Standard Deviation Method for Solving Determined Subsets in Overconstrained Systems

Your Name

December 15, 2024

Abstract

This document presents the **Standard Deviation Method** tailored for solving overconstrained systems by leveraging determined subsets of equations. The method involves random sampling, direct solving of determined systems, and statistical analysis to assess the stability and reliability of the coefficients.

Contents

1 Introduction

In many scientific and engineering applications, one encounters systems of equations where the number of equations exceeds the number of unknowns. Such systems are termed **overconstrained** or **overdetermined**. Traditional methods like Least Squares provide solutions that minimize residuals but may not adequately capture the variability and stability of the coefficients, especially in the presence of noise or inconsistent data.

The **Standard Deviation Method** offers an alternative approach by sampling subsets of equations that form **determined systems**, solving for the unknowns multiple times, and analyzing the statistical properties of the solutions. This method provides insights into the reliability of the coefficients and identifies regions of parameter space that yield stable solutions.

2 Problem Statement

Consider an overconstrained system of equations of the form:

$$\sum_{i=1}^{n} C_i G(z, \Delta_i) = v(z), \quad \forall z \in \{z_1, z_2, \ldots, z_m\} \tag{1}$$

where:

- m is the number of equations, with $m > n$.
- n is the number of unknown coefficients C_i.
- $G(z, \Delta_i)$ are known functions of the variable z and parameters Δ_i.
- $v(z)$ is the known target function.
- z_j are specific points at which the equations are evaluated.

The goal is to determine the coefficients C_i that satisfy the system as closely as possible.

3 Standard Deviation Method

The **Standard Deviation Method** tailored for overconstrained systems involves the following key steps:

1. **Random Sampling of Determined Subsets**: Randomly select subsets of n equations from the total m equations, ensuring each subset forms a determined system.
2. **Solving Determined Subsets**: Solve each sampled subset to obtain estimates of the coefficients.
3. **Statistical Analysis**: Compute the mean and standard deviation of the sampled coefficients.
4. **Reward Function**: Define a reward based on the stability (low standard deviation) of the coefficients.

3.1 Detailed Steps

3.1.1 1. Random Sampling of Determined Subsets

Given an overconstrained system with m equations and n unknowns, randomly select subsets of n equations. Each subset forms a **determined system**, meaning there is exactly one solution for the coefficients C_i. Repeat this sampling process N times to gather a diverse set of solutions.

3.1.2 2. Solving Determined Subsets

For each sampled subset of n equations, solve the linear system:

$$\sum_{i=1}^{n} C_i G(z_{j_k}, \Delta_i) = v(z_{j_k}), \quad k = 1, 2, \ldots, n$$

This can be represented in matrix form as:

$$\mathbf{G} \cdot \mathbf{C} = \mathbf{V}$$

where:

- $\mathbf{G}$ is an $n \times n$ matrix with elements $G(z_{j_k}, \Delta_i)$.
- $\mathbf{C}$ is the vector of unknown coefficients $[C_1, C_2, \ldots, C_n]^T$.
- $\mathbf{V}$ is the vector $[v(z_{j_1}), v(z_{j_2}), \ldots, v(z_{j_n})]^T$.

Since each subset is determined (n equations for n unknowns), the system can be solved using direct methods such as matrix inversion or linear solvers.

3.1.3 3. Statistical Analysis

After obtaining N estimates $\mathbf{C}^{(1)}, \mathbf{C}^{(2)}, \ldots, \mathbf{C}^{(N)}$, compute the statistical measures:

$$\overline{\mathbf{C}} = \frac{1}{N} \sum_{k=1}^{N} \mathbf{C}^{(k)} \tag{2}$$

$$\boldsymbol{\Sigma}_{\mathbf{C}} = \sqrt{\frac{1}{N-1} \sum_{k=1}^{N} (\mathbf{C}^{(k)} - \overline{\mathbf{C}})^2} \tag{3}$$

where $\overline{\mathbf{C}}$ is the mean vector and $\boldsymbol{\Sigma}_{\mathbf{C}}$ is the standard deviation vector of the coefficients.

3.1.4 4. Reward Function

Define a reward function that penalizes high variability in the coefficients relative to their mean values. The reward can be formulated as:

$$R = -\sum_{i=1}^{n} \log \left(\left| \frac{\Sigma_{C_i}}{\overline{C_i}} \right| \right)$$

where:

- $\overline{C_i}$ is the mean of the sampled coefficients for C_i.
- Σ_{C_i} is the standard deviation of the sampled coefficients for C_i.

