Ann Inst Stat Math (2018) 70:275–278
https://doi.org/10.1007/s10463-017-0641-x

INVITED ARTICLE: AKAIKE MEMORIAL LECTURE

# Discussion on the paper by Professor Wu

**Ryo Yoshida**[1]

Received: 5 September 2017 / Published online: 9 February 2018

I express much respect to the great achievements in history of statistical science that have been made by Professor Wu and his coauthors. In conventional experimental design, the factor interactions are often aliased as exemplified for the aliased two-factor interactions in the $2^{4-1}$ design. It has been shown that the CME reparameterization based on Eq 3 or Eq 4 in Wu's paper could be used to de-alias the aliased interaction effects in regular $2^{k-q}$ design.

As discussed by Professor Wu, the CME analysis has the great applicability not only in designed experiments but also in observation studies. Let us focus on the two-factor interaction between $A$ and $B$ with each having two levels, $+$ or $-$, that indicates the presence or absence of the respective factor. Conventionally, the interaction effect is to quantify the product-type influence of $A$ and $B$ on a response variable, which is defined to be the difference of the mean effects between the same signed (both are present or absent) and opposite-signed states. This describes merely one aspect of the interaction in a broader context. The CMEs bring to us another look on the interaction, which provide scientifically more meaningful insights in many applications. Mak and Wu (2017) developed a comprehensive framework of the CME analysis in observation studies with a variable selection procedure based on the effect grouping (Mak and Wu 2017).

Here I try to highlight a great potential of the CME analysis in a study of chemistry. The objective is to uncover the underlying relationship between a physical property and the molecular structure of a chemical compound with given data. Usually, the

The Related Articles are https://doi.org/10.1007/s10463-017-0640-y; https://doi.org/10.1007/s10463-017-0639-4; https://doi.org/10.1007/s10463-018-0646-0.

✉ Ryo Yoshida
yoshidar@ism.ac.jp

[1] The Institute of Statistical Mathematics, 10-3 Midoricho, Tachikawa, Tokyo 190-8562, Japan

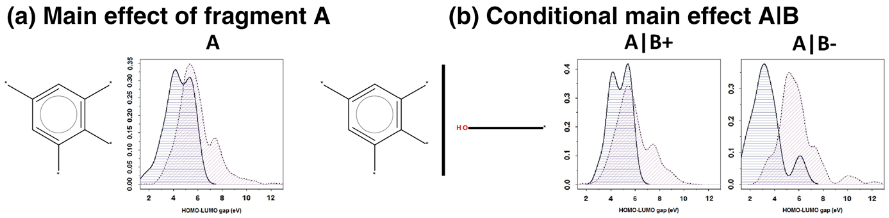

**Fig. 1** Example of the identified chemical fragments in the CME analysis. **a** The fragment $A$ shown in the left exhibited the significant ME. The right figure denotes the distributions of the HOMO–LUMO gaps that correspond to $A+$ (shading with the slope 0°) and $A-$ (45°), respectively. **b** The CME of $A$ given the fragment $B$ exhibited the significance and insignificance for CME($A|B+$) and CME($A|B-$), respectively.

molecular structure is represented by a bit vector through a description procedure called the molecular fingerprinting. An element of the descriptor takes one or zero according to the presence or absence of a specific chemical fragment in which several thousand or more fragments are considered conventionally. In history, various types of fragment sets have been developed to account for structure–property relationships in different applications. For instance, the R package *rcdk* provides ten different fingerprint descriptors (The rcdk package). The structure–property relationship analysis aims to evaluate the effect of each single fragment on the targeted physical property. On the other hand, there has been considerably less progress made in studies on the interaction effects of paired fragments. This might be because of high-dimensionality of the feature space typically comprised of several thousand fragments. I was highly motivated to introduce the CMEs of fragment pairs to the structure–property relationship analysis.

The data that I used were produced in our previous study (Ikebata et al. 2017). For 16,674 organic compounds, the physical properties called the HOMO–LUMO gaps were measured through the quantum chemistry calculation. We prepared the 102 chemical fragments known to be associated with various properties of polymers such as heat capacity, densities. With randomly chosen 1000 samples, I simply run Student's $t$ tests to evaluate the 102 main effects and the 20,604 CMEs which correspond to all possible combinations of two fragments in the given set. The false discovery rate of the multiple testing was controlled at $q$-value $\leq 0.05$ (Storey 2003), then resulting in 30 significant main effects and 3,705 significant CMEs.

