*Proceedings of the 1996 Winter Simulation Conference*
*ed. J. M. Charnes, D. J. Morrice, D. T. Brunner, and J. J. Swain*

# PARAMETER IDENTIFICATION METHODS FOR METAMODELING SIMULATIONS

Don Caughlin

Mission Research Corporation
Colorado Springs, Colorado, 80903
donc@mrccos.com

## ABSTRACT

A metamodel is a mathematical approximation of the system relationships defined by a high fidelity model or simulation. This paper presents methods that support new procedures that expanded the set of available metamodel representations beyond the traditional least squares formulation and added the capability to use dynamical metamodels. These methods compliment a new taxonomy of metamodel structures and procedures that separated the metamodeling process into a set of sequential decisions based on *a priori* information. This work was supported in part by The USAF Rome Laboratory Contract F30602-94-C-0110.

## 1 INTRODUCTION

In Caughlin (1994a) we introduced a framework for the application of System Identification techniques to develop suitable metamodels for tactical combat simulations used by the Department of Defense. We filled in the framework with concrete definitions and identified specific issues associated with the representation of dynamical systems.

Caughlin (1996) presented procedures based on this framework that allowed the separation of the metamodeling process into a set of sequential decisions based on *a priori* information. This paper presents specific methods that are consistent with the new framework and procedures to support the generation of metamodels.

The paper is organized as follows: Section 2 introduces metamodels; Section 3 covers the parameter identification methods; and Section 4 summarizes the paper.

## 2 METAMODELS

A model is a structure that can be used for understanding the behavior of a system Vemuri (1978). The model can be a physical structure such as a wind tunnel model used to determine the aerodynamics of an aircraft, or it could be a conceptual model represented by interactions, a system of equations, or a simulation.

A metamodel is a mathematical approximation of the system relationships defined by another, more detailed model (in our case – a tactical simulation).

There are two general metamodeling techniques: these are the "Direct" and "Inverse" methods. Direct metamodels are developed by applying basic principles to generate a more abstract (approximate) version of the original model. Inverse modeling begins with the input-output data generated by the high fidelity model or simulation and develops the metamodel from the data.

We presented new procedures in Caughlin (1996) that tailor the "Inverse Problem" to generate metamodels of simulations. We separated a complex procedure into two general steps. The first part of the process defined the problem. The second step was the metamodeling process. This step determined the metamodel set and generated the metamodel. We now discuss techniques for identifying the parameters of the selected representation.

## 3 PARAMETER ID METHODS

There are many taxonomies used to categorize identification methods. Methods can be referred to as off-line or on-line. Also, they can be classified as either open-loop or closed-loop methods. Further classification can be made as nonparametric, frequency domain, and as parameter identification methods.

Parameter identification methods are used when the candidate model is defined by a set of parameters.

The combination of model, error criterion, and

numerical method has lead to an overwhelming myriad of "identification methods." Most of the above techniques, however, can be classified by two elements of the identification method: the form of the identifier and the criterion of fit.

The form of the identifier defines the "experimental setup" or the manner in which the estimates are generated and compared. The criterion of fit establishes both the cost function and the method of its minimization. A summary discussion of these elements is included in Caughlin (1996). Details are found in Caughlin (1995).

If we categorize the identification method by the form of the identifier and the criterion of fit, we can reduce the many identification methods to four approaches: Prediction Error and Correlation, Maximum Likelihood, Optimization, and Approximation Techniques. We will now present some of the techniques that result.

## 3.1 Prediction Error and Correlation Approaches

Let the prediction error be given by $\varepsilon(t,\theta) = y(t) - \hat{y}(t|\theta)$ with $y(t)$ the output of the simulation and $\hat{y}(t|\theta)$ the output of the metamodel ($\theta$ is the parameter vector). A "good" model will have small prediction errors. There are two general approaches to define a measure of $\varepsilon$. The first is to define a norm that measures the size of $\varepsilon$ and minimize that norm. This leads to the prediction error method (PEM). Another measure of $\varepsilon$ is to require that $\varepsilon$ be uncorrelated with past data. This is the correlation approach which contains the instrumental-variable (IV) method which we discuss in Section 3.1.4.

### 3.1.1 General Description of The Prediction Error Method

Filter the prediction sequence $\varepsilon(t,\theta)$ using a stable linear filter $L(q)$:

$$\varepsilon_F(t,\theta) = L(q)\varepsilon(t,\theta)$$

where $q^{-1}$ is the backward shift operator defined as $q^{-1}u(t) = u(t-1)$.