This reward function encourages the agent to find regions in the parameter space where the coefficients C_i are stable (i.e., have low standard deviation relative to their mean values).

4 Mathematical Formulation

Given the overconstrained system:

$$\sum_{i=1}^{n} C_i G(z, \Delta_i) = v(z)$$

for m distinct values of z, with $m > n$, the system can be represented in matrix form as:

$$\mathbf{G} \cdot \mathbf{C} = \mathbf{V}$$

where:

$$\mathbf{G} = \begin{bmatrix} G(z_1, \Delta_1) & G(z_1, \Delta_2) & \cdots & G(z_1, \Delta_n) \\ G(z_2, \Delta_1) & G(z_2, \Delta_2) & \cdots & G(z_2, \Delta_n) \\ \vdots & \vdots & \ddots & \vdots \\ G(z_m, \Delta_1) & G(z_m, \Delta_2) & \cdots & G(z_m, \Delta_n) \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} C_1 \\ C_2 \\ \vdots \\ C_n \end{bmatrix}, \quad \mathbf{V} = \begin{bmatrix} v(z_1) \\ v(z_2) \\ \vdots \\ v(z_m) \end{bmatrix}$$

4.1 Solving Determined Subsets

For each sampled subset of n equations, the system is:

$$\mathbf{G}_{\text{sub}} \cdot \mathbf{C} = \mathbf{V}_{\text{sub}}$$

Since $\mathbf{G}_{\text{sub}}$ is an $n \times n$ square matrix, the solution can be found using:

$$\mathbf{C} = \mathbf{G}_{\text{sub}}^{-1} \cdot \mathbf{V}_{\text{sub}}$$

or more robustly using linear solvers that avoid explicit matrix inversion, such as Gaussian elimination or LU decomposition.

4.2 Reward Function

The reward function is designed to penalize high variability in the coefficients:

$$R = -\sum_{i=1}^{n} \log \left(\left| \frac{\Sigma_{C_i}}{\overline{C_i}} \right| \right)$$

This formulation ensures that:

- **Low Variability** (Σ_{C_i}) relative to the mean ($\overline{C_i}$) yields a higher reward.
- **High Variability** results in a lower (more negative) reward, discouraging such regions.

5 Advantages of the Standard Deviation Method with Determined Subsets

- **Computational Efficiency**: Solving determined systems using direct methods is generally faster and more accurate than Least Squares for each subset.
- **Stability and Reliability**: By analyzing the standard deviation of multiple determined solutions, the method provides insights into the stability of the coefficients.

- **Flexibility**: The method can be adapted to different sampling strategies and can handle varying degrees of overconstrained systems by adjusting subset sizes.
- **Avoidance of Singularities**: Random sampling reduces the likelihood of selecting ill-conditioned subsets, enhancing the robustness of the solution.

6 Application Example

Consider applying the Standard Deviation Method to the overconstrained system:

$$\sum_{i=1}^{2} C_i G(z, \Delta_i) = v(z)$$

for $m = 20$ equations and $n = 2$ unknowns C_1 and C_2.

6.1 Step-by-Step Example

6.1.1 1. Random Sampling of Determined Subsets

Randomly select 2 equations out of the 20 available. For instance, equations indexed j_1 and j_2.

6.1.2 2. Solving Determined Subsets

Solve the following system:

$$\begin{cases} C_1 G(z_{j_1}, \Delta_1) + C_2 G(z_{j_1}, \Delta_2) = v(z_{j_1}) \\ C_1 G(z_{j_2}, \Delta_1) + C_2 G(z_{j_2}, \Delta_2) = v(z_{j_2}) \end{cases}$$

Using direct solving methods to obtain $C_1^{(k)}$ and $C_2^{(k)}$.

6.1.3 3. Statistical Analysis

After repeating the sampling and solving process $N = 20$ times, compute:

$$\overline{C_1} = \frac{1}{N} \sum_{k=1}^{N} C_1^{(k)}, \quad \Sigma_{C_1} = \sqrt{\frac{1}{N-1} \sum_{k=1}^{N} (C_1^{(k)} - \overline{C_1})^2}$$

$$\overline{C_2} = \frac{1}{N} \sum_{k=1}^{N} C_2^{(k)}, \quad \Sigma_{C_2} = \sqrt{\frac{1}{N-1} \sum_{k=1}^{N} (C_2^{(k)} - \overline{C_2})^2}$$

6.1.4 4. Reward Calculation

Compute the reward:

$$R = -\left(\log\left(\left| \frac{\Sigma_{C_1}}{\overline{C_1}} \right| \right) + \log\left(\left| \frac{\Sigma_{C_2}}{\overline{C_2}} \right| \right) \right)$$

A higher reward corresponds to lower variability in the coefficients, suggesting more stable solutions.

