm{Y} = (\mathrm{X}^{\mathrm{T}}(\mathrm{A}\mathrm{A}^{\mathrm{T}})^{-1}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}(\mathrm{A}\mathrm{A}^{\mathrm{T}})^{-1}\mathrm{Y} = (\mathrm{X}^{\mathrm{T}}\mathrm{W}^{-1}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{W}^{-1}\mathrm{Y}.$$

Definition: Generalized Least Squares (GLS) Estimator

The estimator defined as:

$$\widehat{\boldsymbol{\theta}} = (\mathrm{X}^{\mathrm{T}}\mathrm{W}^{-1}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{W}^{-1}\mathrm{Y}$$

is called the ***Generalized Least Squares Estimator***.

Properties of the GLS Estimator

- **Covariance Matrix:** The covariance matrix of this estimator is:
$$\mathrm{D}\widehat{\boldsymbol{\theta}} = \sigma^2(\widetilde{\mathrm{X}}^{\mathrm{T}}\widetilde{\mathrm{X}})^{-1} = \sigma^2(\mathrm{X}^{\mathrm{T}}\mathrm{W}^{-1}\mathrm{X})^{-1}.$$
- **Optimality:** According to the ***Gauss–Markov theorem***, this estimator is the ***Best Linear Unbiased Estimator (BLUE)*** under the generalized linear model assumptions.

Comments

In this slide, we formally define the generalized least squares estimator and establish its properties. Recall that after transforming the generalized regression model with covariance matrix $\sigma^2\mathrm{W}$ into the classical form using a spectral factorization, we obtained a model with independent homoscedastic errors. In that classical model, the ordinary least squares estimator is given by the inverse of $\widetilde{\mathrm{X}}^{\mathrm{T}}\widetilde{\mathrm{X}}$, multiplied by $\widetilde{\mathrm{X}}^{\mathrm{T}}\widetilde{\mathrm{Y}}$.

Substituting $\widetilde{\mathrm{X}}$ as $\mathrm{A}^{-1}\mathrm{X}$, and $\widetilde{\mathrm{Y}}$ as $\mathrm{A}^{-1}\mathrm{Y}$, we express the estimator entirely in terms of the original variables X, Y, and the matrix W, which equals $\mathrm{A}\mathrm{A}^{\mathrm{T}}$. After simplification, we arrive at the expression: $\widehat{\theta} = (\mathrm{X}^{\mathrm{T}}\mathrm{W}^{-1}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{W}^{-1}\mathrm{Y}$. This is called the generalized least squares estimator.

It is important to note that this estimator explicitly depends on the covariance structure of the errors via the matrix W. Its variance-covariance matrix, as given on the slide, is $\sigma^2(\mathrm{X}^{\mathrm{T}}\mathrm{W}^{-1}\mathrm{X})^{-1}$.

Finally, the Gauss–Markov theorem ensures that among all linear unbiased estimators, the generalized least squares estimator has the minimal variance. That is, it is the best linear unbiased estimator, or BLUE, for the parameter vector θ.

Motivation

In many practical problems, the researcher has prior knowledge about the model parameters. This can significantly improve estimation accuracy if appropriately incorporated.

Four Typical Types of Prior Information

We will consider four common situations where prior information is available:

1) **Exact Linear Constraints:**

   The parameters satisfy known linear equalities, such as $\mathrm{R}\boldsymbol{\theta} = \mathrm{u}$.

   This reduces to the classical constrained least squares problem.

2) **Noisy Prior Information:**

   The prior knowledge comes from previous measurements or a related model and takes the form:

$$\mathrm{u} = \mathrm{R}\theta + \zeta, \quad \zeta \sim (0, \mathrm{D}), \quad \mathrm{D} > 0, \quad \mathbb{E}[\zeta\epsilon^{\mathrm{T}}] = 0$$

   Here, $\boldsymbol{\zeta}$ is a random error vector independent of the model error $\boldsymbol{\varepsilon}$.

Comments

In many real-world applications, researchers are not working in complete uncertainty — they often possess some prior knowledge about the parameters of the regression model. Incorporating such prior information can improve estimation accuracy and model interpretability.

We will examine four common types of prior information. The first two are introduced here; the remaining two will be considered next.

Case 1 involves exact linear restrictions on the parameter vector — for example, you might know that certain linear combinations of parameters must equal fixed values. In this case, estimation reduces to the familiar problem of least squares with linear equality constraints, which we've already studied.

Case 2 is more nuanced: the researcher has prior data from similar experiments or past studies. This information may not be exact but is still informative. In this setup, the prior information is expressed as a random vector equation: "$\mathrm{u} = \mathrm{R}\theta + \text{noise}$". The noise term, denoted by ζ, is assumed to have some distribution with zero mean and a known positive definite covariance matrix D. It's also uncorrelated with the main model's random error.

This framework allows us to formally include uncertain, yet structured, information in our estimation process.

Four Typical Types of Prior Information (continued)

We now describe the remaining two types of prior information:

3) **Inequality Constraints:**
The parameter vector lies within a known ellipsoid:

$$\Omega = \left\{\boldsymbol{\theta} \in \mathbb{R}^{\mathrm{m}} \;\middle|\; (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^{\mathrm{T}} \mathrm{A} (\boldsymbol{\theta} - \boldsymbol{\theta}_0) \le \mathrm{k}\right\}$$

Here, $\boldsymbol{\theta}_0$ is the center, A is a positive definite matrix, and $\mathrm{k} > 0$ is the size parameter. This leads to a minimax estimation approach.

4) **Probabilistic Prior:**
The parameter vector is assumed to have a known distribution within a region such as Ω. In this case, the problem reduces to classical generalized least squares.

Note. Each type of prior information leads to a different estimation principle: exact constraints yield constrained least squares; linear noisy priors lead to generalized least squares; ellipsoidal constraints require minimax estimators; and probabilistic priors often motivate Bayesian or regularized methods.

Comments

Case 3 involves inequality constraints on the parameter vector. A common example is when it is known that the parameters lie within a certain ellipsoid. This ellipsoid is defined as the set of all parameter vectors such that the quadratic form — $(\theta - \theta_0)^{\mathrm{T}} \mathrm{A} (\theta - \theta_0)$ — is less than or equal to a positive constant k. Here, θ_0 is the center of the ellipsoid, A is a positive definite matrix, and k determines the size. Under such conditions, standard estimation techniques are replaced by minimax procedures, which aim to guard against the worst-case scenario within the given region.

Case 4 reflects a probabilistic form of prior knowledge. In this case, the researcher knows not only that the parameter lies in a certain region, such as the ellipsoid Ω, but also assumes a full probability distribution for the parameter vector within that region. For example, the vector θ may follow a normal distribution centered at θ_0 with covariance matrix proportional to A^{-1}. In such cases, the estimation problem is equivalent to classical procedures such as the generalized least squares or Bayesian estimators, depending on the exact specification.

Let us emphasize: each of the four types of prior information leads to a distinct estimation approach. When constraints are exact equalities, we use constrained least squares. When the prior is linear with noise, generalized least squares applies. Ellipsoidal constraints require minimax estimators, while probabilistic priors motivate Bayesian or regularized methods. These distinctions help select appropriate tools for incorporating prior knowledge in regression analysis.

Let us now consider each of these cases in more detail.

Model with Linear Constraints

Consider a generalized regression model $(\mathrm{Y}, \mathrm{X}\boldsymbol{\theta}, \sigma^2\mathrm{W})$, with additional constraints of the form:

$$\mathrm{u} = \mathrm{R}\boldsymbol{\theta},$$

where R is a known $\mathrm{r} \times \mathrm{m}$ matrix of full row rank $\mathrm{r} < \mathrm{m}$.
By the generalized inverse solution theorem (Theorem 3), the general solution is:

$$\boldsymbol{\theta} = \mathrm{R}^{-}\mathrm{u} + (\mathrm{R}^{-}\mathrm{R} - \mathrm{I})\boldsymbol{\gamma},$$

where $\boldsymbol{\gamma}$ is an arbitrary m-vector.

Transformation and Reduced Model

Define $\mathrm{Z} = \mathrm{Y} - \mathrm{X}\mathrm{R}^{-}\mathrm{u}$. Then $\mathbb{E}[\mathrm{Z}] = \mathrm{X}(\mathrm{R}^{-}\mathrm{R} - \mathrm{I})\boldsymbol{\gamma}$.
This leads to a new regression model: $(\mathrm{Z}, \mathrm{X}(\mathrm{R}^{-}\mathrm{R} - \mathrm{I})\boldsymbol{\gamma}, \sigma^2\mathrm{W})$,
with unknown parameters $\boldsymbol{\gamma}$.

Conclusion. If the matrix $\mathrm{X}(\mathrm{R}^{-}\mathrm{R} - \mathrm{I})$ has full column rank, an unbiased estimate $\widehat{\boldsymbol{\gamma}}$ exists. Then,

$$\widehat{\boldsymbol{\theta}} = \mathrm{R}^{-}\mathrm{u} + (\mathrm{R}^{-}\mathrm{R} - \mathrm{I})\widehat{\boldsymbol{\gamma}}$$

is an unbiased estimate under the constraint $\mathrm{R}\boldsymbol{\theta} = \mathrm{u}$.

Comments

We now consider the case when the parameter vector in a generalized regression model is subject to linear constraints. Specifically, suppose we have a model with observations Y, design matrix X, parameter vector θ, and covariance matrix $\sigma^2\mathrm{W}$. Additionally, suppose that the parameters satisfy a known linear constraint of the form $\mathrm{u} = \mathrm{R}\theta$, where the matrix R has dimensions $\mathrm{r} \times \mathrm{m}$ and full rank r, with r strictly less than m.

According to the theorem on generalized inverse solutions, the general solution of the constraint equation is given by $\theta = \mathrm{R}^{-}\mathrm{u} + (\mathrm{R}^{-}\mathrm{R} - \mathrm{I})\gamma$. Here, γ is an arbitrary m-dimensional vector.

We now substitute this expression into the original model. Let us define a new variable Z, equal to $\mathrm{Y} - \mathrm{X}\mathrm{R}^{-}\mathrm{u}$. Its expectation equals $\mathrm{X}(\mathrm{R}^{-}\mathrm{R} - \mathrm{I})\gamma$. This allows us to consider a new regression model where Z is the new vector of observations, and the parameter to be estimated is γ.

If the new design matrix $\mathrm{X}(\mathrm{R}^{-}\mathrm{R} - \mathrm{I})$ has full column rank, then γ can be estimated without bias. Finally, we reconstruct an estimate for θ using the formula: $\widehat{\theta} = \mathrm{R}^{-}\mathrm{u} + (\mathrm{R}^{-}\mathrm{R} - \mathrm{I})\widehat{\gamma}$. This estimator satisfies the constraint and is unbiased.

Mixed Models: Linear Constraints with Error

Consider a generalized regression model $(\mathrm{Y}, \mathrm{X}\theta, \sigma^2\mathrm{W})$.

- Suppose prior information is given as:

$$\mathrm{u} = \mathrm{R}\theta + \zeta, \quad \zeta \sim (0, \mathrm{D}), \quad \text{with } \mathrm{E}[\zeta\epsilon^{\mathrm{T}}] = 0$$

- This leads to the **mixed model**:

$$(\widetilde{\mathrm{Y}}, \widetilde{\mathrm{X}}\theta, \sigma^2\widetilde{\mathrm{W}}),$$

where:

$$\widetilde{\mathrm{Y}} = \begin{bmatrix} \mathrm{Y} \\ \mathrm{u} \end{bmatrix}, \quad \widetilde{\mathrm{X}} = \begin{bmatrix} \mathrm{X} \\ \mathrm{R} \end{bmatrix}, \quad \widetilde{\mathrm{W}} = \begin{bmatrix} \mathrm{W} & 0 \\ 0 & \frac{1}{\sigma^2}\mathrm{D} \end{bmatrix}$$

- The generalized least squares estimator is:

$$\widehat{\theta} = (\mathrm{X}^{\mathrm{T}}\mathrm{W}^{-1}\mathrm{X} + \mathrm{R}^{\mathrm{T}}\mathrm{D}^{-1}\mathrm{R})^{-1}(\mathrm{X}^{\mathrm{T}}\mathrm{W}^{-1}\mathrm{Y} + \mathrm{R}^{\mathrm{T}}\mathrm{D}^{-1}\mathrm{u})$$

- When σ^2 is unknown, it is replaced by an estimator such as s^2.

Comments

We now consider the case when the prior information about parameters is inexact, expressed with random error. Suppose we have a generalized regression model where the response vector is Y, the design matrix is X, the parameter vector is θ, and the covariance matrix of errors is $\sigma^2\mathrm{W}$. In addition, assume that some auxiliary measurement provides a vector u, related to θ via the equation $\mathrm{u} = \mathrm{R}\theta + \zeta$, where ζ is a random error vector with mean zero and covariance matrix D. This leads to what is called a mixed model, since it combines observation equations with noisy prior constraints.

This mixed model can be rewritten in an extended form with augmented data and design matrices. The extended response vector is the vertical concatenation of Y and u. The extended design matrix stacks X over R. The extended covariance matrix is block-diagonal, with W in the upper-left block and $\frac{1}{\sigma^2}\mathrm{D}$ in the lower-right block.

Using generalized least squares, we obtain the estimator of θ as the inverse of the sum of $\mathrm{X}^{\mathrm{T}}\mathrm{W}^{-1}\mathrm{X}$ and $\mathrm{R}^{\mathrm{T}}\mathrm{D}^{-1}\mathrm{R}$, multiplied by the sum of $\mathrm{X}^{\mathrm{T}}\mathrm{W}^{-1}\mathrm{Y}$ and $\mathrm{R}^{\mathrm{T}}\mathrm{D}^{-1}\mathrm{u}$. This estimator incorporates both the original observations and the prior information with uncertainty. In practice, the value of σ^2 is often unknown and must be replaced by an estimate, such as s^2 from Theorem 5.

Constraints by Region (Set-Based Information)

Let the generalized regression model be given as $(\mathrm{Y}, \mathrm{X}\theta, \sigma^2\mathrm{W})$, and assume that the parameter vector θ is known to lie in a specified region $\Omega \subset \mathbb{R}^{\mathrm{m}}$.

Definition: Minimax Estimator

The **minimax estimator** $\widehat{\theta}$ is defined by

$$\widehat{\theta} = \arg\min_{\widetilde{\theta}\in\Omega} \max_{\theta\in\Omega} \mathrm{g}(\theta, \widetilde{\theta}),$$

where

$$\mathrm{g}(\theta, \widetilde{\theta}) = \mathrm{a}^{\mathrm{T}}\, \mathrm{E}\left[(\widetilde{\theta} - \theta)(\widetilde{\theta} - \theta)^{\mathrm{T}}\right] \mathrm{a},$$

with $\widetilde{\theta}$ ranging over linear estimators of θ, and $\mathrm{a} \in \mathbb{R}^{\mathrm{n}}$ is any fixed nonzero vector.

When $\mathrm{a} = \mathrm{e}_{\mathrm{k}}$, the k-th standard basis vector, the corresponding component $\widehat{\theta}_{\mathrm{k}}$ is the best minimax linear estimator of the coordinate θ_{k}.

Comments

We now consider a generalized regression model where the vector of observations is Y, the design matrix is X, the parameter vector is θ, and the covariance matrix of the errors is σ^2W. Suppose that, in addition to this model, prior information is available in the form of a constraint: the true parameter vector θ is known to belong to some fixed subset Ω of the m-dimensional space.

In such cases, a natural approach is to seek an estimator that is robust under worst-case conditions within Ω. This leads to the minimax principle. A minimax estimator of θ is defined as the linear estimator that minimizes the maximum expected loss over all possible true values of θ in Ω. Specifically, we define the estimator as the argument minimum over all $\widetilde{\theta}$ in Ω of the maximum, over θ in Ω, of the quantity $\mathrm{g}(\theta, \widetilde{\theta})$, where this quantity equals $\mathrm{a}^{\mathrm{T}}\mathrm{E}[(\widetilde{\theta} - \theta)(\widetilde{\theta} - \theta)^{\mathrm{T}}]\mathrm{a}$.

Here, the vector a is any fixed nonzero vector in n-dimensional space. In practice, one is often interested in estimating individual components of θ. If we take a to be the standard basis vector e_{k}, which has one in the k-th coordinate and zeros elsewhere, then the k-th component of the minimax estimator is the best linear estimator for the k-th component of θ in the minimax sense.

Explicit Minimax Estimator under Set Constraints

Let the generalized regression model be given by $(\mathrm{Y}, \mathrm{X}\theta, \sigma^2\mathrm{W})$, and suppose that the parameter vector θ is constrained to a set of the form

$$\Omega = \{\theta \in \mathbb{R}^{\mathrm{m}} : (\theta - \theta_0)^{\mathrm{T}}\mathrm{A}(\theta - \theta_0) \leq \mathrm{k}\},$$

where A is a positive definite matrix and $\mathrm{k} > 0$.

Theorem 6

Under the above model and constraint, the minimax estimator has the explicit form

$$\widehat{\theta}_{\mathrm{mM}} = \left(\frac{\sigma^2}{\mathrm{k}}\mathrm{A} + \mathrm{X}^{\mathrm{T}}\mathrm{W}^{-1}\mathrm{X}\right)^{-1} \mathrm{X}^{\mathrm{T}}\mathrm{W}^{-1}(\mathrm{Y} - \mathrm{X}\theta_0) + \theta_0.$$

A detailed proof of this result can be found in **Ermakov, S. M., Zhiglyavsky, A. A. (1987). Mathematical Theory of Optimal Experiment. Moscow: Nauka** (Theorem 2.1, page 42). The formula also shows that, for sets Ω of the given quadratic form, the minimax estimator $\widehat{\theta}_{\mathrm{mM}}$ does not depend on the choice of vector **a** from the minimax definition.

Comments

We now present an explicit result describing the minimax estimator in the case where the parameter vector θ is known to lie within a closed ellipsoidal region. This region is defined as the set of all θ such that the quadratic form $(\theta - \theta_0)^{\mathrm{T}}\mathrm{A}(\theta - \theta_0)$ is less than or equal to k. The matrix A is symmetric and positive definite, and the constant k is strictly positive.

Under these assumptions, the minimax estimator, which minimizes the worst-case quadratic risk within this region, has a closed-form expression. Specifically, the minimax estimator $\widehat{\theta}_{\mathrm{mM}}$ is equal to the inverse of the matrix $\frac{\sigma^2}{\mathrm{k}}\mathrm{A} + \mathrm{X}^{\mathrm{T}}\mathrm{W}^{-1}\mathrm{X}$, multiplied by the vector $\mathrm{X}^{\mathrm{T}}\mathrm{W}^{-1}(\mathrm{Y} - \mathrm{X}\theta_0)$, and then the result is shifted by θ_0.

This result is formally stated in the theorem and can be found in the book by Ermakov - Zhiglyavsky, page forty-two. Importantly, this formula reveals that, for regions of the given quadratic type, the minimax estimator no longer depends on the choice of the vector a appearing in the general definition of minimaxity. That is, the entire estimator becomes intrinsic to the constraint region, independent of which linear functional of θ we are focusing on.

Special Cases of the Minimax Estimator

Canonical Ellipsoid, Classical Model:
If the ellipsoid is centered at the origin and the model is classical, that is, $\theta_0 = 0$ and $\mathrm{W} = \mathrm{I_N}$, then the minimax estimator takes the form

$$\widehat{\theta}_{\mathrm{mM}} = \left(\frac{\sigma^2}{\mathrm{k}}\mathrm{A} + \mathrm{X^T X}\right)^{-1} \mathrm{X^T Y}.$$

Ridge Estimation Case:
If $\Omega = \{\theta : \theta^{\mathrm{T}}\theta \le \sigma^2/\mathrm{k}\}$, then the minimax estimator becomes

$$\widehat{\theta}_{\mathrm{mM}} = \left(\mathrm{kI_m} + \mathrm{X^T X}\right)^{-1} \mathrm{X^T Y},$$

which is known as the ridge estimator.

Limiting Behavior: As $\mathrm{k} \to \infty$, the minimax estimator converges to the generalized least squares estimator. In other words, the GLS estimator is minimax in the absence of prior information.
Furthermore, for any finite k, the minimax estimator is always uniquely defined, even when the matrix $\mathrm{X^T X}$ is singular. This contrasts with the ordinary least squares estimator, which may not be uniquely defined in degenerate cases.

Comments

Let us consider three notable special cases of the minimax estimator. First, suppose that the center of the ellipsoid is at the origin and the regression model is classical, meaning that $\theta_0 = 0$ and the matrix W is the identity matrix of size N. Then the general minimax formula simplifies to the matrix inverse of $\frac{\sigma^2}{\mathrm{k}}\mathrm{A} + \mathrm{X^T X}$, multiplied by $\mathrm{X^T Y}$. This is a natural shrinkage form centered at zero.

