| Statistica Sinica Preprint No: SS-2015-0377.R1 | |
|---|---|
| **Title** | Statistical Theories for Dimensional Analysis |
| **Manuscript ID** | SS-2015-0377.R1 |
| **URL** | http://www.stat.sinica.edu.tw/statistica/ |
| **DOI** | 10.5705/ss.202015.0377 |
| **Complete List of Authors** | Weijie Shen and<br>Dennis K. J. Lin |
| **Corresponding Author** | Weijie Shen |
| **E-mail** | jayshenwei@gmail.com |
| Notice: Accepted version subject to English editing. | |

Statistica Sinica

# Statistical Theories for Dimensional Analysis

Weijie Shen and Dennis K. J. Lin

*The Pennsylvania State University*

*Abstract:* Dimensional Analysis (DA) is a widely used methodology in physics and engineering. The main idea of DA is to extract dimensionless variables based on physical dimensions. Due to its capability in removing dimensional constraints and reducing the number of variables, its overlooked importance in statistics has been recognized recently. While its properties in physics have been well established, the fundamental statistical theories behind DA remain absent. Such theories are critical in integrating DA into statistical procedures. In this paper, we present a new statistical perspective on DA, which translates the essence of DA into statistical principles. The basis quantities are represented as linear-space bases, while the post-DA variables are formulated as maximal invariant statistics. The proposed statistical properties of DA, the sufficiency and completeness, guarantee the optimality of DA variables. An ocean wave speed example is presented to demonstrate DA methodology. A Meteorology example of planetary boundary layer problem is used to illustrate the proposed statistical properties in a practical context. The proposed representation reveals DA's structural compliance with statistical theories and encourages more appropriate statistical applications.

*Key words and phrases:* Completeness, dimensional reduction, invariance, sufficiency and transformation.

## 1. Introduction

Combining information from both scientific theories and experimental data has always been a challenging problem. Models derived purely from data perspective are often subject to questions on their interpretability, generality and control on sources of errors. It has been widely recognized that incorporating professional knowledge is useful in both guiding the development of valid scientific models and reducing potential sources of random errors. However, such incorporation is often *ad hoc*. A systematic framework is desirable for integrating scientific and empirical information. In this paper, we tackle on physical and engineering problems, and introduce dimensional analysis (DA) as a general approach to unify the information from physical dimensions.

Dimensional analysis (DA) is a variable extraction method that prevails in physics and engineering due to its applicability and effectiveness (see Sonin, 2001; Szirtes, 2007). The principal use of dimensional analysis is to deduce certain limitations and possible relationships among physical quantities from their physical dimensions. The method is of great generality and mathematical simplicity, (Bridgman, 1931). It often serves as a starting point for pilot studies and an ending point for model validation in physical problems. Literature shows its wide usage in various fields, such as Balaguer (2013) in Control Engineering, Islam and Lye (2007) in Hydrodynamics and even Grudzewski and Roslanowska-Plichcinska (2013) in Economics.

Plenty of physical and mathematical theories have contributed to the maturity of DA, supporting its practical applications. In physics, Buckingham (1914) set the foundation for DA. Monin and Obukhov (1954) specialized DA in meteorology, known as Monin-Obukhov similarity theory. In engineering, Zlokarnik (1991) provided a guide on DA and scale-up principles designed for chemical process. In mathematics, group theoretical representations were established by Cariñena, del Olmo, and Santander (1981) and Cariñena, del Olmo, and Santander (1985). An informal yet comprehensive mathematical formalization of DA was provided by Tao (2012). This is only a partial list of many examples.

However, the importance of DA was not recognized by statisticians until very recently. Little effort has been made to actively incorporate such professional knowledge in statistical analyses for long. Recent researches found that

the incorporation of DA in statistical design and analysis greatly increases efficiency and interpretability, as discussed in Albrecht, Nachtsheim, Albrecht, and Cook (2013), and Shen, Davis, Lin, and Nachtsheim (2014). Indeed, DA's physical origin provides an independent way to interpret and summarize the data in addition to statistical approaches. A comprehensive discussion on the combined use of DA in statistics is presented in Albrecht, Nachtsheim, Albrecht, and Cook (2013), Davis (2013), Lin and Shen (2013), Frey (2013), Jones (2013), Piepel (2013), and Plumlee, Joseph, and Wu (2013).

Although practical successes such as those in Albrecht, Nachtsheim, Albrecht, and Cook (2013) and Shen, Davis, Lin, and Nachtsheim (2014) suggest the potential efficacy of DA in statistical applications, the theoretical investigation of DA from a statistical point of view is absent. Compared with other disciplines, statistics seems to be left behind in supporting the formal incorporation of DA. The statistical essence of DA transformation, the critical assumptions and limitations, and how it affects statistical modeling are all inevitable issues to obtain valid analyses. DA is not yet another dimension reduction method; the beauty and value of DA lies in its ability to deduce key features that better characterize a system with intrinsic scaling structure. Treating DA merely as a data preprocessing method is both inefficient and unjustified. In this paper, we establish how DA extracts scale-free information and where the reduction comes from via a statistical perspective. We show its vector space and scaling structure, establish the DA variables as maximal invariant statistics, and derive the optimality of DA variables by their sufficiency and completeness.

The contribution of this paper is by no means to develop new statistical theories on invariant, sufficient or complete statistics, nor on the reduced DA model just because of the reduction. Our contribution is to be the first to connect and frame the unfamiliar DA principles into the rudimentary statistical terminologies, and to articulate the implication of DA transformation. It embeds physical principles into statistical modeling under a general setup, which potentially opens the gate towards a series of innovative methodologies tailored to DA. In fact, DA's fundamental connections with statistical theories further justify its applicability and compatibility in statistical problems. Practically, the developed theories help avoid improper use of DA and promote designs and analyses based

on dimensional constraints and sufficiency. It generalizes DA beyond engineering problems into general scaling systems. We also hope that this work sheds some light on the general problem of how to incorporate information and constraints from scientific background when learning from experimental data.

The rest of the paper is organized as follows. Section 2 introduces the definitions and typical procedures of DA with an illustrative example. Section 3 derives the statistical properties of DA: we represent the unit system as vector space and basis quantities as the basis vectors; define the scaling group of unit changes, and establish DA variables as maximal invariant statistics; finally derive the sufficiency and completeness for DA variables in a general setup. In Section 4, a study of planetary boundary layer problem is presented, focusing on the practical realization of the proposed properties. Issues on the validation of the model and the testing of DA assumption are also discussed. Section 5 provides concluding remarks.

## 2. Dimensional Analysis

### 2.1. Background

In physics, dimension refers to the physical type of a quantity. Based on SI system for classical physics, there are seven fundamental physical dimensions, namely length [$\mathbf{L}$], mass [$\mathbf{M}$], time [$\mathbf{T}$], electrical current [$\mathbf{I}$], absolute temperature [$\mathbf{\Theta}$], amount of substance [$\mathbf{N}$] and luminous intensity [$\mathbf{J}$]. Other physical dimensions can be expressed in terms of these fundamental physical dimensions, and they are called derived dimensions. For example, speed has the dimension length per time [$\mathbf{LT^{-1}}$]. Siano (1985a,b) extends the dimensions above and treats components of vector quantities as different dimensions given a coordinate system. Also in practice, dimensions from separate subsystems are also considered different. Therefore, the practical size of independent dimensions can be much larger than seven.

A measurement system defines units for each dimension. The magnitude of a quantity is expressed as a denominate number: a real number multiple of the unit of measure. In SI system of units for example, we measure length by meters and time by seconds and so on. Other derived physical dimensions are measured accordingly: speed can be measured by the unit meters per second.

Table 1: Dimensions of variables in ocean waves example.

| Variables | Description | Dimensions | SI units | $r_{i,Length}$ | $r_{i,Mass}$ | $r_{i,Time}$ |
|---|---|---|---|---|---|---|
| $v$ | Wave speed | $\mathbf{LT}^{-1}$ | $m/s$ | 1 | 0 | -1 |
| $g$ | Gravity constant | $\mathbf{LT}^{-2}$ | $m/s^2$ | 1 | 0 | -2 |
| $\lambda$ | Wavelength | $\mathbf{L}$ | $m$ | 1 | 0 | 0 |
| $\rho$ | Water density | $\mathbf{ML}^{-3}$ | $kg/m^3$ | -3 | 1 | 0 |
| $H$ | Sea depth | $\mathbf{L}$ | $m$ | 1 | 0 | 0 |

Generally, the magnitude of a physical quantity is characterized by its relative magnitude to the commonly recognized units. Measuring quantities with inappropriate units leads to undefined multipliers. For example, measuring length by seconds is undefined, and measuring area by meters results in infinity (it should be measured by square meters). The comparison of the magnitudes of two quantities is also done through the relative magnitude. It is inappropriate to compare two quantities with different dimensions.

