

DACE (Design and Analysis of Computer Experiments) Model and its Application in the Modeling of Ozone Concentration

Tianji Shi

Department of Statistics
University of British Columbia

January 13, 2011

Statistical Models and Computer Models

Computer or Numerical Models

Computer model, used interchangeably with *numerical model*, are referred as such in the sense that the mathematics involved are mostly scientific equations, and only the input conditions are required to implement a model run that produces a *deterministic* output.

For example, air quality model that simulates the ozone O_3 formation given emission input and meteorological conditions.

Statistical Models

In contrast, a *statistical model* is built around the idea of modeling *randomness* with probability functions, and sample data are needed to fit a model to output.

DACE Model and Notations

Computer model is usually complex and impossible to run at every input value, computationally cheap statistical models are developed to emulate the computer outputs for untried inputs.

Due to the unknown/uncertainties associated with the outputs at untried inputs, we may assume the computer model outputs as realizations of stochastic processes. This assumption enables the statistical modeling.

Common notations used	
\mathbf{x}	vector of covariates
β	vector of regression coefficients
\mathbf{f}	vector of known covariate functions
k	the dimension of \mathbf{f} and β
p	the dimension of \mathbf{x}
Y and y	model output, random variable and its realization
n	number of model runs

Stochastic Model Output and Prediction

The sum of mean and covariance component:

$$Y(\mathbf{x}) = \sum_{j=1}^k \beta_j f_j(\mathbf{x}) + Z(\mathbf{x}). \quad (1)$$

- Random variable $Z(\mathbf{x})$ is assumed to have zero mean and covariance $\sigma^2 R(\mathbf{x}, \mathbf{x}')$ between “input sites” \mathbf{x} and \mathbf{x}' .
- In eq.(1), $Z(\mathbf{x})$ models the random deviation from the regression function $\beta^T \mathbf{f}(\mathbf{x})$, where $\beta = (\beta_1, \dots, \beta_k)^T$ and $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \dots, f_k(\mathbf{x}))^T$.

At untried “input site” \mathbf{x} , one prediction method is called *universal kriging*: $\hat{y}(\mathbf{x}) = \mathbf{c}^T(\mathbf{x})\mathbf{y}_S$.

Where n “input sites” $S = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ we have corresponding output $\mathbf{y}_S = (y(\mathbf{x}_1), \dots, y(\mathbf{x}_n))^T$. These are the data.

In essence, this is a weighted average of *known* outputs.

Stochastic Model Output and Prediction Cont.

The best linear unbiased predictor (BLUP) is the $\mathbf{c}(\mathbf{x})$ that minimizes

$$\text{MSE} [\hat{y}(\mathbf{x})] = \text{E} [\mathbf{c}^T(\mathbf{x})\mathbf{Y}_S - Y(\mathbf{x})]^2 \quad (2)$$

subject to the constraint $\text{E} [\mathbf{c}^T(\mathbf{x})\mathbf{Y}_S] = \text{E} [Y(\mathbf{x})]$, i.e., $\mathbf{c}^T(\mathbf{x})\mathbf{F}\boldsymbol{\beta} = \boldsymbol{\beta}^T \mathbf{f}(\mathbf{x})$ for all $\boldsymbol{\beta}$. \mathbf{F} is the $n \times k$ design matrix

$$\mathbf{F} = \{\mathbf{f}(\mathbf{x}_1), \dots, \mathbf{f}(\mathbf{x}_n)\}^T.$$

Furthermore, introducing correlation matrix on input sites $\mathbf{R} = \{R(\mathbf{x}_i, \mathbf{x}_j)\}$, $1 \leq \{i, j\} \leq n$, and correlations between Z at an untried \mathbf{x} and the Z 's at S represented by the vector $\mathbf{r}(\mathbf{x}) = [R(\mathbf{x}_1, \mathbf{x}), \dots, R(\mathbf{x}_n, \mathbf{x})]^T$. Equation (2) can be expanded into

$$\text{MSE} (\hat{y}(\mathbf{x})) = \sigma^2 [1 + \mathbf{c}^T(\mathbf{x})\mathbf{R}\mathbf{c}(\mathbf{x}) - 2\mathbf{c}^T(\mathbf{x})\mathbf{r}(\mathbf{x})] \quad (3)$$

when accounting for the constraint $\text{E} [\mathbf{c}^T(\mathbf{x})\mathbf{Y}_S] = \text{E} [Y(\mathbf{x})]$.

