Stable Fitting of Nonparametric Models to Predict Complex Human Behaviors

September 2014

Rikiya Takahashi

Stable Fitting of Nonparametric Models to Predict Complex Human Behaviors

Graduate School of Systems and Information Engineering
University of Tsukuba

September 2014

Rikiya Takahashi

Preface

Supporting decision making in the real world is one of the central aims in developing precise descriptive models to explain and predict real human behaviors. Particularly in the environment in which many humans interact to one another, a decision maker must accurately predict or forecast the others' behaviors, and she or he must make rational decisions based on such reliable predictions. While predicting how humans *actually* behave is a central interest in most social sciences, to design explicit mathematical forecasting functions in such descriptive models is an extremely challenging problem. The mathematical hardness in predicting actual human behaviors is a composite outcome of each individual's bounded rationality (e.g., cognitive bias and limited memory), and the interactions among such biased individuals. Each human adopts idiosyncratic decision-making mechanisms that are beyond the scope of the assumptions in simplistic normative models that discuss how each human *should* behave. Then further complex social phenomena could emerge as such bounded-rational decision makers interact to one another with limited information about other players. Considering such complexity of the composite outcome and the practical inefficiency in postulating the functional forms about each individual's decision making mechanisms *a priori*, the author rather adopts experimental approaches to construct statistical predictive models to describe the observed individual or collective behaviors, by introducing highly-expressive probabilistic models whose parameters are fitted with real-world datasets. While the attained statistical models lack parts of the explicit functions to predict microscopic interactions, their predictive powers for the observed behaviors become close to those with the true underlying system.

Throughout this dissertation, the author's essential philosophy to obtain useful descriptive models is to design and fit nonparametric machine learning models. Nonparametric models, which introduce non-linear predictors with interpolating multiple basis functions, possess much higher explanatory powers with lower bias than simple parametric or linear models. The functional flexibility in nonparametrics can greatly help us predict or infer each individual's behavior and the collective social phenomena. While estimation of the parameters of such expressive models involves high variance and over-fitting to the training data, recent advances in machine learning and data mining algorithms allow for suppressing such estimation variance. By effectively fitting a nonparametric model with searching for the optimal bias-variance trade-off, we attain the descriptive human behavior models that have high generalization capabilities, and whose predictions are applicable for lots of out-of-sample datasets. The direction of statistically constructing the descriptive models is promising because of the recent emergence of very large datasets recording real human activities, which further incentivize us to develop more accurate and computationally-efficient nonparametric machine learning algorithms. Beginning with the basic enhancement and its underlying philosophy in stabilizing and accelerating the nonparametric machine learning algorithms, the author develops new algorithms customized with each application domain, in order to help make decisions with predicting the behaviors of other humans.

The author starts with his basic algorithmic contribution in stabilizing and accelerating nonparametric density estimation. Density estimation, which is fitting of probability density function for certain random variables, is one of the most essential machine learning tasks containing the conventional regression and classification problems as its subsets. Unlike the fitting of the exponential family including the well-known Gaussian distributions, a nonparametric density estimator enables the fitting of any (e.g., multi-modal and heavy-tailed) probability density functions, with a unified manner of using a mixture of multiple kernel functions. Yet nonparametric models usually involve high-dimensional parameters and their fitting is easily trapped with poor local optima, because of the high degrees of freedom in determining its structure. For attaining reliable estimates, the author exploits convex optimization, which easily converges into the global optimum merely with gradient ascending or descending. Since optimizing all of the parameters in a nonparametric model is non-convex in most cases, how to relax the fitting problem with a convex-optimization formalism is the essential and practical issue that machine learning researchers must solve. The author realizes a fast and global optimization algorithm for the nonparametric density estimation, with enhancing convex clustering to fit the mixture weights while fixing the parameters in all of the kernel functions. The author introduces a new algorithm to perform the convex clustering and sparse nonparametric density estimation where the true objective is exactly optimized with much faster convergence, while the prior work adopting an Expectation-Maximization (EM) algorithm with an approximate pruning exhibits very slow convergence. The key idea in this study is a new style of the Sequential Minimal Optimization (SMO) approach, inhering the measurement of similar kernel functions and an element-wise Newton-Raphson method. The required number of iterations in the new estimator is drastically reduced because of its accurate and fast pruning of many irrelevant kernels. The much acceleration achieved with the new algorithm makes the sparse nonparametric density estimation applicable to large datasets that the prior work could not handle, and helps application researchers concentrate on designing effective kernel functions depending on the domains of the applications.

Then the author proceeds into the real application parts about the prediction and forecasting of the behaviors of interacting humans. One application the author focuses on is the development of a nonparametric travel-time prediction model, which properly captures the characteristics of interacting vehicle traffic. In automobile transportation, travel time in one route is a composite outcome of the interactions among heterogeneous drivers having their own routing decision mechanisms. Here the interactions are mediated by the digraph structure of the road network, whose vertexes are intersections and whose edges are links between two intersections. Each driver's routing decision is unified as a risk-sensitive travel-time minimization, where each driver evaluates not the risk-neutral but the risk-sensitive utilities for many possible routes with applying each driver's own risk measure. Then one essential information for simulating the complex traffic on the entire road network, or navigating a car along the most desirable route based on the normative decision theory, is the probability density function of travel time for every link. Given the fact that the actual travel-time distribution has a right-skewed heavier tail and depends on the location of each link, the author introduces a new nonparametric conditional density estimation algorithm that precisely fits link-dependent non-Gaussian distributions of travel time using real and large probe-car datasets. The new estimator fully exploits the accelerated convex clustering algorithm that is executed twice inside the entire estimation, given the required properties of kernel functions that must be customized with this application. Here each kernel function, which is interpolated in the final convex clustering, is a mixture of gamma or log-normal distributions whose fitting requires another convex clustering execution. Based on the fact that one congestion diffuses from one link into its connecting links, the optimal mixing weights in the final conditional estimates are given with a diffusion model involving the cascading of the interaction, using

a sparse link similarity matrix constructed from an approximate diffusion kernel on a link connectivity graph. The high predictive accuracies about the travel-time distributions evidence the advantages of the new nonparametric estimator, which adopts the fast and global optimization algorithms.

The last two chapters of this dissertation are spent for another important application designed for economic decision making, which requires both of the descriptive modeling of consumer behaviors and the normative optimization for firms to target such real consumers. For the purpose of marketing implications, the proposed descriptive model explicitly deals with the microscopic behaviors by each individual consumer, based on the author's same philosophy of interpolating multiple kernel functions but with a customized modeling and optimization algorithm adjusted in this domain. Here each consumer's response or purchase propensity is modeled as a composition of her or his own brand attitudes with limited memory, and interaction factors stemming from word-of-mouth, herding, and common biorhythms among humans. The idea of interpolating fixed-parameter kernels is extended for the design of the limited memory, which yields power-law forgetting curves depending on the types of the stimulus events. Considering the computational efficiency in processing variable-interval continuous-time event datasets, the author proposes an effective convex optimization algorithm to precisely fit the event-dependent forgetting curves, which form staircase functions of elapsed time for approximating the power-law decays. The author further introduces a three-step estimator for mining the hidden interactions among consumers, whose records are not observable in the datasets. The fitted forgetting curves provide clear and psychological interpretations about how each consumer perceives each type of the event, and practical implications for firms who want to optimize their marketing investments.

The normative decision support for a firm to target consumers is realized with a new marketing-mix optimization algorithm, which gives the optimal marketing budget allocation across various channels and target segments, along with the timing of each marketing action for maximum profit. This last study focuses on the practical but normative decision making when the fitted descriptive consumer models are assumed to be accurate. The goal in the normative marketing-mix optimization is the maximization of the expected revenue minus the total marketing cost in certain multiple periods. This profit maximization is also constrained with the temporal dynamics about the states of each consumer, where such states are essentially unobservable but can be inferred from the descriptive model fitted. The author introduces a new high-dimensional constrained linear programming formulation, which accounts for both of the temporal dynamics constraints and the complex budget constraints across many periods, segments, and channels required in real marketing operations. Each segment of the consumers is a discrete approximation of an underlying continuous state. Since the attained policy with the discrete-segment approximation is not guaranteed to perform optimally in the true continuous-state system, a continuous-state discrete-event simulation study about the forecasted impacts of the attained policy is also investigated for further implications. This novel combination of machine learning, normative optimization and simulation embodies one of the effective solutions for the ultimate goal targeted in this dissertation, which is the decision making in interacting with other humans whose behaviors are forecasted with statistical descriptive models.

Contents

Acknowledgements

As in most of the scientific work in the world, many of the ideas, implementations, and the perspectives to broadly overview the technical contributions in this dissertation are the outcomes of cooperative interactions between me and my advisors or colleagues. First I would like to thank my supervisor Prof. Setsuya Kurahashi, who has helped me learn many parts of his broad insights about complex systems, and discuss many ways to efficiently handle the interactions and positive feedback with agent-based modeling, machine learning, and data mining techniques. I also would like to thank the committee members for evaluating my dissertation: Prof. Sadaaki Miyamoto, Prof. Kenichi Yoshida, Prof. Kazuaki Tsuda, and Prof. Kiyoshi Izumi. I think that significant improvements in organizing the dissertation are attained after their reviews that focused on both the algorithmic perspectives and the crucial assumptions about decision making.

Many of the technical contributions in this dissertation are fortune outcomes from the challenging projects, during my professional work in IBM. I need to express my sincere gratitudes for Ruby Kennedy and Vincent (Vince) Jeffs, who are the co-authors of my prior publications and who have supported my marketing decision-making research both financially and technically. As well as her characteristic experience as one of the three entrepreneurs who founded the Unica Corporation, Ruby has come up with a lot of advanced and critical questions to improve the quality of our research contributions. Based on his deep expertise in marketing decision making derived from really lots of his customer facing experiences, Vince has provided us how the viewpoints form the academia and the industry become different, and has guided us to take both the efficient and practical way in the improvement. In addition to the names of Ruby and Vince, I appreciate all of the members who spent for our two-year collaboration between IBM Research and IBM Enterprise Marketing Management team that involves many members from the former Unica Corporation. I primarily need to mention the name of David Konopnicki, who created this fortune opportunity of the interactions between the research staff members and lots of managers in the business side. While the sequence of members becomes a bit lengthy, here I would like to at least list the names of Ravi Shah, Robert Crites, Glen Osterhout, Robert Parkin, Michael Leavitt, Abhijeet Warkar, and Aarti Anawalikar in the IBM Enterprise Marketing Management team, Takayuki Yoshizumi and Hideyuki Mizuta at IBM Research - Tokyo, Naoki Abe and Ronny Luss at IBM T.J. Watson Research Center, and Ateret Anaby-Tavor, Ofer Shir, and Oded Margalit at IBM Research - Haifa.

I must declare that I obtained many of my scientific research and writing skills by deeply relying on the fruitful advices from many managers and colleagues in IBM Research. Tsuyoshi Idé, who is my former manager and now is working at IBM T.J. Watson Research Center, has strongly encouraged to do characteristic research and obtain a Ph.D with a bold research agenda. Both in business and personal, I has been helped a lot by Naoki Abe also at IBM T.J. Watson Research Center, who is one of the respected alumni of Senior High School at Komaba, University of Tsukuba, as the high school I graduated from. I still remember the first fortune moment of beginning the marketing optimization

study with Abderrahim (Abdel) Labbi and Cesar Berrospi at IBM Research - Zurich. As well as the original idea of applying Markov decision processes for marketing decision making, Abdel and Cesar really have innovative ideas for marketing applications by using both the academic and industrial knowledge. In addition to these advisors, I must really thank Takayuki Osogami for greatly improving my scientific writing skills. Most parts of my writing skills can have never been acquired unless his adequate supervision when I was a unskilled new-comer to the research community. In setting a challenging research agenda, I learned a lot about the required risk taking, from Fusashi Nakamura who moved from IBM into Wolfram Research Asia.

Both before and after the entrance into the doctoral research program, I learned a lot from external academic communities as well as IBM Research. I wish to thank all of the members belonging to T-PRIMAL as a Tokyo-based interest group focusing on innovative research about machine learning algorithms. I regard the date Prof. Masashi Sugiyama, who is one of the co-founders of T-PRIMAL, visited IBM for his lecture about the basics of machine learning, as the true change point for my academic career. I must never forget to mention the name of Prof. Hisashi Kashima, who is one of my former colleagues in IBM Research and now is working at the Kyoto University. Prof. Kashima has carefully told us how each of the machine learning algorithms has matured or is still immature, and how to find out a new research topic we must address. While I switched from natural language modelng and speech recognition into basic machine learning algorithms, the basic interests for the statistical approaches were established through the discussions with Prof. Nobuaki Minematsu and Prof. Keikichi Hirose, who were the supervisors of my bachelor and master theses. In business side, my expertises in marketing have been greatly uplifted by Ryo Domoto, who manages a research and development team in Hakuhodo and is an exceptionally talented manager I have ever seen.

While I have enjoyed many research achievements in these three years, I never forget the tough days before these achievements, and the name of the friends who have supported the survival as a researcher. Lots of discussions with Tsutomu Aoki, who is one of my best friends since the junior high school, were great mental supports when I repeatedly received the reject notifications in publication trials. Exchange of personal letters with Tetsuo Hayashi, who graduated from the same high school as mine and now manages a small firm in Poland, is another mental support to keep my challenges. Personal talks with Wataru Machida, who was a classmate in the University of Tokyo and with whom I experienced an interesting and coincidental reunion, helped me set an alternative scenario to obtain a Ph.D other than focusing only on the theoretical side of machine learning algorithms. In addition, I believe that the practices of composing music, during the tough days of repeating rejections, have changed my way of thinking and implicitly helped the creation of new research ideas. Many interesting disciplines provided by Tomoumi Omata, who is my teacher of harmony, counterpoint, and orchestration, let me consider the relationship between artificial intelligence and music, and provided a motivation to keep on the studies of machine learning. Still after my change-point publication, I have received an exceptional personal support by a special friend Yasuko Kaneko in these intensive days. The records of daily studies are also the those of her support and patience.

Finally, I would like to thank my family for all their love and encouragement. I know that my tough days in fact affected the personal lives of my parents. Despite the pressures they also feel, parents have always both physically and mentally supported my academic challenge when I return to their home. My brother Ryohei has frequently proposed to set opportunities with which the family gathers in comfortable places. Thank you.

Rikiya Takahashi
University of Tsukuba
July 2014

Chapter 1

Introduction

To accurately predict the behaviors of humans who interact with one another is an essential task in many social science and decision making problems. While standard economists have regarded humans as rational optimizers having consistent objectives, experimental psychologists have clarified the bounded rationality of the real humans and hence we do not have to always regard humans as optimizers. Instead, by adopting statistical *descriptive* models to predict the behaviors of interacting and bounded-rational humans, effective decision making algorithms are derived as we see in the successes of the existing marketing applications. Such statistical models need to possess fat tails and long-range dependence, which are the main outcomes of positive feedback among the interacting humans. By interpolating *multi-scale* basis functions, we introduce nonparametric descriptive models that yield the fat tails and long-range dependence while also allowing for handling more complex functions than the existing parametric functions. By manually fixing the basis functions to cover all of the required scales determined in each application, we stably fit the interpolation weights with global convex optimization algorithms. By starting from the basic enhancement in fitting probability distributions, we proceed into the applications of the proposed methodology for transportation and marketing problems. For the marketing application, we also discuss a rational decision-making algorithm that exploits the predictions provided by the nonparametric descriptive models.

The remainder of this chapter is organized as follows. Section 1.1 introduces the literature of decision making problems involving the interactions among humans. Section 1.2 addresses the examples of the bounded-rational decisions by real humans, and justifies the adoption of statistical descriptive models. Section 1.3 shows how positive feedback among the interacting humans occurs, and addresses the power-law statistical properties of the resulting indicators for such interacting systems. While power-law models have high predictive accuracies, direct statistical modeling of such power-laws becomes impractical for large datasets, as we explain in Section 1.4. As a computationally-efficient alternative to the power-law models, Section 1.5 introduces our new methodology about the statistical descriptive models that adopt the nonparametric mixtures of multi-scale basis functions. Section 1.6 briefly addresses our rational decision making approach to exploit the nonparametric descriptive models, and Section 1.7 summarizes this chapter with introducing the organizing structure of the remainder in this dissertation.

1.1 Decision Making in Interacting Systems

Humans make their own decisions based on interactions with one another, and to accurately predict the behaviors of such interacting humans is an essential problem to realize effective economic deci-

sion making. One important aspect to classify decision making applications involving the interactions among humans is whether or not the main decision maker is a part of the entire interacting system. In transport economics, public policies such as road pricing (Small and Yan, 2001; Verhoef and Small, 2004) should be designed with careful assessments about how such policies impact to the entire traffic, because each vehicle driver would change his own decision of selecting a route by responding to the change of such policies. Given his own origin and destination points, each vehicle driver chooses a proper route whose travel time has a desirable value of some statistics (e.g., expectation (hsin Wu et al., 2004; Idé and Kato, 2009), Entropic Risk Measure (Föllmer and Schied, 2004), and Iterated Conditional Tail Expectation (Hardy and Wirch, 2004)) yielded by a stochastic game against the other drivers (Camponogara and Jr., 2003) to search for good roads that are not occupied by other cars. Design of an effective transportation policy or system is hence a complex decision making problem to control of interactions among lots of drivers, while the policy makers themselves are not the members of the traffic system. In marketing, firms intend effective campaigns involving mass advertising (Kumar et al., 2011; Naik and Raman, 2003; Naik et al., 2005; Raman et al., 2012) and personalized promotions (Elsner et al., 2003; Abe et al., 2004; Tirenni et al., 2007b; Abe et al., 2009) in order to stimulate the purchase decisions by consumers. Here word-of-mouth (Fudenberg, 1995), which causes booming social trends, is a representative example of the interactions among the consumers. When firms plan the optimal campaigns to yield high profits, each firm as the main decision maker is the member of the entire market system in principle.

As we see in the examples of marketing, many of the social science problems are classified with what types of interactions are incorporated or ignored in their predictions. One possible interpretation of the marketing decision making is a stochastic game between the firm and each consumer (e.g., (Ching, 2010)), where the firm tries to find out campaigns yielding high profits by predicting how each consumer perceives and responds to each campaign, and each consumer intends to maximize his own satisfaction instead of naïvely responding to the high-profit campaigns. While the predictions by consumers against the firms actually affect the observed economic activities, in practice these effects have been ignored in most of the successful marketing decision-making algorithms such as optimal direct mailing (Elsner et al., 2003; Abe et al., 2004), design of loyalty program (Tirenni et al., 2007b), and credit assignment (Gómez-Pérez et al., 2009). These algorithms rather have focused on statistical machine learning models to accurately predict the response probabilities or the response amounts without regarding consumers as optimizers of certain objectives. By assuming every consumer to obey this response model, effective marketing strategies to maximize the long-term profits are derived based on Markov Decision Processes (MDPs) (Altman, 1999). Such MDP-based approaches have been theoretically extended to incorporate the interactions among competing firms for shares (e.g., (Naik and Raman, 2003; Naik et al., 2005; Raman et al., 2012)). Here every firm tries to control the total marketing costs and predict aggregate-level sales, with considering a stochastic game against the competitors.

1.2 Bounded Rationality

Another important aspect to classify decision making applications is whether to assume the rationality for each human. Let us consider the main reason behind the successes of existing marketing applications that have simply focused on statistical modeling of the relationship, between input stimuli and output responses for each consumer. In many of the decision-making contexts including game-theoretic situations, each player's optimal choice from several alternatives tends to be very limited such as the ones contained in Nash equilibria (Nash, 1951). Hence strictly assuming the rationality

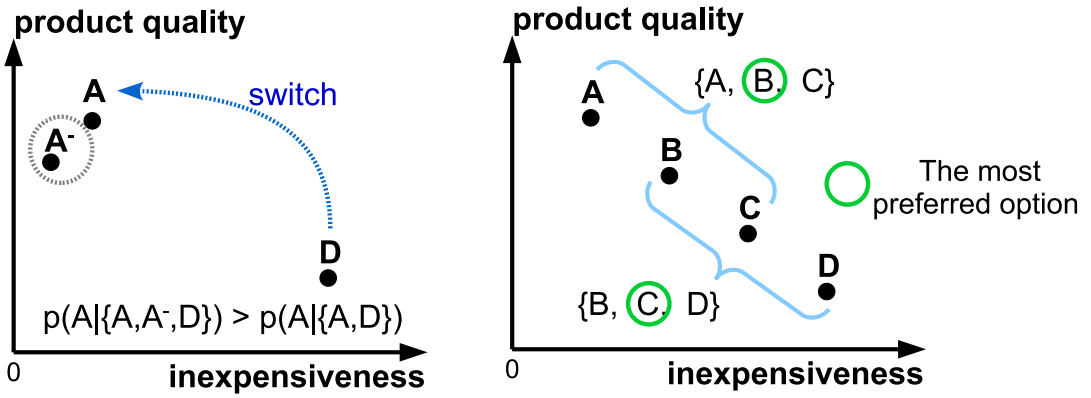
of consumers results in low degrees of freedom in predictive modeling, and loses prediction capabilities when the actual consumers do not follow rational optimization procedures. Unlike the predictions over-confidently relying on the rationality assumption, statistical input-output modeling carefully deals with the actual observations about human behaviors, and hence whether each consumer is rational or irrational is determined based on the experimental evidences. If the statistical predictive formulas cover very broad ranges of forecasting functions, such flexible models precisely fitted with large-scale real datasets provide realistic and accurate predictions adjusted for each consumer's characteristics, such as his degree of rationality. To reinforce this implication from the successes of the statistical approaches, this section provides several experimental evidences about the irrationality of the decisions by real humans, based on the literature of experimental psychology.

There have been lots of debates about whether humans are rational or irrational. By starting from the Von Neumann-Morgenstern utility theorem (von Neumann and Morgenstern, 1953) and Independence from Irrelevant Alternatives (IIA) (Luce, 1959, 1977) as prerequisites for rational decision making, standard economists and management scientists have developed normative decision-making models, with which every decision maker is assumed to be rational by having a consistent objective to be optimized and by protecting himself from intentional exploitation such as money pump (e.g., (Gustafsson, 2010)). Studies about normative models tell us how humans should behave, rather than how they actually behave.

A representative framework to predict the behaviors of such rational humans is the random utility maximization, with which each human has a consistent function to evaluate the attractiveness of an option (e.g., product and service), and the scalar output of such function is called the utility of each option. Econometric forecasts by regarding every consumer as a rational utility maximizer have been adopted in lots of econometric applications, such as product design (Brownstonea et al., 2000), demand forecasting (Train and Winston, 2007; B.D. Frischknecht and Papalambros, 2010), and various marketing problems as summarized in (Chandukala et al., 2007). With the random utility maximization framework, each human is assumed to choose an option having the maximum utility. The utility of an option is a function of only its own attributes, such as costs and benefits, plus random noise (e.g., Gaussian-distributed probit (Louviere, 1988) or Gumbel-distributed logit (McFadden, 1980)). Because the utility of each option is independent from those of the other options, intentional manipulations of the available options do not affect the decision criteria and hence the resulting decision by the utility maximization becomes rational.

