Foundations and Trends$^{®}$ in Marketing
Vol. 9, No. 3 (2014) 175–266

DOI: 10.1561/1700000031

Marketing Dynamics: A Primer on Estimation and Control

Prasad A. Naik
University of California,
Davis, USA
panaik@ucdavis.edu

Contents

Abstract

This primer provides a gentle introduction to the estimation and control of dynamic marketing models. It introduces dynamic models in discrete- and continuous-time, scalar and multivariate settings, with observed outcomes and unobserved states, as well as random and/or time-varying parameters. It exemplifies how various dynamic models can be cast into the unifying state space framework, the benefit of which is to use one common algorithm to estimate all dynamic models.

The primer then focuses on the estimation part, which answers questions such as: how much is the sales elasticity of advertising? How much sales lift can managers expect for a certain level of price promotion? What is the best sales forecast for the next quarter? The estimation relies on two principles: Kalman filtering and the likelihood principle. The Kalman filter recursively infers the means and covariances of an unobserved state vector as the observed outcomes arrive over time. This evolution of moments is then embedded in the likelihood function to obtain parameter estimates and their statistical significance.

Next, the primer elucidates the control part, which answers questions such as: how much should managers spend on advertising over time and across regions? What is the best promotional timing and depth? How should managers optimally respond to competing brands' actions and resulting outcomes? The control part relies on the maximum principle and the optimality principle. Pontryagin's maximum principle allows managers to determine the optimal course of action (for example, the optimal levels and timing of advertising spends or price promotions) to attain a specified goal, such as profit maximization. Bellman's optimality principle, on the other hand, offers insights into optimal course correction when implementing the best plan as the state of a system varies dynamically and/or stochastically. Finally, the primer presents three examples on the application of optimal control, differential games, and stochastic control theory to marketing problems, and illustrates how to discover novel insights into managerial decision-making.

P. A. Naik. *Marketing Dynamics: A Primer on Estimation and Control*.
Foundations and Trends$^{\circledR}$ in Marketing, vol. 9, no. 3, pp. 175–266, 2014.

DOI: 10.1561/1700000031.

1

Introduction

In 1696, Johann Bernoulli posed a challenge to his contemporary "sharpest mathematical minds of the globe" with the following problem: If in a vertical plane two points A and B are given, then what is the trajectory of an object C starting from A to arrive at B in the shortest possible time falling under its own weight? He added that "this problem . . . is not . . . purely speculative and without practical use . . . Rather it even appears . . . that it is very useful also for other branches of science than mechanics." This problem is called *Brachystochrone* problem because, in Greek, $\beta\rho\alpha\chi\iota\sigma\tau\zeta$ = shortest and $\chi\rho o\nu o\zeta$ = time. Leibniz described this problem as "splendid" and furnished the solution (a cycloid) in a letter to Bernoulli, while Newton presented his solution to the Royal Society anonymously.[1] For the quoted text and a definitive account of the intellectual history, see Sussmann and Willems [1997].

[1] Johann Bernoulli ascribes the anonymous solution to Newton because he noted that "you can tell the lion by its claws" (ex ungue leonem). The solution is the cycloid curve for the point C whose coordinates $(x(t),\, y(t))$ evolve, starting from the point A at (0,0) at $t = 0$, according to $x(\theta) = \alpha(\theta - \sin(\theta))$ and $y(\theta) = \alpha(\cos(\theta) - 1)$, where $\theta(t) = t\sqrt{g/\alpha}, g = 9.8\,\mathrm{m/sec}^2$, and the parameters $(\alpha,\, T)$ are determined by the terminal condition $(x(\theta(T)), y(\theta(T))) = B$.

This event gave birth to the mathematical branch, later known as the calculus of variation, due to Leonhard Euler, who was Bernoulli's student and who discovered what is now called the Euler's equation to solve such dynamic optimization problems. Intense mathematical research over the subsequent three hundred years culminated into the modern control theory. Consistent with Bernoulli's prognosis, this theory found use in launching man on Moon, landing Curiosity on Mars, deploying unmanned drones or designing driverless cars, high precision manufacturing using robots, providing navigation guidance turn by turn to users on roads, seas, and air, besides numerous applications in "other branches of science" including Operations Research (for example, Berstekas [2005]), Economics (for example, Stokey and Lucas [1989], Aghion and Howitt [1998], Ljungqvist and Sargent [2004], Weber [2011], Kamien and Schwartz [2012]), Management Science (for example, Sethi and Thompson [2000], Dockner et al. [2000]), and Marketing (for example, Erickson [2003], Jørgensen and Zaccour [2004]).

In Marketing, the point A represents the current state of the company's brand sales or consumer's utility. The point B marks the desired state the decision-maker wants to arrive at. The object C is the decision-maker (for example, CEO, brand managers, consumers), and its use of own weight denotes the set of actions (for example, price, advertising, brand choices) available for transitioning the system from state A to state B. The shortest time specifies the decision-maker's objectives (for example, maximize the stream of future profit or utility). In Section 2, I clarify the terms state, system, transition, actions (or controls), and objectives, but note here that this simple abstraction is "splendid" because it not only unifies diverse problems across many applications, but also offers a systematic approach for solving them.

The purpose of this primer is to impart this systematic approach for solving dynamic marketing problems. To pursue this pedagogical focus, this article does not aim to review dynamic models in the extant marketing literature, for which readers are referred to Bowman and Gatignon [2010], Neslin and van Heerde [2009], Shankar [2008], Hanssens et al. [2003], and Leeflang et al. [2000].

Solving dynamic problems involves two distinct topics: parameter estimation and optimal control. The former refers to *describing* *the system* of relations among current states, past states, and actions via econometric time-series models (for example, Pauwels [2004], Steenkamp et al. [2005]). The latter refers to *managing the system* by determining the optimal course of actions to execute over the future planning horizon (for example, Kumar et al. [2008], Esteban-Bravo et al. [2014]). The foundation for parameter estimation stands on the Kalman Filter and the Likelihood Principle, whereas that for optimal control stands on the Pontryagin's Maximum Principle and Bellman's Principle of Optimality.

This monograph elucidates these four principles and related concepts, focusing on how to estimate dynamic models in Sections 3 and 4, and how to solve the control problems in Sections 5 and 6. But first, I clarify the terms, present the diversity of dynamic models, unify them via the canonical state space model, and highlight the value of unification.

2

Dynamic Models in Marketing

This section clarifies the terms used in the context of dynamic models. Then, I illustrate 10 examples of dynamic marketing models (five discrete-time and five continuous-time). Next, I introduce the state space formulation, which unifies and nests various dynamic models: linear and nonlinear, deterministic and stochastic, discrete- and continuous-time. Finally, I state the main advantages of using state space form.

2.1 Types of variables

Let me first clarify the meanings of the key terms. A dependent variable, denoted by y_t, represents an *observed* outcome at the instant t (for example, a brand's sales level in say January 2015). Whereas a state variable, denoted by x_t or α_t, is not observed directly by measurements, for example, brand's goodwill [Naik et al., 1998], brand's equity [Sriram and Kalwani, 2007, Sriram et al., 2007], consumer's utility and preferences [Lachaab et al., 2006, Teixeira et al., 2010], consumer's interest and conversion [Hu et al., 2014]. But it affects the dependent variable and is affected by a control variable and/or an independent variable. For instance, a brand's goodwill is a state variable that is not directly measured by the company, but it affects the prevailing level of brand sales (an observed dependent variable), and is affected by the brand's advertising (a control variable) and regional seasonality (an independent variable). The concept of state variable is quite broad and context dependent — besides conceptual marketing variables (for example, cognition, affect, experience), even a model's parameter to be estimated or a random error term itself (not just its variance) can be a state variable.

A control variable, denoted by u_t, is a decision variable such as the level of ad spending and/or brand's price determined by a decision-maker. In other words, a marketing manager decides ad spending or price to achieve a certain objective of maximizing profit, for example. In contrast, an independent variable, also known as an exogenous variable or a covariate, is deemed to be pre-determined or not chosen by the decision-maker in a systematic way. This distinction between control and independent variables matters because we want to not only describe the dynamic system, but also manage it to achieve certain objectives.

In sum, a manager decides the values of a control variable (that is, an endogenous variable), which drives the state variable, which in turn affects the observed dependent variable; independent variables are not decision variables but they influence the state and/or dependent variables. Indeed, multiple outcomes (for example, own and competitor's sales), multiple states (for example, goodwill, ad effectiveness), multiple controls (for example, advertising, price), and multiple covariates (for example, seasonality, inflation) are present in a market place. A set of equations relating such variables is called a dynamic system.

Table 2.1 provides a glossary of various variables to be encountered in this monograph.

Table 2.1: Various variables!

Term	Meaning
Adjoint variable	See costate variable.
Continuous variable	A variable that lives on a real number line or its segment. For example, probability of purchase lives on a positive unit interval and potentially takes all fractional values. See discrete variable.
Control variable	A trajectory of the decision variable over time (that is, $\{x_t^* : t = 1, \ldots, T\}$) or a function of state variable (that is, $x_t^* = f(\alpha_t)$). It is often denoted by u or v in the presence of independent variables, which are nondecision variables and denoted by x. See decision variable, pre-determined variable.
Costate variable	A variable, denoted by λ_t or μ_t, associated with each state variable. It captures the "shadow price" of marginal relaxation of the dynamic state constraint. It is a dynamic analog of the Lagrange multiplier in the static optimization theory.
Covariates	See independent variable, pre-determined variable, regressors.
Decision variable	A variable that affects the objective function (for example, brand profit or consumer utility), and a decision-maker seeks to know its "best" value that attains the objective. Suppose advertising expenditure or brand's price x affects brand profit $\Pi(x)$; then x is the decision variable, and $x^* = \operatorname{Arg\,Max} \Pi(x)$ is the best value. See control variable, pre-determined variable.
Dependent variable	A variable that depends on independent variables or regressors, denoted usually by y. See outcome variable.
Discrete variable	A variable that takes the integer values on a real number line or its segment. For example, weeks of the year live on the index set $\{1, 2, \ldots, 52, \ldots\}$. See continuous variable.
Drift variable	A variable that shifts the state vector or its observation and denoted by c_t or d_t in state space models.
Endogenous variable	See dependent variable.

(*Continued*)

Table 2.1: (*Continued*)

Term	Meaning
Exogenous variable	See independent variable, pre-determined variable, and regressors.
Filtered variable	A current mean of an outcome (or state) variable based on the current information set. That is, $\hat{y}_t = E[y_t\|I_t]$. See forecasted variable, smoothed variable.
Forecasted variable	A future mean of an outcome (or state) variable based on the current information set. That is, $\hat{y}_{t+h} = E[y_{t+h}\|I_t]$, where h is the forecast horizon. See filtered variable, smoothed variable.
Independent variable	A variable that does not depend on the outcome or dependent variable, denoted usually by x. In other words, x_t is not a function of y_t. It can be a pre-determined variable based on past $(x_{t-k},\, y_{t-k})$ where $k \geq 1$. A set of independent variables need not be, and often are not, statistically independent of each other. See regressors, pre-determined variable.
Instrumental variable	An instrumental variable z allows one to estimate unbiased $\hat{\beta} = (z'x)^{-1}(z'y)$ when (x, ϵ) are correlated in the scalar regression model $y = \beta x + \epsilon$, where ϵ is an error term. When (x, ϵ) are uncorrelated, the ordinary least squares $\hat{\beta} = (x'x)^{-1}(x'y)$ yields the unbiased estimates.
Interaction variable	Consider a linear model, $y = \beta x + \gamma z + \alpha xz$. Then (x, z) are interacting variables, $(\beta,\, \gamma)$ are their simple effects respectively, and α is the interaction effect. Also $\frac{dy}{dx} = \beta + \alpha z$, so see moderating variable, and synergistic variable.
Intermediate variable	See mediating variable.
Lag variable	A variable whose past values affect the current outcome. For example, $y_t = f(x_{t-1})$. Also x can be y itself.
Latent variable	A theoretical construct not directly observable: *one observes its effects but not itself*. It is operationalized via multiple proxy variables. For examples: goodwill in marketing, inflation in economics, or latent heat in physics.

(*Continued*)

Table 2.1: (*Continued*)

Term	Meaning
Lead variable	A variable whose future values affect the current outcome. For example, $y_t = f(x_{t+1})$. Also x can be y itself. It is incorrect to simply include future values as regressors directly to predict current y. A proper treatment requires using the *expectation* of future values based on current information set, that is, $y_t = f(E[x_{t+1}\|I_t])$. The quantity $E[x_{t+1}\|I_t]$ is the prior mean, and it is readily available from the Kalman filter recursions.
Markovian variable	A variable whose future values depend only on the current values. The "current" period can be augmented to include contiguous past periods. For example, if future depends on past two days, then the "current" state includes today and yesterday. The main point is that the limited past predicts the future; that is, the distant past is irrelevant. Although used in the context of random variables, this concept applies to deterministic variables as well (for example, differential equations).
Mediating variable	Suppose $x \to z \to y$. Then z is the mediating variable. For example, advertising drives awareness that drives liking that drives sales, then the effect of advertising on sales is mediated by awareness and liking, which are mediating variables. See moderating variable as it differs from mediating variable.
Moderating variable	A variable that influences the simple effect of other variables. Suppose x affects y. Then $\beta = dy/dx$ is the simple effect of x on y. A variable z is a moderating variable iff $\frac{dy}{dx} = \beta + \alpha z$, where α is the magnitude of the moderating effect of z on the $x \to y$ effect. See interaction variable and synergistic variable, which are both moderating variables. Also see mediating variable as it differs from moderating variable.
Non-Markovian variable	Its future value depends on the distant past values. See Markovian variable.
Observed variable	A variable with known values.

(*Continued*)

Table 2.1: (*Continued*)

Term	Meaning
Outcome variable	A "variable of interest" in scientific investigations (like the person of interest in police investigations). Also see dependent variable.
Pre-determined variable	A variable whose values are committed in advance. For example, television spots are purchased in the upfront market about 12–18 months in advance. The advertised brand gets to buy the spots at discounted prices, and the networks get to raise dollars needed to fund the creation of program content.
Proxy variable	A variable reflected by the latent variable. See unobserved variable and latent variable.
Random variable	A misnomer: *in a random variable, there is nothing random, and it is not even a variable.* A random variable maps an event space (that is, set of events) to a segment of the real line (that is, a Borel set). Hence, random variable is a set-valued *deterministic function.*
Regressors	A set of variables included in regression models to explain the variation in a dependent variable or predict the mean of a dependent variable. See independent variable.
Smoothed variable	A past mean of an outcome (or state) variable based on the future, current, and past information set. That is, $\hat{y}_{t\|N} = E[y_t\|I_N]$), where $N > t$. See filtered variable, smoothed variable.
State variable	A variable that is an element of the state vector, which reflects the state of the dynamic system denoted by α_t in state space models. It can be (1) an unobserved variable (for example, competitor's market share), (2) latent variable (for example, goodwill or affect), (3) unknown model parameters (for example, as in random-walk models), (4) the error terms (for example, as in AR, MA models), (5) variances of the error terms (for example, ARCH, GARCH models), or whatever you want it to be (for example, heterogeneity in parameters via hierarchical linear models).

(*Continued*)

Table 2.1: (*Continued*)

Term	Meaning
Synergistic variable	A variable that influences an outcome both directly by itself and indirectly by reinforcing other variables. For example, TV advertising directly increases sales, and it indirectly enhances the effects of online advertising by driving the website traffic, search, and engagement. Also see moderating variable.
Transformed variable	Another variable created from a given variable. For example, in the transformation $y = f(x), y$ is the transformed variable and x is the given variable.
Unobserved variable	A variable that is not directly observable. It can be a concrete variable with unavailable data (for example, tracking competitor's market share or altitude measurements of an airplane when tracking its position using radar) or an abstract construct (that is, a latent variable). In the latter case, observed variables serve as the fallible proxies for the unobserved variable. For example, goodwill or affect is not directly observed, but a survey of respondents furnishes measures such as brand awareness to serve as the proxy for goodwill or liking as the proxy for affect. Also see latent variable and state variable.
Variable	A variable is a function of an index such as time, household, brand, or city. For example, the number of units sold varies from week to week, creating the variable "sales" over time.

2.2 Diversity of dynamic models

2.2.1 Autoregressive models

The most commonly used dynamic marketing model is a first-order autoregressive model given by

$$A_t = \lambda A_{t-1} + \beta u_t + \nu_t, \quad \nu_t \sim N(0, \sigma_\nu^2) \tag{2.1}$$

where A_t denotes awareness at time t, u_t is the advertising spending, ν_t is the random error that perturbs the awareness level due to myriad factors not explicitly included in the model. The parameters β and λ denote ad effectiveness and carryover effect, respectively. The change in awareness $\nabla A_t = A_t - A_{t-1} = \beta u_t - (1-\lambda)A_{t-1} + \nu_t$ shows that awareness grows due to advertising, while it wanes in the absence of advertising $(u_t = 0)$ at the rate $(1-\lambda)$.

If managers measure awareness for its brand each week, then the ordinary least squares (OLS) can estimate Equation (2.1) to obtain the estimates of $(\beta, \lambda, \sigma_\nu^2)$. However, awareness is measured by surveying a sample of customers, so it is likely error-prone, which can be expressed as

$$Y_t = A_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma_\varepsilon^2). \tag{2.2}$$

In Equations (2.1) and (2.2), Y_t, A_t, and u_t denote the observed dependent variable, the unobserved state variable, and an independent variable, respectively. If we apply OLS to the observed data (Y_t, u_t), the resulting estimates will be biased due to the presence of measurement errors σ_ε^2. To examine this assertion, Naik et al. [2007] conduct simulation studies and report that, across low to high noise levels, ad effectiveness is over-estimated by 34–147% and carryover effect is underestimated by about 48%. Then they show how to obtain unbiased estimates, which involve controlling for measurement noise in awareness metric and estimating the dynamic model (2.1) simultaneously by using the Kalman filter, which will be described in the next section.

2.2.2 Time-varying parameters models

The ad effectiveness β itself may vary over time. For example, Naik et al. [1998] show that ad effectiveness wears out over the span of continued advertising and restores during the span of hiatus in advertising. Although Naik et al. [1998] incorporate behavioral insights to formulate and estimate the time-varying ad effectiveness, let us instead consider a simpler random walk model:

$$\beta_t = \beta_{t-1} + \nu_t, \quad \nu_t \sim N(0, \sigma_\nu^2). \tag{2.3}$$

Equation (2.3) offers a parsimonious model to capture nonmonotonic variations over time (see Vakratsas and Kolsarici [2008] and Bruce et al. [2012a]).

Alternatively, we specify a smooth cubic spline by letting the second-order variation to be of small magnitude. Specifically, let $\nabla^2\beta_t = \nu_t$, where $\nabla^2\beta_t = \nabla\nabla\beta_t = \nabla(\beta_t - \beta_{t-1}) = \nabla\beta_t - \nabla\beta_{t-1} = (\beta_t - \beta_{t-1}) - (\beta_{t-1} - \beta_{t-2}) = \beta_t - 2\beta_{t-1} + \beta_{t-2}$. Then, we get the cubic spline model:

$$\begin{bmatrix} \beta_t \\ \beta_{t-1} \end{bmatrix} = \begin{bmatrix} 2 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \beta_{t-1} \\ \beta_{t-2} \end{bmatrix} + \begin{bmatrix} \nu_t \\ 0 \end{bmatrix}. \tag{2.4}$$

The second row appends an identity ($\beta_{t-1} = \beta_{t-1}$), which allows us to represent the scalar second-order lag model as the vector first-order lag model, whose benefit will become apparent when I introduce the unifying state space form in Section 2.3. In addition, Equation (2.4) offers a simple way to impute missing values in time-series data; for this application, see Biyalogorsky and Naik [2003].

Finally, if we know the pattern of time-variation based on a theory, for example, periodic variation in ad effectiveness as in Naik et al. [1998] would be $\beta(t) = 1 + \cos(t)$, then we express it as $\frac{\partial^2\beta}{\partial t^2} = -\cos(t) = 1 - \beta$, which implies in discrete-time $\nabla^2\beta_t = 1 - \beta_{t-1}$ to get

$$\begin{bmatrix} \beta_t \\ \beta_{t-1} \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \beta_{t-1} \\ \beta_{t-2} \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \begin{bmatrix} \nu_t \\ 0 \end{bmatrix}. \tag{2.5}$$

Equation (2.5) is also linear in the state vector $(\beta_t, \beta_{t-1})'$ with only a first-order lag, revealing that higher-order lag terms can be represented equivalently as a first-order lag model in an augmented vector space.

2.2.3 VAR models

Many marketing researchers study the impact of control and independent variables on multiple dependent variables simultaneously (for example, see the review article by Dekimpe et al. [2008]). Vector Auto Regressive (VAR) models describe the market response function by combining customer response, competitive reaction, and firm decision rules. When the dynamic system includes exogenous variables, we call it the VAR-X model, where X indicates exogenous independent variables. The VAR-X model is the vector generalization of Equation (2.1), and is represented by

$$\begin{bmatrix} Y_{1t} \\ Y_{2t} \\ \vdots \\ Y_{nt} \end{bmatrix} = [\phi_{ij}] \begin{bmatrix} Y_{1t-1} \\ Y_{2t-1} \\ \vdots \\ Y_{n,t-1} \end{bmatrix} + [\beta_{ij}] \begin{bmatrix} X_{1t} \\ X_{2t} \\ \vdots \\ X_{k,t} \end{bmatrix} + \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \\ \vdots \\ \epsilon_{nt} \end{bmatrix}. \tag{2.6}$$

The elements in the $n \times n$ matrix $\Phi = \{\phi_{ij}\}$ and those in the $n \times k$ matrix $B = \{\beta_{ij}\}$ are estimated using market data, along with the variances of the error terms ϵ_i and the covariances between all pairs of (ϵ_i, ϵ_j). Equation (2.6) is also linear in the state vector $Y_t = (Y_{1t}, \ldots, Y_{nt})'$ with a first-order lag in the vector space. The vector Y_t includes multiple dependent variables such as awareness, liking, sales, prices, promotions, ad spends, distribution coverage, and distribution intensity.