This filtering acts like frequency weighting and can remove or enhance selected properties of the model. Then, using either a fixed or weighted (possibly time varying) norm:

$$V_N(\theta, D) = \frac{1}{N}\sum_{t=1}^{N} l\left(\varepsilon_F(t,\theta),\theta,t\right)$$

define the estimate $\hat{\theta}_N$ by the minimization:

$$\hat{\theta}_N = \hat{\theta}_N(D) = arg\min_{\theta\in D}\{V(\theta, D)\} \qquad (1)$$

where $D$ is the set allowed by the model.

In general, PEM is a technique that approximates (smoothes) the empirical transfer function estimate to the model transfer function with a weighted norm corresponding to the model signal-to-noise at the frequency in question.

### 3.1.2 Specific PEM Methods

While "equation 1" can be solved numerically in the general case, specific methods are obtained as special cases with special selections of the filter $L(q)$ and the scalar valued norm function $l(\cdot)$.

**Least Squares**. If the predictor is linear, the prediction error becomes $\varepsilon(t,\theta) = y(t) - \phi^T(t)\theta$ where $\phi^T(t)$ is the vector of regressors that depends on the selected model structure. Also if $L(q) = 1$ and $l(\varepsilon) = \frac{1}{2}\varepsilon$, then the norm becomes:

$$V_N(\theta, D) = \frac{1}{N}\sum_{t=1}^{N}\frac{1}{2}\left[y(t) - \phi^T(t)\theta\right]^2$$

This is the least squares criterion for linear regression. The performance measure $J = \varepsilon^T\varepsilon$, was based on the view that all errors are equally important. Weighted least squares weights the errors and is based on the criterion $J = \varepsilon^T W\varepsilon$. Other versions of the least squares criterion are the Best Linear Unbiased Estimator where the weight is equal to the inverse of the measurement noise.

If the variance of the parameters is known (or assumed), then we can further improve on the Best Linear Unbiased (Gauss-Markov) Estimator. This improvement is called the minimum variance estimator and includes the variance of the parameters in the normal equations.

**Ridge Regression**. The aim of another modification of ordinary least squares - ridge regression - is the reduction of the mean square error (Press, 1986). This is accomplished by the addition of a symmetric matrix K to the regressor to improve the numerical conditioning of the estimator.

**Chi-Square**. In Chi-Square fitting, we assume that each data point $y_i$ has a measurement error that is independently random and distributed as a normal distribution around the true model. Suppose that the standard deviation is the same for all points; then the probability of the data set is the product of probabilities of each point:

$$P = \Pi\{\exp\left[-\frac{1}{2}\left(\frac{y_i - y(x_i)}{\sigma}\right)^2\right]\Delta y\}$$

Maximizing this is equivalent to maximizing its logarithm, or minimizing the negative of its logarithm:

$$\left[\sum \frac{[y_i - y(x_i)]^2}{2\sigma^2}\right] - N \log \Delta y$$

Since N,$\sigma$, and $\Delta y$ are all constants, minimizing this equation is equivalent to minimizing:

$$\sum_{i=1}^{N} [y_i - y(x_i; \theta_1 \ldots \theta_M)]^2$$

If each data point has its own standard deviation, however, the probability of the data set is modified by considering $\sigma_i$ in place of $\sigma$ (see Press, 1986 for details).

**Eigenstructure Realization Algorithm.** The Eigenstructure Realization Algorithm (ERA) is included under the PEM methods because this algorithm uses the least squares approach to directly identify the Markov parameters of a steady state Kalman filter.

Consider a discrete, time-invariant multivariable linear system:

$$\begin{aligned} x_{t_{i+1}} &= A(\theta)x(t_i) + B(\theta)u(t_i) + M(\theta)w_d(t_i) \\ y(t_i) &= C(\theta)x(t_i) + D(\theta)u(t_i) + \nu(t_i) \end{aligned} \quad (2)$$

An observer for the above system can be developed that will be as stable as desired and the resulting Markov parameters will be the Markov parameters of the observer (Juang, 1993). The system Markov parameters can be extracted from the observer parameters. The major assumption is that of ergoticity.

