2015

Statistical Engineering: A Causal-Stochastic Modeling Research Update

Teddy Steven Cotter
Old Dominion University

Original Publication Citation

Cotter, T. S. (2015). Statistical engineering: A causal-stochastic modeling research update. In S. Long, E-H. Ng, & A. Squire (Eds.), *Proceedings of the American Society for Engineering Management 2015 International Annual Conference* (pp. 1-6). American Society for Engineering Management (ASEM).

Proceedings of the American Society for Engineering Management 2015 International Annual Conference
S. Long, E-H. Ng, and A. Squires eds.

STATISTICAL ENGINEERING: A CAUSAL-STOCHASTIC MODELING RESEARCH UPDATE

T. Steven Cotter
Old Dominion University
tcotter@odu.edu

Abstract
In the ASEM-IAC 2012, Cotter (2012) summarized prior works that led to the proposal for statistical engineering, identified the gaps in knowledge that statistical engineering needs to address, explored additional gaps in knowledge not addressed in the prior works, set forth a working definition of and body of knowledge for statistical engineering, and set forth proposals of potential systems contributions the Engineering Management profession could make toward the development of statistical engineering. In 2014, the ASQ Statistics Division, DOT&E, NASA, and IDA co-sponsored a Statistical Engineering Agreement to jointly research development of the discipline of statistical engineering. The statistics community has continued to frame statistical engineering within the context of the general linear model (GLM). However, incorporating deterministic engineering causal models within the GLM framework leaves missing links of conditional dependencies, yields models that are difficult to fit or that may not converge to a unique solution, and may not increase the understanding of physical causal processes in dynamic stochastic systems. Integration of engineering specific deterministic causal models within stochastic models to provide additional knowledge of the risk of variance from expected response is a key gap in knowledge that must be addressed to realize Statistical Engineering as a discipline. This paper updates research into integrating deterministic engineering models as system dynamic causal components of functional causal Bayesian networks within a state-space framework to model joint deterministic-stochastic dynamic causal effects.

Keywords
Bayesian causal networks, Statistical Engineering

Introduction
On its website, the ASQ Statistics Division (2015) defines statistical engineering as "… the collaborative study and application of the tactical links between statistical thinking and statistical and discipline-specific tools with the objective of guiding better understanding of uncertainty in knowledge and decision-making to generate improved results to benefit the organization and/or society." Cotter (2012) expanded the definition from "discipline-specific" to a general systems definition of "… statistical engineering as the integration of statistical theory with technical, engineering, information systems, managerial, financial, and economic knowledge to solve applied complex organizational and societal problems that involve elements of risk or uncertainty in their outcomes." To model the multivariate nature of such complex problems, Cotter proposed that a general statistical model should reflect the structure and variation of the proposed practical problem being addressed. In general terms, the statistical problem will be,

$$\begin{aligned} &\text{Min } \mathbf{Y}_{\text{Total}} = f(\mathbf{w}'(\mathbf{Y}_{\text{pred}} - \mathbf{T})) \\ &\text{s.t.} \\ &\qquad \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \varepsilon \\ &\qquad \mathbf{L}_X \le \mathbf{X} \le \mathbf{U}_X \\ &\text{possibly} \qquad \mathbf{L}_Z \le \mathbf{Z} \le \mathbf{U}_Z \end{aligned} \tag{1}$$

where f is a generalized transfer function, $\mathbf{Y}_{\text{pred}}$ is the vector or matrix of predicted service, process, and product output characteristics, $\mathbf{T}$ is the vector or matrix of some functional target performance, and $\mathbf{w}$ is a vector of desirability weights that combine the output characteristics in some optimum combination. For each $\mathbf{Y}$ vector, $\mathbf{X}$ is the controllable input variables, $\boldsymbol{\beta}$ the response of $\mathbf{Y}$ to $\mathbf{X}$, $\mathbf{Z}$ is the non-controllable input variables, $\boldsymbol{\gamma}$ is the response

of **Y** to **Z**, and ε is the residual error term. A problem arises with the objective function when some or all of the terms in the vector **T** arises from deterministic mathematical models that represent theoretical economic, information, or physical behavior. In such cases, a strictly linear or even nonlinear statistical model **Y** based on sampled data may not yield a fit that reflects the theoretical behavior. The cause for the lack of fit may be due to any one or a combination of the following reasons.