Second, in the specific case where the constraint set Ω is the ball defined by all θ such that the scalar product $\theta^{\mathrm{T}}\theta$ is less than or equal to σ^2/k, the minimax estimator becomes the inverse of $\mathrm{kI_m} + \mathrm{X^T X}$, multiplied by $\mathrm{X^T Y}$. This is precisely the ridge estimator — a well-known regularized version of the least squares estimator.

Third, let us examine the limiting behavior. As the value of k tends to infinity, the minimax estimator tends to the generalized least squares estimator. This reflects the fact that when there is no prior information, the GLS estimator is minimax. Moreover, for any finite k, the minimax estimator is always well-defined, regardless of whether the matrix $\mathrm{X^T X}$ is invertible. This is a significant advantage over the classical least squares estimator, which requires resorting to generalized inverse matrices when the matrix $\mathrm{X^T X}$ is singular.

Minimax Property in the Class of Unbiased Estimators

Remark

The minimax estimator is generally biased. In the class of unbiased estimators, the generalized least squares (GLS) estimator is minimax.

Key idea: If $\widetilde{\theta} = \mathrm{CY}$ is unbiased, then $\mathrm{CX} = \mathrm{I_m}$. Therefore:

$$\widetilde{\theta} - \theta = \mathrm{CY} - \theta = \mathrm{C}(\mathrm{X}\theta + \varepsilon) - \theta = (\mathrm{CX} - \mathrm{I_m})\theta + \mathrm{C}\varepsilon = \mathrm{C}\varepsilon$$

- The estimation error $\mathrm{C}\varepsilon$ does not depend on θ.
- The risk function becomes:

$$\mathrm{g}(\theta, \widetilde{\theta}) = \sigma^2 \mathrm{a}^{\mathrm{T}} \mathrm{CWC}^{\mathrm{T}} \mathrm{a} = \mathrm{a}^{\mathrm{T}} \mathrm{D}_{\widetilde{\theta}} \mathrm{a}$$

- Thus, g does not depend on θ and is minimized over C.

Conclusion

By the Gauss–Markov theorem, the minimum is attained at:

$$\mathrm{C} = (\mathrm{X}^{\mathrm{T}} \mathrm{W}^{-1} \mathrm{X})^{-1} \mathrm{X}^{\mathrm{T}} \mathrm{W}^{-1}$$

Therefore, the GLS estimator is minimax in the class of unbiased estimators.

Comments

Let us emphasize that the general minimax estimator is, in general, a biased estimator. However, if we restrict attention to the class of unbiased estimators, then the generalized least squares estimator is minimax within this class. This conclusion follows directly from the requirement that any unbiased linear estimator of the form CY must satisfy the constraint that matrix C times matrix X equals the identity matrix of size m.

Under this constraint, the estimation error, which is the vector $\mathrm{C}\epsilon$, no longer depends on the true value of θ. Consequently, the mean squared error expression, which equals $\sigma^2 \mathrm{a}^{\mathrm{T}} \mathrm{CWC}^{\mathrm{T}} \mathrm{a}$, is constant with respect to θ. Then, by the Gauss–Markov theorem for the generalized linear model, the minimum of this error expression is achieved when $\mathrm{C} = (\mathrm{X}^{\mathrm{T}} \mathrm{W}^{-1} \mathrm{X})^{-1} \mathrm{X}^{\mathrm{T}} \mathrm{W}^{-1}$.

That is, when the estimator is the generalized least squares estimator. Therefore, the GLS estimator is minimax among all unbiased linear estimators.

Prior Information

Assume a prior distribution $\mathrm{P}(\mathrm{d}\theta)$ is given on a measurable space $(\Omega, \mathcal{F})$, independent of the noise vector ε, with:

- Mean vector: u
- Covariance matrix: D

This is equivalent to the prior constraint

$$\theta = \mathrm{u} + \zeta, \quad \zeta \sim (0, \mathrm{D}), \quad \mathrm{D} > 0, \quad \mathbb{E}[\zeta \epsilon^{\mathrm{T}}] = 0$$

Definition (Bayesian Estimator)

The posterior mean of θ under the prior $\mathrm{P}(\mathrm{d}\theta)$ is given by:

$$\widehat{\theta} = (\mathrm{X}^{\mathrm{T}}\mathrm{W}^{-1}\mathrm{X} + \sigma^2\mathrm{D}^{-1})^{-1}(\mathrm{X}^{\mathrm{T}}\mathrm{W}^{-1}\mathrm{Y} + \sigma^2\mathrm{D}^{-1}\mathrm{u})$$

This is called the Bayesian estimator.

Comments

Let us now consider the case where prior information about the parameter vector θ is specified in the form of a known probability distribution. Concretely, we assume that there exists a prior distribution $\mathrm{P}(\mathrm{d}\theta)$, defined on a sigma-algebra of subsets of a sample space Ω. This prior is independent of the distribution of the error vector ϵ.

The prior distribution is assumed to have mean vector u and covariance matrix D. This setup corresponds to a situation where the parameter vector θ satisfies a stochastic constraint of the form: $\theta = \mathrm{I}\theta + \zeta$, where ζ follows a normal distribution with mean zero and covariance matrix D.

Under these assumptions, we construct an estimator that minimizes the average prediction error, taking into account both the randomness in the data and the uncertainty in the parameter values. This leads us to the so-called Bayes estimator. The criterion being minimized is the expected squared distance between the estimator and the true parameter vector, where the expectation is taken over both the data and the prior distribution of θ.

The resulting estimator is called the posterior mean. It is computed as follows: the inverse of the matrix $\mathrm{X}^{\mathrm{T}}\mathrm{W}^{-1}\mathrm{X} + \sigma^2\mathrm{D}^{-1}$, multiplied by the vector $\mathrm{X}^{\mathrm{T}}\mathrm{W}^{-1}\mathrm{Y} + \sigma^2\mathrm{D}^{-1}\mathrm{u}$.

This estimator is sometimes referred to as a shrinkage estimator, because it balances between the data-driven estimate and the prior mean u.

PART III. Optimal design theory
(LECTURE 3)

Shpilev Petr Valerievich
Faculty of Mathematics and Mechanics, SPbU

September, 2025

Comments

In this lecture, we begin the formal study of the design of experiments within the framework of optimal design theory. After outlining the historical background and motivation, we introduce the regression model, its standard assumptions, and the role of least squares estimation in experimental planning. We define an experimental design as a probability measure and discuss how approximate designs can be discretized. Central to this theory is the information matrix, whose statistical meaning and key properties are carefully examined, supported by rigorous proofs of foundational theorems and lemmas.

We then develop the geometric underpinnings of design theory, making use of convex hulls, Carathéodory's theorem, and integral representations to describe the structure of feasible designs. Building on this foundation, we introduce optimality criteria for experimental design, including D-, L-, E-, G-, and e_k-optimality, and present their general formulation. The lecture concludes with the equivalence theorem for D-optimality, supported by auxiliary results such as determinant inequalities, concavity of the log-determinant function, and its differentiation, laying the groundwork for modern optimization-based approaches to experimental design.

Design of Experiments: Historical Background and Motivation

Motivation

Up to now, we have studied estimation methods for model parameters, most notably the least squares method. According to the Gauss–Markov theorem, the covariance matrix of the parameter estimates depends on the design matrix X. Thus, the precision of estimation can be improved by adjusting this matrix.

- Rows of the matrix X correspond to experimental conditions or measurement points.
- Choosing these points optimally is the main objective of optimal design theory.

Historical Development

- Early experiments lacked formal planning; setups were chosen intuitively.
- **R. Fisher** pioneered formal design theory in the 1930s using combinatorial tools (e.g., Latin squares).
- In the 1950s, **G. Box** and **J. Kiefer** developed the theory of optimal design for regression models.

Comments

Until now, we have discussed parameter estimation methods for linear models, focusing on the least squares approach. According to the Gauss–Markov theorem, the covariance matrix of the least squares estimator depends on the design matrix X. Therefore, the accuracy of our estimates can be improved by modifying this matrix. In practical terms, the rows of matrix X correspond to specific experimental conditions or measurement points. Selecting these points in a mathematically optimal way becomes a central task of optimal experimental design.

Historically, statistical data collection was performed without systematic planning. Experimenters would decide where, when, and how to conduct measurements largely based on intuition. However, as the cost of experiments grew with scientific and technological progress, the need emerged for a mathematical framework that could justify experimental choices.

This framework is known as the theory of experimental design. Its origin is usually attributed to Ronald Fisher, whose foundational book was published in 1935. His approach, largely based on combinatorial tools such as Latin squares, laid the groundwork for modern design methods. In the 1950s, statisticians George Box and Jack Kiefer introduced a new direction — the theory of optimal design for regression experiments. In this setting, the goal is to model the relationship between observed outputs and controllable conditions, and to choose measurement points that yield the most informative data. Kiefer's contributions are considered the foundation of classical optimal regression design theory.

Model Structure

$$y_j = \eta(t_j, \theta) + \epsilon_j, \quad j = 1, \ldots, N, \tag{1o}$$

- y_j: observed responses,
- $t_j \in \chi$: experimental conditions (design points),
- $\theta = (\theta_0, \ldots, \theta_{m-1})^T$: unknown parameters,
- $\eta(t, \theta)$: known function up to θ,
- ϵ_j: observation errors.

Standard Assumptions

(a) Unbiasedness: $E[\epsilon_j] = 0 \Rightarrow E[y_j] = \eta(t_j, \theta)$
(b) Uncorrelated errors: $E[\epsilon_i \epsilon_j] = 0$ for $i \neq j$
(c) Homoscedasticity: $E[\epsilon_j^2] = \sigma^2 > 0$
(d) Linearity in θ: $\eta(t, \theta) = \theta^T f(t)$
(e) Basis: $f_i(t)$, $i = 0, \ldots, m-1$ continuous and linearly independent on χ
(f) Design space: χ is compact and topological

Comments

We now introduce the general structure of a parametric regression model used in experimental design. Suppose we observe a sequence of real-valued responses, denoted by y_1 through y_N. These are modeled as $y_j = \eta(t_j, \theta) + \epsilon_j$, for j from 1 to N. Here, η is a known function up to a parameter vector θ, which consists of m components. The variables t_1 through t_N represent the design points, that is, the conditions under which each measurement is taken. These belong to a design space denoted by the symbol χ. The terms ϵ_j are random errors of observation.

We make several standard assumptions about the model.

First, the errors are unbiased: the expectation of ϵ_j is zero, which implies that the expected value of y_j is equal to $\eta(t_j, \theta)$. Second, we assume the errors are uncorrelated, meaning the expectation of the product $\epsilon_i \epsilon_j$ is zero whenever i is not equal to j. Third, the errors are homoscedastic — they all have the same variance σ^2, which is strictly positive. Fourth, the model is linear in the parameters, meaning that $\eta(t, \theta)$ equals the scalar product of θ with a known vector function $f(t)$. Fifth, the components of $f(t)$ are continuous and linearly independent on the design space χ.

Finally, we assume that χ is a fixed, compact topological space, meaning that its structure is stable and allows for analysis using limits and continuity.

Note: Assumptions (a)–(e) reflect features of real experiments and can be relaxed.

Experimental Objective

- Estimate parameters $\theta_0, \ldots, \theta_{m-1}$,
- Estimate the regression function $\eta(t, \theta)$,
- Test hypotheses on parameter values.

Estimator and Optimization Criterion

- Estimator $\widehat{\theta} = \widehat{\theta}(t_1, \ldots, t_N,\ y_1, \ldots, y_N)$
- Method of estimation depends on both data and design points
- Under assumptions (a)–(e), least squares is a well-established method:

$$\sum_{j=1}^{N} (y_j - \eta(t_j, \theta))^2 \longrightarrow \min_{\theta}$$

Comments

The assumptions labeled (a) through (e) reflect typical features of real-world experimental setups. While these assumptions can be relaxed — for instance, allowing heteroscedastic or correlated errors — we will not consider such generalizations in our course.

In practice, the goals of experimentation usually fall into three related categories. First, we aim to estimate the unknown parameters θ_0 through θ_{m-1}. Second, we may be interested in estimating the regression function $\eta(t, \theta)$, which is determined by these parameters. And third, we often need to test hypotheses about the values of the individual parameters.

The precision of any estimates obtained depends on two factors: the method used to estimate the parameters, and the choice of design points — that is, the values of t_1 through t_N at which the observations y_1 through y_N are made.

The estimator, denoted $\widehat{\theta}$, is a statistic that depends on both the observation values and the design points. Formally, $\widehat{\theta} = \widehat{\theta}(t_1, \ldots, t_N,\ y_1, \ldots, y_N)$. Choosing an appropriate form of this statistic is a separate task studied in estimation theory.

Under assumptions (a) through (e), the least squares method is a well-established and effective estimation procedure. It defines $\widehat{\theta}$ as the solution to the optimization problem of minimizing, over θ, the sum from $j = 1$ to N of the squared differences $y_j - \eta(t_j, \theta)$, quantity squared.

This least squares criterion was discussed earlier in detail.

Experimental Design as a Probability Measure

Discrete Design

$$\xi_N = \begin{pmatrix} t_1 & \dots & t_N \\ 1/N & \dots & 1/N \end{pmatrix}, \quad t_i \in \chi, \quad i = 1, \dots, N$$

- Total number of observations: N
- t_i may repeat for different i
- Equal weights 1/N

Approximate Design

$$\xi = \begin{pmatrix} t_1 & \dots & t_n \\ \omega_1 & \dots & \omega_n \end{pmatrix}, \quad t_i \in \chi, \quad \omega_i \geq 0, \quad \sum_{i=1}^{n} \omega_i = 1$$

- n = number of distinct support points
- $t_i \neq t_j$ for $i \neq j$
- Weights ω_i represent relative frequencies

Comments

As mentioned earlier, the accuracy of the resulting parameter estimates can be improved by choosing appropriate conditions for conducting the experiment. These conditions are encoded in the structure of the experimental design.

In the discrete setting, the design is represented by a two-row table denoted by ξ_N. The first row consists of the design points t_1 through t_N, where each t_i belongs to the set χ. The second row contains weights, all equal to 1/N, where N is the total number of measurements. It is important to note that in this formulation, individual design points may repeat; that is, t_i can be equal to t_j even when i is not equal to j. This representation reflects the classical setup of repeating measurements at some points.

This two-row array is commonly referred to as a discrete or normalized design.

Building on this concept, J. Kiefer introduced a generalization that allowed a broader class of optimization problems to be addressed. According to Kiefer's idea, the design can be viewed as a probability measure over the design space. This leads to the notion of approximate design, denoted simply by ξ.

In this generalized formulation, the design consists of n distinct support points t_1 through t_n, again chosen from the set χ. Each point is assigned a nonnegative weight ω_i, which indicates its relative importance or frequency. The weights must sum to one, forming a proper probability measure. Unlike in the discrete design, the points in an approximate design must be distinct — that is, t_i is not equal to t_j when i is not equal to j.

This formalism lays the foundation for the theory of optimal design, where the aim is to select the measure ξ that yields the most precise estimates according to a given criterion.

From Approximate to Exact Designs

In practice, only discrete designs can typically be implemented.
Given an approximate design

$$\xi = \begin{pmatrix} t_1 & \dots & t_n \\ \omega_1 & \dots & \omega_n \end{pmatrix},$$

one conducts approximately $N\omega_i$ measurements at each point t_i, for $i = 1, \dots, n$.
Rounding rule (example): Use integer allocations $N_i = \lfloor N\omega_i + \delta_i \rfloor$ where $\delta_i \in [0, 1)$ are chosen so that $\sum_{i=1}^{n} N_i = N$.

Convex Structure of Design Space

Let Ξ_n be the set of approximate designs with exactly n support points. Define the full design space:

$$\Xi = \bigcup_{n=1}^{\infty} \Xi_n.$$

Remark

The set Ξ is convex: if $\xi_1, \xi_2 \in \Xi$, then $\xi = \alpha\xi_1 + (1-\alpha)\xi_2 \in \Xi$ for all $\alpha \in [0, 1]$.

Comments

Although approximate experimental designs are the primary objects of theoretical optimization, in practical settings only discrete designs can be implemented.

Suppose we are given an approximate design ξ with n support points t_1 through t_n and corresponding weights ω_1 through ω_n. If we plan to conduct a total of N measurements, the natural approach is to perform approximately $N\omega_i$ measurements at point t_i. This is called discretization or rounding of the approximate design.

A standard method is deterministic rounding. To do this, we define the integer number of measurements at each support point as $N_i = \lfloor N\omega_i + \delta_i \rfloor$ where $\delta_i \in [0, 1)$ are chosen so that $\sum_{i=1}^{n} N_i = N$. This avoids over- or under-shooting the planned sample size. Specific rounding schemes and corrections are well documented in design literature, such as in the books by Fedorov, Pukelsheim, and Atkinson–Donev.

Next, we consider the geometric structure of the space of all approximate designs. Denote by Ξ_n the set of all designs with exactly n support points with nonzero weights. The total space Ξ is the union over all such Ξ_n.

A key property of this space is convexity. That is, for any two designs ξ_1 and ξ_2 in Ξ, any convex combination of them — meaning $\alpha\xi_1 + (1-\alpha)\xi_2$ — also belongs to Ξ for any α between 0 and 1. This makes the use of convex optimization techniques possible when searching for optimal designs.

The convex combination of designs

Let

$$\xi_1 = \begin{pmatrix} t_1 & t_2 & \cdots & t_n \\ \overline{\omega}_1 & \overline{\omega}_2 & \cdots & \overline{\omega}_n \end{pmatrix}, \quad \xi_2 = \begin{pmatrix} t_1 & t_2 & \cdots & t_n \\ \widetilde{\omega}_1 & \widetilde{\omega}_2 & \cdots & \widetilde{\omega}_n \end{pmatrix},$$

where the supports are extended with zero weights if necessary.
Then their convex combination is

$$\xi = \alpha\xi_1 + (1-\alpha)\xi_2 = \begin{pmatrix} t_1 & t_2 & \cdots & t_n \\ \omega_1 & \omega_2 & \cdots & \omega_n \end{pmatrix}, \quad \text{with } \omega_i = \alpha\overline{\omega}_i + (1-\alpha)\widetilde{\omega}_i.$$

Information Matrix

The information matrix of a design ξ is

$$M(\xi) = \int_{\chi} f(t) f^T(t)\, \xi(dt) \in \mathbb{R}^{m\times m},$$

where $f(t)$ is the regressor vector and χ is the design space.
Define the class of all such matrices:

$$\mathcal{M} = \{M : M = M(\xi),\ \xi \in \Xi\}.$$

Comments

We now clarify what we mean by the convex combination of designs. Let us consider two designs ξ_1 and ξ_2, defined on the same set of support points, possibly after extending them by adding zero-weighted points to match their domains. Their convex combination with coefficient α is defined by combining the weights linearly: the new weight ω_i is equal to $\alpha\overline{\omega}_i + (1-\alpha)\widetilde{\omega}_i$.

This operation defines a new design ξ, whose structure is crucial in convex optimization of experiments.

Next, we define the information matrix of a design. Given a regressor vector function $f(t)$, the information matrix of design ξ is defined as the integral over the design space χ of the outer product $f(t)f^T(t)$ with respect to the design measure ξ. In essence, this matrix summarizes the precision of parameter estimation under design ξ.

We denote by $\mathcal{M}$ the set of all information matrices that correspond to all possible approximate designs ξ from the space Ξ.

Discrete Uniform Design

Consider the discrete design $\xi_N = \begin{pmatrix} x_1, \dots, x_N \\ \frac{1}{N}, \dots, \frac{1}{N} \end{pmatrix}$.

Then the information matrix is $M(\xi_N) = \frac{1}{N} F^T F = \frac{1}{N} \sum_{i=1}^{N} f(x_i) f^T(x_i)$,

where each row of F

is the regressor vector $f^T(x_i)$: $F = \begin{pmatrix} f_0(x_1) & f_2(x_1) & \cdots & f_{m-1}(x_1) \\ f_0(x_2) & f_2(x_2) & \cdots & f_{m-1}(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ f_0(x_N) & f_2(x_N) & \cdots & f_{m-1}(x_N) \end{pmatrix}$.

If $F^T F$ is nonsingular, the least squares estimator is $\widehat{\theta} = (F^T F)^{-1} F^T Y$, with dispersion matrix

$$D\widehat{\theta} = \sigma^2 (F^T F)^{-1} = \frac{\sigma^2}{N} (M(\xi_N))^{-1}.$$

Conclusion: the information matrix $M(\xi_N)$ is proportional to the inverse of the covariance matrix of $\widehat{\theta}$.

