A regime-switching longitudinal model of alcohol lapse-relapse
Chow, S.-M.; Witkiewitz, K.; Grasman, R.P.P.P.; Hutton, R.S.; Maisto, S.A.

Published in:
Handbook of developmental systems theory & methodology

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
A Regime-Switching Longitudinal Model of Alcohol Lapse-Relapse

Sy-Miin Chow
University of North Carolina at Chapel Hill

Katie Witkiewitz
University of New Mexico

Raoul Grasman
University of Amsterdam

Stephen A. Maisto
Syracuse University

Abstract

Contemporary general linear models assume that continuous changes in the predictor variables result in proportionate amounts of (linear) change in the outcome variable. Empirical evidence from the alcohol treatment literature, however, favors the application of nonlinear dynamic models over the general linear model in their ability to capture sudden, discontinuous jumps in individuals’ drinking tendency. One example of such nonlinear models is the cusp catastrophe model, used by Witkiewitz and Marlatt (2004) to represent the complex interplay between different risk factors in triggering sudden shifts in individuals’ tendency to drink. Although the cusp catastrophe model has been promising in capturing some aspects of alcohol use dynamics, current approaches of fitting variations of this model do not address several practical data analytic problems commonly seen in empirical data, including the presence of incomplete data, measurement and/or process noise, the lagged effects of previous drinking on current alcohol use, heterogeneous timing of lapse-relapse within and across subjects, and the large number of abstainers at any given time—commonly referred to as the “zero-inflation” phenomenon. We propose a mixture structural equation model with regime-switching as an alternative approach to account for these data analytic issues while retaining some of the key features of the cusp catastrophe model. The proposed model is illustrated using longitudinal drinking data from the COMBINE study (COMBINE Study Research Group, 2003).
The Cusp Catastrophe Model

Over the last few decades, some researchers’ initial enthusiasm toward nonlinear dynamical systems has gone through various periods of wax and wane. Since the pioneering work of Poincaré (1892) and Lorenz’s (1963) computer simulation of a “chaotic” weather system (for a introduction see Alligood, Sauer, & Yorke, 1996; Gleick, 1987; Hilborn, 1994), some of the properties of nonlinear dynamical systems have continued to intrigue social and behavioral scientists. In particular, the idea that a small difference in initial conditions could lead to drastically discrepant outcomes as time progresses, or the existence of models, broadly classified as catastrophe models, that can capture discontinuous jumps in various systems’ change trajectories, have great conceptual appeal to social scientists (see e.g., Vallacher & Nowak, 1994).

Catastrophe theory (Thom, 1975, 1993) is the study of the many ways in which continuous changes in a system’s parameters can result in discontinuous changes in an outcome variable of interest. Various catastrophe models have been used to represent sudden jumps manifested, for instance, in individuals’ driving speed (Cobb, 1981; Poston & Stewart, 1978; Zeeman, 1976), attitude (Wimmers, Savelsbergh, & Kamp, 1998), affective states (Allen & Carifo, 1995; Strahan & Conger, 1999), alcohol use (Clair, 1999; Witkiewitz & Marlatt, 2004), paranormal beliefs (Lange & Houran, 2000) and human development in discrete stages (Byrne, Mazanov, & Gregson, 2001; Preece, 1980; Molenaar & Nesselroade, 2000, in press; Van der Maas & Molenaar, 1992). Among the seven possible catastrophe models considered by Thom, the cusp catastrophe is one of the simplest and most commonly applied forms in the social sciences. The model posits that changes in the values of an outcome variable (often termed the behavioral parameter) depend on two control parameters: a splitting parameter and a normal parameter (Gilmore, 1981; Stewart & Peregoy, 1983). In practice, these parameters are often expressed as linear, fixed functions of one or more covariates (e.g., Cobb & Zacks, 1985; Grasman, Van der Maas, & Wagenmakers, 2009; Guastello, 1984; Van der Maas & Molenaar, 1992). Thus, they are sometimes referred to as the splitting and normal variables.

Several approaches have been used to fit various catastrophe models. Cobb (Cobb, 1981; Cobb, Koppestein, & Chen, 1983) formulated the cusp catastrophe model as a stochastic partial differential equation model and used a likelihood approach to estimate the parameters. As an alternative to Cobb’s method, Guastello (1982, 1992) presented a simpler approach of using polynomial regression to estimate the parameters in a catastrophe model. This estimation technique has been applied by researchers such as Hufford, Witkiewitz, Shields, Kodva, and Caruso (2003) in modeling the relapse process of alcohol use. Alexander, Herbert, DeShon, and Hanges (1992) pointed out some problems associated with Guastello’s approach and suggested the use of Olivas (1987) multivariate confirmatory

Funding for this study was provided by a grant from NSF (BCS-0826844). We are grateful to Peter Molenaar and Han van der Maas, whose work on catastrophe systems has inspired us to undertake the modeling work in the present chapter. We are also grateful to the comments of R. Shane Hutton in the early stages of model building and to the help of Zhaowei Hua and Ruixin Guo in organizing data from the COMBINE study. Correspondence concerning this article can be addressed to Sy-Miin Chow, University of North Carolina, CB#3270 Davie Hall, Chapel Hill, NC 27599-3270 or by email to symiin@email.unc.edu.
approach as an alternative. In particular, Guastello (1982) and Brown (1995), in a different approach, both used empirical difference scores as the dependent variable in estimating the parameters for a cusp catastrophe model. Brown (1995) used nonlinear least squares together with a Runge-Kutta integration procedure to obtain iterative estimates of the parameters in a model, whereas Guastello (1982) reformulated the cusp catastrophe model into a polynomial regression model and used ordinary least squares to estimate the corresponding parameters.

For the purpose of this chapter, suffice it to say that difference score has a long history of controversy in psychology. In some cases the use of difference score is totally justifiable (e.g., Nesselroade & Cable, 1974; Nesselroade & Bartsch, 1977; Hummel-Rossi & Weinberg, 1975; Maxwell & Howard, 1981), but in others, not necessarily so (Bereiter, 1963; Cronbach & Furby, 1970; Harris, 1963; Horn, 1963; Lord, 1956). Methodological advances in the past few decades have enabled researchers to formulate difference score-related approaches within a latent variable framework, which helps strengthen the appeal of such approaches when used with noisy data (e.g., McArdle & Hamagami, 2001; Molenaar & Newell, 2003). This feature has not be incorporated into the modeling approaches summarized above. One other issue pertaining to these approaches is that transformation functions are typically used to reparameterize the cusp catastrophe model into a simpler form (e.g., polynomial regression model) to facilitate model fitting. However, when the transformation functions are not explicitly specified as constraints in the model specification, there is no guarantee that the fitted, simpler model corresponds in anyway to the original model (for a discussion of this issue in the context of stochastic differential equation modeling see Singer, 1992, 1993).

