

Regression Model for Independent Variables with Uncertainty

ENDO Yasunori[†] MIYAMOTO Sadaaki[†]

[†] Department of Risk Engineering,
Faculty of System Information and Engineering,
University of Tsukuba

1st September, 2009, IIASA, Wien

Outline

- 1 Introduction
 - Background
 - Our Goal
- 2 Preliminaries
 - Hard c -Regression Model
 - Data with Tolerance
 - Data with Tolerance for c -Regression Model
- 3 Theory
 - Outline of Theory
 - Objective Function
 - Optimal Solutions
- 4 Proposed Algorithm
- 5 Numerical Examples
 - GDP Data
 - Calculation Result
 - Consideration
- 6 Conclusion

Regression Model and Hard c -Regression Model

Regression Model

- Regression model is one way to find linear constructure of data.
- We can get a regression line which is calculated by minimizing the total of the dissimilarities between the line and the pair of independent and dependent variables.

Hard c -Regression Model, HCRM

- HCRM is one way to find c number of regression lines based on the regression model.
- While it calculates the regression lines, it find which pair of independent and dependent variables belongs to which regression line.
⇒ It can be regarded as one of the clustering methods.

Regression Model and Hard c -Regression Model

Regression Model

- Regression model is one way to find linear constructure of data.
- We can get a regression line which is calculated by minimizing the total of the dissimilarities between the line and the pair of independent and dependent variables.

Hard c -Regression Model, HCRM

- HCRM is one way to find c number of regression lines based on the regression model.
- While it calculates the regression lines, it find which pair of independent and dependent variables belongs to which regression line.
⇒ It can be regarded as one of the clustering methods.

Regression Model and Hard c -Regression Model

Objective Functions

The algorithms of the regression model and clustering are constructed using optimal solutions of the given objective function.



Performance of algorithms of the regression model and clustering strongly depends on the objective function.



The objective function plays main role.

Uncertainty of Data

In general

Each classified data \Rightarrow A point on a pattern space

However

Measurement error or inherent obscurity of data.



It is more natural that such data are represented as sets than points in the pattern space, for example,

$$[x, \bar{x}] \subset \mathcal{R}.$$

Uncertainty of Data

In general

Each classified data \Rightarrow A point on a pattern space

However

Measurement error or inherent obscurity of data.



It is more natural that such data are represented as sets than points in the pattern space, for example,

$$[\underline{x}, \bar{x}] \subset \mathcal{R}.$$

Problems on Handling Uncertainty

Similarities

Similarities play very important role in the regression model and clustering. From the above veiwpoint, those must be defined as **measures between sets**, not between points.

The Problems

- We can not obtain the exact optimal solutions of objective functions when we introduce the measures between sets into the functions.
- Which dissimilarity between sets should we use ? etc.

Problems on Handling Uncertainty

Similarities

Similarities play very important role in the regression model and clustering. From the above veiwpoint, those must be defined as **measures between sets**, not between points.

The Problems

- We can not obtain the exact optimal solutions of objective functions when we introduce the measures between sets into the functions.
- Which dissimilarity between sets should we use ? etc.

To Solve the Above Problems

We introduce a concept of **tolerance**.



We can naturally formulate clustering problems for data which are represented as sets into optimization problems.

Data with Tolerance

data + tolerance vectors
(with constraints of tolerance vectors)

Conventional c -Regression Model for Uncertain Data

Uncertainty with independent variables has been hardly considered.

The Reasons

In case of independent variables with uncertainty,

- 1 The algorithms have been constructed by not analytical methodology but numerical one, because it is difficult to analytically derive the equations which are based on the algorithms.
- 2 Any regression line is not uniquely determined.

It is better that

- 1 The algorithms are analytically constructed.
- 2 All of the regression lines are uniquely determined.

Conventional c -Regression Model for Uncertain Data

Uncertainty with independent variables has been hardly considered.

The Reasons

In case of independent variables with uncertainty,

- 1 The algorithms have been constructed by not analytical methodology but numerical one, because it is difficult to analytically derive the equations which are based on the algorithms.
- 2 Any regression line is not uniquely determined.

It is better that

- 1 The algorithms are analytically constructed.
- 2 All of the regression lines are uniquely determined.

Conventional c -Regression Model for Uncertain Data

Uncertainty with independent variables has been hardly considered.

