An Exploratory Statistical Cusp Catastrophe Model

Ding-Geng (Din) Chen
School of Social Work &
Department of Biostatistics,
Gillings School of Global Health
University of North Carolina
Chapel Hill, NC 27599, USA
Department of Statistics, University of Pretoria,
Pretoria, South Africa
Email: dinchen@email.unc.edu

Xinguang (Jim) Chen
Department of Epidemiology, College of Public Health &
Health Professions, College of Medicine
University of Florida, Gainesville, FL 32610, USA
Email: jimax.chen@ufl.edu

Kai Zhang
Department of Statistics, University of North Carolina
Chapel Hill, NC 27599, USA
Email: zhangk@email.unc.edu

***Abstract*—The Cusp Catastrophe Model provides a promising approach for health and behavioral researchers to investigate both continuous and quantum changes in one modeling framework. However, application of the model is hindered by unresolved issues around a statistical model fitting to the data. This paper reports our exploratory work in developing a new approach to statistical cusp catastrophe modeling. In this new approach, the Cusp Catastrophe Model is cast into a statistical nonlinear regression for parameter estimation. The algorithms of the delayed convention and Maxwell convention are applied to obtain parameter estimates using maximum likelihood estimation. Through a series of simulation studies, we demonstrate that (a) parameter estimation of this statistical cusp model is unbiased, and (b) use of a bootstrapping procedure enables efficient statistical inference. To test the utility of this new method, we analyze survey data collected for an NIH-funded project providing HIV-prevention education to adolescents in the Bahamas. We found that the results can be more reasonably explained by our approach than other existing methods. Additional research is needed to establish this new approach as the most reliable method for fitting the cusp catastrophe model. Further research should focus on additional theoretical analysis, extension of the model for analyzing categorical and counting data, and additional applications in analyzing different data types.**

Keywords—Cusp Catastrophe Model; bifurcation; asymmetry; bootstrapping; HIV prevention.

I. Introduction

Typically, the statistical models used to examine health outcomes are based on a linear regression approach. However, health outcomes are rarely linear because of the multiple, complex influences of environmental, behavioral, psychological, and biological factors. What might appear to be small and inconsequential changes in these factors can lead to abrupt changes in health outcomes. Under these conditions, a linear approach seriously limits the predictive value of the influence of hypothesized factors on a particular health outcome [16]. To account for nonlinearity in low-dimensional scenarios, statisticians turn to natural extensions of linear regression, which are nonparametric regression methods such as kernel regression or regression/smoothing splines. Other techniques for use with high-dimensional data include additive models, multivariate adaptive regression splines, random forests, neural networks, and support vector machine; these techniques have been discussed extensively elsewhere [15, 17]. However, these nonparametric regressions lack a mechanism to identify and incorporate cusp jumps; the presence of such a mechanism is the fundamental advantage offered by the Cusp Catastrophe Model.

As a complement to traditional analytical approaches, the Cusp Catastrophe Model offers distinct advantages given its capacity to not only simultaneously handle complex linear and nonlinear relationships in a high-order probability density function but also to incorporate sudden jumps in outcome measures, as outlined in Zeeman [24]. Catastrophe theory was proposed in the 1970s [1] to understand a complicated set of behaviors that included gradual, continuous changes as well as sudden and discrete or catastrophic changes. The Cusp Catastrophe Model has been used extensively in a wide range of research, including modeling of accident process [7], adolescent alcohol use [9], changes in adolescent substance use [10], binge drinking among college students [11], sexual initiation among young adolescents [12, 22], nursing turnover [13], HIV prevention [12, 14, 23], therapy and program evaluation [6], and health outcomes [27].

Historically, two approaches or methods have been used to apply the cusp catastrophe theory in the analysis of research data. One method was operationalized by Guastello [6,7] using a polynomial regression approach (PolyCusp). The second method uses a stochastic differential equation Cusp Catastrophe Model (SDECusp) from Cobb and his colleagues [5] with likelihood estimation implemented in an R package [8]. This paper introduces the exploratory development of a third method that casts the Cusp Catastrophe Model in the framework of a statistical nonlinear regression (StatCusp). Our introduction of this third method allows statisticians to fit a cusp model using standard statistical inferences.

To present this approach, Section II first gives an overview of the Cusp Catastrophe Model and then presents the implementation of PolyCusp and SDECusp. In Section III, we describe the novel development of StatCusp and present simulation studies to illustrate the properties of this novel approach. In Section IV, we show the application of StatCusp

978-1-5090-5206-6/16 $31.00 © 2016 IEEE
DOI 10.1109/DSAA.2016.17

to a real-world dataset. Section V provides our discussion and conclusions.

II. Overview of the Cusp Catastrophe Model

A. Cusp Catastrophe Model

Catastrophe theory was proposed in the 1970s by Thom [1] and popularized over the next two decades by several leading statisticians [1-5]. Thom [1] originally proposed the catastrophe theory to understand complicated phenomena that included both gradual, continuous change and sudden, discontinuous or catastrophic change. According to this theory, the presence of catastrophe is defined by five elements, as summarized in Gilmore [28] and others [18-21]:

1. *bimodality* (i.e., existence of two distinctly different behavioral modes);
2. *sudden jump* (i.e., abrupt changes in outcomes between the two modes even with slight changes in the predictors);
3. *inaccessibility* (i.e., an outcome unlikely to be in the area between the two modes);
4. *hysteresis* (i.e., the change of an outcome from one mode to the other cannot be determined by control factors of the same value); and
5. *divergence* (i.e., a slight change in the control factors can lead to substantial change in the outcome and deviation from the linear model).

In summary, the Cusp Catastrophe Model would be a particularly appropriate statistcal approach when an outcome measure has the properties of bimodal distribution (bimodality) with spurts (sudden jumps), an inaccessible middle region between these two modes (inaccessibility), a delay between these transitions (hysteresis), and deviation from a linear relationship between the response outcome measure and the predictors (divergence).

Even though the Cusp Catastrophe Model has been well established theoretically and applied extensively in the physical sciences, the application of this model in the social and behavioral sciences has been criticized [for discussion, see 29, 30].

To apply this theoretical model in research, the deterministic Cusp Catastrophe Model can be specified with three components: two control factors (i.e., α and β) and one outcome variable (i.e., z). This model is defined by a dynamic system:

$$\frac{dz}{dt} = \frac{dV(z;\alpha,\beta)}{dz} \qquad (1)$$

where the potential function V is defined as

$$V(z;\alpha,\beta) = \alpha z + \frac{1}{2}\beta z^2 - \frac{1}{4}z^4$$

In this potential function V, α is the asymmetry or normal control factor, and β is the bifurcation or splitting control factor. Both α and β are linked to determine the outcome variable z in a three-dimensional response surface. When the right side of equation (1) moves toward zero, the outcome z does not change with time; this status is called equilibrium. In general, the behavior of the outcome z (i.e., how z changes with time t) is complicated, but each subject moves toward equilibrium. Figure 1 graphically depicts the equilibrium plane that reflects the response surface of the outcome measure (z) at various combinations of the asymmetry control factor (x as the measure of α in Figure 1) and the bifurcation control factor (y as the measure of β in Figure 1).