More importantly, while some of the features of catastrophe models have proven useful across a wide range of applications, they impose several assumptions that may be overly restrictive in empirical applications. The goal of this study is to present a mixture structural equation model with regime-switching as an alternative approach to account for some of the known data analytic issues associated with the cusp catastrophe model. While the new proposed model only retains some of the features of the original model thought to be particularly pertinent to the modeling of alcohol use dynamics, it provides more flexibility than the cusp catastrophe model in a number of ways. Some of these advantages include: (1) it is suited for use with longitudinal panel data with multiple subjects, (2) missing data can be readily accommodated, (3) model fitting can be performed at the latent variable level if needed, (4) the new model allows researchers to test multivariate extensions involving lagged effects, (5) predictions encompass a greater range of drinking behaviors, including light and moderate drinking levels, as opposed to focusing on the two extreme modes of alcohol use, namely, complete abstinence and heavy drinking, (6) the new approach helps capture heterogeneous timing of relapse within and across subjects and (7) the presence of a large number of abstainers at any given time point—referred to as the zero-inflation phenomenon in statistics—is represented more flexibly using mixture modeling techniques. The proposed model is illustrated using a set of longitudinal drinking data from the COMBINE study (Anton et al., 2006).
Mathematical Background of the Cusp Catastrophe System

Consider a dynamical system whose changes over time are governed by the following deterministic equation:

\[
\frac{\partial y}{\partial t} = -\frac{\partial V(y; \theta)}{\partial y},
\]

where \(y\) is the dependent variable of interest, namely, the behavioral variable in the catastrophe literature; \(-\frac{\partial V(y; \theta)}{\partial y}\) is a potential function that dictates the behavioral dynamics of the system and \(\theta\) is a vector of control parameters. One of the simplest forms of catastrophe is the cusp catastrophe, described by the potential function

\[
-\frac{\partial V(y; \theta)}{\partial y} = \frac{1}{2} \beta y^2 + \alpha y - \frac{1}{4} y^4,
\]

where \(\theta = [\alpha \, \beta]'\) consists of the two control parameters in the system. For reasons that will become clear in a moment, the parameter \(\alpha\) is referred to as the normal variable or asymmetry variable, whereas \(\beta\) is often referred to as the splitting or bifurcation variable (Gilmore, 1981; Grasman et al., 2009; Poston & Stewart, 1978). The equilibrium points of the cusp catastrophe systems are values of \(y\) for which \(\frac{\partial y}{\partial t} = 0\). That is, they are the \(y\) values the system eventually settle into over time. These points are functions of the control parameters, \(\alpha\) and \(\beta\), and they can be obtained by differentiating Equation 2 with respect to \(y\) and setting it to zero as

\[
\beta y + \alpha - y^3 = 0
\]

to obtain the roots or equilibrium points that arise under different values of \(\alpha\) and \(\beta\). It is well known that (see e.g., Poston & Stewart, 1978) the number and nature of the equilibrium points that arise under different values of \(\alpha\) and \(\beta\) can be deduced from the Cardan’s discriminant function, expressed as

\[
D = 4\beta^3 + 27\alpha^2.
\]

The value of \(D\) helps classify the behaviors of the system under different \(\alpha\)-\(\beta\) configurations into 4 broad categories or sets, which are plotted in Figure 1. Specifically, the shaded region labeled as “A” corresponds to values of \(D\) that are < 0 and there are three equilibrium points when \(\alpha\) and \(\beta\) assume values in this region. The four quadrants labeled as “B” outside of the shaded region all correspond to \(D > 0\) and each quadrant is characterized by only one stable equilibrium point. A stable equilibrium is defined as a point toward which a system gravitates and settles into in the long run even when small perturbations are imposed to move the system away from the equilibrium point. The exact value of the equilibrium point differs by quadrant, as is indicated approximately by the modes of the first three conditional densities plotted in the right panel of Figure 1. The two lower quadrants (with \(\beta < 0\)) have the same equilibrium point at \(y = 0\). The third set simply consists of the origin at \(\alpha = 0\) and \(\beta = 0\). At this point, there is only one distinct equilibrium point at \(y = 0\).

The two branches or borders of the shaded region, marked as \(C_1\) and \(C_2\) in Figure 1, constitute the last set of values, for which \(D = 0\). On each of these branches, the system
is characterized by one unstable equilibrium, namely, an inflection point, and one stable equilibrium (indicated by one of the two peaks of the last conditional density in Figure 1). The location of the stable equilibrium differs depending on whether the system shows changes in \( \alpha \) and \( \beta \) values that cross the branches (\( C_1 \) or \( C_2 \)) from left to right, or from right to left. When either one of the two branches is crossed, the original stable minimum is annihilated and a new stable minimum arises. The asymmetry in the values of the equilibria when the branches are crossed from left to right versus from right to left constitutes one of the key catastrophe flags known as \textit{hysterisis}, which we will elaborate further below.

It is often more intuitive to understand the behavioral dynamics of a catastrophe system geometrically in the catastrophe manifold\(^1\), which shows all the possible equilibrium points at different combinations of \( \alpha \) and \( \beta \) values. Witkiewitz and colleagues (Witkiewitz & Marlatt, 2004; Witkiewitz, Van der Maas, Hufford, & Marlatt, 2007; Witkiewitz & Marlatt, 2007) have introduced the cusp catastrophe model as a useful representation of alcohol lapse-relapse dynamics. Their specific example provides a helpful introduction to the various concepts and key constituent variables of the cusp catastrophe model and we adopt the terminologies in their application to illustrate features of the cusp catastrophe model.

\textit{Catastrophe in Alcohol Use Dynamics}

According to the framework presented by Witkiewitz and Marlatt (2004), distal and phasic processes serve as the splitting and normal parameters in the cusp catastrophe model, respectively, to affect drinking outcomes (i.e., the behavioral parameter). Distal risks consist of background variables (e.g., alcohol dependence) and relatively stable pre-treatment characteristics (e.g., alcohol expectancies). Proximal risks, in contrast, include transient precipitants (e.g., stressful situations, negative affect) that change an individual’s tendency to drink on a moment-to-moment basis. The resultant catastrophe manifold depicting the equilibrium values of drinking (represented as the \( z \)-axis) at different values of distal risks (the splitting parameter, shown as the \( y \)-axis) and phasic risks (the normal parameter, shown as the \( x \)-axis) is portrayed in Figure 2. The bottom surface of the model, called the control plane, is defined by the range of possible values for the two control parameters. The top surface, called the behavior plane, is defined as the range of predicted values of the outcome variable (i.e., drinking) at different values of the two control parameters. The folded part of the behavior plane represents the discontinuity that is characteristic of the model, whereby the same values of the control parameters can result in two widely differing values of the behavior variable. While in this region, intermediate outcome values are unlikely to occur or are not sustainable (i.e., they do not constitute a stable equilibrium). This property renders these intermediate outcome values to be called the inaccessible region (Gilmore, 1981; Zeeman, 1978).

Putting the cusp catastrophe model in the context of alcohol use, Witkiewitz and colleagues (Witkiewitz & Marlatt, 2004; Witkiewitz et al., 2007; Witkiewitz & Marlatt, 2007) stipulated that for individuals with low distal risk, drinking changes linearly with changes in phasic risk (path A). However, for those with high distal risk, a small increase in phasic risk (e.g., an argument with a significant other) may push an individual over the

\(^1\)A manifold is simply a higher dimensional analogue of a smooth curve or surface (Poston & Stewart, 1978).
Figure 1. A plot summarizing the the broad categories of behaviors in the cusp catastrophe system as a function of different combinations of the normal (α) and splitting parameters (β). The plot was created using the cusp package in R written by Grasman et al. (2009). The points located on the α-β plane are simulated data generated using the cusp catastrophe model. The conditional densities located in the right panel show the densities of the data points located in different quadrants of the α-β plane.
edge, leading to an episode of lapse (path B). A prolapse (the return from heavy drinking to abstinence; path C) then requires a substantially larger reduction in phasic risk to help the individual regain abstinence. These unique characteristics of the cusp catastrophe model are just some of the flags associated with the cusp catastrophe model. The seven catastrophe flags associated with this model include multimodality, sudden jumps, inaccessibility, hysteresis, anomalous variance, divergence of linear response, and critical slowing down. Each of these flags will be described briefly here.