### 2.2. Illustrative example

Understanding the speed of ocean waves is crucial for the prediction of catastrophe like tsunami. There are multiple sources of driving force that generate wave motions on the sea surface, such as the wind, gravity and rotation of the earth. It is a complicate system whose analytical behavior may not be tractable. Gravity wave is the wave whose restoring force is the gravity of the earth. Suppose the phase speed of the gravity waves ($v$) is of main interest. Our predictors are the gravity constant of the earth ($g$), the wavelength ($\lambda$), the density of water ($\rho$), and the depth of water ($H$). Assuming $v = f(g, \lambda, \rho, H)$, our goal here is to estimate the function $f$.

Table 1 shows the physical dimensions of the variables. By conducting DA, two dimensionless variables $\pi_v = v/\sqrt{g\lambda}$ and $\pi_H = H/\lambda$ are derived. DA claims that the original model $v = f(g, \lambda, \rho, H)$ can be rewritten as $\pi_v = h(g, \lambda, \rho, \pi_H) = h(\pi_H)$, where $h$ is the function to be estimated. Given $\pi_H$, $g, \lambda, \rho$ should not be in the function $h$ due to dimensional homogeneity principle. The number of variables is reduced from 4 to 1. Note that by $\pi_v = v/\sqrt{g\lambda}$ and $\pi_H = H/\lambda$, the DA modeling function is in fact $v = \sqrt{g\lambda} \cdot h(H/\lambda)$.

As given in Socha (2007), the wave speed relationship can be approximated

analytically by $v = \sqrt{g\lambda/2\pi} \times \sqrt{\tanh(2\pi H/\lambda)}$ (considering the case of gravity wave). Compared with the DA models, the true function $h$ has the form $h(x) = \sqrt{(\tanh(2\pi x)/2\pi)}$. Thus, DA helps us remove the nuisance variable $\rho$, and reduce the number of variables to 1 when estimating $h$. Furthermore, the dimensionless ocean depth $\pi_H = H/\lambda$ actually characterizes the feature of ocean waves. When in the deep water, $H >> \lambda$, $\pi_H >> 1$. $\tanh(2\pi\pi_H) \approx 1$ and $v \approx \sqrt{g\lambda/2\pi}$, mainly depends on the wavelength; while in the shallow water (as along the coastline), $H << \lambda$, $\pi_H << 1$. $\tanh(2\pi\pi_H) \approx 2\pi\pi_H$ and $v \approx \sqrt{gH}$, mainly depends on the depth of the water. In short, DA also induces variables and models with better interpretability.

### 2.3. DA principles

The principle of DA is based upon the fact that a physical law must be independent of units used in the measurement, because units are merely a systematic way of recording the physical phenomena. Any meaningful physical equation (or inequality) characterizing the physical laws should remain correct (or incorrect) regardless of the units used on both sides. Otherwise, changing units would lead to contradictory observations.

A foundational theorem of DA is the Buckingham's Π-theorem (Buckingham, 1914). A collection of dimensions is called (a) "independent", if each of them cannot be represented/derived by other dimensions in the collection; and (b) "representable", if they can represent/derive the dimensions associated with any other variables of interest in the experiment. The Buckingham's Π-theorem states that a physically valid equation involving $n$ variables of interest can be reduced to an equation with $p = n - k$ variables, where $k$ is the size of the subset of variables whose dimensions form an independent and representable collection. We call these $k$ variables as basis quantities, as they constitute a basis in terms of dimensions. Note that the size $k$ is uniquely defined while the set of basis quantities is not. Dimensional analysis provides a scheme to select basis quantities and transform the other variables into dimensionless (as will be described below). Equivalently, it is also a scheme to generate dimensionless variables, where each of them is not a function of the others.

### 2.4. DA methodology

A typical procedure of DA in a statistical problem, that is widely used in existing literature, can be formulated as follows. Suppose $Y$ is the response of interest and $X_1, \cdots, X_n$ are potential predictors. $X_1, \cdots, X_n$ can be a mixture of continuous, categorical and discrete variables. For simplicity, we assume $Y, X_1, \cdots X_n$ are positive continuous physical quantities that are bounded in probability. Other types of quantities such as categorical variables and negative values are possible with DA as well. Usually, categorical variables are treated as dimensionless and discrete variables are determined based on their physical dimensions. Physical variables can also have negative values with proper dimensions such as the displacement in harmonic oscillation.

Statistically, the prediction problem can be formulated by a multivariate probability distribution on $(Y, X_1, \cdots, X_n)$ jointly. The conditional distribution of $(Y|X_1, \cdots, X_n)$ is often of main interest. According to the mean squared error criterion, the conditional mean is a good predictor for $Y$. In a regression setting, $E(Y|X_1, \cdots, X_n)$ is modeled as $f(X_1, \cdots, X_n)$. A typical DA modeling found in previous literature takes the following steps, as shown in the previous example of ocean wave speed.

1. Identify the dimensions of all variables involved.

2. Choose the basis quantities such that their dimensions are independent and representable. Denote them by $X_1, \cdots, X_k$.

3. By representability of basis quantities, transform other variables $(Y, X_{k+1}, \cdots, X_n)$ into dimensionless $(\pi_Y, \pi_{X_{k+1}}, \cdots, \pi_{X_n})$ by power law using basis quantities.

4. Rewrite the modeling function as $\pi_Y = h(X_1, \cdots, X_k, \pi_{X_{k+1}}, \cdots, \pi_{X_n}) = h(\pi_{X_{k+1}}, \cdots, \pi_{X_n})$, where $X_1, \cdots, X_k$ are irrelevant because of the independent property of basis quantities. The total number of variables is reduced by $k$.

In the context of ocean wave speed example, the first step of DA is to identify the physical dimensions of variables $v$, $g$, $\lambda$, $\rho$ and $H$. The fundamental dimensions involved in the system are length, mass and time. The dimensions of related variables are presented in Table 1. The second step is to choose the basis quantities. Here, $g$, $\lambda$, $\rho$ are selected. The third step is to transform other

variables into dimensionless using basis quantities. For example, $\pi_v = v/\sqrt{g\lambda}$ and $\pi_H = H/\lambda$ are the dimensionless counterparts. The final step is to rewrite the function of interest: $\pi_v = h(g, \lambda, \rho, \pi_H) = h(\pi_H)$, where $h$ is the function to be estimated. This is in fact $v = \sqrt{g\lambda} \cdot h(H/\lambda)$.

There are several important issues with this DA modeling framework. Firstly, the choices of the basis quantities and the dimensionless variables are not unique. One can always multiply two dimensionless variables and get a third one. In practice, specialists and technicians may have a list of commonly used dimensionless variables in their discipline with specific physical meanings. An alternative is to select variables best explain the systems or the trends in terms of parsimony and significance. Secondly, variables may be ruled out due to the lack of presentation in dimensions (see Albrecht, Nachtsheim, Albrecht, and Cook, 2013). This can be a good feature when it is a valid reduction, or a bad one when the dropped variable is known to be useful. For example, in step 2, if the independent set of basis quantities inevitably include the response $Y$, then $Y$ will be excluded from the model after DA, which is not reasonable. In such cases, additional variables or dimensional constants are recommended to supply the missing dimensions. If not available, relevant basis quantities should be maintained in the model. We think the reverse statement is particularly true and useful: if the basis quantities are highly significant, there is high possibility of missing key variables. That is, significant basis quantities can be good indicators of missing key variables. Further discussions about relevant statistical issues can be found in Lin and Shen (2013) and other discussion papers to Albrecht, Nachtsheim, Albrecht, and Cook (2013).