The Reasons

In case of independent variables with uncertainty,

- 1 The algorithms have been constructed by not analytical methodology but numerical one, because it is difficult to analytically derive the equations which are based on the algorithms.
- 2 Any regression line is not uniquely determined.

It is better that

- 1 The algorithms are analytically constructed.
- 2 All of the regression lines are uniquely determined.

Our Goal

Simply Speaking

We try to solve the above two problems to introduce the concept of tolerance.

Flow

- 1 Independent and dependent variables with uncertainty are formulated by **tolerance**.
- 2 c number of regression lines of which the total of dissimilarities between the pair of the independent and dependent variables is minimum are analytically derived.
- 3 An algorithm of the c -regression model for uncertain data is constructed.
- 4 The algorithm is verified through a numerical example.

HCRM

The Purpose

HCRM determines c number of regression lines H_i ($i = 1 \sim c$):

$$H_i = \{(x, y) \mid y = h_i(x) = \beta_i^T x + \beta_{i,p+1}, \beta_i, x \in \mathbb{R}^p\}$$

Variables

Independent variables: $x_k = (x_{k1}, \dots, x_{kp}) \in \mathbb{R}^p$ ($k = 1 \sim n$)

Dependent variables: $y_k \in \mathbb{R}$

Regression parameters: $\beta_i, \beta_{i,p+1}$

$B = \{\beta_{ij}\}$ ($j = 1, \dots, p+1$)

Membership grades: $u_{ki} \in \{0, 1\}$

$U = \{u_{ki}\}$

HCRM

The Way to Find the Regression Lines

$$J(U, B) = \sum_{k=1}^n \sum_{i=1}^c u_{ki} D_{ki} \rightarrow \min$$

$$\begin{aligned} D_{ki} &= (y_k - \beta_i^T x_k - \beta_{i,p+1})^2 \\ &= (y_k - h_i(x_k))^2 \end{aligned}$$

Data with Tolerance

General Formulation of Uncertainty

$$(x_1, \dots, x_p) \in \mathbb{R}^p \Rightarrow ([\underline{x}_1, \bar{x}_1], \dots, [\underline{x}_p, \bar{x}_p]) \subset \mathbb{R}^p$$

Formulation of Uncertainty by Tolerance

$$x_k + \delta_k \in \mathbb{R}^p$$

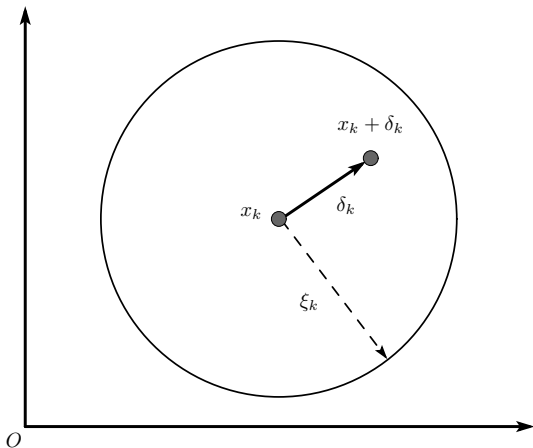
Center vectors: $x_k \in \mathbb{R}^p$

Tolerance vectors: $\delta_k = (\delta_{k1}, \dots, \delta_{kp})^T \in E \subset \mathbb{R}^p$

Constraints:

$$\|\delta_k\|^2 = \sum_{j=1}^p (\delta_{kj})^2 \leq \xi_k^2 \quad (\xi_k \geq 0)$$

An Example of Data with Tolerance on \mathcal{R}^2



Maximum Tolerance Range

Simply Speaking

Each data can move into the following closed ball:

$$CB(x_k; \xi_k) = \{x_k + \delta_k \in \mathbb{R}^p \mid \|\delta_k\|^2 \leq \xi_k^2, \xi_k \geq 0\}$$

Maximum Tolerance Range

$$CB(x_k; \xi_k)$$

Significance of Introducing the Concept of Tolerance

Tolerance vectors δ_k are obtained as **solutions of the optimization problem**. The tolerance vectors can be calculated by iterative optimization.



We can construct algorithms using dissimilarities based on the conventional distance, not on special measures between sets.

Notice

- According to the real data, the tolerance range ξ_k are given in advance.
- Each data is possible to be any point in the tolerance range.

⇒ The tolerance vectors δ_k is calculated not to estimate original position of each data x_k but to derive the suitable clustering results.

Significance of Introducing the Concept of Tolerance

Tolerance vectors δ_k are obtained as **solutions of the optimization problem**. The tolerance vectors can be calculated by iterative optimization.



