Comment: Some Statistical Concerns on Dimensional Analysis

Dennis K. J. Lin and Weijie Shen

Department of Statistics
The Pennsylvania State University
University Park, PA 16802
(*dkl5@psu.edu*; *wxs199@psu.edu*)

We would like to congratulate Albrecht et al. (2013)—henceforth ANAC—on their innovative and inspiring work on design of experiments for DA. The idea of incorporating DA into statistics could be considered as a breakthrough of interweaving physical knowledge and statistics. It is always valuable to formulate the scheme of combining scientific theories and information from data together: DA is a good example. DA has been a popular method to engineers since 1914 when it was first proposed. However, ANAC is clearly among those first few who notices its significance in the context of statistics. They have proposed a way of exploring the physical meanings of covariates rather than the numbers. This work is insightful and inspiring. Our DA experience is somehow different from ANAC. We will first discuss a typical DA example in meteorology to show the merits and limitations of DA from an engineering perspective, then we shall raise some related issues when DA is incorporated in statistical design and analysis.

1. A TYPICAL DIMENSIONAL ANALYSIS CASE IN METEOROLOGY

In meteorology, one of the most important areas is to study a number of boundary layer situations of the atmosphere, where the physical laws to govern the atmosphere's dynamics are clearly nonlinear. The planetary boundary layer, illustrated as the shaded area in Figure 1, is the lowest part of the atmosphere. The planetary boundary layer is categorized into different zones based on the local time and height (the x- and y-axis, respectively). Physical quantities, such as temperature, moisture, and flow velocity in this layer fluctuate rapidly because of its dynamics with the planetary surface. In the convective mixed layer where turbulence is driven by buoyancy and capped at a well-defined height, it is obvious that the convective velocity scale (w_*) and the depth of the boundary layer (z_i) are important scales for all quantities concerned (Stull 1988). Here, we consider developing an expression for vertical velocity variance ($Q_0 = w^2$), as a function of height ($Q_1 = z$) and other variables.

Figure 2(a) displays the scatterplot of w^2 and z, given the measurements from the Phoenix 78 experiment. The purpose of the Phoenix 78 experiment is to study the turbulence of convective boundary layer. During the experiment, the profiles of turbulence statistics from aircraft observations were recorded (see Young 1988, for details). From Figure 2(a), the dependence between w^2 and z is not obvious. This may be attributed to different magnitudes of length ($Q_2 = z_i$) and velocity ($Q_3 = w_*$). Using Buckingham's Π theorem, we can generate dimensionless variables $\pi_0 = Q_0/Q_3 = w^2/w_*^2$ and $\pi_1 = Q_1/Q_2 = z/z_i$. Figure 2(b) is the scatterplot of $\pi_0 = w^2/w_*^2$ and $\pi_1 = z/z_i$. A straightforward LOESS fitting (R Development Core Team 2011) gives the four curves (corresponding to four dates) in Figure 2(b). Each individual curve has the similar shape. We anticipate to build up one empirical model $\pi_0 = f(\pi_1)$, which will be able to describe the common character.

Assuming the function is of power-law form (with some boundary conditions), the empirical model can be built as: $\pi_0 = 1.554\pi_1^{1/2}(1 - 0.866\pi_1^{1/2})$ or $w^2/w_*^2 = 1.554(z/z_i)^{1/2}[1 - 0.866(z/z_i)^{1/2}]$.

It is interesting to compare the empirical model with the conventional model in meteorology (see Stull 1988): $\pi_0 = 1.8\pi_1^{2/3}(1 - 0.8\pi_1)^2$ or $w^2/w_*^2 = 1.8(z/z_i)^{2/3}(1 - 0.8z/z_i)^2$.

Figure 2(c) displays both models. The empirical model is close to the conventional model, but with a better fit. Moreover, they share a similar analytical form. This is a rather typical dimensional analysis (DA) example. Some lessons we have learned here are as follows:

1. DA is a popular and rather mature method in engineering.
2. When DA is used in engineering, typically only a small number of variables (say four or less) are included in the DA procedure.
3. DA can be employed to identify the appropriate dimensionless variables, but it does not provide the functional form of the relationships. Empirical results can be achieved by experiments and this is where statistics could be powerful.
4. For problems with a physical context, the linear model is often not appropriate because of boundary conditions and the fact that many variables are related via products, ratios, and powers.
5. The final empirical model should be tested with independent datasets, ideally over a range of different conditions.