## 3. Statistical Properties of Dimensional Analysis

The prevailing applications of DA lead to full development in the physical theories and properties supporting it. See Sonin (2001) and Szirtes (2007). However, the statistical theory for DA remains primitive. First, the implication of the transformation has rarely been perceived from a statistical point of view. Second, the evaluation of DA assumptions is *ad hoc* and beyond statistical justification. Finally, the absence of adapted modeling and analysis techniques for post-DA variables impedes the development and extension of DA. In this section, we take

the first step to address the above issues by introducing statistical properties of DA and their implication on statistical modeling.

In the DA procedure, information from the variables can be classified into information from the basis quantities and information from the transformed dimensionless quantities. In the following, we first investigate properties of the basis quantities. It turns out that the dimension space is isomorphic to the vector space with basis quantities being the basis vectors. Second, we investigate properties of the transformed dimensionless quantities. It is shown that they are maximal invariant statistics subject to a scaling group of unit changes. Third, since invariant statistics are not unique, a natural question is which one is "optimal". We show that the dimensionless variables produced by DA are both sufficient and complete, and are thus proven to be the optimal invariant statistic. We therefore solve the proposed issues: (1) DA transformation is "principle-driven PCA" that reduces through vector subspace; (2) explicit assumptions are given for DA's invariance and sufficiency property, that are verifiable in statistical problems; (3) we extend the family of DA models based on DA's invariance and sufficiency structure. For simplicity of the presentation, all proofs are given in the Appendix A.

### 3.1. Linear Space Representation

Here, we show that the collection of all dimensions forms a linear space. The basis quantities are interpreted as the basis vectors in the linear space context. Such a structure has been discussed in physical and mathematical literatures (such as Taylor et al. 2008; Drobot 1953; and Cariñena et al. 1985), but these are not straightforward from a statistical point of view.

The physical principle (of "absolute significance of relative magnitude") leads to the fact that physical dimensions can only be generated by fundamental dimensions through power law (Bridgman, 1931). Let $e_1, \cdots, e_m$ be fundamental dimensions, and $\mathcal{F} = \{\mathbf{D} = e_1^{d_1} \cdots e_m^{d_m} : d_1, \cdots, d_m \in \mathbb{Q}\}$. $\mathcal{F}$ is the collection of all dimensions derived from the fundamental dimensions $e_1, \cdots, e_m$. We have the following theorem:

**Lemma 1.** $(\mathbb{Q}, \mathcal{F})$ *is a vector space.*

Lemma 1 shows that the mapping from dimension $\mathbf{D}$ to vector $(d_1, \cdots, d_m)$

is isomorphic. It maps multiplication and scalar power on dimensions in the usual sense into addition and scalar multiplication operators of linear space respectively. Since $V = \{(d_1, \cdots, d_m) : d_1, \cdots, d_m \in \mathbb{Q}\}$ is the $m$-dimensional rational vector space, $\mathcal{F}$ is also an $m$-dimensional vector space, with scalars in $\mathbb{Q}$.

From the linear space interpretation, the basis quantities are merely the analogy of basis for linear space, shown in the following. In a statistical setting, suppose $X_1, \cdots, X_n$ are variables with dimensions $\mathbf{D_1}, \cdots, \mathbf{D_n}$. Denote $e_1, \cdots, e_m$ are relevant fundamental dimensions. Then $\mathbf{D_i} = \prod_{j=1}^{m} e_j^{d_{ij}}$. Therefore, each dimension $\mathbf{D_i}$ can be coded by a vector $v_i = (d_{i1}, \cdots, d_{im})$. For example, suppose $e_1$ is time; $e_2$ is length, then speed (length/time=$e_1^{-1}e_2$) can be coded as $(-1, 1, 0, \cdots, 0)$ and area (length$^2$ = $e_2^2$) can be coded as $(0, 2, 0, \cdots, 0)$. The requirements of independence and representativity for the basis quantities can be interpreted as the same two requirements for the basis vectors in the vector space $V$, as shown below.

Let $D = (d_{ij})$ be the dimensional matrix whose $(i, j)$ element is $d_{ij}$ mentioned above. $k = rank(D)$. Without loss of generality, suppose the first $k$ rows $(v_1, \cdots, v_k)$ are linearly independent and the other rows can be linear represented as $d_{tj} = \sum_{i=1}^{k} b_{ti} d_{ij}$ for $t = k+1, \cdots, n$; $j = 1, \cdots, m$. Then, $(X_1, \cdots, X_k)$ can be taken as basis quantities. These basis quantities are dimensionally independent: if $\mathbf{D}_1 = \mathbf{D}_2^{\alpha_2} \cdots \mathbf{D}_k^{\alpha_k}$, then $v_1 = \alpha_2 v_2 + \cdots + \alpha_k v_k$. $v_1, \cdots, v_k$ are linearly independent, so $\alpha_2, \cdots, \alpha_k$ does not exist. Similar statements apply for $\mathbf{D}_2, \cdots, \mathbf{D}_k$. These basis quantities are also dimensionally representable: $\mathbf{D}_t = \mathbf{D}_1^{b_{t1}} \cdots \mathbf{D}_k^{b_{tk}}$, for $t = k + 1, \cdots, n$. This concludes that the size of basis quantities equals to $k = rank(D)$. It is clear that basis vectors for a linear space is not unique. Correspondingly, the basis quantities are not unique.

Due to representativity, $X_{k+1}, \cdots, X_n$ can be transformed into dimensionless as $\pi_t = X_t X_1^{-b_{t1}} \cdots X_k^{-b_{tk}}$, for $t = k + 1, \cdots, n$. Suppose the original modeling function is $f(X_1, \cdots, X_n) = 0$. Then it can always be rewritten as $g(X_1, \cdots, X_k, \pi_{k+1}, \cdots, \pi_n) = f(X_1, \cdots, X_k, \pi_{k+1} \prod_{i=1}^{k} X_i^{b_{k+1,i}}, \cdots, \pi_n \prod_{i=1}^{k} X_i^{b_{n,i}}) = 0$.

The Buckingham's $\Pi$-theorem indicates that the scales of the coordinate system $(X_1, \cdots, X_k)$ do not contribute to the physical phenomena. It is the "relative

magnitude", i.e., the shape, that matters, which is summarized by dimensionless variables $(\pi_{k+1}, \cdots, \pi_n)$ DA generates. Therefore, $g(X_1, \cdots, X_k, \pi_{k+1}, \cdots, \pi_n) = g(\pi_{k+1}, \cdots, \pi_n) = 0$.

In this representation, DA is closely related to PCA after variables take log transformation. DA constraints variables into a linear subspace by the dimensional requirements, which is similar when we keep $n - k$ largest eigenvalues in PCA and set the rest $k$ to be 0. In the canonical procedure, the basis quantities we drop in the end correspond to the eigenvectors whose eigenvalues are set to 0. From this perspective, we might call DA as "principle-driven PCA", and clearly we can do better by combining data-driven PCA into DA instead of hand-picking basis quantities to drop in the end.

### 3.2. Invariance and Equivariance

In this section, a statistical interpretation of dimensionless variables is established: the dimensionless variables are maximal invariant statistics to scale transformation in fundamental dimensions. In the well established statistical decision theory (see Lehmann and Casella, 2003; Eaton, 1989), invariant decisions are desirable: decisions, such as hypothesis testing results, should not not be influenced by simple transformations on the data. Theoretically, decision $a$ is called completely invariant if it satisfies $a(X) = a(g(X))$, where $X$ is the observations, $g$ is any transformation from a group $\mathcal{G}$. In other cases, equivariant decisions are appropriate: decisions, such as point estimates, should scale in a proper and meaningful way reflecting the transformations on the data. Theoretically $a(X) = \bar{\bar{g}}(a(g(X)))$, where $\bar{\bar{g}}$ is the appropriate transformation on the decision space, also forming a group $\bar{\bar{\mathcal{G}}}$. We believe that the principle of DA (i.e., physical phenomena should be invariant to measurement system), fits well into the context of invariant decisions. Complete invariant decisions are dimensionless; while equivariant decisions are associated with appropriate dimensions. To model a physical system that is intrinsically free from the physical dimensions, it is preferable to implement an invariant probabilistic procedure which does not depend on the units used. We define an invariant probability model as one that satisfies $P_{\bar{\bar{g}}(\theta)}(X' \in g(A)) = P_\theta(X \in A)$, where $X' = g(X)$ is the transformed variables, $A$ and $g(A)$ are some event before and after transformation,

$\theta$ is the parameter and $\bar{g}$ is the corresponding transformation in the parameter space. We define an invariant probabilistic procedure as one that satisfies $L(\bar{\bar{g}}(a), X') = L(a, X)$, where $L$ is some invariant loss function, $a$ is the decision and $\bar{\bar{g}}$ is the corresponding transformation in the decision space.