- Statistical models are based on the recognition that sample data seldom fit simple theoretical functions exactly, because theoretical functions are based partly on restrictive assumptions necessary to explain a particular behavior. The assumptions "assign" unknown degrees of freedom to the structural and random components of theoretical models.
- Some systems exhibit discontinuities and inflection points where the **T** theoretical behavior differs over ranges of **X** predictors. Such change in behavior requires changes in the values of the β or γ coefficients to reflect the change in the response of **Y** to the **T** theoretical behavior.
- The researcher may theorize a relationship $\mathbf{Y} = f(\mathbf{X},\mathbf{Z})$, with its known restrictive assumptions, but be incorrect due to interactions between or among observed **X** variables or due to unobserved latent systemic variables that modify the relationship between **Y** and **X** or **Z**. Both errors arise due to model misspecification or inability to specify the model completely.
- Measurement errors in predictors **X** and **Z** and the response(s) **Y** may not result in a model fit function *f*, even if it is the correct theoretical behavior.
- Sampling error, even from correctly randomized selection processes, will result in samples that do not reflect the known theoretical behavior at the long term type 1 and type 2 error rates. Further, uncontrolled sampling bias may cause variance(s) from the theoretical behavior.

The problem of integrating deterministic theoretical and stochastic models has been addressed sporadically. Mortensen (1969) proposed the following model to account for random environmental forces when modeling deterministic functions.

$$\mathbf{Y}(t) = \boldsymbol{f}(\mathbf{X}(t), t) + \mathbf{G}(\mathbf{X}(t), t)\, \mathbf{v}(t) \tag{2}$$

In the formulation, $\boldsymbol{f}$ is the deterministic gain of the system, **X** is the state of the system at time *t*, **G** accounts for the possibility that noise may influence the gain of the system at time *t*, and **v** a vector of random error at time *t*. Equation (2) is interpreted under Ito-Stratonovich divergence in which **X**(t) becomes a Markov process obtained by a stochastic differential equation. Two problems exist with this approach: (1) The Markov process is a mathematical idealization that only approximates reality. (2) The mathematical approximation cannot accommodate semi-Markovian processes in which errors are correlated and non-Markovian processes which incorporate feedback loops. Shi and Olafsson (1997) presented a general simulation methodology for finite optimization of integrated deterministic and stochastic systems. Miller, Caste, and Temples (2000) integrated deterministic and stochastic methods to characterize contaminated Ecoene aquifers at the Savannah River Site in South Carolina. They modeled scaled gamma-ray values and percent clay deterministically, created multiple equiprobable realizations of lithofacies and grain size, conditioned the lithofacies and grain size to stochastic realizations, and compared models to geological interpretations. The result is a simulation model of contaminated aquifers with limited mathematical interpretation. Min and Zhou (2002) proposed a simulation approach to integrating the deterministic, stochastic, hybrid, and information models of supply chains. As a result of the varying complexity of various supply chain representations, the simulations do not yield mathematical interpretation. Judd, Maliar, and Maliar (2011) presented a generalized stochastic simulation algorithm approach in which precomputation of integrals approximate integrand expressions inside the conditional expectations with parametric basis functions that are separable in endogenous and exogenous state variables.

This work proposes application of functional causal Bayesian networks embedded within a state-space framework for integrating deterministic and stochastic components of dynamic systemic models.

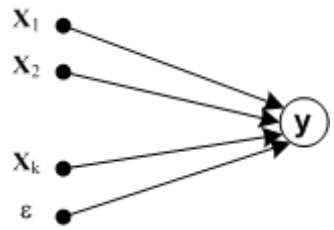
Functional Bayesian Causal Design Topologies

The general linear model (GLM) has been the fundamental working model of statistical regression and current Six Sigma continual improvement. In matrix form, the GLM is constructed for a set of correlated observations, $cor(\mathbf{y}, \mathbf{X})$ as,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{3}$$

where $\mathbf{y}_{n \times j}$ is a matrix of response variables related to a $\boldsymbol{\beta}_{k \times j}$ vector of parameter coefficients through a $\mathbf{X}_{n \times k}$ design matrix of predictor variables, and $\boldsymbol{\varepsilon}_{n \times j}$ is a matrix of random errors with $E[\boldsymbol{\varepsilon}] = 0$ and $cov(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}$. The GLM has been used to model linear, nonlinear, and generalized spatial, hierarchical, and temporal relationships between the **y** response and the **X** predictor variables through the linear covariance between **y** and **β**. The directed acyclic graph (DAG) of the GLM is illustrated Exhibit 1.