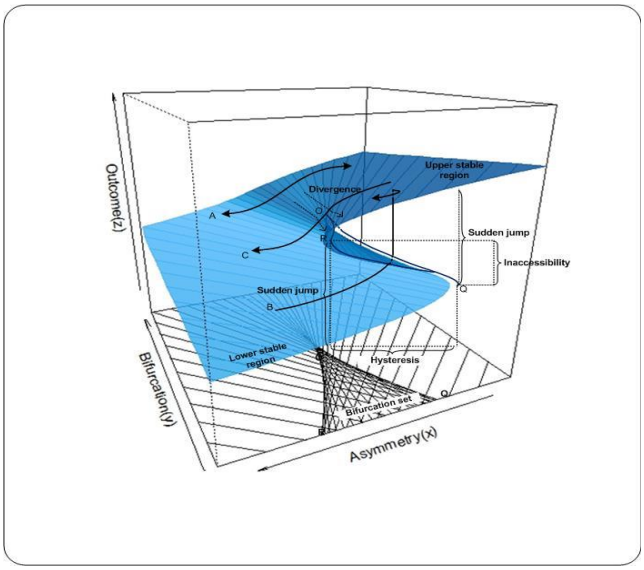

Figure 1. Cusp Catastrophe Model for outcome z in the equilibrium plane with an asymmetry control variable (x as the measure of α) and a bifurcation control variable (y as the measure of β).

As shown in Figure 1, the dynamic changes in z have two stable regions (attractors), which are the lower area in the front left (lower stable region) and the upper area in the front right (upper stable region). Beyond these stable regions, z becomes sensitive to changes in x and y. This unstable region can be projected to the control plane (x, y) as the *cusp region*. The cusp region is characterized by line OQ (the ascending threshold) and line OR (the descending threshold) of the equilibrium surface. In this region, z becomes highly unstable with regard to changes in x and y, jumping between the two stable regions when (x, y) approaches the two threshold lines OQ and OR. In Figure 1, paths A, B, and C depict three typical but distinct pathways of change in the health outcome measure (z). Path A shows that in situations where $y < O$, a smooth relation exists between z and x. Path B shows that in situations when $y > O$, if x increases to reach and pass the ascending threshold link OQ, z will suddenly jump from the low stable region to the upper stable region of the equilibrium plane. Path C shows that a

sudden drop occurs in z as x declines to reach and pass the descending threshold line OR.

The Cusp Catastrophe Model can be used with both qualitative and quantitative research methods to evaluate outcome measures (e.g., behaviors or health outcomes). The qualitative approach focuses on identifying the five catastrophe elements (i.e., catastrophe flags) outlined by Gilmore [28], whereas the quantitative approach uses numerical data to statistically solve the model. Since the introduction of the Cusp Catastrophe Model, two quantitative approaches have been developed and used to implement the model: the PolyCusp approach and the SDECusp approach. To provide a complete overview of the Cusp Catastrophe Model, we briefly outline these two traditional approaches.

B. PolyCusp Model

To operationalize the Cusp Catastrophe Model for research, Guastello [6,7] developed the polynomial regression approach—the PolyCusp method—to implement the concept of the cusp model. Since the first publication of this method, PolyCusp has been widely used in analyzing research data (see Section I). According to Guastello, PolyCusp is derived by reformulating the cusp dynamic system in Equation (1) from a differential equation into a difference system, with change scores $\Delta z = z_2 - z_1$ (the differences in the measurement scores of a behavior assessed at Time 1 and Time 2) as a numerical approximation of dz:

$$\Delta z = \beta_0 + \beta_1 z_1^3 + \beta_2 z_1^2 + \beta_3 y z_1 + \beta_4 x + \beta_5 y + \varepsilon \quad (2)$$

where β_0 is the intercept and ε is the normally distributed error term. Two additional terms, $\beta_2 \times z_1^2$ and $\beta_5 \times y$, are added to the equation to capture potential deviations of the data from the equilibrium plane. When conducting a modeling analysis, a cusp is indicated ***only*** when the estimated β_1 for the cubic term, plus β_3 (for the interaction term) or β_4 (for control variable x) in Equation 2 are statistically significant.

To demonstrate the efficiency of the PolyCusp method in describing behavioral changes that are cusp, Guastello [7] recommended a comparative approach. In this approach, two types and four alternative linear models are constructed and used in modeling the same variables:

- Change scores linear models
  $\Delta z = \beta_0+\beta_1 z_1+\beta_4 x+\beta_5 y$ (3)
  $\Delta z = \beta_0+\beta_1 z_1+\beta_3 y z_1+\beta_4 x+\beta_5 y$ (4)

- Pre-and post- linear models
  $z_2 = \beta_0+\beta_1 z_1+\beta_4 x+\beta_5 y$ (5)
  $z_2 = \beta_0+\beta_1 z_1+\beta_3 y z_1+\beta_4 x+\beta_5 y$ (6)

These alternative linear models add another analytical strategy to strengthen the polynomial regression method. A version of the cusp model (2) with better fit to the data than the alternative linear models (3 thru 6) is often used as additional evidence to support the hypothesis that the dynamics of a study behavior follows the Cusp Catastrophe Model. All procedures for fitting Guastello's PolyCusp regression model and the four alternative models can be conducted with commonly available statistical software, including SAS, SPSS, STATA, and R. Further discussion and applications of theses cusp catastrophe modeling methods are available elsewhere [16, 21]. The validity of the PolyCusp model has been criticized by Alexander et al. [31], who presented extensive simulation studies to demonstrate that the PolyCusp method cannot adequately distinguish between data from a true catastrophe model and data from a true linear model. Given this limitation, this paper does not further examine the PolyCusp method.

