JOURNAL OF AIRCRAFT
Vol. 43, No. 6, November–December 2006

Principal Component Regression for Fitting Wing Weight Data of Subsonic Transports

Humberto Rocha
Old Dominion University, Norfolk, Virginia 23529
and
Wu Li and Andrew Hahn
NASA Langley Research Center, Hampton, Virginia 23681

DOI: 10.2514/1.21934

This paper documents the lessons learned from fitting the wing weight data of 41 subsonic transports by a semi-empirical regression model, least polynomial interpolation, radial basis function interpolation, Kriging interpolation, Gaussian process, and principal component regression using radial basis function interpolation. The paper discusses various aspects of fitting data to a wing weight model: data scaling, variable selection, principal component analysis, subjective choice of input variables, interpolation methods, and verification of constructed wing weight models. The numerical results show that principal component regression using multiquadric radial basis function interpolation is capable of capturing physical trends buried in the wing weight data and generates the most useful wing weight model for conceptual design of subsonic transports among the tested data fitting methods. Even though the benefits of principal component regression are only demonstrated by the wing weight data fitting problem, the methodology could have significant advantages in fitting other historical or sparse data.

Nomenclature

A	=	aspect ratio of wing, i.e., b^2/s
b	=	wing span
c_m	=	mean chord of wing, i.e., s/b
c_r	=	root chord of wing at root
c_t	=	tip chord of wing
f	=	theoretical wing weight function
g	=	approximation of f
N	=	number of data points
n	=	number of input variables
s	=	plan area of wing
t_r	=	thickness of airfoil at root
t_t	=	thickness of airfoil at wingtip
t_r/c_r	=	thickness-to-chord ratio of airfoil at root
t_t/c_t	=	thickness-to-chord ratio of airfoil at wingtip
$[t/c]_m$	=	average thickness-to-chord ratio, i.e., $(t_r/c_r + t_t/c_t)/2$
$\boldsymbol{x}$	=	column vector of input variables $x_1, \ldots, x_n$
x_i	=	ith component of column vector $\boldsymbol{x}$
x_i^j	=	ith component of column vector $\boldsymbol{x}^j$
w	=	actual wing weight
$\bar{w}$	=	estimated or calculated wing weight
w_{to}	=	gross takeoff weight of aircraft
Λ	=	wing sweep angle in radians
λ	=	taper ratio of wing, i.e., c_t/c_r
μ	=	ultimate load factor
σ_i	=	estimated standard deviation of variable x_i
φ	=	radial basis function

Subscripts and Superscripts

i	=	index for ith component of vector
j	=	index for data point
k	=	index for iteration or iterate
T	=	transpose of vector or matrix

Presented at the 11th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, Portsmouth, VA, 6–8 September 2006; received 20 December 2005; revision received 9 May 2006; accepted for publication 15 May 2006.

I. Introduction

This paper uses a real-world wing weight estimation problem as an example to demonstrate how principal component regression (PCR) with radial basis function (RBF) interpolation can be used to build regression models that fit historical data exactly and exhibit desirable trends between and beyond the data points.

Wing weight estimation is a very important step in conceptual design of aircraft (see [1], pp. 467–476). Any significant change in an aircraft concept usually requires resizing of the wings. For example, replacing aluminum by a stronger composite material will reduce the weight of the vehicle, which may use a smaller engine, less fuel, and smaller wings for flying the same mission. Moreover, the vehicle weight also affects the cruise condition for a subsonic transport, and a change in the cruise condition usually requires redesign of wing for better aerodynamic efficiency. The conceptual design process involves constant reevaluations of wing weights, which requires a weight estimation model that yields credible results with a minimum of input burden and execution time.

Every major aircraft manufacturer has its own wing weight estimation models, which are considered proprietary information not to be shared with the public. Also, weight information of existing aircraft is not necessarily available to the public. System analysts at NASA Ames Research Center were able to collect weight information for 41 subsonic transports including Boeing 747, Douglas DC-7C, Fokker F-28 twin engine jet liner, and Lockheed C-130B cargo aircraft (see Fig. 1). This set of weight data allows the current study of benefits and limitations of using general approximation methods for building a wing weight estimation model.

Practical wing weight models for conceptual design are mainly semi-empirical formulas based on historical data. For example, three wing weight formulas for fighters, transports, and general aviation aircraft, respectively, are provided in [1] (see pp. 473–476).

Fig. 1 A variety of subsonic transports in the data set.

However, a very useful wing weight formula for one conceptual design task might be inadequate for another. For example, if a designer wants to understand how bending material weight and aeroelastic effect change the wing weight, then a semi-empirical formula developed by McCullers [2] (see also Eq. (1) in [3]) may be more useful than the ones in [1].

Besides the changes of design parameters, which may make a wing weight model inadequate, the credibility of a wing weight model becomes more questionable as more wing weight data are available in public. It is always desirable to calibrate an existing semi-empirical formula so that the calibrated formula could provide accurate wing weight estimates for the available aircraft data. However, simple semi-empirical formulas might not be able to provide accurate wing weight estimates as more and more wing weight data become available. There are two general methods for improving the wing weight prediction accuracy: one is to use a physics-based structural analysis method and another is to use more complex semi-empirical formulas.

Physics-based wing weight models usually require the development of a structure and load analysis method to estimate the structural weight of the wing (see, e.g., [3,4]), which could then be calibrated to a historical structural weight database by using an average correction factor determined by the wing weight data in the database [4]. A physics-based wing weight model has the potential of using a sparse and poorly distributed data set to generate a credible weight estimation model for conceptual design. However, such a model relies on structural weight analysis that requires more detailed input information (such as specific material properties, point mass values and locations, and load distributions) that may not even be available at an early stage in the design cycle, ignores nonstructural weight contributors (such as fasteners), and increases the computational time by several orders of magnitude. As a result, a physics-based wing weight model does not necessarily provide more credible estimates than semi-empirical models. For example, [4] has to use the following regression model to calibrate the wing weight prediction model based on beam theory structural analysis:

$$w_{\text{calibrated}} = 1.7372 w_{\text{calculated}} \tag{1}$$

where $w_{\text{calculated}}$ is the calculated wing weight based on beam theory structural analysis, $w_{\text{calibrated}}$ is the calibrated wing weight, and the number 1.7372 is the regression coefficient. Fig. 14 in [4] shows that there are still significant differences between $w_{\text{calibrated}}$ and the actual wing weight for six out of eight subsonic transports. Needless to say, the wing weight model based on beam theory alone has more than 70% errors in its prediction for the eight transports in the database. Another finite-element based wing weight model [3] gives wing bending material weight values that are different by up to 40% from the predictions by semi-empirical wing weight models (see Fig. 4 in [3]). Table 4 in [3] shows about 40% difference between the finite-element model and a semi-empirical model for wing bending material weight estimates. Unfortunately, no comparison with real-world weight data was provided in [3].

Complex semi-empirical formulas are hard to derive and might not be more accurate as we will see later. Moreover, a complex semi-empirical formula could lead to unexpected wrong trend predictions as its developer might have misconceptions about the complex interactions among many input variables in the formula.

This paper aims to demonstrate that it is possible to construct a wing weight approximation that could provide credible wing weight estimates by using advanced data fitting methods. In particular, we will show that PCR using multiquadric RBF interpolation is capable of capturing physical trends buried in the wing weight data of 41 subsonic transports and generates the most useful wing weight model for conceptual design of subsonic transports among the tested data fitting methods, including semi-empirical regression methods, least polynomial interpolation, RBF interpolation, Kriging interpolation, and Gaussian process.

The paper is organized as follows. Section II shows how the wing weight data could be used to construct a wing weight estimation model. Subsections of Section II are devoted to various issues in the wing weight data fitting process: knowledge-based wing weight data fitting, forward and backward variable selections, data scaling, principal component analysis, subjective choice of input variables, and PCR using RBF interpolation. In Section III, numerical results are included to show advantages of PCR over other data fitting methods such as knowledge-based wing weight data fitting, Kriging interpolation, least polynomial interpolation, and RBF interpolation. The concluding remarks are given in Section IV. Appendix A includes the wing weight model generated by the multiquadric PCR fitting with eight input variables. Appendix B has the sample data used to construct the RBF interpolant.

II. Wing Weight Data Fitting

A standard data fitting process can usually be decomposed into four steps: 1) data generation and variable screening, 2) fitting the data by a linear or nonlinear function model, 3) tuning intrinsic model parameters by using cross-validation, and 4) verification of constructed approximation. Data generation is mainly for selection

of data sites of the input vector when the corresponding response is calculated by a computer simulation code, and is not needed for wing weight data fitting that uses only historical data.

In the following subsections, we describe the wing weight data for 41 subsonic transports (see Fig. 1), analyze the wing weight data by various methods, give a brief review of PCR using RBF interpolation for wing weight data fitting, and outline some desirable properties of wing weight models for conceptual design.

A. Wing Weight Data

Each wing weight data point consists of the actual wing weight w and relevant configuration parameters: A, b, c_m, c_r, c_t, s, $[t/c]_m$, t_r, t_r/c_r, t_t, t_t/c_t, w_{to}, λ, Λ, and μ. These 15 configuration parameters can be regrouped in three categories: 1) primary wing geometry parameters including chord length at root, chord length at tip, span, reference area, thickness at root, thickness at tip, and sweep angle ($\Lambda = |90^\circ - \Lambda_0|$); 2) derived wing geometry parameters including taper ratio, aspect ratio, mean chord of wing, mean thickness-to-chord ratio, thickness-to-chord ratio at root, and thickness-to-chord ratio at tip; and 3) wing structure parameters including gross takeoff weight of aircraft and ultimate load factor. Figure 2 shows the wing geometry parameters for a trapezoidal approximation of the actual wing.