To understand the invariance structure of DA (especially the transformation of event $g(A)$), it is necessary to investigate the physical dimensions and measurement systems in terms of measures and probabilities. The real line and usual Lebesgue measure should be adjusted by associating them with the unit used, and we call them "*physical Lebesgue measure*". Here a quantity refers to a quantifiable feature of a subject. It stands for an abstract magnitude. A physical Lebesgue measure can be imposed on it to quantifies its relative magnitude to the unit used, acting like a ruler. By the physical Lebesgue measure, the abstract magnitude is mapped into a real value, just as the read on the ruler. Note that different real values can be achieved by imposing different physical Lebesgue measures (Lebesgue measures associated with different units), but the abstract magnitude does not change. We call a collection of physical Lebesgue measures for each dimension a *measurement system*.

For example, define the physical Lebesgue measure $\lambda_u$ with unit $u$ on a unit quantity interval $[0, 1]u$ as $\lambda_u([0, 1]u) = \lambda_u([0, 1u]) = 1$. Define the (measured) value of an abstract quantity $Q$ using unit $u$ as $\lambda_u([0, Q])$. The mapping $\lambda_u$ returns the multiplier in terms of the units for each abstract quantity. Generally, $\lambda_S(E)$ denotes the measurement of quantity interval $E$ using appropriate units in the measurement system $S$, and is abbreviated as $\lambda(E)$. In the previous example, if a physical Lebesgue measure with a different unit $10u$ is used, then by definition $\lambda_{10u}([0, 1]u) = \lambda_{10u}([0, 0.1]10u) = 0.1$. Therefore, when unit changes occur in the measurement system, the physical Lebesgue measure will change correspondingly. So will the measured values of physical quantities. It turns out that, from scale changes in unites, both the induced changes on Lebesgure measure and the induced changes on quantity values form scaling groups. This can be summarized in the following lemma.

**Lemma 2.** *Suppose unit change $T_a$ transforms fundamental units $u_i$ into $u_i' = a_i u_i$. Then all unit changes $\mathcal{T} = \{T_a : a_i > 0, a_i \in \mathbb{R}, a = (a_1, ..., a_m)^T\}$ form a scaling group. The induced changes on Lebesgue measure $\tilde{\mathcal{T}} = \{\tilde{T}_a : a_i >$*

$0, a_i \in \mathbb{R}, a = (a_1, ..., a_m)^T\}$ *is also a scaling group. The induced changes on the measured values of physical quantities* $\hat{\mathcal{T}} = \{\hat{T}_a \circ \cdots \circ \hat{T}_a : a_i > 0, a_i \in \mathbb{R}, a = (a_1, ..., a_m)^T\}$ *is also a scaling group.*

We define a quantity $Q$ to be *dimensionless*, if its value does not depend on the measurement system. The dimension of a dimensionless quantity is the zero vector under the linear space representation of dimensions. Its unit is $w = \prod_{i=1}^m u_i^0 = 1$. Thus its numerical value stays the same $\tilde{T}_a(\lambda_S)(Q) = \lambda_S(Q)$ for any induced change $\tilde{T}_a$ on measure $\lambda_S$ with measurement system $S$ due to unit change $T_a$. Therefore, if $Q$ is dimensionless, its value serves as an invariant statistic to the scale group of unit changes.

Based on the first principle, physical phenomena are consistent regardless of the measurement systems. Therefore, the probability of an event should not depend on the units used. This can be stated as below.

**Lemma 3.** *The probability of an event is dimensionless.*

Lemma 3 suggests DA applications in logistic regression $\ln(p) - \ln(1-p) = \beta X$ where the left hand side of the regression model is log odds and responses are categorical. In this case, it is natural to constrain the right hand side of the regression $\beta X$ to be dimensionless to reflect dimensionless log odds and dimensionless probability.

We define two quantities to have same dimensions, if the ratio of their values is dimensionless. We can conclude in the following, that the value of any measurable quantity set under a physical Lebesgue measure shares the same physical dimension with the quantity itself. In other words, any measurable set for a certain quantity has the same unit.

**Lemma 4.** *Suppose $X$ is a quantity whose dimension is $D$; $E$ is a measurable set of values $X$ can take, then $E$ also has dimension $D$.*

In principle, random variables are usually measured values of an abstract physical quantity by certain measurement system. Unit changes in the measurement system induce a scale change in the random variables. (The variables invariant to the unit changes are dimensionless.) Meanwhile, the probability of the physical events should be invariant to this change. Thus, we prefer a probabilistic model/procedure that compensates the changes in variables. Such an

appropriate modeling measure will guarantee an invariant risk measure provided an invariant loss function, leading to appropriate invariant or equivariant decisions. There are several ways to generate such a probabilistic model. One way is to base the model on equivariant statistics with equivariant parameters.

However, in practice, the equivariant structure of the parameters usually depends on the context. It is often tedious to build a special model (equipped with equivariant parameters) and the corresponding analysis techniques for the specific dimensions of interest case by case. Besides, the equivariant counterparts are difficult to derive, and may involve too many parameters to be practical (for example, in random forest and many other data mining procedures). Furthermore, equivariant parameters have physical dimensions. They complicate the interpretation and extrapolation of the model. These parameters share information about the scales implicitly, and usually are not good characteristics of the physical features of the system that can be compared between platforms. Thus, we resort for another way to construct the model that is applicable to all kinds of scale-change structures in the dimensions. We build it upon the dimensionless variables, i.e. the invariant statistic. Then, the corresponding parameters become completely invariant to unit changes, and so is the distribution. The above complications of the equivariant structures are avoided. We define the probability distribution/model as *invariant* if its form is invariant to the group transformation. Specifically, we call the probabilistic model as *dimensionless model* if its form is invariant to the unit changes in the measurement system.

Such invariant models are especially desirable to practitioners. For arbitrary models on physical quantities, extrapolation to different units or different ranges of variables will be risky. If the considered model is invariant to the joint scaling of variables defined by the measurement system, extrapolation may be achieved. Scaling of original variables may result in DA variables still remain in the experimental domain. Therefore, dimensionless models are of great interest to engineers, particularly in the fields of reliability engineering and accelerated life testings.

Note that DA variables $\pi_{X_{k+1}}, \cdots, \pi_{X_n}$ are dimensionless and therefore invariant statistics. Furthermore, the following lemma shows that DA variables are maximal invariant statistics.

**Lemma 5.** *Suppose $M$ is the DA transformation that satisfies $M(X_1, \cdots, X_n) = (\pi_{X_{k+1}}, \cdots, \pi_{X_n})^T$, where $\pi_t = X_t X_1^{-b_{t1}} \cdots X_k^{-b_{tk}}$ for $t = k+1, \cdots, n$. Then $M$ is maximal invariant over unit change scaling group $\hat{\mathcal{T}}$. $(\pi_{X_{k+1}}, \cdots, \pi_{X_n})^T$ is maximal invariant statistic.*

Therefore, any invariant (dimensionless) statistics is a function of the DA variables. If a model is built upon the dimensionless variables, it suffices to construct the model via DA variables, in order to reduce the parameter space to completely invariant parameters.

### 3.3. Equivalence to Original Models

Invariant statistics are useful in building invariant models. However, there are possibly many other invariant models built upon original variables, instead of invariant statistics. It is possible that the selected invariant statistics lose necessary information about the original variables and thus lead to models that are not equivalent to those on the original statistics. In this section, we directly show that the dimensionless variables derived through DA are sufficient statistics if we consider dimensionless models, and are also complete to the family including all dimensionless models. For this family, the minimal sufficiency of DA variables is proved and the maximal reduction is achieved.

In order to study the probabilistic models on physical quantities with dimensions, it is necessary to investigate the dimensions of the cumulative distribution function (c.d.f.), probability mass function (p.m.f.) and probability density function (p.d.f.). Since the c.d.f. and p.m.f. are actually probability of an event, they are dimensionless (see Lemma 3). We have the following lemma for the p.d.f. of continuous variables.

