Universität zu Köln

Dr. Boqiang Huang

Mathematical Foundations of Data Analysis II — SS 2019

Problem Set 01

Published on 11.04.2019 To be collected on 16.04.2019

Problem 1: (20 points)
This problem comes from a famous method called High Dimensional Model Representation (HDMR) or ANalysis Of VAriance (ANOVA). A simple definition of the HDMR or ANOVA data decomposition model is as follows:
Let $\mathbb{I}$ denote the unit interval $[0,1]$, $\mathbb{I}^n$ the $n-$dimensional unit hypercube and the random variable $\mathbf{x} \in \mathbb{I}^n$. Let $f : \mathbb{I}^n \to \mathbb{R}$ represent a general dynamic system, which depends on the n-D random variable $\mathbf{x} = [\mathrm{x}_1, \cdots, \mathrm{x}_n]^T$. Suppose that the system function $f(\mathbf{x})$ is integrable in a reasonable certain of degree, e.g. square integrable or third-order integrable, it could be represented as a sum of 2^n sub-components defined in all possible sub-spaces:

$$f(\mathbf{x}) = f_\emptyset + \sum_{d=1}^{n} \sum_{i_1 < \cdots < i_d}^{n} f_{\{i_1,\cdots,i_d\}}(\mathrm{x}_{i_1}, \cdots, \mathrm{x}_{i_d}), \tag{1}$$

where the interior sum is over all sets of d integers $i_1, \cdots, i_d$ that satisfy $1 \le i_1 < \cdots < i_d \le n$. Relation (1) also can be understood as

$$f(\mathbf{x}) = f_\emptyset + \sum_i f_{\{i\}}(\mathrm{x}_i) + \sum_{i<j} f_{\{i,j\}}(\mathrm{x}_i, \mathrm{x}_j) + \cdots + f_{\{1,2,\cdots,n\}}(\mathrm{x}_1, \mathrm{x}_2, \cdots, \mathrm{x}_n). \tag{2}$$

For example, taking $n = 3$, we have $\mathbf{x} = [\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3]^T \in \mathbb{I}^3$, and

$$f(\mathbf{x}) = f_\emptyset + f_{\{1\}}(\mathrm{x}_1) + f_{\{2\}}(\mathrm{x}_2) + f_{\{3\}}(\mathrm{x}_3) + f_{\{1,2\}}(\mathrm{x}_1, \mathrm{x}_2) + f_{\{1,3\}}(\mathrm{x}_1, \mathrm{x}_3) + f_{\{2,3\}}(\mathrm{x}_2, \mathrm{x}_3) + f_{\{1,2,3\}}(\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3). \tag{3}$$

The first summand $f_\emptyset$ is the global mean of $f(\mathbf{x})$, e.g.

$$f_\emptyset := \int_{\mathbb{I}^n} f(\mathbf{x}) d\mathbf{x} := \int_{\mathbb{I}} \cdots \int_{\mathbb{I}} f(\mathbf{x}) d\mathrm{x}_1 \cdots d\mathrm{x}_n, \tag{4}$$

while other summands can be obtained recursively

$$\begin{aligned}
&\int_{\mathbb{I}^{n-1}} f(\mathbf{x}) \prod_{k \ne i} d\mathrm{x}_k = f_\emptyset + f_{\{i\}}(\mathrm{x}_i), \quad i = 1, 2, \cdots, n \\
&\int_{\mathbb{I}^{n-2}} f(\mathbf{x}) \prod_{k \ne i,j} d\mathrm{x}_k = f_\emptyset + f_{\{i\}}(\mathrm{x}_i) + f_{\{j\}}(\mathrm{x}_j) + f_{\{i,j\}}(\mathrm{x}_i, \mathrm{x}_j), \quad i \ne j = 1, 2, \cdots, n \\
&\cdots \qquad = \qquad \cdots \\
&f(\mathbf{x}) = f_\emptyset + \sum_{d=1}^{n-1} \sum_{i_1 < \cdots < i_d}^{n} f_{\{i_1,\cdots,i_d\}}(\mathrm{x}_{i_1}, \cdots, \mathrm{x}_{i_d}) + f_{\{1,2,\cdots,n\}}(\mathrm{x}_1, \mathrm{x}_2, \cdots, \mathrm{x}_n).
\end{aligned} \tag{5}$$

If we reconsider the example shown in (3), we can obtain each summand from the given $f(\mathbf{x})$, e.g.

$$\begin{aligned}
f_{\{2\}}(\mathrm{x}_2) &:= \int_{\mathbb{I}} \int_{\mathbb{I}} f(\mathbf{x}) d\mathrm{x}_1 d\mathrm{x}_3 - f_\emptyset \\
f_{\{1,3\}}(\mathrm{x}_1, \mathrm{x}_3) &:= \int_{\mathbb{I}} f(\mathbf{x}) d\mathrm{x}_2 - f_\emptyset - f_{\{1\}}(\mathrm{x}_1) - f_{\{3\}}(\mathrm{x}_3).
\end{aligned} \tag{6}$$