Comments

Let us now clarify the statistical meaning of the information matrix by focusing on the discrete case. Consider the uniform discrete design ξ_N i.e. the design which consists of N equally weighted points x_1 through x_N.

The information matrix of such a design is given by $1/N$ times $F^T F$.

The matrix F generalizes the classical design matrix X. Its rows are formed from the values of the regressor vector f evaluated at each point: the entry in row i, column j of F is equal to function f_j evaluated at point x_i. Thus, F is an N by m matrix, where m is the number of parameters in the model.

If the matrix $F^T F$ is invertible, then the least squares estimator for the parameter vector θ is given by $(F^T F)^{-1} F^T Y$. The dispersion matrix of this estimator is equal to $\sigma^2 (F^T F)^{-1}$, which is also equal to σ^2/N times the inverse of the information matrix.

Hence, the information matrix is proportional to the inverse of the covariance matrix of the estimator vector $\widehat{\theta}$.

Theorem 7: Properties of Information Matrices

The following statements hold:

(1) Every information matrix is positive semidefinite;
(2) If $n < m$, then for any design $\xi \in \Xi_n$, we have $\det M(\xi) = 0$;
(3) The set $\mathcal{M}$ of all information matrices is convex;
(4) Under regularity conditions (a)–(e), the set $\mathcal{M}$, considered as a set of vectors composed of diagonal and upper triangular elements, is compact in $\mathbb{R}^{(m)(m+1)/2}$;
(5) Under the same conditions, for any desing $\xi \in \Xi$, there exists a plan $\widetilde{\xi} \in \Xi_n$ with $n \le (m)(m+1)/2 + 1$, such that

$$M(\widetilde{\xi}) = M(\xi).$$

Proof:

1. Positive Semidefiniteness. For any vector $l \in \mathbb{R}^m$,

$$l^T M(\xi) l = \int_\chi l^T f(t) f^T(t) l \, \xi(dt) = \int_\chi (l^T f(t))^2 \, \xi(dt) \ge 0.$$

Hence, $M(\xi)$ is positive semidefinite.

Comments

This theorem summarizes five key structural properties of the set of information matrices associated with approximate experimental designs. First, every such matrix is positive semidefinite. This reflects the fact that the expression defining the matrix — namely, the integral of the outer product of the regressor vector with itself — yields a nonnegative definite result.

Second, if a design is concentrated in fewer than m points, where m is the dimension of the parameter vector θ, then the determinant of the information matrix equals zero. In such cases, the design does not provide enough independent information to estimate all parameters.

Third, the set of all information matrices, denoted by $\mathcal{M}$, is convex. That is, if we take any two matrices in this set and form their weighted average, the result is again an information matrix.

The fourth point states that under regularity conditions, this set is not only convex but also compact when viewed as a subset of Euclidean space with dimension equal to $m(m+1)/2$. This dimension corresponds to the number of distinct entries in a symmetric matrix of size m.

Finally, any information matrix can be represented by a design supported at a finite number of points — specifically, at most $m(m+1)/2+1$ points. This result is crucial in practice: it justifies searching for optimal designs within a finite-dimensional class.

Let us prove the statements of the theorem.

First, we show that the information matrix is always positive semidefinite. By definition, for any vector l from the m-dimensional real space, the quadratic form $l^T M(\xi) l$ equals the integral over the design space of the square of $l^T f(t)$, all multiplied by the measure $\xi(dt)$. Since squares of real numbers are always nonnegative, the result is greater than or equal to zero. Hence, the matrix $M(\xi)$ is positive semidefinite.

2. Singularity for $n < m + 1$. Using the subadditivity property of matrix rank:

$$\operatorname{rank}(M(\xi)) \le \sum_{i=1}^{n} \operatorname{rank}\bigl(f(x_i)f^{T}(x_i)\bigr).$$

Each term has rank 1, since

$$f(t)f^{T}(t) = \begin{pmatrix} f_0(t)f^{T}(t) \\ f_1(t)f^{T}(t) \\ \vdots \\ f_{m-1}(t)f^{T}(t) \end{pmatrix},$$

i.e., all rows are scalar multiples of $f^{T}(t)$. Thus,

$$\operatorname{rank}(M(\xi)) \le n.$$

If $n < m$, then $\operatorname{rank}(M(\xi)) < m$, so $\det M(\xi) = 0$.

Comments

Here we proceed to the proof of the second statement. Suppose the design ξ is supported on n points. Then, the information matrix is a sum of n matrices of the form $f(x_i)f^{T}(x_i)$. Each such matrix has rank one, because all its rows are proportional to the vector $f^{T}(x_i)$. Therefore, the rank of the total matrix $M(\xi)$ is at most n.

Now, if n is strictly less than m, then the rank of $M(\xi)$ is strictly less than m, meaning that the matrix is singular and its determinant is zero.

3. Convexity of the Set $\mathcal{M}$. Let $\mathrm{M}_1 = \mathrm{M}(\xi_1), \mathrm{M}_2 = \mathrm{M}(\xi_2) \in \mathcal{M}$, and $\alpha \in [0,1]$. Then

$$\begin{aligned}\mathrm{M} &= \alpha \mathrm{M}_1 + (1-\alpha)\mathrm{M}_2 \\ &= \alpha \int \mathrm{f}(\mathrm{t})\mathrm{f}^{\mathrm{T}}(\mathrm{t})\, \xi_1(\mathrm{dt}) + (1-\alpha) \int \mathrm{f}(\mathrm{t})\mathrm{f}^{\mathrm{T}}(\mathrm{t})\, \xi_2(\mathrm{dt}) \\ &= \int \mathrm{f}(\mathrm{t})\mathrm{f}^{\mathrm{T}}(\mathrm{t})\, [\alpha\xi_1(\mathrm{dt}) + (1-\alpha)\xi_2(\mathrm{dt})] \\ &= \int \mathrm{f}(\mathrm{t})\mathrm{f}^{\mathrm{T}}(\mathrm{t})\, \xi(\mathrm{dt}),\end{aligned}$$

where $\xi = \alpha\xi_1 + (1-\alpha)\xi_2 \in \Xi$, since this set is convex. Hence, $\mathrm{M} \in \mathcal{M}$. Thus the set $\mathcal{M}$ of all information matrices generated by designs $\xi \in \Xi$ is convex.

Note: This result is foundational for optimal design theory, enabling use of convex optimization techniques.

Comments

Let us now verify part three of the theorem, which asserts that the set of all information matrices is convex.

To do this, we consider two information matrices, denoted by M_1 and M_2, each corresponding to some designs ξ_1 and ξ_2. We take a convex combination of these matrices, with weights α and $1-\alpha$, where α lies between 0 and 1 inclusive.

By linearity of integration, we can combine the two integrals defining M_1 and M_2 into a single integral involving the convex combination of the design measures. That is, $\alpha\mathrm{M}_1 + (1-\alpha)\mathrm{M}_2$ equals the integral of the outer product $\mathrm{f}(\mathrm{t})\mathrm{f}^{\mathrm{T}}(\mathrm{t})$ with respect to the new measure, which is $\alpha\xi_1 + (1-\alpha)\xi_2$.

According to a previous remark, this new measure is again a valid design measure, and therefore the resulting matrix is also an element of the set of information matrices.

Thus, we have shown that the set of all such matrices is closed under convex combinations, meaning it is convex.

This property is fundamental in optimal design theory. It ensures that optimization problems over information matrices can be tackled using powerful tools from convex analysis.

Let $V \subset \mathbb{R}^k$. We introduce several basic notions:

Convex Combination

A vector $\alpha v_1 + (1-\alpha)v_2$ is called a convex combination of v_1 and v_2 if $0 < \alpha < 1$.

Convex Set

A set $V \subset \mathbb{R}^k$ is called convex if it contains all convex combinations of any two of its vectors:

$$v_1, v_2 \in V \;\Rightarrow\; \alpha v_1 + (1-\alpha)v_2 \in V.$$

Convex Hull

The convex hull $\operatorname{conv} V$ of a set V is the intersection of all convex sets containing V.

We define:

$$\widehat{V}_n = \left\{ v \in \mathbb{R}^k \,\middle|\, v = \sum_{i=1}^{n} \alpha_i v_i,\ \alpha_i > 0,\ \sum \alpha_i = 1,\ v_i \in V \right\}, \quad \widehat{V} = \bigcup_{n=1}^{\infty} \widehat{V}_n.$$

Comments

To proceed with the proof of parts four and five of the theorem, we introduce several auxiliary definitions concerning convexity in Euclidean space.

First, we define a convex combination of two vectors as any linear combination where the coefficients are strictly between zero and one, and add up to one. Specifically, $\alpha v_1 + (1-\alpha)v_2$, with α strictly between 0 and 1.

Second, we define a convex set as a set that contains all convex combinations of any two of its elements. That is, if v_1 and v_2 belong to V, then every convex combination of them also lies in V.

Third, the convex hull of a set V is defined as the smallest convex set that contains V, or equivalently, the intersection of all convex sets containing V.

We now introduce notations that will be used shortly. The set $\widehat{V}_n$ consists of all convex combinations of n elements from V with positive coefficients summing to one. The union over all such n defines the set $\widehat{V}$. These sets are useful for characterizing the convex hull and expressing matrix combinations in vector form.

Lemma 8

The set $\widehat{\mathrm{V}}$ is the convex hull of V:

$$\widehat{\mathrm{V}} = \operatorname{conv} \mathrm{V}.$$

Proof: We split the proof into two parts. First, we show that $\widehat{\mathrm{V}} \subset \operatorname{conv} \mathrm{V}$; then we show the converse inclusion $\operatorname{conv} \mathrm{V} \subset \widehat{\mathrm{V}}$.

1. $\widehat{\mathrm{V}} \subset \operatorname{conv} \mathrm{V}$. We prove this by induction.
Clearly, $\widehat{\mathrm{V}}_1 \subset \operatorname{conv} \mathrm{V}$.
Assume that $\widehat{\mathrm{V}}_n \subset \operatorname{conv} \mathrm{V}$. We show that $\widehat{\mathrm{V}}_{n+1} \subset \operatorname{conv} \mathrm{V}$.
Let $\mathrm{v} \in \widehat{\mathrm{V}}_{n+1}$, then

$$\mathrm{v} = \sum_{i=1}^{n+1} \alpha_i \mathrm{v}_i = \sum_{i=1}^{n} \alpha_i \mathrm{v}_i + \alpha_{n+1} \mathrm{v}_{n+1}.$$

Without loss of generality, we may assume $\alpha_{n+1} > 0$.

Comments

We now state and begin the proof of an important lemma: the set denoted by $\widehat{\mathrm{V}}$ is equal to the convex hull of the set V. That is, every element in $\widehat{\mathrm{V}}$ can be written as a convex combination of vectors from the original set V, and vice versa.

To prove this, we divide the argument into two parts. First, we prove that the set $\widehat{\mathrm{V}}$ is a subset of the convex hull of V. Then we will prove the opposite inclusion.

To establish the first part, we use mathematical induction. In the base case, $\widehat{\mathrm{V}}_1$ clearly lies within the convex hull of V, since it consists of a single point from V.

For the induction step, suppose that $\widehat{\mathrm{V}}_n$ is contained in the convex hull of V. Consider an arbitrary element v from the set $\widehat{\mathrm{V}}_{n+1}$. This vector v can be written as the sum over i from 1 to $n+1$ of $\alpha_i \mathrm{v}_i$. We rearrange this sum by grouping the first n terms together, and separating the term with index $n+1$. That is, we write v as the sum over i from 1 to n of $\alpha_i \mathrm{v}_i$, plus $\alpha_{n+1} \mathrm{v}_{n+1}$.

Without loss of generality, we may assume that the last coefficient α_{n+1} is strictly greater than zero.

Proof of Lemma 8: Inclusion

Let us denote $\alpha = 1 - \alpha_{n+1} = \sum_{i=1}^{n} \alpha_i$.
Define the vectors:

$$\overline{v}_1 = v_{n+1}, \quad \overline{v}_2 = \sum_{i=1}^{n} \frac{\alpha_i}{\sum_{i=1}^{n} \alpha_i} v_i = \sum_{i=1}^{n} \alpha_i' v_i, \quad \text{where } \sum_{i=1}^{n} \alpha_i' = 1.$$

Then:

- $\overline{v}_1 \in V \subset \text{conv}V$,
- $\overline{v}_2 \in \widehat{V}_n \subset \text{conv}V$ (by the induction hypothesis),
- Since convV is convex, we conclude:

$$(1 - \alpha)\,\overline{v}_1 + \alpha\,\overline{v}_2 \in \text{conv}V.$$

Hence, by mathematical induction, we have proved the inclusion $\widehat{V} \subset \text{conv}\, V$.

Comments

To complete the inductive step, we now express the vector v in terms of two auxiliary vectors. Let α be defined as $1 - \alpha_{n+1}$, which is equal to the sum of α_i from i = 1 to n.

We then define two vectors. First, we let $\overline{v}_1$ be equal to v_{n+1}. Second, we define $\overline{v}_2$ as the weighted sum over i from 1 to n of α_i divided by the sum of α_i, times v_i. In other words, $\overline{v}_2$ equals the sum of $\alpha_i' v_i$, where the sum of all α_i' equals one.

By construction, the first vector $\overline{v}_1$ belongs to the original set V, and therefore lies in the convex hull of V. The second vector $\overline{v}_2$ lies in the set $\widehat{V}_n$. By the induction hypothesis, this also lies in the convex hull of V.

Since the convex hull is closed under convex combinations, the weighted sum of these two vectors — namely, $(1 - \alpha)\overline{v}_1 + \alpha\overline{v}_2$ — also lies in the convex hull of V.

This completes the inductive step. We have shown that any element of $\widehat{V}_{n+1}$ also belongs to the convex hull of V. Therefore, by mathematical induction, the entire set $\widehat{V}$ is contained in the convex hull of V.

Proof of Lemma 8: Converse inclusion

Clearly, $\mathrm{V} \subset \widehat{\mathrm{V}}$ since $\widehat{\mathrm{V}}_1 = \mathrm{V}$.
We now verify that $\widehat{\mathrm{V}}$ is convex:

- Let $\widehat{\mathrm{v}}_1, \widehat{\mathrm{v}}_2 \in \widehat{\mathrm{V}}$.
- Suppose $\widehat{\mathrm{v}}_1 = \sum_{\mathrm{i}=1}^{\mathrm{n}_1} \alpha_\mathrm{i} \mathrm{v}_\mathrm{i}$ and $\widehat{\mathrm{v}}_2 = \sum_{\mathrm{i}=1}^{\mathrm{n}_2} \alpha'_\mathrm{i} \mathrm{v}'_\mathrm{i}$.
- Then for any $\alpha \in [0,1]$ we have:

$$\mathrm{v} = \alpha\widehat{\mathrm{v}}_1 + (1-\alpha)\widehat{\mathrm{v}}_2 = \sum_{\mathrm{i}=1}^{\mathrm{n}_1} \alpha\alpha_\mathrm{i}\mathrm{v}_\mathrm{i} + \sum_{\mathrm{i}=1}^{\mathrm{n}_2}(1-\alpha)\alpha'_\mathrm{i}\mathrm{v}'_\mathrm{i} = \sum_{\mathrm{i}=1}^{\mathrm{n}_1+\mathrm{n}_2} \widetilde{\alpha}_\mathrm{i}\widetilde{\mathrm{v}}_\mathrm{i},$$

where:

$$\widetilde{\mathrm{v}}_\mathrm{i} = \mathrm{v}_\mathrm{i}, \quad \widetilde{\alpha}_\mathrm{i} = \alpha\alpha_\mathrm{i}, \quad \mathrm{i} \le \mathrm{n}_1;$$
$$\widetilde{\mathrm{v}}_\mathrm{i} = \mathrm{v}'_{\mathrm{i}-\mathrm{n}_1}, \quad \widetilde{\alpha}_\mathrm{i} = (1-\alpha)\alpha'_{\mathrm{i}-\mathrm{n}_1}, \quad \mathrm{i} > \mathrm{n}_1.$$

- Then $\sum_{\mathrm{i}=1}^{\mathrm{n}_1+\mathrm{n}_2} \widetilde{\alpha}_\mathrm{i} = 1$, and all $\widetilde{\alpha}_\mathrm{i} \ge 0$, hence $\mathrm{v} \in \widehat{\mathrm{V}}$.

Since $\widehat{\mathrm{V}}$ is convex and contains V, we conclude:

$$\operatorname{conv}\mathrm{V} \subset \widehat{\mathrm{V}}.$$

The Lemma is proved. □

Comments

We now complete the proof of the lemma by establishing the reverse inclusion, namely that the convex hull of V is contained in the set $\widehat{\mathrm{V}}$.

First, we observe that the set V is a subset of $\widehat{\mathrm{V}}$. This follows directly from the definition, since $\widehat{\mathrm{V}}_1$ is equal to V.

Next, we show that the set $\widehat{\mathrm{V}}$ is convex. To do this, consider two arbitrary elements of $\widehat{\mathrm{V}}$, denoted $\widehat{\mathrm{v}}_1$ and $\widehat{\mathrm{v}}_2$. Suppose that $\widehat{\mathrm{v}}_1$ is a convex combination of vectors v_i, for i from 1 to n_1, and that $\widehat{\mathrm{v}}_2$ is a convex combination of other vectors v'_i, for i from 1 to n_2.

We now consider a convex combination of $\widehat{\mathrm{v}}_1$ and $\widehat{\mathrm{v}}_2$ with weights α and $1-\alpha$. Expanding this expression, we obtain a single convex combination of the vectors v_i and v'_i, with new weights defined as $\alpha\alpha_\mathrm{i}$ and $(1-\alpha)\alpha'_\mathrm{i}$ respectively.

All these new weights are non-negative, and their sum equals one. Therefore, the resulting vector again belongs to $\widehat{\mathrm{V}}$. This confirms that $\widehat{\mathrm{V}}$ is convex.

Since the convex hull of V is, by definition, the smallest convex set containing V, and since $\widehat{\mathrm{V}}$ is convex and contains V, it must contain the entire convex hull of V.

We thus conclude that the convex hull of V is a subset of $\widehat{\mathrm{V}}$. This completes the proof of the lemma.

Lemma 9

Let $v_1, \ldots, v_{k+2} \in \mathbb{R}^k$. Then there exist scalars $\beta_1, \ldots, \beta_{k+2}$, not all zero, such that

$$\sum_{i=1}^{k+2} \beta_i v_i = 0, \quad \sum_{i=1}^{k+2} \beta_i = 0.$$

Proof: Consider the lifted vectors $\overline{v}_i = (v_i^T,\ 1)^T \in \mathbb{R}^{k+1}$. Since we have $k+2$ vectors in a $(k+1)$-dimensional space, they are linearly dependent. Therefore, there exist scalars $\beta_1, \ldots, \beta_{k+2}$, not all zero, such that

$$\sum_{i=1}^{k+2} \beta_i \overline{v}_i = 0.$$

Writing this out, we have:

$$\sum_{i=1}^{k+2} \beta_i v_i = 0, \quad \sum_{i=1}^{k+2} \beta_i = 0.$$

This proves the lemma. □

Comments

We now prove a basic linear algebra lemma. Suppose we are given $k+2$ vectors in k-dimensional space. Then there exists a nontrivial linear combination of these vectors that sums to the zero vector, and at the same time, the sum of the coefficients in this combination is zero.

To show this, we use the standard lifting technique. For each vector v_i in k-dimensional space, construct a new vector by appending a one at the end. The resulting vector is denoted $\overline{v}_i$ and lives in $k+1$ dimensional space.

Now we have $k+2$ vectors in a space of dimension $k+1$. Therefore, these vectors must be linearly dependent. That means there exist scalars β_1 through β_{k+2}, not all equal to zero, such that the sum over i of $\beta_i \overline{v}_i$ equals zero.

Writing this sum explicitly, we separate it into two parts: the sum over $\beta_i v_i$, and the sum over β_i times the constant one. These two parts must both vanish, which yields the desired result: the sum over i of $\beta_i v_i$ equals zero, and the sum over i of β_i equals zero.

This completes the proof of the lemma.

Carathéodory's Theorem

Let V be a compact subset of $\mathbb{R}^k$. Then every point $v \in \text{conv}(V)$ can be written as a convex combination of at most k+1 points from V:

$$v = \sum_{i=1}^{m} \alpha_i v_i, \quad v_i \in V,\ \alpha_i \geq 0,\ \sum_{i=1}^{m} \alpha_i = 1,\ m \leq k+1.$$

Proof: Let $v \in \text{conv}(V)$ be arbitrary. By Lemma 9, it can be written as a convex combination:

$$v = \sum_{i=1}^{n} \alpha_i v_i, \quad v_i \in V,\ \alpha_i > 0,\ \sum_{i=1}^{n} \alpha_i = 1.$$

If $n \leq k+1$, the statement holds. Suppose instead that $n \geq k+2$.
By Lemma 9, the vectors $v_1, \ldots, v_n$ satisfy a nontrivial relation of the form:

$$\sum_{i=1}^{n} \beta_i v_i = 0, \quad \sum_{i=1}^{n} \beta_i = 0, \quad \text{not all } \beta_i = 0.$$

We will use this relation to eliminate one point from the combination while keeping the result unchanged.