We can construct algorithms using dissimilarities based on the conventional distance, not on special measures between sets.

Notice

- According to the real data, the tolerance range ξ_k are given in advance.
- Each data is possible to be any point in the tolerance range.

⇒ The tolerance vectors δ_k is calculated not to estimate original position of each data x_k but to derive the suitable clustering results.

Tolerance Vectors of Independent and Dependent Variables

c -Regression Model

Independent and dependent variables belong to different spaces.



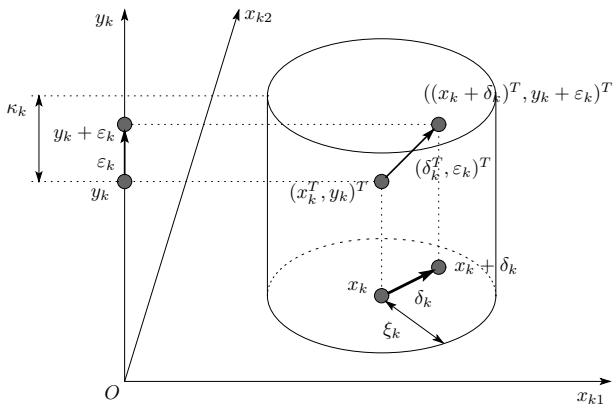
Tolerance ranges are separately given.



Tolerance vectors of independent variables: δ_k

Tolerance vectors of dependent variables: ε_k ($k = 1, \sim, n$)

An Example: Spaces of independent and dependent variables are \mathbb{R}^2 and \mathbb{R} , respectively.



Theory

- 1 We define an objective function for uncertain data using the concept of tolerance.

This objective function is a natural extension of the conventional c -regression model.

- 2 We analytically derive the unique optimal solutions which minimize the objective function.

The most significance of our theory is as follows:

- 1 The optimal solutions are **analytically** derived.
- 2 All of the regression lines are **uniquely** determined.

The reason is attributed to representation of uncertainty of data by the concept of tolerance.

Theory

- 1 We define an objective function for uncertain data using the concept of tolerance.

This objective function is a natural extension of the conventional c -regression model.

- 2 We analytically derive the unique optimal solutions which minimize the objective function.

The most significance of our theory is as follows:

- 1 The optimal solutions are **analytically** derived.
- 2 All of the regression lines are **uniquely** determined.

The reason is attributed to representation of uncertainty of data by the concept of tolerance.

Objective Function and Constraints

The Objective Function

$$J(U, B, \Delta, E) = \sum_{k=1}^n \sum_{i=1}^c u_{ki} D_{ki} \rightarrow \min$$

$$\begin{aligned} D_{ki} &= (y_k + \varepsilon_k - \beta_i^T (x_k + \delta_k) - \beta_{i,p+1})^2 \\ &= (y_k + \varepsilon_k - h_i(x_k + \delta_k))^2 \end{aligned}$$

Constraints

$$\sum_{i=1}^c u_{ki} = 1, \quad \|\delta_k\|^2 \leq \xi_k^2, \quad |\varepsilon_k|^2 \leq \kappa_k^2$$

Optimal Solutions of Membership Grades u_{ki}

$$u_{ki} = \begin{cases} 1 & (i = \arg \min_l D_{kl}) \\ 0 & (\text{otherwise}) \end{cases}$$

Optimal Solutions of Regression Parameters β_i

$$\begin{aligned}\beta'_i &= (\beta_i^T, \beta_{i,p+1})^T \\ &= \left(\sum_{k=1}^n u_{ki} z_k z_k^T \right)^{-1} \sum_{k=1}^n u_{ki} (y_k + \varepsilon_k) z_k \\ z_k &= \begin{pmatrix} x_k + \delta_k \\ 1 \end{pmatrix}\end{aligned}$$

Optimal Solutions of Tolerance Vectors δ_k of x_k

$$\delta_k = \begin{cases} \left(\delta_{k1}, \dots, \delta_{k,p-1}, \frac{\eta_{ki} - \sum_{j=1}^{p-1} \beta_{ij} \delta_{kj}}{\beta_{ip}} \right)^T & (\|\delta_k\|^2 < \xi^2) \\ \eta_{ki} (\beta_i \beta_i^T + \lambda_k \mathbb{I})^{-1} \beta_i & (\|\delta_k\|^2 \geq \xi^2) \end{cases}$$

$$\eta_{ki} = y_k + \varepsilon_k - \beta_i^T x_k - \beta_{i,p+1}$$

$$\lambda_k = \left(\frac{|\eta_{ki}|}{\xi_k} - \|\beta_i\| \right) \|\beta_i\|$$

$$i = \arg_l \{ u_{kl} = 1 \}$$

$$\delta_{kj} = \text{an arbitrary real number with } \|\delta_k\|^2 < \xi^2 \\ (j = 1 \sim p - 1)$$