In contrast, behavioral economists and experimental psychologists have studied how humans actually behave, and have introduced *descriptive* models that predict the actual behaviors of humans with high accuracy. Many of such descriptive models have been introduced for explaining irrational decisions by humans, with modeling the bounded rationality (Simon, 1947) of humans having cognitive bias (e.g., (Townsend and Ashby, 1983; Busemeyer and Townsend, 1993; Roe et al., 2001; Otter et al., 2008; Scheibehenne et al., 2009)) and limited memories involving time decays. Prominent examples of the cognitive bias are the context effects including the similarity effect (Tversky, 1972), the attraction effect (Huber et al., 1982), and the compromise effect (Simonson, 1989; Kivetz et al., 2004). These context effects cause irrational decisions in choosing an option from the set of available options, which is called the choice set. As shown in Figure 1.1, the attraction and compromise effects cause reversal of preference orders, which makes humans vulnerable to intentional manipulations of the choice sets.

Another prominent example of the irrationality is time-inconsistent decision making caused by hyperbolic discounting (Ainslie, 1974; Thaler, 1981). Before understanding the property of hyperbolic discounting, let us first understand the fact that rational dynamic decision making is provided

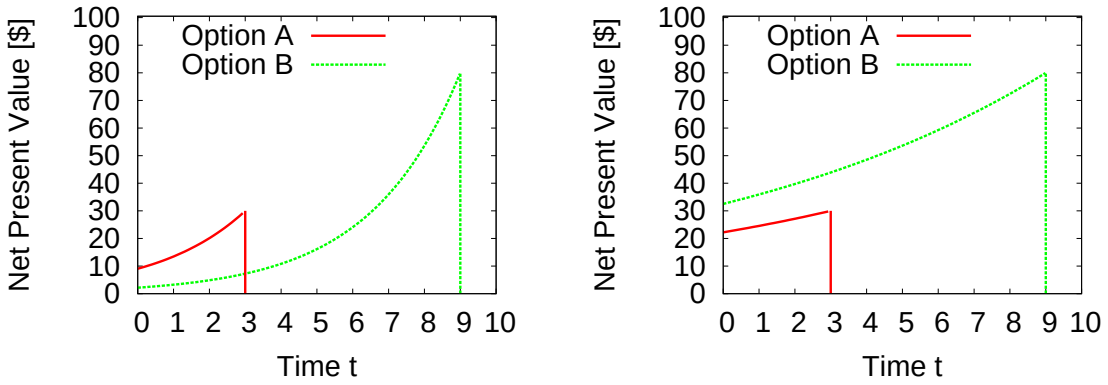


Attraction effect Compromise effect

Figure 1.1: Violation of IIA with the attraction (Huber et al., 1982) and compromise (Simonson, 1989) effects. Options (A,B,C,D) are on a two-dimensional Pareto efficient frontier. For a market in which A and D exist, introduction of option A^-, which is absolutely inferior to A, shifts portions of shares from D to A while A^- is never chosen. In each choice set containing three options, the moderate option obtains the highest share. Preference order for A and D, and that for B and C are reversed with these context effects. Any non-linear utility function cannot explain this preference reversal, as long as the attractiveness of each option is absolutely modeled with such option's own attributes.

by exponential discounting of future payoffs. As illustrated in Figure 1.2, exponential discounting yields time-consistent preferences in comparing multiple options, whose occurrence timings are different. Because of the time-consistency, any exponential discounter does have a gap between what he chooses and what he plans in advance to such choice decision. In contrast, the discounting functions actual humans have take power-law or hyperbolic discounting forms as illustrated in Figure 1.3. With hyperbolic discounting humans discount the value of a future reward based on a power-law function of the remaining time. Unlike the exponential discounting, hyperbolic discounting causes reversal of preferences between small and larger rewards obtained in the near and distant futures, respectively. This preference reversal makes humans' decision making irrational, by producing a gap between what he chooses and what he plans in advance to the choice decision. As well as in inferring the net present value of the future reward, the power-law discounting also appears in forgetting the influences of the past stimuli (Ebbinghaus, 1885; Wixted, 1990; Rubin and Wenzel, 1996; Wixted and Ebbesen, 1997) as shown in Figure 1.4. Here the ratio between the strengths of two stimuli, which have different time-stamps, becomes time-dependent. Such time-dependent relative influence yields time-inconsistent prioritization of each past stimulus, which leads irrational decision making. The power-law forgetting is regarded as a result of interacting information exchange among the short-, mid-, and long-term memories inside the same brain of a human, where only the important events survived in the short-term memory are transferred into the mid-term memory (Benjamin et al., 2008).

Based on the observed violations of the rationality assumptions and the actual successes of the marketing applications in the literature, instead of regarding humans as utility maximizers, we should adopt more flexible descriptive models that adopt statistical approximations on predictive functions whose inputs are the stimuli and information given to each human and whose outputs are the responses and behaviors by the human.



When the discounting rate λ is large

When the discounting rate λ is small

Figure 1.2: Time-consistent choice by exponential discounters. We have two decision makers (left and right), and each of them is asked of choosing one option from two mutually-exclusive monetary options A and B at some time $t \in [0, 3)$. Each figure shows every option's net present value as a function of time. Option A is immediate and small \$30 payoff at time $t = 3$, and option B is more delaying but larger \$80 payoff at time $t = 9$. For discounting the future payoffs, each of the decision makers apply an exponential discounting function $\exp(-\lambda|t - t_R|)$ where t_R is the time-stamp of obtaining each of the payoffs and λ is each decision maker's own discounting rate. The left decision maker, whose value of λ is high, prefers option A to option B at any time $t < 3$. Hence the left decision maker is myopic while his myopia is time-consistent. In contrast, the right decision maker having a lower value of λ always prefers option B, and hence he is regarded as a time-consistent long-term decision maker. As illustrated in these two figures, decision makers using the exponential discounting schemes always have time-consistent preferences, and there is no gap between what they do and what they plan in advance to the occurrence of each payoff.

1.3 Positive Feedback

In addition to the irrationality of human behaviors, positive feedback (e.g., (Zuckerman and Jefferson, 1996)) among such humans makes the predictions in interactive environments further challenging. For example in finance, financial economists and statistical physicists have studied why stock markets boom and bust, by modeling the imitation-based herding among investors (Bala and Goyal, 1998). Here each investor is supposed to balance her/his own knowledge about the market or securities, and the opinions from other investors (Blanchard and Watson, 1982; Roehner and Sornette, 2000). As derived in the rational bubble model (Blanchard and Watson, 1983), even when each individual investor having limited knowledge rationally imitates the investments by others, the irrational bubble occurs as a collective behavior of the entire market system. The rational bubble model has been further extended into the log-periodic model (Sornette et al., 1996; Feigenbaum and Freund, 1996), with assuming a scale-free network among such investors. The log-periodic model is known to possess high predictive powers particularly about the timing of the crash of the market price (Chang and Feigenbaum, 2006). The philosophy to regard the trend following as one of the main causes of the critical phenomena, such as the crash of the market price and huge congestion in traffic systems (Fosgerau and Fukuda, 2012), has been embodied as individual-level epidemic branching (Goffman and Newill, 1964) and information cascading (Bikhchandani et al., 2008) models. In marketing, such trend-following has been regarded as the cause of long-run sales (Deschatres and Sornette, 2005; Crane and Sornette, 2008) that

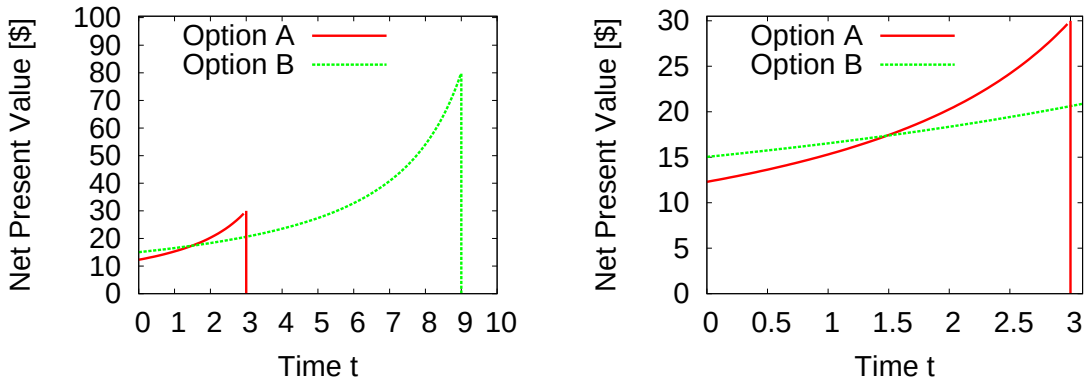


Figure 1.3: Hyperbolic discounting as a time-inconsistent decision making mechanism. The setting of the choice is the same as that in Figure 1.2, while this decision maker uses a hyperbolic power-law discounting function $1/(1+\lambda|t-t_R|)$. The left figure shows the net present values in the same period as that of Figure 1.2, while the right figure provides a magnification when $t \leq 3$. If this decision maker is asked of choosing option A or B at time $t=0$, then he will select the more delaying but large-payoff option B. If the time of the question is $t=2.5$, however, then he will select the immediate and small option A by reversing his own preference. This preference reversal means a gap between what he does choose and what he plans in advance to the choice decision, where he adopts long-term thinking when payoffs do not occur in the near future, whilst changes to be myopic when the moment of reward reaches. In general, hyperbolic discounting yields time-inconsistent preferences and makes decision makers irrational.

decay over time only gradually, because of the repetitious word-of-mouth among consumers.

If we statistically analyze interacting systems inhering positive feedback, distributions and time-series about the main indicators such as stock-price returns often exhibit heavier tails (e.g., (Mandelbrot, 1963) and (Cont, 2001) for more stylized facts) than Gaussian distributions, and long-range dependence (Cont, 2005; Deschatres and Sornette, 2005; Crane and Sornette, 2008; Filimonov and Sornette, 2011). Hence many of the statistical models to predict the outcomes from interacting systems have parametrically incorporated fat-tail distributions (e.g., Student's t-distribution (Blattberg and Gonnedes, 1974) and generalized hyperbolic distribution (Eberlein et al., 1998)) and power-law decays of the autocorrelation (Sornette and Ouillon, 2005; Filimonov and Sornette, 2011). For example in predicting the time-series of volatility, the long-range autocorrelation represents the dynamic interaction among investors, where the strength about the imitation among the investors persists while varies over time. Such explicit power-law decay modeling has also been adopted in forecasting the behaviors of consumers (e.g., Self-excited Hawkes conditional Poisson process (Hawkes, 1971) to predict the sales of books and page-views of online videos (Deschatres and Sornette, 2005; Crane and Sornette, 2008)).

1.4 Limitations of Parametric Power-Law Models

The direct parametrization of the fat tails or the long-range dependence via power-law functions, however, often results in insufficient predictive capabilities, poor local optima of the fitted parameters, and computationally inefficient algorithms that cannot be applied for large datasets obtained in real applications. The main philosophy behind the existing statistical descriptive models, which incorporate

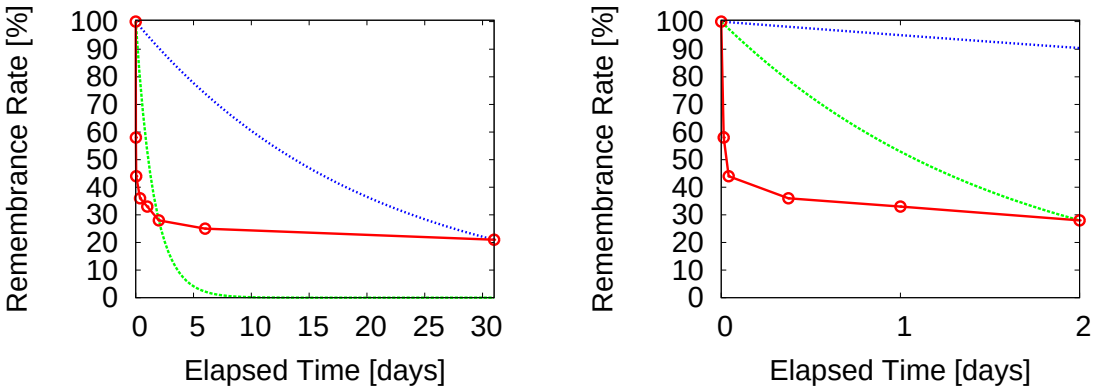


Figure 1.4: Power-law discounting also appeared in forgetting the past events. The left figure shows an observed forgetting curve in Ebbinghaus's experiments (Ebbinghaus, 1885), that measured the time-decaying ratio of subjects who remember nonsense syllables, such as "ZOF", after their initial exposition. The right figure provides a magnification of the left figure in the first 2 days. The red point and line show the observed rate of forgetting and its corresponding curve, respectively. The green and blue lines show exponential forgetting curves, based on the observation of the 2-day rate and that of the 31-day rate, respectively. The low accuracy in approximating the red line by either the green or blue line evidences that humans' forgetting curves cannot be represented by an exponential discounting function having the single discounting rate. Rather, the long tails or long-range dependence in the actual forgetting curves should be modeled by a power-law function, or mixture of exponential discounting functions having multiple heterogeneous scales.

fat tails or long-range dependence, is the fitting of certain power-law functions involving a limited number of tuning parameters. In order to clarify the essences of the underlying computational problems, let us classify the fundamental limitations caused by directly formalizing or fitting of power-law functions into three types: high bias, local optimality, and non-scalability. Here we provide a detailed background for each of the three types of the limitations.

High bias. In typical decision-making tasks, we need to predict lots of heterogeneous indicators, such as prices of many stocks in the market and individual-level responses for each of the many consumers. Depending on the degree of bounded rationality and positive feedback, some indicators obey certain power-law rules, while the others do not. When we always adopt simple power-law formulas in predicting all of the indicators, the resulting predictive accuracies become low, particularly for some indicators that do not obey power-law rules and those involving other types of complexities, such as non-monotonicity or multi-modality of the underlying functions. In other words, parametric statistical models strictly imposing the power-laws have high bias that limits the generalization capability in predictions. For avoiding such high-bias problems, we rather should fit a broader class of statistical predictive models, which contain the power-law functions as its subset. Such flexible models would adaptively yield both power-law and other complex functions, depending on the actual nature of each indicator.

Local optimality. Statistical fitting of probability distributions, which is usually based on the maximum-likelihood principle or Bayes' theorem, involves optimization of the model parameters. The class of probability distributions, whose parameter-fitting is guaranteed to be globally optimal, is very limited. One representative class of such convenient distributions is the exponential family,

whose fitting of the parameters becomes convex optimization (Liese and Miescke, 2008). Unfortunately, power-law distributions such as Student's t-distributions are not contained in the exponential family, and their parameters must be fitted with non-convex optimization algorithms, such as the Expectation-Maximization (EM) algorithms (Dempster et al., 1977; McLachlan and Krishnan, 1997; Karlis, 2002). The fitted parameters with the EM algorithms could be trapped in poor local optima, whose low qualities become crucially problematic particularly when the number of the indicators to be predicted is large.

Non-scalability. This is a characteristic limitation of the autoregressive models for incorporating the long-range dependence. To clarify the problem, let us consider an example of computing a time-varying variable $x_t \in \mathbb{R}$ at discrete time t, which becomes an explanatory variable of some autoregressive model to provide a prediction at every time t. The explanatory variable x_t is determined by a series of the past stimuli $s_{t-1} \in \mathbb{R}, \ldots, s_{t-L} \in \mathbb{R}$.

If the prediction is supplied by an exponential discounting whose per-time rate is $\lambda > 0$, then $x_t = \sum_{\ell=1}^{L} \exp(-\lambda\ell) s_{t-\ell}$ and we get a recurrence formula $x_{t+1} = \exp(-\lambda)(x_t + s_t)$. This recurrence formula makes the autoregressive models computationally efficient, and excludes the necessity of remembering the past stimuli, even when many values in the series $s_{t-1}, \ldots, s_{t-L}$ are non-zero. This memoryless computational efficiency known as the Markov property (Markov, 1954) is a great advantage of adopting exponential discounting in time-series prediction.

The hyperbolic discounting, however, does not yield a useful recurrence formula. One example of the prediction with a hyperbolic discounting is formulated as $x_t = \sum_{\ell=1}^{L} s_{t-\ell}/(1+\lambda'(t-\ell))^{\alpha}$, where $\lambda' > 0$ is another discounting rate and $\alpha > 0$ is a power exponent. Due to the lack of memoryless recurrence formulas, the computation of the variable x_t requires to remember all of the past stimuli. Even when we only store non-zero values of the past stimuli, the computational complexity at each time t linearly grows with respect to the number of non-zero stimuli before time t. This non-Markov property makes hyperbolic discounting inapplicable for large datasets, despite its high predictability for the actual human behaviors.

All of these three limitations imply the necessity of an alternative methodology to design statistical descriptive models, whose modeling capabilities are highly flexible, whose parameters are fitted with global optimization, and whose explanatory variables are efficiently computed without the growth of the computational costs, while such models still must be able to incorporate the fat tails and the long-range dependence. One practical hint to satisfy all of the desirable properties is observable in the well-known exponential-family distributions and exponential discounting functions, which have limited predictive capabilities, but which enjoy the convexity of the optimization and the memoryless Markov property, respectively. As shown in the next Section 1.5, we regard *emulating* the power-laws, via an approximate combination of the useful exponential functions, as the key to realize the desired statistical descriptive models.

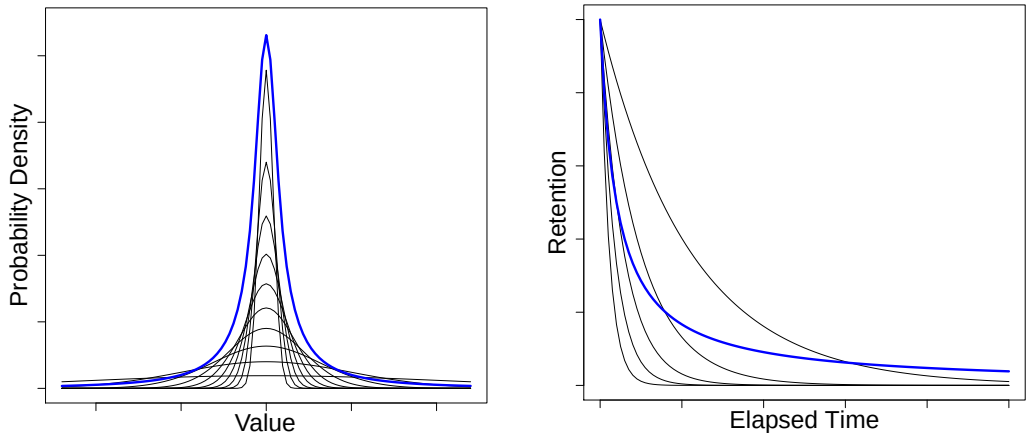
1.5 Nonparametric Descriptive Modeling as the Main Methodology

As a new methodology to predict the behaviors of interacting humans with much higher accuracy than the existing parametric models, we introduce nonparametric descriptive modeling that interpolates *multi-scale* basis functions and whose interpolation weights are optimized with global convex optimization algorithms. Given the evidences of bounded rationality, we follow the literature of statistical functional approximations to directly predict the behaviors of humans, rather than strictly regarding

humans as optimizers. Then we consider a class of nonparametric *mixture* models to formalize probability distributions or discounting functions, where the parametric fat-tail distributions or hyperbolic discounting functions are contained as subsets of the proposed models. Statistical nonparametric models based on a mixture of local basis functions or their corresponding covariance functions, such as Radial Basis Functions (Buhmann, 2003) and Matérn covariance functions (Minasny and McBratney, 2005), possess universal approximation capabilities (Hammer and Gersmann, 2003), which is the statistical consistency in fitting any non-linear functions (e.g., Support Vector Machines (Vapnik, 1995; Platt, 1999), Relevance Vector Machines (Tipping, 2001; Bishop and Tipping, 2000), and Gaussian Process Regression (Simon, 1979; Stein, 1999; Rasmussen and Williams, 2006)). By mixing basis functions whose characteristic scales (e.g., variance and expected decay time) are heterogeneous, we further enforce the nonparametric models to possess fat tails and long-range dependence, in addition to their universal approximation capabilities. The realization of all of the fat tails, long-range dependence, and universal approximation capabilities makes our philosophy of the nonparametric descriptive models applicable for many real-world prediction tasks, which involve the positive feedback among the interacting and bounded-rational humans.

The idea of using multi-scale basis functions is derived from the fact that power-law distributions and hyperbolic discounting are represented as infinite mixtures of exponential families and exponential discounting, respectively. Figure 1.5 illustrates how linear interpolations of exponential families or exponential discounting functions yield fat tails or hyperbolic discounting, when we use the multi-scale basis functions. For one example, Student's t-distribution is an infinite scale-mixture of Gaussian distributions whose variances obey an inverse-Gamma distribution (McLachlan and Krishnan, 1997). For another example, an infinite scale-mixture of exponential discounting functions, whose expected decay times obey an inverse-Gamma distribution, becomes hyperbolic discounting having the power-law decay (Axtell and McRae, 2007). The main factors that impose the resulting functions to obey limited parametric forms are the distributions to generate the multiple scales, i.e., the inverse-Gamma distributions in the examples above. Hence a natural direction to relax the strict parametrics is to replace the parametric distributions, such as the inverse-Gamma distributions, by more flexible nonparametric ones including the empirical distributions (Shorack and Wellner, 1986) and kernel density estimators (Silverman, 1986). For computational efficiency, we approximate the integral in the infinite mixture with a summation using a finite mixture, where we prepare a large number of basis functions for reducing the approximation errors.

The idea of the finite nonparametric mixtures further yields the stability in fitting the model parameters. To clarify the advantage of our philosophy, let us first take the density estimation problem, which is the fitting of probability mass or density function of certain random variables. To solve density estimation problems by using mixture models, machine learners have adopted the methodology to optimize both of the mixture weights and the basis functions, as with the EM algorithms and variational-Bayes methods to fit Gaussian mixture models (Dempster et al., 1977; Blei and Jordan, 2006). Yet these traditional mixture-model methods involve non-convex optimization and hence could be trapped in poor local optima. Our alternative methodology is to *fix* all of the basis functions during the optimization with only updating the mixture weights. This alternative optimization problem is convex, and we are able to obtain the global optimum of the model parameters simply by using gradient ascending methods. Depending on each application domain, we manually design the basis functions to cover all of the heterogeneous scales required in each application. This explicit design of the basis functions guarantees to yield the fat tails and the long-range dependence, unlike the unstable basis functions that are fitted with the traditional approaches without such guarantee. Remember that the main reason of realizing the highly-expressive and stably-fitted models is the replacement of the



Power-law distribution
as a mixture of Gaussian distributions

Power-lay decay
as a mixture of exponential decays

Figure 1.5: Producing a power-law distribution and a power-law decay by mixing Gaussian distributions and exponential decays, respectively. The blue lines represent the probability density function of a power-law distribution and the forgetting curve of a power-law decay, while the black lines represent those of Gaussian distributions and exponential decays. Mixture of Gaussian distributions that share the common mean while having different variances possess a fat tail to generate outliers. Mixture of exponential discounting curves having the different expected decay times produces a power-law discounting curve that declines quickly in short term while persistently survives in long term.

parametric mixtures, whose forms are limited, by the nonparametric ones having more flexible forms.