**Figure 2.** The cusp catastrophe model adopted by Witkiewitz and colleagues to represent the multiple pathways of AUD clinical course. The circles represent possible predictions of this model for individuals with high distal risk, which put an individual into one of two possible modes of behavior. The plot was created using the cusp package in R written by Grasman et al. (2009).

**Multimodality** is indicated by the presence of more than one mode of outcome variable based on the same values of the control parameters (or in the context of alcohol use, distal and phasic risks). The multiple modes arise because of the presence of multiple equilibria points in the system. Thus, continuous changes in the distal risk (the y variable in Figure 2) can yield a sudden, qualitative shift from a single mode of outcome (when an individual is located near path A of the cusp landscape) to the coexistence of two modes of outcome (when an individual is located near the fold). In this case, bifurcation is said to have occurred.
Thus distal risk (or the splitting parameter) is also termed the bifurcation variable. The range of distal risk values (i.e., values along the y-axis) at a specific value of proximal risk $\alpha$ under which bifurcation into two modes of outcome is observed constitutes the bifurcation set.

Individuals may show sudden jumps or abrupt transitions from one mode of behavior to another mode of behavior (i.e., from one equilibrium point to another). Such sudden jumps between the two behavioral modes render any intermediate value between the two modes unlikely to occur, or in other words, inaccessible. The combination of parameter values that gives rise to such behavior constitutes a region of indeterminacy. That is, even if values of the control parameters are known, it is impossible to determine which of the two possible modes the outcome variable would manifest unless previous history of the outcome variable is known.

Hysteresis refers to the asymmetry in the values of the normal parameter (i.e., phasic risk) that trigger a sudden jump in the direction of a prolapce (namely, a transition from heavy drinking to abstinence, as indicated by path C) as opposed to a lapse (namely, a transition from abstinence to heavy drinking, denoted as path B in Figure 2). That is, the points at which the system shows a sudden transition in behavioral mode depend on the previous values of the system and the direction of change. Thus, while a particular value of phasic risk (e.g., stress) is enough to cause an individual to “fall off the wagon” and transition from abstinence to drinking, a much lower level of phasic risk is needed before the same individual can return from drinking to abstinence (i.e., to be back “on the wagon”).

Divergence refers to the characteristic that with increase in the splitting parameter (i.e., distal risk), greater bimodality is evident in the outcome variable. In the context of alcohol use, this means that individuals who have higher levels of distal risk also tend to manifest more extreme transitions between heavy drinking and abstinence, as opposed to the gradual changes in drinking level characterizing those with lower distal risk (see e.g., path A in Figure 2). Closely associated with this flag is the anomalous variance flag, namely, the increase in variance in the outcome variable near the points of sudden transition.

Divergence of linear response and critical slowing down are two catastrophe flags that emerge when a system is perturbed, or in other words, when slight changes are induced in the control parameters, either mathematically or experimentally. Divergence of linear response occurs when perturbations of a system lead to larger fluctuations near the transition points whereas critical slowing down corresponds to the feature that longer time is needed for the system to return to equilibrium if the system is perturbed near the transition points. Such manipulations are difficult to implement in applications in the social and behavioral sciences, and these flags have thus not been discussed as often as the other flags (see e.g., Zeeman, 1978).

Mixture Structural Equation Model with Regime–Switching (MSEM-RS)

Structural Equation Modeling (Jöreskog, 1973) is a statistical technique for representing multivariate relationships among observed and latent variables. Composed of a measurement model that relates the observed indicators to a set of underlying latent variables, and a structural model that specifies the interrelationships among the latent variables,
SEM models and related latent variable models have been used extensively in the social and behavioral sciences. When systematic heterogeneities exist across subpopulations but the grouping variable that distinguishes the subpopulations is unobserved (i.e., latent), mixture SEMs (Arminger & Muthén, 1998; Dolan & Van der Maas, 1998; Jedidi, Jagpal, & DeSarbo, 1997; Vermunt & Magidson, 2005; Yung, 1997) may be employed.

MSEM is an extension of SEM that allows for heterogeneities in the means and covariance structures of a SEM model conditional on a series of latent, unobserved groups, $C_i$, ($i = 1, \ldots, n$), typically referred to as latent classes. More recently, SEM statistical packages such as Mplus also allow users to fit longitudinal extensions of MSEM and estimate how individuals transition among different latent classes over time (Kaplan, 2008; B. O. Muthén & Asparouhov, 2011; Nylund-Gibson, Muthen, Nishina, Bellmore, & Graham, under review). The use of different latent classes over time to represent the different phases of a process is consistent with the notion of regime-switching in the time series and econometric literature (Hamilton, 1994; Kim & Nelson, 1999). Here, we use the terms latent class and regime interchangeable. The resultant model is referred to herein as mixture structural equation model with regime switching (MSEM-RS).

In a MSEM-RS, the measurement model for an individual $i$ in latent class $k$, namely, $C_i = k$, can be expressed as

$$y_i|C_i = k = \nu_{k} + \Lambda_k \eta_i + \epsilon_i,$$

where $y_i$ is a $p \times 1$ vector of continuous observed variables for individual $i$, $\nu_k$ is a $p \times 1$ vector of intercepts, $\Lambda_k$ is a $p \times w$ matrix of factor loadings, $\eta_i$ is a $w \times 1$ vector of latent variables and $\epsilon_i$ is a $p \times 1$ vector of measurement errors. The subscript $k$ highlights the class-specific nature of the parameters. In cases involving ordinal observed variables, the corresponding elements in $y_i$ are unobserved. The $l$th unobserved continuous variable, $y_{ij}$, is then linked to the corresponding manifest ordinal response, $y_{il}^s$, as

$$y_{il}^s = s \iff \tau_{i,s-1} < y_{il} \leq \tau_{i,s}, s = 1, \ldots, S,$$

where $\tau_{i,s}$ is a set of threshold values for variable $l$ that is held invariant across individuals.

Conditional on class $k$, a structural model serves to specify the relationships among a vector of latent variables, $\eta_i$, with

$$\eta_i|C_i = k = \alpha_k + \beta_k \eta_i + \zeta_i,$$

where $\alpha_k$ is a vector of regime-dependent intercepts, $\beta_k$ is a $w \times w$ matrix of regime-dependent regression effects among the latent variables and $\zeta_i$ is a $w \times 1$ vector of disturbances. When multiple time points of the same variables are present, $y_i$ and $\eta_i$ are expanded to include $y_i = [y_{i1}^T \ldots y_{iT}^T]$ and $\eta_i = [\eta_{i1}^T \ldots \eta_{iT}^T]$. Other corresponding modeling components are also expanded accordingly to accommodate the multiple time points.

Multinomial logistic regression models are typically used to represent the initial class (or regime) probabilities as

$$\Pr(C_{i1} = h|x_{i1}) = \pi_{k,i1} = \frac{\exp(a_{k1} + c_{h1}^* x_{i1})}{\sum_{s=1}^{S} \exp(c_{s1} x_{i1})},$$

where $a_{h1}$ is the logit intercept for class $h$ at time 1; $x_{i1}$ is a vector of covariates used to predict initial class membership, and $c_{h1}$ is the associated vector of logit slopes. For
identification purposes, one of the classes has to be designated to be the reference class and the logit intercept and slopes for this class, $a_{K_1}$ and $c_{K_1}$, have to be set to zeroes.