A detailed explanation of these wing parameters can be found in [1]. Note that some of the parameters are related:

$$\begin{gathered}\text{thickness-to-chord ratio at root} = t_r/c_r, \qquad b = \sqrt{As} \\ c_m = s/b, \qquad \lambda = c_t/c_r \\ \text{thickness-to-chord ratio at tip} = t_t/c_t \\ 2[t/c]_m = t_r/c_r + t_t/c_t, \quad \text{and} \quad s = \frac{1}{2}b(c_r + c_t)\end{gathered} \tag{2}$$

However, although all the parameter values were collected independently, some of the formulas in Eq. (2) were violated by the original data set of the 41 subsonic transports. For example, in some cases, the recorded taper ratio is not the same as c_t/c_r and some of the recorded thickness-to-chord ratios at the root were not the same as t_r/c_r. Almost none of the existing data satisfy the relationship $s = \frac{1}{2}b(c_r + c_t)$. A careful examination of the data file reveals the nine source input parameters of the data: A, c_r, c_t, s, t_r, t_t, w_{to}, Λ, and μ. Based on these nine parameter values, six additional parameter values are generated by the first six formulas given in Eq. (2). We could not resolve the inconsistency for the last formula in Eq. (2) because we did not know which parameter among c_t, c_r, and s is responsible for the data inconsistency. To overcome this problem, we assume

$$c_r + c_t \approx \gamma c_m = \gamma s/b \tag{3}$$

where γ is the average of $(c_r + c_t)/c_m$ values for the known wing configurations.

One important application of a wing weight model in conceptual design of aircraft is for trade studies, wherein system analysts try to understand how the wing weight changes if one or more of the parameters b, s, t_r/c_r, λ, and Λ are modified. We shall expand this

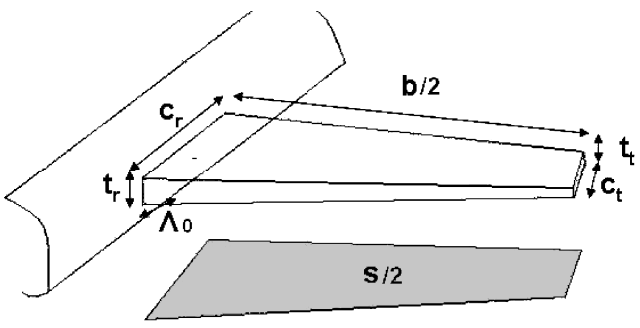

Fig. 2 Airplane wing geometry parameters.

set of five parameters to the following set of eight parameters as the independent variables of wing weight models:

$$\{b, s, t_r/c_r, \lambda, \Lambda, \mu, w_{\text{to}}, t_t/c_t\} \tag{4}$$

The ultimate load factor and total empty weight w_{to} of aircraft determine the structural strength of a wing and can significantly influence the wing weight. In general, t_t/c_t is believed to have a marginal effect on wing weight, which is included just to see whether the additional detail about a wing could lead to better wing weight estimation models.

The default set of independent variables does not exclude the use of other parameters as input variables of a wing weight model. For example, one can use A, s, $[t/c]_m$, w_{to}, λ, Λ, and μ as input variable of a wing weight model $w = w(A, s, [t/c]_m, w_{\text{to}}, \lambda, \Lambda)$. But we shall rewrite the wing weight model as

$$\begin{aligned} w &= w(A, s, [t/c]_m, w_{\text{to}}, \lambda, \Lambda) \\ &= w\left(b^2/s, s, \frac{1}{2}(t_r/c_r + t_t/c_t), w_{\text{to}}, \lambda, \Lambda\right) \end{aligned}$$

Using $\lambda = c_t/c_r$ and Eq. (3), we obtain the following chord approximation formulas for c_r and c_t:

$$c_r \approx \frac{\gamma s}{b(1+\lambda)} \quad \text{and} \quad c_t \approx \frac{\gamma s \lambda}{b(1+\lambda)} \tag{5}$$

In this paper, if c_r or c_t is used as an input variable for a wing weight model, then we will substitute it by the corresponding chord approximation formula in Eq. (5) during trade studies.

B. Knowledge-Based Wing Weight Data Fitting

The current practice in semi-empirical regression for wing weight estimates is based on heuristic regression models that incorporate some engineering understanding of the weight relationship. For example, two current semi-empirical regression models for the given wing weight dataset of 41 subsonic transports are the ratio model:

$$\bar{w} = \bar{\alpha}_1[\mu^{\bar{\alpha}_2} A^{\bar{\alpha}_3} s^{\bar{\alpha}_4} ([t/c]_m)^{\bar{\alpha}_5} (\cos\Lambda)^{\bar{\alpha}_6} (1+\lambda)^{\bar{\alpha}_7} (10^{-3} w_{\text{to}})^{\bar{\alpha}_8}] \tag{6}$$

and the geometry model:

$$\begin{aligned} \bar{w} = \alpha_1[&\mu^{\alpha_2} (0.01b)^{\alpha_3} (10^{-3}s)^{\alpha_4} (t_r)^{\alpha_5} (0.1c_r)^{\alpha_6} (\cos\Lambda)^{\alpha_7} \\ &\times (0.1c_t)^{\alpha_8} (10^{-5} w_{\text{to}})^{\alpha_9}] \end{aligned} \tag{7}$$

where $\alpha_1, \ldots, \alpha_9$ or $\bar{\alpha}_1, \ldots, \bar{\alpha}_8$ are determined by least squares fitting of the data. The engineering intuition behind these wing weight models is that the wing weight is a monotone function with respect to each of the configuration parameters in the model and its range is from 0 to ∞. Note that real-world wing weight has a finite range and it seems unrealistic to allow wing weight to have a range from 0 to ∞. However, during conceptual design, it is not clear what combinations of design variables lead to practical concepts. As a consequence, optimization or design exploration methods for conceptual design will generate unrealistic configurations. A useful wing weight model should assign unrealistically large values to those unrealistic configurations so that they will be eliminated as undesirable concepts during optimization or design exploration process. Of course, if the maximum wing weight for any potential subsonic transport is known, then one could simply limit the wing weight range to the known value even if the estimated wing weight is greater than the known value.

The wing weight models (6) and (7) are based on system analysts' knowledge of subsonic transports and it is nontrivial to derive similar empirical regression models for other types of aircraft. Moreover, the models are not flexible enough to fit the wing weight data for the 41 subsonic transports. Figure 3 shows more than 10% errors in the wing weight estimation by the best fit of each of these two semi-empirical wing weight models.

Fitting the wing weight data by either the ratio model (6) or the geometry model (7) is a nonlinear least squares problem that may

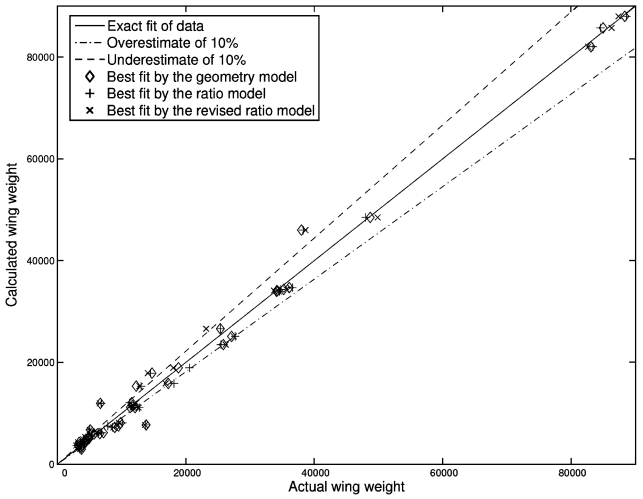

Fig. 3 Regression errors of wing weight fitting by the ratio model (6) and the geometry model (7).

have many local optimal solutions. The best fitting most likely depends on the initial choice of the regression parameters $\alpha_1, \ldots, \alpha_9$ or $\bar{\alpha}_1, \ldots, \bar{\alpha}_8$ and Fig. 3 shows the best fitting computed by the nonlinear optimization code lsqnonlin in MATLAB after several trials with various initial guesses. Even though the geometry model (7) is an attempt to find a more accurate regression model than the ratio model (6) by increasing the degree of freedom in fitting model, there is no decisive advantage of one model over the other model if the goodness-of-fit is the comparison criterion: the maximums of relative fitting errors are 56.44% and 57.55% for the geometry model and the ratio model, respectively. This further illustrates the difficulty of getting accurate regression models of historical data based on expert intuitions.

It is important to provide analysts with useful information about the prediction behaviors of these wing weight models (instead of fitting errors) between and beyond the data points, that helps analysts choose an appropriate prediction model for the conceptual design trade study at hand. Later on, we shall see that the geometry model (7) may provide undesirable trends for trade studies.

C. Variable Selection

A standard variable screening/selection process identifies a subset of the input variables $x_1, \ldots, x_n$ that have significant influences on the response $f(\boldsymbol{x})$. In other words, if the change of $f(\boldsymbol{x})$ with respect to a variable x_i is negligible, then remove x_i from the input vector.

For the wing weight data fitting problem, variable screening methods (see, e.g., [5]) that require the values of the response at specific input sites are not applicable. The main effects estimate method, used by Tu and Jones [6], generally requires a uniform distribution of the existing input data vectors in a rectangular domain of the input space, whereas the forward or backward variable selection method (see, e.g., [7,8]) is mainly to determine the explanatory power of input variables of linear regression models (such as polynomial models). In general, these variable screening methods cannot be applied to identify important input variables for nonlinear wing weight models based on the historical wing weight data. Rech et al. [9] proposed a variable selection technique based on polynomial approximations of the nonlinear regression model, but this method only works well when the response can be approximated by a low-degree polynomial. Moreover, their rule-of-thumb is that there are at least about four times as many observations as the number of the coefficients in polynomials. For the case of 41 wing weight data points and eight variables, this rule-of-thumb leads to poor linear polynomial approximations of the wing weight models (6) and (7), rendering the nonlinear variable selection technique unsuitable for variable selection for the nonlinear models (6) and (7).

However, the forward and backward variable selection methods can be formally applied to the ratio model (6) and the geometry model (7) for variable screening. If the adjusted sample coefficient of determination ($\bar{R}^2$) is used to measure the proportion of the total variation in $y_1, \ldots, y_N$ explained by a given model, then the corresponding forward variable selection procedure starts with no variable in the model, and variables are added one by one into the model in such a way that the added variable yields the most significant increase of the adjusted R^2 value of the "expanded" model. For the backward variable selection, start with the model using all the variables and eliminate variables one by one from the model in such a way that the deleted variable yields the least significant decrease of the adjusted R^2 value of the "reduced" model.

For illustration purpose, we use the backward variable selection to check if any input variable in the geometry model (7) is insignificant for wing weight prediction. For convenience, relabel the eight variables in Eq. (7) as $x_1, \ldots, x_n$ (with $n = 8$).

1. *Backward Variable Selection*

1) Let $g(\boldsymbol{x})$ be the least squares fit of the wing weight data by the geometry model (7).