Our contributions in this dissertation contain both of the enhancement of the basis optimization algorithm to fit the mixture weights, and application-specific designs of the basis functions. While the optimization of nonparametric models having the fixed basis functions is convex, particularly in the travel-time density estimation tasks required for modeling the vehicle traffic, the existing EM algorithm does not provide a satisfiable convergence rate until reaching the global optimum. Hence we first introduce an accelerated optimization algorithm to perform density estimation via the nonparametric mixture model. Then we apply this accelerated algorithm for modeling the vehicle traffic, whose travel-time distributions possess fat tails due to the positive feedback through a large road network. The same philosophy of mixing multi-scale basis functions is also applied for a marketing problem that requires predictions of the responses by many consumers, who exhibit hyperbolic discounting in forgetting the influences of the events in the past. To specialize the algorithm for event prediction tasks, we formalize the response event prediction as a Poisson regression problem with exploiting the convex L_1-regularization methods (Tibshirani, 1994), instead of directly using the general density estimation algorithms. Modeling of the positive feedback by word-of-mouth is also discussed in this marketing work.

1.6 Normative Decision Making exploiting the Nonparametrics

We also introduce a rational decision-making algorithm that exploits the statistical and time-varying predictions about the behaviors of other humans. As an effective and real application example, we op-

timize the marketing-mix investment decisions with assuming that our predictions about the responses from consumers do not change the responses themselves. This assumption or ignorance of the optimizing behaviors by consumers is based on the fact that consumers do not have complete knowledge about the firms' marketing strategies, and remember that such assumption has also been adopted in the existing and successful marketing optimization approaches (Elsner et al., 2003; Abe et al., 2004; Tirenni et al., 2007b; Abe et al., 2009). We derive an effective Linear Programming (LP) algorithm to optimize the marketing-mix by solving a constrained Markov Decision problem, and experimentally confirm the validity of the optimized marketing-mix policy by using a discrete event simulation (e.g., (Delaney and Vaccari, 1998)).

Here the philosophy of decomposing the power-law decay into the exponential decays having multiple characteristic scales plays the essential role in defining the finite-dimensional state vector required in the Markov Decision Processes (MDPs). As the hyperbolic discounting corresponds to the dynamics of humans who are influenced by the events they have experienced in the past, transitions among the state vectors model the temporal dynamics about each consumer's perception toward the sequence of the marketing campaign events. Hence an optimized direct-marketing policy that is a complex function of the state vector allows us for offering the most adequate campaign to each consumer with chasing their time-varying mentality (e.g., loyalty and satisfaction). While our optimization problem is high-dimensional due to the combination of many timings, many segments, and many actions, we are able to obtain the global optimum of the marketing-mix policy because LP problem is convex and we are able to use very fast LP solvers.

1.7 Summary and the Structure of the Dissertation

In this section we newly introduced the nonparametric descriptive modeling to accurately predict the behaviors of interacting humans. The proposed models exploit mixtures of multi-scale basis functions, whose interpolation weights are stably fitted with global convex optimization algorithms while whose characteristic scales are manually chosen in order to suit each application domain. Thanks to the realization of both the flexible functional modeling and the stable fitting, the proposed models successfully incorporate all of the bounded rationality, positive feedback, and more complex functional structures to model the input stimuli and output responses for each human. The philosophy of adopting multiple scales to approximate the power-law functions also eases the implementation of a rational normative decision making algorithm based on constrained Markov Decision Processes.

The remainder of this dissertation is organized as follows. Chapter 2 discusses the related work more specific to each of the problem we discuss. Chapter 3 introduces the accelerated algorithm of the nonparametric density estimation, whose outcome is exploited for modeling the vehicle traffic in Chapter 4 via the travel-time density estimation. Chapter 5 introduces our marketing application to exploit the nonparametric mixture model for predicting responses by consumers. Normative decision making when we rely on the fitted descriptive models are finally addressed in Chapter 6. Chapter 7 concludes the dissertation.

We here comment on the style of mathematical notations in this dissertation. Due to the lots of the parameters required in formalizing nonparametric mixtures, models or optimization problems in all of these chapters involve many parameters and decision variables that waste lots of letters or symbols. Hence every parameter or decision variable is independently defined in each chapter, i.e., vector $\boldsymbol{x}$ in Chapter 3 has a different meaning from that of vector $\boldsymbol{x}$ in Chapter 4.

Chapter 2

Problem Specifications and Related Work

In this chapter, we substantiate each of the specific mathematical problems and discuss the related work. Many statistical prediction problems such as classification and regression problems are solved with supervised learning algorithms, which are unified as conditional density estimation problems. Hence stable and fast algorithms to perform density estimation greatly help us realize an effective economic decision making, which includes transportation decision making requiring travel-time predictions. Among lots of density estimation algorithms, we focus on the enforcement of the convex clustering (Lashkari and Golland, 2008) algorithm, which forms a probability density function as a nonparametric mixture of basis density functions, and which adopts a global optimization of the model parameters. Because the original EM algorithm for the convex clustering is inapplicable for real applications due to its extremely slow convergence, we propose a new accelerated algorithm to make the convex clustering applicable for the real applications including the transportation application. Section 2.1 addresses the prior work of applying nonparametric or mixture models for density estimation.

The original formulation of the convex clustering yields unconditional density estimators whose resulting probability density functions are not conditional on input and explanatory variables. In contrast, many supervised learning and prediction tasks handle input variables that affect the distributions of the output variables. As we show in Chapter 3, via a variable transformation, the convex clustering algorithm is shown to be equivalent with Kullback-Leibler Importance Estimation Procedure (KLIEP; Sugiyama et al. (2008)) that is a globally-optimal conditional density estimator, though the original gradient ascending algorithm in KLIEP also exhibits the slow convergence (Sugiyama et al., 2008). Hence our acceleration of the convex clustering algorithm automatically helps machine learners solve many supervised learning problems, as well as our specific conditional density estimation problem to model the vehicle traffic.

Section 2.2 surveys the travel-time prediction problems in modeling interacting vehicle traffic, and clarifies how the nonparametric density estimators provide advantageous predictions over the existing regression algorithms. A vehicle traffic is a complex outcome of interactions among the drivers in a road network, and a rational route choice decision is a result of certain optimization for the random travel-time associated with each route. One clear choice strategy is to minimize the expected travel time via Dijkstra's algorithm (Dijkstra, 1959), which requires a regression algorithm that mainly predicts the mean of travel time (e.g., (Robinson and Polak, 2005; Idé and Kato, 2009; Idé and Sugiyama, 2011)). In contrast, we focus on the fact that risks about travel time affect the preference order among the routes (Noland and Polak, 2002), and risk-sensitivities are heterogeneous among the drivers (Ulleberg and Rundmo, 2003; Chen, 2009). Because the main component required

for such risk-sensitive route decisions is the probability density function of the travel time for every link in a road network (Miller-Hooks and Mahmassani, 2000; Föllmer and Schied, 2004; Hardy and Wirch, 2004; Osogami, 2011), we formalize a conditional density estimation problem that exploits the accelerated convex clustering algorithm and whose explanatory variables are the identifiers of the links. Our nonparametric-mixture models are able to naturally incorporate the fat tails that are caused by rare but huge congestions.

Another important and practical application of the nonparametric descriptive modeling is the response prediction in marketing. Here the multi-scale mixtures are applied for modeling the long-range dependence of marketing-stimulus events that affect delaying responses and purchase decisions by individual consumers. Since the specific prediction task is formalized as a continuous-time event-spike prediction involving a regression of time-varying response probability, With providing a survey of the experimental facts and the existing approaches about the human memories and positive feedback, Section 2.3 addresses our prediction problem as a statistical fitting of Inhomogeneous Poisson Processes (IPPs) that predict the event spikes by incorporating both of the positive feedback and the long-range dependence.

Optimal marketing-decision making that exploits the event-spike prediction model is implemented as a constrained Markov decision problem, whose state is formed with historical sequence of events each consumer has experienced, and whose action is the type of marketing campaigns to target each individual consumer. Section 2.4 provides the literature of normative marketing-decision making based on the constrained Markov Decision Processes (cMDPs), to maximizing the mid-term revenues minus marketing costs. In addition to the advertising and direct-marketing applications using the cMDPs or optimal control theory, algorithmic studies to handle risk-sensitive constraints in cMDPs are relevant to our problem. The drawbacks in these existing approaches and directions in applying our philosophy of nonparametric mixtures for each of the problems are discussed in Section 2.5. Section 2.6 summarizes this chapter.

2.1 Nonparametric Density Estimation

Density estimation is a representative machine learning task that is deeply related with clustering, and solving the local optimality issue in clustering algorithms is an important problem. One way of grouping many data points into the less number of compact groups is to fit Gaussian mixture models, by k-means (MacQueen, 1967; Anderberg, 1973; Forgy, 1965), Expectation-Maximization (EM) (Dempster et al., 1977) algorithm, or Dirichlet Process Mixtures (DPMs) (Antoniak, 1974; Ferguson, 1983) as a nonparametric Bayesian approach. Estimation in DPMs is done with variational approximations (Blei and Jordan, 2006; Kurihara et al., 2007) or Markov Chain Monte Carlo methods (Ishwaran and James, 2001; Walker, 2007; Papaspiliopoulos and Roberts, 2008). Optimizations in the k-means, the EM algorithm, or the DPMs are non-convex and hence can converge into local optima, as shown in Figure 2.1. Typical strategies to avoid the poor local optima are to execute the fitting many times with random initializations, and Deterministic Annealing (DA) (Rose, 1998; Ueda and Nakano, 1998). In some applications, however, we need a clustering method whose partitioning results are reproducible and random initializations must be avoided in such applications. For example in marketing, clustering has been used to segment customers while unreproducible and initialization-dependent clusters are hard to be exploited in making effective customer-targeting strategies.

Existing clustering approaches that avoid the local optimality issue face the problems of computational inefficiency or unrobustness to noises. The convex clustering algorithm (Lashkari and Golland, 2008) we enforce is globally optimal while its convergence when using the EM algorithm is extremely

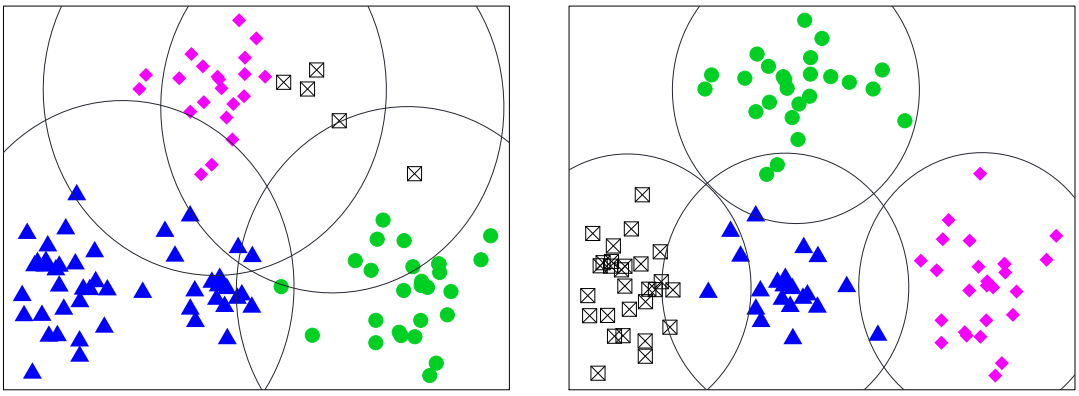


Figure 2.1: Local optima in 4-means clustering examples. Symbols '●', '▲', '♦', and '⊠' represent each sample's cluster assignment. In comparing the quality of clustering, which is measured as the log-likelihood of the fitted Gaussian mixture models, the assignment on the left figure is inferior to on that the right figure.

slow. Other representative algorithms to attain the global optimum in clustering are spanning-tree cutting (Zahn, 1971; Grygorash et al., 2006) and the mean-shift (Yang et al., 2003). Cutting the k-heaviest edges of the Euclidean minimum spanning tree for vector data can yield the global minimum of some loss function and can handle non-elliptic clusters. Yet, spanning-tree cutting is known to be sensitive to the noises, where existence of a few tail samples between the clusters strongly affects the resulting partitions. The mean-shift algorithm finds the local modes of a kernel density estimator and can also detect non-elliptic clusters while automatically determining the number of clusters. Yet, it is essentially an EM algorithm (Carreira-Perpiñán, 2007) that has the slow first-order convergence. Affinity Propagation (Frey and Dueck, 2007) is similar to the convex clustering and also achieves the globally optimal clustering depending on hyperparameter settings, while it is also computationally intensive due to the quadratic computational cost per iteration to the number of data points.

2.2 Travel Time Distributions in Traffic Modeling

Conditional density estimation of travel time is an essential task in modeling and simulating vehicle traffic realistically or navigating a car along the most desirable route. Automobile drivers are heterogeneous in that they tend to take different routes from a common origin to a common destination (Small and Winston, 1999; Wardman, 2001; Small et al., 2005). This heterogeneity partly stems from two facts about the distribution of travel time. First, drivers select their routes based not only on the expected travel time but also on the variability or distribution of the travel time (Noland and Polak, 2002). Second, the sensitivity to the risk associated with the travel time depends on particular drivers (Ulleberg and Rundmo, 2003; Chen, 2009). The effectiveness of traffic services such as personalized route recommendation (Rogers and Langley, 1998), traffic simulation (Ossen and Hoogendoorn, 2007; Barceló, 2010), and road pricing (Small and Yan, 2001; Verhoef and Small, 2004) heavily relies on how well we can model such drivers' heterogeneity.

The distribution of the travel time must be fitted for *every link* of a given road network with high accuracy. Stochastic optimization can then find the optimal route that minimizes the value of a risk measure that corresponds to a particular driver's sensitivity to risk. Minimizing risk measures results in diversified choices of routes, because the risk measures and their parameters vary among drivers,

depending on the individual risk sensitivities. In particular, one can satisfy the needs of drivers who want to avoid the risk of encountering large delays while allowing the increase of the expected travel-time. There is a large body of work on stochastic optimization for route search (e.g., (Miller-Hooks and Mahmassani, 2000)), and recent progress in risk-sensitive Markov decision processes (Osogami, 2011) allows us to find optimal routes with respect to a wide range of risk measures, including the Entropic Risk Measure (Föllmer and Schied, 2004) and Iterated Conditional Tail Expectation (Hardy and Wirch, 2004). The optimal route is found with dynamic programming based on the distributions assigned for each link, where the value of the risk measure is computed solely from the probability density functions of travel-time.

The literature of statistical travel-time prediction is limited to the Ordinary Least-Squares regression of only the expected travel-time, and thus risk-sensitive decisions cannot be properly derived from the existing travel-time prediction algorithms. Because distributions of the travel time are supposed to be functions of the geographic attributes of a route, which are represented as either two-dimensional coordinates or road network structures, we are able to predict the expected travel time with regression methods, such as the k-nearest neighbor method (Robinson and Polak, 2005), artificial neural networks (Wen et al., 2005b,a; Yu et al., 2008; Zhu and Wang, 2009), support vector regression (Dong and Pentland, 2009), and Gaussian process regression (Idé and Kato, 2009; Idé and Sugiyama, 2011). Yet the focus on the expected travel time, stemming from the Gaussian or log-normal parametric assumptions about the distributions, limits the modeling capability toward the actual travel-time distributions that are more complex than Gaussian or log-normal. For example, a small portion of extreme delays in traffic jams yields heavy-tailed or multimodal distributions, which are never be Gaussian or log-normal distributions.

2.3 Response Event-Spike Prediction in Consumer Behavior Modeling

Forecasting how marketing actions will impact consumers' purchase decisions is one of the essential issues in marketing investment decisions across multiple communication channels. Such predictive models should be designed to carefully estimate the indirect effects of actions, in several intermediate phases a consumer experiences before the final purchase decision. Each phase affects the evaluation of the promoted products or services (Townsend and Ashby, 1983; Busemeyer and Townsend, 1993; Roe et al., 2001; Otter et al., 2008; Scheibehenne et al., 2009) and updates the mental state of the consumer (Cannon et al., 2002). Such stimulated consumers exhibit self-active responses such as call center inquiries, web page browsing, and consumer-to-consumer information exchange known as word-of-mouth (Fudenberg, 1995). Word-of-mouth often causes booming trends and its economic impacts could become significant, because total amount of responses is simultaneously increased through coherent behaviors among consumers. Like the imitation-based herding of investors who cause boom and bust in stock market (Bala and Goyal, 1998), each individual is supposed to balance her/his prior knowledge and opinions from others (Blanchard and Watson, 1982; Roehner and Sornette, 2000), as word-of-mouth matters particularly for unfamiliar but interesting new products (Berger and Schwartz, 2011). Thus, marketers are required to clearly understand the complex dependencies among marketing campaigns, responses, and word-of-mouth events where the propensity to convert is modeled with both the individual and collective factors.

In estimating the dependency among marketing actions and responses by consumers, a crucial problem to consider is how to adequately estimate the influences of the unobserved word-of-mouth, by only handling sequences of observable event spikes. Figure 2.2 illustrates two types of the prediction problems involving multiple types of event spikes. In both of these problems, outbound marketing

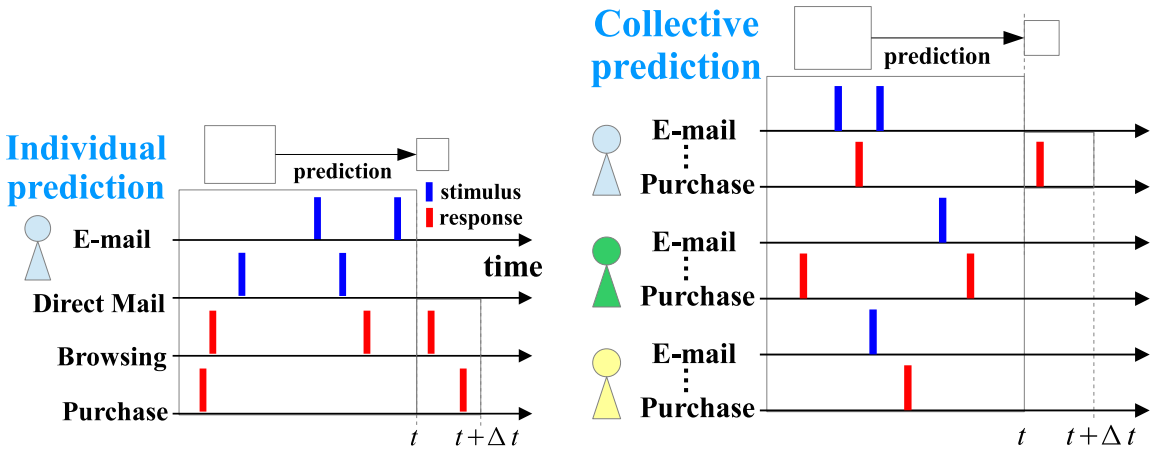


Figure 2.2: Two types of our continuous-time event prediction problems. Using all or parts of the stimulus and response events before time t, we aim to predict the occurrence of response events in time $[t, t+\Delta t)$ for each of the consumers. In predicting the occurrence of each consumer's response, we use either the past events of the same consumer (left) or the past events for a set of consumers (right).

actions and observable responses are recorded as event spikes, where each event is associated with a continuous time-stamp and a real-valued quantity such as a revenue amount. For predicting every consumer's response events in the future, we exploit either the past events of the same consumer or the past events for a set of consumers. Predictive models for practical marketers are expected to incorporate historical spikes as covariates to predict future response spikes. While observable on-line word-of-mouth can be represented as spikes, off-line word-of-mouth among real friends are not recorded in databases.

In the literature of predicting the responses by consumers, Inhomogeneous Poisson Processes (IPPs) have provided accurate predictions of continuous-time event spikes. An IPP introduces a time-varying intensity in the continuous-time Poisson processes and has widely been used for modeling events on consumer activities (Chatfield and Goodhardt, 1975; Schmittlein et al., 1987; Wagner and Taudes, 1986). It is well known that the impact of a stimulus decays over time (Deschatres and Sornette, 2005; Crane and Sornette, 2008), and estimation of some common time-decaying curves can be done with parametric approximations such as using probability density functions about a mixture of gamma distributions (e.g., (Simma and Jordan, 2010)). Beginning from the pioneering work by (Ebbinghaus, 1885), the nature of time decay in human memory has been observed as a power law or hyperbolic discounting with respect to elapsed time (Wixted, 1990; Rubin and Wenzel, 1996; Wixted and Ebbesen, 1997). The hyperbolic discounting possesses a long tail as a compound outcome of interactions among the short-, mid-, and long-term memories of each human. Often with incorporating such power-law time decay, IPPs have been widely used in predicting intermittent activities of consumers (Chatfield and Goodhardt, 1975; Schmittlein et al., 1987; Wagner and Taudes, 1986), because they can naturally model variable-interval sequences of event spikes.

IPPs have also been applied for modeling each consumer's mental process in choosing a brand or product from multiple options. Poisson race models (Townsend and Ashby, 1983; Otter et al., 2008) incorporate intermediate phases to evaluate products and predict real human decisions with high accuracy. A Poisson race model runs multiple IPPs in parallel, where each product is assigned

one IPP whose parameters are functions of the product attributes (e.g., price). A spike generated by each IPP represents not the final purchase, but a hidden trial to evaluate each product. The multiple IPPs represent competition among the products, because each consumer chooses only the first product whose number of trials exceeds a threshold value. The race to exceed the threshold originates from Multiattribute Decision Field Theory (Roe et al., 2001; Scheibehenne et al., 2009), which has been shown to successfully predict the context effects, such as the similarity effect (Tversky, 1972), the attraction effect (Huber et al., 1982), and the compromise effect (Simonson, 1989).

In econophysics studies, IPPs yielding positive feedback have been used for modeling either the persistent responses by a single consumer or word-of-mouth among multiple consumers. Since consumers tend to boost their own activities, series of responses often exhibits positive autocorrelation, called the "self-exciting behavior" (Hawkes, 1971; Sornette and Ouillon, 2005). For instance, time-series about the total count of book sales or online video views exhibits long-range positive autocorrelation (Deschatres and Sornette, 2005; Crane and Sornette, 2008). Self-excited Hawkes conditional Poisson process Hawkes (1971) is useful to model such autocorrelation, with ideas of individual-level epidemic branching (Goffman and Newill, 1964), information cascading (Bikhchandani et al., 2008), and limited human memory. A long-range point process yields a heavy-tail distribution of interval time between consecutive purchase events (e.g., Autoregressive Conditional Durations (Engle and Russell, 1998)), where inter-purchase time distributions provide some marketing implications (Allenby et al., 1999).