A multinomial logistic regression model is, again, used to describe each individual’s transition in class membership from time $t-1$ to time $t$. In the general MSEM-RS framework, the transition in class membership can depend on all previous class membership information (Asparouhov & Muthén, 2011). We adopt a simpler specification and only allow this transition to depend on the class membership at time $t-1$ and the initial class membership at time 1. The corresponding multinomial logistic regression model is written as

$$
\Pr(C_{it} = k | C_{i,t-1} = j, C_{i1} = h, x_{it}) \triangleq \pi_{jk,h,it} = \frac{\exp(a_{h,kt} + b_{h,kt} x_{it})}{\sum_{s=1}^{K_t} \exp(a_{h,st} + b_{h,st} x_{it})}
$$

(9)

where $\pi_{jk,h,it}$ is individual $i$’s transition probability of moving from class $j$ at time $t-1$ to class $k$ at time $t$ conditional on initial class membership in class $h$, $K_t$ denotes the number of classes at time $t$, highlighting the possibility that latent classes may emerge or diminish over time. $a_{h,kt}$ is the logit intercept for the $k$th class at time $t$ given that individual $i$ is in initial class $h$ at time 1, $x_{it}$ is a vector of fixed covariates for predicting the transition probabilities, with an associated vector of logit slopes, $b_{h,kt}$. Included in $x_{it}$ are $K_t - 1$ binary constants reflecting the deviation in logodds of being in latent class $j$ and $k$ at time $t-1$ and time $t$ (i.e., $C_{i,t-1} = j$, $C_{it} = k$), respectively, relative to being in $C_{i,t-1} = j$ and $C_{it} = K_t$, given initial class membership in class $h$. Note that no binary constants are included in $x_{i1}$ in Equation 8 because at time 1, there are no transition probabilities to be predicted, only class probabilities.

Similar to the need to select a reference class at time 1 for identification purposes, a reference class also has to be selected for each time point and all the logit-related parameters for the reference class (including $a_{h,K_t}$ and $b_{h,K_t}$) have to be set to 0 for each $t$. This has the effect of setting each row of person $i$’s $K_t - 1 \times K_t$ conditional transition probability matrix at time $t$ for each initial latent class to sum to 1. For instance, if two regimes exist at time $t-1$ as well as at time $t$,

$$
\pi_{h,it} = \begin{bmatrix}
\pi_{11,h,it} & \pi_{12,h,it} \\
\pi_{21,h,it} & \pi_{22,h,it}
\end{bmatrix},
$$

(10)

where the element $\pi_{jk,h,it}$, as obtained using Equation 9, represents the probability of transitioning from regime $j$ at time $t-1$ to regime $k$ at time $t$ for person $i$ conditional on the initial class membership and the time-varying covariates at time $t$. This ensures that once an individual is assumed to be in one of the $K_{t-1}$ regimes at time $t-1$, the individual can only switch to one of the $K_t$ hypothesized regimes at time $t$.

Maximum likelihood estimates of the parameters in the MSEM-RS can be obtained via the Expectation-Maximization (EM; Dempster, Laird, & Rubin, 1977; Titterington, Smith, & Makov, 1985; Everitt & Hand, 1981) algorithm. The estimation details have been documented elsewhere (see e.g., Asparouhov & Muthén, 2011; B. O. Muthén & Shedden, 1999) and are not reiterated here.

A Cusp Catastrophe-Inspired MSEM-RS Model for Representing Alcohol Use Dynamics

A MSEM-RS is proposed as an alternative to the cusp catastrophe model. Two unobserved initial regimes at time 1 are first obtained from “slicing” the catastrophe landscape
along the direction of the y-axis (i.e., the distal risk axis) to yield two initial regimes: one corresponding to high distal risk and another to low distal risk. That is, the initial regime variable partitions the $\alpha$-$\beta$ plane in Figure 1 into two segments, one with $\beta > 0$ and the other with $\beta < 0$. Then, conditional on whether individuals are in one of these initial regimes, we further describe the catastrophe landscape in terms of a “high drinking” regime and a “low drinking” regime. In contrast to the cusp catastrophe model, we explicitly incorporate some added dynamic functions to represent the within-person, over-time variability in the levels of the behavioral variable while individuals are in these two regimes.

**Initial class membership.** At time 1, the only latent variable of interest is an individual’s latent distal risk based on baseline individual difference measures. We assume that two regimes or classes exist at time 1, with the first class representing a high distal risk class (denoted as $C_{i1} = \text{high distal}$), and the second class representing a low distal risk class (denoted as $C_{i1} = \text{low distal}$). The same measurement model is assumed to be at work for both distal classes, with a series of distal risk measures serving as indicators of an individual’s latent distal risk in the form of Equations 5—6, except that there is no class dependency on the parameters in the measurement model.

A class-specific structural model is then used to define the distribution of individuals’ latent distal risk. For those in the high distal risk class, the initial distribution of latent distal risk is specified to be

$$(\text{Distal}_{it}|C_{i1} = \text{high distal}) \sim N(m_{\text{high distal}}, \psi_{\text{high distal}}),$$

while for those in the low distal risk class, the corresponding initial distribution is specified to be

$$(\text{Distal}_{it}|C_{i1} = \text{low distal}) \sim N(m_{\text{low distal}}, \psi_{\text{low distal}}).$$

(11)

In other words, we allow the two classes that exist at time 1 to differ in their average levels of distal risk (with means $m_{\text{high distal}}$ and $m_{\text{low distal}}$, respectively), as well as variances in distal risk levels (i.e., given by $\psi_{\text{high distal}}$ and $\psi_{\text{low distal}}$, respectively). To ensure that the first class is a high distal risk class, one may impose the constraint that $m_{\text{high distal}} > m_{\text{low distal}}$.

**Time 2 and beyond.** For $t = 2$ and beyond, a measurement model in the form of Equations 5–6 is again used to describe individuals in all latent classes/regimes, with no class-dependency in measurement parameters. In this case, the latent variable vector, $\mathbf{\eta}_{it}$, reduces simply to one latent variable, $\text{Drink}_{it}$, representing individual $i$’s latent drinking tendency at time $t$ and $\mathbf{y}_{it}$ consists of continuous (observed and unobserved) measures of the individual’s underlying drinking tendency.

At the structural level, individuals’ levels of drinking tendency are hypothesized to show distinct changes depending on the regime the person is in at time $t$, expressed as

<table>
<thead>
<tr>
<th>Regime</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1: High drinking</td>
<td>$(\text{Drink}<em>{it} - \text{Drink}^*</em>{\text{high}}) = \alpha_{\text{high}}(\text{Drink}<em>{i,t-1} - \text{Drink}^*</em>{\text{high}}) + \zeta_{\text{drink},it}$ (12)</td>
</tr>
<tr>
<td>R2: Low drinking</td>
<td>$(\text{Drink}<em>{it} - \text{Drink}^*</em>{\text{low}}) = \alpha_{\text{low}}(\text{Drink}<em>{i,t-1} - \text{Drink}^*</em>{\text{low}}) + \zeta_{\text{drink},it}$,</td>
</tr>
</tbody>
</table>

where the two regimes are distinguished by within-person, over-time fluctuations around two different baseline drinking levels, $\text{Drink}^*_{\text{high}}$ and $\text{Drink}^*_{\text{low}}$, with autoregressive parameters, $\alpha_{\text{high}}$ and $\alpha_{\text{low}}$, respectively. That is, while in each of these two regime, deviations in drinking tendency are hypothesized to fluctuate over time following an autoregressive model
of order 1 (i.e., an AR(1) model), as driven by process noise, \( \zeta_{drink,it} \), at time \( t \) and a lag-one autoregression parameter (\( \alpha_{high} \) or \( \alpha_{low} \)). To identify the two regimes, the high drinking regime is constrained to have a higher baseline, \( Drink_{high}^* \), than the low drinking regime baseline of \( Drink_{low}^* \). The process noise variable, \( \zeta_{drink,it} \), often referred to as “random shock” because it encompasses all remaining uncertainties that affect the latent variables in ways that are not modeled explicitly (Browne & Nesselroade, 2005), is assumed to be normally distributed with zero mean and variance, \( \psi \).