TECHNOMETRICS, AUGUST 2013, VOL. 55, NO. 3
DOI: 10.1080/00401706.2013.804445

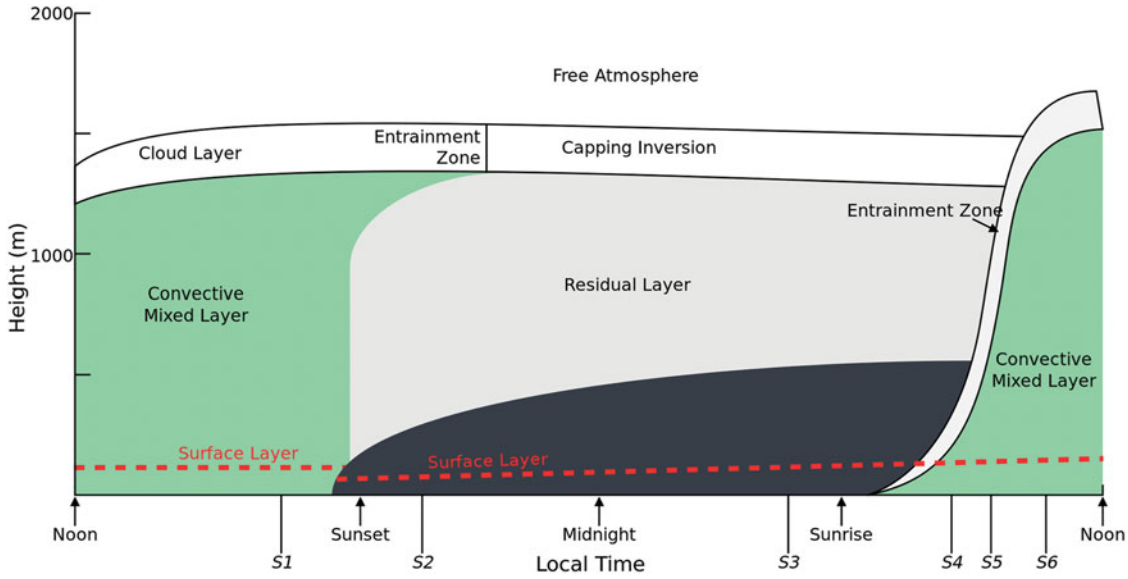

Figure 1. Illustration of planetary boundary layer. The online version of this figure is in color.

The above example shows that the combination of DA and statistical analysis will be of great value to engineering problems. However, in addition to modeling and prediction, the design of experiments and data collection for DA is lacking. Incorporating DA into design of experiments could be beneficial in terms of its efficiency and robustness. This is indeed the key contribution of ANAC's article.

2. IRREGULAR DESIGN SUPPORT AND LOG-TRANSFORMATION

The irregular design support is an important feature for transformed variables after DA. This leads to some problems in design of experiments, as stated by the authors (see Section 5.2). We would like to take a further look at why these problems surface.

In the cylinder drag experiment, the authors state (in Section 5.2) that the common basis quantities contribute to the irregular shapes of the support. While we agree that the irregular support is due to the common basis quantities, we note that the common basis quantities may not always lead to an irregular support. For example, ANAC use $\pi_1 = \rho U d/\mu$ and $\pi_2 = R_S/d$ because π_1 matches the definition of Reynolds number (Section 5.1). The support points in $\{\pi_1, \pi_2\}$, as displayed in Figure 3(a), are indeed irregular. However, direct derivation of DA will give the dimensionless quantity $1/\pi_1 = \mu/(\rho U d)$. The support points in $\{1/\pi_1, \pi_2\}$, as displayed in Figure 3(b), are fairly regular.

In essence, the irregularity (specifically, the hyperbolic design space in Figure 3(a)) occurs because the basis quantities have inverse relationships with different derived quantities. In this case, d is proportional to $\pi_1 = \rho U d/\mu$, but is reciprocal to $\pi_2 = R_S/d$. Under the power law, the relationship between basis quantities and derived quantities can only be either reciprocal or proportional, leading to the design space in either Figure 3(a) or 3(b).