C. SDECusp Model

Another method for applying the Cusp Catastrophe Model in research is the stochastic differential equation approach, or the SDECusp method. In this approach, the deterministic cusp model in equation (1) is first extended with a probabilistic/stochastic Wiener process. With this extension, the modeling process incorporates measurement errors of the outcome variable. Under this approach, the response surface of the cusp catastrophe is modeled as a probability density function where the bimodal nature of the outcome corresponds to the two states of outcome variable. Mathematically, Cobb and his colleagues [3-5] cast the deterministic cusp model in Equation (1) into a stochastic differential equation (SDE) as follows:

$$dz = \frac{\partial V(z,\alpha,\beta)}{\partial z} dt + dW(t) \quad (7)$$

where $dW(t)$ is a white noise Wiener process with variance σ^2. This extension is in fact a special case of general stochastic dynamical system modeling with a constant diffusion function defined by $dW(t)$. The SDECusp model in equation (7) cannot be solved analytically; therefore, computational implementation of SDECusp in health outcomes research is limited. However, at the equilibrium state when time (t) approaches infinity, it is easier to estimate the probability density function of the corresponding limiting stationary stochastic processes. In other words, the probability density function of the outcome measure (z) [19, 32] can be expressed as follows:

$$f(z) = \frac{\Psi}{\sigma^2} exp\left[\frac{\alpha(z-\lambda)+\frac{1}{2}\beta(z-\lambda)^2-\frac{1}{4}(z-\lambda)^4}{\sigma^2}\right] \quad (8)$$

where the parameter ψ is a normalizing constant and λ is used to determine the origin of z. With this probability density function, the regression predictors α and β can be incorporated as linear combinations to replace the canonical asymmetry factor (i.e., x) and bifurcation factor (i.e., y). Note that as a distribution for a limiting stationary stochastic process, this probability density function in equation (8) is independent from time t, thus it can be used to model cross-sectional relationship

with the advantage to detect and quantify its potential cusp nature comprising both sudden and continuous states. Moreover, the probability density function allows the well-known statistical theory of maximum likelihood to be used for model parameter estimation and statistical inference. *R* Package "cusp" has been developed to implement the stochastic Cusp Catastrophe Model [8].

When conducting this modeling analysis, the following criteria are applied to examine the difference in goodness-of-fit between the cusp model and the alternative linear regression and nonlinear logistic regression models: (a) negative log-likelihood values and the associated likelihood-ratio Chi-square test with smaller negative log-likelihood values indicate better model fit; and the associated likelihood-ratio Chi-square test is defined as twice the difference of the negative log-likelihood values between the cusp and the linear/nonlinear logistic regression models with p-value < 0.05 indicating a better fit of the cusp model than either the linear or logistic regression models; (b) R^2 defined as R^2=1-(error variance/variance of z) where the larger the R^2, the better the model fit. Note that in cusp modeling, this R^2 is referred to as a pseudo-R^2 since the value could be negative; (c) Akaike information criterion (*AIC)* [33] and Bayesian information criterion (*BIC*) [34] where smaller values of *AIC* and/or *BIC* indicate better model fit; and (d) at least 10% of the control factor (x, y) data pairs lie within the bifurcation cusp region [19,35]. An alternative and more stringent test for this 10% guideline was proposed by Hartelman [19]; Hartelman et al. [36]; and van der Maas et al. [37] as a nonlinear least squares regression with the logistic curve:

$$z_i = \frac{1}{1+e^{-\alpha_i/\beta_i^2}} + \varepsilon_i \qquad (9)$$

where $i = 1, ..., n$.

This SDECusp model is extremely well-suited for use with cross-sectional data and the R package. We have used this SDECusp model extensively for research and publications [16, 12, 22, 23, 27]. However, we found that this approach is not a viable method when using other types of data such as binary data and counts data. In addition, adequate methods for modeling longitudinal data have not yet been developed, even though modeling longitudinal data is required in real-world research applications. Therefore, extensions or different approaches are needed to fill this gap, which led us to develop the StatCusp model.

III. THE EXPLORATORY STATCUSP MODEL

A. The Model

We first describe the StatCusp model for continuous data as a conceptual model that is guided by the statistical theory of generalized linear models and can be extended to the analysis of binary data, counts data, and mixed-effects models for longitudinal and multi-level data.

Following equation (1), the StatCusp model can be formulated as following:

$$z_i = Z_i + \varepsilon_i \qquad (10)$$

where ε_i (i = 1,…,n) are the residuals from n observations, and are assumed to be normally distributed as $\epsilon_i \sim N(0, \sigma^2)$; Z_i in equation (10) is one of the real roots of the deterministic cusp catastrophe equation:

$$\alpha_i + \beta_i Z_i - Z_i^3 = 0 \qquad (11)$$

where α_i and β_i are two control variables. For any observed data with p independent variables $(x_1, \dots, x_p)$ and the outcome variable z_i, the control variables α_i and β_i are modeled as follows:

$$\alpha_i = a_0 + a_1 x_1 + \cdots + a_p x_p \qquad (12)$$

$$\beta_i = b_0 + b_1 x_1 + \cdots + b_p x_p \qquad (13)$$

With the formulations of equations (10) to (13), a nonlinear regression can be used to estimate the model parameters of $\boldsymbol{a} = (a_0, a_1, \dots, a_p)$, $\boldsymbol{b} = (b_0, b_1, \dots, b_p)$ from equation (12) and (13). The residual variance σ^2 can be estimated using equation (10) to minimize the objective function of the sum of squared errors (SSE). This step can be theoretically formulated as follows:

$$SSE(\boldsymbol{a}, \boldsymbol{b} | data) = \sum_{i=1}^{n} (z_i - Z_i)^2 \qquad (14)$$

Equivalently, these model parameters can be estimated using maximum likelihood estimation, with the likelihood function formulated as follows:

$$L(\boldsymbol{a}, \boldsymbol{b}, \sigma^2 | data) = \left(\frac{1}{\sqrt{2\pi\sigma}}\right)^n exp\left(-\frac{\sum_{i=1}^{n}(z_i - Z_i)^2}{2\sigma^2}\right) \qquad (15)$$

B. Special Properties for Cusp Models

Theoretically, the StatCusp model formulated from equations (10) to (13) should have all the required statistical properties, such as being unbiased and an efficient method of variance estimation. However, this StatCusp model is not the traditional statistical model in which each combination of independent variables is associated with only one outcome value. In fact, the StatCusp model formulated here could have one, two, or three roots from the cusp catastrophe equation (11) based on the Cardan discriminant:

$$\Delta = 27\alpha^2 - 4\beta^3. \qquad (16)$$

It is clear that when $\Delta > 0$, equation (11) has one real root. However, when $\Delta \leq 0$, equation (11) has three real roots. Among these three roots, there are three cases: (a) if $\alpha = \beta =$

$\Delta = 0$, the three roots are the same, which is referred as the *cusp point* (labeled O in Figure 1); (b) if $\Delta = 0$, but $\alpha \neq 0$ *or* $\beta \neq 0$, two roots are the same, which forms the boundary for the cusp region formed by the two lines OQ and OR in Figure 1; and (c) if $\Delta < 0$, and $\alpha \neq 0$ *or* $\beta \neq 0$, the three roots are distinct from one another, which characterizes the cusp region between OQ and OR as illustrated in Figure 1. Therefore, this StatCusp model is no longer within the traditional domain of mathematical and statistical modeling. Further investigations are needed to identify the statistical properties of this StatCusp model.