The hypothesized over-time dynamics of individuals’ drinking levels can be better understood by taking a closer examination of the value of the AR(1) parameter. If the absolute value of the AR(1) parameter is less than 1, any deviations in drinking tendency from baseline are projected to diminish exponentially over time in the absence of new shocks imposed by \( \zeta_{drink,it} \). The speed with which past influences diminish over time is determined by the AR(1) parameter. That is, the closer it is to 0, the faster the rate of return to the regime-specific baseline drinking level. In the special case where a substantial portion of the individuals in a sample is able to maintain abstinence over long periods of time (e.g., in an especially successful treatment program), the structural equation for the low drinking regime can be readily modified by setting \( Drink_{low}^* \) to zero. In addition, the process noise variance in the regime can be freed to yield a “nearly abstinent” regime during which only some fluctuations in drinking level are observed, but always around a baseline of zero. Alternatively, both \( \alpha_{low} \) and the process noise variance in this regime can be set to zero to yield a complete abstinence regime, namely, one with no variability in drinking level while individuals are in this regime. Higher-order AR processes can also be included by allowing drinking tendency from more distant time points (e.g., from time \( t-2 \), \( t-3 \), and so on) to influence the current drinking tendency.

One of the key appeals of the cusp catastrophe mode lies in its ability to allow for the presence of multiple modes of behavioral outcome (i.e., multimodality), as well as the possibility that individuals may show sudden jumps from one mode to another depending on their distal risk levels. In the context of alcohol use, this means that for individuals with low distal risk, the tendency for alcohol use increases proportionately with increase in phasic risk. In contrast, for those with high distal risk, sudden jumps in such tendency may be observed. The proposed regime-switching model captures these properties by allowing the effect of phasic risk on an individual’s transition probabilities to depend on the individual’s distal risk class membership at time 1, \( C_{i1} \). Specifically, a multinomial logistic regression model is specified for the transition probability as

\[
\pi_{jk,h,it} = \frac{\exp[a_{kt} + b_{1,jt} + (b_{2,jt} + b_{C_{i1}=high\;distal,jt})phasic_{it}]}{\sum_{s=1}^{K_{t}} \exp[a_{st} + b_{1,st} + (b_{2,st} + b_{C_{i1}=high\;distal,st})phasic_{it}]}.
\]  

Compared to Equation 9, we can see that the logit intercept, \( a_{h,kt} = a_{kt} \) is constrained to be invariant across both distal risk classes; \( b'_{h,kt} = [b_{1,jt} \; b_{2,jt} \; b_{C_{i1}=high\;distal,jt}] \) with \( b_{C_{i1}=high\;distal,jt} \) freely estimated only for those with \( C_{i1} = \) high distal, and \( x_{it} = [1 \; phasic_{it}]' \). The covariate \( phasic_{it} \) is a composite score indicating individual \( i \)’s phasic risk at a particular time point. The term \( b_{2,jt} \) indicates the deviation in logodds for those in the low distal risk class at time 1 of switching to regime \( k \) from regime \( j \) relative to switching to regime \( k \) from the reference regime with each unit of increase in phasic risk. \( b_{C_{i1}=high\;distal,jt} \) is the corresponding deviation in the logit slope associated with phasic risk for individuals who
are in the high distal risk class at time 1 compared to those who are in the low distal risk class.

Thus, for someone in the high distal risk class at time 1, the effect of phasic risk (as encapsulated in \( b_{2,j}I_{e} + b_{C_{1}=\text{high distal},j} \) in Equation 13) is hypothesized to be more pronounced compared to those in the low distal risk class. That is, we expect \( b_{C_{1}=\text{high distal},j} \) to be greater than zero such that for those in the high distal risk class, the same amount of increase in phasic risk may lead to a disproportionate increase in the probability of “falling off the wagon,” namely, to switch from complete abstinence or low level of drinking to heavy drinking. In a similar vein, we also hypothesize that for those in the high distal risk class, phasic risk would have a disproportionately large effect of increasing the probability for this class of individuals to stay within the heavy drinking regime compared to those who are in the low distal risk class. Thus, once individuals with high distal risk have transitioned into the heavy drinking regime, a prolapse is hypothesized to occur with much lower probability than those in the low distal risk class, even with the same amount of reduction in phasic risk.

Another catastrophe flag of interest is hysteresis, namely, the asymmetry in the values of phasic risk that trigger a lapse versus a prolapse. This property is already inherent to the proposed regime-switching model due to the general inequality between the probability of transitioning from the low to the high drinking regime, and transition in the reverse direction.

Empirical Illustration

The COMBINE Study

The data for this study are from the COMBINE study (Combined Pharmacotherapies and Behavioral Interventions for Alcohol Dependence; COMBINE Study Research Group, 2003), a multi-site randomized trial. A total of 1383 subjects across 11 research sites were randomized into 9 treatment groups. Treatment was provided for 16 weeks, and participants were followed for one year post–treatment.

The sample was recruited from inpatient and outpatient referrals at the study sites and throughout the community. Prior to baseline, 4965 volunteers were screened by telephone to determine whether the individuals met eligibility criteria. Participants were excluded if they were dependent on another drug besides alcohol, nicotine, or cannabis, recently used opioids, had a serious mental illness, had any other medical condition that could disrupt study participation, had taken one of the study medications 30 days prior to baseline, or took medication that could raise the potential risks of the study. To be included in the study, subjects needed to have a minimum of 14 drinks (females) or 21 drinks (males) average per week over 30 days in the 90-day period prior to beginning abstinence. Additionally, participants needed to have two or more days of heavy drinking in the 90-day period, with the last drink being within 21 days of enrollment. A heavy drinking day was defined as 4 drinks for females and 5 drinks for males. Following meeting eligibility criteria, subjects were required to produce a breath alcohol level of zero before completing consent and baseline assessments.

In the COMBINE study, the participants were randomly assigned to one of nine
treatment groups\(^2\). The final sample included 1,383 participants, 31% were female and 69% were male, 23% of the study population were ethnic minorities (76.3% Non-Hispanic White, 11.6% Hispanic American, 7.8% African American, and 4.1% Other). The subjects median age was 44 years, 71% had at least 12 years of education, and 42% were married.

**Measures**

**Outcome measure.** In the COMBINE study, drinking measures were collected at baseline prior to the beginning of treatment (denoted herein as week 0), during treatment (at weeks 1, 2, 4, 6, 8, 10, 12 and 16) and post-treatment (at weeks 26, 36, and 52). Consistent with common practice in the alcohol use disorder (AUD) modeling literature, we used each participant’s Proportion of Heavy Drinking Days (PHD) since the last assessment occasion (or up to 30 days prior to the next occasion, whichever one was shorter) as our key dependent variable of interest. Using information deduced from the Form-90 interview (Miller & Del Boca, 1994), heavy drinking was defined as 4 or more drinks per day for women and 5 or more drinks per day for men. For model fitting purposes, we used data spanning every two weeks from week 0 to week 16, yielding a total of 9 equally-spaced time points, one of which was missing. To evaluate whether information from different points of the treatment process served to indicate longer-term drinking outcomes, the participants’ drinking status at week 52 was used as the target outcome we sought to predict using information available from model fitting.