Figure 3(c) displays the design space after taking logarithm transformation on both π_1 and π_2. It is clear that it is now a regular design space (as discussed in Section 5.2). In fact, such a logarithm transformation works well for both reciprocal and proportional relationships discussed above. However, are there any undesirable consequences by taking log-transformations? For example, how does the log-transformation affect the design space (e.g., uniformity and orthogonality are no longer valid) and the model building? Two issues related to log-transformation are (a) unbiased estimation and (b) design comparison, as described below.

Consider a power-law model with multiplicative errors, in which

$$\pi_0 = \prod \pi_i^{\beta_i} \cdot \epsilon.$$

Taking log on both sides results in a "linear model" of

$$(\log \pi_0) = \sum \beta_i (\log \pi_i) + \log \epsilon.$$

However, $E(\widehat{\log \pi_0}) = \log \pi_0$ does not imply $E(\widehat{\pi_0}) = \pi_0$, because $E(\exp(\widehat{\log \pi_0})) \neq \exp(E(\widehat{\log \pi_0}))$. Is there any direct approach possible for an *unbiased* estimate of $\widehat{\pi_0}$?

Also, Cartesian grid points in Q_i may not be Cartesian (or even uniform) in the transformed π_j. The criteria on these two different design spaces are actually comparing different aspects of performance. How can we measure and compare the performances of designs on the two totally different design spaces in a fair and reasonable fashion?

3. MISSING KEY VARIABLE

The authors stated in Section 4 that "the reverse error—omitting an independent variable when it is active—is usually fatal." Indeed, missing key variable(s) is always an important issue. However, we believe that this is a common issue for almost all scientific investigations. DA is merely one of them. An empirical model will also suffer when key variables are omitted. We thus disagree with the authors for the fourth advantage of empirical strategy in Section 4.1. How can we quantify the loss due to missing key variables for both DA model and