2) Let $g_{-i}(\boldsymbol{x}_{-i})$ be the least squares fit of the data by the simplified geometry model obtained by setting the exponent of the term related to x_i as zero in Eq. (7).

3) Compute the adjusted sample coefficients of determination ($\bar{R}^2$ and $\bar{R}^2_{-i}$) for g and g_{-i} ($i = 1, \ldots, n$):

$$\bar{R}^2 = 1 - \frac{(N-1)\sum_{j=1}^{N}[y_j - g(\boldsymbol{x}^j)]^2}{(N-n)\sum_{j=1}^{N}[y_j - \text{ave}(y)]^2} \quad \text{and}$$

$$\bar{R}^2_{-i} = 1 - \frac{(N-1)\sum_{j=1}^{N}[y_j - g_{-i}(\boldsymbol{x}^j_{-i})]^2}{(N-n+1)\sum_{j=1}^{N}[y_j - \text{ave}(y)]^2}$$

4) If the difference in adjusted sample coefficients of determination $\Delta R^2_{-i} = \bar{R}^2 - \bar{R}^2_{-i}$ is nonpositive for some i, then the corresponding variable x_i could be removed from the input vector and the simplified geometry model would have $(n-1)$ variables.

5) Repeat the process with the simplified geometry model until the number of input variables becomes desirable or all ΔR^2_{-i} is greater than 0.

It is heuristic to use the adjusted R^2 in variable selection. Other metrics can also be used for variable selection (see, e.g., [7,8]). For wing weight fitting by the geometry model, the backward variable selection identifies the ultimate load factor μ as an insignificant variable (see Fig. 4). In fact, if μ is removed from the geometry model (7), then the least squares fitting error only increases to 1.8397×10^8 from 1.8364×10^8. Furthermore, removing both μ and c_t from the geometry model (7) only increases the least squares fitting error to 1.8573×10^8, less than 1% increase of the least squares fitting error by the geometry model (7). In contrast, removing any input variable (except μ and c_t) increases the least squares fitting error by at least 6%. It is interesting that the insignificance of μ is actually due to the fact that the values of μ for civil transports are mandated by FAA regulation, not a design variable determined by engineers. The reason for including μ as a design variable is a legacy inherited from military aircraft weight estimation practices, where μ is a common parameter in the design tradeoff and there is considerably greater variation in the values of μ. Also, as expected, c_t has little impact as it is mostly an aerodynamic, not structural, consideration.

It is worth pointing out that two-dimensional plots of the response with respect to input variables are always helpful to understand how a particular input variable affects the response. Figure 5 shows the relationships of wing weight w with respect to each of the 15 configuration parameters except $[t/c]_m$. Note that 26 subsonic transports have the same μ value of 3.75, also suggesting that the data do not have much information about the relationship between w and μ.

D. Data Scaling

One might wonder why the configuration parameters are scaled in Eqs. (6) and (7). Theoretically, we can rewrite these two equations without scaling. For example,

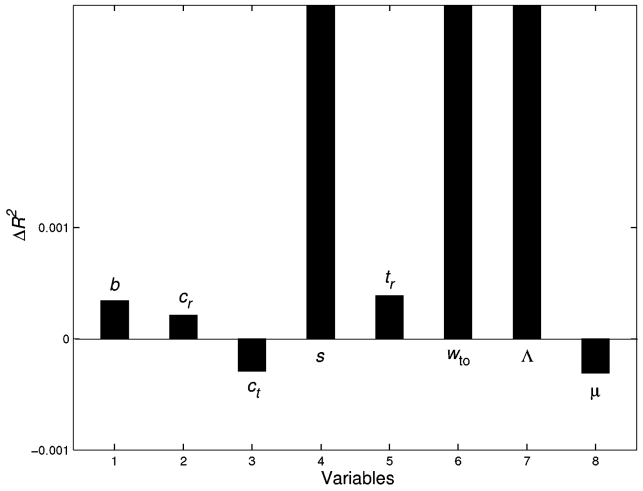

Fig. 4 Backward variable selection for wing weight fitting by the geometry model.

$$\alpha_1[\mu^{\alpha_2}(0.01b)^{\alpha_3}(10^{-3}s)^{\alpha_4}(t_r)^{\alpha_5}(0.1c_r)^{\alpha_6}(\cos\Lambda)^{\alpha_7}(0.1c_t)^{\alpha_8} \times(10^{-5}w_{to})^{\alpha_9}] \equiv \hat{\alpha}_1[\mu^{\alpha_2}b^{\alpha_3}s^{\alpha_4}(t_r)^{\alpha_5}(c_r)^{\alpha_6}(\cos\Lambda)^{\alpha_7}(c_t)^{\alpha_8}(w_{to})^{\alpha_9}]$$

where $\hat{\alpha}_1 = \alpha_1 0.01^{\alpha_3} 10^{-3\alpha_4} 0.1^{\alpha_6} 0.1^{\alpha_8} 10^{-5\alpha_9}$. However, the unscaled geometry model will cause ill-conditioning problems for a nonlinear least squares method to solve the wing weight data fitting problem.

Instead of using heuristic scaling approaches as in Eqs. (6) and (7), we can scale each component x_i by an estimation of its standard deviation σ_i calculated from the data:

$$\sigma_i = \sqrt{\frac{\sum_{j=1}^{N}\left[x_i^j - \text{ave}(x_i)\right]^2}{N-1}}, \quad \text{with ave}(x_i) = \frac{1}{N}\sum_{j=1}^{N} x_i^j$$

Note that users need to be warned of small values of σ_i, say, less than 10% of the mean value ave(x_i). There are two reasons for such a small value of σ_i: 1) the actual range of the variable x_i is about the same magnitude as σ_i or 2) there are not enough data to model the change of the response with respect to x_i. Expert knowledge can be used to determine which case it is. For example, in Table 1, the ultimate load factor has the smallest ratio of $\sigma_i/\text{ave}(x_i)$. The reason is that the μ values for civil transports are usually determined by FAA regulation and do not have much variation. A warning will help analysts to discover that among 41 subsonic transports, 26 of them have the same ultimate load factor of 3.75. Therefore, any relationship between the ultimate load factor and the wing weight based on the given data set might be questionable.

Based on all the analyses performed so far, it seems natural not to include the ultimate load factor as a design variable. The inclusion of μ as a design variable in Eqs. (6) and (7) is a legacy inherited from military aircraft weight estimation practices, where μ is a common parameter in the design tradeoff and there is considerably greater variation in the values of μ. In fact, if we eliminate μ from the geometry model, then the backward variable screening guarantees a marginal increase of fitting error (less than 1% as we mentioned before). The statistical analysis of mean and variance in the data only shows poor distribution of μ values that could adversely impact any attempt to find a meaningful relationship between wing weight and μ. In this paper, we are trying to find an accurate wing weight estimation model that provides desirable trends for each of the sizing design variables, span, reference area, taper ratio, and sweep angle (cf. Subsection II.H), for some nominal value of μ. Because we will use interpolation methods to build wing weight models, there is no theory on whether a particular variable is important for accuracy of the fitting models. For each wing weight interpolation model included in this paper, we also generated the corresponding interpolation model without μ. In almost all the cases, the interpolation models with μ have more desirable trends between and beyond the data points than those without μ. As a result, μ will be included as an input variable for RBF wing weight models in this paper.

Table 1 Estimated means and standard deviations of configuration variables

Index	Variable	Min	Max	Mean [ave(x_i)]	Deviation (σ_i)
1	A	0.3	12.4	8.86	2.3
2	b	26.11	222.7	122.1	40.22
3	c_m	7.78	86.17	16.26	12.83
4	c_r	11.15	54.39	22.2	10.66
5	c_t	3.62	16.16	7.35	2.96
6	s	542.5	8,200	2,019	1,589
7	t_r	1.56	9.75	3.42	1.45
8	t_t	0.34	1.65	0.8	0.25
9	t_r/c_r	0.11	0.22	0.16	0.03
10	t_t/c_t	0.06	0.17	0.12	0.03
11	$[t/c]_m$	0.08	0.18	0.13	0.03
12	w_{to}	26,000	800,000	163,806	175,787
13	λ	0.20	0.61	0.35	0.1
14	Λ	0	55	11.91	15.83
15	μ	3.75	5.3	4.06	0.46

When RBF interpolation method is used for data fitting, scaling each data component by its estimated standard deviation also helps to improve the conditioning of the interpolation matrix. For illustration purpose, consider the multiquadric RBF interpolation of the wing weight data. The multiquadric RBF model is defined by

$$g(\boldsymbol{x}) = \sum_{j=1}^{N} \alpha_j \varphi[d(\boldsymbol{x} - \boldsymbol{x}^j)] \tag{8}$$

where $\varphi(t) = \sqrt{1+t^2}$ is the multiquadric RBF, and $d(\boldsymbol{x} - \boldsymbol{x}^j)$ is a parameterized distance between $\boldsymbol{x}$ and $\boldsymbol{x}^j$ defined as

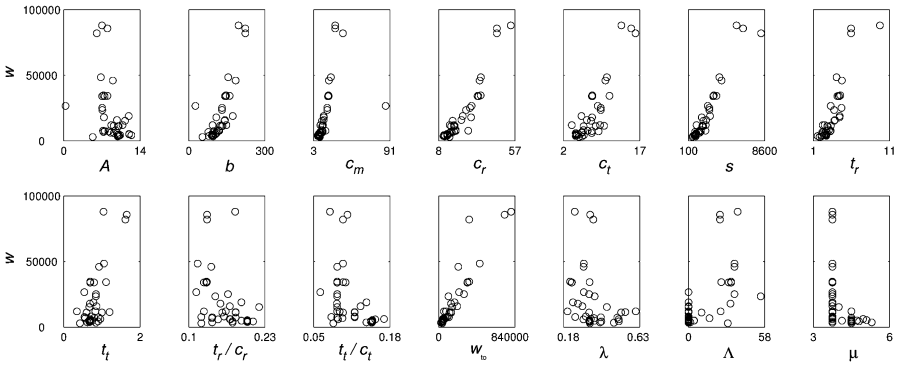

Fig. 5 Two-dimensional plots of wing weight vs configuration parameters.

$$d(\boldsymbol{x}-\boldsymbol{x}^j)=\sqrt{\sum_{i=1}^{n}\theta_i\left(\frac{x_i-x_i^j}{\sigma_i}\right)^2} \tag{9}$$

The coefficients $\alpha_1,\ldots,\alpha_N$ in Eq. (8) are determined by interpolation conditions $g(\boldsymbol{x}^k)=y_k$ $(k=1,\ldots,N)$, where y_k denotes the wing weight of the kth subsonic transport. The scalars $\theta_1,\ldots,\theta_n$ in Eq. (9) are positive scalars that can be determined by minimizing the cross-validation errors for the given data.