**Lemma 6.** *Consider the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Suppose random vector $(X_1, \cdots, X_n)^T$ follows a continuous distribution $F$ with probability density function $f$ with respect to Lebesgue measure $\lambda$, and $X_i$ has dimension $D_i$, then $f(X_1, \cdots, X_n)$ has dimension $(\prod_{i=1}^n D_i)^{-1}$, for each given $\omega \in \Omega$.*

**Example 1.** (Normalization) If random variable $X$ has dimension $D$, then its expectation $\mu = E(X)$ has dimension $D$, its variance $var(X)$ has dimension $D^2$ and its standard deviation $\sigma$ has dimension $D$. Thus, the normalization $(X-\mu)/\sigma$ is dimensionless.

In general, the $k$th moment of $X$ has dimension $D^k$. Here, we retrieve our intuition about the commonly used statistics of random variables with physical dimensions. The expectation and standard deviation of a random variable should maintain the same scale as itself. (It is easy to show that the sample expectation and standard deviation also have the same dimension.) The standardization/normalization of a random variable $(X - EX)/sd(X)$ is dimensionless, which leads us to the great usage of z score and t score: they can be compared across scenarios with different scales but have a common distribution. Similar applications include correlation coefficient and R squares. The concept is very much related to the invariance of scaling group. It is also straightforward to prove that the method-of-moments estimators have the same dimensions as parameters estimated. However, the maximal likelihood estimators depend on the chosen models and may not share the same dimensions.

**Example 2.** (Power-law form) If random variable $X$ has dimension $D$, $f(X)$ is a valid analytic function, then $f(x) = ax^b$ with dimensionless constants $a, b$. Conversely, if $f(X)$ is a valid analytic function but not a power-law form, such as $X + X^2$ and $e^X$, then $X$ should be dimensionless.

This can be derived as follows. The analytic function $f$ has Taylor expansion: $f(X) = f(0)X^0/0! + f^{(1)}(0)X/1! + f^{(2)}(0)X^2/2! + \cdots$. For all $r \geq 0$, $f^{(r)}(0)$ is dimensionless because it is a derivative of an analytic function that does not involve dimensions. However, the power terms $X^0$, $X$, $X^2$, $\cdots$ have distinct dimensions 1, $D$, $D^2$, $\cdots$. In order to make the infinite summation valid, only one derivative $f^{(r)}(0)$ is nonzero. That is $f^{(r)}(0) = 0$ except when $r = b$, which leads to $f(x) = ax^b$. The converse-negative counterpart is obvious.

In the above example, it is assumed that the only part in $f$ having dimension is its argument $X$. We call this type of functions *numerical functions*. It is shown that the power function is the only valid univariate numerical function for variables having dimensions. Similar conclusions can be drawn for numerical functions of several dimensionally independent quantities. For functions of arbitrary physical quantities, DA should be used to derive all the possible forms.

For parametric cases, dimensionless models refers to the models with dimensionless parameters. The following theorem shows that DA variables are sufficient for the family of dimensionless models. Consider the following two assumptions:

Assumption 1: Consider the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Suppose positive random variables $Y, X_1, \cdots, X_n$ have respective dimensions $D_0, D_1, \cdots, D_n$; and suppose $X_1, \cdots, X_k$ are the $k$ basis quantities and $\pi_0, \pi_{k+1}, \cdots, \pi_n$ are dimensionless transformations of $Y, X_{k+1}, \cdots, X_n$ by dimensional analysis procedure; i.e., $\pi_0 = \pi_0(Y, X_1, \cdots, X_k)$, $\pi_{k+1} = \pi_{k+1}(X_{k+1}, X_1, \cdots, X_k)$, $\cdots$, $\pi_n = \pi_n(X_n, X_1, \cdots, X_k)$.

Assumption 2: Suppose $Y|X_1, \cdots, X_n$ follows a distribution with probability density function $f(y; X_1, \cdots, X_n; \theta)$, where $\theta$ is an unknown identifiable parameter. $(X_1, \cdots, X_n)^T$ follows a prior distribution with probability density function $p(x_1, \cdots, x_n)$, and are independent of $\theta$.

**Theorem 1.** *(Sufficient Dimension Reduction for Parametric Case): Under Assumptions 1 and 2, $\theta$ is dimensionless if and only if $(\pi_0, \pi_{k+1}, \cdots, \pi_n)^T$ is a sufficient statistic for $\theta$.*

Theorem 1 shows that if the parametric model is invariant to changes in physical dimensions, DA variables contain all the information needed to infer the parameters. In other words, if the statistical model is independent to the measurement system, we only need the observations based on DA variables. This corresponds to the physical concept that any physical phenomenon should be independent to the measurement system. Therefore, if the statistical model resembles the physical phenomenon, it is necessary to reduce the observations from raw variables to the DA variables. On the other hand, if the DA variables are not sufficient in capturing the behavior of the system, this is a signal to build a statistical model that depends on the dimensions. Conversely, if the DA variables are considered as sufficient information in describing the system, then the statistical model should be built with dimensionless parameters.

Theorem 1 suggests that DA is a sufficient dimension reduction under Assumption 1 and 2, and $\theta$ is dimensionless. Given $(\pi_0, \pi_{k+1}, \cdots, \pi_n)^T$ is a sufficient statistic for $\theta$, it is easy to derive that the distribution of $\pi_0|\pi_{k+1}, \cdots, \pi_n$ is the same as that of $\pi_0|X_1, \cdots, X_n$, thus proving the sufficiency of the reduction $\pi_{k+1}, \cdots, \pi_n$ (Adragni and Cook, 2009). That is, the sufficient reduction proclaimed by Buckingham's Π-theorem – basis quantities should be dropped from the equation – is the result of the sufficiency of DA variables.

With the sufficiency, it is legitimate and necessary to base the models on

DA variables. Some results on DA variables are useful and easy to derive. These include the information matrices of different estimates and the asymptotic of the MLE.

**Corollary 1.** *Under Assumptions 1 and 2,*

*(a) denote $\mathcal{I}_X(\theta)$ as the information matrix of the parameter $\theta$ based on variable set $X$, then $\mathcal{I}_{\pi_0,\pi_{k+1},\cdots,\pi_n}(\theta) \leq \mathcal{I}_{Y,X_1,\cdots,X_n}(\theta)$, with the equality if and only if $\theta$ is dimensionless.*

*(b) let $\delta(Y, X_1, \cdots, X_n)$ be an estimate for $\theta$; and let $\delta_1(\pi_0, \pi_{k+1}, \cdots, \pi_n) = E[\delta(Y, X_1, \cdots, X_n)|\ (\pi_0, \pi_{k+1}, \cdots, \pi_n)]$ be the Rao-Blackwellized version of $\delta$, then $E(\delta_1(\pi_0, \pi_{k+1}, \cdots, \pi_n) - \theta)^2 \leq E(\delta(Y, X_1, \cdots, X_n) - \theta)^2$.*

*(c) let $\hat{\theta}(\pi_0, \pi_{k+1}, \cdots, \pi_n)$ be an Maximum Likelihood Estimate for $\theta$, then under some regularity conditions, $\sqrt{n}(\hat{\theta} - \theta) \to N(0, \mathcal{I}^{-1}_{\pi_0,\pi_{k+1},\cdots,\pi_n})$.*

It is also useful to derive a similar theorem for nonparametric models. The following assumption is needed, however.

Assumption 2′: Let $\mathcal{C}$ be a dominated identifiable family of probability distributions on $\mathbb{R}^{n+1}$, and $Y, X_1, \cdots, X_n$ follows a distribution $\mathcal{P}$ within $\mathcal{C}$.