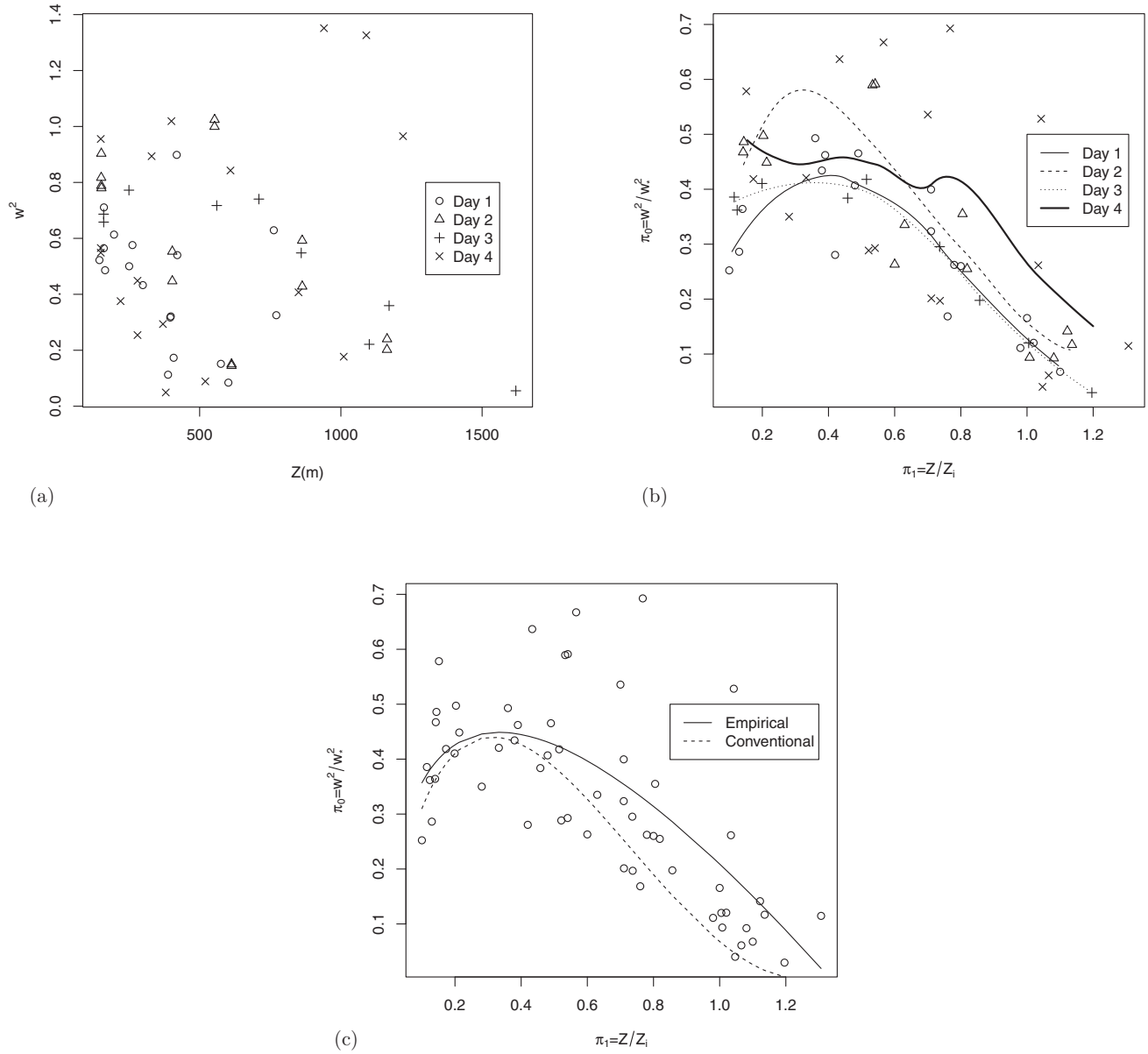

Figure 2. Scatterplots and estimates of Phoenix 78 data. (a) Original dataset. Different symbol stands for different date; (b) Transformed data and LOESS fits for four different dates; (c) Empirical model and conventional model based on the transformed dataset.

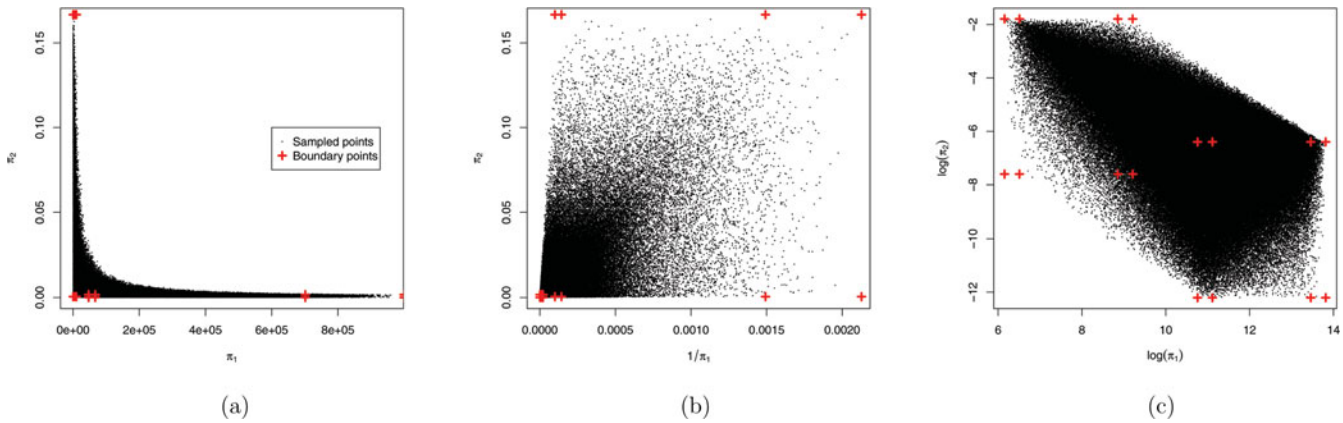

Figure 3. Design spaces for ANAC's cylinder drag example. Several transformations are presented. Uniformly sampled points and boundary points from domains of pre-DA variables are displayed under the above dimensionless transformations. (a) ANAC form; (b) Reciprocal form; (c) Log-transformation.

expectation–maximization (EM) model, so that we can have a legitimate comparison?

Here, we further investigate the consequence of missing key variables for DA procedure. Suppose the response quantity is Q_0 and the predictors are $Q_1, \ldots, Q_p$, and the basis dimensions of all quantities under consideration are $d_1, \ldots, d_k$, where d_i is one of the following seven fundamental physical dimensions: mass M, length L, time T, temperature Θ, electric current I (or charge Q), amount of substance mol, and luminous intensity $\mathrm{I_v}$.

Denote D_i as the set of basis dimensions, which constitutes Q_i for $i = 1, 2, \ldots, p$. Consider the loss when variable Q_a is missing. For an arbitrary basis dimension $d_j \in D_a$, we have the following scenarios.