In VAR-X models, all dependent variables must be observed at every instant. In contrast, for example, in the context of online reverse auctions in industrial markets, only one bidder submits a price quote at a time, while all other bidders do not bid. Different bidders submit price bids at different instants. In other words, at each t, only one element of the vector Y_t is observed, while all its other elements are not observed. To describe such dynamics, Jap and Naik [2008] develop *Partially Observed VAR* model, and they apply the Kalman filter to infer the willingness to pay of all the bidders at all the instants even when their bids were not submitted at each instant. Other extensions of VAR models include higher-order lags to get VAR(p) models with p lag terms or time-varying parameters (TVP) via Equations (2.3)–(2.5) to get *TVP-VAR* models (for such applications, see Section 4 of Koop and Korobilis [2009]). All higher-order VAR models can be expressed equivalently as first-order models in an augmented vector space (for example, by applying the ideas used in arriving at Equation (2.5)).

2.2.4 VARMAX models

The error terms in VAR-X models, namely $\epsilon_t = (\epsilon_{1t}, \ldots, \epsilon_{nt})'$ in Equation (2.6) are assumed uncorrelated over time. To be clear, they are correlated amongst themselves, but temporally independent. To relax this assumption, let the error terms in Equation (2.6) exhibit one-period moving average or MA(1) process as follows:

$$\begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \\ \vdots \\ \epsilon_{nt} \end{bmatrix} = \begin{bmatrix} \nu_{1t} \\ \nu_{2t} \\ \vdots \\ \nu_{nt} \end{bmatrix} + [\theta_{ij}] \begin{bmatrix} \nu_{1t-1} \\ \nu_{2t-1} \\ \vdots \\ \nu_{n,t-1} \end{bmatrix}, \tag{2.7}$$

where the $n \times n$ matrix $\Theta = \{\theta_{ij}\}$ contains parameters to be estimated, and $\nu_t = (\nu_{1t}, \ldots, \nu_{nt})'$ follows multivariate normal with zero mean and finite covariance matrix.

Then we augment the VAR model in (2.6) with Equation (2.7) to get VARMAX model. Other extensions of VARMAX models include higher-order lags in error terms to get MA(q) models with q lag terms (instead of $q = 1$ in Equation (2.7)) or the integrated time-series to get *ARIMAX* models (for this specification, see section 3.3 of Durbin and Koopman [2012]). All such higher-order VARMAX models also can be expressed equivalently as first-order models in an augmented vector space.

2.2.5 Dynamic factor models

Past research on how advertising works [Lavidge and Steiner, 1961, Vaughn, 1980, 1986, Barry and Howard, 1990, Vakratsas and Ambler, 1999] suggests that advertisements nudge consumers along the think–feel–do hierarchy to induce sales. Such intermediate effects — cognition (C), affect (A), and experience (E) — are unobservable constructs. While many market response models explain how advertising grows sales, they ignore the role of intermediate effects in building brands. To introduce the dual roles of advertising to boost sales and build brands, Bruce et al. [2012b] formulate a dynamic factor model of advertising. The classical hierarchy of effects suggests that advertising triggers one of the three intermediate factors to initiate the sequence, and the last in the sequence drives sales.

Classical hierarchical $E \to C \to A$

$$\begin{bmatrix} C_t \\ A_t \\ E_t \\ S_t \end{bmatrix} = \begin{bmatrix} 0 & 0 & \gamma_{13} & 0 \\ \gamma_{21} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & \gamma_{42} & 0 & 0 \end{bmatrix} \begin{bmatrix} C_{t-1} \\ A_{t-1} \\ E_{t-1} \\ S_{t-1} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \beta_3 g\left(u_t\right) \\ 0 \end{bmatrix} + \begin{bmatrix} w_{1t} \\ w_{2t} \\ w_{3t} \\ w_{4t} \end{bmatrix}. \quad (2.8)$$

Consider Equation (2.8) in the context of the $E \to C \to A$ hierarchy, for example. Advertising effect (β_3) triggers the experience E_t, then the prior experience E_{t-1} influences the current cognition C_t (via γ_{13}), the prior cognition C_{t-1} drives the current affect A_t (via γ_{21}), and the prior affect A_{t-1} induces brand sales S_t (via γ_{42}). The error terms $(w_{1t}, w_{2t}, w_{3t}, w_{4t})' \sim N(0, W)$ represent the specification errors in the three factor and brand sales equations. The function $g(u) = \sqrt{u}$ captures the diminishing returns to advertising. Other permutations of experience, cognition, and affect can be empirically tested to discover the operating sequence (see Bruce et al. [2012a]).

In contrast, Vakratsas and Ambler [1999] suggest that advertising ignites all three factors simultaneously via $(\beta_1, \beta_2, \beta_3)'$ then all three factors jointly drive sales via $\gamma_{41}, \gamma_{42}, \gamma_{43}$ and brand purchases reinforce experience through γ_{34}. Equation (2.9) describes this process.

Vakratsas–Ambler model

$$\begin{bmatrix} C_t \\ A_t \\ E_t \\ S_t \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \gamma_{34} \\ \gamma_{41} & \gamma_{42} & \gamma_{43} & 0 \end{bmatrix} \begin{bmatrix} C_{t-1} \\ A_{t-1} \\ E_{t-1} \\ S_{t-1} \end{bmatrix} + \begin{bmatrix} \beta_1 g(u_t) \\ \beta_2 g(u_t) \\ \beta_3 g(u_t) \\ 0 \end{bmatrix} + \begin{bmatrix} w_{1t} \\ w_{2t} \\ w_{3t} \\ w_{4t} \end{bmatrix}. \quad (2.9)$$

For both these views, advertising does not grow sales directly (that is, $\beta_4 = 0$), but only indirectly via the intermediate factors. Hence, the classical hierarchy-of-effects literature presents the “pure” brand-building view of advertising. More importantly, the classical hierarchies and the Vakratsas–Ambler model ignore the dynamic effects (that is, $\gamma_{ii} = 0$).

Hence, Bruce et al. [2012a] integrate the elements of the dynamic sales response model (that is, sales dynamics via γ_{44} and advertising effect via β_4), the classical hierarchy models, and the Vakratsas–Ambler model. In addition, they extend the marketing literature by incorporating:

(i) *Intermediate factor dynamics* ($\gamma_{11} \neq 0$; $\gamma_{22} \neq 0$; $\gamma_{33} \neq 0$; $\gamma_{44} \neq 0$). Like sales carryover effect, current cognition carries over into future periods with a weekly attrition rate $(1 - \gamma_{11})$. Similarly, γ_{22} and γ_{33} capture affect and experience dynamics, respectively.

(ii) *Purchase reinforcement* ($\gamma_{14} \neq 0$; $\gamma_{24} \neq 0$; $\gamma_{34} \neq 0$). For example, γ_{34} measures the purchase reinforcement of current experience due to the link $S \rightarrow E$, as Vakratsas and Ambler [1999] hypothesized. Additionally, they hypothesize the presence of purchase reinforcements on cognition (γ_{14}) and affect (γ_{24}).

(iii) *Advertising grows sales and builds brands simultaneously* ($\beta_1 \neq 0$; $\beta_2 \neq 0; \beta_3 \neq 0$; $\beta_4 \neq 0$) The parameters $(\beta_1, \beta_2, \beta_3, \beta_4)'$ measure the effects of advertising GRPs on all three intermediate factors and brand sales. The first three advertising effects $(\beta_1, \beta_2, \beta_3)'$ ignite all intermediate factors to build brand values; also advertising effect (β_4) grows sales volume directly. Together, brand values and sales volume create the intangible and tangible effects of advertising, respectively.

They refer to this augmented model as the *Integrated Hierarchy* of advertising. Equation (2.10) shows the integrated $E \rightarrow C \rightarrow A$ hierarchy.

Integrated $E \rightarrow C \rightarrow A$ hierarchy

$$\begin{bmatrix} C_t \\ A_t \\ E_t \\ S_t \end{bmatrix} = \begin{bmatrix} \gamma_{11} & 0 & \gamma_{13} & \gamma_{14} \\ \gamma_{21} & \gamma_{22} & 0 & \gamma_{24} \\ 0 & 0 & \gamma_{33} & \gamma_{34} \\ 0 & \gamma_{42} & 0 & \gamma_{44} \end{bmatrix} \begin{bmatrix} C_{t-1} \\ A_{t-1} \\ E_{t-1} \\ S_{t-1} \end{bmatrix} + \begin{bmatrix} \beta_1 g(u_t) \\ \beta_2 g(u_t) \\ \beta_3 g(u_t) \\ \beta_4 g(u_t) \end{bmatrix} + \begin{bmatrix} w_{1t} \\ w_{2t} \\ w_{3t} \\ w_{4t} \end{bmatrix}. \quad (2.10)$$

To link the factor dynamics to observed metrics, they measure a battery of n mindset metrics in each week t, denoted by x_{it}, along with the observed sales volume y_t with the mean sales level S_t as follows:

$$\begin{bmatrix} x_{1t} \\ x_{2t} \\ \\ x_{jt} \\ \\ x_{j't} \\ \\ x_{nt} \\ y_t \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \lambda_{21} & \lambda_{22} & \lambda_{23} & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & 0 \\ & & & \vdots \\ 0 & 0 & 1 & 0 \\ & & & \vdots \\ \lambda_{n,1} & \lambda_{n,2} & \lambda_{n,3} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} C_t \\ A_t \\ E_t \\ S_t \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ \vdots \\ \\ \\ \\ \\ \varepsilon_{nt} \\ \varepsilon_{n+1,t} \end{bmatrix}, \qquad (2.11)$$

where the four rows have only ones and zeros for identification and naming of the factors. Specifically, one of the items for each of the factors (x_{1t} for C_t, x_{jt} for A_t, $x_{j't}$ for E_t, y_t for S_t), usually prototypical variables that unambiguously justify the naming of that factor, serves the purpose of identification, which is achieved by setting unity for corresponding named factor and zeros for other factors (as shown in Equation (2.11)).

Furthermore, if a factor (C_t, A_t, E_t) reflects a measurement item (x_{it}) then it gets a nonzero loading coefficient (λ_{ij}) to be estimated using data. The link matrix in (2.11) permits both the confirmatory factor analysis (that is, known link matrix) and the exploratory factor analysis (that is, unknown link matrix). If factor composition is known, then the values of λ are set according to this knowledge (with zeros for items that do no belong to the factor). If factor composition is unknown, then all the values of λ are estimated (except for the identifying restrictions via the unit row vectors) and significance testing guides the inference on the most probable factor structure that corroborates with market data.

The vector of ε_t denotes measurement errors in the metrics, which serve as fallible proxies for the unobserved factors. This dynamic factor model can be estimated via the Kalman filter–smoother algorithms (see Bruce et al. [2012b], Du and Kamakura [2015], Rutz and Sonnier [2011], and Hasegawa et al. [2012]).

The next five models are formulated in continuous time, where the instant t marches along the positive real line $[0, \infty)$ rather than in the discrete index set $1, 2, \ldots, T$ as in the preceding models.

2.2.6 Nerlove-Arrow model

To obtain the continuous-time version of Equation (2.1), consider its deterministic part, $A_t = \lambda A_{t-1} + \beta u_t$, which equals $A_t - A_{t-1} = \lambda A_{t-1} - A_{t-1} + \beta u_t$, which can be expressed as $\nabla A_t = -(1-\lambda)A_{t-1} + \beta u_t$ and, upon taking limits as the interval between the two time points vanish, we obtain the classical Nerlove and Arrow [1962] model:

$$dA(t)/dt = -\delta A(t) + \beta u(t) \tag{2.12}$$

where $\delta = (1 - \lambda)$ denotes the forgetting rate. Equation (2.12) clarifies the meaning of Equation (2.1). Specifically, awareness decays in the absence of advertising (that is, $u(t) = 0$) and grows due to advertising over time. In what follows, I suppress the time argument in the variables for brevity when its presence is apparent from the context (for example, $dA/dt = -\delta A + \beta u$); parameters are constant over time unless stated otherwise. Mahajan et al. [1984] review this model and its variations in the context of awareness formation of new products.

We can augment the above model to incorporate multiple media. When managers use multi-media communications, the combined effect exceeds the sum of individual effects due to synergies. To incorporate synergies in Equation (2.12), Naik and Raman [2003] formulate the *Integrated Marketing Communications* (IMC) model as follows:

$$dA(t)/dt = -\delta A(t) + \beta_1 u(t) + \beta_2 v(t) + \kappa\, u(t) \times v(t), \tag{2.13}$$

where (β_1, β_2) denote the effectiveness of the two advertising media (u, v) respectively, and κ measures the synergy between them (that is, the joint impact of online and offline advertising, for instance, over and beyond their direct effects via (β_1, β_2)).

To understand the synergy effect, imagine κ increases marginally from zero to a finite number. We can rearrange the right-hand side of (2.13) as follows: $-\delta A + (\beta_1 + 0.5\,\kappa v)u + (\beta_2 + 0.5\,\kappa u)v$. Then we see that the effectiveness of u increases from β_1 to $(\beta_1 + 0.5\,\kappa v)$ and that for

v increases from β_2 to $(\beta_2 + 0.5\,\kappa u)$. Thus the IMC model captures the essence of synergy: each advertising medium not only increases awareness directly, but also enhances the effectiveness of the *other* medium indirectly. In Section 6.1, I will explain how to solve the associated control problem to not only derive the optimal budget allocation, but also discover a counter-intuitive result that managers should allocate *more* than fair share to the *less* effective medium.

If we considered both the deterministic and random terms in Equation (2.1), then the Nerlove–Arrow model becomes

$$dA(t) = (-\delta A(t) + \beta u(t))dt + \sigma_\nu dW(t), \tag{2.14}$$

where $W(t)$ denotes the standard Wiener process, which is also known as the Brownian motion that generalizes the discrete-time random walk model (that is, Equation (2.3)) to continuous-time. Consistent with the role of random errors in Equation (2.1), the increment in random Wiener process, $dW(t)$, perturbs the awareness dynamics due to myriad factors not explicitly included in the model. This perturbation captures the dynamic uncertainty and is normally distributed with mean zero and variance σ_ν^2 at each instant t. This Wiener increment dW is analogous to the normal error term in time series models. $W(t)$ is a continuous function of time at every instant although not differentiable at any instant. Hence the standard calculus does not apply, and Ito's stochastic calculus becomes necessary to analyze models as in Equation (2.14). For such analyses, see Raman and Naik [2004] who incorporate dynamic uncertainty in the IMC model and investigate its effects on the optimal allocation and brand profit. In Section 6, I will explain how to solve the stochastic control problems. (To learn about Brownian motion and Ito's calculus, Jazwinski [1970], Malliaris and Brock [1982], and Grigoriu [2002] provide a gentle introduction to those advanced subjects.)

2.2.7 Vidale–Wolfe model

One of the earliest known continuous-time sales-advertising model, due to Vidale and Wolfe [1957], is given by

$$dS(t)/dt = \beta u(t)[M(t) - S(t)] - \delta S(t), \tag{2.15}$$

where (S, M) denote brand sales and market size. Equation (2.15) says that the growth in brand sales decreases proportional to the sales level at the rate δ in the absence of advertising, whereas advertising influences the untapped market ($M - S$) to generate sales growth.

This model has been used extensively to investigate whether brand managers should use "pulsing" advertising strategy, that is, advertise for a few weeks, then stop advertising for a few weeks, and repeat this pattern over time (see, for example, Sasieni [1971], Sethi [1977], Mahajan and Muller [1986], Park and Hahn [1991], Feinberg [1992], Mesak [1992], Bronnenberg [1998], Naik et al. [1998], Feinberg [2001], Freimer and Horsky [2012]). Because I do not analyze this model in Section 6, I refer the readers to Sethi [1977], who fully characterizes the optimal advertising strategy for Vidale–Wolfe model. His results reveal that pulsing is *not* the optimal strategy. Moreover, Feinberg [2001] proves that the optimal pulsing does not arise in a broad class of models with sales growth $dS/dt = g(u)f(S) - \delta S$, where the advertising response function $g(u)$ can exhibit diminishing returns or an S-shaped threshold like behavior.

Besides the pulsing literature, Vidale–Wolfe model has found applications in analyzing competitive markets because of its property of logical consistency (see, for example, Chintagunta and Vilcassim [1992], Fruchter and Kalish [1997], Fruchter [1999], Naik et al. [2005]). To understand this property, let us normalize the market size to unity so that Equation (2.15) becomes $\dot{x} = dx/dt = \beta u(1 - x) - \delta x$, where x denotes market share. When a competing brand advertises, the focal brand loses its share at a faster rate. To incorporate this effect, let the brand's attrition rate δ be proportional to the competitor's advertising, that is, $\delta_1 = \beta_2 u_2$, where the subscripts refer to the two competing brands. Then, the dynamic competition for market shares is given by the coupled differential equations:

$$\begin{aligned} dx_1(t)/dt &= \beta_1 u_1(t)(1 - x_1(t)) - \beta_2 u_2(t) x_1(t), \\ dx_2(t)/dt &= \beta_2 u_2(t)(1 - x_2(t)) - \beta_1 u_1(t) x_2(t). \end{aligned} \tag{2.16}$$

Summing the left-hand sides, we get $dx_1/dt + dx_2/dt = d(x_1 + x_2)/dt = 0$ because $x_1(t) + x_2(t) = 1$ for every instant $t \in [0, \infty)$. Summing the right-hand sides, we get $[\beta_1 u_1(1-x_1) - \beta_2 u_2 x_1] + [\beta_2 u_2(1-x_2) - \beta_1 u_1 x_2] = \beta_1 u_1 - (\beta_1 u_1 + \beta_2 u_2)x_1 + \beta_2 u_2 - (\beta_1 u_1 + \beta_2 u_2)x_2 = F(1-x_1-x_2)$, where $F(t) = \beta_1 u_1(t) + \beta_2 u_2(t)$ denotes the marketing force at time t. Then, for *any* nonzero function $F(t)$ the right- hand side equals equals zero because $x_1(t) + x_2(t) = 1$, guaranteeing the logical constraint that the shares sum to unity. This property, known as the logical consistency, does not hold for many models (for example, consider extending the Nerlove–Arrow model for two competing brands).

All models thus far are linear in the state variables; the final three are nonlinear dynamic models.

2.2.8 Bass model

Bass [1969] describes how new product categories, innovations, or technologies diffuse over time, and it is one of the 10 most cited papers in the 50-year history of *Management Science*. The model incorporates the interaction between the buyers and the untapped market, due to word of mouth effects, and is specified by the differential equation:

$$\frac{dN(t)}{dt} = \left(p + \frac{q}{M}N(t)\right)(M - N(t)), \tag{2.17}$$

where $N(t)$ is the *cumulative* number of buyers up to time t, and $(p,\, q)$ are the coefficients of innovation and imitation, respectively. By expanding the right- hand side, we observe that the first term represents the effect of untapped market as in Vidale–Wolfe model, and the second term, $N \times (M - N)$ captures the interaction between the buyers and the untapped market. The presence of an interaction term renders the model nonlinear in the state variable N.

Because the Bass model does not contain decision variables, Bass et al. [1994] propose the Generalized Bass model:

$$\frac{dN(t)}{dt} = \left(p + \frac{q}{M}N(t)\right)(M - N(t))F(t), \tag{2.18}$$

where $F(t)$ denotes the marketing force due to a brand's advertising or price inputs. By assuming $F(t) = 1 - \alpha\{[(\dot{p}(t)/p(t))] + \beta[(\dot{a}(t)/a(t))]\}$, where $\dot{z} = dz/dt, (p,\, a)$ are price and advertising inputs, $(\alpha,\, \beta)$ are the sensitivity parameters, Fruchter and Van den Bulte [2011] solve the induced control problem and show that the optimal advertising increases monotonically over time (even in the presence of decreasing prices). This monotonicity result holds for *all* first-order dynamic models with a single state variable regardless of the nature of dynamics [as shown more generally by Hartl, 1987], *unless delayed state variables drive the dynamics* [as shown more recently by Aravindakshan and Naik, 2015].

2.2.9 Sethi model

Sethi [1983] parsimoniously combines both the roles of untapped market and the interaction effect via the differential equation:

$$\frac{dS(t)}{dt} = \beta u(t)\sqrt{M - S(t)} - \delta S(t), \tag{2.19}$$

where the square root introduces nonlinearity in the state variable. To interpret the square root term, let us denote the market penetration by $S/M = x$ and then normalize the market size to unity. Consequently, Equation (2.19) becomes $dx/dt = \beta u(t)\sqrt{1-x} - \delta x$. Sorger [1989] shows that $\sqrt{1-x} \approx (1-x) + x(1-x)$. Also see Erickson [2003, p. 24]. The first term $(1-x)$ represents the untapped market, and the second term $x(1-x)$ captures the interaction effect due to word of mouth.

As Section 6 will show, this square root formulation facilitates the derivation of closed-form analytical solutions to the deterministic control problem, stochastic control problem, and competitive differential games. For example, Prasad and Sethi [2009] extend this model to the IMC context, which is given by

$$\begin{aligned} \frac{dx(t)}{dt} &= F(t)\sqrt{1 - x\,(t)} - \delta x(t) \\ F(t) &= \beta_1 u + \beta_2 v + \kappa\sqrt{u \times v}, \end{aligned} \tag{2.20}$$

where $F(t)$ represents the force of marketing communications due to the two interactive media $(u(t),\, v(t))$ that increases both the market penetration and word of mouth effects embodied in $\sqrt{1-x}$. Prasad and Sethi [2009] derive the optimal closed-loop allocation, which provides the optimal decision rule to spend on each medium based on the prevailing level of $x(t)$. More importantly, they show that managers should allocate more than fair share of the total spending to the less effective medium as synergy increases. This result generalizes Naik and Raman's [2003] result to a different dynamic specification, reassuring us that the inverse allocation arises not due to the choice of model dynamics, but due to the super additivity of marketing communications.

2.2.10 N-brand dynamic oligopoly model

Similar to Equation (2.16), which extends the Vidale–Wolfe model to duopoly markets, Sorger [1989] extends the Sethi [1983] model to duopoly markets. The resulting model satisfies the logical consistency property. Chintagunta and Vilcassim [1992] and Chintagunta and Jain [1995] furnish empirical support for Sorger's model. Naik et al. [2008] further extend Sorger's formulation to incorporate competition amongst N brands to build awareness $A_i(t)$ as follows:

$$\begin{aligned} \frac{dA_i(t)}{dt} &= \underbrace{\beta_i u_i(t)\sqrt{M(t) - A_i(t)}}_{\text{Gain due to own action}} \\ &\quad - \underbrace{\Sigma_{j=1,j\neq i}^{N} \alpha_{ij} u_j(t)\sqrt{M(t) - A_j(t)}}_{\text{Loss or gain from others' actions}}, i = \{1, 2, \ldots, N\}, \end{aligned} \tag{2.21}$$

where β_i is the effectiveness of own action u_i, α_{ij} is the effectiveness of jth brand's action u_j on brand i. Equation (2.21) says that a focal brand's advertising $u_i(t)$ builds its awareness, while other brands' advertising $u_j(t)$ may detract it from achieving its awareness growth when $\alpha_{ij} > 0$. However, if $\alpha_{ij} < 0$, then brand j's advertising (for example, comparative advertisements) creates confusion in consumers' minds and increases awareness of the focal brand. Besides such confusion effects, we get market expansion if $M(t)$ grows over time. Finally, if logical consistency is necessary (for example, when using market share data), researchers can set $\alpha_{ij} = \beta_j/(N - 1)$ and $M = 1$ to ensure $\sum_{i=1}^{N} A_i(t) = 1$ for every t.