To select the correct root for modeling equation (11), the field proposes two modeling conventions; the *delay convention* and the *Maxwell convention*. The delay convention is used to select the root from the cusp surface of $\frac{dV(z;\alpha,\beta)}{dz} = 0$ in equation (1) that are close to the observed *z*. The Maxwell convention is used to select the roots on the cusp surface of $\frac{dV(z;\alpha,\beta)}{dz} = 0$ in equation (1) corresponding to the minimum of the associated potential function $V(z;\alpha,\beta) = \alpha z + \frac{1}{2}\beta z^2 - \frac{1}{4}z^4$.

C. Monte-Carlo Simulation

To verify the novel StatCusp model, we designed simulations with known parameters. Data are simulated from equation (10) with $\sigma = 0.5$ and the number of observations $n = 300$. Two independent variables x_1 and x_2 are simulated independently from the standard normal distribution. To test whether the StatCusp model can correctly distinguish and determine the model variables, we make use of the true parameters of $\boldsymbol{a} = (2, 2, 0)$, $\boldsymbol{b} = (2, 0, 2)$ from equations (12) and (13) where $a_2 = 0$ in equation (12) to represent the correct model selection of x_1 from equation (12) and $b_1 = 0$ to represent the correct model selection of x_2 from equation (13).

Data are simulated in following steps:

1. With $n = 300$, simulate x_1 and x_2 from the standard normal distribution and also simulate ε_i from normal distribution with mean zero and standard deviation σ;
2. With the true parameters of $\boldsymbol{a} = (2, 2, 0)$, $\boldsymbol{b} = (2, 0, 2)$ and the x_1 and x_2 from Step 1, calculate α_i and β_i from equations (12) and (13);
3. With the α_i and β_i from Step 2, solve equation (11) to obtain Z_i and select the one root corresponding to the Maxwell convention, which is the minimum of the associated potential function $V(Z_i, \alpha_i, \beta_i)$;
4. With the selected Z_i from Step 3, the outcome variable z_i can be generated by using equation (10);
5. Using the results from Steps 1 through 4 as data, the objective function can be formed to estimate the parameters $\boldsymbol{a}$ and $\boldsymbol{b}$ based on equation (14) if using nonlinear regression or equation (15) if using maximum likelihood estimation.

We ran these simulations steps 5,000 times and kept a record of the estimated parameters for investigations. The simulation results are summarized in Table 1. Columns in Table 1 labeled "Mean" and "Med" are the mean and median from the 5,000 estimated parameters. These estimates are very close to the true parameters of $\boldsymbol{a} = (2, 2, 0)$ and $\boldsymbol{b} = (2, 0, 2)$, suggesting the estimation equation (14) is unbiased. This unbiased property is true for the estimates of σ^2 where the mean and median from these 5,000 simulations are 0.501 and 0.501, respectively. This unbiasedness is graphically illustrated in Figure 2.

Table 1: Summary of Simulation 1

	Mean	Med	EmpV	EstV	ECP
a_0	2.0094	2.0035	0.0079	0.9525	0.3323
a_1	2.0106	2.0062	0.0134	1.2496	0.2558
a_2	-0.0014	-0.0009	0.0082	0.3232	0.2502
b_0	2.0038	2.0016	0.0048	0.3240	0.3093
b_1	-0.0069	-0.0029	0.0102	0.7649	0.2483
b_2	2.0115	2.0057	0.0169	1.4016	0.2246

The column "EmpV" is the variance of estimated parameters from 5,000 simulations, or the sample variance, which are the consistent estimates of the true variance estimates based on the statistical theory. The column "EstV" is the average of the 5,000 estimated parameters from the Fisher information matrix, and the column "ECP" is the empirical coverage probability. ECP should be close to 95% if the EstV is close to EmpV.

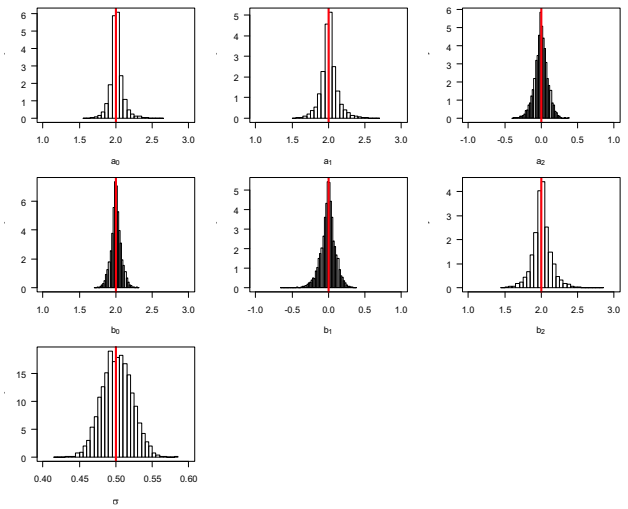

Figure 2. Distributions of simulated parameters from the StatCusp model.

However, the results shown in Table 1 indicate the estimates of ECP for all parameters are less than 95%, which indicates the Fisher information matrix lacks efficiency for variance estimation. Figure 2 depicts this, showing that the distributions from the estimated parameters of $\boldsymbol{a}$ and $\boldsymbol{b}$ are highly leptokurtic. Therefore, a procedure to adjust the variance estimation is needed to validate the StatCusp model. To overcome this issue, we turn to a bootstrapping procedure.

D. Boostrapping Estimation for Variance

Bootstrapping is commonly used in statistics to estimate variance when the typical Fisher information matrix is not correct. For variance estimation, a small number of bootstrap samples (i.e., from 200 to 300) are sufficient; we choose 300 in our simulations.