Case 1: *d_j only appears in Q_a.*
If Q_a is observed, Q_a will be omitted after DA because it cannot be represented by other quantities. Thus, the missing of Q_a will not change the result of DA. There is no loss.

Case 2: *d_j appears in at least three quantities including Q_a (say, Q_a, Q_b, and Q_c).*
If Q_a is observed, select Q_b as the basis quantity. By DA, we will obtain two dimensionless variables: $\Pi_a(Q_a, Q_b)$ and $\Pi_c(Q_c, Q_b)$.
On the other hand, if Q_a is missing, the dimensionless variable will be $\Pi_c(Q_c, Q_b)$. The loss is only one variable, Q_a itself.

Case 3: *d_j only appears in both Q_a and Q_b.*
If Q_a is observed, one of Q_a and Q_b should be the basis quantity. Q_a and Q_b will combine in some way to cancel d_j to get a dimensionless variable $\Pi_a(Q_a, Q_b)$.
On the other hand, if Q_a is missing, Q_b is the only one having d_j in dimension and thus should be omitted after DA. The loss is two variables, Q_a and Q_b. This is the worst scenario.

As a result, only Case 3 leads to a deletion of other relevant quantities. By missing Q_a, the additional number of quantities deleted is at most the number of fundamental dimensions in D_a. This is exactly the case in the ball deformation example with $[E] = \mathrm{ML^{-1}\,T^{-2}}$ and $D_E = \{\mathrm{M,L,T}\}$. For the dimensions within D_E: M only appears in ρ and E, via $[\rho] = \mathrm{ML^{-3}}$; T only appears in V and E, via $[V] = \mathrm{LT^{-1}}$. So the missing of E results in missing of ρ and V. Generally, the total number of quantities deleted will be small.

Case 3 often occurs when one certain basis dimension d_j appears only in one variable Q_b (and the missing variable Q_a—of course, this is unknown to us). If no professional knowledge is available, our recommendation is to include a dimensional constant with dimension d_j so that we will not eliminate Q_b.

4. DEPENDENCY BETWEEN VARIABLES

Consider the questions: (a) will DA change the dependency structure between input variables and response? and (b) will these changes generate spurious features in model building? When transforming variables into dimensionless quantities, we multiply or divide the original variables with basis quantities by a power law. The correlations between transformed response and covariates are changed, indicating that the effects of variables are also changed. For illustration, consider the simplest example, with the response Y, the covariate X, and the basis quantity D. After DA, we have (Y/D) for the response and (X/D) for the covariate. We have the following four potential scenarios:

1. Y and X are uncorrelated; and (Y/D) and (X/D) are also uncorrelated.
2. Y and X are correlated; and (Y/D) and (X/D) are also correlated.
3. Y and X are correlated; but (Y/D) and (X/D) are uncorrelated.
4. Y and X are uncorrelated; but (Y/D) and (X/D) are correlated.

Here, we use "uncorrelated" instead of "independent" to avoid potential confusion. In Scenarios 1 and 2, it is clear that DA does not have impact on the dependency between X and Y.

In Scenario 3, suppose linear regression will yield a significant effect of X on Y, but an insignificant effect of (X/D) on (Y/D). That is, $(Y/D) = a + b(X/D)$ with b insignificant. We could conclude $Y = aD$ and thus D, instead of X, has a significant effect on Y. Because Scenario 3 typically occurs when D explains most of the dependency between Y and X, the conclusion of significant effect of D, in place of X, will not cause any problem.

For Scenario 4, as an illustrative example, consider the data in Figure 4 where the scatterplots and linear fits of response and covariate before and after DA are displayed. It can be seen that the effect of X is insignificant before DA but is significant after DA. A linear regression after DA, $(Y/D) = a + b(X/D)$, implies that $Y = aD + bX$, namely, Y and X are now correlated. This contradicts the fact and a spurious correlation is resulted. Note that Scenario 4 is the only case where DA reaches a contradictory conclusion. An efficient test is needed to identify whether Scenario 4 occurs. This deserves further studies in the future.

5. QUANTITIES IN DA PROCEDURES

In physics, there are many constants that also carry dimensions, such as the Boltzmann constant, the gravitational constant, and the speed of light. These constants (a.k.a., "parameters" in statistics) are to be estimated, not to be controlled. Thus, they cannot be treated as experimental variables. It is important that the constants relative to the physical background should also be included in the model before DA. Otherwise, the DA procedure may falsely rule out significant factors due to their dimensions.