2.4 Normative Marketing-Mix Optimization

The predictions of the response-event spikes conditional on the sequences of marketing-stimulus events enable the assessment about the return on investment provided by a specific marketing-mix policy. Rational marketing-decision making that exploits time-series predictions has been regarded as an optimization problem to maximizing the mid-term revenues minus marketing costs. The underlying common principle is to exploit discrete-state Markov Decision Processes (MDPs) (Altman, 1999) or continuous-state optimal control theory (Bertsekas, 2000; Naik and Raman, 2003; Naik et al., 2005; Raman et al., 2012; Chow and Pavone, 2013), where these two approaches are connected through eigenproblem relaxations (Todorov, 2007, 2009). An MDP is a stochastic control process defined with a set of states, a set of actions, state-transition probability matrix conditional on each action, and stochastic immediate reward conditional on each pair of current state and action. MDPs and optimal control in marketing are classified with whether to target individual consumers, whether to play a stochastic game against competitors optimizing their own objectives, and type of action variables (e.g., boolean choice of direct mailing (Elsner et al., 2003; Abe et al., 2004), multinomial choice (Tirenni et al., 2007b; Abe et al., 2009), integer-valued credit assignment (Gómez-Pérez et al., 2009), and real-valued advertising and promotion budgets (Kumar et al., 2011; Naik and Raman, 2003; Naik et al., 2005; Raman et al., 2012)). In most of the game-theoretic formulations, every firm tries to control the total marketing costs with predicting aggregate-level sales and shares (Naik and Raman, 2003; Naik et al., 2005; Raman et al., 2012) while ignoring the individual-level responses by each consumer. In contrast, personalized targeting (Elsner et al., 2003; Abe et al., 2004; Tirenni et al., 2007b; Gómez-Pérez et al., 2009; Abe et al., 2009; Kumar et al., 2011) aims to lead each individual consumer or client company into high-profit states with controlling the aggregate-level budgets but usually ignoring competitors. While the state of each consumer must be treated as continuous variables, in actual it is approximated with a finite number of discrete states called the segments.

The normative optimization of the marketing-mix investments is algorithmically relevant to the

cMDP-based approaches involving risk-sensitive constraints and target populations as decision variables. It is usual that Linear Programming (LP) or Dynamic Programming (DP) is used in maximizing the expected cumulative reward in multiple periods, subject to certain bounds of the cumulative costs as random variables (Borkar and Jain, 2010; Ruszczynski, 2010; Osogami, 2011, 2012; Elsner et al., 2003; Abe et al., 2004; Tirenni et al., 2007b; Abe et al., 2009). We can classify such approaches into one of two classes, depending on whether the cost bounds accompany only the expectation operators (i.e., risk-neutral constraints (Borkar and Jain, 2010; Osogami, 2012; Elsner et al., 2003)) or more general risk measures (i.e. risk-sensitive constraints (Ross and Varadarajan, 1989, 1991; Borkar and Jain, 2010; Ruszczynski, 2010; Chow and Pavone, 2013)). Problems with risk-neutral constraints are directly solved with either DPs (Piunovskiy, 2006; Chen and Blankenship, 2004; Chen and Feinberg, 2007) or LPs (Djonin and Krishnamurthy, 2007). In contrast, incorporating risk-sensitive constraints has been a hot topic motivating the development of advanced algorithms, with roots in old telecommunication problems (Ross and Varadarajan, 1989, 1991) to maximize the throughput of traffic subject to constraints on delays. Imposing the constraints, not on the expected but on the realized delays, has a similar flavor to our marketing problem addressing the realized marketing costs. For rational decision making, the cMDPs should accompany time-consistent risk measures (Ruszczynski, 2010; Osogami, 2011, 2012; Chow and Pavone, 2013), where (Chow and Pavone, 2013) provides an actual implementation of such approach, based on the theory in (Ruszczynski, 2010). Directions to exploit LP in solving cMDPs are useful when we optimize target populations, as the summation of action probabilities. Optimizing budget over timing, segment, and channel for an airline company (Tirenni et al., 2007b) is one example of the closest problems to ours and is also solved with an LP. Abe et al. introduced another cMDP formulation (Abe et al., 2009) which solves an LP in each iteration of DP with discretizing high-dimensional continuous states, while we cannot exploit their assumption that the maximum or minimum budget remains unchanged for all periods in the future.

2.5 Discussion

Here let us discuss the common issue underlying in the limitations of many existing approaches. Both of the transportation and marketing applications in the prior arts have adopted parametric models, whose forecasting formulas need to strictly obey a limited class of predictive models, such as the Gaussian-distributed regression of travel time and the hyperbolic-decaying IPP models. We regard the limitation of computational powers as the main reason why these parametric models have been used, despite the high discrepancy between the assumed and the true forecasting formulas. In the travel-time prediction problem, the intention to realize the global optimality in fitting parameters limited the regression algorithms to follow the Ordinary Least Squares principle, and the limited predictive powers by the Gaussian or log-normal distributions are outcomes of the compromise to keep the computational efficiency. In the consumer-response prediction problem, even the simple parametric models of emulating the power laws resulted in non-convex optimization and high computational costs in fitting. Hence introduction of a nonparametric model, which has the more complexity than the parametric models, was not sufficiently investigated in designing the covariates of IPPs. Therefore, the main and direct direction to improve the predictive capabilities is a design of new forecasting formulas whose fitting enjoys the global optimality with reasonable computational costs. Our principle of nonparametric multi-scale mixtures is regarded as an effective instance to achieve these desirable properties. Among the several globally-optimal nonparametric density estimation or clustering algorithms, the focus on accelerating the convex clustering is positioned as an essential start point to make the nonparametric approaches applicable for large datasets.

In order to implement the entire algorithms, the adoption of the nonparametric models also necessitates additional statistical algorithms the prior work has not incorporated, in both of the transportation and marketing applications. For improving the predictive accuracies in the travel-time predictions, the literature adopted certain interpolation strategies with assuming similar links or between similar routes to have similar travel times. Replacement from the parametric Gaussian or log-normal distributions into the nonparametric mixtures also requires modification or new introduction of such interpolation strategies, which essentially emulate the propagation of traffic in a road network. The spiking nature of response event predictions in marketing requires us to design an efficient representation of the regression covariates, which both emulate the power-law decays in human memories and enjoy computational efficiency. Such efficient covariates must yield high predictive accuracies, even when the number of actual event spikes in training data is not sufficient. After designing an instance of the efficient covariates, the high-dimensionality nature of the nonparametric models requires us for implementing a special optimization algorithm to solve large-scale constrained Markov Decision Processes. In practice, only linear-programming approaches are feasible to target thousands of segments of consumers with incorporating complex budget constraints.

2.6 Summary

We reviewed the literature about nonparametric density estimation, traffic modeling, consumer-behavior modeling, and rational marketing-decision making that must adequately consider interactions among humans or among the multiple memories in one human. Ones of the crucial problems that have made nonparametric mixture models inapplicable for solving the real decision making problems is the local optimality and slow convergence. Accelerating the global convex clustering algorithm is beneficial for modeling lots of the travel-time distributions to evaluate the risks of a travel in a real road network. Such adequate risk evaluations yield both realistic simulation of the entire vehicle traffic and rational and risk-sensitive route-choice decision making. In marketing as another important decision making problem, one crucial component required in the rational marketing investment decisions is a continuous-time predictive model based on the IPPs, to provide the occurrence probability of a response event spike by each consumer, whose memory exhibits long-range dependence and who causes booming trends through word-of-mouth. Here we need to model the long-term forgetting curves depending on each type of the past events and the effects of the unobservable word-of-mouth, by designing a specialized Poisson regression algorithm with following the same philosophy of nonparametric descriptive modeling to achieve high generalization capabilities. The outcomes of such accurate consumer-response models are finally used for rational marketing-decision making, where the optimization algorithms are relevant to the risk-sensitive cMDPs.

Chapter 3

Global Optimization in Sparse Nonparametric Density Estimation

In this chapter we introduce an accelerated algorithm to solve a convex clustering problem (Lashkari and Golland, 2008), whose parameter fitting yields the global optimum of the fitted probability density function, and whose optimization objective is equivalent with that of a nonparametric conditional density estimation using the Kullback-Leibler Importance Estimation Procedure (KLIEP; Sugiyama et al. (2008)). Hence our goal in accelerating the convex clustering algorithm is to make the globally-optimal nonparametric mixture models applicable for large-scale real-world datasets, such as the probe-car datasets introduced in Chapter 4.

Nonparametric density estimation, which is a statistical estimation of complex probability density functions, is deeply related with the learning of mixture models such as the Gaussian mixture models. In the fitting of Gaussian mixture models whose centroids and bandwidths remain unchanged during the optimization, the negative log-likelihood to be minimized is convex with respect to the mixture weights, and many of the optimal mixture weights become zero (sparse). For n data points, the convex clustering utilizes n kernel distributions whose centers are the n data points themselves. The underlying optimization in the convex clustering algorithms automatically chooses a subset of the n kernel distributions whose cardinality is the number of clusters. The specified bandwidths essentially determine the number of clusters, where narrow bandwidths yield lots of small clusters while wide bandwidths result in a limited number of large clusters.

Though the global optimality in the convex clustering is quite appealing, we experimentally confirmed that the original EM algorithm provided in (Lashkari and Golland, 2008) requires thousands of iterations to converge. The EM algorithm is especially inefficient when applied to the convex clustering, mainly because of the sparsity of the solution. In the EM algorithm, the computational complexity varies in each iteration and is proportional to the number of non-zero elements in the mixture weights. Also, the EM algorithm has a first-order convergence and the updates are small near the sparse optimum. Thus, early pruning of the irrelevant kernels, which is a key to acceleration, conflicts with the nature of the EM algorithm. The actual computational times are sensitive to the specified threshold of the mixture weight.

Our acceleration technique exploits a fast pruning and an element-wise second-order optimization. Instead of using a small threshold, we use a derivative-based conditional expression to accurately judge whether or not a kernel can be trimmed off. By borrowing an idea from the Sequential Minimal Optimization (SMO; Platt (1999)), we regard such judgment as choice of a pair of kernels, and we utilize a nearest-neighbor method in choosing the pairs in each updating step. In optimizing

the non-zero elements of the mixture weights, we use an element-wise Newton-Raphson method, instead of the first-order EM algorithm. While our algorithm's computational complexity per iteration is the same as that of the EM algorithm, the combination of the fast pruning and the second-order Newton-Raphson method drastically reduces the required number of iterations. The element-wise unidimensional Newton-Raphson method instead of the standard multidimensional Newton-Raphson method is essential, because inversions of Hessian matrices that increase the computational complexity are successfully eliminated.

This chapter also provides an additional study to make the accelerated algorithm useful for many types of datasets, by introducing an iterative algorithm to handle high-dimensional data and enabling the model selection. For high-dimensional datasets, when the maximum-likelihood convex clustering is performed with the true bandwidths of the clusters, meaningful clusters cannot be acquired because many irrelevant kernels still remain active after the optimization. While using much larger bandwidths than the true ones often gives some clustering, we observed that the estimated cluster labels are unstable and different from the true labels when the noises are high. In addition, from several bandwidth settings, which clustering should be adopted is an open problem because the acquired mixture model is an improper density estimate and likelihood-based model selection such as likelihood cross-validation cannot be performed. We perform the convex clustering algorithm several times, where bandwidths are initialized by large values and succeedingly updated by smaller values. Since the number of clusters cannot be specified precisely in the convex clustering, practitioners often try multiple settings of the initial bandwidths. The optimal clustering from the multiple results is chosen baed on an empirical-Bayes method that selects appropriate bandwidths if the true clusters are Gaussian. The combination of the repetitions of the convex clusterings and the empirical-Bayes model selection achieves stable prediction performances compared to the existing mixture learning methods.

The main algorithms in this enhancement of nonparametric density estimation were first published in (Takahashi, 2011, 2012). From the conference paper (Takahashi, 2011) in the proceedings of the 11th SIAM International Conference on Data Mining (SDM 2011), which is published by the Society of Industrial and Applied Mathematics (SIAM), we reused formulations, figures, and implications based on the permission for authors[1]. From the journal paper (Takahashi, 2012) in Statistical Analysis and Data Mining, which is published by John Wiley and Sons, the author is permitted to reuse the full article based on a custom contract [2] through Rightslink® system of Copyright Clearance Center.

By beginning with the objective function of the convex clustering and its relationship with the nonparametric conditional density estimation, Section 3.1 address how the original EM algorithm in convex clustering exhibits slow convergence. Our fast pruning with Newton-Raphson updates is introduced in Section 3.2. Section 3.3 introduces the problems when applying the convex clustering to high-dimensional data, and the iterative refitting of the cluster parameters. While different initial bandwidths result in different final clusters, the empirical-Bayes method in Section 3.4 gives a criteria to choose the optimal one. Section 3.5 shows the experimental performances about the convergence rates and the detectability of the hidden clusters. Section 3.6 discusses the general applicability of the proposed algorithms in other tasks, and Section 3.7 summarizes this chapter.

[1] See the SIAM consent to publish `https://www.siam.org/students/siuro/consent.pdf`. Copyright © by SIAM. Unauthorized reproduction of this chapter is prohibited.

[2] The license number 3451120725708. Copyright © 2011 Wiley Periodicals, Inc.

3.1 Harmfully Slow Convergence in Convex Clustering

This section introduces the basics of the convex clustering algorithm. Section 3.1.1 addresses the convex negative log-likelihood and the EM algorithm first introduced in (Lashkari and Golland, 2008). Section 3.1.2 connects the objective of the convex clustering with that of the nonparametric conditional density estimation via a variable transformation. Section 3.1.3 provides an experimental evaluation of the learning speed with the EM algorithm, and discusses the reasons for the slow convergence.

3.1.1 Convexity in Minimizing Negative Log-Likelihood

A clustering of n data points is realized by maximizing the log-likelihood with which a mixture model generates the n data points. By rewriting such optimization problem as a minimization problem, we aim to minimize a negative log-likelihood of the data points. Let $p(\boldsymbol{x}|\boldsymbol{\theta})$ be a probability distribution of the d-dimensional vector $\boldsymbol{x} \in \mathbb{R}^d$, where the vector $\boldsymbol{\theta}$ represents the parameter of the distribution. While this chapter only discusses the case when $p(\boldsymbol{x}|\boldsymbol{\theta})$ is a Gaussian distribution, we allow for the distribution $p(\boldsymbol{x}|\boldsymbol{\theta})$ to belong a broader class of distributions, such as an exponential family whose natural parameter is $\boldsymbol{\theta}$. Let m be a number of clusters and Δ^{m-1} be the simplex in $\mathbb{R}^m$. Given a set of n data points $\mathcal{D} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$, let use minimize a negative log-likelihood

$$-\log p(\mathcal{D}|\boldsymbol{\lambda}, \boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_m) = -\sum_{j=1}^{n} \log \left[\sum_{i=1}^{m} \lambda_i p(\boldsymbol{x}_j|\boldsymbol{\theta}_i) \right], \tag{3.1}$$

where $\boldsymbol{\lambda} \triangleq (\lambda_1, \cdots, \lambda_m)^\top \in \Delta^{m-1}$ is a vector of mixture weights, and $\boldsymbol{\theta}_i$ is a distribution parameter assigned for the cluster i. The standard mixture modeling adopts the fiting of all of the parameters $\{\boldsymbol{\lambda}, \boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_m\}$, which is a non-convex optimization problem in most cases.

We focus on the fact that the optimization only for $\boldsymbol{\lambda}$ on some fixed values of $\{\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_m\}$ is convex (Csiszár and Shields, 2004; Lashkari and Golland, 2008)[3]. The value of the weight vector $\boldsymbol{\lambda}$, computed with gradient descending, hence converges into the global optimum. By using Jensen's inequality, we reach an iterative update rule as

$$\lambda_i^{(t+1)} \leftarrow \frac{1}{n} \sum_{j=1}^{n} \frac{\lambda_i^{(t)} \kappa_{ij}}{\sum_{i'=1}^{m} \lambda_{i'}^{(t)} \kappa_{i'j}},$$

where $\lambda_i^{(t)}$ is the estimate of the weight λ_i at the tth iteration and $\kappa_{ij} \triangleq p(\boldsymbol{x}_j|\boldsymbol{\theta}_i)$. By introducing the auxiliary variables $z_1^{(t)}, \cdots, z_n^{(t)}$ and $\eta_1^{(t)}, \cdots, \eta_m^{(t)}$, the update rules to be iterated are derived as

$$z_j^{(t)} \leftarrow \sum_{i=1}^{m} \lambda_i^{(t)} \kappa_{ij}, \; \eta_i^{(t)} \leftarrow \frac{1}{n} \sum_{j=1}^{n} \frac{\kappa_{ij}}{z_j}, \text{ and } \lambda_i^{(t+1)} \leftarrow \eta_i^{(t)} \lambda_i^{(t)}. \tag{3.2}$$

Since $\forall i, \lim_{t\to\infty} \eta_i^{(t)} = 1$, we are able exploit the values of $\eta_1^{(t)}, \cdots, \eta_m^{(t)}$ for the convergence test in the optimization. The converged values of the parameters provide a hard clustering based on a Naïve Bayes rule with which the cluster that $\boldsymbol{x}_j$ belongs to is $\max_i \lambda_i^{(\infty)} \kappa_{ij}$.

Optimization using the update rules (3.2) provides an automatic determination of the number of clusters. Based on the fact that many elements in the vector $\boldsymbol{\lambda}^{(t)} \triangleq (\lambda_1^{(t)}, \ldots, \lambda_m^{(t)})^\top$ converge into

[3]The proof is simply given with Jensen's inequality. For $\phi \in [0, 1]$ and $\boldsymbol{\lambda}, \boldsymbol{\lambda}' \in \Delta^{m-1}$, $-\log \left[\sum_{i=1}^{m} \left(\phi\lambda_j + (1-\phi)\lambda'_j \right) p(\boldsymbol{x}|\boldsymbol{\theta}_i) \right] \leq -\phi \log \left[\sum_{i=1}^{m} \lambda_j p(\boldsymbol{x}|\boldsymbol{\theta}_i) \right] - (1-\phi) \log \left[\sum_{i=1}^{m} \lambda_j p(\boldsymbol{x}|\boldsymbol{\theta}_i) \right]$.

zero as $t \to \infty$, we define a set of the indexes assigned positive weights, as $\mathcal{A}_t \triangleq \{i; \lambda_i^{(t)} > 0\}$. Let $\pi[i] \in \{1, \ldots, n\}$ be an index assigned for each cluster $i \in \{1, \ldots, m\}$. The index $\pi[i]$ is introduced for the convenience for handling large-scale datasets, i.e., when n is on the order of millions, then we set $m \ll n$ and each $\pi[i]$ is randomly chosen from $\{1, \ldots, n\}$ without duplication. For not so large dataset, we simply set $m = n$ and $\pi[i] = i$. By imposing $\boldsymbol{x}_{\pi[i]}$ to be the centroid (e.g., mean or medoid) of cluster i, we regard the distribution $p(\boldsymbol{x}|\boldsymbol{\theta}_i)$ as a distribution around the exemplar $\boldsymbol{x}_{\pi[i]}$. At each tth step in iterations, we call the distribution $p(\boldsymbol{x}|\boldsymbol{\theta}_i)$ an "active" cluster distribution if $i \in \mathcal{A}_t$, whose the cardinality $|\mathcal{A}_t|$ represents the number of clusters.

A simple but effective choice of $p(\boldsymbol{x}|\boldsymbol{\theta}_i)$ is an isotropic Gaussian kernel $p(\boldsymbol{x}|\boldsymbol{\theta}_i) \propto \exp\left(-\|\boldsymbol{x} - \boldsymbol{x}_{\pi[i]}\|_2^2 / \left(2\sigma^2\right)\right)$, where $\|\cdot\|_2$ denotes the L_2-norm and σ^2 is an isotropic variance. The number of clusters after convergence strongly depends on the value of the isotropic variance σ^2. When σ^2 is large, many elements of the weight vector $\boldsymbol{\lambda}$ become zero and we obtain a sparse nonparametric estimate of the Gaussian mixture model. In contrast, a small value of σ^2 results in lots of small clusters.

When we update the parameters with (3.2), the computational complexity at the tth step is $\mathcal{O}(n|\mathcal{A}_t|)$. Since $\lambda_i^{(t)} = 0$ automatically yields $\lambda_i^{(t+1)} = 0$, a kernel trimmed off at the tth step never becomes active. Therefore, for attaining an accurate and gloally-optimal estimate of the mixture model, the initial values of the mixture weights are required to be dense, i.e., $\forall i, \lambda_i^{(0)} > 0$. We simply adopt an initialization such that $\forall i, \lambda_i^{(0)} \equiv 1/n$. The computational complexity per iteration is first $\mathcal{O}(mn)$, while it gradually decreases as the iterations proceed.

3.1.2 Equivalence with a Nonparametric Conditional Density Estimation

We show that the optimization objective in convex clustering is equivalent to that of a conditional density estimation problem via a variable transformation. Many supervised learning problems, including classification and regression, are instances of conditional density estimation problems that fit $p(\widehat{\boldsymbol{y}}|\widehat{\boldsymbol{x}})$, where $\widehat{\boldsymbol{x}} \in \mathbb{R}^{d_X}$ and $\widehat{\boldsymbol{y}} \in \mathbb{R}^{d_Y}$ are the vectors of input and output variables, respectively. Given a set of n input and output observations $\{(\widehat{\boldsymbol{x}}_1, \widehat{\boldsymbol{y}}_1), \ldots, (\widehat{\boldsymbol{x}}_n, \widehat{\boldsymbol{y}}_n)\}$, let us consider a weighted kernel estimate of the conditional density

$$p(\widehat{\boldsymbol{y}}|\widehat{\boldsymbol{x}}) = \frac{\sum_{i=1}^m w_i K_X(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{x}}_{\pi[i]}) K_Y(\widehat{\boldsymbol{y}}, \widehat{\boldsymbol{y}}_{\pi[i]})}{\sum_{i=1}^m w_i K_X(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{x}}_{\pi[i]})}, \tag{3.3}$$

where $\widehat{\boldsymbol{x}}_{\pi[1]}, \ldots, \widehat{\boldsymbol{x}}_{\pi[m]}$ and $\widehat{\boldsymbol{y}}_{\pi[1]}, \ldots, \widehat{\boldsymbol{y}}_{\pi[m]}$ are $m (\leq n)$ random samples of the training data, $K_X(\cdot, \widehat{\boldsymbol{x}}_{\pi[i]})$ and $K_Y(\cdot, \widehat{\boldsymbol{y}}_{\pi[i]})$ are kernel distributions of input and output variables, respectively. For example, we would assume $K_X(\cdot, \widehat{\boldsymbol{x}}_{\pi[i]})$ and $K_Y(\cdot, \widehat{\boldsymbol{y}}_{\pi[i]})$ to be multivariate Gaussian distributions whose centers are $\widehat{\boldsymbol{x}}_{\pi[i]}$ and $\widehat{\boldsymbol{y}}_{\pi[i]}$.