Without scaling, one could rewrite $d(\boldsymbol{x}-\boldsymbol{x}^j)$ as

$$d(\boldsymbol{x}-\boldsymbol{x}^j)=\sqrt{\sum_{i=1}^{n}\bar{\theta}_i\left(x_i-x_i^j\right)^2},\quad \text{with } \bar{\theta}_i=\frac{\theta_i}{\sigma_i^2} \tag{10}$$

However, in practice, starting without any scaling [i.e., setting $\bar{\theta}_i=1$ in Eq. (10)] may lead to ill-conditioning of the interpolation problem even though the mathematical theory guarantees the existence of a unique multiquadric RBF interpolant for any given data points $(\boldsymbol{x}^1,y_1),\ldots,(\boldsymbol{x}^N,y_N)$. For eight input variables given in (7) and 41 data points, the condition number of the unscaled interpolation matrix is 2.8×10^8, whereas the scaled interpolation matrix [corresponding to $\theta_i=1$ in Eq. (9)] has a condition number of 2.7×10^5. The purpose of using two sets of scaling parameters in Eq. (9) is to allow a nondimensional initial choice of $\theta_i=1$.

E. Principal Component Analysis of Wing Weight Data

For a limited number of historical or measurement data points in a high-dimensional input space, a principal component analysis (PCA) is recommended to check any collinearity of the input attributes of the data points. The importance of using PCA for irregularly distributed data points was discussed in Subsection 3.6 of [10].

For the wing weight data, not all the variables in (4) are truly independent because subsonic wings were not built with randomly generated configuration parameters. As a result, it is important to check whether there is any linear dependency among the parameters for the given data by using PCA. PCA of the data could lead to a potential reduction of the dimension of the input space, resulting in better wing weight models.

The wing configuration of each subsonic transport consists of 15 configuration parameters: b, c_m, c_r, c_t, A, λ, $[t/c]_m$, t_r/c_r, t_t/c_t, s, t_r, t_t, w_{to}, Λ, and μ. Because $[t/c]_m=(t_r/c_r+t_t/c_t)/2$, the three configuration parameters $[t/c]_m$, t_r/c_r, and t_t/c_t are linearly dependent. The redundant parameter $[t/c]_m$ is considered as a potential input variable due to analysts' preference of using $[t/c]_m$ as an input variable instead of t_r/c_r and t_t/c_t [see Eq. (6)].

The PCA of the 15 configuration parameters is done as follows. First, relabel the 15 parameters as variables $x_1,\ldots,x_n$ and the 41 wing configurations as $\boldsymbol{x}^1,\ldots,\boldsymbol{x}^N$. Scale each variable by its estimated standard deviation: $\hat{x}_i^j=x_i^j/\sigma_i$.

Next calculate the covariance matrix $\boldsymbol{C}$ of the scaled input vectors $\hat{\boldsymbol{x}}^1,\ldots,\hat{\boldsymbol{x}}^N$:

$$\boldsymbol{C}=\frac{1}{N-1}\sum_{j=1}^{N}[\hat{\boldsymbol{x}}^j-\text{ave}(\hat{\boldsymbol{x}})][\hat{\boldsymbol{x}}^j-\text{ave}(\hat{\boldsymbol{x}})]^T$$

where $\text{ave}(\hat{\boldsymbol{x}})=\frac{1}{N}\sum_{j=1}^{N}\hat{\boldsymbol{x}}^j$. Then the following spectral decomposition of $\boldsymbol{C}$ can be used to analyze the collinearity of the input parameters: $\boldsymbol{C}=\sum_{i=1}^{n}\tau_i\boldsymbol{u}^i(\boldsymbol{u}^i)^T$, where $\tau_1\ge\tau_2\ge\cdots\ge\tau_n\ge0$ are the eigenvalues of $\boldsymbol{C}$, and $\boldsymbol{u}^1,\ldots,\boldsymbol{u}^n$ are the corresponding unit eigenvectors. In the PCA terminology, the unit vector $\boldsymbol{u}^i$ is called the ith feature vector of the sample set $\hat{\boldsymbol{x}}^1,\ldots,\hat{\boldsymbol{x}}^N$ and the scalar $\hat{\boldsymbol{x}}^T\boldsymbol{u}^i$ is called the ith principal component of $\hat{\boldsymbol{x}}$.

For the 15 wing configuration parameters of the 41 subsonic transports listed in Table 1, the eigenvalues of $\boldsymbol{C}$ are 7.952, 2.757, 1.349, 0.861, 0.676, 0.565, 0.341, 0.187, 0.157, 0.096, 0.031, 0.019, 0.005, 0.003, and 0.000. The last eigenvalue 0 means linear dependence of the 15 configuration parameters due to the linear relationship $[t/c]_m=(t_r/c_r+t_t/c_t)/2$. In fact, the three components of $\boldsymbol{u}^{15}$ corresponding to $[t/c]_m$, t_r/c_r, t_t/c_t are -0.45, -0.43, and 0.78, whereas the remaining components of $\boldsymbol{u}^{15}$ are zero (accurate up to two significant digits). The next two smallest eigenvalues 0.003 and 0.005 also indicate nearly collinear relationships among the 15 configuration parameters because of specific locations of the 41 input vectors in the 15-dimensional space. Such data-specific relationships are most likely to disappear when new wing weight data are added to the existing data set. However, one should exercise caution when using a weight prediction of a wing configuration represented by $\hat{\boldsymbol{x}}$ with relatively large absolute values of $[\hat{\boldsymbol{x}}-\text{ave}(\hat{\boldsymbol{x}})]^T\boldsymbol{u}^{13}$ or $[\hat{\boldsymbol{x}}-\text{ave}(\hat{\boldsymbol{x}})]^T\boldsymbol{u}^{14}$, because the data do not contain much information on how the wing weight changes in terms of these two quantities.

If t_r/c_r and t_t/c_t are excluded from the list of wing configuration parameters, then the 13 eigenvalues of the corresponding $\boldsymbol{C}$ are 7.17, 2.17, 1.33, 0.758, 0.631, 0.312, 0.244, 0.179, 0.123, 0.047, 0.021, 0.016, and 0.004. It also suggests that one nearly collinear relationship could be used to reduce the dimension of the input space to 12, which is the same conclusion from the previous PCA.

If the PCA is applied to analyze the collinearity of the configuration parameters in the two semi-empirical models (6) and (7), then the smallest eigenvalue of $\boldsymbol{C}$ is 0.146 for configuration parameters in model (6). However, for configuration parameters in model (7), the two smallest eigenvalues of $\boldsymbol{C}$ are 0.082 and 0.019, which indicates a higher level of collinearity among the input variables in Eq. (7) than in Eq. (6).

Dimension reduction of the input space can be achieved by choosing the first m $(<n)$ principal components $[\hat{\boldsymbol{x}}-\text{ave}(\hat{\boldsymbol{x}})]^T\boldsymbol{u}^1,\ldots,[\hat{\boldsymbol{x}}-\text{ave}(\hat{\boldsymbol{x}})]^T\boldsymbol{u}^m$ as the significant input variables. If the data points cluster around an m-dimensional subspace of the n-dimensional input space, then the data only contain information on how particular linear combinations of the original input variables might affect the wing weight. See [11] or Subsection 3.6 of [10] for details on reformulating the data fitting problem in the m-dimensional feature space.

F. Subjective Choice of Input Variables

For the wing weight data fitting problem, the initial set of candidates for input variables is a completely subjective choice. There are many different ways to select a set of independent configuration parameters, such as the two sets of parameters in Eqs. (6) and (7). Even though replacing A by b in Eq. (6) yields a mathematically identical model with appropriate choices of regression coefficients $\bar{\alpha}_1,\ldots,\bar{\alpha}_8$, the same replacement will lead to a completely different wing weight model if a general approximation model is used. For example, if a quadratic polynomial P_2 is used as a wing weight model, then $P_2(A,s)=a_0+a_1A+a_2s+a_3A^2+a_4As+a_5s^2$ and $P_2(b,s)=\bar{a}_0+\bar{a}_1b+\bar{a}_2s+\bar{a}_3b^2+\bar{a}_4bs+\bar{a}_5s^2$ are two completely different models no matter what are the values of a_i and $\bar{a}_i$. Similarly, different sets of input variables of a general approximation model (such as polynomials, radial basis functions, and Kriging model) will lead to different wing weight models. Therefore, one must exercise caution in selecting input variables for wing weight models.

To understand the effect of input variables, we shall fit the wing weight data by the following two function models:

$$\begin{aligned}\bar{w}&=g\left(b,c_r,s,\frac{t_r}{c_r},w_{\text{to}},\lambda,\Lambda,\mu\right)\quad\text{or}\\ \bar{w}&=g\left(A,b,c_m,c_r,c_t,s,t_r,t_t,\frac{t_r}{c_r},\frac{t_t}{c_t},w_{\text{to}},\lambda,\Lambda,\mu\right)\end{aligned} \tag{11}$$

The first model uses the input variables in Eq. (7), which are based on system analysts' knowledge of wing weight estimation. The second model uses all the parameters in the wing weight data (except the average of t_r/c_r and t_t/c_t). The second model is used for a comparison study with the first model to understand how useful the expert knowledge is for the wing weight data fitting. To eliminate any dependency among the input variables, the final wing weight estimation formula for conceptual design analysis should be either

$$\bar{w} = g\left(b, \frac{\gamma s}{b(1+\lambda)}, s, \frac{t_r}{c_r}, w_{\text{to}}, \lambda, \Lambda, \mu\right) \tag{12}$$

or

$$\bar{w} = g\left(\frac{b^2}{s}, b, \frac{s}{b}, \frac{\gamma s}{b(1+\lambda)}, \frac{\gamma\lambda s}{b(1+\lambda)}, s, \frac{t_r}{c_r}\frac{\gamma s}{b(1+\lambda)}, \frac{t_t}{c_t}\frac{\gamma\lambda s}{b(1+\lambda)}, \frac{t_r}{c_r}, \frac{t_t}{c_t}, w_{\text{to}}, \lambda, \Lambda, \mu\right) \tag{13}$$

Note that Eqs. (12) and (13) capture the dependency relationships between the wing weight and the input variables in Eq. (11) by using only the independent variables in (4).