Comments

We now state and begin proving Carathéodory's theorem, which is fundamental in convex analysis. It asserts that any point in the convex hull of a compact set in k-dimensional space can be expressed as a convex combination of at most $k+1$ elements from that set.

Let V be a compact subset of k-dimensional Euclidean space. Let v be any point in the convex hull of V. According to the previous lemma, any such point can be written as a convex combination of a finite number of points from V, say v_i with weights α_i, which are strictly positive and sum to one.

If the number of such points is less than or equal to $k+1$, then the claim holds directly. But suppose the number is greater than or equal to $k+2$. Then we can apply the second lemma we proved earlier.

This lemma guarantees that the set of vectors v_1 through v_n satisfies a nontrivial linear dependence, meaning that there exists a set of coefficients β_i, not all zero, such that the sum over i of $\beta_i v_i$ equals zero, and the sum of the β_i equals zero as well.

We will use this relation to eliminate one of the points from the convex combination without changing the result, and we will continue this process iteratively until no more than $k+1$ points remain.

To reduce the number of terms, define the index

$$i_0 = \arg\min_i \left\{ \frac{\alpha_i}{\beta_i} \ : \ \beta_i > 0 \right\}.$$

Let the updated weights be

$$\overline{\alpha}_i = \alpha_i - \frac{\alpha_{i_0}}{\beta_{i_0}} \beta_i, \quad \text{for all } i = 1, \ldots, n.$$

Then clearly:

$$v = \sum_{i=1}^{n} \overline{\alpha}_i v_i, \quad \sum \overline{\alpha}_i = 1, \quad \overline{\alpha}_i \geq 0,$$

and at least one of the new coefficients is zero, namely $\overline{\alpha}_{i_0} = 0$.
Repeating this reduction process, we ultimately obtain a representation

$$v = \sum_{i=1}^{k+1} \widetilde{\alpha}_i v_i, \quad \widetilde{\alpha}_i \geq 0, \quad \sum \widetilde{\alpha}_i = 1.$$

The Theorem is proved. □

Comments

To complete the proof of Carathéodory's theorem, we reduce the number of vectors in the convex combination. Suppose that the point v is represented as a convex combination of n points from the set V, with n greater than or equal to $k + 2$. According to the lemma, there exists a nontrivial linear dependence among any $k + 2$ vectors in Euclidean space of dimension k. That means there exist real numbers β_1 through β_n, not all zero, such that the sum of $\beta_i v_i$ is equal to zero, and the sum of β_i is zero.

We then define the index i_0 as the value of i which minimizes the ratio α_i/β_i among those i for which β_i is strictly positive. Using this, we define new coefficients $\overline{\alpha}_i$ as $\alpha_i - (\alpha_{i_0}/\beta_{i_0})\beta_i$. These modified coefficients remain nonnegative and sum to one.

Importantly, one of the coefficients, namely the one corresponding to i_0, becomes zero. Thus, the number of terms with nonzero weights is reduced by at least one. Repeating this process iteratively, we eventually reduce the representation to a convex combination involving at most $k + 1$ points. This proves the theorem.

To apply Carathéodory's theorem to sets of information matrices, we need one more technical result.

Let

$$\widetilde{\mathrm{V}} = \left\{ \int \mathrm{v}(\mathrm{x})\, \xi(\mathrm{dx}) \,:\, \mathrm{v} \in \mathrm{V} \right\},$$

where ξ is a probability measure on the measurable space $(\mathrm{X}, \mathcal{B})$, and V is a set of continuous vector-valued functions on X.

Lemma 10

If the set X is compact, then $\widetilde{\mathrm{V}} = \operatorname{conv} \mathrm{V}$.

Proof:

- Since the functions $\mathrm{v}(\mathrm{x}) \in \mathrm{V}$ are continuous and X is compact, the convex hull convV is also compact.
- It is easy to see that $\operatorname{conv}\mathrm{V} \subset \widetilde{\mathrm{V}}$, because $\widetilde{\mathrm{V}}$ is convex and contains V.
- Assume, for contradiction, that $\widetilde{\mathrm{V}} \neq \operatorname{conv}\mathrm{V}$. Then there exists a probability measure ξ^* such that

$$\mathrm{v}^* = \int \mathrm{v}\, \xi^*(\mathrm{dv}), \quad \mathrm{v}^* \notin \operatorname{conv}\mathrm{V}.$$

Comments

This part of the lecture prepares a technical result that is essential for applying Carathéodory's theorem to sets of information matrices. The main object introduced here is the set $\widetilde{\mathrm{V}}$, which consists of all integrals of the form: the integral of v(x) with respect to $\xi(\mathrm{dx})$, where v belongs to the set V, and ξ is a probability measure defined on the measurable space X with sigma-algebra $\mathcal{B}$.

The lemma claims that if the set X is compact, then the set $\widetilde{\mathrm{V}}$ coincides exactly with the convex hull of V. In other words, all such integrals can be represented as convex combinations of elements in V.

To prove this, we proceed step by step. First, since the functions v(x) are continuous and the set X is compact, the convex hull of V is also compact. Next, we observe that the convex hull is contained in $\widetilde{\mathrm{V}}$ because $\widetilde{\mathrm{V}}$ is convex and includes the original set V.

Now we suppose, for contradiction, that $\widetilde{\mathrm{V}}$ is strictly larger than the convex hull. Then there must exist a probability measure ξ^* such that the integral of v with respect to ξ^* equals some vector v^* that lies outside the convex hull of V.

End of proofs of Lemma 10 and Theorem 7

To reach a contradiction, we apply the separation theorem:

- Since $v^* \notin \text{conv}V$ and convV is compact, there exists a hyperplane separating v^* from convV.
- That is, there exists a vector α such that

$$\alpha^{\mathrm{T}} v \le C < \alpha^{\mathrm{T}} v^* \quad \text{for all } v \in \text{conv}V.$$

- On the other hand, since $v^* = \int v\, \xi^*(dv)$, we have:

$$\alpha^{\mathrm{T}} v^* = \int \alpha^{\mathrm{T}} v\, \xi^*(dv) \le \int C\, \xi^*(dv) = C.$$

- Contradiction. Thus, $\widetilde{V} = \text{conv}V$.

□

Consequence

Item (4) and (5) of Theorem 7 follows directly from this lemma and Carathéodory's Theorem.

Theorem 7 is proved. □

Comments

To complete the proof, we invoke the separation theorem from convex analysis. This theorem states that if a point lies outside a compact convex set, then there exists a hyperplane that separates the point from the set.

In our case, the vector v^* does not belong to the convex hull of V. Since the convex hull is compact, the separation theorem ensures that there exists a vector α such that, for all vectors v in the convex hull of V, the scalar product $\alpha^{\mathrm{T}} v$ is less than or equal to a constant C, which is strictly less than $\alpha^{\mathrm{T}} v^*$.

This inequality provides the key contradiction. Recall that v^* is defined as the integral of v with respect to the measure ξ^*. Therefore, $\alpha^{\mathrm{T}} v^*$ equals the integral over $\alpha^{\mathrm{T}} v$ with respect to ξ^*.

Because $\alpha^{\mathrm{T}} v$ is at most C for all v in the convex hull, the integral is also bounded above by C. Hence, $\alpha^{\mathrm{T}} v^*$ is less than or equal to C.

But earlier we concluded that $\alpha^{\mathrm{T}} v^*$ is strictly greater than C. This contradiction proves that our initial assumption was false, and thus v^* must belong to the convex hull of V.

This completes the proof of the lemma.

As an immediate consequence, items four and five of Theorem 7 — concerning the compactness of the information matrix and the number of design's support points— follows directly from this lemma and Carathéodory's Theorem. This completes the proof of Theorem 7.

Corollary

If conditions (a)–(e) are satisfied, then the information matrix of any design ξ can be written as a convex combination of at most $m(m+1)/2+1$ matrices of the form $f(x_i)f^T(x_i)$. That is,

$$M(\xi) = \sum_{i=1}^{n} w_i f(x_i) f^T(x_i), \; w_i \geq 0, \; \sum w_i = 1, \quad \text{where } n \leq m(m+1)/2+1.$$

Remark

Since the information matrix is symmetric, it is fully determined by $m(m+1)/2$ elements. In other words, it can be associated with a vector in $\mathbb{R}^{m(m+1)/2}$.

Remark

We have obtained the following important consequence: continuous optimal designs can be constructed from designs of the form

$$\xi = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \\ w_1 & w_2 & \cdots & w_n \end{pmatrix}, \quad n \leq m(m+1)/2+1.$$

Comments

This slide presents a corollary and two remarks that follow from the previously established representation theorem for information matrices under conditions (a) through (e). The corollary states that for any design ξ, the information matrix $M(\xi)$ can be written as a convex combination of at most $m(m+1)/2+1$ rank-one symmetric matrices of the form $f(x_i)f^T(x_i)$. In other words, there exist support points x_1 through x_n and nonnegative weights w_1 through w_n summing to one, such that the information matrix equals the sum over i of $w_i f(x_i)f^T(x_i)$, with the number of support points n not exceeding the stated bound.

The first remark emphasizes a geometric interpretation. Since the matrix M is symmetric, it is uniquely determined by its distinct entries, namely $m(m+1)/2$ parameters. Hence, M can be identified with a point in a Euclidean space of corresponding dimension.

The second remark highlights a constructive implication for optimal design. Any approximate design that is optimal in a given sense can be replaced by a design with no more than $m(m+1)/2+1$ support points. Thus, we have obtained an upper bound for the number of support points of the optimal design.

Key Definitions

- A design ξ is called ***nonsingular*** if its information matrix $M(\xi)$ is invertible.
- For any nonsingular design ξ, the matrix
$$D(\xi) = M(\xi)^{-1}$$
is defined as the ***inverse information matrix***.

- There is generally ***no single optimal design*** ξ^* that minimizes the inverse information matrix for all other designs ξ, i.e.,
$$D(\xi^*) \leq D(\xi) \quad \text{for all } \xi \in \Xi.$$
- This means we cannot find a design that is "better" than all others in every possible way.

- To solve this, we define optimality using specific functions that have a clear statistical meaning.
- These are typically either:
  - ***Concave functions*** on $M(\xi)$.
  - ***Convex functions*** on $D(\xi)$.

Comments

In regression design theory, a fundamental concept is that of a nonsingular design. A design is called nonsingular if its information matrix is invertible. This condition is essential: only for such designs can we uniquely and unbiasedly estimate all the parameters of the regression model. If the information matrix is singular, then the parameter vector is not fully identifiable based on the experiment.

For any nonsingular design ξ, we define the matrix $D(\xi)$ as the inverse of its information matrix, that is, $D(\xi) = M(\xi)^{-1}$. This matrix characterizes the precision of parameter estimation: smaller entries in D correspond to more accurate estimates.

However, in most practical situations, there does not exist a design ξ^* such that $D(\xi^*) \leq D(\xi)$ for every other nonsingular design ξ. In other words, the set of nonsingular designs does not contain a universally best element in terms of the matrix ordering.

Therefore, to compare and rank designs in a meaningful way, one introduces so-called optimality criteria. These are real-valued functions defined either on the set of information matrices or on their inverses. Concave functions on information matrices — or equivalently, convex functions on their inverses — are typically used. These criteria have clear statistical meaning: for example, minimizing the average variance of parameter estimates or minimizing the volume of confidence ellipsoids. They allow us to formally define what we mean by the “best” design under specific optimality goals.

D-criterion

The name derives from "determinant." The D-optimality criterion is written as:

$$\log\det \mathrm{M}(\xi) \to \sup_{\xi\in\Xi} \quad \text{or} \quad \log\det \mathrm{D}(\xi) \to \inf_{\xi\in\Xi}.$$

It corresponds to minimizing the volume of the confidence ellipsoid:

$$\left\{\widetilde{\theta} : (\widetilde{\theta}-\widehat{\theta})^{\mathrm{T}}\mathrm{D}\widehat{\theta}(\widetilde{\theta}-\widehat{\theta}) \le \alpha\right\},$$

where α is a constant depending only on the confidence level.

L-criterion

The L-optimality criterion is:

$$\operatorname{tr}\mathrm{LD}(\xi) \to \inf_{\xi\in\Xi_{\text{н}}},$$

where $\Xi_{\mathrm{NS}} = \{\xi \in \Xi : \det \mathrm{M}(\xi) \neq 0\}$, and $\mathrm{L} \in \mathbb{R}^{(\mathrm{d}+1)\times(\mathrm{d}+1)}$ is a fixed nonnegative definite matrix.
Its statistical meaning is minimization of generalized quadratic loss:

$$\mathbb{E}(\widehat{\theta}-\theta)^{\mathrm{T}}\mathrm{L}(\widehat{\theta}-\theta).$$

Comments

We now turn to specific optimality criteria used to compare and choose among different designs. The most classical of these is the D-optimality criterion. The letter "D" refers to the determinant of the information matrix. According to this criterion, the optimal design is the one that maximizes the logarithm of the determinant of the information matrix $\mathrm{M}(\xi)$ over all admissible designs. Equivalently, one can minimize the logarithm of the determinant of the inverse matrix $\mathrm{D}(\xi)$. In mathematical terms, the criterion is written as: $\log\det \mathrm{M}(\xi) \to \sup_{\xi\in\Xi}$, or, alternatively, $\log\det \mathrm{D}(\xi) \to \inf_{\xi\in\Xi}$.

This criterion has a clear geometric interpretation. It corresponds to minimizing the volume of the confidence ellipsoid for the parameter vector. The ellipsoid is defined as the set of all parameter values $\widetilde{\theta}$ such that the quadratic form of the deviation, with respect to the matrix $\mathrm{D}(\xi)$, does not exceed a fixed constant α. This constant α depends only on the chosen confidence level. Thus, the D-criterion aims to minimize the uncertainty region for the estimated parameters.

Another widely used criterion is the L-optimality criterion. It is defined as the trace of the product of a fixed symmetric nonnegative definite matrix L and the matrix $\mathrm{D}(\xi)$. The goal is to minimize this trace over all nonsingular designs. The statistical meaning of this criterion is minimization of the expected value of the quadratic loss, where the loss is defined with respect to the matrix L. Specifically, it minimizes the expected value of the scalar product of the estimation error with the matrix L. This framework allows tailoring the criterion to emphasize certain parameters or combinations thereof.

e_k-criterion

A design $\xi \in \Xi$ is called e_k-optimal if

$$e_k^T M^-(\xi) e_k \to \inf_{\xi \in \Xi_{e_k}}, \quad \Xi_{e_k} = \{\xi \in \Xi : e_k^T M^-(\xi) M(\xi) = e_k^T\}.$$

This criterion minimizes the variance of the estimator $\widehat{\theta}_k$.

E-criterion

A design $\xi \in \Xi_{NS}$ is called E-optimal if

$$\lambda_{\min}(M(\xi)) \to \sup_{\xi \in \Xi_{NS}} \quad \text{or} \quad \lambda_{\max}(D(\xi)) \to \inf_{\xi \in \Xi_{NS}}.$$

This criterion maximizes the smallest eigenvalue of the information matrix.

G-criterion

A design $\xi \in \Xi_{NS}$ is called G-optimal if

$$\max_{t \in \chi} d(t, \xi) \to \inf_{\xi \in \Xi_{NS}}, \quad d(t, \xi) =$$

This criterion minimizes the maximum prediction variance over the design space.

Comments

On this slide we discuss three important optimality criteria used in the theory of experimental designs: the e_k-criterion, the E-criterion, and the G-criterion.

First, the e_k-criterion focuses on minimizing the variance of the estimator for the single parameter coordinate θ_k. In other words, it aims to reduce the k-th diagonal element of the covariance matrix, which corresponds to the uncertainty in estimating θ_k. It is important to note that e_k-optimal designs are generally degenerate, meaning that not all parameters can be estimated unbiasedly from such designs. However, these designs are useful when only a subset of parameters is of interest. Additionally, the e_k-criterion is a special case of the more general L-optimality criterion when the weighting matrix L is chosen as the outer product of the k-th standard basis vector with itself.

The E-criterion aims to improve the overall worst-case precision by maximizing the smallest eigenvalue of the information matrix. Equivalently, it minimizes the largest eigenvalue of the covariance matrix. This means that the longest axis of the confidence ellipsoid is shortened, leading to a more balanced estimation accuracy across all parameters.

Finally, the G-criterion is concerned with prediction accuracy. It minimizes the maximum variance of predicted values over the entire design space. This criterion ensures that the worst prediction variance at any point is as small as possible, which is crucial for reliable response surface estimation.

Together, these criteria provide different perspectives and tools to design experiments depending on the specific goals: precise estimation of individual parameters, balanced overall estimation, or reliable prediction.

General Form of Optimality Criteria

Optimality criteria in general form

All discussed optimality criteria can be expressed as

$$\Psi(\mathrm{M}^{-}(\xi)) \to \inf_{\xi} \quad \text{or} \quad \Phi(\mathrm{M}(\xi)) \to \sup_{\xi}.$$

Here, the functions Ψ and Φ satisfy the following properties:

(a) Monotonicity:

$$\Phi(\mathrm{M}(\xi_1)) \le \Phi(\mathrm{M}(\xi_2)) \quad \text{if} \quad \mathrm{M}(\xi_1) \le \mathrm{M}(\xi_2).$$

(b) Homogeneity:

$$\Phi(\theta\mathrm{M}(\xi)) \le \gamma(\theta)\Phi(\mathrm{M}(\xi)),$$

where $\gamma(\theta)$ is a non-decreasing function.

(c) Concavity (convexity for Ψ):

$$\Phi(\mathrm{M}((1-\alpha)\xi_1 + \alpha\xi_2)) \ge (1-\alpha)\Phi(\mathrm{M}(\xi_1)) + \alpha\Phi(\mathrm{M}(\xi_2)).$$

Remark

A more complete list of criteria can be found in the experimental design literature.

Comments

This slide presents the general mathematical form in which most optimality criteria for experimental design can be expressed. These criteria are usually written either in terms of the inverse information matrix or the information matrix itself. More precisely, we try to minimize a function Ψ of the inverse information matrix — that is, $\Psi(\mathrm{M}^{-}(\xi))$ — or equivalently, to maximize a function Φ of the information matrix — that is, $\Phi(\mathrm{M}(\xi))$.

For these functions to define reasonable and meaningful criteria, they are typically required to satisfy three important properties.

First is monotonicity. This means that if one design has an information matrix smaller than another in the matrix ordering sense, then the value of the function Φ for that design must also be smaller. In other words, if $\mathrm{M}(\xi_1) \le \mathrm{M}(\xi_2)$, then $\Phi(\mathrm{M}(\xi_1)) \le \Phi(\mathrm{M}(\xi_2))$. This guarantees that better information matrices give better criterion values.

Second is homogeneity. If we scale the information matrix by a positive factor θ, then the criterion Φ scales no faster than some corresponding function $\gamma(\theta)$, which itself is non-decreasing. That is, $\Phi(\theta\mathrm{M}(\xi)) \le \gamma(\theta)\Phi(\mathrm{M}(\xi))$.

Third is concavity. This means that if we mix two designs — say, ξ_1 and ξ_2 — with weights $1-\alpha$ and α, then the criterion value for the mixed design is at least as large as the weighted average of the criterion values for the individual designs. Mathematically, $\Phi(\mathrm{M}((1-\alpha)\xi_1 + \alpha\xi_2)) \ge (1-\alpha)\Phi(\mathrm{M}(\xi_1)) + \alpha\Phi(\mathrm{M}(\xi_2))$. This property is crucial for optimization because it guarantees that local maxima are also global.

Taken together, these properties ensure that the function Φ behaves in a predictable and mathematically nice way, which allows us to formulate and solve design optimization problems efficiently. There are many other optimality criteria that are used in practice, but their study is beyond the scope of our course.

Kiefer–Wolfowitz Equivalence Theorem

For the model (1o) under assumption (a) – (f) if the set of information matrices is compact, then the following conditions are equivalent for a design ξ^*:

(a) ξ^* is ***D-optimal***;

(b) ξ^* is ***G-optimal***;

(c) $\mathbf{max}_{x\in\mathcal{X}} d(x, \boldsymbol{\xi^*}) = m$,

where $d(x, \xi^*) = f^T(x)M^{-1}(\xi^*)f(x)$.