The optimization of the mixture weights $\boldsymbol{w} = (w_1, \ldots, w_m)^\top$ in (3.3), while K_X and K_Y remain unchanged, is essentially the same as that in the convex clustering. Kullback-Leibler Importance Estimation Procedure Sugiyama et al. (2008) performs an optimization

$$\max_{\boldsymbol{w}} \sum_{j=1}^n \log\left[\sum_{i=1}^m w_i K_X\left(\widehat{\boldsymbol{x}}_j, \widehat{\boldsymbol{x}}_{\pi[i]}\right) K_Y\left(\widehat{\boldsymbol{y}}_j, \widehat{\boldsymbol{y}}_{\pi[i]}\right)\right] \text{ subject to } \sum_{j=1}^n \sum_{i=1}^m K_X\left(\widehat{\boldsymbol{x}}_j, \widehat{\boldsymbol{x}}_{\pi[i]}\right) w_i = n, \tag{3.4}$$

which is derived as the maximization of an empirical approximation of the density ratio between the joint distribution $p(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}})$ and the marginal distribution $p(\widehat{\boldsymbol{x}})$.

Let us introduce a variable transformation $\lambda_i \triangleq \frac{w_i}{n} \sum_{j=1}^{n} K_X\left(\widehat{\boldsymbol{x}}_j, \widehat{\boldsymbol{x}}_{\pi[i]}\right)$. Then Optimization (3.4) is modified as

$$\max_{\boldsymbol{\lambda}} \sum_{j=1}^{n} \log \left[\sum_{i=1}^{m} \lambda_i \frac{n K_X\left(\widehat{\boldsymbol{x}}_j, \widehat{\boldsymbol{x}}_{\pi[i]}\right)}{\sum_{j'=1}^{n} K_X\left(\widehat{\boldsymbol{x}}_{j'}, \widehat{\boldsymbol{x}}_{\pi[i]}\right)} K_Y\left(\widehat{\boldsymbol{y}}_j, \widehat{\boldsymbol{y}}_{\pi[i]}\right) \right] \text{ subject to } \sum_{i=1}^{m} \lambda_i = 1,$$

which adopts an equivalent objective to that of the convex clustering, when we set

$$\kappa_{ij} = \frac{n K_X\left(\widehat{\boldsymbol{x}}_j, \widehat{\boldsymbol{x}}_{\pi[i]}\right)}{\sum_{j'=1}^{n} K_X\left(\widehat{\boldsymbol{x}}_{j'}, \widehat{\boldsymbol{x}}_{\pi[i]}\right)} K_Y\left(\widehat{\boldsymbol{y}}_j, \widehat{\boldsymbol{y}}_{\pi[i]}\right).$$

3.1.3 Extreme Number of Iterations Required in the EM Updates

Despite the appealing global optimality and easy implementability of the EM algorithm using (3.2), we experimentally confirmed that lots of iterations are required until the mixture weights become sparse. (Lashkari and Golland, 2008) recommends a pruning of component i that satisfies $\lambda_i^{(t)} < 10^{-3}/n$, for reaching the sparse solution within finite time. Figure 3.1 demonstrates an experimental evaluation of the learning rate when applying this thresholding rule for an artificial 2D dataset. For n data points, even n iterations are not sufficient for pruning all of the irrelevant kernels.

We must note that the slow pruning of the irrelevant kernels is the nature of the EM algorithm. Since EM algorithms to learn mixture models have first-order convergence (Redner and Walker, 1984; Xu and Jordan, 1995), the update amounts become small near the optimum, which is a sparse vector in our convex clustering problem. Hence the small updates nearby the optimum implies the prevention of the pruning for irrelevant kernels.

An alternative approach to realize reasonable computational costs is to use a loose threshold, which is larger than the recommendation$10^{-3}/n$, but this naïve acceleration does not provide adequate clusters. In Figure 3.1, the updates of the mixture weights are not monotonic for the number of iterations. The weight of a component grows in some iterations and shrinks in other iterations. Hence pruning with a loose threshold involves a risk of trimming off the relevant kernels and makes the optimizations unstable.

3.2 The Accelerations

To replace the slow and first-order EM algorithm combined with the heuristic thresholding, we introduce a fast and second-order Newton-Raphson method supported by an exact pruning rule. A notable property of the optimization objective in the convex clustering exists in its sub-problem focusing on a *pair* of kernels. In optimizing the mixture weights only for such pair of kernels, we are able to exactly determine whether or not the selected kernels should be pruned. We regard the idea of focusing on a pair of kernels as similar to that in Sequential Minimal Optimization (Platt, 1999), which is originally a fast optimization technique to train Support Vector Machines (SVMs; Vapnik (1995)).

While the original SMO train SVMs exploits the Karush-Kuhn-Tucker conditions (Kuhn and Tucker, 1951) for picking up the pairs of kernels, we more carefully analyze the characteristics of the pairs to be pruned. Our optimization strategy is implemented as a combination of fast pruning using a derivative-based rule to determine whether or net one of the two mixture weights should become zero, and fast updating to optimize the non-zero elements via Newton-Raphson updates. The element-wise updating in our strategy successfully excludes the necessity of computationally-intensive evaluation

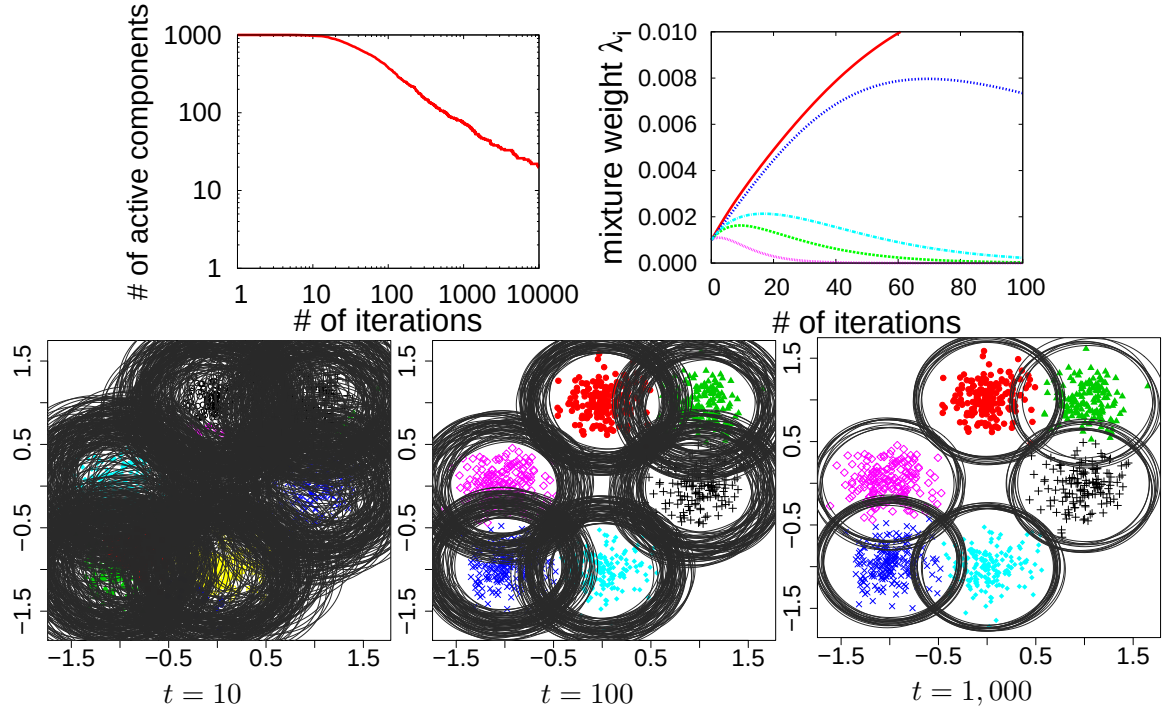
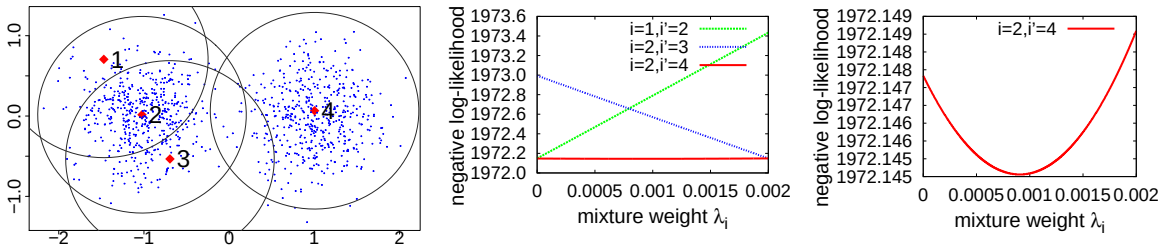


Figure 3.1: Slow convergence of the EM algorithm when applying a thresholding rule $\lambda_i^{(t+1)} \leftarrow 0$ if $\lambda_i^{(t)} < 10^{-3}/n$. 1,000 samples of $\mathbb{R}^2$ points are distributed from a 6-cluster Gaussian mixture model. The centroids of the clusters are $(1,0)^\top$, $(-1,0)^\top$, $(0,1)^\top$, $(0,-1)^\top$, $(1,1)^\top$, and $(-1,-1)^\top$. Every cluster has its mixture weight 1/6 and standard deviation 0.2, while we specified a learning bandwidth $\sigma := 0.3$ in optimization. The top-left figure plots the numbers of the active clusters for each step of the iterations, and we should confirm the fact that even 1,000 or 10,000 iterations are not sufficient for reaching the convergence. The top-right figure shows the dynamics in updating the mixture weights $\lambda_i^{(0)}, \lambda_i^{(1)}, \ldots, \lambda_i^{(100)}$ for several samples of the cluster indexes $\{i\}$. The updates are not monotonic, because many of the mixture weights first increase and later decrease. The slow convergence can also be confirmed in the bottom three figures, where the resulting partitions at 10, 100, or 1,000 iterations are shown.

of the Hessian matrix that multivariate Newton-Raphson methods require to compute, and keeps the computational complexity per iteration in our algorithm to be the same as that of the EM algorithm. Based on a specific analysis to our problem, we use a nearest-neighbor method for choosing the pair of kernels. The analysis in Section 3.2.1 clarifies a direction to accelerate the convex clustering. Section 3.2.2 derives the fast and exact pruning conditions, and Section 3.2.3 addresses the element-wise Newton-Raphson updating. Section 3.2.4 mentions several issues to be cared in implementing the proposed algorithm.

3.2.1 Analysis for a Pair of Kernels

Let (i, i') be a pair of components we focus on, and let us introduce a sub-problem of the entire optimization (3.1) with fixing the mixture weight for every other component $u \neq i, i'$. We are able to regard the values $\lambda_{\backslash i \backslash i'} \triangleq 1 - \lambda_i - \lambda_{i'}$ and $c_{j \backslash i \backslash i'} \triangleq \sum_{u \neq i, i'} \lambda_u p(\boldsymbol{x}_j | \boldsymbol{\theta}_u)$ as constants in optimization. By



Selected kernels

Shapes of the negative conditional log-likelihood

Figure 3.2: The three types of the shapes of the constrained negative log-likelihood with respect to one mixture weight λ_i, by taking an example case when $\forall u \neq i, i', \lambda_u \equiv 1/n$. For the selected points 1, 2, 3, and 4 in the left-most figure, we provide the shapes of the negative log-likelihood functions, by picking up pairs $(i, i') \in \{(1,2),(2,3),(2,4)\}$. The case when $(i, i') = (2,4)$ is magnified in the right-most figure, where the optimum is located in an intermediate point and does not correspond to the sparse solution.

exploiting the fact that $\lambda_i + \lambda_{i'} \equiv 1 - \lambda_{\backslash i \backslash i'}$, we rewrite (3.1) by a unidimensional function $f_{ii'}(\lambda_i)$ with respect to one scalar parameter $\lambda_i \in [0, 1 - \lambda_{\backslash i \backslash i'}]$.

An analysis based on the derivative of the unidimensional function yields an exact pruning rule. Figure 3.2 shows three typical shapes of the function $f_{ii'}(\lambda_i)$ depending on the choice of the pair (i, i'). We are able to immediately set $\lambda_i = 0$ or $\lambda_{i'} = 0$, when $f_{ii'}(\cdot)$ is monotonically decreasing or increasing, respectively. A generalized fact is that when two kernels $p(\boldsymbol{x}|\boldsymbol{\theta}_i)$ and $p(\boldsymbol{x}|\boldsymbol{\theta}_{i'})$ have similar probability density functions, then only one of the pair (i, i') becomes an active kernel and the other should be eliminated in optimization.

3.2.2 Fast and Exact Pruning

By assessing the monotonicity of $f_{ii'}(\cdot)$, we derive a pruning condition for the irrelevant kernels, whose computation is fast. Let us take the gradient of $f_{ii'}(\lambda_i)$ at $\lambda_i = 0$, as

$$f'_{i0i'} \triangleq \left. \frac{\partial f_{ii'}}{\partial \lambda_i} \right|_{\lambda_i = 0} = -\sum_{j=1}^{n} \frac{\kappa_{ij} - \kappa_{i'j}}{\left(1 - \lambda_{\backslash i \backslash i'}\right) \kappa_{i'j} + c_{j \backslash i \backslash i'}}. \tag{3.5}$$

If $f'_{i0i'} > 0$, then $f_{ii'}(\cdot)$ is monotonically increasing within $[0, 1 - \lambda_{\backslash i \backslash i'}]$, and hence $\lambda_i = 0$ and $\lambda_{i'} = 1 - \lambda_{\backslash i \backslash i'}$. In the same way, we evaluate the gradient in another boundary, as

$$f'_{ii'0} \triangleq \left. \frac{\partial f_{ii'}}{\partial \lambda_i} \right|_{\lambda_i = 1 - \lambda_{\backslash i \backslash i'}} = -\sum_{j=1}^{n} \frac{\kappa_{ij} - \kappa_{i'j}}{\left(1 - \lambda_{\backslash i \backslash i'}\right) \kappa_{ij} + c_{j \backslash i \backslash i'}}. \tag{3.6}$$

If $f'_{ii'0} < 0$, then $f_{ii'}(\cdot)$ is monotonically decreasing within $[0, 1 - \lambda_{\backslash i \backslash i'}]$, and hence $\lambda_i = 1 - \lambda_{\backslash i \backslash i'}$ and $\lambda_{i'} = 0$. If $f'_{i0i'} < 0 \wedge f'_{ii'0} > 0$ where $\wedge$ denotes the logical product ("and" operator), then the optimum of the parameter λ_i is an interior point in the interval $[0, 1 - \lambda_{\backslash i \backslash i'}]$. Note that there is no case such that $f'_{i0i'} > 0 \wedge f'_{ii'0} < 0$, because $f_{ii'}$ is a convex function.

Let us defer the optimization for such non-zero values in Section 3.2.3 and keep on focusing on the fast pruning. For each iteration, we update the mixture weights for all of the active components. At the tth iteration, since the number of active components is $|\mathcal{A}_t|$, the computational cost to evaluate (3.5) and (3.6) is $\mathcal{O}(n|\mathcal{A}_t|)$, which is the same as in the EM algorithm.

One practical way to make the optimization efficient is supplied by a useful rule to pick up the pair (i, i'). Our idea based on the analysis in Section 3.2.1 is is the choice of index i' as one of the neighbors of the index i. The optimum of the parameter λ_i or $\lambda_{i'}$ tends to be zero when $p(\boldsymbol{x}|\boldsymbol{\theta}_i)$ and $p(\boldsymbol{x}|\boldsymbol{\theta}_{i'})$ are similar. Hence taking pairs $\{(i, i')\}$, such that i' is a neighbor of i, increase the chances to prune the irrelevant kernels. In advance of the main updating steps, we pre-compute a sequence of indexes $\varepsilon[i, 1], \varepsilon[i, 2], \ldots, \varepsilon[i, m-1]$ such that $\varepsilon[i, k]$ is the k-nearest neighbor of i, in terms of a similarity measure (e.g., Euclid distance) for the space of the parameters $(\boldsymbol{\theta}_i)_{i=1}^m$. For each step in iterations, we set $i' = \varepsilon[i, k]$ with minimum possible k such that $\varepsilon[i, k]$ is an active component. Also, because i is eliminated with high probability when the weight λ_i is small, the pruning judgment based on (3.5) and (3.6) is performed on the ascending order of λ_i.

3.2.3 Element-Wise Newton-Raphson Updating

Another essential technique to accelerate the updates is the second-order optimization of the non-zero elements. For the unidimensional function $f_{ii'}(\cdot)$, introduced in Section 3.2.2, the first-order and second-order derivatives $h_{ii'}^{(1)} \triangleq \dfrac{\partial f_{ii'}}{\partial \lambda_i}$ and $h_{ii'}^{(2)} \triangleq \dfrac{\partial^2 f_{ii'}}{\partial \lambda_i^2}$ are given as

$$h_{ii'}^{(1)} = -\sum_{j=1}^{n} \frac{\kappa_{ij} - \kappa_{i'j}}{\lambda_i \kappa_{ij} + \left(1 - \lambda_{\backslash i \backslash i'} - \lambda_i\right) \kappa_{i'j} + c_{j \backslash i \backslash i'}} \text{ and}$$

$$h_{ii'}^{(2)} = \sum_{j=1}^{n} \left[\frac{\kappa_{ij} - \kappa_{i'j}}{\lambda_i \kappa_{ij} + \left(1 - \lambda_{\backslash i \backslash i'} - \lambda_i\right) \kappa_{i'j} + c_{j \backslash i \backslash i'}} \right]^2 > 0,$$

respectively. We derive an element-wise Newton-Raphson update rule

$$\lambda_i^{(t+1)} \leftarrow \lambda_i^{(t)} - h_{ii'}^{(1)} / h_{ii'}^{(2)},$$

whose execution is stable thanks to the convexity of the function $f_{ii'}(\cdot)$. The computational complexity per iteration to evaluate $h_{ii'}^{(1)}$ and $h_{ii'}^{(2)}$ for every index $i \in \mathcal{A}_t$ is also $\mathcal{O}(n|\mathcal{A}_t|)$ that is essentially the same as our fast pruning.

3.2.4 Implementation Notes

Let us complement the points in implementing the algorithm. By using matrix notations, we summarized the new accelerated convex clustering algorithm in Algorithm 1 that introduces several auxiliary variables for quick computations.

The vector $\boldsymbol{z} = (z_1, \ldots, z_n)^\top$ caches the current value of $\sum_{i=1}^m \lambda_i \kappa_{ij}$ for every index $j \in \{1, \ldots, n\}$. In judging whether a component i should be pruned or not, a modified value of the summation $\sum_{i=1}^m \lambda_i \kappa_{ij}$ for each j is calculated by setting $\lambda_i = 0$ or $\lambda_{i'} = 0$. Because only λ_i and $\lambda_{i'}$ are tried to be updated, always summing the values $\{\lambda_i \kappa_{ij}\}$ for all of $i \in \{1, \ldots, m\}$ is inefficient. We utilize the values of $\boldsymbol{z}$ and another vector $\boldsymbol{v} = (v_1, \ldots, v_n)^\top$, which caches the gradient values at the corner $\lambda_i = 0$ or $\lambda_{i'} = 0$.

The value of $\sum_{i=1}^m \lambda_i \kappa_{ij}$ is quickly computed when the modification $\lambda_i = 0$ or $\lambda_{i'} = 0$ is applied, as $v_j = z_j - \lambda_i \kappa_{ij} + \lambda_i \kappa_{i'j}$ or $v_j = z_j + \lambda_{i'} \kappa_{ij} - \lambda_{i'} \kappa_{i'j}$, respectively. If the pruning condition is satisfied, then we substitute each element of the vector $\boldsymbol{v}$ into that of the vector $\boldsymbol{z}$. Otherwise, the values of the vector $\boldsymbol{v}$ are ignored and the algorithm proceeds into the next step. For each kernel i,

the calculations of the vectors $\boldsymbol{v}$ and $\boldsymbol{z}$ require $\mathcal{O}(n)$ computational costs. Hence the computational complexity per iteration in Algorithm 1 is $\mathcal{O}(n|\mathcal{A}_t|)$.

The repeating additions and subtractions for the vector $\boldsymbol{z}$ might integrate small numerical errors. Before the pruning trials in each iteration, we recommend to re-calculate the vector $\boldsymbol{z}$ using all of the active mixture weights and kernel values. This re-calculation only requires $\mathcal{O}(n|\mathcal{A}_t|)$ and does not essentially increase the computational complexity.

For high-dimensional data, we must care the fact that every value of the probability density function $p(\boldsymbol{x}_j|\boldsymbol{\theta}_i)$ is low due to the exponential decay with respect to the dimension d. We are easily able to avoid the resulting underflow, by taking the logarithms of the densities. Using a value $lp_j \triangleq \max_i \log p(\boldsymbol{x}_j|\boldsymbol{\theta}_i)$ that can be stably calculated, we use a normalized kernel matrix $\overline{\boldsymbol{K}} = (\overline{\kappa}_{ij})$ such that

$$\overline{\kappa}_{ij} = \frac{p(\boldsymbol{x}_j|\boldsymbol{\theta}_i)}{\sum_{i=1}^m p(\boldsymbol{x}_j|\boldsymbol{\theta}_i)} = \frac{\exp\left(\log p(\boldsymbol{x}_j|\boldsymbol{\theta}_i) - lp_j\right)}{\sum_{i=1}^m \exp\left(\log p(\boldsymbol{x}_j|\boldsymbol{\theta}_i) - lp_j\right)},$$

instead of the original kernel matrix $\boldsymbol{K}$. The optimum of the mixture weight vector $\boldsymbol{\lambda}$, with the normalized kernel matrix $\overline{\boldsymbol{K}}$, is the same as that with the original kernel matrix $\boldsymbol{K}$, because

$$\arg\max_{\boldsymbol{\lambda}} \sum_{j=1}^n \log \sum_{i=1}^m \lambda_i \overline{\kappa}_{ij} = \arg\max_{\boldsymbol{\lambda}} \sum_{j=1}^n \log \sum_{i=1}^m \lambda_i \kappa_{ij} - \left(\sum_{j=1}^n lp_j\right),$$

and $\left(\sum_{j=1}^n lp_j\right)$ is a constant that does not affect the outcome of the optimization.

3.3 Bandwidth Choice and Iterative Refitting for High-Dimensional Data

This section discusses how to efficiently choose appropriate bandwidths of each cluster to yield highly-predictive distributions. Our first heuristic introduced in Section 3.3.1 is a local Maximum-Likelihood (ML) estimation of the adaptive bandwidths. The local ML estimation is an example of implementing the multi-scale basis density functions to handle heteroscedastic data, while it is practical only for low-dimensional random variables. For high-dimensional data, we show the fact that the mixture weights acquired with Algorithm 1 are not sparse even when the correct bandwidths are used. Section 3.3.2 discusses the resulting inconsistency between the objective of clustering and that of density estimation, even when we adopt the same convex clustering formalisms. With addressing the necessity of refitting the cluster parameters, we discuss an annealing process to stably obtain clustering in Section 3.3.3. Based on the consideration about the refitting, Section 3.3.4 introduces an iterative algorithm that repeats convex clustering several times.