G. Interpolation of Wing Weight Data

For the given wing weight data, it is desirable to fit all the data points exactly, i.e., interpolation of the wing weight data is desirable. Therefore, we shall use four different interpolation methods for the wing weight data fitting: least polynomial interpolation (LPI) [12–14], RBF interpolation [15–17], Kriging interpolation [18], and Gaussian process [19]. Moreover, PCR using RBF interpolation [11] is also used for the wing weight data fitting.

Let $\varphi_1(\boldsymbol{x}), \ldots, \varphi_N(\boldsymbol{x})$ be functions of $\boldsymbol{x}$. Then the multivariate interpolation of the data $\{(\boldsymbol{x}^1, y_1), \ldots, (\boldsymbol{x}^N, y_N)\}$ by linear combinations of $\varphi_1(\boldsymbol{x}), \ldots, \varphi_N(\boldsymbol{x})$ is to find a function $\sum_{j=1}^N \alpha_j \varphi_j(\boldsymbol{x})$ that satisfies the following interpolation conditions:

$$\sum_{j=1}^{N} \alpha_j \varphi_j(\boldsymbol{x}^k) = y_k \quad \text{for } k = 1, \ldots, N \tag{14}$$

For RBF interpolation methods, we could use any of the following RBF: cubic polynomial RBF $\varphi(t) = t^3$, thin plate spline RBF $\varphi(t) = t^2 \log t$, multiquadric RBF $\varphi(t) = \sqrt{1+t^2}$, and Gaussian RBF $\varphi(t) = \exp(-t^2)$, which can be used to model cubic, (approximately) quadratic, and linear growth rates of the response as well as exponential decay of the response for trend predictions. For multiquadric RBF interpolation, $\varphi_j(\boldsymbol{x}) = \varphi[d(\boldsymbol{x} - \boldsymbol{x}^j)]$ [see Eq. (8)]. Except the least polynomial interpolation, all other interpolation models use RBFs and have intrinsic model parameters $\theta_1, \ldots, \theta_n$ in $d(\boldsymbol{x} - \boldsymbol{x}^j)$ [see Eq. (9)]. The Gaussian process uses the maximum likelihood for determining the optimal set of intrinsic model parameters in Gaussian RBF, whereas cross-validation error minimization is used to find the optimal model parameters for RBF interpolation and Kriging interpolation.

One could also use PCA to transform the data into a feature space and use an interpolation method in the feature space. The principal component regression using RBF interpolation [11] is such an example. Specifically, let $\boldsymbol{v}$ be the vector with m components $[\hat{\boldsymbol{x}} - \text{ave}(\hat{\boldsymbol{x}})]^T \boldsymbol{u}^1, \ldots, [\hat{\boldsymbol{x}} - \text{ave}(\hat{\boldsymbol{x}})]^T \boldsymbol{u}^m$. Then the N data points in the $\boldsymbol{x}$-space can be mapped to N points in the $\boldsymbol{v}$-space. The corresponding problem of fitting N data points $(\boldsymbol{v}^1, y_1), \ldots, (\boldsymbol{v}^N, y_N)$ can be solved by using a standard interpolation method. If the solution of the data fitting problem in the $\boldsymbol{v}$-space is $\hat{g}(\boldsymbol{v})$, then the corresponding fitting of the data points in the $\boldsymbol{x}$-space is

$$\bar{w} = \hat{g}\{[\hat{\boldsymbol{x}} - \text{ave}(\hat{\boldsymbol{x}})]^T \boldsymbol{u}^1, \ldots, [\hat{\boldsymbol{x}} - \text{ave}(\hat{\boldsymbol{x}})]^T \boldsymbol{u}^m\} \tag{15}$$

where $\hat{x}_i = x_i/\sigma_i$ are scaled variables.

By using different interpolation methods and different input spaces, we can generate hundreds of wing weight interpolation models for the wing weight data. Each of these wing weight models fits the given data exactly, i.e., its value at $\boldsymbol{x}^k$ is exactly y_k for $k = 1, \ldots, N$. Therefore, engineering criteria for desirable trends of a wing weight model between and beyond the data points shall be used to verify which wing weight model is useful for conceptual design of subsonic transports.

H. Verification of Wing Weight Models

In general, a desirable wing weight model should have the following properties: the estimated wing weight w is an increasing function with respect to each of b, s, λ, and Λ; and w is a decreasing function with respect to t_r/c_r. These properties are derived from simple engineering rules on the relationships between the wing weight and each of configuration parameters b, s, λ, Λ, and t_r/c_r. All the constructed wing weight models will be evaluated against these empirical rules. Of course, a more meaningful validation of the usefulness of a constructed approximation is to see whether conceptual aircraft design using the constructed approximation for weight estimation can produce better aircraft concepts, but such a task is beyond the scope of this paper.

III. Numerical Results

The numerical experiments are designed to demonstrate how input space, interpolation model, and PCR would affect the behaviors of the resulting wing weight model between and beyond the data points. Two-dimensional plots are used to compare various wing weight estimation models including the geometry model (7), least polynomial interpolation, Kriging interpolation, Gaussian process, Gaussian RBF interpolation, multiquadric RBF interpolation, and multiquadric PCR.

Each curve is generated by fixing all the input variables in Eqs. (12) or (13) to the corresponding values of one transport configuration (labeled as baseline in figures) except the variable used in the horizontal axis of the plot. For each of the five input variables b, s, t_r/c_r, λ, and Λ, we examine 41 curves corresponding to the 41 transports for wing weight trend vs the input variable. The selected plot shows a representative trend in these 41 curves. In some plots, we also include the 41 data points. Because each plotted point shows only the wing weight and the value of the specified input variable, two points close to each other do not necessarily mean similar configurations. Note that each curve is the intersection of the graph of a wing weight estimation function and a two-dimensional plane (called plotting plane hereafter). To give a perception on how far a data point is from the plotting plane, we divide the data points into three groups: 1) the baseline, 2) distance to the plotting plane ≤ 0.8, and 3) distance to the plotting plane >0.8. Here the distance from a data point to the plotting plane is computed by using the following normalized distance between two input vectors of seven components:

$$\left[\left(\frac{b-b_j}{\sigma(b)}\right)^2 + \left(\frac{s-s_j}{\sigma(s)}\right)^2 + \left(\frac{t_r/c_r - t_{r,j}/c_{r,j}}{\sigma(t_r/c_r)}\right)^2 + \left(\frac{w_{\text{to}} - w_{\text{to},j}}{\sigma(w_{\text{to}})}\right)^2 + \left(\frac{\lambda-\lambda_j}{\sigma(\lambda)}\right)^2 + \left(\frac{\Lambda-\Lambda_j}{\sigma(\Lambda)}\right)^2 + \left(\frac{\mu-\mu_j}{\sigma(\mu)}\right)^2\right]^{1/2}$$

where the standard deviation $\sigma(\cdot)$ of each variable is estimated by using the 41 data points.

A. Avoiding Chord Approximation Errors

The chord approximation (5) is not accurate and could have an adverse effect on wing weight models if improperly used. Here we use the multiquadric PCR fitting to show what might go wrong when Eq. (11) is used in trade studies.

For an interpolation model $g(\boldsymbol{x})$ with the eight input variables in Eq. (7) as the components of $\boldsymbol{x}$, it is natural to define the data points $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^N$ as follows:

$$x_1^j = b_j, \quad x_2^j = c_{r,j}, \quad x_3^j = s_j, \quad x_4^j = \frac{t_{r,j}}{c_{r,j}}$$
$$x_5^j = w_{\text{to},j}, \quad x_6^j = \lambda_j, \quad x_7^j = \Lambda_j, \quad \text{and} \quad x_8^j = \mu_j \tag{16}$$

The corresponding interpolation conditions are

$$g\left(b_j, c_{r,j}, s_j, \frac{t_{r,j}}{c_{r,j}}, w_{\text{to},j}, \lambda_j, \Lambda_j, \mu_j\right) = w_j, \quad \text{for } 1 \leq j \leq N \tag{17}$$

However, in a trade study of w vs s near an existing subsonic transport, say the kth transport, the two-dimensional plot shows the following relationship:

$$\bar{w} = g\left(b_k, \frac{\gamma s}{b_k(1+\lambda_k)}, s, \frac{t_{r,k}}{c_{r,k}}, w_{\text{to},k}, \lambda_k, \Lambda_k, \mu_k\right) \tag{18}$$

where c_r is replaced by the chord approximation formula in Eq. (5).

Equation (5) is not accurate. For the 41 subsonic transports, the standard deviation of $|c_{r,j} - \gamma s_j/[b_j(1+\lambda_j)]|$ is 73% of the average of $c_{r,j}$. In the worst case, $|c_{r,j} - \gamma s_j/[b_j(1+\lambda_j)]|$ is 3.5 times of $c_{r,j}$, which leads to a significant difference between $w_k = g(b_k, c_{r,k}, s_k, t_{r,k}/c_{r,k}, w_{\text{to},k}, \lambda_k, \Lambda_k, \mu_k)$ and $g(b_k, \gamma s_k/[b_k(1+\lambda_k)], s_k, t_{r,k}/c_{r,k}, w_{\text{to},k}, \lambda_k, \Lambda_k, \mu_k)$.

To avoid chord approximation errors in trade studies, we shall use the following interpolation conditions if the variables in Eq. (7) are used as the input variables:

$$g\left(b_j, \frac{\gamma s_j}{b_j(1+\lambda_j)}, s_j, \frac{t_{r,j}}{c_{r,j}}, w_{\text{to},j}, \lambda_j, \Lambda_j, \mu_j\right) = w_j, \quad \text{for } 1 \le j \le N$$

i.e., the chord approximation formula shall be used before the data fitting. Similarly, if all the variables in Table 1 except $[t/c]_m$ are used as the input variables, then the following interpolation conditions shall be used to construct wing weight estimation models:

$$g\left(\frac{b_j^2}{s_j}, b_j, \frac{s_j}{b_j}, \frac{\gamma s_j}{b_j(1+\lambda_j)}, \frac{\gamma \lambda_j s_j}{b_j(1+\lambda_j)}, s_j, \frac{t_{r,j}}{c_{r,j}}\frac{\gamma s_j}{b_j(1+\lambda_j)}, \times \frac{t_{t,j}}{c_{t,j}}\frac{\gamma \lambda_j s_j}{b_j(1+\lambda_j)}, \frac{t_{r,j}}{c_{r,j}}, \frac{t_t}{c_t}, w_{\text{to},j}, \lambda_j, \Lambda_j, \mu_j\right) = w_j$$

for $1 \le j \le N$. In other words, we shall use either Eq. (12) or (13) directly for wing weight data fitting instead of Eq. (11).