This bootstrapping procedure is implemented with an additional bootstrapping step from Section III (C) as follows:

1. With n = 300, simulate x_1 and x_2 from the standard normal distribution and also simulate ε_i from normal distribution with mean zero and standard deviation σ;
2. With the true parameters of $\boldsymbol{a}$ = (2, 2, 0), $\boldsymbol{b}$ = (2, 0, 2) and the x_1 and x_2 from Step 1, α_i and β_i are calculated using equations (12) and (13);
3. With the α_i and β_i from Step 2, solve equation (11) to obtain Z_i and select the one root corresponding to the Maxwell convention: the minimum of the associated potential function $V(Z_i, \alpha_i, \beta_i)$;
4. With the selected Z_i from Step 3, the outcome variable z_i can be generated by using equation (10);
5. With the generated data from Steps 1 to 4, the objective function is formulated in equation (14) if using nonlinear regression, or equation (15) if using maximum likelihood estimation to estimate $\boldsymbol{a}$ and $\boldsymbol{b}$;
6. (The bootstrapping step) Bootstrap the data from Step 4 and re-run the estimation Step 5 300 times to generate a bootstrapping sample. The bootstrapping sample (n = 300) can be used for two purposes: (a) to estimate the variances (denoted by "ECP1"), and (b) to construct 95% confidence intervals (CI) for each estimate (denoted by "ECP2") from Step 5 to construct the empirical coverage probability.

Table 2: Summary of Simulation 2 with Bootstrapping

	Mean	Med	EmpV	EstV	ECP1	ECP2
a_0	2.023	2.014	0.037	0.059	0.980	0.988
a_1	2.031	2.009	0.0614	0.098	0.977	0.980
a_2	-0.014	0.002	0.0363	0.050	0.977	0.967
b_0	2.005	2.009	0.0195	0.029	0.982	0.982
b_1	-0.023	-0.009	0.0467	0.073	0.982	0.980
b_2	2.027	2.010	0.0787	0.126	0.980	0.990

The results from this bootstrapping procedure are summarized in Table 2 where the column "EstV" is the estimated variance obtained from the 300 bootstrapping samples, which are much closer to the empirical variance in the column "EmpV". The column "ECP1" is the coverage probability using the bootstrapping variance, and the "ECP2" is the coverage probability using the bootstrapping CI, which Table 2 shows are much closer to 95% than the values obtained without bootstrapping.

IV. Real Data Analysis

A. Data

Data used to demonstrate the application of the new method were derived from an NIH funded project (Award #: R01 MH069229, PI: Stanton and Chen) designed to provide HIV prevention education to adolescents in the Bahamas. Students in Grade 9 were randomly selected to receive the intervention. Data were collected in classroom settings using paper and pencil questionnaires. Data used for this analysis were collected when the students participating in the study advanced to Grade 12 (n = 1,790, 40.6% male, mean age = 16.9 years, SD = 0.74).

The outcome variable z = self-efficacy for condom use (mean score = 4.36, SD = 0.80). This variable was measured using six survey items (Cronbach alpha = 0.81). A typical item read, "I could convince my partner that we should use a condom even if he (she) doesn't want to." Individual items were measured using a 5-point Likert scale (1 = *no, not at all*; 2 = *probably not;* 3 = *don't know/unsure;* 4 = *probably yes;* 5 = *certainly yes*). Mean scores were computed for analysis, with higher scores indicating higher levels of self-efficacy in condom use.

The asymmetry variable x_1 is HIV knowledge (mean = 14.29, SD=2.39). Student knowledge of HIV transmission and prevention was measured using 18 items. An example item read, "A woman can get HIV if she has anal sex with a man who has HIV." Correct answers were scored as 1 point, with higher total scores indicating a higher level of HIV-related knowledge. The total scores had a possible range from 0 (no HIV knowledge) to 18 (fully knowledgeable).

The bifurcation variable x_2 = response efficacy (mean score = 4.36, SD = 0.88). This variable is defined as the perceived effectiveness of condom use in preventing HIV infection, and was assessed using three items (Cronbach alpha = 0.80). An example item is, "Condoms are an important way to prevent you from getting a sexually transmitted disease (STD)." All three items used the same the 5-point Likert scale that ranged from 1 = *strongly disagree*, to 3 = *neutral*, to 5 = *strongly agree*).

In theory, as x_1 or HIV knowledge increases, adolescents are more confident that they will use a condom during sex to prevent the transmission of HIV. However, this process can be bifurcated by x_2, the response efficacy or perceived effectiveness of condom use. That is, when x_2 is below the bifurcation point, the positive relationship between HIV knowledge and condom use self-efficacy will be continuous; however, when x_2, the perceived condom efficacy, is greater than the bifurcation point, changes in condom use self-efficacy will manifest as a discrete process with two z values distributed at all (x_1, x_2) combination points on the cusp surface.

B. *Linear Regression Analysis*

The linear regression was used first to fit the data as in conventional statistical analysis. The following are the modeling results:

```
            Estimate   Std. Error t value Pr(>|t|)
(Intercept) 2.876649   0.118648   24.245  < 2e-16 ***
A(x1)       0.047284   0.007967    5.935 3.46e-09 ***
B(x2)       0.203343   0.020204   10.065  < 2e-16 ***

Residual standard error: 0.7677 on 1992 degrees of freedom
Multiple R-squared:  0.07986,  Adjusted R-squared:  0.07894
F-statistic: 86.45 on 2 and 1992 DF,  p-value: < 2.2e-16
```

Results from the multiple linear regression analysis indicate a positive relationship between the two predictor variables x_1 (HIV knowledge, $\beta = .0080$, $p < .01$) and x_2 (the perceived condom efficacy, $\beta = .2033$, $p < .01$) with the outcome variable z (condom use self-efficacy). R^2, including adjusted R^2 was less than 8%.

C. *SDECusp Modeling Analysis*

We modeled the same data based on the theory of the SDECusp model and the R package "cusp" developed by Grasman [8]. The package can be freely downloaded and the description of this package can be found at http://cran.r-project.org/web/packages/cusp/vignettes/Cusp-JSS.pdf. All variables were standardized before analysis based on the suggestion from the package. The following are the results from the analysis conducted using the R package "cusp":

```
               Estimate Std. Error z value Pr(>|z|)
a[(Intercept)]  1.07607    0.04898  21.967  < 2e-16 ***
a[tA]           0.17599    0.02574   6.839 8.00e-12 ***
b[(Intercept)]  2.24280    0.08206  27.332  < 2e-16 ***
b[tB]           0.21467    0.03535   6.073 1.26e-09 ***
w[(Intercept)]  1.35941    0.02117  64.199  < 2e-16 ***
w[tY]           0.79771    0.01286  62.038  < 2e-16 ***

  Null deviance:  1.2689e+03  on 1994  degrees of freedom
Linear deviance:  1.8348e+03  on 1991  degrees of freedom
Logist deviance: -9.5683e-11  on 1990  degrees of freedom
 Delay deviance:  9.7105e+02  on 1989  degrees of freedom

             R.Squared logLik npar  AIC    AICc     BIC
Linear model 0.0798  -2747.25 4   5502.51 5502.53 5524.91
Cusp model   0.3381  -2192.024 6  4396.05 4396.09 4429.64

Note: R.Squared for cusp model is Cobb's pseudo-R^2. This
value can become negative.
     Chi-square test of linear vs. cusp model
     X-squared = 1110, df = 2, p-value = 0
```

First, similar to the linear regression model presented in Section IV(B), results from SDECusp indicate that both the asymmetry variable x_1 (HIV knowledge, $a = .1760$, $p < .01$) and the bifurcation variable x_2 (perceived condom efficacy, $b = .2147$, $p < .01$) were highly significant in predicting the outcome variable z (condom use self-efficacy). Both coefficients were positive, which is consistent with the substantive theory.