Stochastic Model Output and Prediction Cont.

Since we are minimizing with a constraint, we would add Lagrangian term $\lambda^T(\mathbf{x})[\mathbf{F}^T \mathbf{c}(\mathbf{x}) - \mathbf{f}(\mathbf{x})]$ to equation (3) and obtain the BLUP criteria

$$\begin{pmatrix} 0 & \mathbf{F}^T \\ \mathbf{F} & \mathbf{R} \end{pmatrix} \begin{pmatrix} \lambda(\mathbf{x}) \\ \mathbf{c}(\mathbf{x}) \end{pmatrix} = \begin{pmatrix} \mathbf{f}(\mathbf{x}) \\ \mathbf{r}(\mathbf{x}) \end{pmatrix}.$$

Solving for $\mathbf{c}(\mathbf{x})$ results in the predictor

$$\hat{y}(\mathbf{x}) = \mathbf{f}^T(\mathbf{x})\hat{\beta} + \mathbf{r}^T(\mathbf{x})\mathbf{R}^{-1}(\mathbf{y}_S - \mathbf{F}\hat{\beta}), \quad (4)$$

with $\hat{\beta}$ being the generalized least-square estimate. One might notice that this expression is akin to the conditional mean of $y(\mathbf{x})$ given \mathbf{y}_S under a multivariate Gaussian distribution.

- This is one path that arrives to predictive eq.(4), geo-statisticians uses similar predictive equations.
- This is an interpolation method, where a strong emphasis is placed on the correlation structure.

Correlation Function

There are several means of estimating parameters, typical method is to maximize the profile-likelihood:

1. Assume *Gaussian Process* on Y_S and derive likelihood function.
2. Substitute in *MLE* expressions for mean and variance to find the profile-likelihood with only correlation parameters in $R(\cdot, \cdot)$ as unknown.

We are left to decide the form of correlation function $R(\mathbf{x}, \mathbf{x}')$. One specification is called *power-exponential correlation function*, which has expression

$$R(\mathbf{x}, \mathbf{x}') = \exp\left(-\sum_{j=1}^p \theta_j |x_j - x'_j|^{\nu_j}\right), \quad \theta_j > 0 \text{ and } 1 \leq \nu_j \leq 2, \quad (5)$$

where θ_j measures the correlation strength or “activity” of x_j and ν_j is the smoothness parameter.

Some Popular References

"Design and Analysis of Computer Experiments (with Discussion)" (1989), Sacks, J and Welch, W. J. and Mitchell, T. J. and Wynn, H.P., *Statistical Science*, 4, 409-423.

Model optimization techniques

"Screening, Predicting, and Computer Experiments" (1992), Welch, W. J. and Buck, R. J. and Sacks, J. and Wynn, H. P. and Mitchell, T. J. and Morris, M. D., *Technometrics*, 34, 15-25

"Efficient Global Optimization of Expensive Black-Box Functions" (1998), Jones, D. R. and Schonlau, M. and Welch, W. J., *Journal of Global Optimization*, 13, 455-492.

Fundamentals in Kriging

"The Origin of Kriging" (1990), Cressie, N., *Mathematical Geology*, 22, 239-251.

More Popular Reference

Melding physical observation and model output

"Bayesian Calibration of Computer Models (with Discussion)" (2001), Kennedy, M. C. and O'Hagan, A., *Journal of the Royal Statistical Society, Series B*, 63, 425-464.

Improving with model output downscaling

"Model Evaluation and Spatial Interpolation by Bayesian Combination of Observations with Outputs from Numerical Models" (2005), Fuentes, M. and Raftery, A. E., *Biometrics*, 61, 36-45. [Non-stationary correlation](#)

"A Spatio-Temporal Downscaler for Output From Numerical Models" (2009), Berrocal, V. J. and Gelfand, A. E. and Holland, D. M., *Journal of Agricultural, Biological, and Environmental Statistics*, 15, 176-197 [Temporal effect](#).

Direct Application into Ozone Modeling I

Using the methodology of Sacks et al (1989), Gao et al (1996) showed that the Gaussian process model can be applied to model observational ozone concentrations. They modelled daily data from Chicago for the period 1981 to 1991.