Moreover, if ξ^* has finite support, this maximum is attained at the support points x_i^*. All D-optimal designs share the same information matrix. Under the conditions of the theorem $\xi^* \in \Xi$.

Notation

$D(\xi) = M(\xi)^{-1}$ is the inverse information matrix; m is the number of parameters.

Comments

This slide introduces one of the central results in the theory of optimal experimental design — the Kiefer–Wolfowitz Equivalence Theorem. It establishes a deep connection between different optimality criteria and provides a powerful tool for verifying the optimality of a given design.

According to the theorem, if the set of possible information matrices is compact — which is a standard regularity assumption — then three conditions are equivalent for a design denoted ξ^*.

First, condition (a): the design ξ^* is D-optimal. This means it maximizes the determinant of the information matrix, or equivalently, minimizes the volume of the confidence ellipsoid for the parameter estimates.

Second, condition (b): the same design ξ^* is also G-optimal. That is, it minimizes the maximum variance of the predicted response over the entire experimental region.

Third, condition (c): the maximum value of the function $d(x, \xi^*)$, taken over all points x in the design space, is equal to m, where m is the number of parameters in the model. Here, the function $d(x, \xi)$ is defined as the transpose of the regression vector at x times the inverse information matrix times the regression vector at x. This quantity represents the variance of the predicted response at point x.

Importantly, if the optimal design ξ^* is concentrated on a finite set of points, then this maximum value of the function d is achieved exactly at the support points of the design. These are the points where the design assigns positive weight.

Another significant implication of the theorem is that all D-optimal designs — even if they differ in their specific support points — share the same information matrix. This means the precision of estimation is the same across all such designs.

This theorem is extremely useful because it allows us to check D-optimality by verifying a simpler G-optimality condition or by calculating the function d and checking whether its maximum equals m.

Lemma 11

Let A be an arbitrary positive definite matrix of size $m \times m$. Then the following integral representation for its determinant holds:

$$(\det A)^{-1/2} = \frac{1}{\pi^{m/2}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-x^T Ax}\, dx_1 \dots dx_m.$$

Proof:

- We use the ***classical identity*** from real analysis:
$$\int_{-\infty}^{\infty} e^{-z^2}\, dz = \sqrt{\pi}.$$
- First, assume that A is a ***diagonal matrix***, where $A = \mathbf{\Lambda}$ with nonzero diagonal elements λ_i. Then, the integral becomes:
$$\int_{\mathbb{R}^m} e^{-x^T Ax}\, dx = \int_{\mathbb{R}^m} e^{-\sum_{i=1}^m \lambda_i x_i^2}\, dx = \prod_{i=1}^m \int_{-\infty}^{\infty} e^{-\lambda_i x_i^2}\, dx_i.$$
- Each one-dimensional integral evaluates to $\sqrt{\pi/\lambda_i}$, so the result is:
$$\int_{\mathbb{R}^m} e^{-x^T Ax}\, dx = \prod_{i=1}^m \sqrt{\frac{\pi}{\lambda_i}} = \pi^{m/2} (\det A)^{-1/2}.$$

Comments

To prove the Kiefer–Wolfowitz equivalence theorem, we need several auxiliary results. This slide presents the first such lemma — a classic identity from multivariate calculus.

The lemma states that for any positive definite matrix A of size $m \times m$, we can express the reciprocal square root of the determinant of A as a Gaussian integral over $\mathbb{R}^m$. Specifically, the formula involves the exponential of minus one-half times $x^T Ax$, integrated over the whole space.

This result is not just an elegant analytical identity — it plays an important role in optimal design theory because it connects determinants with Gaussian densities, which frequently appear in statistical estimation, particularly in maximum likelihood and Bayesian contexts.

The proof is straightforward for diagonal matrices. We start by recalling the classical Gaussian integral: the integral of e^{-z^2} over the real line equals $\sqrt{\pi}$. Then, if the matrix A is diagonal, say with elements λ_1 through λ_m, the quadratic form reduces to the sum of $\lambda_i x_i^2$ for each component. The multivariate integral then splits into a product of independent one-dimensional integrals, each of which evaluates to $\sqrt{\pi/\lambda_i}$. Multiplying these together yields the final formula.

In the next step, we will generalize this to arbitrary positive definite matrices.

Proof of the Lemma (continued)

- If A is an arbitrary ***positive definite matrix***, we can use the ***spectral decomposition theorem*** to write:

$$A = P^T \mathbf{\Lambda} P, \quad \text{where } P^T P = I \text{ and } \mathbf{\Lambda} \text{ is diagonal.}$$

- Let's make a change of variables: $x = Py$. This simplifies the quadratic form:

$$x^T A x = (Py)^T A P y = y^T \mathbf{\Lambda} y = \sum_{i=1}^{m} \lambda_i y_i^2.$$

- Since the columns of P form an orthonormal basis, the ***Jacobian*** of this transformation is equal to one. Therefore, the integral becomes:

$$\int_{\mathbb{R}^m} e^{-x^T A x} dx = \int_{\mathbb{R}^m} e^{-y^T \mathbf{\Lambda} y} dy = \pi^{m/2} (\det A)^{-1/2}.$$

This completes the proof. □

Comments

To complete the proof of the lemma, we now consider the case when matrix A is arbitrary but still positive definite. In this more general setting, we apply the spectral decomposition theorem. According to this result, matrix A can be written as $P^T \Lambda P$, where P is an orthogonal matrix ($P^T P = I$) and Λ is a diagonal matrix containing the eigenvalues of A on its main diagonal.

We now make a change of variables: let vector $x = Py$. This transformation corresponds to a rotation of the coordinate system. Since P is an orthogonal matrix, this change preserves volume, and therefore the Jacobian of the transformation is equal to one.

Substituting this new variable into the quadratic form $x^T A x$, we get: $y^T \Lambda y$. Because Λ is diagonal, this reduces to a sum of $\lambda_i y_i^2$, where i runs from 1 to m.

Thus, the multivariate integral of the exponential of $-x^T A x$ becomes the same as in the diagonal case. The integral over y is then the product of m one-dimensional Gaussian integrals, each evaluating to $\sqrt{\pi/\lambda_i}$. Therefore, we again arrive at the expression $\pi^{m/2} (\det A)^{-1/2}$.

This confirms that the lemma holds for any positive definite matrix A.

Lemma 12

Let A and B be arbitrary positive definite matrices of size $m \times m$. Then the following inequality holds:

$$\det(\alpha A + (1-\alpha)B) \geq (\det A)^{\alpha}(\det B)^{1-\alpha}, \quad \text{for all } \alpha \in [0,1],$$

with equality if and only if $A = B$.

Proof: The proof is based on Hölder's inequality.

- Let $p > 1$ and define $q = \frac{p}{p-1}$. Let $f \in L^p$, $g \in L^q$ be functions on a measurable set $X \subset \mathbb{R}^n$.
- Then:

$$\int_X f(x)g(x)dx \leq \left(\int_X f(x)^p dx\right)^{1/p} \left(\int_X g(x)^q dx\right)^{1/q},$$

with equality if and only if

$$\frac{f(x)^p}{\int_X f(x)^p dx} = \frac{g(x)^q}{\int_X g(x)^q dx}.$$

Comments

To advance the proof of the Kiefer–Wolfowitz Equivalence Theorem, we need another auxiliary result: for any two positive definite matrices A and B, both of size $m \times m$, the determinant of their convex combination, $\alpha A + (1-\alpha)B$, is at least the product of $(\det A)^{\alpha}$ and $(\det B)^{1-\alpha}$, for any α between 0 and 1.

Equality holds only when $A = B$. This inequality will allow us to further establish concavity of the determinant function, which is critical for linking D-optimality, maximizing the determinant of the information matrix, to G-optimality, minimizing the maximum prediction variance.

The proof hinges on Hölder's inequality. For functions f in L^p and g in L^q, where $q = p/(p-1)$ and $p > 1$, the integral of $f(x)g(x)$ over a measurable set X is bounded by the p-norm of f times the q-norm of g. Equality holds only when the normalized p-th power of f equals the normalized q-th power of g. Applying Hölder's inequality, we derive the determinant inequality by expressing the determinant in terms of integrals, building on the Gaussian integral from the previous lemma.

Completion of the Proof (Determinant Inequality)

- Let $p = \frac{1}{\alpha}, \quad q = \frac{1}{1-\alpha}, \quad f(x) = e^{-\alpha x^T A x}, \quad g(x) = e^{-(1-\alpha)x^T B x}.$
- Then:
$$\int_{\mathbb{R}^n} e^{-z^T(\alpha A+(1-\alpha)B)z}\,dz = \int f(z)g(z)\,dz$$
- Applying Hölder's inequality:
$$\int f(z)g(z)\,dz \le \left(\int f^{1/\alpha}(z)\,dz\right)^{\alpha}\left(\int g^{1/(1-\alpha)}(z)\,dz\right)^{1-\alpha}$$
- Using the Gaussian integral formula:
$$\frac{\pi^{n/2}}{\sqrt{\det(\alpha A+(1-\alpha)B)}} \le \left(\frac{\pi^{n/2}}{\sqrt{\det A}}\right)^{\alpha}\left(\frac{\pi^{n/2}}{\sqrt{\det B}}\right)^{1-\alpha}$$
- Therefore:
$$\det(\alpha A+(1-\alpha)B) \ge (\det A)^{\alpha}(\det B)^{1-\alpha}$$

Equality holds if and only if $A = B$. □

Comments

To complete the proof of the determinant inequality, we apply Hölder's inequality directly to the Gaussian integral involving the convex combination of matrices A and B. First, we define two functions: $f(z) = e^{-\alpha z^T A z}$, and $g(z) = e^{-(1-\alpha)z^T B z}$. We also set the exponents $p = 1/\alpha$ and $q = 1/(1-\alpha)$.

With these definitions, the integral of the exponential of $-z^T(\alpha A + (1-\alpha)B)z$ becomes the integral of $f(z)g(z)$. Hölder's inequality now tells us that this integral is less than or equal to the product of the L^p-norm of f raised to α and the L^q-norm of g raised to $1-\alpha$.

The norms of f and g correspond to Gaussian integrals. From the previous lemma, we know that the integral of $e^{-z^T Az}$ over all of Euclidean space is equal to $\pi^{n/2}/\sqrt{\det A}$. Applying this fact to f and g, we arrive at an inequality for the integral involving the convex combination of A and B.

Finally, rearranging the inequality gives us the determinant inequality: $\det(\alpha A + (1-\alpha)B) \ge (\det A)^{\alpha}(\det B)^{1-\alpha}$. Equality holds only if $A = B$. This concludes the proof of the lemma.

Lemma 13

The function $f(A) = \ln\det A$ is strictly concave on the set of positive definite matrices of size $m \times m$.

Proof: By definition, it suffices to show that for any $0 < \alpha < 1$ and $A \neq B$, the inequality

$$\ln\det(\alpha A + (1-\alpha)B) > \alpha \ln\det A + (1-\alpha)\ln\det B$$

holds. This follows immediately by taking the logarithm of the inequality from the previous lemma. □

Lemma 14

For any differentiable matrix function $A(\alpha)$ with $A(\alpha) > 0$ (the matrix of size $m \times m$), the following identity holds:

$$\frac{\partial}{\partial\alpha}\ln\det A(\alpha) = \operatorname{tr}\left(A^{-1}(\alpha)\frac{\partial A(\alpha)}{\partial\alpha}\right)$$

Comments

In this slide, we present two important lemmas that play a central role in the analysis of optimality criteria and the proof of the equivalence theorem.

The first lemma states that the function $f(A) = \ln\det A$ is strictly concave on the set of positive definite matrices of size $m \times m$. That is, if we take any two distinct positive definite matrices A and B, and form their convex combination weighted by α and $1-\alpha$, then the logarithm of the determinant of this combination is strictly greater than the weighted sum of the logarithms of the determinants. This is a classic result in matrix analysis, and it follows directly by taking the logarithm of the determinant inequality we proved earlier. That inequality stated that the determinant of the convex combination of A and B is at least the product of their determinants raised to the corresponding weights. Taking the logarithm transforms the multiplicative inequality into an additive one, and the strict inequality holds whenever A and B are not equal.

The second lemma provides a useful identity for differentiating the log-determinant function. If we have a matrix-valued function $A(\alpha)$, and this matrix is positive definite for all values of α, then the derivative of $\ln\det A(\alpha)$ with respect to α is equal to the trace of $A^{-1}(\alpha)$ times the derivative of $A(\alpha)$. This result comes from matrix calculus and is widely used in optimization, especially in contexts where we differentiate likelihood functions or objective functions involving log determinants. It provides an efficient way to compute gradients without directly differentiating the determinant itself, which would be more complicated.

These lemmas will be used in the next steps of the proof.

Proof of Lemma on Derivative of log-det

Proof: We use the rule for the derivative of the logarithm:

$$\frac{\partial}{\partial\alpha}\ln\det \mathrm{A}(\alpha) = \frac{1}{\det \mathrm{A}(\alpha)}\cdot\frac{\partial}{\partial\alpha}\det \mathrm{A}(\alpha)$$

Using the expansion of the determinant:

$$\det \mathrm{A} = \sum_{i,j=1}^{m}(-1)^{i+j}a_{ij}A_{ij}, \quad \text{where } A_{ij} \text{ are cofactors}$$

Since cofactors do not depend on the element a_{ij}, we get:

$$\frac{\partial \det \mathrm{A}}{\partial a_{ij}} = (-1)^{i+j}A_{ij}$$

Using the identity:

$$a^{ij} = \frac{(-1)^{i+j}A_{ji}}{\det \mathrm{A}} \quad (\text{entries of } \mathrm{A}^{-1})$$

We obtain:

$$\frac{\partial}{\partial\alpha}\ln\det \mathrm{A}(\alpha) = \sum_{i,j=1}^{m}\underbrace{\frac{(-1)^{i+j}A_{ij}}{\det \mathrm{A}}}_{a^{ji}}\frac{\partial a_{ij}(\alpha)}{\partial\alpha} = \operatorname{tr}\left(\mathrm{A}^{-1}\cdot\frac{\partial \mathrm{A}}{\partial\alpha}\right) \qquad \square$$

Comments

To prove the identity for the derivative of the log determinant, we begin by applying the standard calculus rule for differentiating a logarithmic function. The derivative of $\ln\det \mathrm{A}(\alpha)$ with respect to α is equal to $1/\det \mathrm{A}(\alpha)$ times the derivative of $\det \mathrm{A}(\alpha)$ with respect to α.

Next, we recall the classical formula for the determinant of a square matrix. It can be written as a sum over all elements a_{ij} multiplied by their corresponding cofactors A_{ij}, each weighted by the sign factor $(-1)^{i+j}$. This expansion allows us to compute the partial derivative of the determinant with respect to any element a_{ij}, which equals the signed cofactor.

Then, we recall the standard expression for the entries of the inverse matrix. The entry in position (i,j) of the inverse matrix A^{-1} is equal to the signed cofactor A_{ji} divided by the determinant of A. Using this identity, we rewrite the derivative of the determinant as a sum over i and j of the derivative of a_{ij} times the corresponding entry of the inverse matrix.

Combining all these steps, the derivative of $\ln\det \mathrm{A}(\alpha)$ becomes the trace of the product of A^{-1} and $\partial \mathrm{A}/\partial\alpha$. This completes the proof of the lemma.

PART III. Optimal design theory
(LECTURE 4)

Shpilev Petr Valerievich

Faculty of Mathematics and Mechanics, SPbU

September, 2025

Санкт-Петербургский государственный университет

Comments

In this lecture we continue the study of optimal design theory by focusing on the equivalence theorems. We begin with the sensitivity function and continue the proof of the Kiefer–Wolfowitz theorem, the fundamental result which provides the key condition for verifying D-optimality of a design. Several examples will illustrate how these conditions can be applied in practice, including discrete designs.

We then extend the equivalence framework to other criteria — E-, L-, and e_k- optimality — covering both nonsingular and singular cases. Special attention is given to extremal polynomials and block-diagonalization techniques, which make the verification of optimality more tractable.

Finally, we discuss analytical solutions of optimal design problems, showing how the theory leads to explicit D-optimal designs in polynomial and trigonometric regression models. This lecture thus connects the abstract equivalence principles with concrete methods for constructing and testing optimal experimental designs.

Lower Bound for the supremum of the sensitivity function

Lemma 15

For any nonsingular design ξ, the following inequality holds:

$$\sup_{x} d(x, \xi) \geq m.$$

Proof:

$$\sup_{x} d(x, \xi) \geq \int d(x, \xi)\, \xi(dx) = \operatorname{tr}\left[M^{-1}(\xi) \int f(x) f^{T}(x)\, \xi(dx)\right]$$

$$= \operatorname{tr}\left(M^{-1}(\xi) M(\xi)\right) = \operatorname{tr} I_m = m. \quad \square$$

Explanation

- The function $d(x, \xi) = f^{T}(x) M^{-1}(\xi) f(x)$ is known as the ***sensitivity function***. It represents the variance of the predicted response at point x.
- This lemma establishes a ***fundamental lower bound*** (m) for the maximum variance over the entire design space.
- This result is a crucial component of the ***Kiefer–Wolfowitz Equivalence Theorem***, which uses the condition $\sup d(x, \xi) = m$ as a test for D-optimality.

Comments

This lemma gives us a simple but important inequality: for any nonsingular design, the maximum of the sensitivity function is always at least equal to the number of parameters in the model, denoted by m. The sensitivity function is defined as the transpose of the regression vector at point x multiplied by the inverse of the information matrix times the regression vector again.

In the proof, we use Jensen's inequality for the supremum: the supremum of a function is always greater than or equal to its integral with respect to any probability measure. We apply this to the sensitivity function integrated over the design measure. The integral of the outer product of the regression vector gives us back the information matrix. Then we simply take the trace of the identity matrix, which gives us m.

This inequality tells us that no matter how we choose the design, we cannot reduce the maximum prediction variance below the number of parameters. This is especially important in the context of D-optimal designs. In fact, one of the central results of the Kiefer–Wolfowitz Equivalence Theorem is that D-optimality is achieved when this supremum actually equals m. So this lemma provides a necessary lower bound and builds the foundation for what follows.

Proof of the Kiefer–Wolfowitz Theorem:

The scheme of the proof is: (a) ⇒ (c) ⇒ (b) ⇒ (a)

1. (a) ⇒ (c) : Let ξ_D be a D-optimal design. Consider the design $\xi_\alpha = (1-\alpha)\xi_D + \alpha\xi_x$, where $\xi_x = \{x, 1\}$ is a design concentrated at point x, and $0 < \alpha < 1$.

Due to the D-optimality of ξ_D, the following inequality holds:

$$\ln\det M(\xi_\alpha) \le \ln\det M(\xi_D) \Leftrightarrow \frac{\ln\det M(\xi_\alpha) - \ln\det M(\xi_D)}{\alpha} \le 0$$

Passing to the limit as $\alpha \to 0+$, we obtain: $\left.\frac{\partial}{\partial\alpha}\ln\det M(\xi_\alpha)\right|_{\alpha=0+} =$

$$= \operatorname{tr} M^{-1}(\xi_\alpha)\left.\frac{\partial M(\xi_\alpha)}{\partial\alpha}\right|_{\alpha=0+} = \operatorname{tr} M^{-1}(\xi_\alpha)\left.\frac{\partial((1-\alpha)M(\xi_D)+\alpha M(\xi_x))}{\partial\alpha}\right|_{\alpha=0+} =$$

$$= \operatorname{tr} M^{-1}(\xi_D)(f(x)f^T(x) - M(\xi_D)) = d(x,\xi_D) - m \le 0$$

Hence, $d(x,\xi_D) \le m$ for all x, and therefore, $\sup_x d(x,\xi_D) = m$.

Comments

We begin the proof of the Kiefer–Wolfowitz equivalence theorem by showing that condition (a) implies condition (c). The proof is constructed as a cycle of implications, and here we consider the first step.

We assume that the design denoted by the Greek letter ξ with subscript D is optimal in the sense of the D-criterion. This means that it maximizes the logarithm of the determinant of the information matrix over all possible designs.

To test whether this property implies a certain inequality involving the sensitivity function, we perturb the optimal design slightly. We construct a new design by combining the original D-optimal design with a design concentrated at an arbitrary point x. This mixture is controlled by a small positive parameter α.

Since the original design is optimal, any such perturbation should not increase the logarithm of the determinant. Dividing the difference in the log-determinants by α and taking the limit as α approaches zero gives us the directional derivative of the objective function with respect to this perturbation.

Using the formula for the derivative of the logarithm of the determinant and applying the linearity of the matrix derivative, we arrive at an expression for the derivative in terms of the trace of the product of the inverse information matrix and the matrix difference between the outer product of the regression function and the original information matrix.