However, further review of the results indicated that the estimated model coefficient a_1 for the asymmetry variable and b_1 for the bifurcation variable had less difference in the SDECusp model (0.1760 and 0.2147) than the two corresponding coefficients in the linear regression model (0.0080 and 0.2033), as shown in Section IV(B).

Most important, the estimated R^2 for the cusp modeling was 34%, and substantially greater than the 7.8% in the linear regression model, indicating the superiority of the cusp model to the linear regression model in quantifying the relationship between the two predictors and the outcome variable. A Chi-square test of linear regression model to the SDECusp model gave the Chi-squared as 1110 with $df = 2$, which yielded a p-value < 0.0001, which also indicated a better model fit.

The estimated density distribution of the data in Figure 3 illustrates the data point distribution, including the cusp region (shaded area) with the control plane (α = asymmetry, β = bifurcation). Most data points ($n = 1{,}723$) were located within the cusp region (whose condom use self-efficacy subjects to rapid change), with the remaining data points ($n = 272$) located in the upper stable region (with high condom use self-efficacy); No data points were observed in other regions.

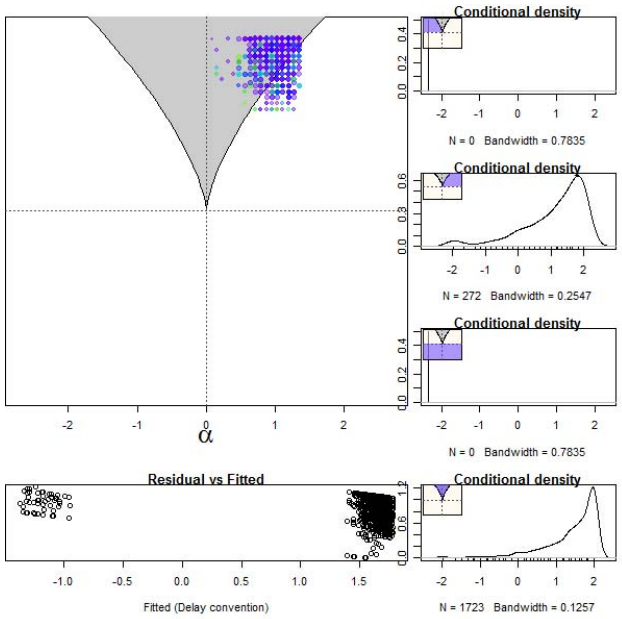

Figure 3. Cusp region from SDECusp model fitting.

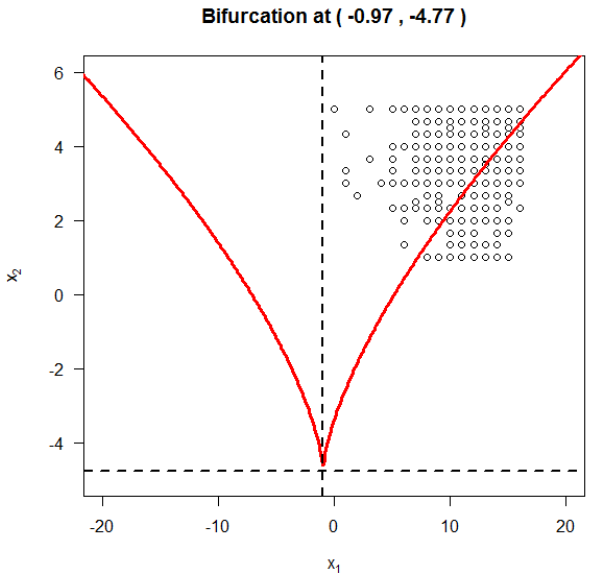

Figure 4. Estimated bifurcation point from the SDECusp modeling.

Based on the results of the SDECusp modeling, we re-constructed the cusp region to the original measurement scale of the data (1-5) (see Figure 4). In Figure 4, the two red lines represent the threshold lines for sudden changes in students' condom use self-efficacy. The bifurcation point (α, β) obtained by backsolving from the constructed cusp region was x_1 (HIV knowledge) = -0.97 and x_2 (response efficacy) = -4.77, as shown in Figure 4. However, the solved bifurcation point cannot be explained logically, since it is not possible for a high school student to suddenly believe that he or she can correctly use condoms during sex to prevent HIV transmission while also having negative HIV knowledge and negative perceptions of the effectiveness of condoms to prevent HIV infection.

D. StatCusp

When applying our StatCusp model to the same data used for the SDECusp modeling (with z, x_1 and x_2 also standardized), the estimated residual variance is 0.982, indicating an excellent model fit to the data. For the asymmetry variable, the two parameters estimated were $a_0 = -0.083$, $a_1 = 0.094$ ($p < .01$ for both); for the bifurcation variables, the two parameters estimated were $b_0 = 1.568$; $b_1 = 0.672$ ($p < .01$).

In the StatCusp model, the two predictor variables, x_1 (HIV knowledge) and x_2 (response efficacy), significantly and positively predicted the outcome variable z (condom use self-efficacy): this prediction is consistent with both linear regression modeling and SDECusp modeling. Furthermore, as shown numerically, the differences in the estimated model coefficients between the two predictor variables in StatCusp are more similar to the two estimated model parameters from the linear regression than the from the SDECusp.

The most exciting finding from the StatCusp model is the estimate of the cusp point (see Figure 5). The solid red lines in Figure 5 indicate the bounds of the cusp region. However, the estimated bifurcation point (α, β) is located at (x_1 = HIV knowledge = 14.55, and x_2 = response efficacy = 2.33), which is scientifically reasonable. This result suggests the following:

1. When the response efficacy (perceived effectiveness of condom use for HIV prevention) is below 2.33 (slightly smaller than the average of 2.5), the relationship between HIV knowledge and condom use self-efficacy is continuous and gradual; students with greater HIV knowledge are more likely to believe that they can correctly use condoms during sex to prevent transmission of HIV.
2. When the response efficacy is above average, increases in HIV knowledge might result in totally opposite results: (a) a sudden jump in confidence in their ability to correctly use condoms, or (b) remain unconfident in their ability to correctly use condoms, with the trigger point for the confidence jumps determined by the red threshold lines. For example, when the perceived efficacy of condoms to prevent HIV infection increased from 2.3 (*somewhat ineffective*) to 4.0 (*somewhat effective*) or 5.0 (*very effective*), students with limited HIV knowledge (i.e., knowledge scores less than 5) might trigger a sudden jump among students with low levels of confidence to become confident about condom use during sex.