Exhibit 1. General Linear Model DAG.



The **β** coefficients are estimated using least squares or maximum likelihood under the assumptions that:

- The $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_k$ predictor variables are fixed, exogenous (independent) and measured with no error.
- Each β_i is estimated as a constant.
- The **β** parameter matrix is considered to be independent of the matrix of errors **ε**.
- The covariance matrix **Σ** of the parameters is assumed to be constant with $var(\mathbf{X}_k) = \sigma^2$ on the diagonal and $cov(\mathbf{X}_i, \mathbf{X}_j) = 0$ on the off diagonal.

Two major problems exist in estimating the **β** parameter coefficients. The first is non-normal variances of the **X** predictors. When the observations **y** in the linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ are normally distributed, the method of least squares yields errors that are independent, $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}\boldsymbol{\sigma}^2)$, and the estimate of the **β** parameters is the maximum likelihood estimate. However, when the **y** observations follow some non-normal distribution, particularly one that has longer or heavier tails or leverage data points, the method of least squares may not be appropriate for estimating the **β** coefficients. Heavy-tailed distributions and leverage data points usually generate outliers, and these outliers may act as influence points on the **β** estimates. Robust regression procedures have been developed to "dampen" the effect of outlier observations that would be highly influential on **β** estimates if least squares estimation is applied. Two primary approaches have been taken to robust regression: $\boldsymbol{L}_p$-norm estimators and ***M***-Estimators. The second and more troublesome problem with no definitive solution is that of collinearity among the **X** predictor variables. Two X_k predictors are collinear if they have a large covariance. In this case, they are said to be confounded in their effect on **y**. Confounding arises from either the **X** predictor variables errors being correlated or from a third, unidentified, exogenous lurking variable **U** not included in the model that affects or predicts the collinear $\mathbf{x}_k$ and $\mathbf{x}_{k+l}$ predictor variables specified in the model. High collinearity causes instability or nonconvergence in the estimate of the **β** coefficients.

In order to address the two major problems in fitting GLMs and integrate deterministic engineering models as system dynamic causal components, this research focuses on building functional causal Bayesian networks within a state-space framework to model joint deterministic-stochastic dynamic causal effects. The **X** controllable and **Z** noncontrollable input variables become endogenous variables of the form

$$\begin{aligned} x_i &= f_i(pa_i, u_{xi}) \quad i = 1 \text{ to } k \text{ predictors} \\ z_j &= f_j(pa_j, u_{zj}) \quad j = 1 \text{ to } l \text{ covariates} \end{aligned} \tag{4}$$

where $f_i(\bullet)$ and $f_j(\bullet)$ take on any linear or nonlinear and constant, temporal, or instantaneous or short-term inflection inducing physical model that accurately represents the dynamics of the process, pa_i and pa_j are the exogenous and possibly endogenous parents of x_i and z_j respectively whose functional form and current values determine the *a priori* Bayesian state of each x_i and z_i respectively, and u_{xi} and u_{zi} are structural and random errors associated with each x_i predictor and z_j covariate respectively (notation taken from Pearl, 2009). The random component of each u_{xi} and each u_{zi} is not restricted to being $N(0, \sigma^2)$ distributed. Deterministic physical models are incorporated in their

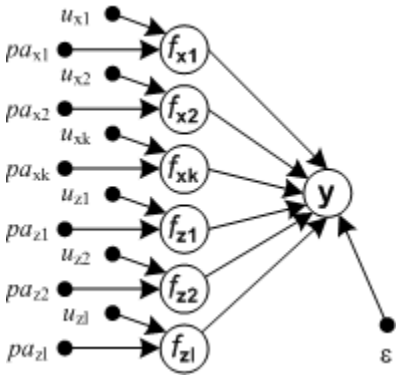
functional form as $\mathbf{x}_i$ controllable and z_j noncontrollable input variables with respective u_{xi} and u_{zj} error terms to reflect structural and random lack of fit. With this functional notation, the GLM of equation (1) now becomes