This expression equals the sensitivity function evaluated at point x minus the number of parameters. Since this must be less than or equal to zero, it follows that the sensitivity function does not exceed the number of parameters anywhere. Therefore, its supremum equals the number of parameters, as claimed in condition (c).

2. (c) ⇒ (b)
Suppose the design ξ^* satisfies: $\sup_x d(x, \xi^*) = m$, then by Lemma 15, the design ξ^* is G-optimal.
3. (b) ⇒ (a)
We prove by contradiction. Suppose that a G-optimal design exists which is not D-optimal. Let this design be ξ^* and let ξ_D be some D-optimal design.

Consider the design:

$$\xi_\alpha = (1-\alpha)\xi^* + \alpha\xi_D.$$

By Lemma 13, we have:

$$\ln\det M(\xi_\alpha) \geq (1-\alpha)\ln\det M(\xi^*) + \alpha\ln\det M(\xi_D).$$

$$\Rightarrow \quad \ln\det M(\xi_\alpha) - \ln\det M(\xi^*) \geq \alpha\left(\ln\det M(\xi_D) - \ln\det M(\xi^*)\right) > 0,$$

since ξ^* is not D-optimal.

$$\Rightarrow \quad \frac{\ln\det M(\xi_\alpha) - \ln\det M(\xi^*)}{\alpha} > 0.$$

Comments

In the second part of the proof, we show that condition (c) implies condition (b). If the supremum of the sensitivity function for a design equals the number of parameters, then, according to the referenced lemma, this design is optimal in the sense of the G-criterion. That completes the second implication.

Next, we establish the implication from condition (b) to condition (a) by contradiction. Assume there exists a design that is G-optimal but not D-optimal. Denote this G-optimal design by ξ^*, and let ξ_D be a design that is D-optimal.

We construct a new design as a convex combination of ξ^* and ξ_D, controlled by a small positive parameter α.

From the earlier lemma about the strict concavity of the logarithm of the determinant, it follows that the logarithm of the determinant of the information matrix of the combined design is greater than or equal to the convex combination of the logarithms of the determinants of the individual designs.

Subtracting the logarithm corresponding to the G-optimal design and using the assumption that it is not D-optimal, we conclude that the difference is strictly positive. Dividing this difference by α yields a positive quantity.

From the previous step, we get:

$$\frac{\partial}{\partial\alpha}\ln\det \mathrm{M}(\xi_\alpha)\Big|_{\alpha=0+} = \mathrm{tr}\left(\mathrm{M}^{-1}(\xi^*)\mathrm{M}(\xi_\mathrm{D})\right) - \mathrm{m} > 0, \Rightarrow \mathrm{tr}\left(\mathrm{M}^{-1}(\xi^*)\mathrm{M}(\xi_\mathrm{D})\right) > \mathrm{m}.$$

On the other hand, since the design ξ^* is G-optimal, we have:

$$\mathrm{f}^\mathrm{T}(\mathrm{x})\mathrm{M}^{-1}(\xi^*)\mathrm{f}(\mathrm{x}) \le \mathrm{m} \quad \forall \mathrm{x},$$

which implies:

$$\mathrm{tr}\left(\mathrm{M}^{-1}(\xi^*)\mathrm{M}(\xi_\mathrm{D})\right) = \int \mathrm{f}^\mathrm{T}(\mathrm{x})\mathrm{M}^{-1}(\xi^*)\mathrm{f}(\mathrm{x})\,\xi_\mathrm{D}(\mathrm{dx}) \le \mathrm{m}.$$

This is a contradiction $\Rightarrow$ G-optimal ξ^* is also D-optimal.
Therefore, all three conditions (a), (b), and (c) are equivalent.

- The maximum of $\mathrm{d}(\mathrm{x},\xi)$ is attained at the support points of an optimal design (by Lemma 15).
- All optimal designs have the same information matrix (by Lemma 13).

The Kiefer–Wolfowitz Theorem is proved. □

Comments

We continue the proof by differentiating the logarithm of the determinant of the information matrix for the convex combination of the G-optimal and D-optimal designs. The derivative at zero from the right turns out to be strictly positive, which implies that the trace of the product of the inverse information matrix of the G-optimal design and the information matrix of the D-optimal design is strictly greater than the number of parameters.

However, because the original design is G-optimal, the sensitivity function is bounded above by the number of parameters for all points in the design space. Integrating this bound with respect to the D-optimal design gives a contradiction: the same trace must be less than or equal to the number of parameters.

This contradiction shows that any G-optimal design must also be D-optimal. As a result, we conclude the equivalence of all three conditions stated in the theorem.

The fact that the maximum of the sensitivity function is attained at the support points of the optimal design (in the case of finite support) follows directly from the earlier lemma, which becomes an equality in this case.

Finally, the fact that all optimal designs have identical information matrices follows from the concavity property of the logarithm of the determinant.

Example: an approximate design

Design and model

- Regression vector: $\mathrm{f}(\mathrm{x}) = (1,\ \mathrm{x},\ \mathrm{x}^2)^{\mathrm{T}}$
- Design space: $\chi = [-1,\ 1]$
- Candidate design:

$$\xi^* = \begin{pmatrix} -1 & 0 & 1 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}$$

Information matrix and its inverse

- Compute $\mathrm{M}(\xi^*) = \sum \omega_{\mathrm{i}} \mathrm{f}(\mathrm{x}_{\mathrm{i}}) \mathrm{f}^{\mathrm{T}}(\mathrm{x}_{\mathrm{i}})$ and its inverse:

$$\mathrm{M}(\xi^*) = \begin{pmatrix} 1 & 0 & \frac{2}{3} \\ 0 & \frac{2}{3} & 0 \\ \frac{2}{3} & 0 & \frac{2}{3} \end{pmatrix} \qquad \mathrm{M}^{-1}(\xi^*) = \mathrm{D}(\xi^*) = \begin{pmatrix} 3 & 0 & -3 \\ 0 & \frac{3}{2} & 0 \\ -3 & 0 & \frac{9}{2} \end{pmatrix}$$

Comments

This example demonstrates how we can use the Kiefer–Wolfowitz equivalence theorem to verify whether a specific design is D-optimal. We consider a quadratic regression model, where the regression vector consists of one, x, and x^2. The design space is the closed interval from -1 to 1.

Our candidate design is symmetric, with support points at -1, 0, and 1, each assigned equal weight — one third. We aim to verify whether this design satisfies the condition for D-optimality.

To do this, we start by computing the information matrix for this design. This matrix is obtained as the weighted sum of the outer products of the regression vectors at each design point. The resulting matrix is symmetric, with off-diagonal elements reflecting the mixed terms.

Next, we compute the inverse of this matrix, which we denote as $\mathrm{D}(\xi^*)$. This inverse will be used in the next steps, where we evaluate the sensitivity function to test the equivalence condition.

Checking the equivalence condition

D-optimality criterion

To verify D-optimality of ξ^*, check:

$$d(x, \xi^*) = f^T(x)\, D(\xi^*)\, f(x) \le m = 3, \quad \forall x \in [-1, 1]$$

Evaluation of the sensitivity function

$$d(x, \xi^*) = \begin{pmatrix} 1 & x & x^2 \end{pmatrix} \begin{pmatrix} 3 & 0 & -3 \\ 0 & 1.5 & 0 \\ -3 & 0 & 4.5 \end{pmatrix} \begin{pmatrix} 1 \\ x \\ x^2 \end{pmatrix} = 3 + \frac{9}{2} x^2 (x^2 - 1)$$

- The term $\frac{9}{2} x^2 (x^2 - 1)$ is nonpositive for $x \in [-1, 1]$
- Therefore, $d(x, \xi^*) \le 3$ for all x in $[-1, 1]$
- The equality is attained at $x = -1,\ 0,\ 1$
- By the K–W Theorem, ξ^* is D-optimal

Remark

For discrete designs, the equivalence between D-optimality and G-optimality does not generally hold.

Comments

To check D-optimality of the design, we apply the equivalence theorem. According to it, we need to verify that the sensitivity function — that is, the transpose of the regression vector multiplied by the inverse information matrix and then by the regression vector again — does not exceed the number of parameters in the model. In our case, the number of parameters is three.

We compute this function explicitly. The result is a constant term three plus an additional term equal to $\frac{9}{2}$ times x^2 times the difference between x^2 and one. Since this second term is nonpositive over the interval from -1 to 1, the whole expression is maximized when x is -1, 0, or 1. At each of these points, the expression equals three.

Therefore, the function never exceeds three, and the equivalence condition is satisfied. This confirms that the design we are testing is indeed D-optimal.

Finally, a brief remark: for discrete designs, the equivalence between D-optimality and G-optimality does not generally hold. That is, a D-optimal design may fail to be G-optimal if it is discrete.

Example: a discrete design

Model setup

- Regression vector: $\mathrm{f}(\mathrm{x}) = (1\ \mathrm{x})^{\mathrm{T}}$
- Design space: $\chi = [-1, 1]$
- Number of trials: $\mathrm{N} = 3$

Candidate design structure

$$\xi^* = \begin{pmatrix} \mathrm{x}_1 & \mathrm{x}_2 & \mathrm{x}_3 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}$$

- At least two distinct support points to avoid degeneracy (by Theorem 7)

Information matrix

$$\mathrm{M}(\xi^*) = \sum_{\mathrm{i}=1}^{3} \frac{1}{3} \mathrm{f}(\mathrm{x}_\mathrm{i}) \mathrm{f}^{\mathrm{T}}(\mathrm{x}_\mathrm{i}) = \begin{pmatrix} 1 & \frac{\mathrm{x}_1+\mathrm{x}_2+\mathrm{x}_3}{3} \\ \frac{\mathrm{x}_1+\mathrm{x}_2+\mathrm{x}_3}{3} & \frac{\mathrm{x}_1^2+\mathrm{x}_2^2+\mathrm{x}_3^2}{3} \end{pmatrix}$$

Comments

In this example, we work with a simple linear model where the regression vector consists of one and x. The design space is the interval from -1 to 1. We are allowed to perform three measurements, and therefore we look for a D-optimal design with three equally weighted support points.

To avoid degeneracy of the information matrix, we assume that at least two of the support points are distinct. This requirement comes from the earlier theorem on properties of the information matrix, which states that when the number of a design support points is fewer than the number of parameters, its information matrix becomes rank-deficient.

Let us compute the information matrix corresponding to this design. It takes a compact symmetric form with the average of the support points appearing in the off-diagonal entries, and the average of the squared points appearing in the lower-right element. This matrix characterizes the total information collected under this design and will be the basis for our optimization.

Example (continued): a discrete design

Determinant of the information matrix

$$\det(\mathrm{M}(\xi^*)) = \frac{2}{9}(x_1^2 + x_2^2 + x_3^2 - x_1x_2 - x_2x_3 - x_1x_3) =$$
$$= \frac{1}{9}\left[(x_1 - x_2)^2 + (x_1 - x_3)^2 + (x_2 - x_3)^2\right]$$

Optimization over design points

- Without loss of generality: $x_1 \le x_2 \le x_3$
- Maximum over x_1 attained at $x_1 = -1$; over x_3 — at $x_3 = 1$
- Resulting quadratic in x_2 achieves maximum at endpoints: $x_2 = -1$ or $x_2 = 1$

D-optimal designs

$$\xi_1^* = \begin{pmatrix} -1 & 1 \\ \frac{2}{3} & \frac{1}{3} \end{pmatrix}, \quad \xi_2^* = \begin{pmatrix} -1 & 1 \\ \frac{1}{3} & \frac{2}{3} \end{pmatrix}, \quad \det(\mathrm{M}) = \frac{8}{9}$$

Comments

To identify the optimal design, we maximize the determinant of the information matrix. The expression for the determinant can be transformed into a sum of squared differences between the design points. This shows that the determinant grows when the design points are spread farther apart.

Assuming the design points are ordered from smallest to largest, we analyze the derivative of the determinant with respect to the outer points. It turns out the maximum is reached when the first point is -1 and the last point is 1.

Substituting these values reduces the optimization to a function of the middle point. Since this function is quadratic and concave, the maximum occurs at either endpoint, which means the second point must also be either -1 or 1.

This gives us two distinct but equivalent D-optimal designs, each supported at the endpoints of the interval but with different weight distributions. The determinant of the information matrix is the same for both designs and equals $\frac{8}{9}$.

Example (final): a discrete design

D-optimal designs: matrices and d-functions

$$M(\xi_1^*) = \begin{pmatrix} 1 & -\frac{1}{3} \\ -\frac{1}{3} & 1 \end{pmatrix}, \quad D(\xi_1^*) = \begin{pmatrix} \frac{9}{8} & \frac{3}{8} \\ \frac{3}{8} & \frac{9}{8} \end{pmatrix}$$

$$d(x, \xi_1^*) = f^T(x) D(\xi_1^*) f(x) = \frac{1}{8}(3x+1)^2 + 1$$

$$\max_{x \in [-1,1]} d(x, \xi_1^*) = 3, \quad \max_{x \in [-1,1]} d(x, \xi_2^*) = \frac{1}{8}(3x-1)^2 + 1 = 3$$

Comparison with a uniform plan

$$\bar{\xi} = \begin{pmatrix} -1 & 0 & 1 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}, \quad M(\bar{\xi}) = \begin{pmatrix} 1 & 0 \\ 0 & \frac{2}{3} \end{pmatrix}, \quad D(\bar{\xi}) = \begin{pmatrix} 1 & 0 \\ 0 & \frac{3}{2} \end{pmatrix}$$

$$d(x, \bar{\xi}) = 1 + \frac{3}{2}x^2, \quad \max_{x \in [-1,1]} d(x, \bar{\xi}) = 2.5 < 3$$

Thus the discrete D-optimal design ξ_1^* and ξ_2^* are not G-optimal.

Comments

To analyze whether the D-optimal designs are also G-optimal, we compute the information matrix and its inverse for one of the D-optimal plans. Then we compute the sensitivity function d, which is the quadratic form involving the inverse information matrix and the regression vector. For the first D-optimal design, this function is a quadratic polynomial in the design variable, and its maximum over the interval equals three. The same maximum is obtained for the second D-optimal design.

We then compare these values with those of an alternative, uniformly weighted design supported at three equally spaced points. In this case, the information matrix and its inverse yield a simpler d-function, which is also quadratic. However, the maximum value of this function is only two and a half, strictly less than three.

This shows that although the D-optimal designs provide maximum information in terms of determinant, they do not minimize the maximum prediction variance across the design space. Hence, they are not G-optimal.

Equivalence theorem for the E-optimality criterion

- There are many versions of the Kiefer–Wolfowitz theorem for various optimality criteria.
- They are usually called equivalence theorems.
- Such theorems reduce the extremum problem to verifying a specific conditions for a function, simplifying the task and enabling design optimality checking.

Equivalence theorem for the E-criterion

Let $\mathcal{A}$ be the class of all positive semidefinite matrices A with $\operatorname{tr} A = 1$.
A design $\xi^* \in \Xi$ is E-optimal if and only if there exists $A^* \in \mathcal{A}$ such that for all $x \in \chi$,

$$f^T(x)A^*f(x) \le \lambda_{\min}(M(\xi^*)).$$

Moreover,

$$\min_{A\in\mathcal{A}} \max_{x\in\chi} f^T(x)Af(x) = \max_{\xi} \lambda_{\min}(M(\xi)),$$

$$f^T(x_i^*)A^*f(x_i^*) = \lambda_{\min}(M(\xi^*)),$$

where $x_i^*, i = 1, \dots, n$ are the support points of the E-optimal design.

This result is given in Melas V.B., *E-optimal experimental designs*, SPbU, 1997.

Comments

There exists a general family of results known as equivalence theorems, which play a central role in the theory of optimal experimental designs. These theorems, stemming from the work of Kiefer and Wolfowitz, establish necessary and sufficient conditions for a design to be optimal under a given criterion. Rather than directly maximizing or minimizing the criterion, they reformulate the task as checking for certain conditions related to a specific function, which are called "extremal polynomials" in the literature on experimental design. This approach not only simplifies the optimization problem but also provides a powerful way to check whether a candidate design is indeed optimal.

We will consider several versions of the equivalence theorem for different optimality criteria, but without proofs. The logic and structure of the proof is similar in all cases; only technical details differ. We start with the E-optimality criterion, which minimizes the length of the longest axis of the confidence ellipsoid for the estimated parameters — equivalently, it maximizes the smallest eigenvalue of the information matrix.

The theorem states that an E-optimal design exists if and only if there is a matrix with unit trace, which majorizes all values of the corresponding quadratic form over the design space, and matches the minimum eigenvalue at the support points.

Setup for L-optimality criterion

Let $L = \sum_{i=1}^{k} l_i l_i^T$, where $l_i \in \mathbb{R}^m$. Define the class Ξ_L as the set of all designs for which the linear combinations $l_i^T \theta$, $i = 1, \ldots, k$, are estimable.

Definition: class Ξ_L^*

A approximate design η belongs to the class Ξ_L^* if $\eta \in \Xi_L$ and for any approximate design ξ the limit

$$\lim_{\alpha \to 0} f^T(t) M^+(\xi_\alpha) L M^+(\xi_\alpha) f(t) = f^T(t) M^+(\eta) L M^+(\eta) f(t)$$

exists, where $\xi_\alpha = (1 - \alpha)\eta + \alpha\xi$, $\alpha \in [0, 1]$, and $M^+(\xi_\alpha)$ is the Moore–Penrose pseudoinverse of $M(\xi_\alpha)$.

Definition: L-optimal design

An approximate design $\xi^* \in \Xi_L^*$ is L-optimal if

$$\xi^* = \arg \min_{\xi \in \Xi_L^*} \operatorname{tr}(L M^+(\xi))$$

where L is a fixed nonnegative definite matrix. If ξ^* is nonsingular, then $M^+(\xi^*) = M^{-1}(\xi^*)$.

Comments

We now introduce the definitions necessary to formulate the equivalence theorem for the L-optimality criterion. First, we assume a symmetric nonnegative definite matrix L, defined as the sum of outer products of given vectors — that is, L equals the sum from $i = 1$ to k of the vector l_i times its transpose. The set Ξ_L consists of all continuous experimental designs for which the linear combinations of the parameters, namely $l_i^T \theta$, are estimable for all i.

Next, we define an extended class of designs, denoted Ξ_L^*. A design η belongs to this class if it is in Ξ_L and satisfies a specific continuity property: for any other design ξ, the limit as α goes to zero of the function $f^T(t) M^+(\xi_\alpha) L M^+(\xi_\alpha) f(t)$ exists and equals the corresponding expression evaluated at η. Here, the pseudoinverse is understood in the Moore–Penrose sense.

Finally, we define an approximate L-optimal design as one that minimizes the trace of L times the pseudoinverse of the information matrix over the class Ξ_L^*. If the design is nondegenerate, then the pseudoinverse coincides with the usual inverse of the information matrix.

Equivalence theorem for L-optimality (nonsingular case)

We will separately consider the nonsingular and singular cases.

Comments

- Assume the existence of a nonsingular L-optimal design ξ^*
- If $\operatorname{rank} L = m$, then all L-optimal designs are nonsingular

Equivalence theorem for L-optimality

The following statements are equivalent:

(a) $\xi^* = \arg\min_{\xi \in \Xi_L} \operatorname{tr} LD(\xi)$

(b) $\max_{x \in \Xi_L} q(x, \xi^*) = \operatorname{tr} LD(\xi^*)$, where $q(x, \xi) = f^\top(x) D(\xi) LD(\xi) f(x)$.

Moreover, equality holds for all support points x_i of ξ^*:

$$q(x_i, \xi^*) = \operatorname{tr} LD(\xi^*)$$

This result can be found in Ermakov S.M., Zhiglyavsky A.A., Mathematical Theory of Optimal Experiment, Nauka, 1987. (Theorem 2.4, p. 112).

Comments

This theorem is a version of the equivalence theorem tailored for the L-optimality criterion, under the assumption that the optimal design is nonsingular. The assumption is reasonable whenever the rank of L equals m, the number of parameters in the model, since in that case all optimal designs are automatically nonsingular.

The equivalence theorem states that two conditions are equivalent: first, that the design minimizes the trace of L times the inverse of the information matrix, and second, that the maximum value of the function q, defined as the transpose of f times the inverse information matrix times L times the inverse again times f, over all design points equals that trace. Furthermore, this maximum is attained exactly at the support points of the optimal design.

It is important to highlight a major difference from the D-optimality criterion. In the D-optimal case, we always require nonsingular designs, since all parameters must be estimable. In contrast, for L-optimality, the focus is on estimating only certain linear combinations of the parameters. Therefore, we may allow singular designs, where fewer measurements than parameters are used. This flexibility makes the theory more subtle. That is why we will treat the nonsingular and singular cases separately: the singular case requires a more advanced treatment and leads to complications not present in the nonsingular setting.