Choose p such that $mp > n$ (where n is the number of states and m is the number of outputs) and, beginning at the p+1 measurement, let:

$$y = [y(p+1)\ y(p+2)\ y(p+3) \ldots y(k-1)]$$

From the definition of the Kalman Filter we have:

$$\overline{Y} = \left[ D\ C\overline{B}\ C\overline{AB} \ldots C\overline{A}^{k-1}\overline{B} \right]$$

with

$$\begin{aligned} \overline{A} &= A + MC \\ \overline{B} &= [B + MD, -M] \end{aligned}$$

and

$$U = \begin{vmatrix} u(p+1) & u(p+2) & \cdots & u(k-1) \\ \begin{bmatrix} u(p) \\ y(p) \end{bmatrix} & \vdots & \cdots & \begin{bmatrix} u(k-2) \\ y(k-2) \end{bmatrix} \\ \vdots & \begin{bmatrix} u(p) \\ y(p) \end{bmatrix} & \cdots & \begin{bmatrix} u(k-3) \\ y(k-3) \end{bmatrix} \\ \vdots & \ddots & \vdots & \vdots \\ \begin{bmatrix} u(0) \\ y(0) \end{bmatrix} & \begin{bmatrix} u(1) \\ y(1) \end{bmatrix} & \cdots & \begin{bmatrix} u(k-p-2) \\ y(k-p-2) \end{bmatrix} \end{vmatrix}$$

When $C\overline{A}^k\overline{B} \approx 0$ for $k > p$, the system $y = \overline{Y}U$ can be solved for $\overline{Y}$ using a weighted least squares. Once the observer Markov parameters are determined, the system parameters must be extracted. After extracting the system Markov parameters from the observer, we can recover the state space model by the ERA. Define the following $r_1 \times s$ block data matrix:

$$H(\tau) = \begin{bmatrix} Y_\tau & Y_{\tau+1} & \cdots & Y_{\tau+s-1} \\ Y_{\tau+1} & Y_{\tau+2} & \cdots & Y_{\tau+s} \\ Y_{\tau+2} & Y_{\tau+3} & \cdots & Y_{\tau+s+1} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{\tau+r_1-1} & Y_{\tau+r_1} & \cdots & Y_{\tau+r_1+s-2} \end{bmatrix}$$

The order of the system is determined by the singular value decomposition of $H(0)$:

$$H(0) = U\Sigma V^T = U_1 S_1 V_1^T$$

where $\Sigma$ are all of the singular values. $S_1$ is an $n \times n$ diagonal matrix of positive singular values that are retained and n will become the order of the system:

$$\begin{aligned} A &= S_1^{-1/2} U_1^T H(1) V_1 S_1^{-1/2} \\ B &= S_1^{-1/2} V_1 E_m \\ C &= E_1^T U_1 S_1^{-1/2} \end{aligned}$$

where $E_r^T = [I_{r\times r}\ 0_{r\times(r_1-m)m}]$ and $E_m^T = [I_{m\times m}\ 0_{m\times(r_1-m)m}]$. The observer gain can be extracted in a similar fashion.

### 3.1.3 Correlation Approaches

Ideally the prediction error $\varepsilon(N, \hat{\theta})$ for a "good" model should be independent of past data $Z^{N-1}$. If $\varepsilon(N, \hat{\theta})$ is correlated with past data, then there is more information available in the data. A true test of the correlation of $\varepsilon(N, \hat{\theta})$ and $Z^{N-1}$ requires testing every nonlinear transformation of $\varepsilon(N, \hat{\theta})$ with all possible functions of $Z^{N-1}$. This is not feasible.

We can, however, select a finite dimensional vector sequence $\{\zeta(t)\}$ derived from $Z^{N-1}$ and force a certain transformation of $\varepsilon(N, \hat{\theta})$ to be uncorrelated with this sequence (Ljung, 1987). In general we can accomplish this by filtering the prediction errors:

$$\varepsilon_F(N, \hat{\theta}) = L(q)\varepsilon(N, \hat{\theta})$$

choosing a sequence of correlation vectors:

$$\{\zeta(t, \hat{\theta})\} = \{\zeta(t)(t, Z^{N-1}, \hat{\theta})\}$$

and a function: $\alpha\left(\varepsilon_F(N, \hat{\theta})\right)$ for computing:

$$f_N(\hat{\theta}, Z^{N-1}) = \frac{1}{N}\sum_{t=1}^{N} \zeta(t, \hat{\theta})\alpha\left(\varepsilon_F(N, \hat{\theta})\right)$$

and then finding $\hat{\theta}_N$ such that $f_N(\hat{\theta}, Z^{N-1}) = 0$ .