Naik et al. [2008] design an extended Kalman filter (EKF) to estimate this N-brand dynamic oligopoly model using market data for five car brands over time. In addition, for any feasible parameter value (that is, not just the estimated parameters), they derive the optimal closed-loop Nash equilibrium strategies for every brand. The empirical results furnish strong support for the proposed model in terms of both goodness-of-fit in the estimation sample and cross-validation in the out-of-sample data. More importantly, the estimation method offers managers a systematic way to estimate ad effectiveness and forecast awareness levels for their particular brands as well as competitors' brands. Most importantly, the analytical solution to the associated control problem reveals — contrary to the proportional-to-sales or competitive parity heuristics — the counter-intuitive insight: large (small) brands should advertise proportionally less (more) than small (large) brand's advertising.

2.3 Unifying framework: state space models

All Markovian dynamic models are unified within the state space form. To illustrate this point, I present the linear state space model:

$$Observation\ equation: \quad Y_t = Z_t\alpha_t + c_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, H_t) \tag{2.22}$$

$$Transition\ equation: \quad \alpha_t = T_t\alpha_{t-1} + d_t + \nu_t, \quad \nu_t \sim N(0, Q_t) \tag{2.23}$$

Equation (2.23) is called the transition equation, which represents the dynamics of the state vector α_t. Equation (2.22) is known as the observation equation, which connects the state vector α_t to the observation vector Y_t. The number of observed variables in Y_t and those in the state vector α_t need not be the same. Hence the link matrix Z_t is rectangular in general. Let m denote the number of observed variables, and n be the number of unobserved state variables. Then the various vectors and matrices in (2.22) and (2.23) have the names and dimensions as given in Table 2.2.

In principle, all vectors and matrices in (2.22) and (2.23) can vary over time as direct functions of time or through covariates X_t or via past observations Y_{t-1}. For example, covariates can enter via the observation

Table 2.2: Names and dimensions for vectors and matrices in state space models.

Notation	Vector or matrix	Name	Dimension
Y	Vector	Observation vector	m × 1
α	Vector	State vector	n × 1
T	Matrix	Transition matrix	n × n
$T(\alpha)$	Vector-valued function	Transition function	n × 1 outputs; n × 1 arguments
c	Vector	Drift vector (in observation)	m × 1
d	Vector	Drift vector (in transition)	n × 1
Z	Matrix	Link matrix (from state to observation)	m × n
ε	Vector	Observation errors	m × 1
ν	Vector	Transition errors	n × 1
H	Matrix	Observation noise matrix	m × m
Q	Matrix	Transition noise matrix	n × n

drift $c_t = X_t\gamma$ or heteroscadesticity $H_t = \exp(X_t\gamma)$. Or we formulate conditionally Gaussian models by specifying the elements of link or transition matrices (Z, T) to depend on past observations; that is, $T_{ij} = g(Y_{t-1})$. Furthermore, when we generalize the link and transition matrices to the link and transition *functions*, respectively, we get the nonlinear state space model:

$$Observation\ equation: \quad Y_t = Z_t(\alpha_t) + c_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, H_t) \tag{2.24}$$

$$Transition\ equation: \quad \alpha_t = T_t(\alpha_{t-1}) + d_t + \nu_t, \quad \nu_t \sim N(0, Q_t) \tag{2.25}$$

If necessary, we can further generalize the above nonlinear state space model by specifying the observation equation via $Y_t = Z_t(\alpha_t, c_t, \varepsilon_t)$, state transition via $\alpha_t = T_t(\alpha_{t-1}, d_t, \nu_t)$, and the error terms $\varepsilon_t \sim WS(0, H_t)$ and $\nu_t \sim WS(0, Q_t)$. That is, the arguments need not be additively separable nor the error terms normally distributed. The distribution WS$(\cdot)$ stands for a "wide-sense" distribution with finite second moments (H, Q) and its density exhibits any shape (that is, need not be symmetrical or Gaussian).

Similar to the above discrete-time state space models, we can specify their continuous-time versions. For example, consider the deterministic part of (2.23); then its continuous-time linear state transition is given by

$$\frac{d\alpha(t)}{dt} = T(t)\alpha(t) + B(t)u(t), \tag{2.26}$$

where $\alpha(t)$ is a n dimensional state vector, $u(t)$ is k dimensional vector of controls or covariates, and (T, B) are time-varying conformable matrices. The unobserved state vector connects to the observations via Equation (2.22). The term $d(t) = B(t)u(t)$ represents the drift vector. In case we need to incorporate continuous-time uncertainty, we extend Equation (2.26) as $d\alpha = [T(t)\alpha(t) + B(t)u(t)]dt + Q(t)^{1/2}dW$, where the vector $W(t)$ contains n standard Wiener process, and $Q(t)^{1/2}$ represents the Cholesky factor of $Q(t)$. Accordingly, we obtain further nonlinear extensions.

The above state space forms subsume or nest diverse time-series models. To exemplify this claim, I will cast one discrete-time TVP model given by Equations (2.1)–(2.3) into state space form (2.22) and (2.23) and one continuous-time Vidale–Wolfe model given by Equation (2.15) into state space form (2.26).

Consider first the time-varying parameter model specified by Equations (2.1)–(2.3). Equation (2.2) is the observation equation, which is scalar, with $Z = (1, 0)$ as a row vector, $\alpha_t = (A_t, \beta_t)'$ as a column vector, $c_t = 0$, and $H_t = \sigma^2_\varepsilon$. Equations (2.1) and (2.3) form a system of equations as follows:

$$\begin{bmatrix} A_t \\ \beta_t \end{bmatrix} = \begin{bmatrix} \lambda & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} A_{t-1} \\ \beta_{t-1} \end{bmatrix} + \begin{bmatrix} \beta_t u_t \\ 0 \end{bmatrix} + \begin{bmatrix} \nu_{1t} \\ \nu_{2t} \end{bmatrix},$$

where the transition matrix T = diag$(\lambda, 1)$, the vector drift $d_t = (\beta_t u_t, 0)'$, and the covariance matrix Q$_t$=diag$(\sigma^2_{\nu 1}, \sigma^2_{\nu 2})$. Thus, the time-varying parameter model Equations (2.1)–(2.3) are nested in the state-space form stated in Equations (2.22) and (2.23).

Next consider the continuous-time IMC model. By comparing Equations (2.15) and (2.26), we observe that $\alpha(t) = S(t), T(t) = -\beta u(t) - \delta$, and $B(t) = \beta M(t)$. Thus, the continuous-time Vidale–Wolfe model is nested in the state-space form stated in Equation (2.18). The other eight models — indeed any other Markovian dynamic model — can be cast into the state space form, including dynamic *spatial* models (see, for example, Aravindakshan et al. [2012]).

I close this subsection by presenting an example of a dynamic model that does *not* belong to the class of Markovian dynamics. Standard awareness formation models (for example, Equation (2.1) or (2.12)) imply that awareness decline commences instantaneously. In contrast, Aravindakshan and Naik [2011] investigate memory effects by allowing the possibility that awareness decline can be delayed due to the memory for ads. This change converts an ordinary differential equation (ODE) to a *delayed* differential equation (DDE). DDEs are a special class of differential equations where the argument is allowed to be "delayed," and the resulting models exhibit non-Markovian dynamics. In Aravindakshan and Naik [2011], awareness evolves according to the delay differential equation

$$\frac{dA(t)}{dt} = \beta\sqrt{u(t)} - \delta A(t - \tau), \tag{2.27}$$

with the initial function $A_0(t) = A_0$ over the interval $[-\tau, 0]$. In Equation (2.27), awareness $A(t)$ and advertising $u(t)$ are non-negative, the square root function captures the diminishing returns to advertising, and the parameters (β, δ, τ) belong to the non-negative octant. When $\tau \neq 0$, the awareness dynamics becomes non-Markovian; that is, the future and the past are *not* independent given the present.

By analyzing the induced non-Markovian control problem, Aravindakshan and Naik [2015] show that the optimal trajectory $u^*(t)$ exhibits periodicity over time, namely, $u^*(t) = u^*(t+t')$ for some future time t'. Specifically, for different values of (β, δ), Figure 2.1 presents truly cyclic patterns of optimal advertising with different levels of ad spending in different weeks. This result — the optimality of multi-level pulsing — is

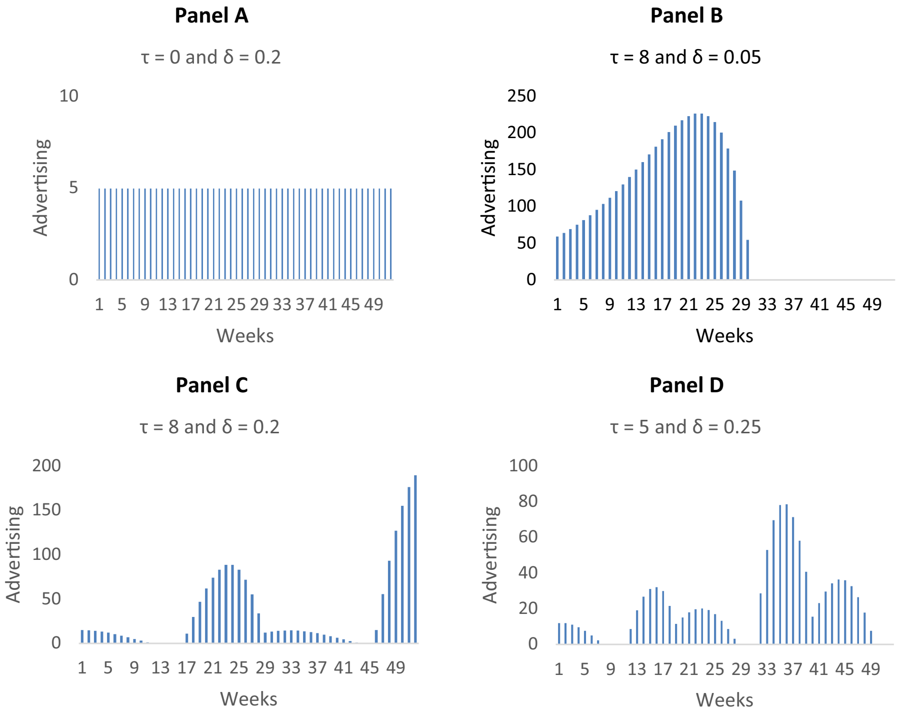

Figure 2.1: Cyclic optimal advertising policies.

novel especially because only on–off pulsing patterns have been shown to be optimal (for example, Naik et al. [1998], Dubé et al. [2005], Freimer and Horsky [2012]) in the extant literature over the past four decades since Sasieni [1971]. It is important to recognize that even such non-Markovian dynamic models can be estimated by transforming them into a proper state space form (see Aravindakshan and Naik [2011]).

2.4 Advantages of state space models

Statistical and econometric estimation approaches make strong assumptions on data or model. For example, they require no unit root (that is, stationary time series), no missing values, no irregularly spaced observations, no unobserved or partially observed sequences, constancy of parameters over time, among others. Recent literature, developed by frequentists and Bayesians, relaxes these assumptions and offers specialized algorithms to handle missing observations, irregularly spaced observations, unobserved or partially observed sequences, time-varying coefficients, and so on. As a result, the literature consists of a *plethora* of methods: one for estimating AR models, another one for MA models, yet another for time-varying coefficients, and so on. Seemingly as many methods as there are models.

In contrast, rooted in dynamic systems theory in mathematics and used in engineering, the state space model unifies all of these special cases over and beyond the 10 examples I presented. The value of this unification is that *ad hoc* algorithms are not needed on a model-by-model basis. A *single* algorithm (rather more correctly, a theorem), known as the Kalman filter, given by one set of recursive equations (namely Equations (3.1), (3.2), (3.3), (3.5), and (3.6) to be presented later) can estimate all of these diverse models because they can be shown to be the special cases of the state space form presented in Equations (2.22) and (2.23) or more generally Equations (2.24) and (2.25)
.

Consequently, state space models offer several practical advantages:

(i) *exact* likelihood function can be computed to obtain parameter estimates, infer statistical significance, and select among model specifications;

(ii) *multivariate* outcomes are handled as easily as univariate time-series;

(iii) *missing values* do not require special algorithms for imputation;

(iv) *unequally spaced* time-series observations pose no additional challenges;

(v) *unobserved* variables such as goodwill or brand equity can be incorporated;

(vi) *time-varying coefficients* and *nonstationarity* can be specified;

(vii) *heterogeneity* via random coefficients can be introduced;

(viii) *coupling* and correlations across equations can be specified;

(ix) *normative decision-making* can be integrated with the econometric analyses;

(x) a *common* algorithm, based on Kalman filter recursions, can be used to analyze and estimate diverse model specifications.

The last feature offers the most compelling advantage for using state space models. Because state space models nest any dynamic specification — not just the above 10 models — we need only one algorithm, known as the Kalman filter, to estimate them. This unification not only offers a mathematical harmony underlying the diversity of dynamic models, but also furnishes a common estimation algorithm via the Kalman filter, which we next present.

3

State Estimation

Dynamic models can be estimated using the maximum likelihood (see Harvey [2001]), expectation–maximization (EM) algorithm (see Shumway and Stoffer [2011]), or Bayesian estimation (see Harrison and West [2013]). The maximum likelihood estimation requires the Kalman filter recursions, whereas EM or Bayesian estimations require the Kalman filter recursions and the Kalman smoother recursions. Hence, in this section, I derive both the Kalman filter and smoother recursions by first considering the linear state space model in Equations (2.22) and (2.23) and then its nonlinear extension in Equations (2.24) and (2.25).

3.1 Derivation of the Kalman filter

Consider Equations (2.22) and (2.23) in which the system matrices depend on the parameter vector θ. That is, the system matrices are $Z(\theta)$, $T(\theta)$, $c(\theta)$, $d(\theta)$, $H(\theta)$, and $Q(\theta)$. Note that θ can be time-varying; other elements of the system matrices Z, T, c, d, H, and Q also can be time-varying; even the dimensions of the system matrices can be time-varying (in a conformable manner). For the sake of exposition, however, we assume θ is fixed and ignore these two generalizations, but note that the filter recursions we derive in this section remain valid for the general case. Our goal is to derive the best estimate of the state vector α_t given the sequential (that is, one at a time) arrival of the observation vectors $\{Y_1, Y_2, \ldots, Y_t, \ldots, Y_N\}$, where N is the sample size.

Let us denote $a_{t|t-1} = E[\alpha_t|I_{t-1}]$ to be the mean of the state vector at time t based on information up to and including time $t - 1$. In other words, I_{t-1} denotes observations up to $t - 1$, which can be expressed as the information set $I_{t-1} = \{Y_1, Y_2, \ldots, Y_{t-1}\}$. Also let $a_t = E[\alpha_t|I_t]$ denote the mean of the state vector at time t based on information at time t, which includes Y_t. In other words, I_t denotes observations up to time t, representing the information set $I_t = I_{t-1} \cup \{Y_t\} = \{Y_1, \ldots, Y_{t-1}, Y_t\}$. Similarly, let $P_{t|t-1} = \mathrm{Cov}[\alpha_t|I_{t-1}]$ and $P_t = \mathrm{Cov}[\alpha_t|I_t]$ denote the covariance matrices of the state vector based on information up to $t - 1$ and t, respectively.

At time $t - 1$, taking expectation of Equation (2.23), we obtain $E[\alpha_t|I_{t-1}] = T_t E[\alpha_{t-1}|I_{t-1}] + d_t + E[\nu_t|I_{t-1}]$, which yields

$$a_{t|t-1} = T_t a_{t-1} + d_t. \tag{3.1}$$

Similarly, we apply the variance operator to Equation (2.23) to get $\mathrm{Cov}[\alpha_t|I_{t-1}] = T_t \mathrm{Cov}[\alpha_{t-1}|I_{t-1}] T_t' + \mathrm{Cov}[\nu_t|I_{t-1}]$, which yields

$$P_{t|t-1} = T_t P_{t-1} T_t' + Q_t. \tag{3.2}$$

Suppose we get the new observation Y_t. Now let us update the above mean and covariance expressions using the new information set $I_t = I_{t-1} \cup \{Y_t\}$. Specifically, we update the mean vector proportional to the forecasting errors:

$$a_t = a_{t|t-1} + K_t(Y_t - \hat{Y}_t), \tag{3.3}$$

where $\hat{Y}_t = E[Y_t|I_{t-1}] = E[(Z_t\alpha_t + c_t + \varepsilon_t)|I_{t-1}] = Z_t a_{t|t-1} + c_t$, and the optimal time-varying matrix K_t is to be determined.

Equation (3.3) embodies the following rationale: (1) the posterior mean a_t is estimated by a linear combination of the observed Y_t (that is, $a_t \propto K_t Y_t$); and (2), more importantly, the prior mean $a_{t|t-1}$ is updated by the amount proportional to the forecast error $(Y_t - \hat{Y}_t)$ to arrive at an estimate of the posterior mean a_t.

To determine the "best" matrix K_t, define the loss function as the squared differences between the true and estimated values of the n elements of the state vector. This loss function is given by

$$\begin{aligned} J_t &= E\left[(\alpha_{1t} - a_{1t})^2 + (\alpha_{2t} - a_{2t})^2 + \cdots + (\alpha_{nt} - a_{nt})^2\right] \\ &= E[\nu_t'\nu_t] = E[\mathrm{Tr}(\nu_t'\nu_t)] = \mathrm{Tr}(E[\nu_t\nu_t']) \qquad (3.4) \\ &= \mathrm{Tr}(P_t), \end{aligned}$$

where the third equality follows by noting $\nu_t'\nu_t$ is a scalar; the fourth one applies the cyclic permutation within a trace operation (that is, $\mathrm{Tr}(ABC) = \mathrm{Tr}(BCA) = \mathrm{Tr}(CAB)$) and then switches the trace and expectation operators; and the last one sums the diagonal of the matrix P_t (to be determined).

To determine P_t we first express the transition error vector as follows:

$$\begin{aligned} \nu_t &= \alpha_t - a_t \\ &= \alpha_t - (a_{t|t-1} + K_t(Y_t - \hat{Y}_t)) \\ &= (\alpha_t - a_{t|t-1}) - K_t(Y_t - (Z_t a_{t|t-1} + c_t)) \\ &= \nu_{t|t-1} - K_t(Z_t\alpha_t + c_t + \varepsilon_t - Za_{t|t-1} - c_t) \\ &= \nu_{t|t-1} - K_tZ_t(\alpha_t - a_{t|t-1}) - K_t\varepsilon_t \\ &= \nu_{t|t-1} - K_tZ_t\nu_{t|t-1} - K_t\varepsilon_t \\ &= (I - K_tZ_t)\nu_{t|t-1} - K_t\varepsilon_t, \end{aligned}$$

where $v_{t|t-1} = \alpha_t - a_{t|t-1}$. Then, we evaluate$P_t = E[\nu_t\nu_t']$ by simplifying the cross product terms as follows:

$$\begin{aligned} P_t &= E[\nu_t\nu_t'] \\ &= E[\{(I - K_tZ_t)\nu_{t|t-1} - K_t\varepsilon_t\}\{(I - K_tZ_t)\nu_{t|t-1} - K_t\varepsilon_t\}'] \\ &= (I - K_tZ_t)E[\nu_{t|t-1}\nu_{t|t-1}'](I - K_tZ_t)' \\ &\quad -(I - K_tZ_t)E[\nu_{t|t-1}\varepsilon_t']K_t' - K_tE[\varepsilon_t\nu_{t|t-1}'](I - K_tZ_t)' \\ &\quad +K_tE[\varepsilon_t\varepsilon_t']K_t' \\ &= (I - K_tZ_t)P_{t|t-1}(I - K_tZ_t)' + K_tH_tK_t', \qquad (3.5) \end{aligned}$$

where the last equality follows because the second and third terms in the third equality vanish due to independence across periods (that is, $E[\nu_{t|t-1}\epsilon_t'] = 0$). Finally, to determine the "best" gain matrix K_t^* across infinitely many $n \times m$ time-varying K_t matrices, we minimize the loss function in (3.4). Recalling that $\partial \mathrm{Tr}(ABA') = 2AB\partial A$ for symmetric B, we obtain the first-order condition:

$$\frac{\partial J_t}{\partial K_t} = 2(I - K_t Z_t)P_{t|t-1}(-Z_t)' + 2K_t H_t,$$

which when equated to zero and solved for yields the optimal gain matrix K_t^*:

$$\begin{aligned}
&(I - K_t^* Z_t)P_{t|t-1}Z_t' = K_t^* H_t \\
&P_{t|t-1}Z_t' = K_t^*(Z_t P_{t|t-1} Z_t' + H_t) \\
&K_t^* = P_{t|t-1}Z_t'(Z_t P_{t|t-1} Z_t' + H_t)^{-1}. \qquad (3.6)
\end{aligned}$$

Thus the gain factor K_t^* is optimal in the sense it yields the minimum variance state estimator.

3.2 Summary of the Kalman filter

Let the partially (or fully) observed linear-in-state dynamic stochastic system be specified by Equations (2.22) and (2.23) re-stated below:

$$\textit{Observation equation}: \qquad Y_t = Z_t\alpha_t + c_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, H_t), \qquad (2.22)$$

$$\textit{Transition equation}: \qquad \alpha_t = T_t\alpha_{t-1} + d_t + \nu_t, \quad \nu_t \sim N(0, Q_t). \qquad (2.23)$$

Then, based on information available up to and including time t, the unique and optimal time-path of the distribution of the state vector — that is, the evolution of the state vector's joint density function — is given by the following recursions that constitute what is known as "the celebrated Kalman filter":

Prior means (time update)

$$a_{t|t-1} = T_t a_{t-1} + d_t, \qquad (3.1)$$

Prior covariances (time update)

$$P_{t|t-1} = T_t P_{t-1} T_t' + Q_t, \tag{3.2}$$

Kalman gain factor

$$K_t = P_{t|t-1} Z_t' (Z_t P_{t|t-1} Z_t' + H_t)^{-1}, \tag{3.6}$$

Posterior means (measurement update)

$$a_t = a_{t|t-1} + K_t(Y_t - Z_t a_{t|t-1} - c_t)), \tag{3.3}$$

Posterior covariances (measurement update)

$$P_t = (I - K_t Z_t) P_{t|t-1} (I - K_t Z_t)' + K_t H_t K_t'. \tag{3.5}$$

The above filter recursions run forwards from the given initial state distribution $\alpha_0 \sim N(a_0, P_0)$. For exact initialization of the Kalman filter, see Osinga et al. [2010].