Equivalence theorem for L-optimality (singular case)

Model setup

- Let $L = \sum_{i=1}^{k} l_i l_i^T$ with fixed $l_i \in \mathbb{R}^m$
- Assume an optimal design $\xi^* \in \Xi_L^*$ exists

Equivalence theorem for L-optimality (singular case)

For model (1o):

1) A design ξ belongs to Ξ_L if and only if for all $i = 1, \ldots, k$:

$$l_i^T M^-(\xi) M(\xi) = l_i^T$$

2) A design $\xi^* \in \Xi_L^*$ is L-optimal if and only if:

(a) $\max_{t \in \mathcal{X}} \varphi(t, \xi^*) = \mathrm{tr}(LM^+(\xi^*)), \quad \varphi(t, \xi) = f^T(t) M^+(\xi) L M^+(\xi) f(t)$

(b) $\varphi(t_i, \xi^*) = \mathrm{tr}(LM^+(\xi^*)), \quad t_i \in \mathrm{supp}(\xi^*)$

This and the following result can be found in Shpilev P.V. *Equivalence Theorem for Singular L-Optimal Designs*, Vestnik St. Petersburg University. Mathematics, Vol. 48, No. 1, pp. 29–34, (2015).

Comments

We now turn to the equivalence theorem for the singular case. This is a more general and technically more complex situation, in which the information matrix may be singular. Such cases naturally arise when we are interested not in estimating all parameters, but only certain linear combinations — for instance, specific contrasts in regression.

We start with the structure of L, which is assumed to be the sum of outer products of fixed vectors l_i. These vectors specify the combinations of parameters that we care about. Under this structure, a design belongs to the class Ξ_L if and only if the identity $l_i^T M^-(\xi) M(\xi) = l_i^T$ holds for all i. This ensures that the design provides enough information to estimate the desired linear combinations.

The second part of the theorem characterizes L-optimal designs within the narrowed class Ξ_L^*. Such a design is optimal if and only if the maximum value of the function φ, which is the transpose of f times the pseudoinverse of M times L times the same pseudoinverse again times f, equals the trace of L times the pseudoinverse of the information matrix. Moreover, this maximum must be attained at all support points of the design.

This formulation generalizes the equivalence result to situations where the information matrix may be singular, and only certain linear estimators are required. It allows for designs with fewer support points than parameters, which makes the class of admissible designs significantly richer.

Notation for perturbed designs

Let $\xi \in \Xi_L$ be a singular design and let $\xi_\alpha = \alpha\eta + (1-\alpha)\xi$ with any design η such that ξ_α is nonsingular. Define:

$$\tilde{D}_{ij}(\xi) := \begin{cases} y_{ij}, & \text{if } 0 < |\lim_{\alpha\to 0} M_{ij}^{-1}(\xi_\alpha)| < \infty \\ 0, & \text{otherwise} \end{cases}$$

$$\tilde{D}_{ij}^{+}(\xi) := \begin{cases} M_{ij}^{+}(\xi), & \text{if } \lim_{\alpha\to 0} M_{ij}^{-1}(\xi_\alpha) = M_{ij}^{+}(\xi) \\ \tilde{D}_{ij}(\xi), & \text{otherwise} \end{cases}$$

Matrix $\tilde{D}^{+}(\xi)$ is symmetric, y_{ij} is unknow variables.

Theorem 8(Extremal polynomial in singular case)

Let $L \in \mathbb{R}^{m\times m}$ be fixed and positive semidefinite, and let $\xi^* \in \Xi_L \setminus \Xi_L^*$. Then the coefficients of the extremal polynomial $\overline{\varphi}(t,\xi^*) = f^T(t)\tilde{D}^{+}(\xi^*)L\tilde{D}^{+}(\xi^*)f(t)$ are the solution to the system:

$$\begin{cases} \overline{\varphi}'(t_i,\xi^*) = 0 \\ \overline{\varphi}(t_i,\xi^*) = \mathrm{tr}(LM^{+}(\xi^*)) \end{cases} \quad \text{for } t_i \in \mathrm{supp}(\xi^*)$$

Comments

We now address the most technically challenging scenario: when the optimal design is singular and does not belong to Ξ_L^*. This means that for design ξ_α from definition some specific elements of the Moore–Penrose pseudoinverse of the information matrix — namely, those involved in computing the coefficients of the extremal polynomial — do not coincide with the limiting values of the corresponding elements in the Moore–Penrose pseudoinverse for optimal design. Or in other words, relevant elements of pseudoinverse matrix are not uniquely determined by the limiting process.

To handle this ambiguity, we define a matrix with entries denoted by $\tilde{D}_{ij}(\xi)$. Each such entry equals y_{ij} — a symbolic unknown — if the absolute value of the limit of the ij element of the inverse of ξ_α exists, is nonzero, and finite, where ξ_α is defined as $\alpha\eta + (1-\alpha)\xi$, with α between zero and one, and η any design for which ξ_α is nonsingular. Otherwise, the corresponding entry is set to zero. The symbols y_{ij} are symmetric by construction, meaning y_{ij} equals y_{ji}.

We then define another matrix, $\tilde{D}^{+}(\xi)$, which selects either the corresponding element of the Moore–Penrose pseudoinverse or falls back to $\tilde{D}$ depending on whether the limiting value exists and matches. With this definition, we construct a generalized extremal polynomial — denoted by $\overline{\varphi}(t)$ — as the quadratic form $f^T(t)\tilde{D}^{+}(\xi^*)L\tilde{D}^{+}(\xi^*)f(t)$. The coefficients of this polynomial are to be determined as the solution of a system of equations evaluated at the support points of ξ^*.

Remark

Let $\xi^* \in \Xi_L \setminus \Xi_L^*$. Then ξ^* is L-optimal if and only if the extremal polynomial $\overline{\varphi}(t, \xi^*)$ (as defined in Theorem 8) satisfies:

(a) $\max_{t \in \mathcal{X}} \overline{\varphi}(t, \xi^*) = \mathrm{tr}(LM^+(\xi^*))$

(b) $\overline{\varphi}(t_i, \xi^*) = \mathrm{tr}(LM^+(\xi^*)), \quad t_i \in \mathrm{supp}(\xi^*)$

The trigonometric regression model

Let us consider the third-order trigonometric regression model on $[-\pi, \pi]$:

$$y = \beta^T f(t) = \beta_0 + \beta_1 \sin(t) + \beta_2 \cos(t) + \beta_3 \sin(2t) + \beta_4 \cos(2t) + \\ + \beta_5 \sin(3t) + \beta_6 \cos(3t) + \varepsilon$$

Example: Singular design optimal for estimating the linear combination $\beta_2 + \beta_3$

The candidate to optimal design

Let us consider a symmetric design: $\xi_\alpha = \begin{pmatrix} -\pi + x & -x & 0 & x & \pi - x \\ \frac{1-\alpha}{4} & \frac{1-\alpha}{4} & \alpha & \frac{1-\alpha}{4} & \frac{1-\alpha}{4} \end{pmatrix}, \quad \alpha \in [0, 1]$

Comments

In the singular case, when the design ξ^* belongs to the set Ξ_L but not to Ξ_L^*, the equivalence theorem remains applicable via the generalized extremal polynomial from Theorem 8. In this situation, the extremal polynomial defined using the matrix $\tilde{D}^+(\xi^*)$ must satisfy the same two conditions that appear in the equivalence theorem for the nonsingular case: its maximum over the design space must be equal to the trace of L times M^+, and it must attain this maximum at all support points of the design.

Below the remark, we begin an example that illustrates how this works in practice. We consider a third-order trigonometric regression model on the interval from $-\pi$ to π. The regression function includes the constant term, the sine and cosine of t, the sine and cosine of 2t, and the sine and cosine of 3t, making a total of seven parameters.

We consider an optimal design within a class of symmetric designs dependent on parameter α. The design's support points are symmetric about zero, with the central point assigned weight α, while the two outer pairs equally share the remaining weight. Since this design comprises only five support points, whereas the regression model has seven parameters, the information matrix remains singular for any value of α.

Structure of the information matrix

From the definition of the information matrix, it follows that for the design ξ_α we have:

$$M(\xi_\alpha) = \begin{pmatrix} 1 & 0 & 0 & 0 & m_{0,4} & 0 & 0 \\ 0 & m_{1,1} & 0 & 0 & 0 & m_{1,5} & 0 \\ 0 & 0 & m_{2,2} & 0 & 0 & 0 & m_{2,6} \\ 0 & 0 & 0 & m_{3,3} & 0 & 0 & 0 \\ m_{0,4} & 0 & 0 & 0 & m_{4,4} & 0 & 0 \\ 0 & m_{1,5} & 0 & 0 & 0 & m_{5,5} & 0 \\ 0 & 0 & m_{2,6} & 0 & 0 & 0 & m_{6,6} \end{pmatrix}$$

where the nonzero entries are given by:

$$m_{i,j} = \sum_{k=1}^{5} f_i(t_k) f_j(t_k) \omega_k$$

for t_k, ω_k being the support points and weights of the design ξ_α.

Comments

From the definition of the information matrix, we observe that due to the symmetry of the design, many cross-products of basis functions vanish when integrated with respect to the design measure. In particular, the basis function corresponding to the constant term interacts only with itself and with the second cosine term, that is, cosine of double argument. This is a direct consequence of the orthogonality relations between the trigonometric basis functions over symmetric support.

The resulting information matrix has a block-sparse structure, in which nonzero elements appear only for specific pairs of basis functions. The positions of the nonzero elements reflect which function pairs have nonzero scalar products on the design support. These scalar products are explicitly given by a weighted sum over the support points of the design: the nonzero entry in position i, j is the sum of the product of the ith and jth basis functions evaluated at those points, each weighted by the corresponding design weight.

The sparsity and structure of this matrix reflect the properties of the design and play a crucial role in computing its generalized inverse and the associated extremal polynomial.

Optimality criterion

To estimate β_2 and β_3, the optimal design minimizes $\operatorname{tr} \mathrm{LM}^{+}(\xi)$

$$\text{with} \quad \mathrm{L} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Block-diagonalization

The information matrix $\mathrm{M}(\xi_\alpha)$ can be block-diagonalized using a nonsingular symmetric matrix P:

$$\mathrm{PM}(\xi_\alpha)\mathrm{P} = \begin{pmatrix} \bar{\mathrm{M}}(\xi_\alpha) & 0 \\ 0 & \bar{\mathrm{M}}_1(\xi_\alpha) \end{pmatrix}, \quad \bar{\mathrm{M}}(\xi_\alpha) = \begin{pmatrix} \mathrm{m}_{2,2} & 0 & \mathrm{m}_{2,6} \\ 0 & \mathrm{m}_{3,3} & 0 \\ \mathrm{m}_{2,6} & 0 & \mathrm{m}_{6,6} \end{pmatrix}$$

Comments

The goal is to estimate the parameters β_2 and β_3. According to the general theory, the optimal design for this purpose minimizes the trace of the product of a selection matrix L and the Moore–Penrose inverse of the information matrix. Matrix L is constructed to extract the sum of variances of the least-squares estimators for β_2 and β_3, which correspond to the third and fourth diagonal elements of the pseudoinverse.

The specific structure of the information matrix corresponding to the design ξ_α makes it possible to construct a nonsingular symmetric matrix P that transforms the information matrix into a block-diagonal form. In this representation, the upper-left block is a three-by-three matrix that includes the elements corresponding to the third, fourth, and seventh columns and rows of the original matrix. These entries relate to the basis functions sine of t, sine of 2t, and cosine of 3t. These basis functions are precisely the ones involved in the estimation of β_2 and β_3, as well as any interactions between them and other components.

The remaining part of the information matrix, after transformation, corresponds to the components of the model that are not involved in the estimation of β_2 and β_3. As a result, they do not contribute to the trace of the selection matrix L times the pseudoinverse.

This block-diagonalization allows us to reduce the dimensionality of the optimization problem by focusing only on the relevant submatrix.

Equivalence theorem test for the design ξ^*

- For the symmetric design ξ_α, the optimization problem $\min\limits_{\xi} \operatorname{tr}(\mathrm{L}\mathrm{M}^{+}(\xi))$

$$\text{reduces to the problem } \min_{\xi} \operatorname{tr}(\bar{\mathrm{L}}\bar{\mathrm{M}}^{+}(\xi_\alpha)),\ \ \bar{\mathrm{L}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

- Test the design

$$\xi^* = \begin{pmatrix} -\frac{5\pi}{6} & -\frac{\pi}{6} & \frac{\pi}{6} & \frac{5\pi}{6} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{pmatrix}$$

for optimality using the equivalence theorem.

- Let $\xi_\alpha^* = (1-\alpha)\xi^* + \alpha\eta$ where $\eta = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$. Then

$$\lim_{\alpha\to 0} \mathrm{f}^{\mathrm{T}}(\mathrm{t})\bar{\mathrm{M}}^{+}(\xi_\alpha^*)\bar{\mathrm{L}}\bar{\mathrm{M}}^{+}(\xi_\alpha^*)\mathrm{f}(\mathrm{t}) = \frac{16}{9}\cos^2(\mathrm{t}) - \frac{32}{9}\cos(\mathrm{t})\cos(3\mathrm{t}) +$$

$$+\frac{16}{9}\sin^2(2\mathrm{t}) + \frac{16}{9}\cos^2(3\mathrm{t}) \neq \mathrm{f}^{\mathrm{T}}(\mathrm{t})\bar{\mathrm{M}}^{+}(\xi^*)\bar{\mathrm{L}}\bar{\mathrm{M}}^{+}(\xi^*)\mathrm{f}(\mathrm{t}) = \frac{16}{9}\cos^2(\mathrm{t}) + \frac{16}{9}\sin^2(2\mathrm{t})$$

- Therefore, $\xi^* \notin \Xi_{\mathrm{L}}^*$ $\Rightarrow$ we cannot apply the equivalence theorem to check its optimality directly.

Comments

Having reduced the criterion to a trace expression involving only the relevant block of the information matrix, we now test whether a given symmetric design is optimal. The candidate design ξ^* is supported at four symmetric points with equal weights. To verify the optimality of the candidate design ξ^*, we construct a perturbed design ξ_α by mixing it with a point mass at zero. The aim of this construction is to ensure that the reduced information matrix remains nonsingular. This allows us to replace the Moore–Penrose pseudoinverse with the usual inverse and compute the limiting form of the extremal polynomial uniquely.

A comparison of the limiting extremal polynomial with the expression obtained by directly substituting ξ^* into the pseudoinverse matrix formula shows that the two results do not coincide. This discrepancy means that the design ξ^* does not belong to the class Ξ_{L}^*. As a result, we cannot apply the equivalence theorem to assess its optimality directly. We need to use theorem 8 first.

Matrices involved

For the candidate design ξ^*, the matrices $\tilde{D}^+(\xi^*)$ and $\bar{M}^+(\xi^*)$ are:

$$\tilde{D}^+(\xi^*) = \begin{pmatrix} \frac{4}{3} & 0 & y_{31} \\ 0 & \frac{4}{3} & 0 \\ y_{31} & 0 & 0 \end{pmatrix}, \quad \bar{M}^+(\xi^*) = \begin{pmatrix} \frac{4}{3} & 0 & 0 \\ 0 & \frac{4}{3} & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Extremal polynomial

The extremal polynomial from Theorem 8: $\overline{\varphi}(t, \xi^*) = f^T(t)\tilde{D}^+(\xi^*)\bar{L}\tilde{D}^+(\xi^*)f(t)$

$$= \frac{16}{9}\cos^2(t) - \frac{8}{3}y_{31}\cos(t)\cos(3t) + \frac{16}{9}\sin^2(2t) + y_{31}^2\cos^2(3t)$$

- From the condition $\overline{\varphi}'(\frac{\pi}{6}, \xi^*) = 0$ we find $y_{31} = \frac{2}{9}$.
- Then $\overline{\varphi}(t)$ achieves its maximum value $\frac{8}{3}$ at all support points of ξ^*:

$$\max_{t\in[-\pi,\pi]} \overline{\varphi}(t, \xi^*) = \overline{\varphi}\left(\pm\frac{\pi}{6}, \xi^*\right) = \overline{\varphi}\left(\pm\frac{5\pi}{6}, \xi^*\right) = \mathrm{tr}(\bar{L}\,\bar{M}^+(\xi^*)) = \frac{8}{3}$$

- Hence, ξ^* satisfies conditions (a) and (b) of the equivalence theorem and is L-optimal.

Comments

Now let us examine how we can still apply the equivalence theorem in this case. For this purpose, we use Theorem 8 and construct the extremal polynomial $\overline{\varphi}$ using the general inverse matrix $\tilde{D}^+(\xi^*)$. This matrix differs from the standard pseudoinverse in that it may contain symbolic variables—in our case, a single unknown variable denoted y_{31}. Using the system from Theorem 8, we determine the value of this variable. We find that y_{31} equals $\frac{2}{9}$.

It remains to verify that the constructed polynomial satisfies conditions (a) and (b) of the equivalence theorem i.e. that the resulting polynomial attains its maximum at all support points of the design and that this maximum matches the trace of the matrix L multiplied by the Moore–Penrose pseudoinverse of the reduced information matrix. Both conditions are fulfilled: the extremal polynomial peaks at $\pm\frac{\pi}{6}$ and $\pm\frac{5\pi}{6}$, and the maximum value equals $\frac{8}{3}$. Thus, the equivalence theorem confirms that this design is indeed L-optimal—even though its information matrix is singular.

Therefore, the design ξ^* is L-optimal.

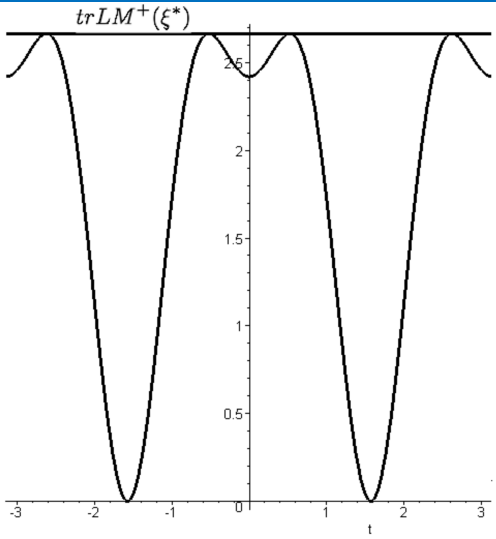

Figure: The behavior of the extremal polynomial $\overline{\varphi}(t, \xi^*)$ on the interval $[-\pi, \pi]$.

Comments

Here we visualize the extremal polynomial for the optimal design ξ^*. The horizontal axis represents the design space — in this case, the interval from $-\pi$ to π. The vertical axis shows the value of the extremal polynomial over this interval.

Now, according to the general equivalence theorem, a design is optimal if and only if the extremal polynomial achieves its maximum exactly at the support points of the design. In our case, the support points are $\pm\frac{\pi}{6}$ and $\pm\frac{5\pi}{6}$. We can immediately spot these four points on the graph — they are precisely where the curve touches its highest level.

In practical terms, this plot provides visual confirmation that the design ξ^* is indeed optimal under the L-criterion, even though its information matrix is singular. This makes the extremal polynomial a powerful diagnostic tool — it translates an abstract matrix condition into a concrete and interpretable shape on the real line.

Equivalence theorem for the e_k-criterion

Let $\bar{f}_k(t)$ denote the vector obtained by removing $f_k(t)$ from $f(t) = (f_1(t), \ldots, f_m(t))^T$, for some $k = 1, \ldots, m$. Then a design ξ^* is optimal for estimating θ_k in model (1o) under assumptions (a)–(e) if and only if there exist $h > 0$ and $q \in \mathbb{R}^{m-1}$ such that the extremal polynomial

$$\varphi(t) = f_k(t) - q^T \bar{f}_k(t)$$

satisfies:

1) $h\varphi^2(t) \le 1$ for all $t \in \chi$;
2) $\operatorname{supp}(\xi^*) \subset \{t \in \chi \mid h\varphi^2(t) = 1\}$;
3) $\int_\chi \varphi(t)\bar{f}_k(t)\,\xi^*(dt) = 0 \in \mathbb{R}^{m-1}$.

In this case, $h = e_k^T M^-(\xi^*) e_k$.

Source: Dette, Melas, Pepelyshev (2004), *Optimal designs for estimating individual coefficients in polynomial regression – a functional approach*. JSP&I, 118, 201–219.

Comments

We now present the equivalence theorem for the so-called e_k-criterion. This criterion focuses on the precision of estimating a single component of the parameter vector — namely, the k-th coordinate of θ. Instead of optimizing the estimation of all parameters together, as in most common design criteria, this approach focuses on just one parameter and tries to estimate it as precisely as possible.