Time series plot and histogram of the PHD scores of 300 randomly selected participants over the course of the treatment program are shown in Figures 3A-B. The presence of a large number of abstainers at any single time point is evident in both of the plots. Two other prominent characteristics of the data were (1) the highly positive skewed nature of alcohol data, and (2) the considerable variability in drinking levels among the non-abstainers. To account for the positive skew of the data, we performed model fitting using the log transformed PHD scores (plotted in Figure 3C). PHD scores that were exactly zero (corresponding to complete abstinence) was fixed to a small non-zero constant (i.e., .01) to enable us to take log transformation. It can be seen from the plot that the log transformed PHD scores still deviated from normality due to the presence of a large number of abstainers and right censored responses. As noted earlier, the first issue is commonly referred to as the zero-inflation phenomenon. The latter arose because the participants could not have a PHD score that exceeded 1.0 (i.e., with greater than 100% of heavy drinking days for a particular assessment period). Thus, the scores were “censored” at the upper value of 1.0. Ways of handling these deviations from normality will be described in the context of our modeling results.

**Measures of phasic risk.** Negative emotion, perceived stress, and craving were incorporated as measures to derive a composite indicator of phasic risk. The Profile of Mood States–Brief (POMS; McNair, Lorr, & Droppleman, 1971) was administered as a measure

---

\(^2\)The Medical Management groups \((n = 607)\) included: Naltrexone, Acamprosate, Naltrexone + Acamprosate, and Placebo. The Combined Behavioral Intervention (CBI) groups \((n = 776)\) consisted of: Naltrexone + CBI, Acamprosate + CBI, Naltrexone + Acamprosate + CBI, Placebo + CBI, and CBI-only (COMBINE Study Research Group, 2003). Subjects received treatment for a total of 16 weeks; participants receiving study medication were offered 9 Medical Management visits and those who received CBI had a maximum of 20 sessions.
Figure 3. Data from 300 randomly selected participants from the COMBINE data set, with (A) a plot of the biweekly PHD scores over the course of the treatment program, (B) a histogram showing the frequencies of the PHD scores, (C) histogram of the log transformed PHD scores. Scores corresponding to 100% abstinence (i.e., zeros) were converted into a small non-zero constant (i.e., .01) prior to being subjected to log transformation.

of negative emotions. Participants were asked to provide ratings on 30 adjectives describing feelings and moods (e.g., Tense, Angry, Annoyed, etc.), with possible responses ranging from 0 (not at all) to 4 (extremely). Ratings on the 30 items were combined into six subscales: Tension, Depression, Anger, Vigor, Fatigue, and Confusion. For the present study, a composite score of negative emotions was obtained by averaging each individual’s ratings on the Tension, Depression, Anger, and Fatigue subscales.

The four-item version of the Perceived Stress Scale (PSS; Cohen, 1983) was used to measure the participants’ levels of perceived stress (over the week prior to each assessment) on a five-point Likert-type scale (from never to very often). The Obsessive Compulsive Drinking Scale (OCDS; Anton, Moak, & Latham, 1995) was used as a measure of craving. The OCDS is a 14-item self-report instrument assessing drinking-related thoughts, urges to drink, and the ability to resist thoughts and urges to drink. Items were rated on a five-point Likert-type scale, with lower ratings indicating less craving. Two of the 14 items, assessing quantity and frequency of drinking in the past week, were excluded from all analyses and a total score was calculated by summing the remaining 12 items at each time point.

For model fitting purposes, we computed a composite phasic risk score for each individual and each time point using the measures described earlier. Because the different
measures of phasic risk all conformed to very diverse scales, we first standardized each phasic risk indicator using each individual’s within-person, over-time mean and standard deviation on that indicator prior to computing a composite score using all the variables. This standardization procedure was appropriate in the current context because we were primarily interested in using phasic risk as a time-varying covariate. That is, we were not interested in the between-person differences in overall phasic risk level; rather, we sought to investigate how within-person, over-time variations in phasic risk would influence individuals’ probabilities of manifesting lapses and prolapses on an occasion-by-occasion basis.

Measures of distal risk. Alcohol dependence measures based on self-reported symptoms and a structured clinical interview were used as potential indicators of distal risk at baseline. The measures included self-reported alcohol dependence severity, the number of alcohol dependence symptoms manifested by each participant, and a clinician’s rating of the participant’s severity of alcohol dependence. Self-reported alcohol dependence severity was based on scores from the Alcohol Dependence Scale (ADS; Skinner & Horn, 1984), a 25-item measure of alcohol withdrawal symptoms, impaired control over drinking, awareness of compulsions to drinking, tolerance, and drink-seeking behavior. The number of alcohol dependence symptoms was determined according to the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV; Association, 1994) criteria, using Module E of the Structured Clinical Interview for DSM-IV (SCID; First, Spitzer, Gibbon, & Williams, 1997). This measure is a count variable and it is denoted herein as $ALCDEP$. A final measure was an ordinal item from the SCID interview (item 22 from the alcohol evaluation module of the SCID) indicating a clinician’s rating on the participant’s severity of alcohol dependence during “the worst week of the past month” using one of three response categories: mild, moderate and severe. This ordinal item is referred to herein as $ORD-DEPSEV$.

In sum, to identify a continuous latent variable representing individual $i$’s distal risk prior to the initiation of treatment, we used three observed indicators, two of which were discrete in nature (including a count variable, $ALCDEP$, and an ordinal item, $ORD-DEPSEV$), and one was continuous in nature (i.e., scores on the ADS). Compared to specifying $ALCDEP$ as a continuous indicator, we found that specifying it as a count variable with a Poisson link function led to slower computational speed and estimation difficulties in fitting some of the more complex models, while yielding only negligible differences in the modeling results. We thus proceeded to treating it as a continuous observed indicator in all subsequent model fitting. In addition, we were not interested in the scaling differences between these measures. Thus, $ALCDEP$ and $ADS$ were both standardized using the means and standard deviations of all participants at week 0 prior to model fitting.

Model Fitting Results

The MSEM-RS model summarized in Equations 11—13 was fitted to the COMBINE data. Based on the plots of the data in Figure 3C, there was a substantial portion of abstainers in the sample. This data characteristic is commonly encountered in empirical AUD data. To accommodate this data feature, we modified the structural model for regime 2 (the low drinking regime) to define a complete abstinence regime by setting $\alpha_{low}$, $Drink_{low}^*$, and the process noise variance in this regime (i.e., variance of $\zeta_{drink,it}$) to be zero. In addition, for our modeling purposes, the logit parameters at $t = 2$ were allowed to differ
from the corresponding parameters at the remaining time points (i.e., for \( t = 3, \ldots, T \)). From \( t = 3 \) and beyond, we constrained these parameters to be invariant over time. This is because at \( t = 2 \), the transition probabilities captured the probabilities of transitioning from different distal risk classes into the two drinking regimes (i.e., complete abstinence vs. drinking). These transition probabilities were thus conceptually distinct from the transition probabilities for the remaining time points, where they reflected the probabilities of transitioning between the two drinking regimes. Another caveat noted earlier was that the log transformed PHD scores were right censored at the value of \( \log(1) \). Here, we relied on the use of the sandwich-type standard error estimator in Mplus to circumvent the non-normality of the log transformed PHD scores while individuals were in the drinking regime.

