Appendix: Model-selection strategy in DSDApp

In the framework of DSD, possible second-order models are created after experimenting with the chosen design. One of the common ways of building models is the use of the Akaike information criterion with finite correction (AICc). AICc (or generally AIC) is an estimator to select a "good" model that can explain the given data and avoid overfitting by including as few terms as possible in the model. Assuming that the errors of all factors follow independent and identical normal distributions N(0, σ^2) with variance σ^2, AICc is expressed for the least square estimation[10] as:

$$\mathrm{AICc} = n\ln\left(\hat{\sigma}^2\right) + 2k + \frac{2k(k+1)}{n-k-1} \quad (1) \quad ,$$

where n is the number of runs and k is the number of the factors in the model. Here, the estimated variance $\hat{\sigma}^2$ is calculated by:

$$\hat{\sigma}^2 = \frac{\sum\left(y_i - \hat{y}_i\right)^2}{n} \quad (2) \quad ,$$

where y_i and $\hat{y}_i$ are observed and predicted values, respectively, of the i^{th} factor. As can be seen from equation (1), a "good" model has a smaller value of AICc because $\hat{\sigma}^2$ should be small, and k should be also small to avoid overfitting. In DSDApp, the two-step approach similar to the effective model selection in reference[9] is employed. Algorithm 1 describes how the model is selected in DSDApp.

First of all, the error s_{e} is estimated based on either or both of the N_{c} center runs and N_{F} fake factors. The error based on the center runs is calculated by:

$$s_{\mathrm{c}} = \sqrt{\frac{1}{N_c - 1}\sum_{i=1}^{N_c}(y_i - \bar{y_c})} \quad (3) \quad ,$$

and the error based on the fake factors:

$$s_{\mathrm{F}} = \sqrt{\frac{Y^{\mathrm{t}} P_{\mathrm{F}} Y}{N_{\mathrm{F}}}}, \qquad P_{\mathrm{F}} = X_{\mathrm{F}}(X_{\mathrm{F}}^{\mathrm{t}} X_{\mathrm{F}})^{-1} X_{\mathrm{F}}^{\mathrm{t}} \quad (4) \quad ,$$

where X_{F} is the matrix of the fake factors. In the case of Table 1, the rows and columns of X_{F} consist of those of fake1 and fake2. Considering both s_{c} and s_{F}, the overall error s_{e} is estimated by:

$$s_{\mathrm{e}} = \sqrt{\frac{(N_{\mathrm{c}} - 1)s_{\mathrm{c}}^2 + s_{\mathrm{F}}^2}{N_{\mathrm{c}} + N_{\mathrm{F}} - 1}} \quad (5) \quad .$$

Note that the further procedure is not performed if there are no further center runs or no fake factors.

Based on observation Y (the vector of obtained data $[y_1,\ldots y_{\mathrm{n}}]^{\mathrm{t}}$), the first-order regression model $\widehat{Y_{1\mathrm{st}}}$ is built. If the coefficient of a factor in the model $\widehat{Y_{1\mathrm{st}}}$ exceeds the error s_{e} multiplied by the t-value for an 80% confidence interval, the factor is included as one of the main factors. Then, the

first-order model $\widehat{Y_{\mathrm{main}}}$ is made by using these main factors. Afterwards, the active quadratic terms are selected. Here, the quadratic terms are comprised of the main factors. This is to say that if factors A and B are included in the main factors, AA, BB, and AB are the candidate terms. Figure 1 shows the model selection process in the case of the main factors A and B. To select the effective quadratic terms, the candidate term is evaluated by forward regression with the initial model $\widehat{Y_{\mathrm{main}}}$ adding the term whose p-value is minimum among the candidate. The same procedure is continued for the model with added quadratic terms until no more effective quadratic terms (with p-value <0.2) are found. Each time a quadratic term is added, the aforementioned AICc is calculated for the model, and the model with minimum AICc is selected as the final model. A more detailed explanation of this algorithm is given by reference[9].

Algorithm 1

1) Estimate error s_e (see equations (9) and (10)).
2) Calculate b_i (i=1,…,k), the coefficients of the first-order regression model $\widehat{Y_{1\mathrm{st}}}$=$\boldsymbol{bX}$.
3) If b_i / $t(N_c+N_F\text{-}1, 0.8)$> s_e, add i^{th} factor into main factors.
4) Make a first-order model $\widehat{Y_{\mathrm{main}}}$ using the main factors.
5) Make a candidate group of second-order terms comprised of main factors.
6) Select effective quadratic terms among the candidate group.
   I. Add quadratic terms by forward regression with the initial model $\widehat{Y_{\mathrm{main}}}$.
   II. Calculate AICc for the added model.
   III. Stop adding quadratic terms if the p-value of the added term exceeds 0.2.
   IV. Select the final model with minimum AICc
7) Make a multi-regression model by using main factors and effective quadratic terms.

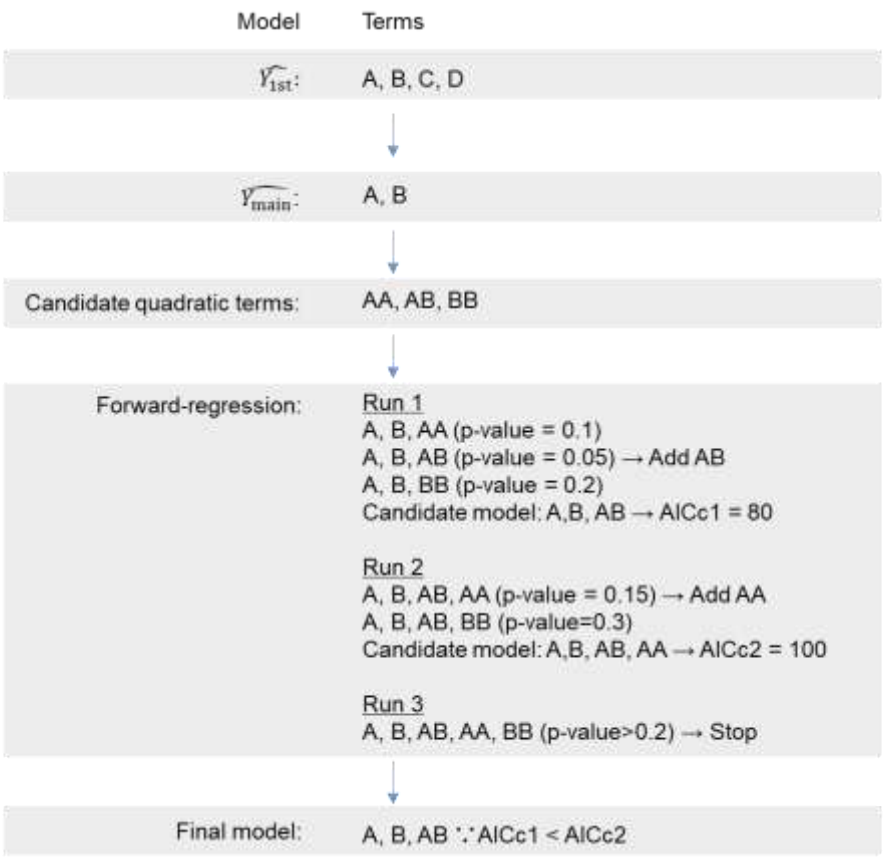

Figure 1 An example of the model selection procedure. In the example, A, B, C, and D are the initial factors. A and B are selected as "main factors." Among the candidate terms AA, AB, and BB, only AB is incorporated in the final model. The values of p-value and AICc are assumed just for explanation.