Now, if we define the full sub-index set $\mathbb{S} := \{1, 2, \cdots, n\}$, its subset $\mathbb{U} \subseteq \mathbb{S}$, and the size of the set $|\mathbb{U}|$, e.g. $|\mathbb{S}| = n$, the whole HDMR or ANOVA can be compactly represented as

$$\begin{aligned}
f(\mathbf{x}) &= \sum_{\mathbb{U} \subseteq \mathbb{S}} f_{\mathbb{U}}(\mathbf{x}_{\mathbb{U}}), \\
f_{\mathbb{U}}(\mathbf{x}_{\mathbb{U}}) &= \int_{\mathbb{I}^{n-|\mathbb{U}|}} f(\mathbf{x}) d\mathbf{x}_{\mathbb{S}\setminus\mathbb{U}} - \sum_{\mathbb{V} \subset \mathbb{U}} f_{\mathbb{V}}(\mathbf{x}_{\mathbb{V}}).
\end{aligned} \tag{7}$$

Now, with the basic definitions in the probability and statistics, please prove following nice properties:
(1) Exclude $f_\emptyset$, all other summands $f_{\mathbb{U}}(\mathbf{x}_{\mathbb{U}})$ has the zero mean (7 points)

$$\int_{\mathbb{I}} f_{\mathbb{U}}(\mathbf{x}_{\mathbb{U}}) d\mathrm{x}_k = 0, \quad \mathbb{U} \subseteq (\mathbb{S} \setminus \emptyset), \ k \in \mathbb{U}; \tag{8}$$

(2) Include $f_\emptyset$, every two distinct summands $f_{\mathbb{U}}(\mathbf{x}_{\mathbb{U}})$ and $f_{\mathbb{V}}(\mathbf{x}_{\mathbb{V}})$ are orthogonal (7 points)

$$\int_{\mathbb{I}^n} f_{\mathbb{U}}(\mathbf{x}_{\mathbb{U}}) f_{\mathbb{V}}(\mathbf{x}_{\mathbb{V}}) d\mathbf{x} = 0, \quad \mathbb{U}, \mathbb{V} \subseteq \mathbb{S}, \ \mathbb{U} \neq \mathbb{V}; \tag{9}$$

(3) Define the variance of each summand and the total variance of $f(\mathbf{x})$ as

$$\begin{aligned} D_{\mathbb{U}} &:= \int_{\mathbb{I}^{|\mathbb{U}|}} f^2_{\mathbb{U}}(\mathbf{x}_{\mathbb{U}}) d\mathbf{x}_{\mathbb{U}}, \quad \mathbb{U} \subseteq (\mathbb{S} \setminus \emptyset) \\ D &:= \int_{\mathbb{I}^n} (f(\mathbf{x}) - f_\emptyset)^2 \, d\mathbf{x} \end{aligned} \tag{10}$$

the following equality exists (6 points)

$$D := \sum_{\mathbb{U} \subseteq (\mathbb{S} \setminus \emptyset)} D_{\mathbb{U}}. \tag{11}$$

(Additional Bonus: 10 points)
If we define the third-order moment, the skewness, and the fourth-order moment, the kurtosis, as follows

$$\begin{aligned} \gamma &:= \int_{\mathbb{I}^n} (f(\mathbf{x}) - f_\emptyset)^3 \, d\mathbf{x}, \\ \kappa &:= \int_{\mathbb{I}^n} (f(\mathbf{x}) - f_\emptyset)^4 \, d\mathbf{x}, \end{aligned} \tag{12}$$

Can you derive the corresponding equality between γ and $\gamma_{\mathbb{U}}$, and the one between κ and $\kappa_{\mathbb{U}}$?

Remarks:
The reason that I changed my original mind and design such big problem as the only problem in set 01 is:
(a) this is a very famous data decomposition method based on the concepts in statistics, which is unfortunately not introduced in many statistical data analysis textbooks;
(b) if we divide the both sides of (11) by D, and define that

$$1 := \sum_{\mathbb{U} \subseteq (\mathbb{S} \setminus \emptyset)} \frac{D_{\mathbb{U}}}{D} =: \sum_{\mathbb{U} \subseteq (\mathbb{S} \setminus \emptyset)} S_{\mathbb{U}}, \tag{13}$$

the $S_{\mathbb{U}}$ could be understood as the sensitivity index of those random variables $\mathbf{x}_{\mathbb{U}}$ to the dynamic system $f(\mathbf{x})$. The larger $S_{\mathbb{U}}$, the more sensitive of the system to $\mathbf{x}_{\mathbb{U}}$. In other words, those $\mathbf{x}_{\mathbb{U}}$ will be more important to the system. In the future, if you are facing to an unfamiliar big system or black box, and would like to understand it or control it, hope that you shall take some benefits from here.

Two useful references:
[1] I.M. Sobol, Theorems and examples on high dimensional model representation, Reliability Engineering and System Safety, vol. 79, pp. 187-193, 2003.
[2] M. Holtz, Sparse grid quadrature in high dimensions with applications in finance and insurance, Chapter 2, Springer-Verlag Berlin Heidelberg, 2011.