Parameters from fitting the proposed model wherein only the statistically significant parameters are summarized in Table 1. Based on the value of the logit intercept for the high distal risk class (i.e., \( a_{12} \)), 87% of the participants were classified to be in the high distal risk class whereas the remaining 13% were classified to be in the low distal risk class. To help shed light on the between-regime differences in transition patterns for the remaining time points, the logit intercept and slope parameters in Table 1 were used to compute the transition probabilities associated with different periods of the study. The parameter \( b_{1,12} \) was not significantly different from zero, indicating that on the days with average phasic risk, there was no significant difference in the log odds of the two distal risk classes to switch into the drinking regime at \( t = 2 \). Thus, from pre-treatment to the first assessment occasion during the treatment phase (i.e., \( t = 1 \) to \( t = 2 \), corresponding to 2 weeks after the initialization of treatment), the transition probability matrix on the days with an average level of phasic risk was governed by

\[
\begin{array}{ccc}
\text{Drink}_{t=2} & \text{Abs}_{t=2} \\
\text{High distal risk}_{t=1} & .27(0.02) & .73(0.02) \\
\text{Low distal risk}_{t=1} & .27(0.02) & .73(0.02) \\
\end{array}
\]

The patterns of transition probabilities indicate that within the first 2 weeks of the treatment program, a large proportion of the participants were able to maintain abstinence on the days with average phasic risk. There was no significant difference in the two distal risk classes in their probabilities for a lapse. In contrast, on the days with high phasic risk (i.e., 1 standard deviation above the mean), the transition probability matrix on the days with an average level of phasic risk was given by

\[
\begin{array}{ccc}
\text{Drink}_{t=2} & \text{Abs}_{t=2} \\
\text{High distal risk}_{t=1} & .64(0.04) & .36(0.04) \\
\text{Low distal risk}_{t=1} & .27(0.02) & .73(0.02) \\
\end{array}
\]

That is, on the days with high phasic risk, individuals with high distal risk showed a substantially higher probability for a lapse compared to those who were low in the distal risk class. For those in the low distal risk class, the parameter \( b_{2,22} \) was not significantly different from zero, suggesting that phasic risk did not have a significant effect on the probability for the low distal risk class to manifest a lapse during the first two weeks of the treatment program. On the contrary, the parameter \( b_{2,12} \) was positive and significantly different from zero. Thus, for those in the high distal risk class, a one unit increase in phasic risk led to an abrupt increase in the probability of a lapse from .27 to .64.
Table 1: Parameter Estimates Obtained from Fitting the Regime-Switching Alcohol Use Model. R1 = Drinking Regime; R2 = Complete Abstinence Regime.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Meanings</th>
<th>Estimates (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m_{high distal}$</td>
<td>Average distal risk of high distal risk class</td>
<td>-0.01 (0.03)</td>
</tr>
<tr>
<td>$m_{low distal}$</td>
<td>Average distal risk of high distal risk class</td>
<td>-0.21 (0.09)</td>
</tr>
<tr>
<td>$\psi_{high distal}$</td>
<td>Variance in distal risk of high distal risk class</td>
<td>0.37 (0.03)</td>
</tr>
<tr>
<td>$\psi_{low distal}$</td>
<td>Variance in distal risk of low distal risk class</td>
<td>0.31 (0.05)</td>
</tr>
<tr>
<td>$a_{1}$</td>
<td>Logit intercept at time 1</td>
<td>1.90 (0.22)</td>
</tr>
<tr>
<td>$m_{drink 1}$</td>
<td>Initial average alcohol use tendency at $t = 2$</td>
<td>-1.24 (0.05)</td>
</tr>
<tr>
<td>$\psi_{1}$</td>
<td>Initial variance in alcohol use tendency at $t = 2$</td>
<td>0.47 (0.03)</td>
</tr>
<tr>
<td>$Drink_{high}^{*}$</td>
<td>Baseline drinking level in R1</td>
<td>-1.32 (0.05)</td>
</tr>
<tr>
<td>$\alpha_{high}$</td>
<td>AR(1) parameter in R1 starting from $t = 3$</td>
<td>0.35 (0.02)</td>
</tr>
<tr>
<td>$\psi$</td>
<td>Random shock variance in R1 starting from $t = 3$</td>
<td>0.56 (0.02)</td>
</tr>
<tr>
<td>$a_{12}$</td>
<td>Logit intercept at $t = 2$</td>
<td>-1.00 (0.09)</td>
</tr>
<tr>
<td>$b_{1,12}$</td>
<td>Deviation in log odds of being in R1 at $t = 2$</td>
<td>0</td>
</tr>
<tr>
<td>$b_{2,12}$</td>
<td>Effect of phasic risk on the logodds of transitioning from $C_{i1} = high distal class to $C_{i2} = drinking (R1)$</td>
<td>1.57 (0.17)</td>
</tr>
<tr>
<td>$b_{2,22}$</td>
<td>Effect of phasic risk on the log odds of transitioning from $C_{i1} = low distal class to $C_{i2} = drinking (R1)$</td>
<td>0</td>
</tr>
<tr>
<td>$a_{1t}, t = 3, \ldots T$</td>
<td>Logit intercept for other $t$</td>
<td>-1.71 (0.08)</td>
</tr>
<tr>
<td>$b_{1,1t}, t = 3, \ldots T$</td>
<td>Deviation in log odds of being in R1 for other $t$</td>
<td>3.09 (0.13)</td>
</tr>
<tr>
<td>$b_{2,1t}, t = 3, \ldots T$</td>
<td>Effect of phasic risk on the logodds of staying in the drinking regime from $t-1$ to $t$ given $C_{i1} = low distal class$</td>
<td>1.03 (0.16)</td>
</tr>
<tr>
<td>$b_{2,2t}, t = 3, \ldots T$</td>
<td>Effect of phasic risk on the logodds of transitioning from abstinence at $t-1$ to drinking at $t$ given $C_{i1} = low distal class$</td>
<td>-2.12 (0.39)</td>
</tr>
<tr>
<td>$b_{C_{i1} = high distal, 1t}, t = 3, \ldots T$</td>
<td>Deviation in the effect of phasic risk on $b_{2,1t}$ given $C_{i1} = high distal risk class$</td>
<td>0</td>
</tr>
<tr>
<td>$b_{C_{i1} = high distal, 2t}, t = 3, \ldots T$</td>
<td>Deviation in the effect of phasic risk on $b_{2,2t}$ given $C_{i1} = high distal risk class$</td>
<td>3.97 (0.37)</td>
</tr>
</tbody>
</table>
Figure 4. (A) A plot of the posterior probability of being in the abstinence regime (R2) at each time point from a random subsample of 500 participants; (B) a plot of the proportion of participants with higher posterior probability of being in the abstinence regime than in the drinking regime at each time point.
The transition probabilities characterizing the rest of the study span helped indicate the relative stability of the drinking and abstinence regimes once individuals transitioned into these regimes. While individuals were in the drinking regime, the AR(1) parameter of 0.35 indicated that when individuals deviated from the baseline drinking level of \( Drink_{high}^* = -1.32 \), they tended to return to it relatively quickly. The closer the AR(1) parameter was to 0 in absolute value, the quicker the return.