$$\begin{aligned} &\text{Min } \mathbf{Y}_{\text{Total}} = f(\mathbf{w}'(\mathbf{Y}_{\text{pred}} - \mathbf{T})) \\ &\text{s.t.} \\ &\mathbf{Y} = \mathbf{F}(pa_i, u_{xi})\boldsymbol{\beta} + \mathbf{F}(pa_j, u_{zj})\boldsymbol{\gamma} + \varepsilon \\ &\quad \mathbf{L}_X \le \mathbf{F}(pa_i, u_{xi}) \le \mathbf{U}_X \\ \text{possibly} \quad &\quad \mathbf{L}_Z \le \mathbf{F}(pa_j, u_{zj}) \le \mathbf{U}_Z \end{aligned} \tag{5}$$

where $\mathbf{F}(\bullet)$ is a matrix of functional relationships of the $\mathbf{X}$ predictors and $\mathbf{Z}$ covariates respectively. Where the functional relationship has an unknown form, $f_i(pa_i, u_{xi}) = x_i$ observed data and $f_j(pa_j, u_{zj}) = z_j$ observed covariate values with the residual error accumulating in the ε term. The $\boldsymbol{\beta}$ response parameters of $\mathbf{Y}$ to $\mathbf{X}$ and the $\boldsymbol{\gamma}$ response parameters of $\mathbf{Y}$ to $\mathbf{Z}$ are still constant slope coefficients. Improved fit may be attained by decomposing the deterministic functional forms into systems dynamics elements in the $\mathbf{F}(\bullet)$ functional relationships.

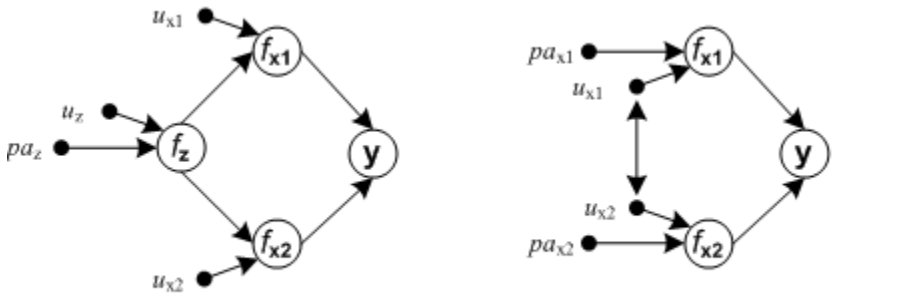
Under this functional causal Bayesian network modeling approach, the GLM of Exhibit 1 is now represented in Exhibit 2.

Exhibit 2. Functional Causal Bayesian Network GLM DAG.



If, as shown in Exhibit 2, the model diagram is acyclic, the model is semi-Markovian, and the values of the $\mathbf{X}$ and $\mathbf{Z}$ variables will be uniquely determined by the pa_i, u_{xi}, pa_j, and u_{zj}. Additionally, if the u_{xi} and u_{zj} are jointly independent and the $f_{xi} = x_i$ and the $f_{zi} = z_i$, the model is Markovian and the GLM results. If the diagram is not acyclic, as is the case for non-Markovian models, the values of $\mathbf{X}$ and $\mathbf{Z}$ cannot be uniquely determined but can be only bounded. Similarly, under this functional causal Bayesian network modeling approach, lurking variable confounding and correlated errors confounding are explicitly modeled as illustrated in Exhibit 3.

Exhibit 3. Functional Causal Bayesian Network Confounding DAG.



(a) Lurking Variable Confounding (b) Correlated Errors Confounding

State Space Dynamics

Transitioning from a strictly spatial GLM to a dynamic functional causal Bayesian network modeling approach necessitates implementation within a state space dynamic modeling framework in order to track the $\mathbf{F}(pa_i, u_{xi})$, $\mathbf{F}(pa_j, u_{zj})$, and $\mathbf{Y}$ state matrices. This requires the state space equations to take the form

$$\begin{aligned} d/dt\,[\mathbf{X}|\mathbf{Z}](t) &= F([\mathbf{F}(pa_i, u_{xi}) \mid \mathbf{F}(pa_j, u_{zj})](t),[\mathbf{U}_i(t)|\mathbf{U}_j(t)], t) \\ \mathbf{Y}(t) &= G([\mathbf{X}(t) \mid \mathbf{Z}(t)](t), [\mathbf{U}_i(t)|\mathbf{U}_j(t)], t) \end{aligned} \quad (6)$$

This requires only restating the traditional state space $\mathbf{X}(t)$ and $\mathbf{U}(t)$ matrices into partitioned forms $[\mathbf{X} \mid \mathbf{Z}](t)$ and $[\mathbf{U}_i \mid \mathbf{U}_j](t)$, which can be implemented in existing state space modeling software. If a given model diagram is semi-Markovian with only statistically non-significant confounding or the model diagram is strictly Markovian and the $f_{xi} = x_i$ and the $f_{zi} = z_i$, existing state space modeling software's matrix multiplication can be used without further modification. Conversely, if the model diagram is not acyclic or statistically significant confounding exists or dynamic causal components of $f(pa_i, u_i)$ and $f(pa_j, u_j)$ must be modeled, the regression model must be updated sequentially from input exogenous to endogenous **X** and **Z** variables to predicted **Y** response, and simulation or systems dynamics software will be required.