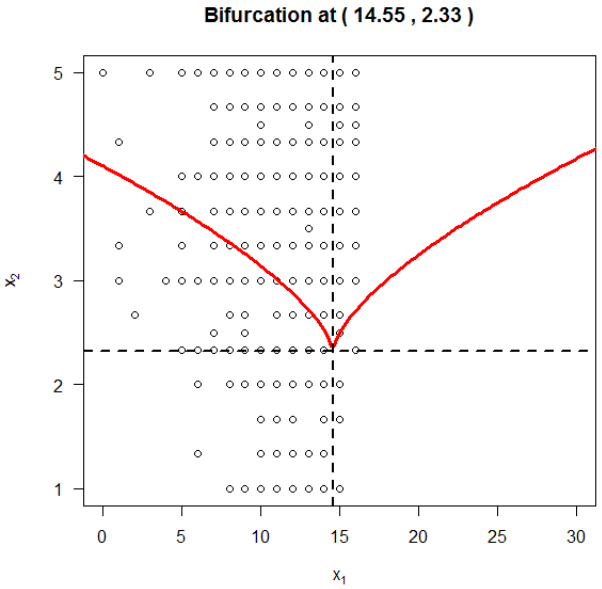

Figure 5. Cusp region from StatCusp model.

V. Conclusions and Discussions

A. Summary

Cusp catastrophe has a strong theoretical base and the promising results from the application of this theory in research make the cusp catastrophe model an extremely useful tool for resarchers interested in broadening the horizons of their data analysis [16]. Decades of past research that sought statistical solutions for the deterministic Cusp Catastrophe Model provide useful experience and insight. Such experience includes the development of the polynomial cusp regression approach that mimics a polinomial regression, as discussed in Guastello [6, 7, 21]. In addition, prior research contributed the stochastic cusp model that capitalizes on the power of likelihood function in statistics, as discussed in Cobb [3,4,5] and Grassman [8]. However, these methods have been challenged by theoretical analysis and practical application [14, 23, 24].

This paper has illustrated the novel development of the statistical Cusp Catastrophe Model (i.e., StatCusp), demonstrated the unbiasedness of the parameter estimation, and demonstrated estimation efficiency through simulation studies as illustrated in Tables 1 and 2 as well as Figure 2. Moreover, we have described an innovative use of a bootstrapping procedure to resolve the estimation of the variance, recognizing the limitations of the Fisher information matrix in estimating variance of model residuals. Last, we applied our StatCusp method to real-world data and compared the results to those obtained with other methods. This comparison clearly demonstrated the advantages of our StatCusp approach in obtaining model parameters that are meaningful.

B. Primary conclusions

Through this explorative study, we conclude that this research has advanced the field and previously published methods in several respects:

1. Theoretical strengths: We have established the StatCusp method based on a nontraditional, nonlinear regression that capitalizes on maximum likelihood estimation. This novel method is the first of such endeavors with great potential to open a new approach to solving the Cusp Catastrophe Model and to promote the application of the StatCusp method in practical research,
2. Supported with simulation studies: The validity of our novel StatCusp is supported with extensive analysis of simulated data. Further, this simulation analysis demonstrated two key points: (a) the parameter estimation is unbiased, and (b) the variance estimation using a bootstrapping method is efficient.
3. Supported with real-world data: By using real study data to illustrate StatCusp, we demonstrated that our method provides reasonable estimates of the key cusp model parameters, particularly the bifurcation points.

Correct estimation of the bifurcation point is essential to determining the threshold lines that are critical in health and behavioral research. In previous studies, very few, if any, researchers have used empirical data to investigate the cusp region. Despite claims that the cusp region can be estimated using the SDECusp approach, our application of SDECusp with empirical data yielded results that could not be explained by data. However, when our StatCusp method is applied to empirical data, the results are reasonable.

C. Future Research

Although our studies to date have produced highly encouraging findings, more research is needed to fully establish the StatCusp method. Theoretically, we need to further investigate the statistical properties of the StatCusp model that we have proposed. Methodologically, we need to investigate why the Fisher information matrix does not provide a good estimate of the variance. Computationally, we need to establish protocols to derive a set of data-driven initial values for the estimation outlined in equation (14) or (15). These protocols are needed because although our novel StatCusp model works satisfactorily in model fitting, the model is sensitive to the initial values. Although this issue is universal in all nonlinear modeling methods, we need to establish an objective approach and a guideline to solve the sensitivity issue. We are currently working on a procedure with data-driven initial values, which will complete the StatCusp model development.

Practically, we need to test the StatCusp method with more data, including the standardized data such as cusp machine data and data from experimental research, including data used in previous research conducted by others as well as our own research.

Last, work is also needed to expand our StatCusp model to the analysis of different data types, including binary data, count data, and longitudinal data. Such expansion of our model holds promise for health and behavioral research.

ACKNOWLEDGMENT

This research was support in part by National Institute of Health (NIH) Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD, R01HD075635, PIs: Chen X and Chen D).

REFERENCES

[1] Thom, R. (1975). *Structural stability and morphogenesis*. New York, NY: Benjamin-Addison-Wesley.

[2] Thom, R., & Fowler, D. H. (1975). *Structural stability and morphogenesis: An outline of a general theory of models*. New York, NY: W. A. Benjamin. .

[3] Cobb, L., & Ragade, R. K. (1978). Applications of catastrophe theory in the behavioral and life sciences. *Behavioral Science, 23*, 291-419. doi:10.1002/bs.3830230511

[4] Cobb, L., & Watson, B. (1980). Statistical catastrophe theory: An overview. *Mathematical Modelling, 1*(4), 311-317. doi:10.1016/0270-0255(80)90041-X

[5] Cobb, L., & Zacks, S. (1985). Applications of catastrophe theory for statistical modeling in the biosciences. *Journal of the American Statistical Association, 80*(392), 793-802. doi:10.1080/01621459.1985.10478184

[6] Guastello, S. J. (1982). Moderator regression and the cusp catastrophe: Application of two-stage personnel selection, training, therapy and

program evaluation. *Behavioral Science, 27*, 259-272. doi:10.1002/bs.3830270305

[7] Guastello, S. J. (1989). Catastrophe modeling of the accident processes: Evaluation of an accident reduction program using the occupational hazards survey. *Accident Analysis and Prevention, 21*(17-28). doi:10.1016/0001-4575(89)90049-3

[8] Grasman, R. P., van der Mass, H. L., & Wagenmakers, E. (2009). Fitting the cusp catastrophe in R: A cusp package primer. *Journal of Statistical Software, 32*(8), 1-27. doi:10.18637/jss.v032.i08

[9] Clair, S. (1998). A Cusp Catastrophe Model for adolescent alcohol use: An empirical test. *Nonlinear Dynamics, Psychology, and Life Sciences, 2*(3), 217-241. doi:10.1023/A:1022376002167

[10] Mazanov, J., & Byrne, D. G. (2006). A Cusp Catastrophe Model analysis of changes in adolescent substance use: Assessment of behavioural intention as a bifurcation variable. *Nonlinear Dynamics, Psychology, and Life Sciences, 10*(4), 445-470.