Optimal Solutions of Tolerance Vectors ε_k of y_k

$$\begin{aligned}\varepsilon_k &= -\alpha_k (y_k - \beta_i^T (x_k + \delta_k) - \beta_{i,p+1}) \\ \alpha_k &= \min \left\{ \frac{\xi_k}{|y_k - \beta_i^T (x_k + \delta_k) - \beta_{i,p+1}|}, 1 \right\} \\ i &= \arg_l \{u_{kl} = 1\}\end{aligned}$$

Proposed Algorithm

Algorithm (Hard c -Regression Model for Data with Tolerance in Independent Variables, HCRTI)

HCRTI1 *Set the initial values of B , Δ and E .*

HCRTI2 *Calculate $U = \arg \min J$ on fixing B , Δ and E .*

HCRTI3 *Calculate $B = \arg \min J$ on fixing U , Δ and E .*

HCRTI4 *Calculate $\Delta = \arg \min J$ on fixing B , U and E .*

HCRTI5 *Calculate $E = \arg \min J$ on fixing U , B and Δ .*

HCRTI6 *If the stop criterion satisfies, the algorithm is finished. Otherwise, go back to **HCRTI2**.*

Numerical Examples

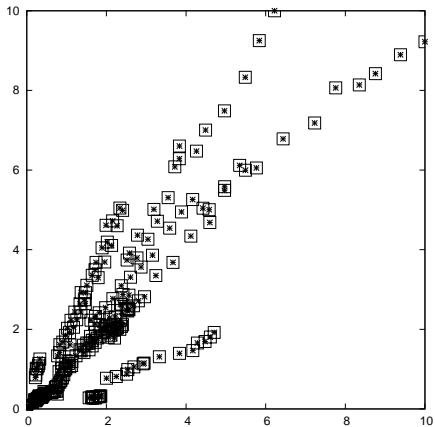
GDP Data

- Pairs of GDP and Energy Consumption of 240 Regions of Asia from 1973 to 1992. This data set is regularized into $[0, 10]$.
- We try to classify the data set into three clusters.

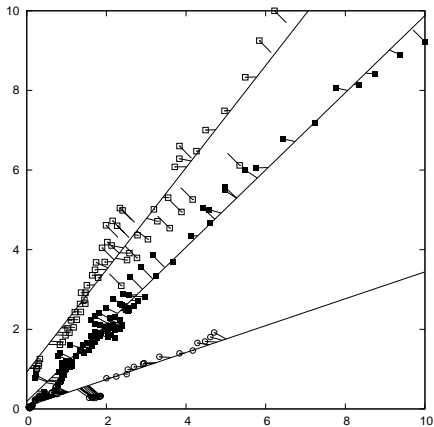
Symbols

- Each circle, black box and white box: Each cluster
- Each line: Regression line
- Segment between each center of data and each regression line: Tolerance vector

Data



Calculation Result



Consideration

Although the uncertainty which is represented as hyper-rectangle is given to each data,

- all of the regression lines are uniquely calculated,
- the calculation result is very natural as compared to the conventional hard c -regression model for data without uncertainty.

Conclusion

The regression model for independent variables with uncertainty has been hardly considered because of the following reasons.

- 1 It is difficult to analytically derive the optimal solutions of the objective functions of the regression model.
- 2 The regression line is not uniquely determined.

In this presentation

- 1 We have shown one way to solve the problems by formulating uncertainty of data using the concept of **tolerance**.
- 2 We have constructed **an alternative optimization algorithm** of hard c -regression model for independent and dependent variables with uncertainty.

Acknowledgment

We thank Professor UCHIYAMA Yoji of University of Tsukuba for presentation of the GDP data.

This study has partly been supported by the Grant-in-Aid for Scientific Research (C) (No.21500212) and (B) (No.19300074), Japan Society for the Promotion of Science.

Thank you for your kind attention.

endo@risk.tsukuba.ac.jp