3.3 Properties of the Kalman filter

Below I state the important properties of the Kalman filter.

1. The mean and covariance of the state vector, $(a_{t-1},\, P_{t-1})$, contain complete information up to time $t-1$. To obtain the mean vector a_t, we only need to know the present data Y_t without access to any subset of the past observations $Y_k, k < t$. In other words, the history contained in $I_{t-1} = \{Y_1,\, Y_2, \ldots, Y_{t-1}\}$ can be discarded. This property follows from the Markov structure of the transition Equation (2.23).

2. To obtain the covariance matrix P_t, we need neither the history $I_{t-1} = \{Y_1,\, Y_2, \ldots, Y_{t-1}\}$ nor the current information Y_t. All matrices $P_k, k = 1, \ldots, N$ can be pre-computed before observing any data. They only depend on the system matrices given by the dynamic model (that is, Equations (2.22) and (2.23)). To see this, observe that Equations (3.2), (3.5), and (3.6) do not depend on Y_k for any k. This property does not hold in general for nonlinear models (for example, Equations (2.24) and (2.25)).

3. Equation (3.5) for the covariance matrix evolution is known as the Joseph stabilized version. It guarantees that P_t remains symmetric and positive definite if $P_{t|t-1}$ is symmetric and positive definite. We can also use another equivalent expression: $P_t = (I - K_t, Z_t)P_{t|t-1}$. Although compact, this latter expression need not preserve symmetry or positive definiteness over time.

4. The Kalman filter is both unique and optimal across all possible estimators of any linear state space model with normal errors (ε_t, ν_t). This property holds even when the errors (ε_t, ν_t) are correlated over time or between themselves.

5. The Kalman filter is optimal across all *linear* estimators even in the presence of *non*-normal errors (ε_t, ν_t). This property holds even when errors (ε_t, ν_t) are correlated over time or between themselves. This property follows from the fact that the above derivation did not rely on the normality of the error terms — just that their second moments (H, Q) are finite regardless of the shape of the density function. In other words, if the state space model is linear, and the error terms follow *any* non-normal distribution (for example, Gamma, uniform, mixture of normals) with finite two moments, then the Kalman filter recursions provide the best linear updating of the state vector.

3.4 Kalman smoother

3.4.1 What is smoothing?

Kalman filter provides the mean a_t and the covariance P_t of the state vector based on the past and present information up to time t, which can be expressed as $I_t = \underbrace{I_{t-1}}_{\text{past}} \cup \underbrace{Y_t}_{\text{present}}$. Now suppose we are given additional information from the future $I_k = \{Y_1, \ldots, Y_{t-1}, Y_t, \underbrace{Y_{t+1}, \ldots, Y_k}_{\text{future}}\}$, where $k > t$. Then how do we optimally update the present mean and covariance using future information? Smoothing answers this question. Because we use more information, the resulting means yield "smoother" time plots than those based on the forecast estimates $\{a_{t|t-1} : t = 1, 2, \ldots, N\}$ or the filtered estimates $\{a_{t|t} : t = 1, 2, \ldots, N\}$. In general, three kinds of smoothing estimators exist: fixed-point smoother, fixed-lag smoother, and fixed-interval smoother.

In fixed-point smoothing, we improve the estimate of the state vector at a fixed time point s as future information arrives. In notation, at any time s, the Kalman filter provides the filtered estimate $a_{s|s}$; then denoting $a_{s|k} = E[\alpha_s|I_k], k > s$, we update the filtered mean $a_{s|s}$ to obtain $a_{s|s+1}, a_{s|s+2}, a_{s|s+3}$ as the new information $Y_{s+1}, Y_{s+2}, Y_{s+3}$ rolls in. Here the time point s remains fixed.

In fixed-lag smoothing, we seek improvement in the filtered estimates τ periods ago. In notation, at any time s, we obtain the estimate $a_{s-\tau|s}$ for $s = \tau, \tau + 1, \ldots$ so on. Here the lag window τ remains fixed, but the time point s marches on.

Lastly, in fixed-interval smoothing, we seek improvement in all the filtered estimates over the observation span N. In notation, at every time t, we seek the estimate $a_{0|N}, a_{1|N}, \ldots, a_{t|N}, \ldots, a_{N|N}$. Here the time span N remains fixed, but the time point t covers the entire period from the initial to the terminal value.

In marketing, the first two smoothing concepts have not been used as yet, whereas the fixed-interval smoothing is necessary for estimating model parameters using either Bayesian or EM estimation (but smoothing is not required in maximum likelihood estimation and inference). Hence, I next derive the optimal recursions for fixed-interval smoothing.

3.4.2 Derivation of the Kalman smoother

Similar to the derivation of Kalman filter, we can apply the optimization theory to obtain the Kalman smoother recursions without requiring normality (that is, property 5 holds for the Kalman smoother as well). But we will sacrifice this generality here in the interest of learning from statistical theory the two important results:

Result 1. The Law of Iterated Expectation states that

- $E[x_1] = E[E[x_1|x_2]]$
- $\text{Cov}[x_1] = E[\text{Cov}[x_1|x_2]] + \text{Cov}[E[x_1|x_2]]$.

Result 2. If $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}' & \Sigma_{22} \end{bmatrix}\right)$, then the conditional mean follows $x_1|x_2 \sim N(\mu, \Sigma)$, where $\mu = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)$ and $\Sigma = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}'$.

Armed with these two results, we derive the fixed-interval Kalman smoother. Let us first stack the posterior state vector $\alpha_t|I_t$ above the next period's prior state vector $\alpha_{t+1}|I_t$ as follows:

$$\begin{bmatrix} \alpha_t|I_t \\ \alpha_{t+1}|I_t \end{bmatrix} = \begin{bmatrix} \alpha_{t|t} \\ \alpha_{t+1|t} \end{bmatrix} \sim N\left(\begin{bmatrix} a_t \\ a_{t+1|t} \end{bmatrix}, \begin{bmatrix} P_t & P_t T_{t+1}' \\ T_{t+1}P_t & P_{t+1|t} \end{bmatrix}\right),$$

where we obtained the off-diagonal matrices by noting that $\text{Cov}(\alpha_{t+1|t}, \alpha_{t|t}) = \text{Cov}((T_{t+1}\alpha_t), (\alpha_t)|I_t) = T_{t+1}P_t$. Applying Result 2, we then get

$$\begin{aligned} E[\alpha_t|\alpha_{t+1|t}] &= a_t + P_t T_{t+1}' P_{t+1|t}^{-1}(\alpha_{t+1|t} - a_{t+1|t}) && (3.7) \\ \text{Cov}[\alpha_t|\alpha_{t+1|t}] &= P_t - P_t T_{t+1}' P_{t+1|t}^{-1} T_{t+1} P_t. && (3.8) \end{aligned}$$

Now we use Result 1 to derive the smoothed mean:

$$\begin{aligned} a_{t|N} &= E[\alpha_t|I_N] \\ &= E[E[\alpha_t|\alpha_{t+1|t}]|I_N] \\ &= E[a_t + L_t(\alpha_{t+1|t} - a_{t+1|t})|I_N] \\ &= a_t + L_t(a_{t+1|N} - a_{t+1|t}), && (3.9) \end{aligned}$$

where we denote $L_t = P_t T_{t+1}' P_{t+1|t}^{-1}$ obtained in (3.7). In Equation (3.9), the first equality defines the smoothed mean as the expected value of the state vector based on the entire sample (past, present, and future) up to N. The second equality applies the law of iterated expectation and conditions on the random variable $\alpha_{t+1|t}$. The third equality substitutes the result obtained in Equation (3.7), and it depends on the random variable $\alpha_{t+1|t}$. The final equality re-applies the definition of the smoothed mean (that is, $a_{t+1|N} = E[\alpha_{t+1|t}|I_N]$).

To gain intuition, note that the smoothing occurs backwards, starting from $t = N$, $N-1$, $\ldots, 1$, 0. The final equality in (3.9) incorporates the effect of future information on the filtered mean a_t obtained from the Kalman filter. It reveals that the correction to the filtered mean a_t is proportional to the change in the expected values of α_{t+1} based on the complete information at N and the available information at t. The proportionality factor, L_t, depends on the ratio of the posterior covariance $P_{t|t}$ to the prior covariance $P_{t+1|t}$.

Similarly, we derive the smoothed covariance matrix:

$$\begin{aligned}
P_{t|N} &= \operatorname{Cov}[\alpha_t|I_N] \\
&= E[\operatorname{Cov}[\alpha_t|\alpha_{t+1|t}]|I_N] + \operatorname{Cov}[E[\alpha_t|\alpha_{t+1|t}]|I_N] \\
&= (P_t - L_t P_{t+1|t} L_t') + \operatorname{Cov}[a_t + L_t(\alpha_{t+1} - a_{t+1|t})|I_N] \\
&= P_t - L_t P_{t+1|t} L_t' + L_t P_{t+1|N} L_t' \\
&= P_t - L_t(P_{t+1|t} - P_{t+1|N})L_t'. \qquad (3.10)
\end{aligned}$$

In Equation (3.10), the first equality defines the smoothed mean as the covariance of the state vector based on the entire sample (past, present, and future) up to N. The second equality applies the law of iterated expectation and conditions on the random variable $\alpha_{t+1|t}$. In the third equality, the first term on the right-hand side substitutes the result obtained in Equation (3.8), which does not involve a random variable; the second term substitutes the result obtained in Equation (3.7), which involves the random variable $\alpha_{t+1|t}|I_N$. The final equality re-applies the definition of the smoothed covariance (that is, $P_{t+1|N} = \operatorname{Cov}[(\alpha_{t+1|t} - a_{t+1|t})|I_N]$). As before, the correction to the filtered covariance P_t is proportional to the change in the expected covariances of α_{t+1} based on the available information at t and the complete information at N. Because information reduces variance, we expect $(P_{t+1|t} - P_{t+1|N})$ to be positive definite; hence the smoothed covariance is smaller than its filtered estimate.

3.4.3 Summary of the Kalman smoother

Let the partially (or fully) observed linear-in-state dynamic stochastic system be specified by Equations (2.22) and (2.23) re-stated below:

$$\textit{Observation equation}: \qquad Y_t = Z_t\alpha_t + c_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, H_t) \qquad (2.22)$$

$$Transition\ equation: \qquad \alpha_t = T_t\alpha_{t-1} + d_t + \nu_t, \quad \nu_t \sim N(0, Q_t) \tag{2.23}$$

Then, based on *all* information available up to and including time $t = N$, the unique and optimal time-path of the distribution of the state vector — that is, the evolution of the state vector's joint density function — is given by the following recursions that constitute what is known as "the Kalman smoother":

Smoothed means (temporal update)

$$a_{t|N} = a_t + L_t(a_{t+1|N} - a_{t+1|t}), \tag{3.9}$$

Smoothed covariances (temporal update)

$$P_{t|N} = P_t - L_t(P_{t+1|t} - P_{t+1|N})L_t', \tag{3.10}$$

Smoother gain factor

$$L_t = P_tT_{t+1}'P_{t+1|t}^{-1}. \tag{3.7}$$

The above smoother recursions run backwards from the terminal state distribution $\alpha_N \sim N(a_N, P_N)$, which is furnished by the Kalman filter at $t = N$.

3.5 Nonlinear filters

The above development completes the solution to the linear state estimation problem: what is the best estimate of the state vector and its precision given the set of observations up to time $t-1$ before the present data Y_t arrives; the best state estimate and its precision upon arrival of the present data; and the best state estimate and its precision after future observations up to N roll in.

But the results assume linear dynamics in Equations (2.22) and (2.23). Arguably, the market dynamics need not be linear. Then why did we spend so much time developing the linear theory? In *The Feynman Lectures on Physics* (see Chapter 25), Richard Feynman offers the insight, "*The answer is simple: because we can solve them! . . . if we understand linear equations, we are ready, in principle, to understand a lot of things.*" See Feynman et al. [2006, p. 25-4].

Accordingly, I describe how the linear theory helps us tackle nonlinear filtering. Consider the nonlinear state space model in Equations (2.24) and (2.25). Let the nonlinear link function $Z(\alpha_t) = \alpha_t^2$, for example. Then the expected value $E[Y_t] = E[Z(\alpha_t) + c_t + \varepsilon_t] = E[\alpha_t^2] + c_t \neq (E[\alpha_t])^2 + c_t$. In words, the mean of the nonlinear function of random variables is not the same as the nonlinear function of the means of random variables. In general, $E[Z(\alpha_t)] \neq Z(E[\alpha_t])$ and $E[T(\alpha_{t-1})] \neq T(E[\alpha_{t-1}])$. These facts exacerbate the updating of the means and covariance of the state vector as observations arrive sequentially. The linearity of state space models circumvents this problem, leading to the *exact* Kalman filter and smoother recursions.

To derive the optimal nonlinear filter, not only do we need to update the means and covariance of the state vector, but its entire probability density at each instant — that includes the third moment, fourth moment, and all higher-order moments. The good news is that the exact optimal nonlinear filter exists. Analogous to the Kalman and Bucy [1961] filter for continuous-time linear state space models, the Kushner equation [Kushner, 1964] and the Zakai equation [Zakai, 1969] provide the exact updating of the un-normalized density function of the state vector over time for continuous-time nonlinear state space models. Kushner's equation follows a nonlinear stochastic partial differential equation, while Zakai's equation is a linear stochastic partial differential equation.

Both the solutions are infinite dimensional in general and would require finite dimensional approximations to implement. Hence, I describe a simpler approach known as the previously mentioned EKF, which is an approximate solution obtained via the Taylor series expansion of the nonlinear $Z(\cdot)$ and $T(\cdot)$ functions. The earliest application of EKF was by NASA for nonlinear re-entry dynamics to bring back the rocket after moon landing [Jazwinski, 2007]. Over the years EKF is "... undoubtedly the most widely used nonlinear state estimation technique ... applied in the past few decades" [Simon, 2006, p. 396] and considered the *de facto* standard in nonlinear state estimation (see Julier and Uhlmann [2004]).

3.5.1 Extended Kalman filter (EKF)

The idea underlying EKF is simple. It uses the *linear* Kalman filter and smoother recursions, except we replace the link matrix Z and the transition matrix T by the Jacobian of the corresponding nonlinear functions. Specifically, denote $\tilde{Z}_t = \frac{\partial Z(\alpha)}{\partial \alpha'}\Big|_{\alpha=a_{t|t-1}}$ as the $m \times n$ Jacobian matrix of the nonlinear link function $Z(\alpha)$ with respect to the vector α, which is evaluated at the prior mean $a_{t|t-1}$. Similarly, let $\tilde{T}_t = \frac{\partial T(\alpha)}{\partial \alpha'}\Big|_{\alpha=a_{t|t-1}}$ be the $n \times n$ Jacobian matrix of the nonlinear transition function $T(\alpha)$ evaluated at the prior mean $a_{t|t-1}$.

To summarize, let the partially (or fully) observed nonlinear state dynamic stochastic system be specified by Equations (2.24) and (2.25) re-stated below:

$$\textit{Observation equation}: \quad Y_t = Z_t(\alpha_t) + c_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, H_t) \tag{2.24}$$

$$\textit{Transition equation}: \quad \alpha_t = T_t(\alpha_{t-1}) + d_t + \nu_t, \quad \nu_t \sim N(0, Q_t) \tag{2.25}$$

Then, based on information available up to and including time t, the time-path of the distribution of the state vector — that is, the evolution of the state vector's joint density function — is given by the following recursions:

Prior means (time update)

$$a_{t|t-1} = T_t(a_{t-1}) + d_t,$$

Prior covariances (time update)

$$P_{t|t-1} = \tilde{T}_t P_{t-1} \tilde{T}_t' + Q_t,$$

Kalman gain factor

$$K_t = P_{t|t-1} \tilde{Z}_t' (\tilde{Z}_t P_{t|t-1} \tilde{Z}_t' + H_t)^{-1},$$

Posterior means (measurement update)

$$a_t = a_{t|t-1} + K_t(Y_t - Z_t(a_{t|t-1}) - c_t),$$

Posterior covariances (measurement update)

$$P_t = (I - K_t\tilde{Z}_t)P_{t|t-1}(I - K_t\tilde{Z}_t)' + K_t H_t K_t'.$$

Besides linearization, approximation also occurs in the predicted $\hat{Y}_t = E[Y_t|I_{t-1}] = E[Z(\alpha_t) + c_t + \varepsilon_t)|I_{t-1}] \approx Z(a_{t|t-1}) + c_t$, which is approximate because $E[Z(\alpha_t)|I_{t-1}] \neq Z(a_{t|t-1})$. To mitigate the resulting bias, we can further refine the expansion point at which the Jacobians are evaluated. In the *iterated* EKF, at time t, we first apply the EKF to obtain the posterior mean a_t^0, which is then refined iteratively by evaluating the Jacobians $\tilde{Z}_t^l = \frac{\partial Z(\alpha)}{\partial \alpha'}\Big|_{\alpha=a_{t|t-1}^l}$ and $\tilde{T}_t^l = \frac{\partial T(\alpha)}{\partial \alpha'}\Big|_{\alpha=a_{t|t-1}^l}$ for $l = 0, 1, 2, \ldots, L$. Using the refined link and transition matrices, we compute the new prior mean, new prior covariance, new Kalman gain K_t^l to obtain the next refined posterior mean $a_t^{l+1} = a_{t|t-1}^l + K_t^l(Y_t - Z(a_{t|t-1}^l) - c_t)$. We continue these iterations till the consecutive refined posterior means do not change appreciably. Then, we process the next observation using the EKF within which we repeat the L iterations. For marketing applications of the nonlinear Kalman filter, see Naik et al. [2008] or Kolsarici and Vakratsas [2010].

3.5.2 Other nonlinear filters

The EKF is simple to understand and implement: replace the link and transition matrices in the Kalman filter by the Jacobians of the nonlinear functions. Other nonlinear filters entail varying degrees of conceptual complexity and computational effort in the pursuit of marginal improvement in accuracy. More importantly, the resulting improvements are model- and data-dependent with limited generalizability. Hence we briefly present two types of nonlinear filters: particle filter and unscented Kalman filter (UKF).

Particle filter represents the posterior density of the state vector, $p(\alpha_t|I_t)$, using a set of weights and points $\{w_{t,i}, \alpha_{t,i}^+\}$, so that

$$p(\alpha_t|I_t) = \operatorname*{Lim}_{j\to\infty} \sum\nolimits_{i=1}^{J} w_{t,i}\delta(\alpha_{t,i}^+),$$

where $\delta(\alpha_{t,i}^+)$ is nonzero at the point $\alpha_{t,i}^+$ and zero elsewhere, and it integrates to unity. Particle filtering is approximate because the number of points J is finite in practice. Typically, $J = 1000$ for state dimension $n < 5$. These points are known as draws or particles. The positive and negative superscripts on $\alpha_{t,i}$ denote the posterior and prior particles, respectively. In particle filtering, the particles $\{w_{t,i}, \alpha^+_{t,i}\}$ propagate to new set $\{w_{t+1,i}, \alpha^+_{t+1,i}\}$ at the next time period based on the observed data and model dynamics. Instead of using the filtering recursions to obtain the mean and covariance of the state vector, a simulation step generates the set of particles and weights to characterize the distribution of the state vector.

To convey the basic ideas of particle filtering, let us randomly generate J state vectors from the initial density $p(\alpha_0|I_0)$, which is assumed to be known. At time 0, we denote the posterior particles by $\alpha^+_{0,i}, i = 1, \ldots, J$. Then, at each time step $t = 1, 2, \ldots, N$, we obtain the prior particles $\alpha^-_{t,i} = T_t(\alpha^+_{t-1,i}) + d_t + \nu^i_t$, where ν^i_t are random draws from the error distribution $\nu_t \sim N(0, Q_t)$ Next, using the observed data Y_t, we obtain the relative likelihood given by $q_{t,i} \propto \exp(-0.5(Y_t - Z(\alpha^-_{t,i}) - c_t)' H_t^{-1}(Y_t - Z(\alpha^-_{t,i}) - c_t))$. We normalize the relative likelihoods to sum to unity via $w_{t,i} = q_{t,i}/\sum_i q_{t,i}$. Subsequently, to obtain the posterior particles, we resample from the likelihood distribution in two-steps. First, we generate a random number r from a uniform distribution [0,1]. Second, for each particle i, we accumulate the likelihood until it exceeds r. In other words, $\sum_{m=1}^{j-1} q_m < r$ and $\sum_{m=1}^{j} q_m \geq r$. Here m is an index of summation, and we seek the smallest value of j such that the accumulated sum just exceeds the random draw r. See Figure 15.2 in Simon [2006] for an illustration. Then assign the prior particle $\alpha^-_{t,j}$ to the new posterior particle $\alpha^+_{t,i}$. That is, $\alpha^+_{t,i} = \alpha^-_{t,j}$ with probability q_j for $(i, j) = 1, \ldots, J$. This completes the loop to obtain the new set $\{w_{t,i}, \alpha^+_{t,i}\}$. Repeat the above steps across N time periods. See Andrieu and Doucet [2002] for further details. For a marketing application of particle filtering, see Bruce [2008].

The above approach may lead to degenerate weights when the regions of the state space with significant mass under $p(Y_t|\alpha_t)$ do not overlap with $p(\alpha_t|I_{t-1})$. In such cases, resampling selects a few prior particles to become posterior particles, leading to the collapse of all particles to the same value. This phenomenon is called the “black hole" of particle filtering. Drawing a larger set J by brute force not only increases the computational burden, but also delays the inevitable sample impoverishment. To remedy this problem, we consider roughening, prior editing, regularized particle filtering, MCMC resampling, and auxiliary particle filtering. For these details, readers should consult Simon [2006, Chapter 15] and references therein.