A key step in DA (see Section 2.4) is to identify a complete, dimensionally independent subset of the input variables (as the chosen basis quantities). However, such a subset is not unique. Will a different choice of such a subset lead to a different consequence, especially in design stage? If so, what

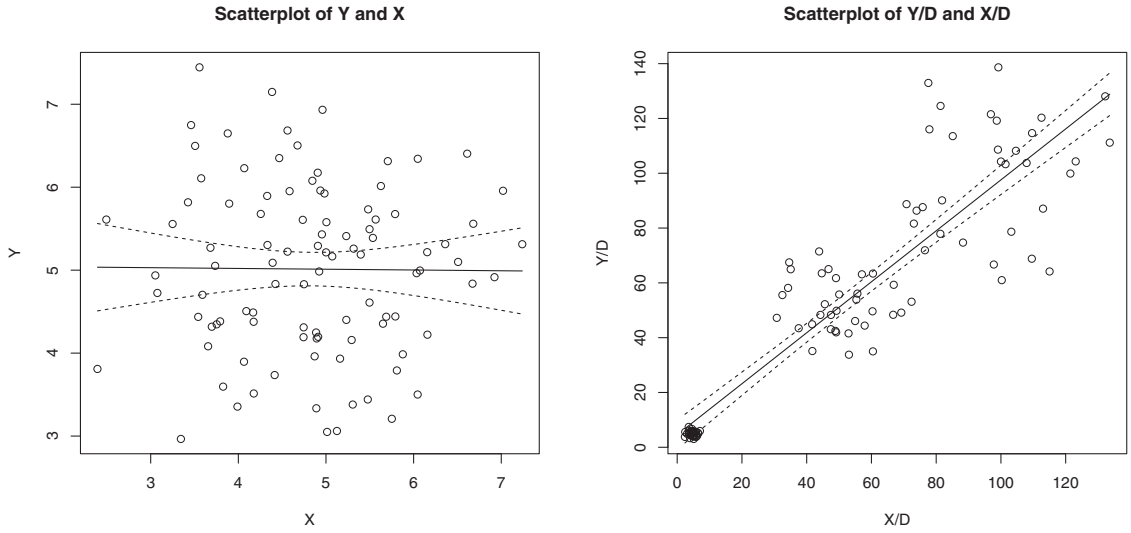

Figure 4. Different relationships before and after DA with linear fits and confidence intervals.

will be the "optimal" choice of the basis quantities and what will be an appropriate optimization criterion? Note that, in canonical engineering, basis quantities are usually the baseline scales of the quantities in the system; in our meteorology example, these are the velocity scale w_* and the length scale z_i.

One problem with the power-law form of the derived quantities is that the denominator can be zero. It is quite common to have zeros for physical quantities, resulting some dimensionless quantities to be undefined. This problem may be avoided by properly designing the experiments in certain situations., but does occur in general.

Our experience indicates that polynomial fitting may not be realistic in DA modeling. Unless there is strong (physical) support for any specific model (and consequently D-optimality can be defined), we believe that U-optimality is more appropriate here. A U-optimal design has more power, especially when a nonlinear model is employed. Furthermore, other design criteria may also be useful and worth exploring (for example, I-optimality).

6. CONCLUSION

Albrecht, Nachtsheim, Albrecht, Cook's article pertains a fresh perspective in combining DA and design of experiments. The authors introduce DA to statisticians and propose relative difficulties and solutions. They emphasize the use of experimental designs for the improvement of the DA procedure. Congratulations again for this outstanding work. We are grateful for this opportunity to be part of the discussion.

ADDITIONAL REFERENCES

Albrecht, M. C., Nachtsheim, C. J., Albrecht, T. A., and Cook, R. D. (2013), "Experimental Design for Engineering Dimensional Analysis," *Technometrics*, 55, 251–271. [281]

R Development Core Team (2011), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. [281]

Stull, R. B. (1988), *An Introduction to Boundary Layer Meteorology*, Dordrecht: Kluwer Academic Publishers. [281]

Young, G. (1988), "Turbulence Structure of the Convective Boundary Layer. Part I: Variability of Normalized Turbulence Statistics," *Journal of the Atmospheric Sciences*, 45, 719–726. [281]