### 3.1.4 Instrumental-Variable (IV) Method

If we define $\varepsilon(N, \hat{\theta})$ above to be $\varepsilon(N, \hat{\theta}) = \left[y(t) - \phi^T(t)\hat{\theta}\right]$ we can expand the sequence of the correlation vectors to include model dependent parameters by:

$$\{\zeta(t, \hat{\theta})\} = K_u(q, \hat{\theta})u(t)$$

where $K_u(q, \hat{\theta})$ is a $d \times m$ matrix filter and $L(q)$ is of dimension $p \times p$. With $dim\zeta(t) = dim\hat{\theta} = d \times p$, we have the instrumental-variable (IV) method:

$$\hat{\theta}_{IV} = \left[\zeta(t, \hat{\theta})^T X\right]^{-1} \zeta(t, \hat{\theta})^T y$$

If we allow $dim\zeta(t) > d$ and a minimum norm solution for $f_N(\hat{\theta}, Z^N)$, we have the extended IV method.

## 3.2 Maximum Likelihood Approaches

If we consider independent, identically distributed measurements, and if an efficient estimate (unbiased estimate with finite covariance such that no other unbiased estimate has a lower covariance) exists, it can always be found through maximum likelihood approaches. Although the maximum likelihood estimate will be biased for small samples, it will provide the unique minimum variance estimate attaining the Cramr-Rao lower bound if this is possible (Maybeck, 1982).

The objective is to provide a parameter estimator that does not require complete *a priori* parameter statistics yet still allows the inclusion of *a priori* knowledge. Unlike the best linear unbiased estimate provided by appropriately weighted least squares, this method propagates the probabilistic information in time and directly allows the inclusion of known statistical information.

The key to the identification algorithm will be the residuals of the state estimator, and the most significant drawback of the maximum likelihood approaches is the lack of theoretical knowledge on the behavior of the estimates for small sample sizes.

The following discussions are limited to linear-time invariant (discrete time) systems. Nonlinear effects can be included by appropriately modifying the prediction equations in either of two ways. First, nonlinear system effects can be directly included in the propagation of the state. Second, nonlinear measurements (with linear propagation) can be handled with an extended Kalman filter model.

Beginning with a linear time-invariant discrete state space model (see equation 2), there are a number of conditional probability destiny functions that could be used for the likelihood function. Variations include fixed length versus growing length functions, specification of *a priori* statistics, use of the initial conditions, and the sensitivity of the estimate on the identified parameters. The most appropriate density function is:

$$\begin{aligned} f_{x(t_i),Z(t_i)|\theta} &= f_{x(t_i)|Z(t_i),\theta} f_{Z(t_i)|\theta} \\ &= f_{x(t_i)|Z(t_i),\theta} \Pi_{j=1}^{i} f_{x(t_i)|Z(t_i),\theta} \end{aligned}$$

Minimization of the likelihood function with this density results in the state predicted by the Kalman-Bucy filter, but there is no closed form solution to compute the partial derivatives.

### 3.2.1 Full Scale Estimator

A full scale estimator can be derived that minimizes the likelihood function in an iterative process. This estimator uses the last N observations to identify v uncertain parameters in the system and input matrices A and B. (Note: Uncertainty in these parameters could not be separated from uncertainties in C and D. Consequently, the assumption is that C and D are known and the uncertainty is A and B.)

The iterative estimator for minimization of the likelihood equation:

$$\left.\frac{\partial L\left[\hat{\theta}, Z^N\right]}{\partial \hat{\theta}}\right|_{\hat{\theta}(t_i)=\theta_*(t_i)}$$

using the method of "steepest descent" is:

$$\hat{\theta}(t_i) = \hat{\theta}(t_i) - \left[\frac{\partial^2 L\left[\hat{\theta}, Z^N\right]}{\partial \hat{\theta}^2}\right]^{-1} \left[\frac{\partial L\left[\hat{\theta}, Z^N\right]}{\partial \hat{\theta}}\right]$$

To use this algorithm, the Hessian (second derivative matrix) must be of full rank. Using a technique called "scoring," we can approximate the Hessian with the conditional information matrix. However, considering the propagation of the values in time, incorporation of measurements, and the summation over the last N residuals, the implementation of the above equations is quite complex. Even with the approximations, the full scale estimator requires a large number of calculations (see Maybeck, 1982).

### 3.2.2 Modified Maximum Likelihood (MMLE)

In the modified maximum likelihood formulation, A, B, C, D, and M are estimated and used with the error covariance, P, to determine the Kalman gain, K, from an approximation based on the Riccatti equation. To provide a parameter estimator, we consider the measurement equation. Since we have assumed a Gaussian error model, the Conditional Probability Density Function (CPDF) for the measurement becomes:

$$P(z_i|z_{i-1},\theta) = \frac{1}{[(2\pi)^m detP]^{1/2}} \exp\left\{-\frac{1}{2}\tilde{z}_i^T (P)^{-1}\tilde{z}_i\right\}$$

where $P = E\{\tilde{z}\tilde{z}^T\}$ with dimension $m \times m$ and $\tilde{z} = z_i - \hat{z}$ is the innovations process (residuals) computed by the Kalman filter (where all of the matrices could be functions of $\theta$).