Particle filters draw a large set of random particles and then relies on a swarm of these particles to zero in on an appropriate region of the state space to characterize the posterior density $p(\alpha_t|I_t)$, given the model dynamics and observed data. In contrast, another class of nonlinear filters designs a filter based on a small set of *intelligently* located points in the state space (as opposed to randomly drawn). These points are known as sigma points. Recall that, in nonlinear filtering, $E[Z(\alpha_t)] \neq Z(E[\alpha_t])$ and $E[T(\alpha_{t-1})] \neq T(E[\alpha_{t-1}])$, which leads to inaccurate expected values which, in turn, injects inaccuracy in the covariance matrices. These inaccuracies accumulate over time $t = 1, \ldots, N$ causing the filter to diverge from the true state evolution.

To mitigate this problem, we seek to obtain better estimates of the means and covariance matrix of the nonlinear function $y = h(x)$ via the *unscented transform* [Julier and Uhlmann, 2004]. Chapter 14.2 by Simon [2006] provides a lucid explanation of unscented transformations. To cover it briefly here, let x denote a random vector of dimension $n \times 1$ with mean $\bar{x}$ and covariance matrix S. The distribution of x can be non-normal. Then find the Cholesky matrix $\sqrt{nS}$ such that $(\sqrt{nS})'\sqrt{nS} = nS$. Next select the specific $2n$ sigma points as follows:

$$\begin{array}{rclrcl} x^i &=& \bar{x} + \tilde{x}^i, & i &=& 1, \ldots, 2n, \\ \tilde{x}^i &=& (\sqrt{nS})'_i, & i &=& 1, \ldots, n, \\ \tilde{x}^{(n+i)} &=& -\,(\sqrt{nS})'_i, & i &=& 1, \ldots, n, \end{array}$$

where $(\sqrt{nS})_i$ is the ith row of $\sqrt{nS}$. Finally, compute the expected value and covariance using $\mu_y = \frac{1}{2n}\Sigma_{i=1}^{2n} y^i$ and $\Sigma_y = \frac{1}{2n}\Sigma_{i=1}^{2n}(y^i - \mu_y)(y^i - \mu_y)'$, where $y^i = h(x^i)$ for $i = 1, \ldots, 2n$. It can be shown (see Simon [2006, p. 444]) that μ_y is accurate up to the third-order Taylor series expansion of $h(\cdot)$.

To run the unscented Kalman filter, let (a_0, P_0) denote the initial mean vector and covariance matrix. At time $t = 1, 2, \ldots, N$ specify the sigma points as follows:

$$\begin{aligned} a^i_{t-1} &= a_{t-1} + \tilde{a}^i, & i &= 1, \ldots, 2n, \\ \tilde{a}^i &= (\sqrt{nP_{t-1}})'_i, & i &= 1, \ldots, n, \\ \tilde{a}^{(n+i)} &= -(\sqrt{nP_{t-1}})'_i, & i &= 1, \ldots, n. \end{aligned}$$

Obtain the prior sigma points using the transition Equation (2.25). That is, $a^i_{t|t-1} = T_t(a^i_{t-1}) + d_t$ for $i = 1, \ldots, 2n$. Average across those sigma points to find the prior mean $a_{t|t-1} = \frac{1}{2n}\Sigma^{2n}_{i=1}\ a^i_{t|t-1}$. Similarly, compute the prior covariance $P_{t|t-1} = \frac{1}{2n}\Sigma^{2n}_{i=1}(a^i_{t|t-1} - a_{t|t-1})$ $(a^i_{t|t-1} - a_{t|t-1})' + Q_t$. Now use this new prior mean vector and covariance matrix to find the new sigma points:

$$\begin{aligned} a^i_t &= a_{t|t-1} + \tilde{a}^i, & i &= 1, \ldots, 2n, \\ \tilde{a}^i &= \left(\sqrt{nP_{t|t-1}}\right)'_i, & i &= 1, \ldots, n, \\ \tilde{a}^{(n+i)} &= -\left(\sqrt{nP_{t|t-1}}\right)'_i, & i &= 1, \ldots, n. \end{aligned}$$

Using the observation equation get $Y^i_t = Z_t(a^i_t) + c_t$ for $i = 1, \ldots, 2n$. Then average across those sigma points to predict the observation $\hat{Y}_t = \frac{1}{2n}\Sigma^{2n}_{i=1}Y^i_t$. Next compute the observation covariance $P_{y,t} = \frac{1}{2n}\Sigma^{2n}_{i=1}(Y^i_t - \hat{Y}_t)(Y^i_t - \hat{Y}_t)' + H_t$ and the cross-covariance $P_{\alpha y,t} = \frac{1}{2n}\Sigma^{2n}_{i=1}(a^i_t - a_{t|t-1})(Y^i_t - \hat{Y}_t)'$. Thus obtain the Kalman gain factor $K_t = P_{\alpha y,t}P^{-1}_{y,t}$. Finally update the posterior mean $a_t = a_{t|t-1} + K_t(Y_t - \hat{Y}_t)$ and the posterior covariance $P_t = P_{t|t-1} - K_t P_{y,t} K'_t$. Repeat the above steps across N time periods.

Because the UKF relies on a small set of design points, located intelligently in the state space rather than drawn randomly by brute force, it is computationally more efficient than particle filters. An application of UKF in marketing literature does not exist as yet. We close this section by noting that there are other ways to design sigma points; for example, see cubature Kalman filter [Arasaratnam and Haykin, 2008].

4

Parameter Estimation

In the previous section, we learnt how to estimate the state vector and its precision, $(a_t P_t)$, using observed data, but assuming that the parameters defining the system matrices, namely, $Z(\theta), c(\theta), H(\theta)$, $T(\theta), d(\theta), Q(\theta)\}$, are known. For example, consider Equations (2.1) and (2.2), which depend on the parameters $(A_0, \lambda, \beta, \sigma_\varepsilon^2, \sigma_\nu^2)$, whose values are not known and needs to be determined based on the observed sample. This section explains how to estimate the parameter vector θ and its precision using observed data. To this end, I present the maximum likelihood estimation, inference, and model selection.

4.1 What is the likelihood principle?

Suppose we observe a sample $\{Y_1 \ldots, Y_N\}$ from the normal distribution $N(\mu, \Sigma)$. What is the probability of this event? We compute it via the joint density function $f(Y_1, \ldots, Y_N|\theta)$, where $\theta = (\mu, \text{vech}(\Sigma))'$. We now reverse our perspective to imagine the joint density function as a function of θ rather than the random variables $\{Y_1 \ldots, Y_N\}$. Let us denote the resulting function as $L(\theta|Y_1, \ldots, Y_N)$ and call it the likelihood function. In other words, the likelihood function $L(\theta|Y_1, \ldots, Y_N)$ equals the joint density function $f(Y_1, \ldots, Y_N|\theta)$, except that the domain of $L(\cdot)$ is the parameter space in which θ lives, whereas the domain of $f(\cdot)$ is the space of random variables $\{Y_t\}$. So the parameter θ is the variable that changes in $L(\theta|Y_1, \ldots, Y_N)$ for the fixed data, whereas the random variables change in $f(Y_1, \ldots, Y_N|\theta)$ for the fixed θ.

The likelihood principle states that the likelihood function, which equals the joint density function viewed as a function of θ given the data, contains all relevant information about θ for the purposes of estimation and inference. Consequently, if $L(\theta_1|Y_1, \ldots, Y_N) > L(\theta_2|Y_1, \ldots, Y_N)$ then θ_1 is a better estimate than θ_2 because the likelihood of the sample arising from the joint density indexed by $f(Y_1, \ldots, Y_N|\theta_1)$ is larger. If so, then the "best" estimate is the one that corresponds to the largest value of the likelihood function. Hence the practical application of the likelihood principle leads us to find the maximum of the likelihood function and consider the corresponding θ^* as the maximum-likelihood estimate. We express symbolically "the argument that maximizes the likelihood function" by $\theta^* = \text{ArgMax}\, L(\theta|Y_1, \ldots, Y_N)$. Figure 4.1 illustrates this idea of maximum likelihood estimation.

To assess the precision of θ^*, we look at the curvature of the likelihood function at the maximum. See Figure 4.1. Recall from basic geometry that a straight line has zero curvature; sharp bend has a large curvature. The larger the curvature, the tighter the range θ^* varies. The smaller the curvature, the flatter the likelihood function, the broader the range of θ^*. Hence an inverse of the curvature provides the estimate of $Var(\theta^*)$. We formalize this intuition in Section 4.3.

4.2 State space model estimation

To estimate the parameters of state space models, we begin by writing the likelihood function for observing the data $\{Y_1, \ldots, Y_N\}$ sequentially over time. From the above discussion, the likelihood function equals the

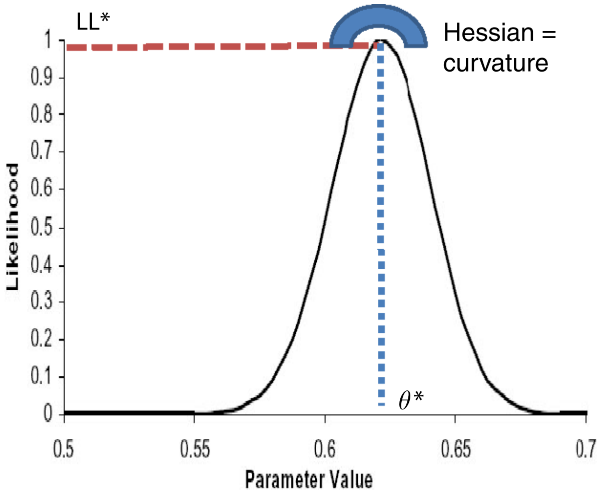

Figure 4.1: Likelihood Principle.

joint density of observing the realized sequence $\{Y_1, \ldots, Y_N\}$. Formally,

$$\begin{aligned} L(\theta|\{Y_1, Y_2, \ldots, Y_N\}) &= f(Y_1, Y_2, \ldots, Y_N; \theta) \\ &= f(Y_N|\{Y_1, Y_2, \ldots, Y_{N-1}\}; \theta) \\ &\quad \times f(\{Y_1, Y_2, \ldots, Y_{N-1}\}; \theta) \\ &= f(Y_N|I_{N-1}, \theta) \times f(Y_{N-1}|I_{N-2}, \theta) \\ &\quad \times f(\{Y_1, Y_2, \ldots, Y_{N-2}\}; \theta) \\ &= f(Y_N|I_{N-1}, \theta)\, f(Y_{N-1}|I_{N-2}, \theta) \ldots \\ &\quad f(Y_1|I_0, \theta) \\ &= \prod_{t=1}^{N} f(Y_t|I_{t-1}, \theta). \end{aligned} \tag{4.1}$$

The first equality expresses the equivalence of the likelihood function and the joint density function. The second equality decomposes the joint density into the product of the conditional density of observing $Y_N|I_{N-1}$ with the joint density of the event $\{Y_1, \ldots, Y_{N-1}\}$. The next two equalities recursively apply this logic to arrive at the last equality that expresses the likelihood function as the product of conditional densities.

Because $Y_t|I_{t-1}$ is normally distributed, that is, $Y_t|I_{t-1} \sim N(\mu_t, \Sigma_t)$, we can further simplify Equation (4.1) to obtain the log-likelihood function:

$$\begin{aligned}
\mathrm{LL}(\theta) &= \ln(L(\theta|\{Y_1, Y_2, \ldots, Y_N\})) \\
&= \ln\left(\prod_{t=1}^{N} f(Y_t|I_{t-1}, \theta)\right) \\
&= \sum_{t=1}^{N} \ln(f(Y_t|I_{t-1}, \theta)) \\
&= -0.5 \sum_{t=1}^{N} \Big[\ln(\det(\Sigma_t)) + (Y_t - \mu_t)'\Sigma_t^{-1}(Y_t - \mu_t)\Big]. \quad (4.2)
\end{aligned}$$

In Equation (4.2), we evaluate the natural log of the likelihood function because (1) the monotonicity of logarithm ensures the locations of the maximized $L(\theta)$ and $\mathrm{LL}(\theta)$ remain the same, and (2) the sums in the last two equalities are easier to maximize (than the product form in (4.1)). We obtain the last equality in (4.2) based on the result $f(Y_t|I_{t-1}) = (2\pi)^{-m} \det(\Sigma_t)^{-0.5} \exp(-0.5(Y_t - \mu_t)'\Sigma_t^{-1}(Y_t - \mu_t))$, where $\det(\cdot)$ denotes the determinant of a matrix, and the constant $(2\pi)^{-m}$ is ignored because it does not depend on θ.

In linear state space models, $\mu_t = E[Y_t|I_{t-1}] = E[Z_t\alpha_t + c_t + \varepsilon_t|I_{t-1}] = Z_t a_{t|t-1} + c_t$. Similarly, $\Sigma_t = \mathrm{Cov}[Y_t|I_{t-1}] = Z_t P_{t|t-1} Z_t' + H_t$. Using these expressions, we obtain μ_t and Σ_t recursively to build the log-likelihood function in (4.2). Because the system matrices are potentially nonlinear functions of θ, both μ_t and Σ_t become complicated nonlinear functions of θ even though the state space model is linear-in-state and the Kalman filter offers closed-form recursions. In nonlinear state space models, $\mu_t = E[Y_t|I_{t-1}] = Z_t(a_{t|t-1}) + c_t$ and $\Sigma_t = \mathrm{Cov}[Y_t|I_{t-1}] = \tilde{Z}_t P_{t|t-1} \tilde{Z}_t' + H_t$, where $\tilde{Z}_t$ is the Jacobian used in the EKF (see Section 3.3.1). Hence, for both the linear and nonlinear state space models, the maximization of (4.2) requires numerical solvers, which are based on quasi-Newton algorithms (for example, BFGS, BHHH) and are available in commercial software (for example, Matlab's fminunc or Gauss' Optmum) and free open source software (for example, optimx[1] in R). Gill et al. [1982] and Nocedal and Wright [2006] serve as good references to quasi-Newton optimization methods.

By applying numerical optimization, we obtain the maximum-likelihood estimates:

$$\theta^* = \text{ArgMaxLL}(\theta | Y_1, \ldots, Y_N). \tag{4.3}$$

In numerical optimization, the main challenges are the lack of convergence or the convergence to local maxima. If convergence is failed, (1) relax the tolerance on the allowable maximum slope till you find good starting values θ_0 that lead to the neighborhood of θ^* and then tighten it back. (2) Specify different θ_0 randomly. (3) Rescale (Y_t, X_t) so that the elements of θ have similar magnitudes; **this fix is the most important one.** (4) Use DFP or BHHH algorithm instead of BFGS to maximize LL (see Gill et al. [1982] and Nocedal and Wright [2006]). (5) Use derivative-free methods to maximize LL and use its solution as θ_0. (6) Use EM algorithm to maximize $E[\text{LL}]$ as explained in Shumway and Stoffer [2011] and then use the resulting EM estimates as the starting values θ_0. That is because EM does not provide the Hessian at convergence, so standard errors are unavailable, and hence this ML step is required for statistical inference. Finally, ensure that multiple starting values do not yield larger LL*; if multiple (local) maxima are attained, select θ^* corresponding to the largest value of LL*.

4.3 Statistical inference

As mentioned, intuitively, $\text{Var}(\theta^*)$ depends on the curvature of the likelihood function at the maximum. We measure curvature using the second derivative of the likelihood function. Suppose θ is a $K \times 1$ vector. Then we compute the $K \times K$ matrix of second partial derivatives,

[1] For documentation and download, see http://cran.r-project.org/web/packages/optimx/optimx.pdf.

known as the Hessian matrix and defined by $\mathcal{H} = \frac{\partial^2 L(\theta)}{\partial\theta\partial\theta'}$. Note that each element of Y_t in the sample $\{Y_1, \ldots, Y_t, \ldots Y_N\}$ traces a curve as a function of time (for example, $\{Y_{j1}, \ldots, Y_{jt}, \ldots Y_{jN}\}$). Different realizations of Y_t would generate different random curves; collectively they represent an ensemble of infinitely many curves. Hence, we take the expectation of the sample H to obtain the Fisher information matrix:

$$\Im = E[-\mathcal{H}]. \tag{4.4}$$

Why the negative sign? Because the likelihood function $L(\theta)$ increases at a decreasing rate to attain the maximum, its Hessian $\mathcal{H}$ is negative definite at the maximum point. (As an aside, this fact satisfies the second-order condition, thereby ensuring θ^* in (4.3) locates the maximum of $L(\theta)$.) The negative sign thus renders $\Im$ as a positive definite matrix, thus conforming to the properties of covariance matrices. (Recall from basic calculus that a Hessian is symmetric by construction.)

Having obtained the Fisher information matrix, the likelihood theory proves that, asymptotically as the sample size N tends to infinity,

$$\mathrm{Cov}(\theta^*) = \Im^{-1}, \tag{4.5}$$

which formalizes the intuition that the inverse of the curvature measures the variability of the estimates. Furthermore, the likelihood theory proves that the maximum-likelihood estimators are consistent (that is, $\theta^*_N \to \theta$ as $N \to \infty$) and asymptotically normally distributed (that is, $\theta^* \sim N(\theta, \Im^{-1})$).

We implement this theory in practice as follows. We compute the Hessian matrix $\mathcal{H} = \frac{\partial^2 L(\theta)}{\partial\theta\partial\theta'}$, invert and multiply by negative unity, to get $\mathrm{Cov}(\theta^*) = -\mathcal{H}^{-1}$, which represents the variance–covariance matrix of the estimated parameters. Using it, we can conduct any statistical inference and joint hypotheses tests. For example, to obtain standard errors, we extract the diagonal of $\mathrm{Cov}(\theta^*)$ and take the square roots of its elements (that is, $se(\theta^*) = \mathrm{sqrt}(\mathrm{diag}(-\mathcal{H}^{-1}))$), which can then provide the t-values and confidence intervals.

The above statistical inference assumes that the model specification is correct. But the estimated model specification could be wrong (that is, different from the true data generating process). To hedge for such misspecification errors, we conduct robust inferences. Specifically, we compute a robust estimator of the variance-covariance matrix given by the sandwich estimator [Huber, 1967, White, 1980]:

$$V = (-\mathcal{H}^{-1})M(-\mathcal{H}^{-1}), \tag{4.6}$$

where M is a $K \times K$ matrix of the gradients of the log-likelihood function. That is, $M = G'G$ and G is the $N \times K$ matrix obtained by stacking the $1 \times K$ vector of the gradient of $\mathrm{LL}(\theta)$ for each of the N observations. In correctly specified models, $M = -\mathcal{H}^{-1}$ and so both the Equations (4.5) and (4.6) yield exactly the same standard errors (as they should); otherwise, we use the robust standard errors given by the square roots of the diagonal elements of V.

4.4 Model selection

Applying the maximum-likelihood approach to a given data set, we can estimate model parameters and assess their statistical significance. Indeed, we can estimate multiple models: Model 1, Model 2, Model 3, and so on. Then how do we decide which model to retain? To answer this question, model selection literature offers metrics to evaluate "good" models from a set of alternative models.

Good models exhibit (1) simplicity to facilitate understanding and (2) fidelity to furnish accurate forecasts. These two goals exert opposing forces: simplest models (for example, intercept only) do not predict well, whereas complex models may generate good forecasts but obfuscate insights. To balance the opposing forces of simplicity (measured by parsimony) and fidelity (measured via goodness-of-fit), statisticians developed a few metrics, known as information criteria, such as AIC (Akaike information criterion), AIC$_\mathrm{C}$ *(corrected AIC)*, and BIC (Bayesian information criterion):

$$\begin{aligned} \mathrm{AIC} &= -2\mathrm{LL}^* + 2K, \\ \mathrm{AIC_C} &= -2\mathrm{LL}^* + \tfrac{N(N+K)}{N-K-2}, \\ \mathrm{BIC} &= -2\mathrm{LL}^* + K\ln(N). \end{aligned} \tag{4.7}$$

Table 4.1: Flowchart for the ML-KF estimation, inference, and selection.

Step 1	• Load data (Y_t, X_t) for $t = 1, \ldots, N$ • Setup the parameter vector to be estimated, θ • Create time-invariant system matrices $\{Z, T, c, d, H, Q\}$ using θ
Step 2	• Initialize the state mean a_0 using one of the elements of θ • Specify $P_0 = \kappa I$, where κ is a constant (10–100 times a_0), and I is an identity matrix
Step 3	**Kalman filter** • For $t = 1, \ldots, N$ ◦ Create time-varying system matrices $\{Z_t, T_t, c_t, d_t, H_t, Q_t\}$ using θ and (t, Y_{t-k}, X_t), $k \geq 1$. Don't use Y_t as it is not observed as yet. **Time update** ◦ Compute $a_{t\|t-1}$ using (3.1) ◦ Compute $P_{t\|t-1}$ using (3.2) ◦ Compute the Kalman gain factor using (3.6) **Measurement update** ◦ Obtain a_t using (3.3) ◦ Obtain P_t using (3.5) **Likelihood contribution** ◦ Compute $\mu_t = Z_t a_{t\|t-1} + c_t$ ◦ Compute the forecast error $e_t = Y_t - \mu_t$. *Use* Y_t *in this step only.* ◦ Compute $\Sigma_t = Z_t P_{t\|t-1} Z_t' + H_t$ ◦ Compute $l_t = -0.5[\ln(\det(\Sigma_t) + e_t' \Sigma_t^{-1} e_t)]$ • Do next t • Return LL $= \Sigma_{t=1}^{N} l_t$. It is the likelihood function in (4.2) for state space models.

(*Continued*)

Table 4.1: (*Continued*)

Step 4 **Maximum likelihood Estimation**

- Specify the starting values θ_0
- Specify the tolerance 10^{-5}
- Use BFGS algorithm to maximize LL
- BFGS returns the solution: the maximized LL*, the parameter values θ^*, and the Hessian H at convergence
- Check convergence
    - ○ Is the largest slope $\frac{\partial \text{LL}}{\partial \theta_i}$ across all the element of θ smaller than the tolerance? If so, convergence is attained.
    - ○ If convergence failed,
        - ▪ Relax tolerance till you find good θ_0 and then tighten it
        - ▪ Specify different θ_0 randomly
        - ▪ Rescale $(Y_t,\, X_t)$ so that the elements of θ have similar magnitudes; **this fix is the most important one.**
        - ▪ Use DFP or BHHH algorithm instead of BFGS to maximize LL
        - ▪ Use derivative-free methods to maximize LL and use its solution as θ_0
        - ▪ Use EM algorithm to maximize E[LL] (see Shumway and Stoffer [2011]) and use its solution as θ_0. That is because EM does not provide Hessian at convergence, so standard errors unavailable, and hence this ML step is required.
- Ensure different starting values do not yield larger LL*

Inference

- Compute se$(\theta^*) = \text{sqrt}(\text{diag}(-\mathcal{H}^{-1}))$
- Use se(θ^*) to obtain confidence intervals, t-values, and p-values.
- Use (4.6) to obtain robust standard errors, if desired

(*Continued*)

Table 4.1: (*Continued*)

Selection

- Across multiple models (need not be nested),
    - ○ Repeat Kalman filter ML steps above
    - ○ Compute AIC, $\mathrm{AIC_C}$, BIC in (4.7)
    - ○ Retain the model with the smallest score on a criterion
        - ▪ If AIC, $\mathrm{AIC_C}$, BIC all retain the same model, you attained convergent validity
        - ▪ If not, use BIC when $\dim(\theta)/N$ is small; AIC when $\dim(\theta)/N$ is moderate; $\mathrm{AIC_C}$ when $\dim(\theta)/N$ is large.