3.3.1 Local Maximum-Likelihood with a k-Nearest Neighbor Method

Real data exhibit heteroscedasticity, with both sparse and dense regions existing in the same dataset. For such data, using a globally-common bandwidth for every kernel causes over-partitioning in the sparse regions and under-partitioning in the dense regions, as shown in Figure 3.3. A natural improvement is to incorporate adaptive bandwidths, where each cluster has its own local bandwidth.

Let σ_i^2 be an isotropic variance assigned for kernel i, whose center is the exemplar $\boldsymbol{x}_{\pi[i]}$. We consider a heteroscedastic sparse mixture model formalized as

$$p\left(\boldsymbol{x}\right) = \sum_{i=1}^m \lambda_i \mathcal{N}\left(\boldsymbol{x}; \boldsymbol{x}_{\pi[i]}, \sigma_i^2 \boldsymbol{I}_d\right) \equiv \sum_{i=1}^m \frac{\lambda_i}{\sqrt{2\pi\sigma_i^2}^d} \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}_{\pi[i]}\|_2^2}{2\sigma_i^2}\right),$$

Algorithm 1 Fast Convex Clustering

input An $(n \times m)$ kernel matrix $\boldsymbol{K} = (\boldsymbol{\kappa}_1, \ldots, \boldsymbol{\kappa}_m)$
such that $\boldsymbol{\kappa}_i \triangleq (\kappa_{i1} \equiv p(\boldsymbol{x}_1|\boldsymbol{\theta}_i), \ldots, \kappa_{in} \equiv p(\boldsymbol{x}_n|\boldsymbol{\theta}_i))^\top$

function $\widehat{\boldsymbol{\lambda}} = FastConvexClustering(\boldsymbol{K})$
  $\boldsymbol{\lambda}^{(0)} \leftarrow (1/m, \ldots, 1/m)^\top$ and $\mathcal{S} \leftarrow \{1, \ldots, m\}$
  Allocate temporary variables $\boldsymbol{v} = (v_1, \ldots, v_n)^\top$ and $\boldsymbol{z} = (z_1, \ldots, z_n)^\top$
  for $i = 1$ **to** m **do**
    Sort the indexes $(\varepsilon[i,1], \ldots, \varepsilon[i,m-1])$ to satisfy $\varepsilon[i,k]$ is the k-nearest neighbor of i.
  end for
  $t \leftarrow 0$
  repeat
    $\boldsymbol{z} \leftarrow \boldsymbol{K}\boldsymbol{\lambda}^{(t)}$
    for $i \in \mathcal{S}$ (ascending order of λ_i) **do**
      $i' \leftarrow \min_k \varepsilon[i,k]$ s.t. $\varepsilon[i,k] \in \mathcal{S}$
      $\boldsymbol{v} \leftarrow \boldsymbol{z} + \lambda_i^{(t)} (\boldsymbol{\kappa}_{i'} - \boldsymbol{\kappa}_i)$.
      $f'_{i0i'} \leftarrow -\sum_{j=1}^n \left(\kappa_{ij} - \kappa_{i'j}\right) / v_j$
      if $f'_{i0i'} > 0$ **then**
        $\lambda_i^{(t+1)} \leftarrow 0$, $\lambda_{i'}^{(t+1)} \leftarrow \lambda_i^{(t)} + \lambda_{i'}^{(t)}$, $\boldsymbol{z} \leftarrow \boldsymbol{v}$
        Remove i from $\mathcal{S}$
      else
        $\boldsymbol{v} \leftarrow \boldsymbol{z} + \lambda_{i'}^{(t)} (\boldsymbol{\kappa}_i - \boldsymbol{\kappa}_{i'})$
        $f'_{ii'0} \leftarrow -\sum_{j=1}^n \left(\kappa_{ij} - \kappa_{i'j}\right) / v_j$
        if $f'_{ii'0} < 0$ **then**
          $\lambda_i^{(t+1)} \leftarrow \lambda_i^{(t)} + \lambda_{i'}^{(t)}$, $\lambda_{i'}^{(t+1)} \leftarrow 0$, $\boldsymbol{z} \leftarrow \boldsymbol{v}$
          Remove i' from $\mathcal{S}$
        else
          $h_{ii'}^{(1)} \leftarrow -\sum_{j=1}^n \left(\kappa_{ij} - \kappa_{i'j}\right) / z_j$
          $h_{ii'}^{(2)} \leftarrow \sum_{j=1}^n \left(\kappa_{ij} - \kappa_{i'j}\right)^2 / z_j^2$
          $\lambda_i^{(t+1)} \leftarrow \lambda_i^{(t)} - h_{ii'}^{(1)} / h_{ii'}^{(2)}$
          $\lambda_{i'}^{(t+1)} \leftarrow \lambda_i^{(t)} + \lambda_{i'}^{(t)} - \lambda_i^{(t+1)}$
          $\boldsymbol{z} \leftarrow \boldsymbol{z} + \left(\lambda_i^{(t+1)} - \lambda_i^{(t)}\right)\boldsymbol{\kappa}_i + \left(\lambda_{i'}^{(t+1)} - \lambda_{i'}^{(t)}\right)\boldsymbol{\kappa}_{i'}$
        end if
      end if
    end for
    $t \leftarrow t+1$
  until $\boldsymbol{\lambda}^{(t)}$ converges
end function

output The converged mixture weight vector $\boldsymbol{\lambda}^{(t)}$.

where $\boldsymbol{I}_d$ is the d-dimensional identity matrix and $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the probability density function of the multivariate Gaussian distribution whose mean is $\boldsymbol{\mu}$ and whose variance-covariance matrix is $\boldsymbol{\Sigma}$.

Let ξ be a rough value of the number of clusters. When we assume that the converged values of the non-zero mixture weights, after the iterative optimization, are similar to one another, each of the

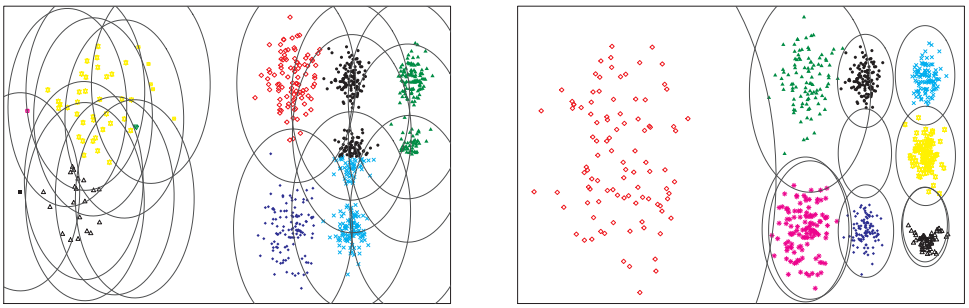


Figure 3.3: An inappropriate partitioning of heteroscedastic data using only a single bandwidth. The datasets in the two figures are the same, with low-density, middle-density, and high-density regions located at the left, the center, and the right, respectively. The left figure shows a clustering with a single bandwidth that is adjusted to the center clusters, where the low-density regions are over-partitioned while the high-density regions are under-partitioned. As shown in the right figure, convex clustering using adaptive bandwidths is able to provide a desirable partitioning and a probability density function.

cluster has about (n/ξ) members distributed around its centroid. Because each local bandwidth should be determined with these (n/ξ) members, a local Maximum-Likelihood (ML) estimate of σ_i^2 is given as

$$\widehat{\sigma}_i^2 = \frac{1}{kd} \sum_{\ell=1}^{k} \|\boldsymbol{x}_{\varepsilon[i,\ell]} - \boldsymbol{x}_{\pi[i]}\|_2^2, \tag{3.7}$$

where k is a rounded integer of (n/ξ) and $\varepsilon[i,\ell]$ is the ℓ-nearest neighbor index for $\boldsymbol{x}_{\pi[i]}$.

Eq. (3.7) provides one practical implementation of the multi-scale basis density functions, which target the modeling of multivariate data. The estimates with (3.7) adopt larger bandwidths in sparse regions while smaller ones in dense regions, and hence realize appropriate modeling particularly for the real datasets exhibiting fat tails. The distributions involving fat tails generate outliers, and the nearest-neighbor distances around these outliers become large. By following the multi-scale nature in setting multiple basis density functions, the local ML method realizes a natural adaptation to incorporate fat tails.

3.3.2 Inconsistency between Clustering and Density Estimation

For high-dimensional data, the exemplar-based convex clustering does not induce sparsity of the mixture weights, even when the correct bandwidths are used. Let $\boldsymbol{\mu}_i \in \mathbb{R}^d$ and σ_i^2 be the true centroid and the bandwidth of the cluster the exemplar $\boldsymbol{x}_{\pi[i]}$ belongs to. When an exemplar $\boldsymbol{x}_j$ belongs to the cluster whose centroid is $\boldsymbol{\mu}_i$, the value of $(\|\boldsymbol{x}_j - \boldsymbol{\mu}_i\|_2^2/\sigma_i^2)$ obeys a chi-square distribution $\chi^2(d)$, whose degrees of freedom is the input dimensionality d. Since the mean and the standard deviation of the chi-square distribution $\chi^2(d)$ are d and $\sqrt{2d}$, respectively, probability mass of the distribution $\chi^2(d)$ is strongly concentrated around its mean value d, particularly when d is high. Then values of the probability density function $\mathcal{N}\left(\boldsymbol{x}_j; \boldsymbol{\mu}_i, \sigma_i^2 \boldsymbol{I}_d\right)$ become close to $(2\pi e \sigma_i^2)^{-d/2}$ for any data point $\boldsymbol{x}_j$.

In contrast, when we assume $\boldsymbol{\mu}_i = \boldsymbol{x}_{\pi[i]}$, the value of the kernel $\kappa_{ii} = (2\pi\sigma_i^2)^{-d/2}$ is $e^{d/2}$ times larger than the true value. Let $\widehat{\boldsymbol{\lambda}} \triangleq (\widehat{\lambda}_1, \ldots, \widehat{\lambda}_m)^\top$ be the converged value of the mixture weight vector provided by Algorithm 1. Because of the extremely large multiplier $e^{d/2}$, most of the elements in the vector $\widehat{\boldsymbol{\lambda}}$ remain positive in order to hold the ability to explain the training data $\mathcal{D}$. Consequently, the ML density estimation, which uses the correct bandwidths, does not produce meaningful clusters.

due to the non-sparsity of the weight vector $\widehat{\boldsymbol{\lambda}}$. In short, the approximation $\boldsymbol{\mu}_i = \boldsymbol{x}_{\pi[i]}$ to justify the exemplar-based clustering is significantly inaccurate when the dimensionality d is high, and we need an alternative approach to apply the convex clustering algorithm to cluster high-dimensional datasets.

We can slightly relax the curse of dimensionality by discounting the value of the kernel κ_{ii} not by strictly assuming $\boldsymbol{\mu}_i = \boldsymbol{x}_{\pi[i]}$ but by regarding $\boldsymbol{x}_{\pi[i]}$ as the closest point to $\boldsymbol{\mu}_i$. Focusing on $\boldsymbol{x}_{\varepsilon[i,1]}$, which is the 1-nearest neighbor of $\boldsymbol{x}_{\pi[i]}$, we calculate the value of the kernel κ_{ij} as

$$\kappa_{ij} = \begin{cases} \dfrac{1}{\sqrt{2\pi\widehat{\sigma}_i^2}^d} \exp\left(-\dfrac{\|\boldsymbol{x}_{\varepsilon[i,1]} - \boldsymbol{x}_{\pi[i]}\|_2^2}{2\widehat{\sigma}_i^2}\right) & \text{if } j = \pi[i] \\ \dfrac{1}{\sqrt{2\pi\widehat{\sigma}_i^2}^d} \exp\left(-\dfrac{\|\boldsymbol{x}_j - \boldsymbol{x}_{\pi[i]}\|_2^2}{2\widehat{\sigma}_i^2}\right) & \text{otherwise,} \end{cases} \tag{3.8}$$

where $\widehat{\sigma}_i^2$ is the local ML estimate introduced in Section 3.3.1. Since the squared 1-nearest neighbor distance $\|\boldsymbol{x}_{\varepsilon(i,1)} - \boldsymbol{x}_{\pi[i]}\|_2^2$ in (3.8) is also influenced by the distribution $\chi^2(d)$, the ratio $e^{d/2}$ between the values of κ_{ii} and κ_{ij} such that $j \neq i$ is cancelled. While various modifications of distance structures to set $\boldsymbol{\mu}_i \neq \boldsymbol{x}_i$ can relax the curse of dimensionality, Eq. (3.8) is one of the simplest manipulations that still guarantee the condition that $\boldsymbol{x}_{\pi[i]}$ is the closest point to $\boldsymbol{\mu}_i$.

Yet heuristic modifications of the distance structures do not fundamentally solve the problem. Even when using (3.8), the differences among the squared distances $\{\|\boldsymbol{x}_j - \boldsymbol{x}_{\pi[i]}\|_2^2\}_{i \neq j}$ largely fluctuate on the scale of $\sqrt{2d}$. Hence the relative ratio among $\{\kappa_{ij}\}_{i=1}^m$ involves multipliers of the order $\exp(\sqrt{d})$, which is highly volatile when d is high.

3.3.3 Using Large Bandwidths as an Annealing Process

One way to directly cancel the effects of the extremely-large multiplier $\exp(\sqrt{d})$ is to multiply some discounting factor to each of the squared distances. Let β $(0 < \beta \leq 1)$ be a relaxation factor for the kernel distributions. Because $\exp(-\beta\|\cdot\|_2^2) \equiv \exp(-\|\cdot\|_2^2)^\beta$, we consider a relaxed optimization

$$\max_{\boldsymbol{\lambda}} \sum_{j=1}^n \log \sum_{i=1}^m \lambda_i \kappa_{ij}^\beta \text{ subject to } \sum_{i=1}^m \lambda_i = 1. \tag{3.9}$$

The optimum of $\boldsymbol{\lambda}$ in (3.9) becomes sparser as we use the lower value of β. Hence applying a low value of β for the optimization (3.9) yields some clustering. At the same time, we should assess the meaning of the fitted result for such relaxation to yield the clustering. Because the probability density function with $\beta \neq 1$ is improper and different from the standard likelihood function, we should never regard the fitted result as a density estimation, but only exploit it for clustering.

We note the similarity of (3.9) with the mixture learning via the Deterministic Annealing (DA) (Rose, 1998; Ueda and Nakano, 1998), that solves an optimization problem

$$\max_{\boldsymbol{\lambda}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m} \sum_{j=1}^n \log \sum_{i=1}^m \lambda_i^\beta p(\boldsymbol{x}_j | \boldsymbol{\theta}_i)^\beta \text{ subject to } \sum_{i=1}^m \lambda_i = 1.$$

In the DA, the hyperparameter β is called the inverse temperature. As β becomes close to zero, the objective function gets more relaxed and the number of local optima becomes smaller. With

introducing a latent probability q_{ij} with which the ith cluster contributes to the generation of the jth exemplar, the DA-EM algorithm iterates update rules

$$\textbf{E-step:} \quad q_{ij} \leftarrow \frac{\lambda_i^\beta p(\boldsymbol{x}_j|\boldsymbol{\theta}_i)^\beta}{\sum_{i=1}^m \lambda_i^\beta p(\boldsymbol{x}_j|\boldsymbol{\theta}_i)^\beta} \text{ for } i \in \{1,\dots,m\} \text{ and } j \in \{1,\dots,n\}$$

$$\textbf{M-step:} \quad \lambda_i \leftarrow \frac{1}{n}\sum_{j=1}^n q_{ij} \text{ and } \boldsymbol{\theta}_i \leftarrow \arg\max_{\boldsymbol{\theta}_i} \sum_{j=1}^n q_{ij} \log p(\boldsymbol{x}_j|\boldsymbol{\theta}_i) \text{ for } i \in \{1,\dots,m\}.$$

The DA-EM algorithm first computes an optimum for a low value of β. Then the acquired optimum becomes the initial parameter in a slightly harder optimization using higher β. The value of β is gradually increased and we finally perform the original optimization with $\beta = 1$. The continuity of the global optimum with respect to β is guaranteed.

While DA-EM algorithm is designed to reduce the number of local optima in the parameter space, our relaxation reduces the number of local modes of the input space. When β is small, for many settings of the mixture weights $\boldsymbol{\lambda}$, the relaxed density $\sum_{i=1}^m \lambda_i p(\boldsymbol{x}|\boldsymbol{\theta}_i)^\beta$ has a less number of local modes (peaks) in the input space $\mathbb{R}^d$. Hence the convex clustering algorithm prunes the irrelevant kernels that do not strongly contribute for the local modes, and relevant kernels are located near the local modes of the relaxed density.

The analogy to link our approach with the DA tells us the limitation that the convex clustering result with low β is not guaranteed to be a good clustering. In DA, the acquired local optimum using $\beta < 1$ is not the final solution of the original objective and needs to *move* in harder optimizations using higher β. Similarly, in the convex clustering, each mixture weight and relevant kernel fitted with low β are not guaranteed to be optimal and relevant in modeling the original input density. The results with low β should be used as initial values, as in the DA-EM algorithm. Figure 3.4 demonstrates a fitting result with a low value of the inverse-temperature β, by projecting a high-dimensional dataset into a unidimensional space. We should emphasize the importance of refitting the cluster parameters in practice.

The difference between the proposed optimization and DA comes from the sparsity requirement. Instead of (3.9), one can consider an optimization

$$\max_{\boldsymbol{\lambda}} \sum_{j=1}^n \log \sum_{i=1}^m \lambda_i^\beta \kappa_{ij}^\beta \text{ subject to } \sum_{i=1}^m \lambda_i = 1. \tag{3.10}$$

Optimization (3.10) is also convex when $0 < \beta \leq 1$, thanks to the concavity of x^β. When $\beta < 1$, the first derivative of (3.10) always becomes infinite at the corner $\lambda_i = 0$, and this infinite derivative makes the optimum of (3.10) always non-sparse. Consequently, using $\beta < 1$ is inappropriate for performing clustering.

3.3.4 Repeating the Convex Clustering Algorithms

In refitting the cluster parameters, we repeat the convex clustering algorithms several times. One can perform such repetitions without updating the kernel distributions, where a high value of β is applied only for further pruning the irrelevant kernels. Yet the relevant kernels found with a low value of the inverse-temperature β are often not close to the true cluster centroids, and hence we refit the centroids and bandwidths as in the EM algorithm.

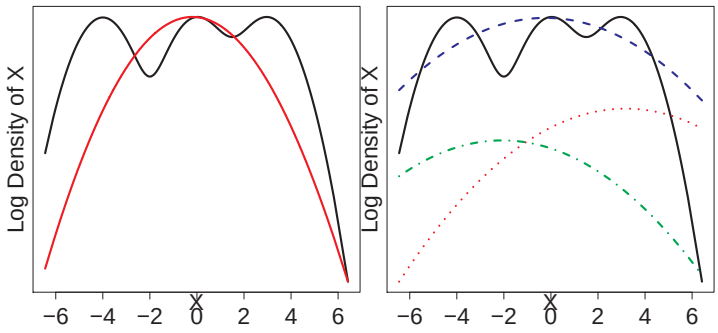


Figure 3.4: Insufficiency of the fitted result using a low value of the inverse-temperature β. $1,000$ samples of 100-dimensional data points were distributed from an equally-sized 3-mixture of isotropic Gaussian distributions, whose centroids are $(-4, 0, \ldots, 0)^\top$, $(0, 0, \ldots, 0)^\top$, and $(3, 0, \ldots, 0)^\top$. Every cluster's covariance matrix is $\boldsymbol{I}_d$. We applied the accelerated convex clustering algorithm for this dataset, by using the bandwidth $\boldsymbol{I}$ and $\beta = 0.13$. The figures show the values of the true or relaxed density functions for samples $\{\boldsymbol{x} = (x, 0, \ldots, 0)^\top\}$. The left figure superimposes the true log-probability density function $\log p(\boldsymbol{x})$ and the fitted relaxed log-density $\log \sum_{i=1}^m \widehat{\lambda}_i p(\boldsymbol{x}|\boldsymbol{\theta}_i)^\beta$ with a rescaling. The right figure shows the relaxed log-density $\log \widehat{\lambda}_i p(\boldsymbol{x}|\boldsymbol{\theta}_i)^\beta$ for each fitted component. While the convex clustering algorithm captures the latent three components, the summed density is not three-modal in the projected subspace, because the largest component dominates the value of relaxed density. In this condition, the three clusters cannot be detected with the Bayes' rule and refitting of both the mixture weights and the kernel centroids are suggested.

Let $u \in \{0, 1, 2, \ldots\}$ be the step number in the refitting process and $\widehat{\sigma}_i^{(0)} \leftarrow \widehat{\sigma}_i$, and $\kappa_{ij}^{(0)} \leftarrow \kappa_{ij}^\beta$. Using Algorithm 1, we first compute the mixture weight as

$$\widehat{\boldsymbol{\lambda}}^{(u)} = \arg\max_{\boldsymbol{\lambda}} \sum_{j=1}^n \log \sum_{i=1}^m \lambda_i \kappa_{ij}^{(u)}. \tag{3.11}$$

Then the latent cluster assignment probability is estimated as

$$q_{ij}^{(u)} \leftarrow \frac{\widehat{\lambda}_i^{(u)} \kappa_{ij}^{(u)}}{\sum_{i=1}^m \widehat{\lambda}_i^{(u)} \kappa_{ij}^{(u)}} \text{ for each } j \in 1, \ldots, n \text{ and } i : \widehat{\lambda}_i^{(u)} > 0. \tag{3.12}$$

For cluster i such that $\lambda_i^{(u)} > 0$, we refit the centroid $\overline{\boldsymbol{x}}_i^{(u)}$ and the bandwidth $\widehat{\sigma}_i^{(u)}$ as

$$\overline{\boldsymbol{x}}_i^{(u)} \leftarrow \frac{\sum_{j=1}^n q_{ij}^{(u)} \boldsymbol{x}_j}{\sum_{j=1}^n q_{ij}^{(u)}} \text{ and } \left(\widehat{\sigma}_i^{(u)}\right)^2 \leftarrow \frac{d\left(\widehat{\sigma}_i^{(u)}\right)^2 + \sum_{j=1}^n q_{ij}^{(u)} \|\boldsymbol{x}_j - \overline{\boldsymbol{x}}_i\|_2^2}{d + d\sum_{j=1}^n q_{ij}^{(u)}}. \tag{3.13}$$

Using $\overline{\boldsymbol{x}}_i^{(u)}$ and the bandwidth $\widehat{\sigma}_i^{(u)}$, a new kernel matrix $\boldsymbol{K}^{(u+1)} = (\kappa_{ij}^{(u+1)})$ is given as

$$\kappa_{ij}^{(u+1)} = \mathcal{N}\left(\boldsymbol{x}_j; \overline{\boldsymbol{x}}_i^{(u)}, (\widehat{\sigma}_i^{(u)})^2 \boldsymbol{I}_d\right). \tag{3.14}$$

We repeat (3.11)-(3.14) in $u = 0, 1, 2, \ldots,$ until no more kernels are pruned. Experimentally, 5 iterations were sufficient because most irrelevant kernels are pruned in the first and second iterations.