Assuming a constant innovations covariance, use of a steady state filter results in a constant filter gain. This allows the CPDF to be written as:

$$P(z|\theta) = \Pi_{i=1}^N \frac{1}{[(2\pi)^m detP]^{1/2}} \exp\left\{-\frac{1}{2}\tilde{z}_i^T (P)^{-1}\tilde{z}_i\right\}$$

There are two approaches to the solution depending on whether *a priori* information is used.

**Maximum Likelihood (ML) Estimation.** Given the above CPDF, the ML LLF becomes:

$$\begin{aligned} LLF(\hat{\theta}) &= \frac{1}{2}\sum_{i=1}^N \{\tilde{z}_i^T (P)^{-1}\tilde{z}_i\} \\ &\quad + \frac{N}{2}\log\det(P) + \frac{Nm}{2}\log 2\pi \end{aligned}$$

A necessary condition at the minimum is that $P = E\{\tilde{z}\tilde{z}^T\}$ must equal the sample innovations covariance (Goodwin, 1977). Therefore, since P has dimension $m \times m$, the first term in the LLF becomes $Nm/2$, and the minimization is reduced to a minimization of the determinant of the sample innovations covariance matrix.

When P is known, the LLF can be minimized by minimizing the following cost function:

$$J(\hat{\theta}) = \frac{1}{2}\sum_{i=1}^N \{\tilde{z}_i^T (P)^{-1}\tilde{z}_i\}$$

This minimization is usually carried out using a Gauss-Newton method using the first and second gradients of the cost function.

**Maximum A Posteriori (MAP) Estimation.** In the MAP estimator, we continue to require that $\hat{P} = \frac{1}{N}\sum_{i=1}^N \tilde{z}\tilde{z}^T$ but add the term $-\log P(\hat{\theta})$. Assuming that $\theta$ is normally distributed with a covariance $\Sigma$:

$$\begin{aligned} -\log P(\hat{\theta}) &= \frac{1}{2}\left(\hat{\theta}-\theta\right)^T \Sigma^{-1}\left(\hat{\theta}-\theta\right) + \\ &\quad \frac{1}{2}\log\left((2\pi)^m \det\Sigma\right) \end{aligned}$$

the $LLF_{MAP}$ becomes:

$$\begin{aligned} LLF_{MAP}(\hat{\theta}) &= \frac{1}{2}\sum_{i=1}^N \left\{\tilde{z}_i^T (\hat{P})^{-1}\tilde{z}_i\right\} \\ &\quad + \frac{1}{2}(\hat{\theta}-\theta)\Sigma^{-1}(\hat{\theta}-\theta) \end{aligned}$$

which adds a quadratic term that biases the estimates toward *a priori* values.

## 3.3 Optimization

Often we are unable to formulate the problem such that a suitable prediction equation is available. Therefore we must resort to either a "nonlinear state space model" or a "simulation model." In these situations, where we are unable or unwilling to consider a linearized or perturbation approach, the best we can do is take the output of the model, incorporate it into a "cost" function, and adjust the model parameters to optimize (minimize) that function.

There are several "standard" numerical procedures that are used to search for the minimum of a function. These are the iterative optimization methods: successive approximation, Newton's method, the Gauss-Newton algorithm, to name a few.

In addition, there are several programs that are specifically designed to perform parameter estimation.

**pEst.** A minimum mean square error parameter estimator, pEst is an interactive program for the parameter estimation of nonlinear dynamic systems (Maine and Iliff, 1981). This program solves a vector set of time-varying, finite-dimensional, ordinary differential equations that are separated into a continuous-time state equation and a discrete-time measurement equation:

$$\begin{aligned} \dot{x} &= f[x(t), u(t), \theta] \\ z(t_i) &= g[x(t_i), u(t_i), \theta] \end{aligned}$$

pEst uses three separate minimization algorithms (steepest descent, modified Newton-Raphson, and

Davidod-Fletcher-Power) to minimize the following cost function:

$$J(\hat{\theta}) = \frac{1}{2n_N n_z} \sum_{i=1}^{n_t} [z(t_i) - \hat{z}(t_i)]^T W [z(t_i) - \hat{z}(t_i)]$$

where $n_N$ equals the number of data points, and $n_z$ is the number of response variables.