7 Conclusion

The **Standard Deviation Method** effectively addresses the challenges posed by overconstrained systems by breaking them down into determined subsets. By solving these subsets directly and analyzing the statistical properties of the solutions, the method ensures both accuracy and stability in estimating the coefficients C_i.

This approach is particularly beneficial in scenarios where data may be noisy or inconsistent, as it provides a mechanism to evaluate and prioritize solutions based on their statistical robustness. By focusing on the variability of the coefficients across multiple determined systems, the method enhances the reliability of the solutions obtained from inherently overconstrained systems.

8 Incorporating Measurement Errors

Assume that the target function $v(z_j)$ is subject to additive measurement noise:

$$v(z_j) = v_{\text{true}}(z_j) + \epsilon_j \tag{4}$$

where:

- $v_{\text{true}}(z_j)$: True value of the target function at z_j.
- ϵ_j: Measurement noise, modeled as $\epsilon_j \sim \mathcal{N}(0, \sigma_v^2)$, i.e., Gaussian noise with zero mean and variance σ_v^2.

9 Sampling Determined Subsets

Each sampled subset of n equations forms a determined system:

$$\mathbf{G}_{\text{sub}} \cdot \mathbf{C} = \mathbf{V}_{\text{sub}} \tag{5}$$

where:

- $\mathbf{G}_{\text{sub}}$: $n \times n$ matrix comprising the coefficients $G(z_j, \Delta_i)$ of the sampled equations.
- $\mathbf{C}$: $n \times 1$ vector of unknown coefficients.
- $\mathbf{V}_{\text{sub}}$: $n \times 1$ vector of target values $v(z_j)$ for the sampled equations.

Solving this system yields:

$$\mathbf{C}^{(k)} = \mathbf{G}_{\text{sub}}^{-1}\mathbf{V}_{\text{sub}} = \mathbf{C}_{\text{true}} + \mathbf{G}_{\text{sub}}^{-1}\boldsymbol{\epsilon}_{\text{sub}}^{(k)} \tag{6}$$

where:

- $\mathbf{C}_{\text{true}}$: True coefficients vector.
- $\boldsymbol{\epsilon}_{\text{sub}}^{(k)}$: Noise vector for the k-th subset, $\boldsymbol{\epsilon}_{\text{sub}}^{(k)} \sim \mathcal{N}(\mathbf{0}, \sigma_v^2\mathbf{I}_n)$.

10 Distribution of Coefficient Estimates

From Equation (6), each coefficient estimate $\mathbf{C}^{(k)}$ can be expressed as:

$$\mathbf{C}^{(k)} = \mathbf{C}_{\text{true}} + \mathbf{G}_{\text{sub}}^{-1}\boldsymbol{\epsilon}_{\text{sub}}^{(k)} \tag{7}$$

Assuming that each subset $\mathbf{G}_{\text{sub}}$ is sampled uniformly and independently, the covariance matrix of the coefficient estimates is:

$$\operatorname{Cov}(\mathbf{C}^{(k)}) = \sigma_v^2 \left(\mathbf{G}_{\text{sub}}^{-1}\mathbf{G}_{\text{sub}}^{-T}\right) = \sigma_v^2(\mathbf{G}_{\text{sub}}\mathbf{G}_{\text{sub}}^{T})^{-1} \tag{8}$$

11 Statistical Analysis of Coefficient Estimates

After performing N samplings, we obtain N estimates of the coefficients $\mathbf{C}^{(k)}$. We compute the following statistical measures for each coefficient C_i:

11.1 Mean of Coefficients

$$\overline{C_i} = \frac{1}{N}\sum_{k=1}^{N} C_i^{(k)} \tag{9}$$

11.2 Standard Deviation of Coefficients

$$\Sigma_{C_i} = \sqrt{\frac{1}{N-1}\sum_{k=1}^{N}\left(C_i^{(k)} - \overline{C_i}\right)^2} \tag{10}$$

12 Reward Function

The reward function is defined as:

$$R = -\sum_{i=1}^{n} \log\left(\left|\frac{\Sigma_{C_i}}{\overline{C_i}}\right|\right) \tag{11}$$

This function penalizes high variability in the coefficients relative to their mean values, thereby encouraging stable and reliable solutions.