Daily maximum ozone concentration y is treated as a realization of a stochastic process

$$Y(\text{yr}, \mathbf{met}, d) = \beta_{\text{yr}} + Z(\mathbf{met}, d) + \varepsilon_{\text{yr},d} \quad (6)$$

where $d = 1, \dots, n$ is the day, $\text{yr} = 1981, \dots, 1991$ is the year and \mathbf{met} is a vector of meteorological conditions for day d .

- Instead of representing numerical model output, $Y()$ here is the observed daily ozone maximum. In addition, the modeling is based on correlation across time rather than location.
- The parameter β_{yr} is the additive annual effect, $Z(\mathbf{met}, d)$ is a zero-mean GP indexed by day and its covariates \mathbf{met} , and $\varepsilon_{\text{yr},d}$ follows $iid N(0, \sigma_\varepsilon^2)$.

Direct Application into Ozone Modeling II

Predictive equation adjusted from eq.(4) has expression

$$\hat{y}(yr, \mathbf{x}) = \hat{\beta}_{yr} + \mathbf{v}^T(\mathbf{x})\mathbf{C}^{-1}(\mathbf{y} - \mathbf{F}\hat{\beta}). \quad (7)$$

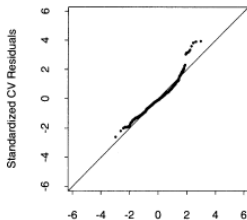
$\mathbf{x} = \{\text{met}, d\}$, $\hat{\beta} = (\hat{\beta}_{1981}, \dots, \hat{\beta}_{1991})^T$ is GLM estimate,
 $\mathbf{C} = (\sigma_Z^2/\sigma^2)\mathbf{R} + (\sigma_\epsilon^2/\sigma^2)\mathbf{I}$, $\sigma^2 = \sigma_Z^2 + \sigma_\epsilon^2$ and $\mathbf{v}(\mathbf{x}) = (\sigma_Z^2/\sigma^2)\mathbf{r}(\mathbf{x})$. Design matrix is

$$\mathbf{F} = \begin{pmatrix} \vec{\mathbf{1}}_{n1981} & \vec{\mathbf{0}} & \dots & \vec{\mathbf{0}} \\ \vec{\mathbf{0}} & \vec{\mathbf{1}}_{n1982} & \dots & \vec{\mathbf{0}} \\ \vdots & \vdots & \ddots & \vdots \\ \vec{\mathbf{0}} & \vec{\mathbf{0}} & \dots & \vec{\mathbf{1}}_{n1991} \end{pmatrix}.$$

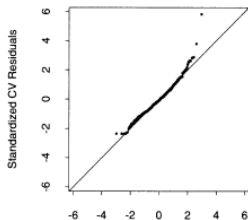
It can be shown that when \mathbf{x} is a column of data matrix, the prediction is the observation $y(\mathbf{x})$ itself with zero prediction variance. One way of assessing the model performance is through cross-validation.

Direct Application into Ozone Modeling III

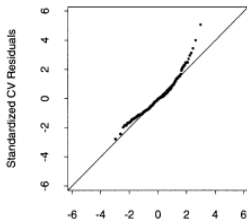
From Gao et al (1996):



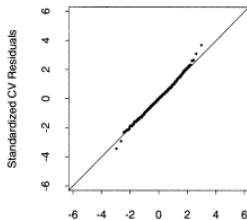
Standard Normal Quantiles
May 15 -- June 15



Standard Normal Quantiles
June 15 -- July 15



Standard Normal Quantiles
July 15 -- August 15



Standard Normal Quantiles
August 15 -- September 15

Conclusion of Ozone Modeling

Bloomfield et al (1996) implemented traditional statistical approach of building regression models. Compared to the “automatic” procedure shown, the much more complex and time-consuming methods were worse off.

References mentioned:

”Predicting Urban Ozone Levels and Trends with Semiparametric Modeling” (1996), Gao, F. and Sacks, J. and Welch, W. J., *Journal of Agricultural, Biological and Environmental Statistics*, 1, 404-425.

”Accounting for Meteorological Effects in Measuring Urban Ozone Levels and Trends” (1996), Bloomfield, P. and Royle, A. J. and Steinberg, L. J. and Yang, Q., *Atmospheric Environment*, 30, 3067-3077.