Algorithm 2 Convex Clustering Repetitions with Refitting the Cluster Parameters

input Set of training data points $\mathcal{D}=\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$, rough number of clusters ξ, inverse temperature β.

function $\{q_{ij}, \sigma_i^2\} = FastConvexClustering(\mathcal{D}, \xi, \beta)$
  $u \leftarrow 0$, $k \leftarrow int(n/\xi)$
  for $i=1$ **to** m **do**
    Calculate the nearest neighbor indexes $\varepsilon[i,1], \varepsilon[i,2], \ldots, \varepsilon[i,k]$
    $(\widehat{\sigma}_i^{(0)})^2 \leftarrow \frac{1}{kd}\sum_{\ell=1}^{k} \|\boldsymbol{x}_{\varepsilon[i,\ell]} - \boldsymbol{x}_{\pi[i]}\|_2^2$
    for $j=1$ **to** n **do**
      $j' \leftarrow (j=\pi[i])\ ?\ \varepsilon[i,1]\ :\ j$, $\kappa_{ij}^{(0)} \leftarrow \mathcal{N}\left(\boldsymbol{x}_{j'}; \boldsymbol{x}_{\pi[i]}, (\widehat{\sigma}_i^{(0)})^2 \boldsymbol{I}_d\right)$
    end for
  end for
  repeat
    $\widehat{\lambda}^{(u)} \leftarrow FastConvexClustering(\boldsymbol{K}^{(u)})$
    for $i \in \{1, 2, \ldots, m\}$ s.t. $\widehat{\lambda}_i^{(u)} > 0$ **do**
      for $j=1$ **to** n **do** $q_{ij}^{(u)} \leftarrow \dfrac{\widehat{\lambda}_i^{(u)} \kappa_{ij}^{(u)}}{\sum_{i=1}^{m} \widehat{\lambda}_i^{(u)} \kappa_{ij}^{(u)}}$ **end for**
      $\overline{\boldsymbol{x}}_i^{(u)} \leftarrow \dfrac{\sum_{j=1}^{n} q_{ij}^{(u)} \boldsymbol{x}_j}{\sum_{j=1}^{n} q_{ij}^{(u)}}$, $\left(\widehat{\sigma}_i^{(u)}\right)^2 \leftarrow \dfrac{d\left(\widehat{\sigma}_i^{(0)}\right)^2 + \sum_{j=1}^{n} q_{ij}^{(u)} \|\boldsymbol{x}_j - \overline{\boldsymbol{x}}_i\|_2^2}{d + d\sum_{j=1}^{n} q_{ij}^{(u)}}$
      $\kappa_{ij}^{(u+1)} \leftarrow \mathcal{N}\left(\boldsymbol{x}_j; \overline{\boldsymbol{x}}_i^{(u)}, (\widehat{\sigma}_i^{(u)})^2 \boldsymbol{I}_d\right)$
    end for
    $u \leftarrow u+1$
  until $\boldsymbol{\lambda}^{(u)}$ converges
end function

output The converged cluster assignment probability $\{q_{ij}^{(u)}\}$ and cluster bandwidth $\{\widehat{\sigma}_i^{(u)}\}_{i=1}^{m}$.

Algorithm 2 summarizes the iterative refitting procedures. We note that the design of refitting algorithm is not needed to be the maximum-likelihood estimator. While Eq. (3.13) involves a regularization term using $\widehat{\sigma}_i^{(0)} \equiv \widehat{\sigma}_i$, we are able to use more sophisticated Bayesian regularizations including variational-Bayes method. Because we finally apply more precise empirical-Bayes regularization in Section 3.4, here we made the refitting algorithm simple.

3.4 Empirical-Bayes Model Selection

As an example to automatically determine the number of clusters and resulting probability density function, we discuss an automatic selection of the clustering by adopting the maximum marginal-likelihood criteria. Different settings of the rough number of clusters ξ, as well as the inverse-temperature β, result in different clustering results. When the desirable number of clusters is determined *a priori*, then we can simply choose the corresponding result. Otherwise, we often need to choose the appropriate clustering from several candidates. Since the resulting numbers of clusters are different among the initial parameter settings, we need a clustering quality measure even when the numbers of clusters are different among the results.

By following the tradition in Bayesian formalisms, we formulate and solve a maximization of the marginal likelihood, for choosing the optimal bandwidths that provide accurate density estimation. When the true clusters are isotropic Gaussian distributions, then the chosen number of active components becomes close to the true number of clusters. The marginal likelihood function in our Gaussian mixture model is derived in Section 3.4.1. Section 3.4.2 shows a closed-form empirical-Bayes estimate that maximizes the marginal likelihood.

3.4.1 Deriving Approximate Marginal Likelihood

For a training likelihood function with a general Gaussian mixture model

$$p\left(\mathcal{D}|\Theta\right)=\prod_{j=1}^{n}\sum_{i=1}^{m}\phi_i\mathcal{N}\left(\boldsymbol{x}_j;\boldsymbol{\mu}_i,\sigma_i^2\boldsymbol{I}_d\right) \text{ where } \Theta\triangleq\{\boldsymbol{\phi}\triangleq(\phi_1,\ldots,\phi_m)^\top,\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_m,\sigma_1^2,\ldots,\sigma_m^2\},$$

we take its lower bound with Jensen's inequality as

$$p\left(\mathcal{D}|\Theta\right)\geq\prod_{j=1}^{n}\prod_{i=1}^{m}\left[\frac{\phi_i\mathcal{N}\left(\boldsymbol{x}_j;\boldsymbol{\mu}_i,\sigma_i^2\boldsymbol{I}_d\right)}{q_{ij}}\right]^{q_{ij}}\triangleq q\left(\mathcal{D}|\Theta,\{q_{ij}\}\right).$$

We regard $q\left(\mathcal{D}|\Theta,\{q_{ij}\}\right)$ as an approximation of the data likelihood. The values of q_{ij} and σ_i^2 are given as $q_{ij}^{(u)}$ and $\widehat{\sigma}_i^{(u)}$ in Section 3.3.4, respectively. We simply denote the converged values of $q_{ij}^{(u)}$ and $\widehat{\sigma}_i^{(u)}$ by q_{ij} and σ_i^2, respectively. In addition, we denote the converged value of the centroid $\overline{\boldsymbol{x}}_i^{(u)}$ by $\overline{\boldsymbol{x}}_i$.

Let us place an m-dimensional symmetric Dirichlet distribution prior for the mixture weight vector $\boldsymbol{\phi}$, and a heteroscedastic isotropic Gaussian prior for the centroid $\boldsymbol{\mu}_i$ as

$$\begin{aligned}p(\boldsymbol{\phi}|\alpha) &= \frac{\Gamma\left(\alpha\right)}{\Gamma^m(\alpha/m)}\prod_{i=1}^{m}\phi_i^{\alpha/m-1} \text{ and}\\ p(\boldsymbol{\mu}_i|\omega_i) &= \mathcal{N}\left(\boldsymbol{\mu}_i;\boldsymbol{\mu}_0,\omega_i\sigma_i^2\boldsymbol{I}_d\right) \text{ for each } i\in\{1,\ldots m\},\end{aligned}$$

where α is a concentration hyperparameter, $\boldsymbol{\mu}_0=\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{x}_j$ is the global mean among the n data points, and $\{\omega_1,\ldots,\omega_m\}$ is a set of prior bandwidths. The entire prior takes a factorial form $p(\Theta|\Psi)=p(\boldsymbol{\phi}|\alpha)\prod_{i=1}^{m}p(\boldsymbol{\mu}_i|\omega_i)$ where $\Psi=\{\alpha,\omega_1,\ldots,\omega_m\}$.

The approximate data likelihood $q\left(\mathcal{D}|\Theta,\{q_{ij}\}\right)$ is analytically marginalized with the prior $p(\Theta|\Psi)$, because $q\left(\mathcal{D}|\Theta,\{q_{ij}\}\right)$ is a product of exponential families. We obtain closed-forms of the approximate marginal likelihood $q\left(\mathcal{D}|\{q_{ij}\},\Psi\right)\triangleq\int_\Theta q\left(\mathcal{D}|\Theta,\{q_{ij}\}\right)p(\Theta|\Psi)d\Theta$ and the posterior $q\left(\Theta|\mathcal{D},\{q_{ij}\},\Psi\right)\triangleq q\left(\mathcal{D}|\Theta,\{q_{ij}\}\right)p(\Theta|\Psi)/q\left(\mathcal{D}|\{q_{ij}\},\Psi\right)$ as

$$\begin{aligned}q\left(\mathcal{D}|\{q_{ij}\},\Psi\right)=&\left[\prod_{j=1}^{n}\prod_{i=1}^{m}q_{ij}^{-q_{ij}}\right]\frac{\Gamma(\alpha)}{\Gamma(\alpha+n)}\prod_{i=1}^{m}\frac{\Gamma(\alpha/m+\widehat{n}_i)}{\Gamma(\alpha/m)}\cdot\\ &\prod_{i=1}^{m}\left[\left(2\pi\widehat{\sigma}_i^2\right)^{-\frac{d\widehat{n}_i}{2}}\left(1+\widehat{n}_i\omega_i\right)^{-\frac{d}{2}}\exp\left(-\frac{\widehat{n}_i\|\overline{\boldsymbol{x}}_i-\boldsymbol{\mu}_0\|_2^2}{2(1+\widehat{n}_i\omega_i)\widehat{\sigma}_i^2}\right)\right] \text{ and}\end{aligned} \tag{3.15}$$

$$q\left(\Theta|\mathcal{D},\{q_{ij}\},\Psi\right)=\frac{\Gamma\left(\alpha+n\right)}{\prod_{i=1}^{m}\Gamma(\alpha/m+\widehat{n}_i)}\prod_{i=1}^{m}\left[\phi_i^{\widehat{n}_i+\alpha/m-1}\mathcal{N}\left(\boldsymbol{\mu}_i;\frac{\boldsymbol{\mu}_0+\widehat{n}_i\omega_i\overline{\boldsymbol{x}}_i}{1+\widehat{n}_i\omega_i},\frac{\omega_i}{1+\widehat{n}_i\omega_i}\sigma_i^2\right)\right],$$

respectively, where $\widehat{n}_i \triangleq \sum_{j=1}^{n} q_{ij} \equiv n\widehat{\lambda}_i$.

Eq. (3.15) is an unsupervised clustering quality score that enables a comparison among the clustering results having different numbers of clusters. The fitness to the training data points is represented in the term $\prod_{i=1}^{m} \left(2\pi\widehat{\sigma}_i^2\right)^{-\frac{d\widehat{n}_i}{2}}$ that expresses the total variances of the clusters. The term $\frac{\Gamma(\alpha)}{\Gamma(\alpha+n)} \prod_{i=1}^{m} \frac{\Gamma(\alpha/m+\widehat{n}_i)}{\Gamma(\alpha/m)}$ is the marginal likelihood stemming from the mixture weights, and naturally favors a low number of clusters thanks to the Dirichlet prior. For each cluster i, the term $(1+\widehat{n}_i\omega_i)^{-\frac{d}{2}} \exp\left(-\frac{\widehat{n}_i\|\overline{\boldsymbol{x}}_i-\boldsymbol{\mu}_0\|_2^2}{2(1+\widehat{n}_i\omega_i)\widehat{\sigma}_i^2}\right)$ incorporates the prior loss in distributing its centroid from the global mean. When there are many clusters, the penalty related with the Gaussian prior also increases, as well as those related with the Dirichlet prior. For avoiding the numerical underflow, we should take the logarithm of (3.15) and exploit the log-gamma function in implementing the evaluation of the marginal likelihood.

While we are able to evaluate the logarithm of the marginal likelihood (3.15) by fixing the values of α and $\{\omega_i\}_{i=1}^{m}$, optimization of these prior hyperparameters, as we show in next Section 3.4.2. Therefore, the value of the clustering quality for each setting should be measured after the optimization of the prior hyperparameters.

3.4.2 Closed-Form Empirical-Bayes Estimate

Empirical-Bayes method (e.g., (Robbins, 1956)), or type-II maximum-likelihood method, is a class of Bayesian statistical estimation methods that select the prior hyperparameters maximizing the marginal likelihood. For the concentration hyperparameter α, we compute the maximum marginal-likelihood estimate
$\widehat{\alpha} = \max_\alpha \log \frac{\Gamma(\alpha)}{\Gamma(\alpha+n)} \prod_{i=1}^{m} \frac{\Gamma(\alpha/m+\widehat{n}_i)}{\Gamma(\alpha/m)}$ with a fast modified-Newton method (Minka, 2003). Since we prefer a sparse optimum of $\boldsymbol{\phi}$, we adopt an initialization $\alpha=1$, based on the fact that settings such that $\alpha < m$ correspond to the preferences of sparse solutions. While we are able to globally optimize the value of α by starting from several initial values, we confirmed that the optimum of α is unique in many experimental cases. For the prior bandwidth ω_i, the maximum marginal-likelihood estimate is given as $\widehat{\omega}_i = \left(\frac{\|\overline{\boldsymbol{x}}_i-\boldsymbol{\mu}_0\|^2}{d\widehat{\sigma}_i^2} - \frac{1}{\widehat{n}_i}\right)^+$ where $(\cdot)^+ \triangleq \max\{\cdot, 0\}$.

With computing each clustering's best combination of the prior hyperparameters $(\widehat{\alpha}, \widehat{\omega})$, we compare the maximized value of the marginal likelihood associated with each clustering. Then we compute the point estimates of the mixture weights and the centroids, which are associated with the prior hyperparameters providing the best clustering. Since the posterior Dirichlet distribution's mode often becomes sparse while its mean is always dense, we adopt the Maximum A Posteriori (MAP) estimate of the vector $\boldsymbol{\phi}$ for obtaining the sparse estimate. Let n_i^*, $\overline{\boldsymbol{x}}_i^*$, α^*, σ_i^*, and ω_i^* be the values of $\widehat{n}_i$, $\overline{\boldsymbol{x}}_i$, $\widehat{\alpha}$, σ_i, and $\widehat{\omega}_i$ corresponding to the best clustering. The MAP solutions of $\boldsymbol{\phi}$ and $\boldsymbol{\mu}_i$ are given as

$$\phi_i^* = \frac{(\alpha^*/m+n_i^*-1)^+}{\sum_{i'=1}^{m}\left(\alpha^*/m+n_{i'}^*-1\right)^+} \text{ and } \boldsymbol{\mu}_i^* = \frac{\boldsymbol{\mu}_0+\omega_i^* n_i^* \overline{\boldsymbol{x}}_i^*}{1+\omega_i^* n_i^*}.$$

The procedures to evaluate the maximized log marginal likelihood, and the corresponding MAP estimates of the mixture weights and the centroids, are finally summarized in Algorithm 3.

Our model selection provides an appropriate number of clusters only when the true clusters are isotropic Gaussian distributions. When each cluster is a non-Gaussian distribution, relatively many kernels remain active to cover each cluster's non-Gaussianity. We should also note that large noises

Algorithm 3 Empirical-Bayes Model Selection with Maximum Marginal Likelihood

input Set of training data points $\mathcal{D}=\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$, estimates of the bandwidths $\sigma_1^2, \ldots, \sigma_m^2$, and cluster assignment probabilities $\{q_{ij}; j \in \{1, \ldots, n\}, i \in \{1, \ldots, m\}\}$

function $(\widehat{\mathcal{L}}, \mathcal{A}, \{\widehat{\phi}_i, \widehat{\boldsymbol{\mu}}_i; i \in \mathcal{A}\}) = EmpiricalBayes(\mathcal{D}, \sigma_1^2, \ldots, \sigma_m^2, \{q_{ij}\})$

$\boldsymbol{\mu}_0 \leftarrow \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{x}_j$

$\widehat{\mathcal{L}} \leftarrow -\sum_{j=1}^{n}\sum_{i=1}^{m} q_{ij} \log q_{ij}$

for $i = 1$ **to** m **do** $\widehat{n}_i \leftarrow \sum_{j=1}^{n} q_{ij}$ **end for**

for $i \in \{1, 2, \ldots, m\}$ s.t. $\widehat{n}_i > 0$ **do**

$\overline{\boldsymbol{x}}_i \leftarrow \frac{\sum_{j=1}^{n} q_{ij}\boldsymbol{x}_j}{\widehat{n}_i}$, $\widehat{\omega}_i \leftarrow \left(\frac{\|\overline{\boldsymbol{x}}_i - \boldsymbol{\mu}_0\|_2^2}{d\sigma_i^2} - \frac{1}{\widehat{n}_i}\right)^+$

$\widehat{\mathcal{L}} \leftarrow \widehat{\mathcal{L}} - \frac{d\widehat{n}_i}{2}\log\left(2\pi\sigma_i^2\right) - \frac{d}{2}\log\left(1+\widehat{n}_i\widehat{\omega}_i\right) - \frac{\widehat{n}_i\|\overline{\boldsymbol{x}}_i - \boldsymbol{\mu}_0\|_2^2}{2(1+\widehat{n}_i\widehat{\omega}_i)\sigma_i^2}$

end for

$\widehat{\alpha} \leftarrow \max_\alpha \mathcal{L}_{DCM}(\alpha)$ where $\mathcal{L}_{DCM}(\alpha) \triangleq \left[\log\frac{\Gamma(\alpha)}{\Gamma(\alpha+n)} + \sum_{i:\widehat{n}_i>0}\log\frac{\Gamma(\alpha/m+\widehat{n}_i)}{\Gamma(\alpha/m)}\right]$

$\widehat{\mathcal{L}} \leftarrow \widehat{\mathcal{L}} + \mathcal{L}_{DCM}(\widehat{\alpha})$

for i s.t. $\widehat{n}_i > 1 - \frac{\widehat{\alpha}}{m}$ **do**

$\mathcal{A} \leftarrow \{\mathcal{A}, i\}$, $\widehat{\phi}_i \leftarrow \frac{(\widehat{\alpha}/m+\widehat{n}_i-1)^+}{\sum_{i'=1}^{m}(\widehat{\alpha}/m+\widehat{n}_{i'}-1)^+}$, $\widehat{\boldsymbol{\mu}}_i \leftarrow \frac{\boldsymbol{\mu}_0 + \widehat{\omega}_i\widehat{n}_i\overline{\boldsymbol{x}}_i}{1+\widehat{\omega}_i\widehat{n}_i}$

end for

end function

output Log marginal likelihood $\widehat{\mathcal{L}}$, set of active kernels $\mathcal{A}$, and empirical-Bayes MAP estimates $\{\widehat{\phi}_i, \widehat{\boldsymbol{\mu}}_i; i \in \mathcal{A}\}$

might decrease the estimated number of clusters, because separating high-noise clusters are often not beneficial in stabilizing the estimates of the cluster parameters. Even in such more challenging cases, the resulting density estimate is highly accurate in forming the probability density function for test data.

3.5 Experimental Evaluations

Our aims in experiments are three-fold. The first one is the evaluation of the acceleration, by comparing the actual computational costs in the proposed algorithm with those in the original EM algorithms, by using the same kernel parameters. The second one is the evaluation of our repetitious algorithm, by seeing the influences of the initial bandwidths and the temperature settings in predictions. The third one is to quantitatively evaluate the detectability of the hidden clusters, for inferring the practicality of the proposed algorithms for real-world datasets. Section 3.5.1 presents the convergence rates using artificial datasets. Section 3.5.2 examines the sensitivity to the initial bandwidths and temperatures. Sections 3.5.3 shows the detectability of the hidden clusters, using higher-dimensional datasets.

3.5.1 Convergence Rate and Computational Time

We compared the convergence rate and the actual CPU time between the proposed and the original EM algorithms. The CPU time was calibrated in a Debian GNU/Linux x86_64 PC with an Intel® Xeon™ Processor 2.80 GHz and 16 GB main memory. Both the EM and the proposed algorithms were

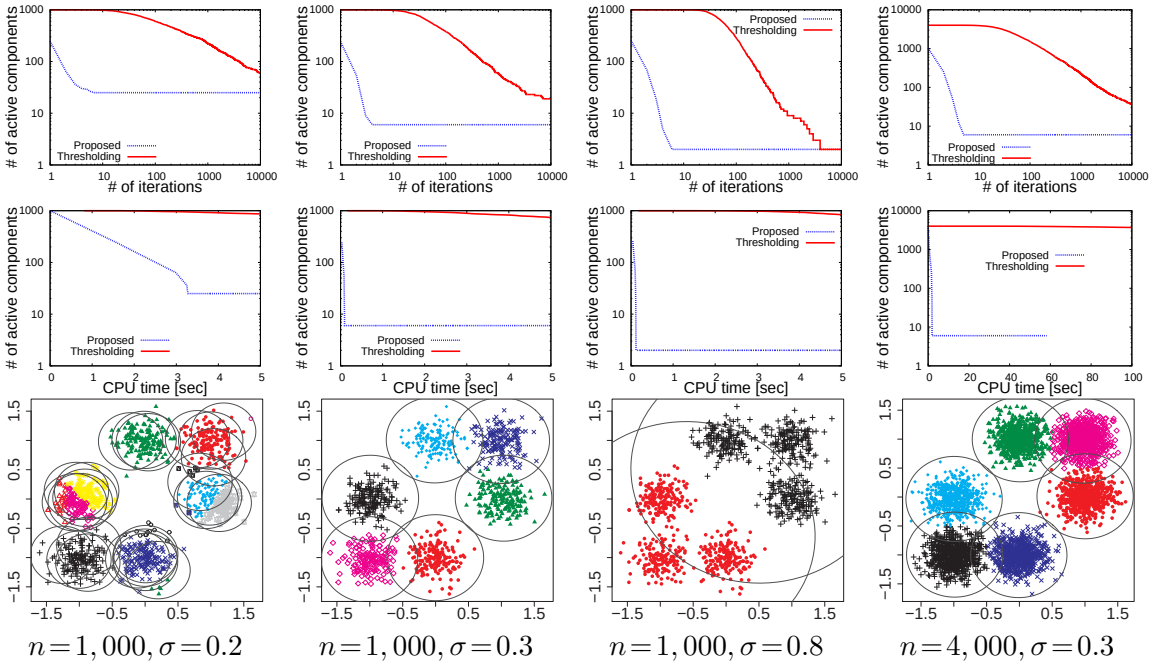


Figure 3.5: Experimental comparisons between the proposed method and the EM algorithm combined with the thresholding rule $\lambda_i \leftarrow 0$ if $\lambda_i < 10^{-3}/n$ (Lashkari and Golland, 2008), for a dataset involving a globally-common single bandwidth. The numbers of active components as the iterations proceed, and the final partitions with the proposed method are shown. The horizontal lines to confirm the convergence rates are the number of iterations (top), or CPU time (middle). We generated 1,000 or 4,000 samples in which the cluster distributions are the same as those in Figure 3.1. While the thresholding rule gradually prunes each component, the proposed method achieves significant reductions of the components in the first 10 steps. Since the centroid of each cluster is constrained to be one of the observed data points and is slightly apart from the true center of the underlying density, a marginally larger bandwidth than the true bandwidth, such as $\sigma = 0.3 > 0.2$, yields a compact grouping.

implemented as Java™ programs using the sparse matrix classes in Apache Commons Mathematics Library version 2.2[4].