**Simulated Annealing.** Using statistical mechanical theories, an optimization technique called "simulated annealing" provides a new option to directly process nonlinear, discontinuous, stochastic functions (Ingber, 1993). Given data and a cost function, it will globally optimize that function by emulating the physical annealing process to arrive at a global minimum. See Ingber (1990) and Caughlin (1994b) for a description on how to use Adaptive Simulated Annealing.

## 3.4 Approximation Techniques for Identification

### 3.4.1 Stochastic Approximation

Stochastic approximation may be regarded as the application of gradient methods to stochastic problems. It is a scheme for successive approximation of a sought quantity when the observations involve random errors due to the stochastic nature of the problem. The main advantage is the simplicity of the implementation and the fact that prior knowledge of the noise statistics are not necessary.

Stochastic approximation can be applied to any problem which can be formulated as a regression in which repeated observations are made. This approach is an exact analog of the deterministic gradient procedure.

### 3.4.2 Spline Approximation

Polynomials are excellent approximating functions when a smooth function is to be approximated locally. Any such smooth piecewise polynomial function is called a spline, and they are commonly used for fitting data.

The typical use for the spline approximation is to construct a piecewise polynomial to fit data. An exact fit involves interpolation; an approximate fit uses least squares (minimum mean square error) approximation. To explain the structure and advantages of the spline, consider a truncated Taylor series (expanded about $x_0$ where $D^i$ is the $i^{th}$ derivative):

$$\sum_{i=0}^{n} \frac{(x - x_0)^i}{i!} D^i f(x_0)$$

This polynomial should provide a satisfactory approximation for $f(x)$ if the function is sufficiently smooth and $x$ is sufficiently close to $x_0$. But, if the function must be approximated over a larger interval, the degree of the polynomial may have to be unacceptably large.

The alternative to a higher order polynomial is to subdivide the interval into sufficiently small intervals such that, on each interval, a polynomial with a relatively low degree can provide an adequate approximation.

The construction of a series of splines over an interval is a stable and straightforward mathematical procedure. At the breakpoints, derivatives are continuous. At the end points, two conditions are possible. In the "natural" cubic spline, the second derivative is zero. In the "not-a-knot" end condition, the jump in the third derivative is zero.

Once developed, the spline can be evaluated, integrated, differentiated, augmented, or cut.

### 3.4.3 Canonical Variate Analysis

Another approximation technique is canonical variate analysis. The canonical variate method is a prediction error approximation technique that optimally predicts future responses based on a reduced order state space system (Larimore, 1989).

In the statistical literature, the canonical variate problem is one of maximizing the correlation between two sets of variables. Here we will use the technique to chose nonlinear combinations of past data to predict the future data by considering the fact that the conditional expectation is an optimal projection in Hilbert space. We optimally select k linear combinations of the past data for prediction of the future.

Observations coming from the behavior we desire to model are separated into the past p(t) of a vector process and the future f(t) of another vector process. They are assumed to be jointly stationary:

$$\begin{aligned} p^T &= (y^T(t), y^T(t-1), \ldots, u^T(t), \ldots)^T \\ f^T &= (y^T(t+1), y^T(t+2), \ldots, y^T(t-l))^T \end{aligned}$$

where the vector process $p(t)$ can include both inputs and outputs.

The optimal $k^{th}$ order linear predictor $\hat{f}(t)$ of the past is measured by the prediction error:

$$E\left\{ \| f - \hat{f} \|^2_{\Lambda^{-1}} \right\} \equiv \left\{ (f - \hat{f})^T \Lambda^{-1} (f - \hat{f}) \right\}$$

where $\Lambda$ is arbitrary positive semidefinite, so that $\Lambda^{-1}$ is a quadratic weighting matrix that is possibly singular. The CVA problem is to determine

$c(t) = J_k p(t)$ and $d(t) = L_k f(t)$ (a function of reduced order memory) such that the prediction error is minimized.

The connection between CVA and metamodeling is not direct and much of the literature is very confusing or misleading. First, recall that the metamodel is a reduced order model that is the result of an optimal projection of the higher order model onto a subspace of reduced dimensions. It can be shown that projection operators on a Hilbert Space of nonlinear functions can be expressed as a conditional expectation (Larimore, 1989). It can also be shown that eigenvectors of this conditional expectation have a common eigenvalue which is equal to the squared maximal correlation. If a process has a rational power spectrum (i.e., is a finite order Markov process) then there are a finite number of nonzero canonical correlations between the past and future outputs (Larimore and Baillieul, 1990).