In (4.7), LL^* denotes the maximized log-likelihood value, and K and N are the number of parameters and sample size, respectively. All criteria share the common tenet: improve goodness-of-fit and employ fewer parameters. They operationalize goodness-of-fit by the maximized log-likelihood value (that is, the first term) and impose a penalty for excessive parameters (that is, the second terms in (4.7)).

In practice, we compute a score for each model and then the "best" model is the one that attains the smallest score on an information criterion. If the difference in scores on a criterion for any two models exceeds 2, then the model with the larger score is rejected [Burnham and Anderson, 2002, p. 70]. If all three criteria select the same model as the best one, then we achieve convergent validity, which enhances our confidence in the retained model. But if various criteria select different models as the best one, we need to know which criterion to rely on to retain a model. To this end, we note that AIC and $\mathrm{AIC_C}$ possess the efficiency property; hence they perform better when models involve many parameters and small sample sizes (that is, large K/N ratio). Furthermore, AIC_C outperforms AIC as the K/N ratio increases. In contrast, BIC possesses the consistency property; it performs better when models involve few parameters and large sample sizes (that is, small K/N ratio). Thus use BIC for small K/N ratio; AIC for moderate K/N ratio; and AIC_C for large K/N ratio. For further discussions on the efficiency versus consistency properties, see McQuarie and Tsai [1998, p. 3] and Naik et al. [2007, Remark 2].

Table 4.1 provides the flowchart for KF-ML estimation, inference, and model selection. It also states the convergence criterion and suggests ways to tackle the lack of convergence.

5

Optimal Control Theory

In the previous section, we learnt how to estimate the model parameters, assess their statistical significance, and select an appropriate model describing the dynamic system. Using this information, we can forecast the future outcomes of the unobserved states and the observable dependent variables assuming that we know the course of actions planned for the future. In this section, we will focus on how to determine the "best" future course of actions to drive the dynamic system from the current states to the desired ones that attain certain goals such as maximizing brand profit or consumer utility.

To fix ideas, consider a brand with current sales level $x(t)$ that seeks to maximize total profit. Let $u(t)$ denote the advertising expenditure at each time t. (Although I use "advertising" as the control variable for exposition, state space models have been applied to other marketing mix variables such as price, promotions, salesforce size, distribution intensity, detailing, sampling, paid search advertising, company-owned Websites, or earned-media advertising (for example, Facebook likes), or experience goods or logit demand models — (see Ataman et al. [2008], Ataman et al. [2010], Montoya et al. [2010], Rutz and Bucklin

[2011], Aravindakshan et al. [2015], Chintagunta and Rao [1996], Chintagunta et al. [1993]). A brand manager can spend advertising dollars over the planning horizon $[t_0, T]$ in infinitely many ways, each plan resulting in a different stream of profits. The best course of action — the optimal advertising plan amongst numerous alternative ones — is the one that maximizes the net present value of profit streams, given by $J(u(t)) = \int_{t0}^{T} e^{-\rho(t-t_0)}[mx(t) - u(t)]dt + S(x(T))$, where ρ denotes the discount rate, m is the price-cost margin, and $S(\cdot)$ represents the salvage value function. The term $[mx(t) - u(t)]$ measures the net profit at each instant t; the integral computes its discounted sum; and the last term evaluates the salvage value of the terminal sales level. Because *infinitely* many alternative plans exist, the brand manager cannot enumerate all of them and sort over the net present value $J(u)$ to find the optimal ad spending plan that yields the largest J. (Although I use “profit maximization” as the objective for exposition, not-for-profit firms seek to attain other objectives, for example, balancing the shortage and excess of blood collection (see Aravindakshan et al. [2015]). Hence, we need a framework that offers constructive steps to obtain $u^*(t)$ that maximizes *any* $J(u(t))$ over all admissible trajectories of $u(t)$. To derive such a best course of actions, the optimal control theory provides the constructive steps via Pontryagin's maximum principle or the Bellman's optimality principle.

5.1 Pontryagin's maximum principle

For the sake of exposition, we consider an infinite rather than finite horizon problem, let $x(t)$ denote the current state, $u(t)$ be any admissible control trajectory, and $F(t, x(t), u(t))$ be a general objective function that also explicitly depends on time, the control applied at time t, and the resulting state. In the above sales example, $F(t, x(t), u(t)) = e^{-\rho t}[mx(t) - u(t)]$. The decision-maker seeks a solution to the following optimization problem:

$$\underset{u(t)}{\text{Max}}\, J(u(t)) = \int_0^\infty F(t, x(t), u(t))dt, \tag{5.1}$$

subject to the state dynamics

$$\frac{dx(t)}{dt} = g(t, x(t), u(t)), \tag{5.2}$$

starting with $x(0) = x_0$.

In the absence of control theory, one would need to solve the differential Equation (5.2), substitute the solution $x(t) = p(t, u(t), x_0)$ in (5.1) to eliminate $x(t)$, then integrate $F(t, p(t, u(t), x_0), u(t))$ from zero to infinity so as to obtain $J(u(t))$ as a function of the control trajectory, and then maximize it across infinitely many possible time paths of $u(t)$. Rather than resorting to such a direct maximization, we invoke the Pontraygin's Maximum Principle, which is simple to learn and implement.

Let us abstract some important details for the sake of presenting the essence of the maximum principle. Using the objective function $F(\cdot)$ and the state dynamics $g(\cdot)$, we first construct the Hamiltonian function as follows:

$$H = F + \lambda g, \tag{5.3}$$

where λ is known as the "co-state" or "adjoint" variable similar to the Lagrange multiplier in static optimization. Then, the maximum principle offers the following three first-order conditions:

$$\frac{\partial H}{\partial u} = 0, \tag{5.4}$$

$$\frac{\partial H}{\partial \lambda} = \frac{dx(t)}{dt}, \tag{5.5}$$

$$\frac{\partial H}{\partial x} = -\frac{d\lambda(t)}{dt}. \tag{5.6}$$

Equation (5.4) is the usual first-order condition with respect to the decision variable (that is, the control variable u); Equation (5.5) recovers the state dynamics in (5.2) because the left-hand side $\frac{\partial H}{\partial \lambda} = g(\cdot)$ upon differentiating (5.3); and Equation (5.6) specifies the evolution of the co-state variable $\lambda(t)$. The Hamiltonian in (5.3) and the triplet Equations (5.4)–(5.6) solve the dynamic maximization problem stated in (5.1) and (5.2).

The elegance of the maximum principle lies in the construction of the Hamiltonian function. The principle requires us to change our perspective: we do not view the dynamic optimization problem as dynamic; rather we focus on the instantaneous static optimization of (5.3). It's analogous to viewing a video (dynamic) as a sequence of still shots (static).

To emphasize this view, we suppress the time argument in (3.9) even though the Hamiltonian is a function of time (both directly due to t in F and indirectly due to $(x(t), u(t), \lambda(t))$. Then, for each instant t (that is, a particular still shot), we construct a function — the Hamiltonian viewed as frozen in time — by adjoining the dynamic constraint $g(\cdot)$ to the objective function $F(\cdot)$. Then, we apply the static optimization principles to maximize the objective $F(\cdot)$ at the instant t, taking into account the constraint that represents instantaneous increase in the "quantity" of the state variable due to $g(\cdot)$, and valuing this incremental state's contribution by the associated "shadow price" λ. In other words, the interpretation of the Hamiltonian function is that Hdt is the total contribution to J when $x(t) = x$ and $u(t) = u$ over the small interval $(t, t + dt)$. And $\lambda(t)dx$ is the valuation of the state's contribution to J Finally, we join the series of static optimization solutions using the time path of the adjoint variable $\lambda(t)$, which is obtained by solving the differential equation in (5.6). Thus, this construction of the Hamiltonian function decouples the overall dynamic optimization into a series of static optimization problems (which are easier to solve).

The implementation of the maximum principle is easy. We first construct the Hamiltonian function from the given objective function and state dynamics, introducing an unknown variable $\lambda(t)$. Focusing on (5.3), we differentiate the Hamiltonian with respect to the control, set it to zero, and solve for the "optimal control" as a function of the state variable $x(t)$, the co-state variable $\lambda(t)$, and the model parameters arising from (5.1) and (5.2). The resulting optimal control is not fully characterized as yet because it depends on the unknown co-state variable. To this end, we solve simultaneously the two differential equations — state and costate dynamics — induced by the optimal control trajectory. The state dynamics (forward pass) begins from the initial condition $x(0) = x_0$; the costate dynamics (backward pass) begins from the terminal point $\lambda(\infty) = 0$. Because the initial and terminal conditions are fixed, the resulting problem is known as Two-Point Boundary Value Problem (TPBVP). Analytical closed-form solutions to TPBVPs are rarely available, requiring the use of numerical methods (for which readers are directed to Lapidus and Pinder [1999], Smith [2003], and Naik et al. [2005]). Furthermore, if the planning horizon were finite (for example, Raman [2006] and Bass et al. [2013]), then the terminal value $\lambda(T) = \frac{\partial S(x(T))}{\partial x}$, where $S(x(T))$ furnishes the salvage value of the terminal state. Substituting the state and costate solutions in the Hamiltonian maximizing optimal control, we fully characterize the best course of action to attain the goal in (5.1) subject to the state dynamics in (5.2). Finally, the second-order conditions for optimality requires the maximized Hamiltonian to be concave in x for each λ and t, and salvage function $S(x)$ be concave in x (for weaker requirements, see Seierstad and Sydsaeter [1977]).

The above exposition abstracted important details such as continuity of the functions, transversality conditions, state and/or control constraints among others, which can be found in the lucid tutorial by Sethi and Thompson [1981] and the classic books by Kamien and Schwartz [2012] and Sethi and Thompson [2000]. These latter books also provide several applications of the optimal control theory to economics, management science, finance, and marketing.

5.2 Bellman's principle of optimality

The Hamiltonian approach splits the overall dynamic optimization in (5.1) and (5.2) into static optimization problems one for each instant t, constructs the Hamiltonian function, and determines the optimal solution by solving simultaneously the state and costate differential equations as a two-point boundary value problem. In contrast, the Bellman approach splits the overall dynamic optimization into two periods from $(t_0, t_0+\Delta t)$ and $(t_0+\Delta t, \infty)$ — rather than a sequence of instantaneous static problems — constructs the Value function, and determines the optimal solution by solving one partial differential equation known as the Hamilton–Jacobi–Bellman (HJB) equation.

For the sake of presenting the essence of the Bellman approach, let us define the Value function as follows:

$$V(t_0, x_0) = \operatorname*{Max}_{u(t)} \int_{t_0}^{\infty} F(t, x(t), u(t))dt, \tag{5.7}$$

subject to the state dynamics in (5.2). Equation (5.7) denotes the *maximized* value of $J(u^*(t))$ in (5.1) under the optimal control $u^*(t)$ starting from the initial time t_0 when the state level is x_0 We can split this overall maximization into two periods to obtain

$$\begin{aligned} V(t_0, x_0) \quad &= \quad \operatorname*{Max}_{u(t)\in(t_0, t_0+\Delta t)} \int_{t_0}^{t_0+\Delta t} F(t, x, u)dt \\ &\quad + \operatorname*{Max}_{u(t)\in(t_0+\Delta t, \infty)} \int_{t_0+\Delta t}^{\infty} F(t, x, u)dt, \end{aligned} \tag{5.8}$$

Then, the principle of optimality states that the solution in the second period starting from $(t_0 + \Delta t, \infty)$ remains optimal starting from any state arrived at $(x_0 + \Delta x)$ from the actions in the first period $(t_0, t_0 + \Delta t)$. Consequently, the solution for the second period problem is given by the value function,

$$V(t_0 + \Delta t, x_0 + \Delta x) = \operatorname*{Max}_{u(t)\in(t_0+\Delta t, T)} \int_{t_0+\Delta t}^{\infty} F(t, x, u)dt$$

Applying this principle, we re-express (5.8) as

$$V(t_0, x_0) = \operatorname*{Max}_{u(t)\in(t_0, t_0+\Delta t)} \left\{ \int_{t_0}^{t_0+\Delta t} F(t, x, u)dt + V(t_0 + \Delta t, x_0 + \Delta x) \right\}, \tag{5.9}$$

and focus on finding the optimal course of actions in the first period. (This notion — solve the last period first and then solve the earlier periods sequentially — is known as the Backward Induction.)

To solve the first period problem, for small $\Delta t \to 0$, note that the first term on the right-hand side of (5.9) can be approximated by $\int_{t_0}^{t_0+\Delta t} F(t, x, u)dt \approx F(t_0, x_0, u)\Delta t$. The second term $V(t_0 + \Delta t, x_0 + \Delta x)$ can be approximated by Taylor series as $V(t_0 + \Delta t, x_0 + \Delta x) \approx V(t_0, x_0) + V_t(t_0, x_0)(t_0 + \Delta t - t_0) + V_x(t_0, x_0)(x_0 + \Delta x - x_0)$, where $V_t(t_0, x_0) = \partial V(t_0, x_0)/\partial t$ and $V_x(t_0, x_0) = \partial V(t_0, x_0)/\partial x$. Substituting them in (5.9) and canceling $V(t_0, x_0)$ from both sides, we get $0 =$

$\text{Max}_u\{F(t_0, x_0, u)\Delta t + V_t(t_0, x_0)\Delta t + V_x(t_0, x_0)\Delta x\}$. By rearranging the last expression, dividing throughout by Δt and taking limits as $\Delta t \to 0$, and noting that $\Delta x/\Delta t = dx/dt = g(t, x, u)$ from (5.2), we obtain the HJB equation, which is also known as the Bellman equation:

$$-V_t(t, x) = \underset{u}{\text{Max}}\{F(t, x, u) + V_x(t, x)g(t, x, u)\}. \tag{5.10}$$

We dropped the zero subscript in (5.10) since we are at the initial condition looking forward at the vanishingly small first period $\Delta t \to 0$. Because Equation (5.10) involves time and state derivatives, it is a partial (as opposed to ordinary) differential equation that the value function $V(t, x)$ obeys.

To implement the Bellman approach in practice, we take the following steps. First, apply the static optimization principles to maximize the parenthetical term in (5.10) with respect to u; then obtain the optimal u^* as a function of (t, x, V_x) and substitute it back in (5.10) so as to remove the max operator (because the HJB is now driven by the maximizing control); next solve the partial differential equation: $-V_t(t, x) = F(t, x, u^*(t, x, V_x)) + V_x(t, x)g(t, x, u^*(t, x, V_x))$; and finally differentiate the resulting solution $V(t, x)$ with respect to x to obtain $V_x(t, x)$ and substitute this result in $u^*(t, x, V_x)$ to yield the optimal control $u^*(t, x)$. In most economics and marketing problems, the objective function $F(t, x, u)$ takes the separable form $e^{-\rho t}f(x, u)$; consequently the optimal control depends only on the state such as $u^*(x)$.

As before, this exposition abstracted important details that can be found in Kamien and Schwartz [2012], Sethi and Thompson [2000], Dockner et al. [2000], and Jørgensen and Zaccour [2004]. They also provide several applications in the context of discrete-time, stochastic dynamics, or competitive markets.

5.3 When to use which approach?

The answer depends on the context of the problem. But first, we need to appreciate that the Hamiltonian and Bellman approaches are intimately connected with each other and with the classical calculus of variation problem alluded in Section 1. Specifically, the "splendid" problem as Leibnitz calls it and proposed by Bernoulli can be expressed as maximizing the integral $\int_0^\infty F(t, x(t), x'(t))dt$ with respect to the trajectory $x(t)$ subject to the initial and terminal conditions. Note that the integrand involves both the level and slope $x' = dx/dt$. This formidable problem is a special case of the Hamiltonian approach when we replace Equation (5.2) with $dx/dt = g(t, x, u) = u(t)$. Then by applying the maximum principle via (5.4)–(5.6), we recover the optimal solution, which is identical to the one obtained via the celebrated Euler equation. In other words, the calculus of variation problem is nested in the Hamiltonian approach with state dynamics driven by the control linearly.

Next, to see the equivalence between the Hamiltonian and Bellman approaches, compare Equation (5.3) with the right-hand side of (5.10) to equate $\lambda(t)$ and $V_x(t, x)$. Now differentiate the HJB equation $-V_t(t, x) = F(t, x, u^*) + V_x(t, x)g(t, x, u^*)$ with respect to x to obtain

$$\begin{aligned} -V_{tx}(t, x) &= F_x(t, x, u^*) + V_{xx}(t, x)g(t, x, u^*) + V_x(t, x)g_x(t, x, u^*) \\ -[V_{tx}(t, x) + V_{xx}(t, x)g(t, x, u^*)] &= F_x(t, x, u^*) + V_x(t, x)g_x(t, x, u^*), \end{aligned} \tag{5.11}$$

where the double subscripts denote the second partial derivatives. Equation (5.11) reveals an insight: its left-hand side equals $V_{tx}(t, x) + V_{xx}(t, x)dx/dt = \frac{\partial V_x(t,x)}{\partial t} = \frac{d\lambda}{dt}$, whereas its right-hand side equals $F_x(t, x, u^*) + \lambda(t)g_x(t, x, u^*) = \frac{\partial H}{\partial x}$ since $H = F + \lambda g$ by construction. Thus Equation (5.11) implies $-\frac{d\lambda}{dt} = \frac{\partial H}{\partial x}$, which is none other than the co-state equation in (5.6)! In other words, by equating the costate variable $\lambda(t)$ with the marginal value function $V_x(t, x)$, we arrive at the first-order condition of the maximum principle (Equation (5.6)). In addition, we reaffirm the interpretation of costate variable $\lambda(t)$ as "shadow price" because it equals the marginal contribution due to incremental state changes (that is, V_x). Thus the Hamiltonian and Bellman approaches lead us to the same summit of the Everest, but from two different sides: the paths differ, but attain the same peak.

The paths differ conceptually. The Hamiltonian approach provides the optimal control $u^*(t)$ as a function of time, a solution referred to as "open-loop." Whereas the Bellman approach provides the optimal control $u^*(x)$ as a function of state, a solution referred to as "closed-loop." The open-loop solution is invaluable for *planning* the future course of actions; it implies commitment to stay the course; and it is proactive. In contrast, the closed-loop solution is invaluable for *responding* to alternative states; it implies flexibility to accommodate feedback from whatever the current state is (optimal or not); but it is reactive.

To exemplify the open loop strategy, envision the associate dean's task of scheduling courses/instructors over various quarters of the next year; or hundreds of millions of dollars that companies spend on media buys in the upfront market a year in advance [Tellis, 1998, Belch and Belch, 2004]. To illustrate the closed loop strategy, imagine the dean's task of making a retention offer in wake of counter-offers to a faculty member from competing schools; or hundreds of millions of drivers who use global positioning system like Garmin to navigate via the initial route guidance (that is, open-loop plan) and then "re-calculating" a revised route (that is, closed-loop) based on the current state resulting from a missed exit or turns in the original plan.

In practice, consider using the Hamiltonian approach for making proactive plans over time, for example pricing strategy [Kalish, 1983] or media plans [Naik and Raman, 2003]. The Bellman approach is better suited to tackle competition in dynamic markets [Naik et al., 2005, 2008] or dynamic continuous uncertainty [Raman and Chatterjee, 1995, Raman and Naik, 2004] or dynamic discrete uncertainty [Rubel et al., 2011]. Both the approaches handle continuous-time (as above) and discrete-time formulations (for details, see Appendix B in Sridhar et al. [2011]). We next illustrate how to apply the two approaches to solve dynamic marketing problems.

6

Marketing Applications

We present the applications of Optimal Control Theory, Differential Games Theory, and Stochastic Control Theory using three examples from the recent marketing literature. The first example solves an optimal control problem for deterministic dynamics under single state and two controls; the second analyzes dynamic competition between two brands that use advertising to control awareness dynamics; the third example examines the impact of discrete uncertainty due to product harm crisis looming on the horizon.

6.1 Multimedia allocation: optimal control theory

This example is based on Naik and Raman [2003], who study the allocation of advertising budget to N-media activities. For brevity, consider advertising on two media $u_1(t)$ and $u_2(t)$ that grows brand sales over time as follows: $\frac{dx}{dt} = \beta_1\sqrt{u_1} + \beta_2\sqrt{u_2} + \kappa\sqrt{(u_1 u_2)} - \delta x$. The square root function captures the notion of diminishing returns to advertising, which means the impact of advertising increases at a decreasing rate. The parameter κ denotes the synergy between two media: the effectiveness of each medium increases in the presence of the other media. The parameter δ represent the attrition of sales in the absence of advertising.

The brand manager wants to maximize the net present value of the profit stream over infinite horizon, which is given by $J(u_1, u_2) = \int_0^\infty e^{-\rho t}(mx(t) - u_1(t) - u_2(t))dt$. The parameter m indicates the price-cost margin in each unit of sales, and ρ denotes the discount rate of the forward-looking manager. A large (small) ρ corresponds to a manager who is more (less) impatient and present-oriented (future-oriented). There are infinitely many admissible trajectories of $u_i(t) \in (0, \infty), i = 1, 2$, each resulting in some net present value $J(u_1, u_2)$ of the discounted profit stream. Our task is to find dynamically optimal trajectories that yield the largest net present value $J(u_1^*, u_2^*)$, taking into account cross-media synergy, ad effectiveness, sales carryover effect, product margin, and discount rate.