[11] Guastello, S. J., Aruka, Y., Doyle, M., & Smerz, K. E. (2008). Cross-cultural generalizability of a Cusp Catastrophe Model for binge drinking among college students. N*onlinear Dynamics, Psychology, and Life Sciences, 12*(4), 397-407.

[12] Chen, X., Lunn, S., Harris, C., Li, X., Deveaux, L., Marshall, S., … Stanton B. (2010). Modeling early sexual initiation among young adolescents using quantum and continuous behavior change methods: Implications for HIV prevention. *Nonlinear Dynamics, Psychology, and Life Sciences, 14*(4), 491-509.

[13] Wagner, C. M. (2010). Predicting nursing turnover with catastrophe theory. *Journal of Advanced Nursing 66*(9), 2071-2084. doi:10.1111/j.1365-2648.2010.05388.x

[14] Chen, X., Lunn, S., Deveaus, L., Li, X., Brathwaite, N., Cottrell, L., & Stanton, B. (2008). A cluster randomized controlled trial of an adolescent HIV prevention program among Bahamian youth: Effect at 12 months post-intervention. *AIDS and Behavior, 13*, 495-508

[15] Faraway, J.J. (2009). *Linear model with R*. Abingdon Oxfordshire, UK: Taylor & Francis.

[16] Chen, X., & Chen, D. (2015). Cusp catastrophe modeling in medical and health research. In D. Chen & J. Wilson (Eds.), *Innovative statistical methods for public health data* (pp. 265-290). ICSA book series in statistics. New York, NY: Springer. doi: 10.1007/978-3-319-18536-1

[17] Chen, D. G., & Peace, K. E. (2011). Clinical *Trial Data Analysis Using R.* Boca Raton, FL: Chapman and Hall/CRC.

[18] Saunders, P. T. (1980). *An introduction to catastrophe theory*. Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9781139171533

[19] Hartelman, A. I. (1997). *Stochastic catastrophe theory*. Amsterdam, The Netherlands: University of Amsterdam.

[20] Iacus, S. M. (2008). Simulation and Inference for Stochastic Differential Equations with R Examples. New York, NY:Springer. doi:10.1007/978-0-387-75839-8

[21] Guastello, S. J., & Gregson, A. M. (2011). Nonlinear dynamic systems analysis for the behavioral sciences using real data. Boca Raton, FL: Chapman and Hall/CRC Press.

[22] Gong, J., Stanton, B., Lunn, S., Devearus, L., Li, X., Marshall, S., … Chen, X. (2009). Effects through 24 months of an HIV/AIDS prevention intervention program based on protection motivation theory among preadolescents in the Bahamas. *Pediatrics, 123*, 917-928. doi:10.1542/peds.2008-2363

[23] Chen, X., Stanton, S., Chen, D.G., & Li, X. (2013). Is intention to use condom a linear process? Cusp modeling and evaluation of an hiv prevention intervention trial. *Nonlinear Dynamics, Psychology, and Life Sciences, 17*(3), 385 -403.

[24] Zeeman, E.C. (1976, April) Catastrophe theory. Things that change suddenly, by fits and starts, have long resisted mathematical and analysis. A method derived from topology describes these phenomena as examples of seven "elementary catastrophes." *Scientific American.* (April), 65-70; 75-83.

[25] Mazanov, J., & Byrne, D. G. (2008). Modelling change in adolescent smoking behaviour: Stability of predictors across analytic models. *British Journal of Health Psychology, 13*, 361-379. doi:10.1348/135910707X202490

[26] West, R., & Sohal, T. (2006). "Catastrophic" pathways to smoking cessation: Findings from national survey. *British Medical Journal, 332*, 458-460. doi:10.1136/bmj.38723.573866.AE

[27] Chen, D. G., Lin, F., Chen, X., Tang, W., & Kitzman, H. (2014). Cusp Catastrophe Model: A nonlinear model for health outcomes research. *Nursing Research, 63*(3), 211-220. doi:10.1097/NNR.0000000000000034

[28] Gilmore, R. (1993). *Catastrophe theory for scientists and engineers.* New York, NY: Dover.

[29] Sussmann, H. J., & Zahler, R. S. (1978). Catastrophe theory as applied to the social and biological sciences: A critique. *Synthese, 37,* 117-216. doi:10.1007/BF00869575

[30] Rosser, J. B., Jr. (2007). The rise and fall of catastrophe theory applications in economics: Was the baby thrown out with the bathwater? *Journal of Economic Dynamics and Control, 31,* 3255-3280. doi:10.1016/j.jedc.2006.09.013

[31] Alexander, R.A., Herbert, G. R., DeShon, R. P., & Ranges, P. J. (1992). An examination of least-squares regression modeling of catastrophe theory. *Psychological Bulletin 111*(2), 366-374. doi:10.1037/0033-2909.111.2.366

[32] Honerkamp, J. (1994). *Stochastic dynamical system: Concepts, numerical methods, data analysis*. New York, NY: VCH Publishers.

[33] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19,* 716–723. doi:10.1109/TAC.1974.1100705

[34] Gelfand, A. E., & Dey, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society. Series B (Methodological) 56*, 501-514. Retrieved from http://www.jstor.org/stable/2347123

[35] Cobb, L. (1998). *An introduction to cusp surface analysis*. [Technical report]. Louisville, CO: Aetheling Consultants. Retrieved from http://www.aetheling.com/modes/cusp/Intro.htm

[36] Hartelman, P. A. I., van der Maas, H. L. J., & Molenaar, P. C. M. (1998). Detecting and modelling developmental transitions. *British Journal of Developmental Psychology, 16,* 97-122. doi:10.1111/j.2044-835X.1998.tb00751.x

[37] van der Maas, H. L. J., Kolstein, R., & van der Pligt, J. (2003). Sudden Transitions in attitudes. *Sociological Methods Research, 32,* 125-152. doi:10.1177/0049124103253773