The solution to the canonical variate problem is expressed by putting the covariance structure of the past and future data in a canonical form such that in this new basis the norm of the weighted prediction error is the sum of squares. This is equivalent to finding J and L such that:

$$\begin{aligned} J\Sigma_{pp}J^T &= I_m \\ L\Lambda L^T &= I_n \end{aligned}$$

$$J\Sigma_{pf}L^T = Diag\{\gamma_1 \geq \gamma_2 \geq, \ldots, \geq \gamma_q \geq 0, \ldots, 0\}$$

where $\Sigma_{pp}, \Sigma_{ff}, and \Sigma_{pf}$ are the covariance matrices of past, future, and cross covariance of the past and future data defined by:

$$\Sigma = \begin{pmatrix} \Sigma_{pp}\Sigma_{pf} \\ \\ \Sigma_{fp}\Sigma_{ff} \end{pmatrix}$$

with $Diag\{\gamma_1 \geq \gamma_2 \geq, \ldots, \geq \gamma_q \geq 0, \ldots, 0\}$ a diagonal matrix with the singular values on the diagonal. Since the past and future basis in the new basis are orthonornal and uncorrelated, the singular values are also the correlations between the canonical variates p and f.

In a linear system, independent variables are orthogonal. For nonlinear systems, stochastic independence is required. The maximal correlation is defined by:

$$\rho(p,f) = \sup_{p,f} \rho\left(p(y), f(y)\right) = \sup_{p,f} E\left\{p(y), f(y)\right\}$$

with $\| p \| = 1$ and $\| f \| = 1$.

If $\rho(p,f) = 0$, then $p(y), f(y)$ are statistically independent. Therefore, to find the optimal combination of past data to predict the future, we want the maximal correlation.

Determining this structure requires multiple steps. First, given the past and future vectors, the mean is removed to meet the constraints of the alternating conditional expectation (ACE) algorithm that will be used to determine the maximum correlation between transformed input and output variables c and d (Breiman and Friedman, 1985). Then a $(\Sigma_{pp}, \Lambda)$ singular value decomposition of $\Sigma_{pf}$ will determine a J and L such that after the transformations $c(t) = J_k p(t)$ and $d(t) = L_k f(t)$ and the covariances $\Sigma_{cc} = \Sigma_{dd} = I$.

### 3.4.4 State Space Reconstruction

Our final approximation technique, state space reconstruction generates a state space model from an optimal prediction of the future states from linear combinations of the past. Given the data from CVA, or any other identification method, we can use these predictions to parameterize a state space system for any order $k < q$ via a least squares regression.

Assume the following state space system:

$$\begin{aligned} x_{t_{i+1}} &= A(\theta)x(t_i) + B(\theta)u(t_i) + M(\theta)w_d(t_i) \\ y_{t_i} &= C(\theta)x(t_i) + D(\theta)u(t_i) + \\ &\quad O(\theta)w_d(t_i) + \nu(t_i) \end{aligned}$$

Define $m_{t_{i+1}} = J_k p(t_{i+1})$ and $M_t = J_k p(t)$. The state space system above expresses $(x_{ti+1}\ y_t)$ as a linear combination of $(x_t\ u_t)$. We can replace the predicted value of $x_{t_{i+1}}$ and $m_{t_{i+1}}$ with $x_t$ and $m_t$ and express $(m_{t_{i+1}}\ y_t)$ as a linear combination of $(m_t\ u_t)$. With this substitution, all of the data is available for a least squares fit of the two data sets leading to:

$$\begin{pmatrix} AB \\ CD \end{pmatrix} = cov\left\{\begin{pmatrix} m_{t_{i+1}} \\ y_t \end{pmatrix}, \begin{pmatrix} m_{t_i} \\ u_t \end{pmatrix}\right\} cov^{-1}\left\{\begin{pmatrix} m_t \\ u_t \end{pmatrix}, \begin{pmatrix} m_t \\ u_t \end{pmatrix}\right\}$$

with the prediction error given by:

$$S = \begin{pmatrix} S_{11}S_{12} \\ S_{21}S_{22} \end{pmatrix} = cov\left\{\begin{pmatrix} m_{t_{i+1}} \\ y_t \end{pmatrix}, \begin{pmatrix} m_{t_{i+1}} \\ y_t \end{pmatrix}\right\}$$