On the days with average phasic risk, the same transition probability matrix was assumed to govern individuals’ drinking patterns regardless of initial distal risk status. The corresponding transition probability matrix was observed to be

**Average phasic risk, all distal risk classes:**

\[
\begin{array}{cc}
\text{Drink}_t & \text{Abs}_t \\
\text{Drink}_{t-1} & .80(0.01) & .20(0.01) \\
\text{Abs}_{t-1} & .15(0.01) & .85(0.01)
\end{array}
\]

That is, on the days with average phasic risk, the probability for a lapse was relatively low \( \text{Pr}(\text{Drinking at time } t \mid \text{Abstinence at time } t-1) = .15 \), but the probability for a prolapse was also very low \( \text{Pr}(\text{Abstinence at time } t \mid \text{Drinking at time } t-1) = .20 \). On the days with high phasic risk (i.e., 1 SD above the mean), some differences in transition probability patterns were observed. Specifically, although there was no significant difference in the log odds for the two distal risk classes to manifest a prolapse on the days with high phasic risk (i.e., \( b_{C_{11}}=high \text{ distal}_1 \), was not significantly different from zero; see Table 1), those with high distal risk were much more likely show a lapse when phasic risk was high. Consequently, for individuals in the low distal risk class, their transition probability matrix was given by

**High (+1SD) phasic risk, low distal risk:**

\[
\begin{array}{cc}
\text{Drink}_t & \text{Abs}(t) \\
\text{Drink}_{t-1} & .92(0.02) & .08(0.02) \\
\text{Abs}_{t-1} & .02(0.01) & .98(0.01)
\end{array}
\]

whereas for those in the high distal risk class, the corresponding transition probability matrix was given by

**High (+1SD) phasic risk, high distal risk:**

\[
\begin{array}{cc}
\text{Drink}_t & \text{Abs}(t) \\
\text{Drink}_{t-1} & .92(0.02) & .08(0.02) \\
\text{Abs}_{t-1} & .54(0.05) & .47(0.05)
\end{array}
\]

Thus, our hypothesis concerning the moderating effect of distal risk on individuals’ probabilities for a lapse when phasic risk was high was confirmed. Specifically, individuals with high distal risk were found to show a significantly greater increase in the probability for a
lapse when phasic risk was high (i.e., from .15 to .54), as indicated by the statistical significance of the parameter $b_{C_1}=\text{high distal}_2$. In contrast, for those with low distal risk, the probability for a lapse actually decreased slightly on the days with high phasic risk. This may due in part to the modeling constraint that the two distal risk classes were specified to show the same transition probability patterns on the days with average phasic risk. That is, based on postulates of the cusp catastrophe model, we did not include the main effects of distal risk on the transition probabilities in Equation 13. Instead, distal risk was posited to only play a moderating role in the effects of phasic risk on the transition probabilities. Regardless, the detrimental effects of high phasic risk in triggering a lapse among those with high distal risk is clear throughout the treatment phase.

Contrary to our initial hypothesis, we did not find a significant difference between the two distal risk classes in their probability to bounce back from a lapse (i.e., to manifest a prolapse). Whereas the estimated prolapse rate was generally low (.20) for both distal risk classes even on the days with average phasic risk, the probability of a prolapse reduced by more than half (i.e., the probability decreased from .20 to .08) when phasic risk was high.

A plot of the posterior probabilities of 500 randomly selected participants is shown in Figure 4A. These were the estimated probabilities that the participants were in the abstinence regime conditional on the data. The plot indicates that the participants continued to show ongoing transitions into and out of the abstinence regime throughout the 16-week treatment period. The proportion of participants assigned to the abstinence regime at each time point based on their highest posterior probabilities is plotted in Figure 4B. The plot indicates that overall, approximately 60% of the participants managed to maintain abstinence at each time point. However, a substantial portion of the participants continued to switch frequently between the drinking regime and the abstinence regime. The low stability of the abstinence regime was particularly salient among those with high distal risk, whose probability for a lapse depended heavily on the level of phasic risk experienced in the moment (see transition probabilities in Equation 17).

Overall, some of our key results indicated that individuals with high distal risk were especially susceptible for a lapse when phasic risk was high within the first two weeks of the treatment program. Specifically, individuals who were high in distal risk were more likely to transition into the drinking regime when faced with a heightened phasic risk level within the first two weeks of treatment ($P = .64$) than individuals with low levels of distal risk ($P = .27$). In fact, the probability for the low distal risk class to manifest a lapse during the initial treatment phase appeared unrelated to the fluctuations in phasic risk reported by the participants. It is thus critical to monitor the progress of the high distal risk group closely during the early stages of a treatment program.

Discussion

The MSEM-RS model summarized in Equations 11–13 was proposed as an alternative to the cusp catastrophe model in the context of alcohol use. While it may not be as mathematically elegant as the cusp catastrophe model (e.g., the cusp catastrophe model is capable of producing the rich array of dynamics described earlier with only one nonlinear potential function), it relaxes some of the constraints of the cusp catastrophe model that may be overly restrictive in empirical analytic situations. First, the proposed model can be fit using existing structural equation modeling programs such as Mplus (L. K. Muthén &
The corresponding estimation procedures implemented in these programs can readily handle longitudinal panel data with multiple subjects as well as the presence of missing data hypothesized to be missing at random (Little & Rubin, 2002). Because the broader modeling framework assumes that a structural equation model underlies each latent regime, the proposed model can fully capitalize on the flexibility of the structural equation modeling framework in handling (linear) structural relations among latent variables. Furthermore, the proposed model incorporates the lagged effects of previous drinking on current drinking (while individuals are in the drinking regime) and it can be expanded, as needed, to test other multivariate extensions involving other variables of interest. In addition, by using person- and time-specific covariates to predict the probabilities of transitioning between regimes, the new approach also allows for heterogeneous timing of lapse and prolapse within and across subjects. Finally, the zero-inflation phenomenon often observed in alcohol use data is now explicitly modeled by specifying a complete abstinence regime for the abstainers.

Despite some of the promises of the proposed regime-switching dynamics models, some methodological challenges remain. For instance, the increase in computational costs was substantial when more than 5 time points were included in the model. The increase in computational time was even more substantial when we attempted to fit variations of the proposed model with three drinking regimes. This imposes added burden in fitting mixture models because users are often advised to fit mixture models using multiple starting values to check the sensitivity of the estimation results to different starting values, which is a computationally intensive pursuit in itself. To this end, some of the alternative formulations in the time series and state–space literature as well as related estimation procedures can be adopted (Hamilton, 1994; Kim & Nelson, 1999; Yang & Chow, 2010).

Although the proposed model was inspired by the cusp catastrophe model and was able to retain some of its features, it should by no means be regarded as a replacement of the cusp catastrophe or other related catastrophe models. For instance, through the use of one nonlinear potential function, the cusp catastrophe model is capable of generating a wide array of interesting dynamic properties such as hysteresis, sudden jumps and the co-existence of multiple equilibria. Our use of two different sets of transition probability patterns to characterize the differing pathways of change manifested by individuals with different degrees of distal risk also did not fully capture the distinction between the linear and the nonlinear pathways of change posited in the cusp catastrophe model. In addition, as opposed to treating distal risk as a continuous variable (e.g., the way the splitting parameter is utilized in the cusp catastrophe model), we only distinguished between the trajectories of those in the low versus high distal risk class. In other words, distal risk was treated in our proposed model as a categorical, as opposed to a continuous latent variable. Thus, as emphasized earlier, the proposed model is intended to serve as an alternative to the cusp catastrophe in the context of modeling alcohol use. Whereas it helps to circumvent several data analytic issues commonly encountered in the analysis of AUD data, we still acknowledge the general appeal of the cusp and other related catastrophe models as a representation of change.
References


Asparouhov, T., & Muthén, B. O. (2011). C on C and X. (Mplus technical appendix)


Muthén, B. O., & Asparouhov, T. (2011, July). LTA in Mplus: Transition probabilities influenced by covariates. (Mplus Web Notes: No. 13)


and instability of peer victimization during middle school: Using latent transition analysis with covariates, distal outcomes, and modeling extensions.