Using the chord approximation before the data fitting not only makes the final wing weight model in Eq. (12) or (13) reproduce the exact wing weight for each existing subsonic transport, but also leads to wing weight models with more desirable two-dimensional trends. See Fig. 6 for typical plots of the wing weight vs the reference area or span, where the wing weight model generated by applying the chord approximation before multiquadric PCR fitting has the most desirable trends. In fact, both span and reference area plots presented in Fig. 6 show that the wing weight model generated by applying the chord approximation before multiquadric PCR (solid line) has the most desirable trends among the three models: it is an increasing function of reference area and it is also an increasing function of span except for very small unrealistic span values (less than 40 ft). (Note that for a baseline with span of about 125 ft, a new configuration with the same reference area and span of 40 ft will have to increase the chord lengths of the baseline by 200%, which is absurd.) In contrast, the geometry model has the desired trend for reference area but a wrong trend for span; and the chord approximation after multiquadric PCR just does not have desirable trends for these two sizing design variables.

B. Effect of Subjective Variable Selection

It would be ideal if PCR would not be affected by inclusion of nonessential variables as input variables of a wing weight model. The reality is that expert knowledge does make a difference in the generated wing weight model. The following plots compare the multiquadric PCR fittings corresponding to wing weight models (12) and (13), as well as the multiquadric RBF interpolants. In almost all 2-D plots, the fittings generated with the eight input variables have more desirable trends between and beyond the data points than the fittings generated with the 14 input variables. Two plots in Fig. 7 show the wing weight trends with respect to the reference area and span, respectively. For the wing weight data fitting problem, if a wrong set of input variables is used in data fitting, then the generated wing weight model might have poor trend predictions no matter which method is used. Among all the models in Fig. 7, only the multiquadric RBF fitting with eight variables (solid line) has the desired trends for both sizing design variables: wing weight increases as either reference area or span increases in a reasonable range.

C. Comparison of Interpolation Models

Figure 8 shows two-dimensional sample plots of wing weight models generated by different interpolation methods. For the wing weight fitting problem, Kriging and Gaussian RBF interpolants tend to create unnecessary oscillations between data points. One reason is that the basis functions $\varphi[d(\boldsymbol{x} - \boldsymbol{x}^i)]$ decreases at the exponential rate as $d(\boldsymbol{x} - \boldsymbol{x}^i)$ increases (or the configuration is away from the existing

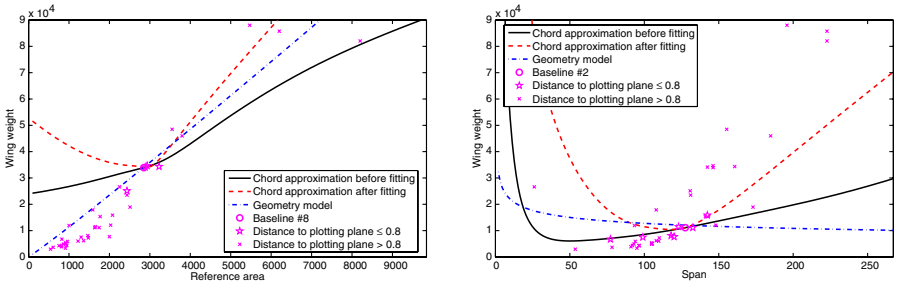

Fig. 6 Differences between using chord approximation formula (3) before and after multiquadric PCR fitting.

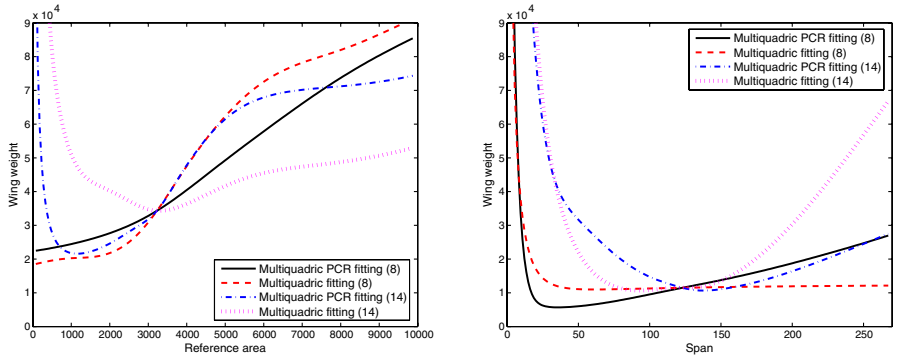

Fig. 7 Differences between various approximations generated by using two different sets of input variables.

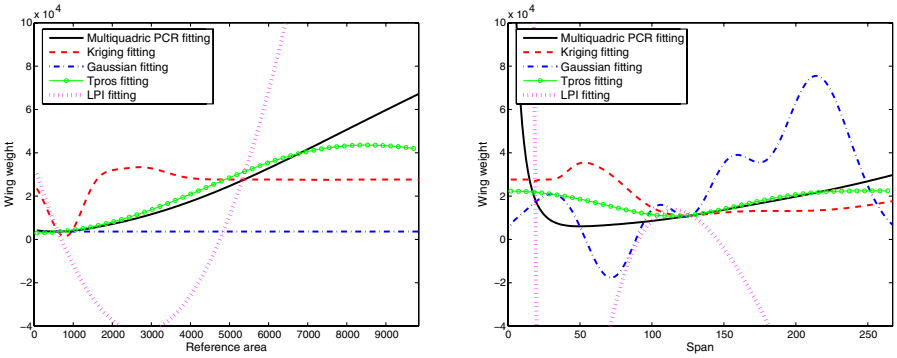

Fig. 8 Two-dimensional plots of different approximation models.

configurations). LPI is extremely sensitive with respect to data points and LPI of the 41 wing data points is very oscillatory. In comparison, an appropriate choice (multiquadric) of RBFs leads to nonoscillatory trend predictions of the approximation. Gaussian process generates a Gaussian RBF interpolant as the wing weight model, which is labeled as Tpros fitting in Fig. 8 because the software code, Tpros, by Gibbs and MacKay is used to generate the solution [19]. The only difference between Tpros fitting and the standard Gaussian RBF interpolant (labeled as Gaussian fitting in Fig. 8) is that the former uses maximum likelihood whereas the latter uses minimization of cross-validation error for model parameter tuning. For the wing weight data fitting problem, Tpros fitting is better than the Gaussian fitting; however, one could see the eventual decay of Tpros fitting as the reference area or span is at the upper end of the plotted range. For the given wing weight fitting problem, by inspecting the generated plots, the multiquadric PCR fitting has the most desirable trends with respect to reference area or span.

D. Benefits of Principal Component Regression

The motivation of using PCR is to customize the fitting model to capture the trends in the wing weight data correctly. In most cases we examined, the multiquadric PRC fitting has more desirable trend predictions than the other wing weight models, as shown in the following four plots in Fig. 9, as well as those in Figs. 6–8.

Note that Fig. 9 presents a typical plot selected from a large set of plots inspected. Almost all the 2-D plots for the multiquadric PCR fitting exhibit the desirable properties specified in Subsection II.H, at least in a neighborhood of the baseline data point. In few cases, the wing weight is not a decreasing function of thickness-to-chord ratio at root. In many cases, the wing weight is not an increasing function of taper ratio. No generated wing weight model is capable of showing an increase of wing weight as taper ratio decreases. Perhaps the wing weight data do not have sufficient information about the relationship between wing weight and taper ratio. Note that the 2-D plot for wing weight vs taper ratio in Fig. 5 shows a random relationship between wing weight and taper ratio. In general, the multiquadric PRC fitting has more desirable 2-D trends than the geometry model (7), with respect to the criteria given in Subsection II.H. For example, in the 2-D plot of wing weight vs span, the geometry model (7) has the wrong trend whereas the multiquadric PRC fitting has the desired trend before span becomes unrealistically small.

IV. Concluding Remarks

It is very easy to fit a set of data exactly by various interpolation methods no matter how the data points are distributed, but interpolation models can be drastically different between and beyond the data points, even if they are identical at the given data points. For a set of wing weight data of 41 subsonic transports, we compare the wing weight models generated by a semi-empirical regression method, radial basis function interpolation, Kriging interpolation, Gaussian process, least polynomial interpolation, and principal component regression using multiquadric RBF. All of the intrinsic model parameters for RBFs are determined by minimizing the cross-validation errors, except that the Gaussian process uses the maximum likelihood to determine the intrinsic model parameters for Gaussian RBF.

In the conceptual design of subsonic transports, a generally desired feature for a useful wing weight model is monotonicity (i.e., either increasing or decreasing) of the predicted wing weight with respect to each of wing configuration parameters $b, s, t_r/c_r, \lambda$, and Λ. If we relabel the input variables as $\boldsymbol{x}$, then a wing weight model is a multivariate function $g(\boldsymbol{x})$. To check monotonicity of $g(\boldsymbol{x})$ with respect to an input variable x_i, we first fix all the other variables x_j $(j \neq i)$ and then plot $g(\boldsymbol{x})$ as a function of x_i. However, it is impossible to check all the 2-D curves of $g(\boldsymbol{x})$ with respect to x_i for infinitely many different values of x_j $(j \neq i)$. Therefore, we only check those 2-D curves that pass through one of the 41 data points

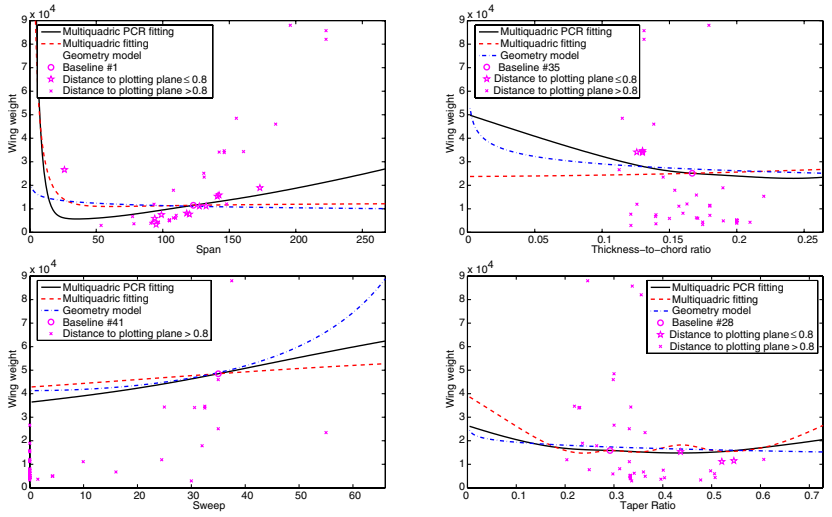

Fig. 9 Differences between multiquadric PCR fitting and multiquadric fitting.