$$- \; cov\left\{\begin{pmatrix} m_{t_{i+1}} \\ y_t \end{pmatrix}, \begin{pmatrix} m_t \\ u_t \end{pmatrix}\right\} \cdot$$
$$cov^{-1}\left\{\begin{pmatrix} m_t \\ u_t \end{pmatrix}, \begin{pmatrix} m_t \\ u_t \end{pmatrix}\right\} \cdot$$
$$cov\left\{\begin{pmatrix} m_{t_{i+1}} \\ y_t \end{pmatrix}, \begin{pmatrix} m_t \\ u_t \end{pmatrix}\right\}$$

so that

$$\begin{aligned} Q &= S_{11} \\ O &= S_{21}S_{11}^{-1} \\ R &= S_{22} - S_{21}S_{11}^{-1}S_{12} \end{aligned}$$

## 4 SUMMARY

This paper presented methods that support new procedures to expand the set of available metamodel representations beyond the traditional least squares formulation and added the capability to use dynamical metamodels. These methods compliment a new taxonomy of metamodel structures and procedures that allow separation of the metamodeling process into a set of sequential decisions based on *a priori* information.

## REFERENCES


Breiman, L and J.H. Friedman. 1985. Estimating Optimal Transformations for Multiple Regression and Correlation, pp 580-598, *Journal of the American Statistical Association*, Vol. 80, No. 391, September.

Caughlin, D. 1994a. A Metamodeling Approach to Model Abstraction, *Proceedings of the 1994 IEEE Dual-Use Technologies & Application Conference.* SUNY Institute of Technology, Utica/Rome, New York.

Caughlin, D. 1994b. An Evaluation of Simulated Annealing for Modeling Air Combat Simulations, *Proceedings of the 1994 IEEE Dual-Use Technologies & Application Conference.* SUNY Institute of Technology, Utica/Rome, New York.

Caughlin, D. 1995. *Final Report, Modeling Techniques and Applications, Volume I.* USAF Contract F30602-94-0110, Rome Laboratory/IRAE, 32 Hangar Rd, Griffis AFB, NY 13441-4114.

Caughlin, D. 1996. New Procedures to Metamodel Simulations, *Proceedings of the 6th Annual Conference on AI, Simulation, and Planning in High Autonomy Systems*, March 1996.

Goodwin, Payne. 1977. *Dynamic System Identification.* New York: Academic Press.

Ingber, A, L. 1990 Draft of Statistical Mechanical Aids to Calculating Term Structure Models. *Journal Phys. Rev.* Vol 42.

Ingber, A, L. 1993. Simulated Annealing: Practice versus Theory. Reprint from *Journal Mathl. Comput. Modeling.* December.

Juang, J, et.al. 1993. Identification of Observer/Kalman Filter Markov Parameters: Theory and Experiments. pp 320-329, *Journal of Guidance*, Vol 16, No. 2.

Larimore, W. E. 1989. System Identification and Filtering of Nonlinear Controller Markov Processes by Canonical Variate Analysis. Final Report for AF Office of Scientific Research, October 27, 1989.

Larimore,W. E and J. Baillieul. 1990. Identification and Filtering of Nonlinear Systems Using Canonical Variate Analysis. pp 635-640. *Proceedings of the 29$^{th}$ Conference on Decision and Control.*

Ljung, L. 1987. *System Identification: Theory for the User.* New Jersey: Prentice-Hall.

Maine, R. E. and K.W. Iliff. 1981. Formulation and Implementation of a Practical Algorithm for Parameter Estimation with Process and Measurement Noise. *SIAM Journal of Applied Mathematics*, Vol. 41, No. 3,

Maybeck, P. S. 1982. *Stochastic Models, Estimation, and Control, Vol 2*, New York: Academic Press.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. 1986. *Numerical Recipes in FORTRAN.* New York: Cambridge University Press.

Vemuri, V. 1978 *Modeling of Complex Systems*, New York: Academic Press.


## AUTHOR BIOGRAPHY

**DON CAUGHLIN** is Chief Scientist at Mission Research Corporation, Colorado Springs. He received a B.S. in Physics from the Air Force Academy, an MBA from the University of Utah, and M.S. and Ph.D. degrees in Electrical Engineering from the University of Florida. His research interests include system identification, pattern recognition, and intelligent control. Dr. Caughlin has over 28 years experience as an experimental test pilot, research scientist, program manager, and was also Associate Dean of the School of Engineering at the Air Force Institute of Technology. He currently holds an appointment as Research Associate Professor at the University of Colorado at Colorado Springs. He is a senior member of IEEE and AIAA and a member of the Society of Experimental Test Pilots.