**Theorem 2.** *(Sufficient Dimension Reduction for Nonparametric Case): Under Assumptions 1 and 2′, any distribution $\mathcal{P}$ in family $\mathcal{C}$ is invariant to changes in physical dimensions if and only if $T = (\pi_0, \pi_{k+1}, \cdots, \pi_n)^T$ is a sufficient statistic for $\mathcal{C}$.*

Denote $\mathcal{P}_S$ as the joint distribution of $\vec{X}_S = (Y, X_1, \cdots, X_n)^T$, when the values of variables $Y, X_1, \cdots, X_n$ are recorded using measurement system $S$. It is worth noting that $\mathcal{P}_S$ is merely a nominal measure on the measured values of variables of interest, not on the abstract physical quantities themselves. It may not always be dimensionally invariant like the probability of an event $\mathbb{P}$ (in fact, $\mathbb{P} = \mathcal{P}_{\mathcal{S}} \circ \vec{X}_S = \mathcal{P}_{S'} \circ \vec{X}_{S'}$). The changes in dimensions from $S$ to $S'$ lead to the change of measure from $\mathcal{P}_S$ to $\mathcal{P}_{S'}$ and the change of variable values from $\vec{X}_S$ to $\vec{X}_{S'}$. On the other hand, similar interpretation holds for Theorem 2: in case of capturing physical phenomena that are invariant to dimensional changes, the nonparametric statistical model should be built upon DA variables. If DA

variables are not adequate to describe the system, models that do depend on the measurement system are suggested. Similar to the previous parametric case, if we assume models $\mathcal{P}_S$ are invariant to changes in measurement system, $\pi_{k+1}, \cdots, \pi_n$ is then a sufficient dimension reduction from $X_1, \cdots, X_n$ for regressing on $\pi_0$ under Assumption 1 and $2'$.

In addition to the sufficiency, which displays the capability of DA variables in retaining full information, to some family they are actually the smallest in size. The completeness of DA variables indicates that unbiased dimensionless statistics are unique and optimal. In these cases, the DA variables are the optimal statistics to work with. The result is summarized in the following theorem.

Assumption 3: Let $\mathcal{C}$ be the dominated identifiable family of all probability distributions on $\mathbb{R}^{n+1}$ that are invariant to dimensional changes (i.e., the collection of dimensionless models). Let $(Y, X_{k+1}, \cdots, X_n)^T \sim \mathcal{P} \in \mathcal{C}$.

**Theorem 3.** *(Completeness)*
*Under Assumptions 1 and 3, $(\pi_0, \pi_{k+1}, \cdots, \pi_n)$ is complete for family $\mathcal{C}$. i.e., for measurable function $h$,*

$$\forall F \in \mathcal{C}, E_F h(\pi_0, \pi_{k+1}, \cdots, \pi_n) = 0 \Rightarrow \forall F \in \mathcal{C}, \mathbb{P}_F(h(\pi_0, \pi_{k+1}, \cdots, \pi_n) = 0) = 1.$$

The completeness of DA variables within family $\mathcal{C}$ provides many theoretical results that are useful in practice. It guarantees the optimality and uniqueness of estimates based on DA variables. Some corollaries are shown below.

**Corollary 2.** *Under Assumptions 1 and 3, we have*

*(a) (Lehmann-Scheffe) suppose $\hat{\theta} = \hat{\theta}(\pi_0, \pi_{k+1}, \cdots, \pi_n)$ is an unbiased estimator for $\theta$. Then $\hat{\theta}$ is the unique best unbiased estimator (UMVUE).*

*(b) (Basu) $(\pi_0, \pi_{k+1}, \cdots, \pi_n)$ is independent of ancillary statistics of family $\mathcal{C}$.*

*(c) (Bahadur) $(\pi_0, \pi_{k+1}, \cdots, \pi_n)$ is the minimal sufficient statistics for distributions in family $\mathcal{C}$.*

Based on the theorems and corollaries above, we can conclude that if we consider dimensionless models, then DA variables are the optimal choice to construct estimators with smallest variance given the bias.

In addition to studies of families where DA variables are sufficient and complete as shown above, previous literature also investigated the preservation of (minimal) sufficiency and completeness under invariance structure. Following their notations, the main result of the relationship between sufficiency and invariance by Hall et al. (1965) stated that $\mathcal{B} \cap \mathcal{A}_G$ is sufficient for $\mathcal{A}_G$ if (i) $\mathcal{B}$ is a sufficient and G-stable ($g(\mathcal{B}) = \mathcal{B}$) $\sigma$-field, and (ii) $\mathcal{B} \cap \mathcal{A}_G \sim \mathcal{B} \cap \mathcal{A}_A(\mathcal{P})$ for G-invariant family $\mathcal{P}$ ($\mathcal{P}g^{-1} \subset \mathcal{P}$), where $\mathcal{A}_G$ is the $\sigma$-field of G-invariant sets ($g^{-1}(A) = A$); $\mathcal{A}_A$ is the $\sigma$-field of almost-G-invariant sets ($g^{-1}(A) \sim A(\mathcal{P})$). In our context, $\mathcal{A}_G$ is the induced $\sigma$-field $M^{-1}(\mathcal{R}^n)$ of $M$, due to the maximal invariant property of $M$ in Lemma 5. By Hall et al. (1965), it can be inferred that the dimensionless version of a sufficient statistic for the original model is still sufficient for DA invariant models. Furthermore, the completeness of model families is also inherited through invariance reduction by DA. However, minimal sufficiency is not inherited: if $\mathcal{B}$ is a minimal sufficient $\sigma$-field, the dimensionless version $\mathcal{B} \cap \mathcal{A}_G$ is not guaranteed to be minimal sufficient. Counterexamples were provided by Hall et al. (1965) and Chacón et al. (2006). (Note that an equivalent statement of (ii) was given by Berk (1972): the ancillary invariant $\sigma$-field is independent of an appropriate sufficient $\sigma$-field. Landers and Rogge (1973) proved the necessity of condition (ii) and a substitution of G-stability condition in (i), $g(\mathcal{B}) = \mathcal{B}$, by dominated $\mathcal{P}$.)

Implications from the above theories in our context are as follows. For the probability family $\mathcal{P}$ that is G-invariant with $\mathcal{B} \cap \mathcal{A}_G \sim \mathcal{B} \cap \mathcal{A}_A(\mathcal{P})$, dimensionless version of sufficient and complete statistics is still sufficient and complete (implying minimal sufficient) for induced models based on DA variables, which is particularly applicable to exponential families. However, in general, the dimensionless version of the minimal sufficient statistics for the original model may not be minimal sufficient for the induced model on DA variables. Our theory articulates the complete family and thus the condition for minimal sufficiency.

In summary, DA generates variables that are maximal invariant, sufficient and complete under an appropriate probability family. Although these proposed properties can be perceived easily in the DA procedure described in Section 2.4, the proofs given here are proceeded with minimal assumptions as DA requires. The procedure in Section 2.4 is merely a special case that satisfies the condi-

tions. Through the proposed representation, DA can be properly incorporated into a more general probabilistic approach, without restricting to such special form. On the other hand, the proofs are more direct and specialized compared to the general theories of invariance and sufficiency, which induces the following advantages include that (i) it is not necessary to establish a probability model to prove the maximal invariant property; (ii) sufficiency is given as an "if and only if" statement; (iii) the probability family to which DA variables are complete is the most generic family; (iv) practitioners do not need to verify the textbook invariance setting case by case.

## 4. Case Study: A Meteorology Example

One of the most important topics in Meteorology is to investigate the dynamics and processes at the atmospheric boundary layer. The planetary boundary layer, illustrated as the shaded area in Figure 1, is the lowest part of the atmosphere, starting from the surface layer where we live to the cloud layer. Its behavior is categorized into different zones based on the local time and height (the x- and y-axises in Figure 1, respectively). Modeling difficulties arise in such region because the physical laws governing the atmosphere's dynamics are complex and non-linear. Physical quantities such as temperature, moisture and flow velocity in this layer fluctuate rapidly because of its interactive dynamics with the planetary surface. Extensive progress has been made in theoretical, numerical and experimental studies.

Here, we consider developing the relationship between the vertical velocity variance ($Y = w^2$) and the height where it is measured ($X_1 = z$). Similar problems can be found in Young (1988); Lin and Shen (2013). 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 ($X_2 = w_*$), and the depth of the boundary layer ($X_3 = z_i$), are important scales for all quantities concerned (Stull, 1988). Thus we intend to model an numerical expression of the velocity variance $w^2$ based on the height $z$, as well as the scales $w_*$ and $z_i$.