(corresponding to the 41 subsonic transports). Except for the relationship between w and λ, almost all 2-D curves for the PCR using multiquadric RBF exhibit more desirable monotone trends than the other wing weight models. No wing weight model (including the semi-empirical regression model) shows a monotone relationship between w and λ, which might indicate a lack of information on the relationship between w and λ in the data.

Principal component analysis that identifies the feature variables for the given data is a powerful tool to customize a wing weight model in an appropriate feature space for better trend predictions. PCR using multiquadric RBF interpolation incorporates data mining in a standard approximation process so that the resulting approximation is more likely to capture the correct trends in the data. Least polynomial interpolant is more likely to oscillate between the data points than RBF interpolants. Among RBF interpolants, Gaussian RBF and Kriging interpolants are more likely to perform poorly due to the exponential rate of decay of the Gaussian RBF as the distances between the estimation point and the data points increase.

Variable screening could be a powerful tool for reducing the dimension of the input space of approximation models if applicable. For wing weight data fitting, most of the commonly used variable screening methods cannot be applied. If a parametric regression model (such as a semi-empirical regression model) is used for data fitting, backward variable selection could be useful to identify insignificant variables. In general, expert knowledge on significant input variables could have a significant effect on whether the constructed approximation model captures physical trends buried in the data.

The significance of this study is that PCR using RBF interpolation method is able to generate a general approximation model for wing weight estimation that is more accurate and has more desirable properties than the best semi-empirical wing weight model available. Even though the benefits of using PCR are only demonstrated by the wing weight data fitting problem, the methodology could have significant advantages in fitting other historical or sparse data.

Appendix A: Multiquadric PCR Fitting Formula

The wing weight model based on the multiquadric PCR fitting can be expressed as follows:

$$\begin{aligned}\bar{w} &= g\left(b, \frac{\gamma s}{b(1+\lambda)}, s, \frac{t_r}{c_r}, w_{\text{to}}, \lambda, \Lambda, \mu\right)\\ &= g(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8) = \hat{g}\{[(\hat{\boldsymbol{x}} - \text{ave}(\hat{\boldsymbol{x}})]^T \boldsymbol{u}^1, \ldots, [\hat{\boldsymbol{x}} - \text{ave}(\hat{\boldsymbol{x}})]^T \boldsymbol{u}^7\} = \hat{g}(v_1, v_2, v_3, v_4, v_5, v_6, v_7) = \hat{g}(\boldsymbol{v})\\ &= \sum_{j=1}^{41} \alpha_j \varphi[d(\boldsymbol{v} - \boldsymbol{v}^j)] = \sum_{j=1}^{41} \alpha_j \sqrt{1 + \sum_{i=1}^{7} \theta_i \left(v_i - v_i^j\right)^2} \quad \text{(A1)}\end{aligned}$$

The variables, parameters, and fitting coefficients in Eq. (A1) are defined as follows.

Choice of Input Variables: $x_1 = b$ (span), $x_2 = \gamma s/[b(1+\lambda)]$ (the approximation of chord length at root) with $\gamma = 1.9739$ (the average of $(c_r + c_t)/c_m$ values for the 41 existing subsonic transports), $x_3 = s$ (reference area), $x_4 = t_r/c_r$ (thick-to-chord ratio at root), $x_5 = w_{\text{to}}$ (gross weight), $x_6 = \lambda$ (taper ratio), $x_7 = \Lambda$ (sweep angle in radius), and $x_8 = \mu$ (ultimate load factor).

Mean and Standard Deviation of Input Variables: Let ave($\boldsymbol{x}$) be the mean of the sample vectors. That is, ave($\boldsymbol{x}$) denotes the column vector whose i component is the mean of $x_i^1, \ldots, x_i^{41}$. Then

$$\begin{aligned}&\text{ave}(\boldsymbol{x})_1 = 122.10, \qquad \text{ave}(\boldsymbol{x})_2 = 24.124\\ &\text{ave}(\boldsymbol{x})_3 = 2019.0, \qquad \text{ave}(\boldsymbol{x})_4 = 0.1614\\ &\text{ave}(\boldsymbol{x})_5 = 163806, \qquad \text{ave}(\boldsymbol{x})_6 = 0.3519\\ &\text{ave}(\boldsymbol{x})_7 = 11.915, \qquad \text{ave}(\mathbf{x})_8 = 4.0612\end{aligned}$$

Let σ_i be the standard deviation of the sample points $x_i^1, \ldots, x_i^{41}$:

$$\begin{aligned}&\sigma_1 = 40.225, \qquad \sigma_2 = 19.710, \qquad \sigma_3 = 1589.4\\ &\sigma_4 = 0.0299, \qquad \sigma_5 = 175787, \qquad \sigma_6 = 0.0953\\ &\sigma_7 = 15.834, \qquad \sigma_8 = 0.4601\end{aligned}$$

Note that ave($\boldsymbol{x}$)$_2$ and σ_2 are different from the sample mean and standard deviation of the chord length at root (cf. Table 1) because the chord approximation formula is used.

Normalized Input Variables: Let $[\hat{\boldsymbol{x}} - \text{ave}(\hat{\boldsymbol{x}})]$ be the component scaling of $[\boldsymbol{x} - \text{ave}(\boldsymbol{x})]$ by $\sigma_1, \ldots, \sigma_8$, i.e.,

$$[\hat{\boldsymbol{x}} - \text{ave}(\hat{\boldsymbol{x}})]_i = [x_i - \text{ave}(\boldsymbol{x})_i]/\sigma_i \quad \text{for } 1 \le i \le 8 \qquad \text{(A2)}$$

Selected Feature Vectors: Let $\boldsymbol{u}^1, \ldots, \boldsymbol{u}^7$ be the seven eigenvectors corresponding to the largest seven eigenvalues of the covariance matrix $\boldsymbol{C}$ defined by the sample vectors $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^{41}$. Then

$$(\boldsymbol{u}^1, \boldsymbol{u}^2, \ldots, \boldsymbol{u}^7) = \begin{pmatrix} -0.2291 & 0.3122 & -0.0018 & 0.0134 & -0.1162 & 0.5704 & -0.3528\\ -0.2574 & -0.2633 & -0.2429 & -0.1802 & 0.6830 & -0.3455 & -0.2907\\ -0.3545 & 0.1703 & 0.0332 & -0.1499 & 0.2003 & 0.3035 & -0.4406\\ -0.1801 & -0.5427 & -0.5800 & -0.0019 & -0.0748 & 0.4800 & 0.3083\\ 0.8148 & -0.1938 & -0.1289 & -0.1883 & 0.0763 & 0.2182 & -0.4381\\ 0.0810 & -0.2525 & 0.6057 & 0.2670 & 0.5161 & 0.3986 & 0.2557\\ -0.2290 & -0.6363 & 0.4211 & -0.0897 & -0.4494 & -0.1199 & -0.3717\\ -0.0160 & 0.0585 & 0.2065 & -0.9111 & 0.0101 & 0.1093 & 0.3262 \end{pmatrix}$$

Input Variables in Feature Space: The ith component of the input vector $\boldsymbol{v}$ in the feature space is defined as follows:

$$v_i = [\hat{\boldsymbol{x}} - \text{ave}(\hat{\boldsymbol{x}})]^T \boldsymbol{u}^i \quad \text{for } 1 \le i \le 7 \qquad \text{(A3)}$$

Sample Vectors in Feature Space: The corresponding sample vectors in the $\boldsymbol{v}$-space are denoted by $\boldsymbol{v}^1, \ldots, \boldsymbol{v}^{41}$, i.e., the ith component of $\boldsymbol{v}^j$ is $[\hat{\boldsymbol{x}}^j - \text{ave}(\hat{\boldsymbol{x}})]^T \boldsymbol{u}^i$ for $1 \le i \le 7$. The sample vectors $\boldsymbol{v}^j$ ($1 \le j \le 41$) are listed as the rows in Appendix B.

Coefficients for RBF Interpolation Formula: Let $\varphi(t) = \sqrt{1+t^2}$. Then the basis functions for interpolation in the $\boldsymbol{v}$-space are $\varphi[d(\boldsymbol{v}-\boldsymbol{v}^1)], \ldots, \varphi[d(\boldsymbol{v}-\boldsymbol{v}^{41})]$, where

$$d(\boldsymbol{v}-\boldsymbol{v}^j) = \sqrt{\theta_1\left(v_1 - v_1^j\right)^2 + \theta_2\left(v_2 - v_2^j\right)^2 + \ldots + \theta_7\left(v_7 - v_7^j\right)^2}$$

The multiquadric PCR fitting of the 41 subsonic transport data has the following expression:

$$\hat{g}(\boldsymbol{v}) = \sum_{j=1}^{41} \alpha_j \varphi[d(\boldsymbol{v}-\boldsymbol{v}^j)] = \sum_{j=1}^{41} \alpha_j \sqrt{1 + \sum_{i=1}^{7} \theta_i (v_i - v_i^j)^2}$$