An Integrated Stochastic-Causal Modeling Framework

Current research is directed toward developing an integrated stochastic-causal modeling framework. In order for the framework to be effective and efficient in deriving functional causal Bayesian models of joint causal-stochastic dynamic effects, the following issues must be addressed:

- Model design decision rules must be developed to guide the model building process. These rules must provide guidance on developing the correct functional forms of $f(pa_i, u_i)$ and $f(pa_j, u_j)$ and their graphical relationships. This is the most critical modeling step, and it is currently left to the modeler's knowledge of the process and his or her intuition.
- Simulation and systems dynamics software capabilities must be identified in terms of modeling worst-case non-Markovian models.
- Interoperative coding must be developed to integrate the simulation or systems dynamics software within existing state space modeling software.
- In cases where non-Markovian models must be used to represent dynamic processes, guidance must be developed on the correct coding sequence to update the simulation or systems dynamic model. Coding sequence will be particularly critical in non-acyclic models that exhibit high confounding among the **X** and **Z** variables.
- Systems dynamics and causal Bayesian modeling rules must be integrated into the design and modeling code to alert the modeler of rule violations.
- Causal Bayesian bounding algorithms must be integrated into the modeling code to guide identification of **X** and **Z** bounds in non-Markovian models.

Continuing Research into Stochastic-Causal Modeling

Once a cogent functional causal Bayesian modeling framework is worked out, research can be initiated toward testing whether complete differential forms of deterministic and stochastic casual effects or decomposition of the differential forms into systems dynamics elements provide improved modeling accuracy. Again, if it is found that decomposition of differential forms into systems dynamics elements provided improve modeling accuracy, model design rules will have to be developed to guide the decomposition and modeling processes.

References

American Society for Quality Statistics Division. (2015) Statistical Engineering. Retrieved from http://asq.org/statistics/quality-information/statistical-engineering.

Cotter, T. (2012) Engineering Management Contributions to Statistical Engineering. *Proceedings from the 2012 ASEM National Conference*, (pp. 108-115). Virginia Beach.

Judd, K, Maliar, L, and Maliar, S. (2011) How to Solve Dynamic Stochastic Models Computing Expectations Just Once. *NBER Working Paper No. 17418*.

Miller, R., Castle, J. and Temples, T. (2000) Deterministic and Stochastic Modeling of Aquifier Stratigraphy, South Carolina. *GROUND WATER*, 38(2), 284-295.

Min, H. and Zhou, G. (2002) Supply chain modeling: past, present, and future. Computers and Industrial Engienering, 43, 231-249.

Mortensen, R. (1969) Mathematical Problems of Modeling Stochastic Nonlinear Dynamic Systems. Journal of Statistical Physics, 1(2), 272-296.

Pearl, J. (2013) Causality: Models, Reasoning, and Inference. New York: Cambridge University Press.

Shi, L. and Olaffsson, S. (1997) An Integrated Framework for Determistic and Stochastic Optimization. *Proceedings of the 1997 Winter Simulation Conference*, (pp. 358-365), Atlanta.

About the Author

T. Steven Cotter is a Lecturer with the Engineering Management and Systems Engineering department at Old Dominion University. He earned a Ph.D. in Engineering Management and Systems Engineering from Old Dominion University, a Master of Science in Engineering Management with a concentration in quality/reliability engineering from the University of Massachusetts at Amherst, a Master of Business Administration with a concentration in finance and a Bachelor of Science both from the University of South Carolina, and a diploma in Electronic Technology from Graff Area Vocational and Technical School (now Ozarks Technical Community College). He is a certified Quality Engineer and Reliability Engineer with the American Society for Quality. His research interests are in engineering analytics design, human-machine intelligent socio-technical organizations, quality systems design, and statistical engineering.