Figure 2(a) displays the scatterplot of $w^2$ and $z$, based on the measurements from the Phoenix 78 experiment (see Young, 1988). The relevant data set is presented in Appendix B in the supplementary material. The purpose of the

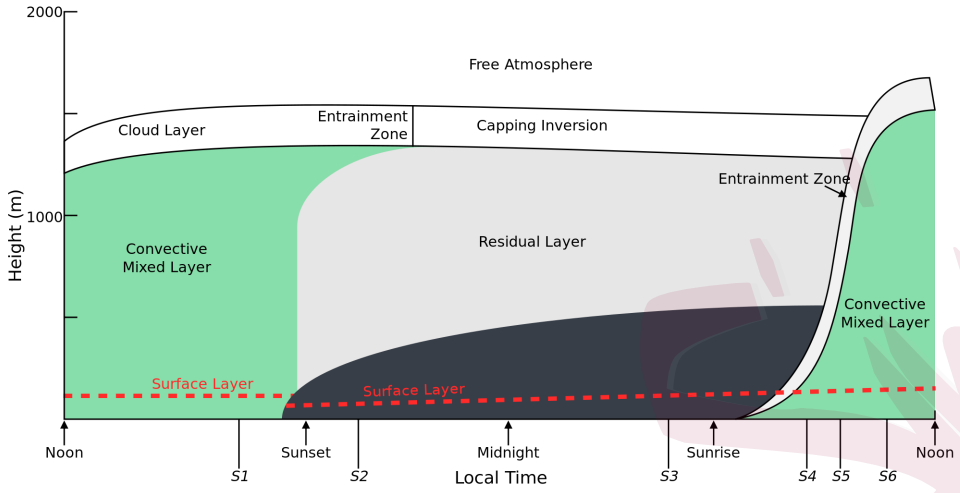
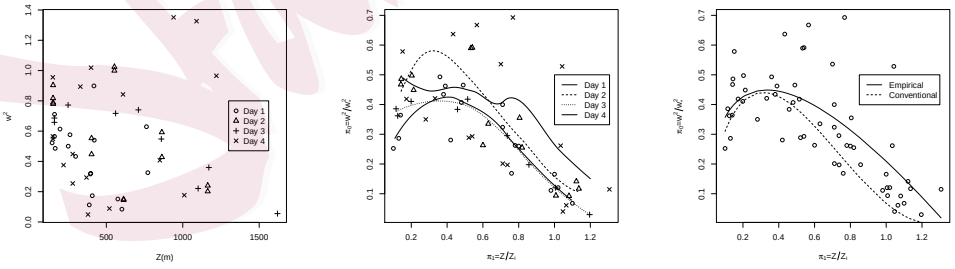



Figure 1: Illustration of planetary boundary layer.

Phoenix 78 experiment was to study the turbulence of convective boundary layer. During the experiment, the profiles of turbulence statistics were recorded through aircraft observations. From Figure 2(a), the dependence between $w^2$ and $z$ is not obvious. Data points scatter apart quite randomly. This may be attributed to different magnitudes of the velocity scale ($X_2 = w_*$) and the boundary layer depth ($X_3 = z_i$). Statistical models on raw data would conclude insignificant dependence. Furthermore, predictions on $w^2$ generate unreasonable (negative) results when extrapolating to low $w_*$ and $z_i$, which is not desirable.




(a) Original data set. Different symbols stand for different dates.

(b) Transformed data and LOESS fits for four different dates.

(c) Empirical and conventional model based on the transformed data set.

Figure 2: Scatter plots and estimates of Phoenix 78 data.

Table 2: Dimensions of Variables from the Phoenix 78 Experiment.

| Variables | $Y = w^2$ | $X_1 = z$ | $X_2 = w_*$ | $X_3 = z_i$ |
|---|---|---|---|---|
| Dimensions | $\mathbf{L^2T^{-2}}$ | $\mathbf{LT^0}$ | $\mathbf{LT^{-1}}$ | $\mathbf{LT^0}$ |
| Vector Representation | $(2, -2)$ | $(1, 0)$ | $(1, -1)$ | $(1, 0)$ |

### 4.1. Dimensional analysis and statistical properties

Following the DA procedure in Section 2.1.3, first we need to identify the dimensions of variables involved. The corresponding physical dimensions of the vertical velocity variance $w^2$, the height $z$, the convective velocity scale $w_*$ and the depth of boundary layer $z_i$ are listed in Table 2.

The two fundamental dimensions involved are the length [**L**] and the time [**T**]. Based on Section 3.1, these two dimensions generate a group of dimensions $\mathcal{F} = \{D = \mathbf{L}^{d_1}\mathbf{T}^{d_2} : d_1, d_2 \in \mathbb{Q}\}$ and the 2-dimensional vector space $(\mathbb{Q}, \mathcal{F})$ with multiplication and power as two valid operations. Note that the dimensions of the variables of interest are elements of $\mathcal{F}$. They can be coded as vector forms as in Table 2. Therefore the dimensional matrix $D$ is

$$D = \begin{pmatrix} 2 & -2 \\ 1 & 0 \\ 1 & -1 \\ 1 & 0 \end{pmatrix} \begin{matrix} \cdots & [Y] \\ \cdots & [X_1] \\ \cdots & [X_2] \\ \cdots & [X_3] \end{matrix}$$

The rank of $D$ is 2. The last two rows ($X_2$, $X_3$) are selected to be the basis; then the other two row can be represented as $[Y] = [X_2]^2$ and $[X_1] = [X_3]$. Consequently, dimensionless variables are $\pi_0 = Y/X_2 = w^2/w_*^2$ and $\pi_1 = X_1/X_3 = z/z_i$. The original model is $w^2 = f(z, w_*^2, z_i)$, where $f$ is to be estimated. The DA model is $\pi_0 = g(\pi_1)$, i.e., $w^2 = w_*^2 g(z/z_i)$. Our task is to estimate function $g$ (instead of $f$).

In SI measurement system, length [**L**] has unit meter and time [**T**] has unit second. Imperial system is another popular alternative to the metric system, where the unit for [**L**] is 1 feet = 0.3048 meters. According the statistical decision theory, we advocate statistical methods that yield same result, no matter which measurement system is used. Based on Section 3.2, dimensionless variables help us build such invariant statistical methods. Without dimensionless variables, the

transition of results between different measurement systems sometimes can be difficult. For instance, suppose one decides to use a local polynomial regression type method to estimate the function $f$ of the original $z, w_*^2, z_i$. Between different platforms, the bandwidth/smoothing parameter and the weight function may need to be adjusted corresponding to different scales in order to obtain the same result. But if DA is used and $g$ is estimated, both $\pi_0 = Y/X_2 = w^2/w_*^2$ and $\pi_1 = X_1/X_3 = z/z_i$ will have the same numerical value regardless whether metric system or imperial system is used. The subsequent procedure will thus be invariant to the scale changes of dimensions as well.

Consequently, we reduce the number of variables of interest from 4 to 2. According to Section 3.3, actually $\pi_0$ and $\pi_1$ are sufficient and complete statistics to the family of all invariant statistical models of original variables. If finding an appropriate model from the invariant family is of interest, it is sufficient and optimal to focus/condition on the two dimensionless $\pi_0$ and $\pi_1$. As stated by Corollary 1(b), estimators based on $\pi_0$ and $\pi_1$ have less mean squared errors. By Corollary 2(a), the unbiased estimators are unique UMVUE.

### 4.2. Further remarks on model building and scalability

Figure 2(b) is the scatterplot of $\pi_0 = w^2/w_*^2$ and $\pi_1 = z/z_i$. To capture the nonparametric relationship between $\pi_0$ and $\pi_1$, we fit a local linear regression via LOESS (R Development Core Team, 2011). Based on Theorem 2, $(\pi_0, \pi_1)$ is sufficient. It gives four curves in Figure 2(b) (corresponding to four different dates). Each individual curve shares a similar shape. We anticipate to build up a common empirical model $\pi_0 = f(\pi_1)$ that is adequate to describe the common feature.

Now we switch to the parametric case in Theorem 1. Assuming the function is of power law form (with some boundary conditions at $\pi_1 = 0$ and multiplicative log-normal errors of $\mu = 0$ and constant $\sigma$), the empirical model can be built as: $\pi_0 = 1.554\pi_1^{1/2}(1 - 0.866\pi_1^{1/2})$ or $w^2/w_*^2 = 1.554(z/z_i)^{1/2}\,[1 - 0.866(z/z_i)^{1/2}]$, using maximum likelihood estimate. By Corollary 1(c), estimates are asymptotically Normal under some regularity conditions. It is interesting to compare the empirical model with the conventional model in meteorology (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) dis-

plays both models. The empirical model is close to the conventional model, but with a better fit. Moreover, they share a similar analytical form.