To this end, let us apply the Maximum Principle. We first construct the Hamiltonian $H = F + \lambda g$ and re-express it as the "current value" Hamiltonian $\tilde{H} = e^{\rho t}H$. Then $\tilde{H} = (mx - u_1 - u_2) + \tilde{\lambda}(\beta_1\sqrt{u_1} + \beta_2\sqrt{u_2} + \kappa\sqrt{u_1 u_2} - \delta x)$, where the current value co-state $\tilde{\lambda} = e^{\rho t}\lambda$. We then use Equation (5.4) to find the Hamilton-maximizing controls by differentiating $\tilde{H}$ with respect to the two controls $(u_1 u_2)$. Specifically, $\frac{\partial \tilde{H}}{\partial u_1} = -1 + \frac{\tilde{\lambda}(\beta_1 + \kappa\sqrt{u_2})}{2\sqrt{u_1}}$, which when set to zero yields $2\sqrt{u_1} = \lambda\beta_1 + \tilde{\lambda}\kappa\sqrt{u_2}$. Similarly, $\frac{\partial H}{\partial u_2} = 0$ yields $2\sqrt{u_2} = \lambda\beta_2 + \tilde{\lambda}\kappa\sqrt{u_1}$. By solving these two linear simultaneous equations in $(\sqrt{u_1}, \sqrt{u_2})$, we find the optimal controls as $u_i^* = [2\tilde{\lambda}\beta_i + \tilde{\lambda}^2\kappa\beta_j)/(4 - 2\tilde{\lambda}^2\kappa^2)]^2$.

Next, to eliminate the current costate variable $\tilde{\lambda}$, we do a little bit of algebra and translate Equation (5.6) to $\frac{d\tilde{\lambda}}{dt} = \rho\tilde{\lambda} - \frac{\partial \tilde{H}}{\partial x} = \rho\tilde{\lambda} - [m + \tilde{\lambda}(-\delta)] = (\rho + \delta)\tilde{\lambda} - m$. A stable solution is given by $\tilde{\lambda} = m/(\rho + \delta)$ Finally, we substitute it in $u_i^* = [2\tilde{\lambda}\beta_i + \tilde{\lambda}^2\kappa\beta_j)/(4 - 2\tilde{\lambda}^2\kappa^2)]^2$ to obtain the optimal spending on each of the two media as a function of all model parameters in closed-from. Thus we determine the total budget $\sum u_i^*$ and the optimal allocation ratio u_1^*/u_2^*. For details, see Naik and Raman [2003].

They further generalize these results to N media. But more importantly, first, using market data on sales and advertising, they furnish evidence on the existence of synergy κ empirically. Second, using comparative statics, they discover the counter-intuitive result: as synergy increases, the optimal allocation ratio tilts in the favor of the *less* effective medium. That is, the marginal dollar should be allocated to the weak medium (rather than the stronger one). This result in the presence of synergy is in stark contrast with the one in the absence of synergy that suggests a larger budget should be allocated to the more effective medium (see Dorfman and Steiner [1954]).

6.2 Competitive models: differential games theory

This example is based on Naik et al. [2008], who study the role of competition when mature brands advertise to build awareness. They study a general model of N brands competing to build awareness over time in the presence of marketing expansion and confusion effects (own advertising builds awareness of other brands), providing both empirical parameter estimation and analytical closed-form solutions. For illustrating the Bellman approach, however, let us examine a simple case of two symmetric brands (that is, equally strong brands with identical parameters), who compete for market share, as in Sorger [1989] for example, where own advertising $u_1(t)$ and competitive advertising $u_2(t)$ influence market shares over time as follows: $\frac{dm_1}{dt} = \beta u_1(t)\sqrt{1 - m_1(t)} - \beta u_2(t)\sqrt{m_1(t)}$, and $\frac{dm_2}{dt} = \beta u_2(t)\sqrt{1 - m_2(t)} - \beta u_1(t)\sqrt{m_2(t)}$. These equations state that own brand's share grows because own advertising acts on the consumers of the other brand, and own brand share decreases because competitive advertising steals share proportional to own market share. The parameter β denotes ad effectiveness. When we add both the equations, we see that $\frac{dm_1}{dt} + \frac{dm_2}{dt} = 0$ and hence the logical consistency property $m_1(t) + m_2(t) = 1$ for every t holds.

A brand manager wants to maximize the net present value $J(u_i) = \int_0^\infty e^{-\rho t}(Rm_i(t) - c(u_i(t)))dt$. The parameter ρ denotes the discount rate of the forward-looking manager; R indicates the category revenues per unit share; and $c(u_i) = u_i^2$ is the convex cost function. The convex costs in the objective function are equivalent to diminishing returns in the state dynamics; see details in Naik et al. [2008, p. 135].

Unlike the previous example, the brand manager $i, i = 1, 2$, has only one control (own advertising), and s/he competes with the other brand manager to gain a larger market share. Each one seeks to maximize own net present value $J(u_i)$ taking into account the other brand's best course of action u^*_{-i} subject to the market share dynamics resulting from the mutual decisions. There are infinitely many admissible trajectories of $u(m_i) \in (0, \infty)$, and our task is to find the dynamically optimal response such that neither brand manager has any unilateral incentive to alter the course of action. This solution concept is known as the Markov Perfect Nash equilibrium [Başar and Olsder, 1999], and it explores all closed-loop strategies in continuous time and continuous state.

To discover the optimal strategy, let us apply the Bellman's principle of optimality. We first construct the value function $V(m_{0i}) = \text{Max } J(u_i)$ starting from any initial state m_{0i}. The objective function is separable in the time argument, that is, $V(m) = \text{Max} \int_0^\infty F(t, m, u)dt = \text{Max} \int_0^\infty e^{-\rho t} f(m, u)dt$. Hence, upon differentiating with respect to time, $\frac{\partial V}{\partial t} = (-\rho)V(m)$. Furthermore, we only need one co-state variable in the value function because $m_1(t) + m_2(t) = 1$. Consequently, for brand 1, the HJB equation in (5.10) becomes

$$\rho V_1 = \underset{u_1}{\text{Max}} \left\{ \left[Rm_1 - u_1^2 \right] + V_{m_1} \left[\beta u_1 \sqrt{1 - m_1} - \beta u_2 \sqrt{m_1} \right] \right\}.$$

Similarly, for brand 2,

$$\rho V_2 = \underset{u_2}{\text{Max}} \left\{ \left[Rm_2 - u_2^2 \right] + V_{m_2} \left[\beta u_2 \sqrt{1 - m_2} - \beta u_1 \sqrt{m_2} \right] \right\}.$$

Here $V_{m_i} = \partial V / \partial m_i$.

Next, to eliminate the max operator, let us differentiate the term in the curly brackets with respect to u_1 and obtain $(-2u_1 + V_{m_1} \beta \sqrt{1 - m_1})$, which when set to zero yields the "Hamiltonian" maximizing control $u_1^* = 0.5 V_{m_1} \beta \sqrt{1 - m_1}$. Similarly, we get $u_2^* = 0.5 V_{m_2} \beta \sqrt{1 - m_2}$.

If we knew V_{m_i} as a function of the model's parameters, then we would have determined the feedback control that depends on the current state (that is, share m_i). To this end, we conjecture that a linear value function satisfies the Bellman equation. (Although we skip it here, this conjecture can be verified as true.) Given symmetric brands, we let the value function $V(m_i) = a + bm_i$ so that $V_{m_i} = b$, where (a, b) depend on the parameters of model dynamics and profit function.

Then, we substitute the resulting $u_1^* = 0.5b\beta\sqrt{1-m_1}, u_2^* = 0.5b\beta\sqrt{1-m_2}$, and $V(m_i) = a + bm_i$ in the HJB equation $\rho V_1 = [Rm_1 - u_1^2] + V_{m_i}[\beta u_1\sqrt{1-m_1} - \beta u_2\sqrt{m}_1]$. Upon simplification and equating the coefficients on both sides of the value function, we find that $a = \beta^2 b^2/(4\rho)$, and the positive root of the quadratic equation in b is given by $b = (-2\rho + \sqrt{4\rho^2 + 12R\beta^2})/(3\beta^2)$, thereby fully characterizing the conjectured value function $V(m_i) = a + bm_i$. Thus, the dynamically optimal closed-loop (that is, feedback) Nash equilibrium strategy for each brand i is given by $u^*(m_i) = \alpha\sqrt{1-m_i}$, where $\alpha = (-2\rho + \sqrt{4\rho^2 + 12R\beta^2})/(6\beta)$.

Because optimal advertising in competitive and dynamic markets follows $u^*(m_i) = \alpha\sqrt{1-m_i}$, we learn the counter-intuitive result known as the Inverse Allocation Principle: *The smaller the market share, the larger the advertising.*

Jones [1986, 1990] furnishes empirical evidence to corroborate this inverse allocation principle, noting that "... for large brands, the market share normally exceeds the advertising share; for smaller brands, the opposite is true" Jones [1986, p. 100]. Furthermore, the inverse allocation principle is not the consequence of our simplifying assumption of symmetric brands; it holds in more general settings with asymmetric brands as well as in the presence of N-brand oligopoly of asymmetric brands, market expansion, and confusion effects (see the closed form analytical results in Naik et al. [2008]).

Substantively, this allocation principle is opposite of the competitive parity method recommended in textbooks. Specifically, the competitive parity method requires matching the ratio of own advertising to market share with that of the other brand, thereby suggesting that larger (smaller) brands should spend more (less) on advertising. In contrast, this inverse allocation principle advises managers to build dominant brands because they would face less competitive resistance in the long run and thus will able to reduce advertising. From a life-cycle perspective, small up-and-coming brands should spend more on advertising, whereas mature brands may “fly on auto pilot” without much advertising and instead rely on the momentum of brand purchases and positive consumption experience.

6.3 Dynamic uncertainty: stochastic control theory

Both the above examples assume that the state dynamics are deterministic, that is, not perturbed by random events over time. To introduce uncertainty in continuous time, we perturb the state dynamics using the Brownian motion, known as the Wiener process and denoted by $W(t)$, whose time increment dW follows a normal distribution with zero mean and unit variance. The resulting stochastic state dynamics is expressed by $dx = g(t, x, u)dt + \sigma dW$. When $\sigma = 0$, we recover the usual ordinary differential equation as in (5.2). To incorporate the effect of uncertainty in optimal control strategy, the standard HJB equation in (5.10) becomes

$$-V_t(t, x) = \underset{u}{\mathrm{Max}}\{F(t, x, u) + V_x(t, x)g(t, x, u) + 0.5\sigma^2 V_{xx}\},$$

where $V_{xx} = \partial^2 V/\partial x^2$. Thus the presence of uncertainty transforms the usual HJB Equation (5.10) from a first-order partial differential equation to a second-order partial differential equation.

In Marketing, applying this stochastic HJB equation, Raman and Chatterjee [1995] study optimal pricing under demand uncertainty; Prasad and Sethi [2009] examine the optimal budgeting and allocation for integrated marketing campaigns in the presence of cross-media synergy and multi-brand competition; Raman and Naik [2004] investigate the long-term profit impact of integrated marketing communications programs; among others. For discrete-time stochastic control problems, Esteban-Bravo et al. [2014] study how to allocate budgets optimally in a customer relationship management application using stochastic dynamic programming approach.

The continuous-time uncertainty via the Wiener process $W(t)$ represents small shocks at each instant whose net impact, on average, is zero. In contrast, the possibility of product harm crisis induces uncertainty of a discrete event, whose net impact is often not zero (see, for example, van Heerde et al. [2007], Liu and Shankar [2015], and Kalaignanam et al. [2013]). To study uncertainty from such rare catastrophic events, Rubel et al. [2011] present a general framework to incorporate shocks that are large and discrete rather than small and continuous. The resulting dynamic model augments the above control-theoretic models by introducing a random stopping problem, thereby paving the way to study optimal decisions in the presence of rare but catastrophic events.

Consider a product harm crisis that strikes at an unknown instant T in the future with probability χ given that it has not occurred as yet. When it occurs, the planning horizon splits into two regimes: pre-crisis regime $[0, T)$ and post-crisis regime $[T, \infty)$. Brand sales grow according to the dynamics $dS/dt = \beta_j \sqrt{u_j(t)} \sqrt{M(t) - S(t)} - \delta_j S(t)$, where $j = (1, 2)$ denote the pre- or post-crisis regimes, $M(t)$ is the time-varying market potential for the brand, and $(\beta_j,\ \delta_j)$ are ad effectiveness and attrition rate in regime j, respectively. This equation says that advertising acts on the untapped market $(M - S)$, ad spending via $\sqrt{u}$ captures the diminishing returns, and brand sales decay proportional to its level in the absence of advertising.

At $t = 0$ the initial sales is S_0, When the crisis occurs at $t = T$, the prevailing sales level incurs a "damage." Consequently, brand sales right after the crisis is $S(T^+) = (1 - \phi)S(T^-)$, where $S(T^+)$ and $S(T^-)$ represent sales just after and before the crisis, and the fraction ϕ denotes the damage rate: the larger the damage rate, the sharper the drop in the baseline sales.

The net present value of the stream of profits from the pre-crisis regime is $J(u_1) = \int_0^T e^{-\rho t} \pi(S(t), u_1(t)) dt$ and from the post-crisis regime is $J(u_2) = \int_T^\infty e^{-\rho t} \pi(S(t), u_2(t)) dt$, where ρ is the discount rate, and $\pi(S, u) = mS - u$ yields the instantaneous profit with margin m per unit sales. Because the crisis event occurs at the random time T, the profit integrals $J(u_1)$ and $J(u_2)$ are random variables, rendering this formulation a stochastic control problem. The brand manager wants to maximize the total profit $J(u_1, u_2) = E[J(u_1) + e^{-\rho T} J(u_2)]$, where the expectation is evaluated over all possible crisis times $T \in (0, \infty)$ after discounting $J(u_2)$ back to the initial time $t = 0$ because the post-crisis net "present" value accrues at time T.

By applying integration by parts to this random stopping problem (because the first regime "stops randomly" when the crisis occurs), Rubel et al. [2011] evaluate the profit expectation analytically to show that $J(u_1, u_2) = \int_0^\infty e^{-(\rho+\chi)t}\{\pi(S, u_1) + \chi J(u_2)\}dt$. Note that the discount factor increases from ρ to $(\rho+\chi)$, revealing a novel insight: *crisis anticipation enhances impatience.*

Having evaluated the profit expectation in closed-form, we can now apply the standard control theory to maximize $J(u_1, u_2)$ with respect to the decision variables (u_1, u_2). We apply the Bellman approach to find optimal feedback advertising strategies, which possess the desirable property of subgame perfectness. That is, the optimal advertising spending is optimal not only when sales evolve along the optimal state trajectory, but also when sales depart from the optimal trajectory at any time. This property is also known as Markov perfectness or strong time consistency (see Başar and Olsder [1999]). Appendix in Rubel et al. [2011] provides the details of solving the HJB equation to obtain closed-form results. The resulting optimal feedback advertising strategy in each regime j is given by $u_j^*(S) = (M - S)(0.5\beta\lambda_j)^2$, where $\lambda_2 = (-2(\rho + \delta_2) + 2\sqrt{(\rho + \delta_2)^2 + m_2\beta_2^2})/\beta_2^2$ and $\lambda_1 = (-2(\rho+\delta_1+\chi)+2\sqrt{(\rho + \delta_1 + \chi)^2 + (m_1 + \chi(1 - \phi)\lambda_2)\beta_1^2})/\beta_1^2$.

Because $u^*(S) \propto (M - S)$, contrary to the proportional-to-sales heuristic in textbooks, the optimal strategy recommends the inverse allocation principle: *managers should spend more when sales are low and less when sales are high.* That is, advertise intensively when the untapped market is large.

More importantly, to gain understanding of the effects of crisis likelihood χ, Figure 6.1 illustrates the optimal sales and advertising trajectories in the presence of low and high crisis likelihood. First, it shows that the pre-crisis advertising is small for high χ at every $t < T$. Second, at the crisis time T, advertising increases to recover the drop in

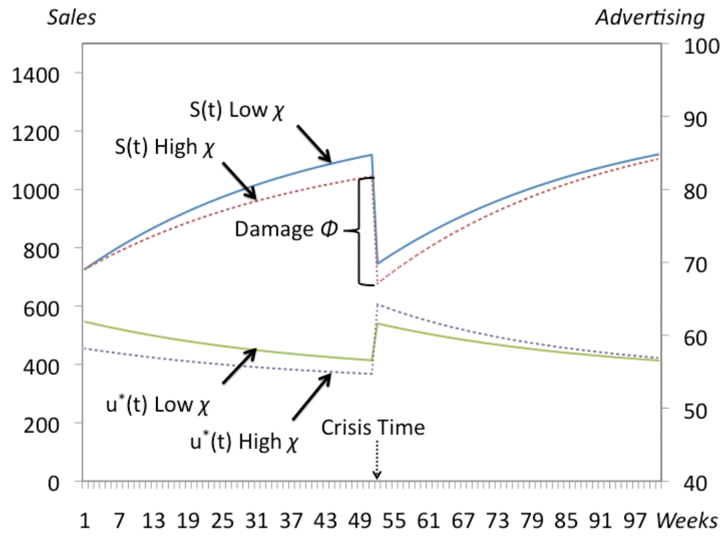

Figure 6.1: Crisis likelihood effects on sales and advertising trajectories.

baseline sales. Third, the post-crisis advertising is large for high χ at every $t > T$. Thus, we analytically discover the *crossover interaction*: u_1 decreases but u_2 increases when managers anticipate crisis. Why? Because they should conserve resources now to recover sales later.

7

Conclusion

The purpose of this primer is to introduce broad principles for solving *any* dynamic marketing problem. To this end, I elucidated four principles: the Kalman Filter, Fisher's Likelihood Principle, Pontryagin's Maximum Principle, and Bellman's Optimality Principle. These principles are general and apply in all dynamics contexts (for example, linear or nonlinear, deterministic or stochastic, discrete- or continuous-time, discrete- or continuous-state) because they do not depend on a specific model specification or data sets. Hence they hold not only across all dynamic marketing models — extant and future — but also across disciplines such as physics, engineering, economics, or management science. Given this generality, readers should learn to master the use of these principles.

The first two principles — the Kalman filter and the Likelihood Principle — allow us to estimate the parameters of the dynamic system that specifies relations among current states, past states, actions, and outcomes. The Kalman filter provides the means and covariances of the unobserved state vector recursively, that is, as the observed data on outcomes and actions unfold over time. When the dynamic system is linear, the Kalman filter is the optimal filter not just when the errors terms are normally distributed, but also when the errors follow any non-normal distribution with finite first two moments. When the dynamic system is non-linear, the Kalman filter is still the optimal filter within the class of all linear estimators. Other nonlinear estimators, for example EKF, Particle Filter, or Unscented KF, are sub-optimal but may improve performance at the expense of extra computational costs.

Using the means and covariances of the unobserved state vector from the Kalman filter, one can next compute the likelihood function for scalar or multivariate time-series, with or without missing values, equally or irregularly spaced observations, random and/or time-varying parameters, and discrete- or continuous-time models. To apply Fisher's likelihood principle, we maximize the likelihood function with respect to the model parameters to obtain the best parameter estimates, where the "best" refers to the most precise estimates (that is, least variance) in the class of all unbiased estimators. Not only the precision of the parameter estimates, but also their cross-correlations across estimated parameters are obtained from the curvature of the likelihood function (that is, negative inverse of the Hessian matrix).

If multiple dynamic systems are estimated, then we score them using information criteria (for example, AIC, $\mathrm{AIC}_{\mathrm{C}}$, BIC) and retain the model associated with the smallest score. Information criteria balance the trade-off between fidelity (that is, improving the goodness-of-fit) and parsimony (that is, employing fewer parameters). If these criteria point to the same model as the best one, then the resulting convergent validity enhances our confidence in the retained model. But if these criteria indicate different models as the best one, then use BIC for small K/N ratio; AIC for moderate K/N ratio; and $\mathrm{AIC}_{\mathrm{C}}$ for large K/N ratio.

Having estimated and retained the best dynamic system that corroborates with market data (that is, descriptive phase), we next seek prescriptive or normative answers to managerial questions: how should managers optimally spend hundred million dollars over time to advertise a brand, or whether 100 million is the "right" sum, how to allocate it across geographic regions or multiple media? To this end, we need to specify the objective function in addition to the dynamic system and then apply either the Maximum Principle or the Optimality Principle to obtain the best course of action or competitive response, respectively.

The Pontryagin's Maximum Principle yields open-loop plans, which are proactive and require commitment over time. It views the dynamic optimization as a sequence of static problems, one for each instant t, constructs the Hamiltonian function, and determines the optimal solution by solving simultaneously the state and costate differential equations as a two-point boundary value problem. It yields the best course of action in terms of total marketing budget and its allocation over time.

In contrast, the Bellman's Optimality Principle yields closed-loop plans, which are reactive and offer flexibility over time. It splits the dynamic optimization into current period and distant future — rather than a sequence of static problems as in the maximum principle — constructs the Value function, and determines the optimal solution by solving a partial differential equation known as the Hamilton–Jacobi–Bellman (HJB) equation. It yields the best competitive response or course correction in terms of total marketing budget and its allocation over time.

Both principles are powerful because they reveal insights that cannot be established numerically. For example, consider the counter-intuitive insights from Section 6:

- as synergy increases, more than fair share of marginal advertising dollar should be allocated to the weaker medium rather than the stronger one;
- in competitive markets, advertising spending is inversely proportional to own market share;
- when anticipating crisis, managers should decrease pre-crisis advertising and increase post-crisis advertising to conserve resources now and recover sales later.

These three insights are analytically proven, which means they hold true for every permissible value of all the model parameters (not just the estimated parameter values based on a specific data set). Numerical solutions can never be verified "for every parameter value of all the model parameters" because the parameter space is infinite dimensional. Hence control theory complements the estimation theory and, together, they enable researchers to discover general marketing insights.