The theorem provides necessary and sufficient conditions for a design ξ^* to be optimal for estimating θ_k. It states that such a design exists if and only if we can construct a function $\varphi(t)$, which is a linear combination of the components of $f(t)$, such that three conditions are met.

First, the function $h\varphi^2(t)$ must stay below or equal to one throughout the entire design space. Second, the support of the design must be concentrated exactly where this function reaches its maximum value of one. And third, a certain orthogonality condition must be satisfied — namely, the weighted average of $\varphi(t)$ times $\bar{f}_k(t)$ over the design must be zero.

The scalar h that appears in this condition turns out to be exactly the k-th diagonal entry of the generalized inverse information matrix. That is, h equals $e_k^T M^- e_k$. The function φ is called the extremal polynomial and plays the same certifying role as in the classical case.

This theorem offers a constructive way to verify optimality and to visualize it using extremal polynomials — especially in polynomial regression, where this form becomes explicit.

Comments

- In many classical models, optimal designs can be found analytically.
- We will now look at how to construct a D-optimal design for two widely used models:
  - Polynomial regression
  - Trigonometric regression

Polynomial regression model:

$$y_j = \theta^T f(x_j) + \varepsilon_j, \quad j = 1, \dots, N, \qquad f(x) = (1, x, \dots, x^{m-1})^T, \quad \chi = [-1, 1]$$

Theorem 9

For polynomial model on $[-1, 1]$, an unique approximate D-optimal design exists.

- It is supported with equal weights on m points.
- These points are the roots of the polynomial $(x^2 - 1)P'_{m-1}(x)$, where $P_{m-1}(x)$ is the Legendre polynomial of degree $m - 1$.

Comments

In many practical applications, equivalence theorems play a crucial role in simplifying the search for optimal designs. Instead of exploring the entire space of possible approximate designs, a more efficient approach is to focus on a well-structured subclass. For example, in the case we examined earlier, we restricted our search to 5-point symmetric designs as potential candidates. This constraint dramatically reduced the problem's complexity and made the optimization process more manageable. By confirming that the resulting design meets the equivalence conditions, we can be certain that it represents a global optimum—not just a local one.

Equivalence theorems often enable analytical derivation of optimal designs. As an illustration, consider D-optimal design construction for two commonly used models: polynomial and trigonometric regression. Let's begin with the polynomial case.

In this model, the regression functions consist of monomials: one, x, x^2, and so on, up to x^{m-1}. Under standard error assumptions, the D-optimal design on the interval is unique and has several remarkable properties. First, it is an approximate design supported at exactly m points. Second, these support points coincide with the roots of the polynomial $(x^2 - 1)P'_{m-1}(x)$ where $P_{m-1}(x)$ is the Legendre polynomial of degree $m - 1$. Finally, all m points receive equal weights.

This elegant result provides researchers with a powerful analytical tool for constructing optimal designs in polynomial regression settings.

Proof: D-optimal design for polynomial regression

Proof:

- Existence: Follows from the continuity of regressors and the compactness of the set of information matrices.
- Kiefer-Wolfowitz Theorem: We use this theorem to prove the remaining statements about the design.

Kiefer–Wolfowitz condition: the function $d(x, \xi^*)$ must satisfy:

$$\max_{x \in [-1,1]} d(x, \xi^*) = f^T(x) M^{-}1(\xi^*) f(x) = \sum_{i,j=1}^{m} d_{ij}(\xi^*) x^{i+j-2} = m$$

- $d(x, \xi^*)$ is a polynomial of degree $2m - 2$.
- The maxima occur at the support points of the optimal design.
- The design must have exactly m support points to avoid a singular information matrix.

The candidate for D-optimal design has the form:

$$\xi^* = \begin{pmatrix} x_1 & x_2 & \dots & x_m \\ w_1 & w_2 & \dots & w_m \end{pmatrix}$$

where $-1 = x_1 < x_2 < \dots < x_m = 1$ and $\sum_i w_i = 1$, $w_i > 0$ for all i.

Comments

Let's prove this theorem. The existence of a D-optimal design is guaranteed by the continuity of the regressors and the compactness of the design space. To characterize the structure of the optimal design, we apply the Kiefer–Wolfowitz equivalence theorem.

According to this theorem, a design ξ^* is D-optimal if and only if the maximum of the extremal polynomial $d(x, \xi^*)$, which is also known as a sensitivity function in the literature, does not exceed m, the number of model parameters, and this maximum is attained exactly at the support points of the design.

Due to the fact that the vector of regression functions $f(x)$ has the form: one, x, x^2, and so on, up to x^{m-1}, and the inverse information matrix for a given design ξ^* is simply a numerical matrix defining the coefficients of the polynomial d, the degree of this polynomial is equal to $2m - 2$. Since such a polynomial has at most m real extrema, the number of support points of the design cannot exceed m. At the same time, having fewer than m points leads to a singular information matrix, which is not allowed.

Hence, the D-optimal design must have exactly m support points, located strictly inside the interval and at its boundaries. Thus the candidate for the D-optimal design has the following form: It consists of m support points, x_1 through x_m, and their corresponding weights, w_1 through w_m. The points are ordered from -1 to 1, and all the weights are positive and sum up to one.

The information matrix of design ξ^* can be written as:

$$M(\xi^*) = \sum_{i=1}^{m} f(x_i) f^T(x_i) w_i$$

$$= \underbrace{\begin{pmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_m \\ \vdots & & \vdots \\ x_1^{m-1} & \dots & x_m^{m-1} \end{pmatrix}}_{F} \underbrace{\begin{pmatrix} w_1 & 0 & \dots & 0 \\ 0 & w_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & w_m \end{pmatrix}}_{W} \underbrace{\begin{pmatrix} 1 & x_1 & \dots & x_1^{m-1} \\ 1 & x_2 & \dots & x_2^{m-1} \\ \vdots & \vdots & & \vdots \\ 1 & x_m & \dots & x_m^{m-1} \end{pmatrix}}_{F^T}$$

From this factorization, we obtain:

$$\det(M(\xi^*)) = \det(FWF^T) = \left[\det(F^T)\right]^2 \cdot \prod_{i=1}^{m} w_i$$

Comments

By the definition, the information matrix $M(\xi^*)$ is equal to the sum over i from one to m of $f(x_i)f^T(x_i)w_i$.

We can write this as a matrix product. First, we define the matrix F, whose columns consist of the vectors $f(x_1)$ through $f(x_m)$. Since f consists of monomials one, x, x^2, and so on, F is the transposed Vandermonde matrix. Then we define a diagonal matrix W whose diagonal entries are the weights w_1 through w_m. Finally, F^T is the transpose of the first matrix. So the matrix $M(\xi^*)$ equals FWF^T.

By applying this identity, we compute the determinant of $M(\xi^*)$. The determinant evaluates to the square of the determinant of F^T, multiplied by the product of the weights. This result reveals an important property of D-optimal designs: the optimization problem naturally decomposes into two independent subproblems. Specifically, we can separately maximize the determinant with respect to (1) the support points and (2) their corresponding weights.

Proof: D-optimal design for polynomial regression (continued)

- The determinant $\det(F^T)$ is called the Vandermonde determinant and is calculated recursively. Let's denote

$$\triangle_{m-1} = \det(F^T).$$

- Then

$$\triangle_{m-1} = \begin{vmatrix} 1 & 0 & \dots & 0 \\ 1 & x_2 - x_1 & \dots & x_2^{m-1} - x_2^{m-2}x_1 \\ \vdots & \vdots & & \vdots \\ 1 & x_m - x_1 & \dots & x_m^{m-1} - x_m^{m-2}x_1 \end{vmatrix} = \prod_{i=2}^{m}(x_i - x_1) \cdot \triangle_{m-2}.$$

- Thus,

$$\triangle_{m-1} = \prod_{1 \le i < j \le m} (x_i - x_j).$$

- We find the values of x_i that maximize the expression $\triangle_{m-1}$. To do this, we differentiate with respect to the variable x_i, where $i = 2, \dots, m-1$, and set the result to zero:

$$\frac{\partial \triangle_{m-1}}{\partial x_i} = 0 \quad \Leftrightarrow \quad \frac{1}{x_i - x_1} + \dots + \frac{1}{x_i - x_{i-1}} + \frac{1}{x_i - x_{i+1}} + \dots + \frac{1}{x_i - x_m} = 0.$$

Comments

Let us now examine the structure of the determinant of F^T. This determinant is called the Vandermonde determinant and is denoted by $\triangle_{m-1}$. That is, $\triangle_{m-1}$ is equal to the determinant of F^T.

Using elementary row operations, we can express this determinant recursively. Specifically, $\triangle_{m-1}$ is equal to the determinant of a matrix where the first row is one, zero, up to zero, and each subsequent row contains one, followed by $x_i - x_1$, then $x_i^{m-1} - x_i^{m-2}x_1$, and so on. This allows us to factor out the product over i from two to m of the differences $x_i - x_1$. What remains is a smaller determinant, namely $\triangle_{m-2}$.

By recursively applying this reduction, we eventually obtain the well-known closed formula for the Vandermonde determinant. That is, $\triangle_{m-1}$ is equal to the product over all pairs $i < j$ from one to m of $x_i - x_j$. This product is symmetric and strictly positive whenever the x-values are distinct.

To find the support points that maximize this expression, we differentiate with respect to each inner point x_i, for i ranging from two to $m-1$. Setting the derivative equal to zero gives a system of equations. Each equation takes the form: the sum over all $j \neq i$ of one divided by $x_i - x_j$ equals zero. This condition characterizes the so-called electrostatic equilibrium of the support points, and the solution to this system will lead us to the optimal support for the D-optimal design.

Differential equation for support polynomial

- Define the function:
$$\varphi(x) = (x - x_2)\dots(x - x_{m-1})$$
- The equilibrium conditions can be rewritten as:
$$\frac{1}{x_i + 1} + \frac{1}{x_i - 1} + \frac{\varphi''(x_i)}{2\varphi'(x_i)} = 0, \quad i = 2, \dots, m-1$$
- Equivalent form:
$$(x_i^2 - 1)\varphi''(x_i) + 4x_i\varphi'(x_i) = 0, \quad i = 2, \dots, m-1$$
- This expression has the same degree and roots as $\varphi(x)$, so:
$$(x^2 - 1)\varphi''(x) + 4x\varphi'(x) = \text{const} \cdot \varphi(x)$$
- Matching the coefficient at x^{m-2} yields:
$$\text{const} = m(m-1) - 2$$
- The differential equation:
$$(x^2 - 1)\varphi''(x) + 4x\varphi'(x) - (m(m-1) - 2)\varphi(x) = 0$$

Comments

In order to reduce this system to a differential equation we introduce a new function: $\varphi(x)$ equals the product over j from 2 to $m-1$ of $(x - x_j)$. This function captures the internal support points of the D-optimal design, excluding the boundary points -1 and 1.

Using this notation, we can rewrite the equilibrium conditions in a more compact form. For each i from 2 to $m-1$, the equation becomes: $\frac{1}{x_i+1} + \frac{1}{x_i-1} + \frac{1}{2}\frac{\varphi''(x_i)}{\varphi'(x_i)} = 0$.

This expression can be simplified algebraically to: $(x_i^2 - 1)\varphi''(x_i) + 4x_i\varphi'(x_i) = 0$.

This equation has the same degree and the same roots as $\varphi(x)$ itself. Therefore, the left-hand side must be equal to a constant times $\varphi(x)$. To determine this constant, we match the coefficients of x^{m-2} on both sides. This gives us that the constant is equal to $m(m-1) - 2$.

Thus, we obtain the desired differential equation for $\varphi(x)$: $(x^2-1)\varphi''(x) + 4x\varphi'(x) - (m(m-1) - 2)\varphi(x) = 0$.

Completion of the proof

- Consider the differential equation:

$$(x^2-1)P''(x)+2xP'(x)-m(m-1)P(x)=0,$$

where $P(x)$ is a polynomial of degree $m-1$.

- The unique solution to this is the Legendre polynomial of order $m-1$:

$$P_{m-1}(x)=\frac{1}{2^{m-1}(m-1)!}\frac{d^{m-1}}{dx^{m-1}}(x^2-1)^{m-1}$$

- Note that $P'_{m-1}(x)$ satisfies the derivative of the previous equation:

$$\left[(x^2-1)P''(x)+2xP'(x)-m(m-1)P(x)\right]'=0$$

$$\Leftrightarrow (x^2-1)P'''(x)+4xP''(x)-(m(m-1)-2)P'(x)=0$$

- Hence, $\overline{\varphi}(x)=P'_{m-1}(x)$ is a solution of the desired differential equation.
- All we have to do is find the weights. So we solve the system:

$$\begin{cases}\frac{\partial\prod\omega_i}{\partial\omega_i}=0, & i=1,\dots,m-1\\ \sum_{i=1}^{m}\omega_i=1\end{cases}\quad\Longrightarrow\quad \omega_1=\dots=\omega_m=\frac{1}{m}$$

Theorem is proved. □

Comments

We now consider the differential equation: $(x^2-1)P''(x)+2xP'(x)-m(m-1)P(x)=0$. Here, $P(x)$ is a polynomial of degree $m-1$.

According to the theory of orthogonal polynomials, the unique solution up to a multiplicative constant is the Legendre polynomial of order $m-1$. This polynomial is given by $\frac{1}{2^{m-1}(m-1)!}\frac{d^{m-1}}{dx^{m-1}}(x^2-1)^{m-1}$.

Differentiating the original differential equation, we get a new equation involving the third derivative of P. This leads to: $(x^2-1)P'''(x)+4xP''(x)-(m(m-1)-2)P'(x)=0$.

This implies that the derivative of the Legendre polynomial, P'_{m-1}, satisfies the differential equation we previously obtained for φ.

Finally, solving the system of equations for the weights shows that all ω_i must be equal, since the only solution satisfying the product-maximization condition and the sum-to-one constraint is $\omega_i=\frac{1}{m}$ for all i.

The theorem is proved.

D-optimal designs for trigonometric model

Let's consider D-optimal designs for the trigonometric regression model.

- The trigonometric regression model has the function:

$$\eta(x,\theta) = \theta_0 + \sum_{j=1}^{m} \theta_{2j-1}\sin(jx) + \sum_{j=1}^{m} \theta_{2j}\cos(jx),$$

  where $\theta = (\theta_0, \theta_1, \ldots \theta_{2m})^T$ is the vector of unknown parameters.
- The regression function vector is $f(t) = (1, \sin t, \cos t, \ldots, \sin(mt), \cos(mt))^T$.
- The design interval is $\chi = [-\pi, \pi]$.

Theorem 10

Let $\chi = [-\pi, \pi]$. An approximate D-optimal design for the trigonometric regression model is any design:

$$\xi_N^* = \begin{pmatrix} t_1^* & \ldots & t_N^* \\ 1/N & \ldots & 1/N \end{pmatrix},$$

where $t_i^* = \frac{i-1}{N}2\pi - \pi$ for $i = 1, \ldots, N$, and $N \geq 2m+1$ (m is the order of the regression model).

The uniform design is also D-optimal: $\xi^* = \frac{1}{2\pi}dx$.

Comments

Let us now examine the D-optimal design for a different class of models — trigonometric regression.

In this case, the regression function is a trigonometric polynomial: it consists of the constant term θ_0, followed by alternating sine and cosine terms of increasing frequency. Specifically, we have $\sin(x)$, $\cos(x)$, $\sin(2x)$, $\cos(2x)$, and so on, up to $\sin(mx)$ and $\cos(mx)$. Altogether, there are $2m+1$ parameters to estimate, so the design must be able to support at least that many.

The design domain is the interval from $-\pi$ to π, which reflects the periodic nature of trigonometric functions.

The theorem on the slide gives a simple and elegant solution: if we take N equally spaced points on the interval from $-\pi$ to π, where N is at least $2m+1$, and assign equal weights to each point, then the resulting design is D-optimal. These support points are computed using the formula: $t_i^* = \frac{i-1}{N}2\pi - \pi$. So the points are uniformly spaced and symmetric.

This design distributes the observations uniformly over the interval, and thanks to the orthogonality of trigonometric functions on this domain, the information matrix turns out to be diagonal.

Moreover, the continuous uniform distribution on the interval from $-\pi$ to π is also D-optimal. That is, the design with density $\frac{1}{2\pi}$ also leads to the maximum determinant of the information matrix.

Let's prove this theorem.

Proof:

Due to the orthogonality of trigonometric functions, for any integers $i, j > 0$:

$$\int_{-\pi}^{\pi} \sin ix \, dx = \int_{-\pi}^{\pi} \cos jx \, dx = \int_{-\pi}^{\pi} \sin ix \cos jx \, dx = 0$$

The orthogonality conditions also yield:

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \sin^2 ix \, dx = \frac{1}{2\pi} \int_{-\pi}^{\pi} \cos^2 jx \, dx = \frac{1}{2}$$

Using these properties, the information matrix of the uniform design $\xi^* = \frac{1}{2\pi} dx$ is diagonal:

$$M(\xi^*) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) f^T(x) dx = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & \frac{1}{2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{2} \end{pmatrix}.$$

Comments

Let's now go through the proof of this result. The key idea here is to exploit the orthogonality of trigonometric functions on the interval from $-\pi$ to π.

We begin with three basic integral identities. First, for any positive integer i, the integral from $-\pi$ to π of sin(ix) is zero. Second, for any positive integer j, the integral from $-\pi$ to π of cos(jx) is also zero. Third, the integral from $-\pi$ to π of the product of sin(ix) and cos(jx) is again zero.

These three equalities follow from the fact that sine and cosine functions are orthogonal over the interval from $-\pi$ to π. In particular, sine and cosine are orthogonal not just to each other, but also to the constant function, which explains why these integrals vanish.

Next, the orthogonality properties also tell us how the squared terms behave. If we integrate $\sin^2(ix)$ over the interval from $-\pi$ to π and divide by 2π, we get $\frac{1}{2}$. The same is true for $\cos^2(jx)$: its average over this interval is also $\frac{1}{2}$. This again holds for any positive integers i and j.

From this we automatically get that the information matrices of the designs defined in the conditions of the theorem are diagonal. The first element on the diagonal corresponds to the constant function and equals one. All the remaining diagonal elements correspond to sine and cosine terms and equal $\frac{1}{2}$. So the matrix has a one in the top-left corner and $\frac{1}{2}$ along the rest of the diagonal.

Proof (continued)

- The Kiefer–Wolfowitz theorem requires showing that the extremal polynomial $d(x, \xi^*) = f^T(x)D(\xi^*)f(x)$ satisfies $\max d(x, \xi^*) = 1 + 2m$.
- The inverse information matrix is:

$$D(\xi^*) = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 2 \end{pmatrix},$$

which gives the extremal polynomial

$$d(x, \xi^*) = 1 + 2(\sin^2 x + \cos^2 x + \dots + \sin^2(mx) + \cos^2(mx)) = 1 + 2m.$$

- The optimality of the discrete design ξ_N^* follows from the fact that for any $j = 1, \dots, 2m$, the sums of the trigonometric functions at the design points are zero:

$$\sum_{i=1}^{N} \sin jt_i^* = 0, \quad \sum_{i=1}^{N} \cos jt_i^* = 0.$$

The theorem is proved. □

Comments

To complete the proof, we apply the Kiefer–Wolfowitz equivalence theorem. We must show that the extremal polynomial, denoted $d(x, \xi^*)$, defined as $f^T(x)D(\xi^*)f(x)$, reaches its maximum value equal to $1 + 2m$.

Recall that $D(\xi^*)$ is the inverse of the information matrix. Since the original matrix is diagonal with one as the first entry and $\frac{1}{2}$ on all others, its inverse is also diagonal with entries one and two. So, $D(\xi^*)$ has one in the top-left corner, and twos along the remaining diagonal positions.

Substituting this into the definition of the extremal polynomial, we compute $d(x, \xi^*)$ as: one plus two times $\sin^2(x)$ plus two times $\cos^2(x)$, and so on, up to two times $\sin^2(mx)$ and two times $\cos^2(mx)$.

Each pair $\sin^2(jx) + \cos^2(jx)$ equals one. There are m such pairs, so the total becomes $1 + 2m$, as required. This proves that the continuous uniform design satisfies the optimality condition.

Finally, we show that the approximate design ξ_N^* with support points t_i^* — spaced uniformly on the interval from $-\pi$ to π — is also D-optimal. This follows from the fact that for every integer j from one to 2m, the sum of $\sin(jt_i^*)$ over all i equals zero, and similarly for the cosine terms.

Thus, the information matrix for the discrete design coincides with the continuous one. The theorem is proved.