It is also possible to test the Buckingham's Π-theorem. One can build up a model with both the dimensionless ones $\pi_0$ and $\pi_1$ and basis quantities $w_*$, $z_i$, and test the significance of the latter two. If Buckingham's Π-theorem holds, they should not be involved in the model. As pointed out in Section 2.4, the significance of basis quantities is an indicator of missing key variables. In case of the significance of $w_*$ and $z_i$, we should maintained them in the model while searching for other related quantities.

The importance of DA certainly lies beyond the convenience of transiting results between systems. More importantly, the dimensionless variables better characterize the intrinsic shape and comparative magnitude of the system rather than the scales. Dimensionless variables constitute dimensionless models with good extrapolation capability. This property is essential for engineering problems. For example, in order to study the product reliability in the real scale, an accelerated laboratory testing is usually conducted with much less cost. Engineers use wind tunnels and small-scale experiments as pilot studies. Modeling the relative magnitude could help us generalize the experimental results to the real scale. From the pilot forecast, it is also easier to design the follow-up real scale experiment, such as determining how many data points are necessary for controlling the errors. Hence, in order to maintain both the physical insight from the physical dimensions and the simplicity of the empirical analysis, dimensionless models based on DA dimensionless variables are recommended.

## 5. Conclusion

Dimensional Analysis has been a popular and well-developed methodology in the physical sciences and engineering disciplines. It removes dimensional constrains and generate scientifically valid models. It utilizes the physical information and produces interpretable and scalable results. Its compatibility with statistical procedures proves itself to be a good framework to incorporate professional knowledge in physical and engineering problems.

While there are plenty of theoretical development of DA in other disciplines such as physics and mathematics, the missing counterpart in statistics inhibits

its implementation and popularization among statistical practitioners. In this paper, we review DA from a statistical perspective and derive some fundamental statistical properties: the vector space structure of physical dimensions; the maximal invariance of dimensionless quantities, and the sufficiency and completeness of DA variables. Although the derivation might seem intuitive after specifying the structure, the results are of great significance in guiding future development of statistical methods based on DA. DA's close connection with the fundamental statistical concepts justify its compatibility and relevance in statistical problems.

Furthermore, we hope this paper could shed some light on the proper usage of DA and the post-DA modeling. We believe that when conducting DA, we utilize the inherent scaling structure and assume that the scales do not affect the physical outcomes. It is the "absolute significance of relative magnitude", that characterizes the physical system. Thus, the probability models ought to provide invariant and equivariant decisions under such dimensional scaling. DA transforms variables into dimensionless, while preserving the sufficiency and completeness. Therefore, estimates based on DA variables are automatically invariant and optimal under squared loss. Besides, modeling on the relative magnitude generates good scalability, that is essential in engineering problems such as accelerated life testing.

## References

Adragni, K. P. and R. D. Cook (2009). Sufficient Dimension Reduction and Prediction in Regression. *Philosophical Transactions of the Royal Society A: Physical, Mathematical and Engineering Sciences 367*, 4385–4405.

Albrecht, M. C., C. J. Nachtsheim, T. A. Albrecht, and R. D. Cook (2013). Experimental Design for Engineering Dimensional Analysis. *Technometrics 55*(3), 257–270; with Rejoinder 292–295.

Balaguer, P. (2013). *Application of Dimensional Analysis in Systems Modeling and Control Design*. IET Control Engineering. Stevenage: The Institution of Engineering and Technology.

Berk, R. H. (1972). A note on sufficiency and invariance. *The Annals of Mathematical Statistics 43*(2), 647–650.

Bridgman, P. (1931). *Dimensional Analysis* (2nd ed.). Yale University Press.

Buckingham, E. (1914). On physically similar systems; illustrations of the use of dimensional equations. *Physical Review 4*(4), 345–376.

Cariñena, J. F., M. A. del Olmo, and M. Santander (1981). Kinematic groups and dimensional analysis. *Journal of Physics A: Mathematical and General 14*, 1–14.

Cariñena, J. F., M. A. del Olmo, and M. Santander (1985). A new look at dimensional analysis from a group theoretical viewpoint. *Journal of Physics A: Mathematical and General 18*, 1855–1872.

Chacón, J. E., J. Montanero, A. G. Nogales, and P. Pérez (2006). A note on minimal sufficiency. *Statistica Sinica 16*, 7–14.

Davis, T. P. (2013). Comment: Dimensional Analysis in Statistical Engineering. *Technometrics 55*(3), 271–274.

Drobot, S. (1953). On the Foundations of Dimensional Analysis. *Studia Mathematica 14*(1), 84–99.

Eaton, M. L. (1989). Group Invariance Applications in Statistics. *Regional Conference Series in Probability and Statistics 1*, i–v+1–133.

Frey, D. D. (2013). Comments: Dimensional Analysis and Experimentation as a Catalyst to Learning From Data. *Technometrics 55*(3), 271–274.

Grudzewski, W. and K. Roslanowska-Plichcinska (2013). *Application of Dimensional Analysis in Economics*. Amsterdam: IOS Press.

Hall, W. J., R. A. Wijsman, and J. R. Ghosh (1965). The relationship between sufficiency and invariance with applications in sequential analysis. *The Annals of Mathematical Statistics 36*, 575–614.

Islam, M. and L. M. Lye (2007). "Combined Use of Dimensional Analysis and Statistical Design of Experiment Methodologies in Hydrodynamics Experiments". 8th Canadian Marine Hydromechanics and Structures Conference.

Jones, B. (2013). Comments: Enhancing the Search for Compromise Designs. *Technometrics 55*(3), 278–280.

Landers, D. and L. Rogge (1973). On sufficiency and invariance. *The Annals of Statistics 1*(3), 543–544.

Lehmann, E. and G. Casella (2003). *Theory of Point Estimation* (2nd ed.). Springer Texts in Statistics. New York: Springer.

Lin, D. K. J. and W. Shen (2013). Comments: Experimental Design for Engineering Dimensional Analysis. *Technometrics 55*(3), 281–285.

Monin, A. S. and A. M. Obukhov (1954). Basic laws of turbulent mixing in the surface layer of the atmosphere. *Tr. Akad. Nauk. SSSR Geophiz. Inst.*.

Piepel, G. F. (2013). Comments: Spurious Correlation and Other Observations on Experimental Design for Engineering Dimensional Analysis. *Technometrics 55*(3), 286–289.

Plumlee, M., V. R. Joseph, and C. F. J. Wu (2013). Comments: Alternative Strategies for Experimental Design. *Technometrics 55*(3), 289–292.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Shen, W., T. Davis, D. K. J. Lin, and C. J. Nachtsheim (2014). Dimensional Analysis and Its Applications in Statistics. *Journal of Quality Technology 46*(3), 185–198.

Siano, D. (1985a). Orientational Analysis - A Supplement to Dimensional Analysis - I. *Journal of the Franklin Institute 320*(6), 267–283.

Siano, D. (1985b). Orientational Analysis, Tensor Analysis and the Group Properties of SI Supplementary Units - II. *Journal of the Franklin Institute 320*(6), 285–302.

Socha, K. (2007). Circles in circles: Creating a mathematical model of surface water waves. *The American Mathematical Monthly 114*(3), 202–216.

Sonin, A. A. (2001). *The Physical Basis of Dimensional Analysis* (2nd ed.). web.mit.edu/2.25/www/pdf/DA_unified.pdf.

Stull, R. B. (1988). *An Introduction to Boundary Layer Meteorology*. Kluwer Academic Publishers.

Szirtes, T. (2007). *Applied Dimensional Analysis and Modeling, 2nd ed.* Elsevier Butterworth-Heinemann.

Tao, T. (2012). A mathematical formalisation of dimensional analysis. http://terrytao.wordpress.com/2012/12/29/a-mathematical-formalisation-of-dimensional-analysis/.

Taylor, M., A. I. Diaz, L. A. Jodar-Sanchez, and R. J. Villanueva-Mico (2008). A Matrix Generalisation of Dimensional Analysis Using New Similarity Transforms to Address the Problem of Uniqueness. *Adv. Studies Theor. Phys. 2*(20), 979–995.

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

Zlokarnik, M. (1991). *Dimensional Analysis and Scale-up in Chemical Engineering*. Berlin: Springer-Verlag.

Department of Statistics, the Pennsylvania State University, University Park, PA, 16802, U.S.A.

E-mail: jayshenwei@gmail.com

Department of Statistics, the Pennsylvania State University, University Park, PA, 16802, U.S.A.

E-mail: dkl5@psu.edu