In this primer I did not discuss dynamic models with *discrete-valued* unobserved states. For example, consider the advertising driven sales evolution $S_t = \lambda S_{t-1} + \beta u_t + \nu_t$ as in Equation (2.1), but now imagine that the ad effectiveness and carryover effect (β, λ) differ during economic expansions and contractions. The two discrete states of the economy — expansion and contraction — not only are unobserved in the data, but they also exhibit dynamics: expansion in period t is likely to be followed by an expansion in period $t + 1$ with probability p, while recessionary economy stays in the contraction state with probability q. Indeed, expansion switches to contraction with probability $(1 - p)$ and so does contraction to expansion with probability $(1 - q)$. To estimate the parameters (β_1, λ_1) during expansion, (β_2, λ_2) during contraction, the likelihoods of recession $(1 - p)$ and recovery $(1 - q)$, we need a continuous-valued observation equation in the state space model Equation (2.22) for the observed sales and a discrete-valued transition Equation (2.23) for the unobserved states. The resulting model is called the Hidden Markov Model (HMM). The model parameters can be estimated using either maximum-likelihood or EM algorithm or Bayesian estimation (see Zucchini and MacDonald [2009] for details). For the application of the EM estimation for this advertising example, see Smith et al. [2006]. For an application of the Bayesian estimation in a customer relationship context, see Netzer et al. [2008]. Besides parameter estimation and inference in HMMs, three central problems in HMMs are as follows: (i) determining the number of states (for example, are the hidden states boom and bust, or boom, slump and bust?); (ii) determining the optimal sequence of state evolution (for example, was it boom $\rightarrow$ bust $\rightarrow$ slump or boom $\rightarrow$ slump $\rightarrow$ bust); and (iii) determining the optimal control of HMMs (for example, how should we optimally advertise u_t^* in each of the boom, bust, slump regimes?) Problem (i) is tackled in Smith et al. [2006], who derived the Markov Switching Criterion to not only select the number of hidden states optimally, but also retain the variables parsimoniously in a regression model in each of the selected states. Problem (ii) is classic and its solution is provided by the Viterbi algorithm. General techniques to solve Problem (iii) are presented in Elliott et al. [1995].

I close by recommending further readings beyond this primer. For Kalman filtering, see Jazwinski [2007], Anderson and Moore [2012], and Simon [2006]. For parameter estimation, see Harvey [2001] for ML estimation, Harrison and West [2013] for Bayesian estimation, and Shumway and Stoffer [2011] for EM estimation. For optimal control theory, see Bryson and Ho [1975], Sethi and Thompson [2000], Kamien and Schwartz [2012], and Weber [2011]. For differential games, see Jørgensen and Zaccour [2004] and Dockner et al. [2000]. For stochastic calculus, see Malliaris and Brock [1982] and Grigoriu [2002]. For stochastic control, Dixit and Pindyck [1994, Part II] and Astrom [2006]. Finally, for numerical solutions to control problems, see Lapidus and Pinder [1999] and Smith [2003] who present powerful tools to solve partial differential equations using finite difference methods.

References

P. Aghion and P. Howitt. *Endogenous Growth Theory.* MIT Press, Cambridge, MA, 1998.

B. D. O. Anderson and J. Moore. *Optimal Filtering.* Dover Publications, Mineola, NY, 2012.

C. Andrieu and A. Doucet. Particle filtering for partially observed gaussian state space models. *Journal of the Royal Statistical Society, Series B*, 64 (4):827–836, 2002.

I. Arasaratnam and S. Haykin. Squareroot quadrature kalman filtering. *IEEE Transactions on Signal Processing*, 56(6):2589–2593, 2008.

A. Aravindakshan and P. A. Naik. How does awareness evolve when advertising stops? the role of memory. *Marketing Letters*, 22(3):315–326, 2011.

A. Aravindakshan and P. A. Naik. Understanding the memory effects in pulsing advertising. *Operations Research*, 63(1):35–47, 2015.

A. Aravindakshan, K. Peters, and P. A. Naik. Spatiotemporal allocation of advertising budgets. *Journal of Marketing Research*, 49(1):1–14, 2012.

A. Aravindakshan, O. Rubel, and O. Rutz. Managing blood donations with marketing. *Marketing Science*, 34(2):269–280, 2015.

K. J. Astrom. *Introduction to Stochastic Control Theory.* Dover Publications, Mineola, NY, 2006.

M. B. Ataman, C. F. Mela, and H. J. van Heerde. Building brands. *Marketing Science*, 27(6):1036–1054, 2008.

M. B. Ataman, H. J. van Heerde, and C. F. Mela. The long-term effect of marketing strategy on brand sales. *Journal of Marketing Research*, 47(5): 866–882, 2010.

T. Başar and G. J. Olsder. *Dynamic Noncooperative Game Theory.* Society for Industrial and Applied Mathematics, Philadelphia, 2nd edition, 1999.

T. E. Barry and D. J. Howard. A review and critique of the hierarchy of effects in advertising. *International Journal of Advertising*, 9(2):121–135, 1990.

F. M. Bass. A new product growth model for consumer durables. *Management Science*, 15(5):215–227, 1969.

F. M. Bass, T. V. Krishnan, and D. C. Jain. Why the bass model fits without decision variables. *Marketing Science*, 13(3):203–223, 1994.

F. M. Bass, A. Krishnamoorthy, A. Prasad, and S. P. Sethi. Advertising competition with market expansion for finite horizon firms. *Journal of Industrial and Management Optimization*, 1(1):1–19, 2013.

G. Belch and M. Belch. *Advertising & Promotion: An Integrated Marketing Communications Perspective.* McGraw-Hill Companies, 6th edition, 2004.

D. P. Berstekas. *Dynamic Programming and Optimal Control,* vol. 1 and 2. Athena Scientific, Belmont, MA, 2005.

E. Biyalogorsky and P. A. Naik. Clicks and mortar: The effect of online activities and offline sales. *Marketing Letters*, 14(1):21–32, 2003.

D. Bowman and H. Gatignon. Market response and marketing mix models: Trends and research opportunities. *Now Publishers, Boston, USA*, 4(3): 129–207, 2010.

B. J. Bronnenberg. Advertising frequency decisions in a discrete markov process under a budget constraint. *Journal of Marketing Research*, 35(3): 399–406, 1998.

N. Bruce, N. Z. Foutz, and C. Kolsarici. Dynamic effectiveness of advertising and word of mouth in sequential distribution of new products. *Journal of Marketing Research*, 49(4):469–486, 2012a.

N. I. Bruce. Pooling and dynamic forgetting effects in multi-theme advertising: Tracking the advertising sales relationship with particle filters. *Marketing Science*, 27(4):659–673, 2008.

N. I. Bruce, K. Peters, and P. A. Naik. Discovering how advertising grows sales and builds brands. *Journal of Marketing Research*, 49(6):793–806, 2012b.

A. E. Bryson and Y. C. Ho. *Applied Optimal Control: Optimization, Estimation and Control.* Taylor and Francis, Boston, MA, 1975.

K. P. Burnham and D. R. Anderson. *Model Selection and Multimodel Inference: A Practical Information-theoretic Approach.* Springer Science & Business Media, New York, NY, 2002.

P. K. Chintagunta and D. C. Jain. Empirical analysis of a dynamic duopoly model of competition. *Journal of Economics & Management Strategy*, 4 (1):109–131, 1995.

P. K. Chintagunta and V. R. Rao. Pricing strategies in a dynamic duopoly: A differential game model. *Management Science*, 42(11):1501–1514, 1996.

P. K. Chintagunta and N. Vilcassim. An empirical investigation of advertising strategies in a dynamic duopoly. *Management Science*, 38(9):1230–1244, 1992.

P. K. Chintagunta, V. R. Rao, and N. J. Vilcassim. Equilibrium pricing and advertising strategies for nondurable experience products in a dynamic duopoly. *Managerial and Decision Economics*, 14(3):221–234, 1993.

M. G. Dekimpe, P. H. Franses, D. M. Hanssens, and P. A. Naik. Time-series models in marketing. In B. Wierenga, editor, *Handbook of Marketing Decision Models*, pages 373–398, Berlin, 2008. Springer Verlag.

A. K. Dixit and R. S. Pindyck. *Investement Under Uncertainty.* Princeton University Press, Princeton, NJ, 1994.

E. J. Dockner, S. Jørgensen, N. V. Long, and G. Sorger. *Differential Games in Economics and Management Science.* Cambridge University Press, Cambridge, UK, 2000.

R. Dorfman and P. O. Steiner. Optimal advertising and optimal quality. *The American Economic Review*, 44(5):826–836, 1954.

R. Du and W. Kamakura. Improving the statistical performance of tracking studies based on repeated cross-sections with primary dynamic factor analysis. *International Journal of Research in Marketing*, 32(1):94–112, 2015.

J. P. Dubé, G. J. Hitsch, and P. Manchanda. An empirical model of advertising dynamics. *Quantitative Marketing and Economics*, 3(2):107–144, 2005.

J. Durbin and S. J. Koopman. *Time Series Analysis by State Space Methods.* Oxford University Press, Oxford, UK, 2012.

R. Elliott, L. Aggoun, and J. Moore. *Hidden Markov Models: Estimation and Control.* Springer, New York, N.Y., 1995.

G. M. Erickson. *Dynamic Models of Advertising Competition.* Kluwer, Norwell, MA, 2nd edition, 2003.

M. Esteban-Bravo, J. M. Vidal-Sanz, and G. Yildirim. Valuing customer portfolios with endogenous mass and direct marketing interventions using a stochastic dynamic programming decomposition. *Marketing Science*, 33 (5):621–640, 2014.

F. M. Feinberg. Pulsing policies for aggregate advertising models. *Marketing Science*, 11(3):221–234, 1992.

F. M. Feinberg. On continuous-time optimal advertising under s-shaped response. *Management Science*, 47(12):1476–1487, 2001.

R. Feynman, R. Leighton, and M. Sands. *The Feynman Lectures on Physics,* Vol. 2. Chapter 25-2, http://www.feynmanlectures.caltech.edu/ (HTML edition), 2006.

M. Freimer and D. Horsky. Periodic advertising pulsing in a competitive market. *Marketing Science*, 31(4):637–648, 2012.

G. E. Fruchter. The many-player advertising game. *Management Science*, 45: 1609–1611, 1999.

G. E. Fruchter and S. Kalish. Closed-loop advertising strategies in a duopoly. *Management Science*, 43(1):54–63, 1997.

G. E. Fruchter and C. Van den Bulte. Why the generalized bass model leads to odd optimal advertising policies. *International Journal of Research in Marketing*, 28(3):218–230, 2011.

P. Gill, W. Murray, and M. H. Wright. *Practical Optimization.* Academic Press, London, UK, 1982.

M. Grigoriu. *Stochastic Calculus: Applications in Science and Engineering.* Birkhauser, Boston, MA, 2002.

D. M. Hanssens, L. J. Parsons, and R. L. Schultz. Market response models: Econometric and time series analysis. In *International Series in Quantitative Marketing*, Massachusetts, USA: Kluwer, 2003. Academic Publishers.

J. Harrison and M. West. *Bayesian Forecasting and Dynamic Models.* Springer, New York, NY, 2013.

R. F. Hartl. A simple proof of the monotonicity of the state trajectories in autonomous control problems. *Journal of Economic Theory*, 41:211–215, 1987.

A. C. Harvey. *Forecasting, Structural Time Series Models and the Kalman Filter.* Cambridge University Press, Cambridge, UK, 2001.

S. Hasegawa, N. Terui, and G. M. Allenby. Dynamic brand satiation. *Journal of Marketing Research*, 49(6):842–853, 2012.

Y. Hu, R. Y. Du, and S. Damangir. Decomposing the impact of advertising: Augmenting sales with online search data. *Journal of Marketing Research*, 51(3):300–319, 2014.

P. J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1(1):221–233, 1967.

S. D. Jap and P. Naik. Bid analyzer: A method for estimation and selection of dynamic bidding models. *Marketing Science*, 27(6):949–960, 2008.

A. H. Jazwinski. *Stochastic Processes and Filtering Theory.* Dover Publications, Mineola, N.Y., USA, 2007.

J. P. Jones. *What's in a Name: Advertising and the Concept of Brand.* D. C. Heath and Company, Lexington, MA, 1986.

J. P. Jones. Ad spending: Maintaining market share. *Harvard Business Review*, 68(1):38–42, 1990.

S. Jørgensen and G. Zaccour. Differential games in marketing. In *International Series in Quantitative Marketing.* Kluwer Academic Publishers, USA, 2004.

S. J. Julier and J. K. Uhlmann. Unscented filtering and nonlinear estimation. *Proceedings of the IEEE*, 92(3):401–422, 2004.

K. Kalaignanam, T. Kushwaha, and M. Eilert. The impact of product recalls on future product reliability and future accidents: Evidence from the automobile industry. *Journal of Marketing*, 77(2):41–57, 2013.

S. Kalish. Monopolist pricing with dynamic demand and production cost. *Marketing Science*, 2(2):135–159, 1983.

R. E. Kalman and R. S. Bucy. New results in linear filtering and prediction theory. *Journal of Basic Engineering*, 83(1):95–108, 1961.

M. I. Kamien and N. L. Schwartz. *Dynamic Optimization: The Calculus of Variations and Optimal Control in Economics and Management.* Dover Publications, USA, 2nd edition, 2012.

C. Kolsarici and D. Vakratsas. Category-versus brand-level advertising messages in a highly regulated environment. *Journal of Marketing Research*, 47(6):1078–1089, 2010.

G. Koop and D. Korobilis. Bayesian multivariate time-series methods for empirical macroeconomics. *Foundations and Trends in Econometrics*, 3(4): 267–358, 2009. Now Publishers, Boston, USA.

V. Kumar, R. Venkatesan, T. Bohling, and D. Beckmann. The power of CLV: Managing customer lifetime value at IBM. *Marketing Science*, 27(4): 585–599, 2008.

H. J. Kushner. On the differential equations satisfied by conditional probability densities of markov processes with applications. *Journal of SIAM Control, Series A*, 2(1):106–119, 1964.

M. Lachaab, A. Ansari, K. Jedidi, and A. Trabelsi. Modeling preference evolution in discrete choice models: A bayesian state-space approach. *Quantitative Marketing and Economics*, 4(1):57–81, 2006.

L. Lapidus and G. F. Pinder. *Numerical Solution of Partial Differential Equations in Science and Engineering*. John Wiley and Sons, New York, NY, 1999.

R. J. Lavidge and G. A. Steiner. A model for predictive measurements of advertising effectiveness. *Journal of Marketing*, 25(6):59–62, 1961.

P. S. H. Leeflang, D. R. Wittink, M. Wedel, and P. A. Naert. Building models for marketing decisions. In *International Series in Quantitative Marketing vol. 9*. Springer, USA, 2000.

Y. Liu and V. Shankar. *The Dynamic Effects of Product Harm Crises on Brand Preference and Advertising Effectiveness: An Empirical Analysis of the Automobile Industry*. forthcoming, 2015.

L. Ljungqvist and T. J. Sargent. *Recursive Macroeconomic Theory*. MIT Press, Cambridge. MA, 2nd edition, 2004.

V. Mahajan and E. Muller. Advertising pulsing policies for generating awareness for new products. *Marketing Science*, 5(2):89–111, 1986.

V. Mahajan, E. Muller, and S. Sharma. An empirical comparison of awareness forecasting models of new product introduction. *Marketing Science*, 3(3): 179–197, 1984.

A. G. Malliaris and W. A. Brock. *Stochastic Methods in Economics and Finance*. North Holland, 1982.

A. McQuarie and C.-L. Tsai. *Regression and Time Series Model Selection*. World Scientific, Singapore, 1998.

H. I. Mesak. An aggregate advertising pulsing model with wearout effects. *Marketing Science*, 11(3):310–326, 1992.

R. Montoya, O. Netzer, and K. Jedidi. Dynamic allocation of pharmaceutical detailing and sampling for long-term profitability. *Marketing Science*, 29 (5):909–924, 2010.

P. A. Naik and K. Raman. Understanding the impact of media synergy in multimedia communications. *Journal of Marketing Research*, 40(4):375–388, 2003.

P. A. Naik, M. K. Mantrala, and A. Sawyer. Planning pulsing media schedules in the presence of dynamic advertising quality. *Marketing Science*, 17(3): 214–235, 1998.

P. A. Naik, K. Raman, and R. Winer. Planning marketing-mix strategies in the presence of interactions. *Marketing Science*, 24(1):25–34, 2005.

P. A. Naik, P. Shi, and C.-L. Tsai. Extending the akaike information criterion to mixture regression models. *Journal of the American Statistical Association*, 102(477):244–254, 2007.

P. A. Naik, A. Prasad, and S. P. Sethi. Building brand awareness in dynamic oligopoly markets. *Management Science*, 54(1):129–138, 2008.

M. Nerlove and K. Arrow. Optimal advertising policy under dynamic conditions. *Economica*, 29(114):129–142, 1962.

S. A. Neslin and H. J. van Heerde. *Promotion Dynamics.* Now Publishers, Boston, USA, 2009.

O. Netzer, J. M. Lattin, and V. Srinivasan. A hidden markov model of customer relationship dynamics. *Marketing Science*, 27(2):185–204, 2008.

J. Nocedal and S. J. Wright. *Numerical Optimization.* New York, NY: Springer-Verlag, 2006.

E. C. Osinga, P. S. H. Leeflang, and J. E. Wieringa. Early marketing matters: A time-varying parameter approach to persistence modeling. *Journal of Marketing Research*, 47(1):173–185, 2010.

S. Park and M. Hahn. Pulsing in a discrete model of advertising competition. *Journal of Marketing Research*, 28(4):397–405, 1991.

K. Pauwels. How dynamic consumer response, competitor response, company support, and company inertia shape long-term marketing effectiveness. *Marketing Science*, 23(4):596–610, 2004.

A. Prasad and S. P. Sethi. Integrated marketing communications in markets with uncertainty and competition. *Automatica*, 45(3):601–610, 2009.

K. Raman. Boundary value problems in stochastic optimal control of advertising. *Automatica*, 42(8):1357–1362, 2006.

K. Raman and R. Chatterjee. Optimal monopolist pricing under demand uncertainty in dynamic markets. *Management Science*, 41(1):144–162, 1995.

K. Raman and P. A. Naik. Long-term profit impact of integrated marketing communications program. *Review of Marketing Science*, 2(1), 2004. Article 8, DOI: 10.2202/1546-5616.1014.

O. J. Rubel, P. A. Naik, and S. Srinivasan. Optimal advertising when envisioning a product-harm crisis. *Marketing Science*, 30(6):1048–1065, 2011.

O. J. Rutz and R. Bucklin. From generic to branded: A model of spillover dynamics in paid search advertising. *Journal of Marketing Research*, 48(1): 87–102, 2011.

O. J. Rutz and G. P. Sonnier. Modeling the evolution of internal market structure. *Marketing Science*, 30(2):274–289, 2011.

M. Sasieni. Optimal advertising expenditure. *Management Science*, 18:64–72, December 1971.

A. Seierstad and K. Sydsaeter. Sufficient conditions in optimal control theory. *International Economic Review*, 18(2):367–391, 1977.

S. P. Sethi. Optimal advertising for the nerlove-arrow model under a budget constraint. *Operations Research Quarterly*, 28(3):683–693, 1977.

S. P. Sethi. Deterministic and stochastic optimization of a dynamic advertising model. *Optimal Control Applications and Methods*, 4(2):179–184, 1983.

S. P. Sethi and G. L. Thompson. A tutorial on optimal control theory. *INFOR*, 19(4):279–291, 1981.

S. P. Sethi and G. L. Thompson. *Optimal Control Theory: Applications to Management Science and Economics*. Kluwer Academic Publishers, Boston, 2nd edition, 2000.

V. Shankar. Strategic allocation of marketing resources: Methods and insights. In Roger Kerin and Rob O'Regan, editors, *Marketing Mix Resource Allocation and Planning: New Perspectives and Practices*, pages 154–183. American Marketing Association, 2008.

R. H. Shumway and D. S. Stoffer. *Time Series Analysis and Its Applications: With R Examples*. Springer, New York, NY, 2011.

D. Simon. *Optimal State Estimation*. John-Wiley and Sons, Hoboken, NJ, USA, 2006.

A. Smith, P. A. Naik, and C.-L. Tsai. Markov-switching model selection using kullback–leibler divergence. *Journal of Econometrics*, 134(2):553–577, 2006.

G. D. Smith. *Numerical Solution of Partial Differential Equations: Finite Difference Methods*. Oxford University Press, Oxford, UK, 2003.

G. Sorger. Competitive dynamic advertising: A modification of the case game. *Journal of Economics Dynamics and Control*, 13:55–80, 1989.

S. Sridhar, M. K. Mantrala, P. A. Naik, and E. Thorson. Dynamic marketing budgeting for platform firms: Theory, evidence, and application. *Journal of Marketing Research*, 48(6):929–943, 2011.

S. Sriram and M. U. Kalwani. Optimal advertising and promotion budgets in dynamic markets with brand equity as a mediating variable. *Management Science*, 53(1):46–60, 2007.

S. Sriram, S. Balachander, and M. U. Kalwani. Monitoring the dynamics of brand equity using store-level data. *Journal of Marketing*, 71(2):61–78, 2007.

J.-B. E. M. Steenkamp, V. R. Nijs, D. M. Hanssens, and M. G. Dekimpe. Competitive reactions to advertising and promotion attacks. *Marketing Science*, 24(1):35–54, 2005.

N. L. Stokey and R. E. Lucas. *Recursive Methods in Economic Dynamics.* Harvard University Press, Cambridge, MA, 1989.

H. J. Sussmann and J. C. Willems. 300 years of optimal control: From the brachystochrone to the maximum principle. *IEEE Control Systems Magazine*, pages 33–44, June 1997.

T. S. Teixeira, M. Wedel, and R. Pieters. Moment-to-moment optimal branding in TV commercials: Preventing avoidance by pulsing. *Marketing Science*, 29(5):783–804, 2010.

G. J. Tellis. *Advertising and Sales.* Reading, MA: Addison-Wesley, 1998.

D. Vakratsas and T. Ambler. How advertising works: What do we really know? *Journal of Marketing*, 63(1):26–43, 1999.

D. Vakratsas and C. Kolsarici. A dual market diffusion model for a new prescription pharmaceutical. *International Journal of Research in Marketing*, 25(4):282–293, 2008.

H. van Heerde, K. Helsen, and M. G. Dekimpe. The impact of product-harm crisis on marketing effectiveness. *Marketing Science*, 26(2):230–245, 2007.

R. L. Vaughn. How advertising works: A planning model. *Journal of Advertising Research*, 20(5):27–33, 1980.

R. L. Vaughn. How advertising works: A planning model revisited. *Journal of Advertising Research*, 26(1):57–66, 1986.

M. L. Vidale and H. B. Wolfe. An operations research study of sales response to advertising. *Operations Research*, 5:370–381, 1957.

T. A. Weber. *Optimal Control Theory with Applications in Economics.* MIT Press, Cambridge, MA, 2011.

H. White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838, 1980.

M. Zakai. On the optimal filtering of diffusion processes. *Probability Theory and Related Fields*, 11(3):230–243, 1969.

W. Zucchini and I. L. MacDonald. *Hidden Markov Models for Time Series: An Introduction Using R.* Boca Raton, FL: Chapman & Hall, 2009.