Here are the values of coefficients θ_i and α_j:

$$\theta_1 = 1.7613, \quad \theta_2 = 0.9916, \quad \theta_3 = 2.0715,$$
$$\theta_4 = 3.0030, \quad \theta_5 = 0.0013, \quad \theta_6 = 0.3279,$$
$$\theta_7 = 1.6528$$

and

$$\alpha_1 = -3913.7, \quad \alpha_2 = 4888.9, \quad \alpha_3 = -3349.1$$
$$\alpha_4 = -938.81, \quad \alpha_5 = -48944, \quad \alpha_6 = -4336.6$$
$$\alpha_7 = -6396.4, \quad \alpha_8 = 13653, \quad \alpha_9 = 41055$$
$$\alpha_{10} = 3309.5, \quad \alpha_{11} = -1204.0, \quad \alpha_{12} = 5276.5$$
$$\alpha_{13} = 4820.6, \quad \alpha_{14} = -6161.0, \quad \alpha_{15} = 4146.2$$
$$\alpha_{16} = -6110.6, \quad \alpha_{17} = -1889.2, \quad \alpha_{18} = -1248.6$$
$$\alpha_{19} = 4113.9, \quad \alpha_{20} = 11449, \quad \alpha_{21} = 7762.5$$
$$\alpha_{22} = 153645, \quad \alpha_{23} = -168989, \quad \alpha_{24} = -961.38$$
$$\alpha_{25} = -6400.0, \quad \alpha_{26} = 4174.1, \quad \alpha_{27} = -3945.2$$
$$\alpha_{28} = -6414.3, \quad \alpha_{29} = -9.2846, \quad \alpha_{30} = 1527.4$$
$$\alpha_{31} = 6609.7, \quad \alpha_{32} = 3838.7, \quad \alpha_{33} = -3023.2$$
$$\alpha_{34} = -9912.2, \quad \alpha_{35} = 1029.0, \quad \alpha_{36} = 10770$$
$$\alpha_{37} = 345.13, \quad \alpha_{38} = 4849.0, \quad \alpha_{39} = -2145.5$$
$$\alpha_{40} = 5201.7, \quad \alpha_{41} = 121.28$$

Wing Weight Estimation for Given Design Variables: For any given b, s, t_r/c_r, w_{to}, λ, Λ, and μ, set $x_1 = b$, $x_3 = s$, $x_4 = t_r/c_r$, $x_5 = w_{to}$, $x_6 = \lambda$, $x_7 = \Lambda$, $x_8 = \mu$, and calculate $x_2 = \gamma s/[b(1+\lambda)]$. Then use Eqs. (A2) and (A3) to get the values for $v_1, \ldots, v_7$, which can be substituted into the last formula in Eq. (A1) to get a wing weight estimate $\bar{w}$.

Appendix B: Sample Data in Feature Space

Table B1 Feature vectors of 41 subsonic transports

Index	v_1	v_2	v_3	v_4	v_5	v_6	v_7
1	0.2821	−0.3211	0.5107	1.3519	1.0303	1.1564	1.0329
2	0.0546	0.2705	−0.1641	0.6815	−0.3009	−0.1528	−0.0787
3	−0.1699	0.4908	−0.0041	−1.5064	−0.0082	0.0224	1.6621
4	0.2261	−0.0719	0.6079	0.9109	−0.3922	−1.2644	0.6031
5	0.1488	0.1504	0.0742	−0.1464	−0.7024	−0.8031	−2.2813
6	0.9398	−1.1394	−1.1243	−0.9925	−0.1566	1.7183	−4.4064
7	−0.0957	0.3665	0.0802	0.3906	−1.3973	−1.9795	−0.3881
8	0.0854	0.0440	0.0454	−0.1451	−0.8322	−0.6898	−2.2883
9	0.1423	0.0382	−0.0391	−0.2106	−0.8223	−0.7055	−2.4037
10	0.4129	0.6750	1.2043	−1.8930	0.0541	−0.6066	2.3086
11	0.1300	−0.0985	0.9569	−0.6266	0.6524	0.5956	2.3686
12	0.1199	−0.1598	0.3030	1.3808	0.4659	0.1852	1.5867
13	−0.4217	0.6821	−1.7347	0.3230	−0.4834	0.8294	−0.5581
14	−0.1330	−0.5891	−0.8563	0.5289	0.3165	1.6930	1.1062
15	0.0933	0.0187	0.8535	−0.0953	0.7159	0.6952	1.8461
16	0.0261	−0.1555	0.5816	−0.1914	0.7958	0.8317	1.7481
17	−0.0754	0.2219	1.6175	0.0749	1.3617	1.6791	1.3740
18	−0.2046	−0.2366	−0.5746	−0.9931	−0.4920	0.3568	1.8988
19	−0.1654	−0.0132	−0.5169	−1.7190	−0.3625	0.4770	2.2744
20	−0.1348	−0.0917	−0.7876	−0.5331	−0.3748	0.3370	1.8438
21	−0.1953	−0.2597	−0.6765	−0.5568	−0.4934	0.3054	1.7457
22	−0.0953	−0.2008	−0.7828	−0.5160	−0.3447	0.0974	2.0251
23	−0.0679	−0.1994	−0.7870	−0.5229	−0.3467	0.1225	1.9987
24	0.0137	−0.1377	−0.5676	0.2886	−0.1506	0.0205	1.6877
25	0.2411	0.6853	−0.4202	0.8100	−0.3256	−0.6705	0.7148
26	0.0804	0.6238	−0.5825	0.7403	−0.1587	−0.3228	0.3660
27	−0.2457	−0.5890	0.0106	0.3349	−1.0982	−1.0073	−0.6732
28	0.1676	0.5850	−1.0124	0.4890	−0.1132	0.2835	−0.2821
29	0.1059	−0.1514	0.3690	1.2962	0.8718	1.1871	0.9002
30	0.4782	0.5746	0.3637	0.9893	0.4088	−0.5538	0.5838
31	−0.9439	−1.3704	1.4083	0.4049	−1.0293	−0.5203	−1.4615
32	0.2178	−0.3327	1.5302	−0.6804	−0.8494	−1.9863	0.8840
33	−1.5118	0.8732	0.4282	−0.4440	1.3318	1.6690	−4.2690
34	−0.8315	0.1176	0.4907	0.0695	−0.5822	0.3274	−2.4225
35	−0.0243	−1.0734	0.1186	0.2360	−0.5904	0.0389	−1.2920
36	−0.0308	1.4762	−0.3011	0.4621	−0.0768	−1.2243	−0.4474
37	−0.1826	−0.7509	−1.1894	0.5530	4.2036	−4.0953	−1.5312
38	0.1434	0.2216	0.5328	0.1776	−0.0950	0.1336	−2.0311
39	0.8943	0.4436	0.1375	−0.6112	0.7163	1.9151	−4.5534
40	−0.0598	−0.5916	−0.8368	1.1935	−0.1079	0.4788	1.7832
41	0.5861	−0.0250	0.7332	−0.1973	−0.2380	−0.2741	−2.9745

References

[1] Raymer, D., *Aircraft Design: A Conceptual Approach*, 3rd ed., AIAA, Reston, VA, 1999.

[2] McCullers, L., "Aircraft Configuration Optimization Including Optimized Flight Profiles," *Proceedings of a Symposium on Recent Experiences in Multidisciplinary Analysis and Optimization*, edited by J. Sobieski, CP-2327, NASA, April 1984, pp. 395–412.

[3] Huang, X., Haftka, R., Grossman, B., and Mason, W., "Comparison of Statistical Weight Equation with Structural Optimization for Supersonic Transport Wings," AIAA Paper 94-4379, Sept. 1994.

[4] Ardema, M., Chambers, M., Patron, A., Hahn, A., Miura, H., and Moore, M., "Analytical Fuselage and Wing Weight Estimation of Transport Aircraft," NASA Technical Memorandum 110392, May 1996.

[5] Campolongo, F., Keijnen, J., and Andres, T., "Screening Methods," *Sensitivity Analysis*, edited by A. Saltelli, K. Chan, and E. M. Scott, John Wiley & Sons, New York, 2000, pp. 65–80.

[6] Tu, J., and Jones, D., "Variable Screening in Metamodel Design by Cross-Validated Moving Least Squares Method," AIAA Paper 2003-1669, April 2003.

[7] Hockings, R., "A Biometrics Invited Paper. The Analysis and Selection of Variables in Linear Regression," *Biometrics*, Vol. 32, March 1976, pp. 1–49.

[8] Rawlings, J., Pantula, S., and Dickey, D., *Applied Regression Analysis: A Research Tool*, 2nd ed., Springer, New York, 2001, Chap. 7.

[9] Rech, G., Teräsvirta, T., and Tschernig, R., "A Simple Variable Selection Technique for Nonlinear Models," *Communications in Statistics: Theory and Methods*, Vol. 30, No. 6, 2001, pp. 1227–1241.

[10] Li, W., and Padula, S., "Approximation Methods for Conceptual Design of Complex Systems," *Approximation XI*, edited by C. Chui, M. Neamtu, and L. Schumaker, Nashboro Press, Brentwood, TN, 2005, pp. 241–278.

[11] Rocha, H., Li, W., and Hahn, A., "Principal Component Regression Using Radial Basis Function Interpolation," *Wavelets and Splines: Athens 2005*, edited by G. Chen and M. J. Lai, Nashboro Press, Brentwood, TN, 2006, pp. 402–415.

[12] de Boor, C., and Ron, A., "The Least Solution for the Polynomial Interpolation Problem," *Mathematische Zeitschrift*, Vol. 210, No. 3, 1992, pp. 347–378.

[13] de Boor, C., and Ron, A., "Computational Aspects of Polynomial Interpolation in Several Variables," *Mathematics of Computation*, Vol. 58, No. 198, 1992, pp. 705–727.

[14] de Boor, C., "Polynomial Interpolation in Several Variables," *Studies in Computer Science*, edited by R. de Millo and J. R. Rice, Plenum Press, New York, 1994, pp. 87–119.

[15] Hardy, R., "Theory and Applications of the Multiquadric-Biharmonic Method (20 Years of Discovery 1968–1988)," *Computers and Mathematics with Applications (1975-) / Computers & Mathematics with Applications*, Vol. 19, Nos. 8–9, 1990, pp. 163–208.

[16] Powell, M., "Radial Basis Function Methods for Interpolation to Functions of Many Variables," *Proceedings of the 5th Hellenic–European Conference on Computer Mathematics and Its Applications*, edited by E. A. Lipitakis, LEA Press, Athens, 2002, pp. 2–24.

[17] Buhmann, M., "Radial Basis Functions: Theory and Implementations," Cambridge Univ. Press, Cambridge, England, U.K., 2003.

[18] Sacks, J., Welch, W., Mitchell, T., and Wynn, H., "Design and Analysis of Computer Experiments," *Statistical Science*, Vol. 4, No. 4, 1989, pp. 409–435.

[19] Gibbs, M., and MacKay, D., "Efficient Implementation of Gaussian Processes," Cavendish Laboratory Technical Report, Cambridge, England, U.K., May 1997, http://citeseer.ist.psu.edu/gibbs97efficient.html [Accessed 6 Oct. 2005].