The CME analysis exhibits increased scientific values as the effect significances are interpreted in various combinations. For example, when the main effect of the chemical fragment $A$ is present, we are more concerned with detailed circumstances in which the fragment $A$ retains the function of increasing or decreasing the HOMO–LUMO gap. According to the CME reparameterization, the main effect ME($A$) can be represented as ME($A$) $= \{$CME($A|B+$) $+$ CME($A|B-$)$\}/2$ with any conditional factor $B$. For each of the statistically significant main effects, I identified paired conditionals in which ME($A$) is significant, but one of CME($A|B+$) and CME($A|B-$) is found to be insignificant as shown in Fig. 1. A chemist can find clues to chemical structure manipulation as investing the effect of $A$ depending on whether or not the current compound contains the fragment $B$.

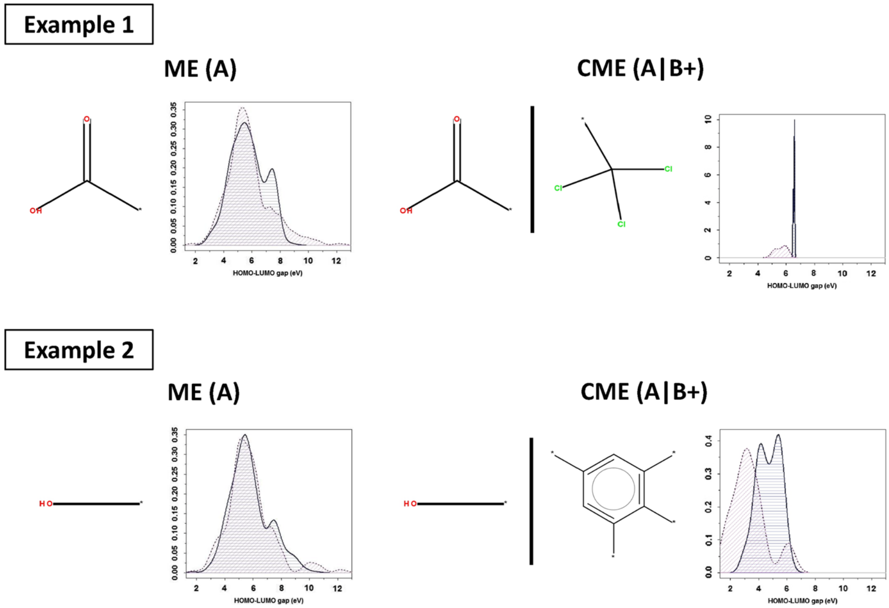

**Fig. 2** Two examples of fragment pairs in which the tests on ME($A$) and CME($A|B-$) resulted in to be insignificant, but the effect of $A$ turned to be significant, i.e., CME($A|B+$), when $B$ is present

There are many other ways to the effect grouping that provide chemically valuable insights to the understanding of structure–property relationships. The CME analysis identified 264 pairs of two factors in which ME($A$) and CME($A|B-$) are insignificant, but CME($A|B+$) is significant. Figure 2 shows some examples. The chemical implication is that the fragment $A$ is effective to increase or decrease the HOMO–LUMO gap whenever $B+$ is present. In the case where ME($A$) and CME($A|B-$) are both significant but CME($A|B+$) is insignificant, $B$ is implicated as a silencer of the factor $A$. In addition, we should consider higher-order CMEs such as ternary or quadruplet factors.

In this preliminary analysis, I focused only on the HOMO–LUMO gaps of the small number of compounds. If the CME analysis is performed with more comprehensive data sets that consist in various properties of millions of compounds and a larger fragment set, it is expected to obtain an overall grasp of the structure–property relationship. Such comprehensive studies would lead to a lot of new things in chemistry as briefly demonstrated here. One difficulty we then encounter is the computational issue in considering a quite large number of higher-order CMEs. For example, we need to run 686,800 tests when considering the CMEs with the third order only for the 102 fragments. In addition, it is required to establish a systematic way of interpreting and summarizing testing results on such many effects in combination.

## References

Ikebata, H., Hongo, K., Isomura, T., Maezono, R., Yoshida, R. (2017). Bayesian molecular design with a chemical language model. *Journal of Computer-Aided Molecular Design*, *31*(4), 379–391.

Mak, S., Wu, C. F. J. (2017). cmenet: A new method for bi-level variable selection of conditional main effects. arxiv.1701.05547.

Storey, S. D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *Annals of Statististics*, *31*(6), 2013–2035.

The rcdk package: R interface to the CDK libraries. https://cran.r-project.org/web/packages/rcdk/index.html.