First we generated 1,000 or 4,000 data points in $\mathbb{R}^2$ from Gaussian mixture models as shown in Figures 3.5 and 3.6. Figure 3.5 shows the behaviors of the learning algorithms when we adopt the common single bandwidth, while Figure 3.6 shows the results when we adopt the adaptive bandwidths, respectively. For the results with the adaptive bandwidths, we attached several results with the local ML estimation using the k-nearest neighbor method. Here we did not use the iterative refitting algorithm and the empirical-Bayes model selection, because our focus is the evaluation of computational costs.

In both of the single- and adaptive-bandwidth datasets, the combination of the fast pruning and the Newton-Raphson method drastically reduced the required numbers of iterations. In practice, about 10 iterations were sufficient in both settings while hundreds or thousands of iterations were required in the EM algorithm.

Let us complement the information about the CPU time. For easy comparisons, in all of the

[4] `http://commons.apache.org/math/`

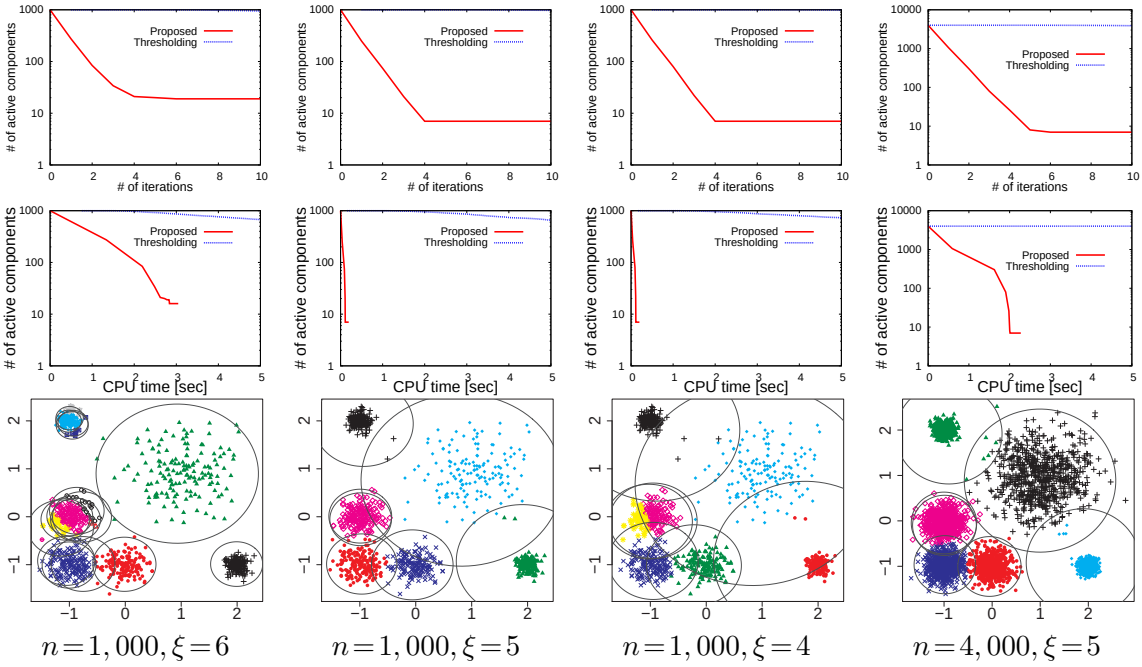


Figure 3.6: Experimental comparisons between the proposed method and the EM algorithm using the same thresholding rule as that in Figure 3.5, when the bandwidths are adaptive. 1,000 or 4,000 samples of $\mathbb{R}^2$ points are distributed from a 6-cluster heteroscedastic Gaussian mixture model. The mixture weight is 1/6 in each cluster. The standard deviation for each of the three clusters centered at $(-1,-1)^\top$, $(-1,0)^\top$, and $(0,-1)^\top$ is 0.2. The standard deviation for each of the two clusters centered at $(-1,2)^\top$ $(2,-1)^\top$ is 0.1. The largest standard deviation assigned to the cluster centered at $(1,1)^\top$ is 0.5. The results in the first 10 iterations are magnified.

results, the times to calculate the kernel matrix and nearest-neighbor indexes are not included in the figures. Computations of the kernel matrices are required in both the single- and adaptive-bandwidth settings. In the single-bandwidth setting, the proposed algorithm additionally needs to calculate the nearest neighbor indexes based on the sorting of the kernel matrix elements. The computational costs to calculate the kernel matrix were 1.07±0.07 seconds when $n = 1,000$, and 15.15 seconds when $n = 4,000$. The additional costs to sort the elements were 1.81±0.05 seconds when $n = 1,000$ and 20.34 seconds when $n = 4,000$. Therefore, even when we consider the CPU costs of additional sorting, the proposed algorithm is evidenced to work sufficiently faster than the EM algorithm. In computing the adaptive bandwidths, since both of the EM and the proposed algorithms need the sorting of the pairwise distances, the initialization costs are the same. The costs to sort the pairwise distances and compute the adaptive bandwidths were 1.60±0.02 seconds when $n = 1,000$ and 29.00 seconds when $n = 4,000$. The costs to calculate the kernel matrix were 0.77±0.02 seconds when $n = 1,000$ and 13.44 seconds when $n = 4,000$.

Next we evaluated the convergences when the dimensionality gets higher. Here we summarize the properties of the higher-dimensional artificial datasets that are also used in the later experiments. 2,000 data points in $\mathbb{R}^d$ were generated from c-cluster heteroscedastic Gaussian mixture models in which all of the mixture weights were $1/c$. The c cluster centroids were distributed from $\mathcal{N}\left(\mathbf{0}, 5^2\boldsymbol{I}_d\right)$ and each cluster variance is the inverses of the values sampled from chi-square distribution whose degrees of freedom is ν, and whose mean is σ_0^2. The scale σ_0 determines the basic noise level of the

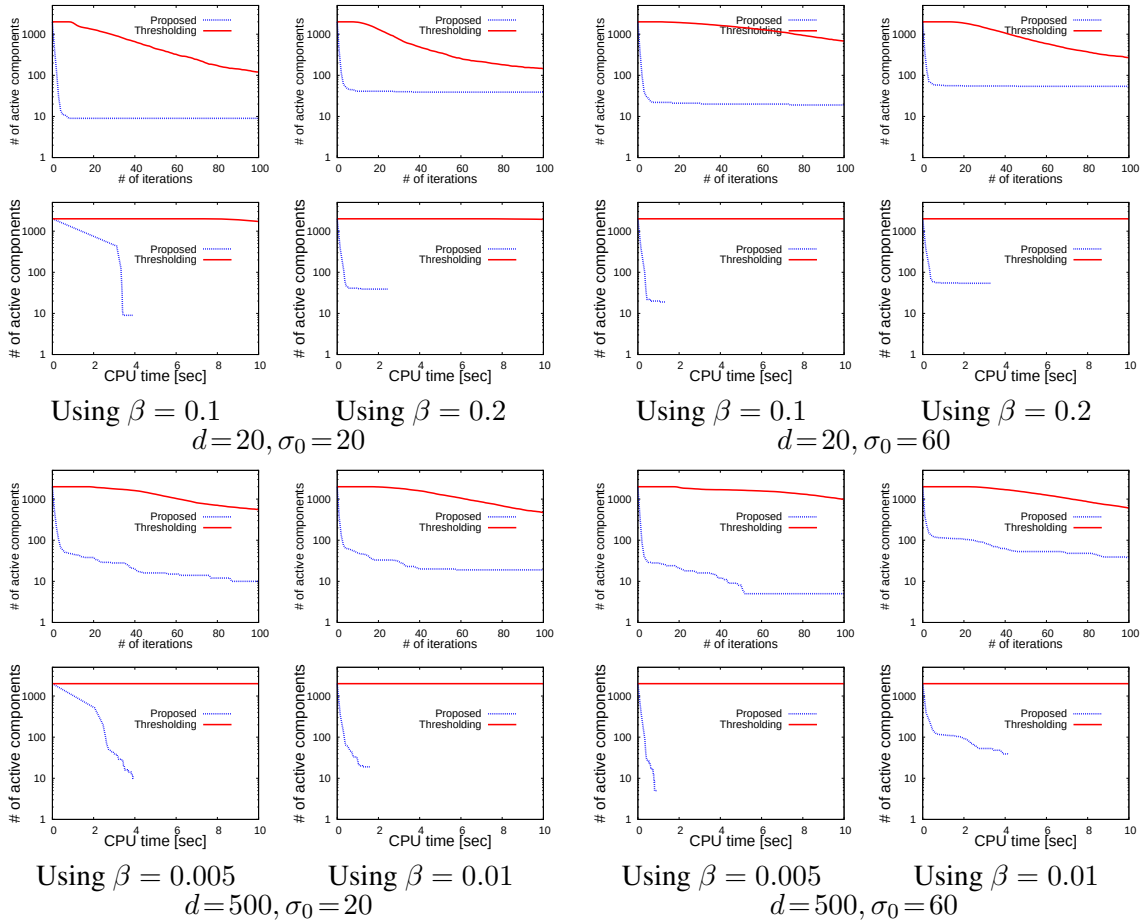


Figure 3.7: Experimental comparisons between the proposed method and the EM algorithm the same thresholding rule as that in Figure 3.5, when datasets are high-dimensional. The number of clusters is $c = 20$ and the degrees of freedom to set the heteroscedasticity is $\nu = 5$. In all of the settings, the proposed algorithm outperformed the EM algorithm. When $d = 500$, the relative large number of iterations needed for the convergence is caused by the difficulty of the pruning judgement, where every kernel vector is similar to each other.

data while ν represents the degrees of heteroscedasticity, where lower values of ν make the datasets more heteroscedastic.

By experimenting using several datasets, we illustrated the calibrated computational costs in the initial convex clustering using large bandwidths in Figure 3.7. The proposed algorithm outperformed the EM algorithm even when the dimensionality of the datasets is high. Note that refitting was not applied in the results of Figure 3.7, while we briefly explain the computational costs of refitting. Since the second convex clustering starts from the condition whose number of clusters is given by the initial convex clustering, the computational costs in the second convex clustering is less than the initial convex clustering. Usually, refitting the cluster parameters is more costly for high-dimensional datasets. The essential computational complexity is $\mathcal{O}(nd|\mathcal{A}_t|)$, where $\mathcal{A}_t$ is the set of active components at tth iteration of the initial clustering. Remember that its order is still the same as that of the standard c-means algorithm when $c = \mathcal{A}_t$.

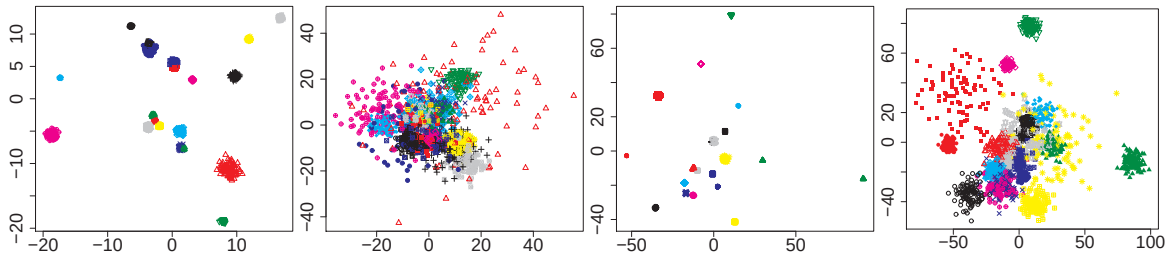


Figure 3.8: Example of the input density in the artificial datasets. Each figure shows a 2-dimensional projection of an artificial high-dimensional dataset, based on a principal component analysis. Each sample's color and symbol represents its belonging cluster. For all of the datasets, the number of clusters is $c = 20$ and the degrees of freedom to set the heteroscedasticity is $\nu = 5$. Though we perform both the proposed and the reference clustering algorithms without dimensionality reduction, the setting $\sigma_0 = 1$ makes the separation of the samples too easy. Hence we should use noisier datasets than in the prior work.

3.5.2 Dependence on the Initial Bandwidths

This experiment investigates how the prediction performances depend on the initial bandwidths, for establishing a rule of thumb to reduce the computational costs in searching for the appropriate hyper-parameters. The artificial high-dimensional datasets, which were introduced in Section 3.5.1, include the information of the true cluster labels. Hence we were able to evaluate the detectability of these hidden cluster labels with unsupervised convex clustering, for various settings of the rough number of clusters ξ and the inverse temperature β. In addition, we studied the characteristics of the hyperparameter selection with the proposed empirical-Bayes method.

The detectability of the hidden clusters is measured as the Normalized Mutual Information (NMI) between the estimated and the true cluster labels. For two disjoint sets $\Psi = \{\psi_1, \ldots, \psi_{|\Psi|}\}$ and $\Omega = \{\omega_1, \ldots, \omega_{|\Omega|}\}$, NMI between Ψ and Ω is given as

$$NMI(\Psi, \Omega) \triangleq \frac{2I(\Psi, \Omega)}{H(\Psi) + H(\Omega)},$$

where

$$I(\Psi, \Omega) = \frac{1}{n} \sum_{a=1}^{|\Psi|} \sum_{b=1}^{|\Omega|} |\psi_a \cap \omega_b| \log \frac{n|\psi_a \cap \omega_b|}{|\psi_a||\omega_b|} \text{ and } H(\Omega) = -\sum_{b=1}^{|\Omega|} \frac{|\omega_b|}{n} \log \frac{|\omega_b|}{n}.$$

Higher NMI indicates a superior predictive performance. The factor $2/[H(\Psi) + H(\Omega)]$ to normalize the entropy enables a comparison of clustering results having different numbers of clusters.

In generating the datasets, we took care of the basic noise level σ_0. The reference value $\sigma_0 = 1$ used in (Lashkari and Golland, 2008) makes the datasets unrealistically clean, as we show in Figure 3.8. Since we think more noisy data should be tested for assessing the predictive capabilities, we adopt larger values of the noise level σ_0.

We performed a grid search for the rough number of clusters ξ and the inverse temperature β. The range of β was determined with the dimension d. Since the inverse temperature β was introduced to relax the multipliers $\exp(d/2)$ and $\exp(\sqrt{d})$ in Section 3.3.3, we set the minimum value of β by $1/d$, where $\beta = 1/d$ makes most of the datasets be clustered into only a single or a few components. We

prepared the candidates of β as $\beta \in \{d^{-1}, d^{-0.9}, \ldots, d^{-0.1}, 1\}$. The rough number of clusters ξ was chosen from $10, 20, \ldots, 100$ and totally 10×10 grid search was performed for each dataset.

As the value of the inverse-temperature β decreases, the influences of the absolute scale of the local bandwidths are weakened, because every element of the kernel matrix has a similar value to each other, while the relative ratio among the local bandwidths is still important. For the high-dimensional datasets, because the squared distance $\|\boldsymbol{x}_j - \boldsymbol{\mu}_i\|_2^2$ becomes close to the mean $d\sigma_i^2$, the local bandwidth estimate $\widehat{\sigma}_i^2$ does not strongly depend on the number of neighbors determined with ξ. We conclude that only the ratios among the values $\{\sigma_i^2\}_{i=1}^m$ are meaningful, with expecting that ξ does not affect to the performances so strongly as β.

Figure 3.9 shows contour plots of the NMI scores depending on the values of ξ and β, where the refitting of the cluster parameters are incorporated. The initial convex clustering was performed with 10 iterations using the SMO. Because we tried the refitting 5 times and 10 iterations were involved in the convex clustering after each refitting, $10 + 5 \times 10 = 60$ iterations were repeated during the entire optimization. As we expected, the results were more sensitive to the inverse temperature β than the rough number of clusters ξ, whereas sometimes ξ affected the performances with specific β.

The empirical-Bayes method gave almost the optimal clustering when the noise level is not so high, while it tended to select more parsimonious models than the best models. When the noise level is middle, the optimal β was around $1/\sqrt{d}$, which is a scale of the squared distances after the discounting in (3.8).

3.5.3 Unsupervised Classification

In this experiment, the detectability of the hidden clusters by the proposed algorithm was compared to those by the reference algorithms. The proposed algorithm consists of the convex clustering, refitting of the cluster parameters, and the empirical-Bayes hyperparameter selection. For each parameter setting of the artificial datasets, 20 independent sets were randomly generated and we evaluated the average performances among the 20 sets. The rough number of clusters was set as $\xi = c$ and the inverse temperature β was chosen from $\{d^{-1}, d^{-0.9}, \ldots, 1\}$ inside the fitting algorithm.

One reference method is the adaptive-bandwidth soft c-means algorithm using the DA, in which an equally-sized Gaussian mixture model $c^{-1}\sum_{i=1}^c \mathcal{N}\left(\boldsymbol{x}; \boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2 \boldsymbol{I}_d\right)$ is fitted. In the DA, the parameters $\left\{\boldsymbol{\mu}_1, \sigma_1^2, \ldots, \boldsymbol{\mu}_c, \sigma_c^2\right\}$ were optimized with regularizing the variances as $\sigma_i^2 \geq 0.01$. In the internal steps in the DA, we maximized a relaxed log-likelihood $\sum_{j=1}^n \log c^{-1}\sum_{i=1}^c \mathcal{N}^\rho\left(\boldsymbol{x}_j; \boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2 \boldsymbol{I}_d\right)$, where the annealing was performed first with $\rho = 0.1$, second with $\rho = 0.5$, and finally with $\rho = 1$. We also implemented another version of the soft c-means algorithm that repeats random initializations 20 times, and that picks the best clustering.

Another reference method is to fit the Dirichlet Process Mixtures (DPM) inferred with a variational method (Blei and Jordan, 2006), where both of the bandwidths and the number of clusters were automatically chosen. In applying the DPM for the datasets, we fit fully-parametrized Gaussian mixture models $\sum_i \lambda_i \mathcal{N}\left(\boldsymbol{x}; \boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2 \boldsymbol{I}_d\right)$, and adopt the posterior mean values of the parameters $\left\{\lambda_i, \boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2\right\}_i$ for the post-processing hard-clustering. In the sampling scheme $G|\alpha, G_0 \sim DP\left(\alpha, G_0\right)$ underlying in each DPM model, we placed a vague hyperprior $\alpha \sim Gamma(1, 1)$. For the mean $\boldsymbol{\mu}$ and the covariance matrix $\sigma^2 \boldsymbol{I}_d$ of each cluster, a product measure $\mathcal{N}\left(\boldsymbol{\mu}; \boldsymbol{\mu}_0, \sigma^2 \boldsymbol{I}_d\right) Gamma(1/\sigma^2; 0.1, 0.1)$ was used as the base distribution G_0. In the automatic choice of the number of clusters, we limited the number of mixtures less than or equal to 150, for reducing the computational costs.

Figure 3.10 shows the comparison results, where the performance stability of the convex clustering is prominent especially in predicting high-dimensional data points. Basically, performances by the

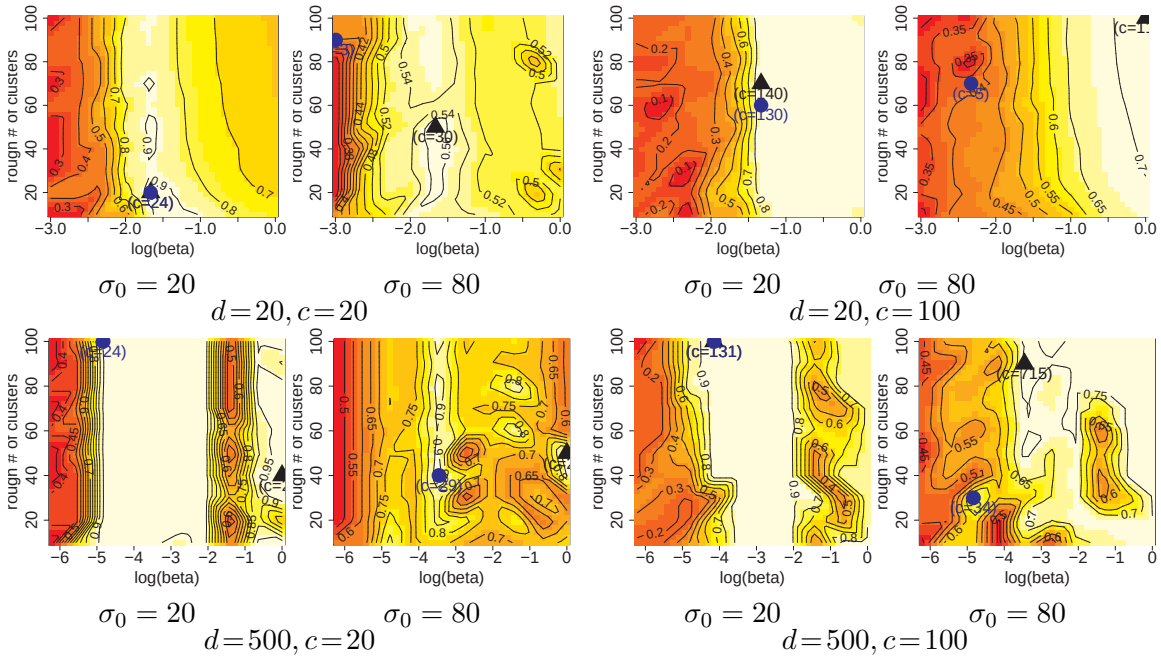


Figure 3.9: Comparisons of the obtained Normalized Mutual Information (NMI) between the true and the estimated cluster labels, among several setting of the prior hyperparameters. The black symbol '▲' is located at the best hyperparameter to most predict the true labels, while the blue symbol '●' is located at the hyperparameter chosen with the unsupervised empirical-Bayes method. The NMI score was more sensitive to the inverse-temperature β than the rough number of clusters ξ, whereas the rough number of clusters was also meaningful in partitioning high-noise data. For middle noise datasets, the maximum marginal-likelihood hyperparameters achieved the highest NMI scores in many cases. For high noise datasets, the empirical-Bayes method preferred parsimonious models having lower numbers of clusters than those in the true densities. When the dimensionality is high, the optimal hyperparameter tended to be located near $\beta = 1/\sqrt{d}$, which is the middle point for each x-axis.

DA or those by many random initializations were highly volatile. While the variances of the DPMs' performances were smaller than those of the DA and random initializations, DPMs struggled to find the true cluster structure in the underlying density, because of its nature to automatically determine the number of clusters. In contrast, the exemplar-based nature of the convex clustering provided robust detections of the underlying structures. While we utilized the true number of clusters as a rough number of clusters ξ, we already confirmed that ξ does not so affect to the performance as to β. The automatic selection with the empirical-Bayes methods were able to recover the true clusters in middle noise conditions. The improvement in the high-noise condition would be the future work.

3.6 Discussion

Let us discuss the applicability of the proposed accelerated algorithms and the possible future work to handle more complex datasets. Adoption of the locally-adaptive bandwidths is a multivariate extension of introducing the multi-scale basis density functions, for precisely handling data involving either of much dense regions, much sparse regions, or outliers. Given the experimental high performance of our algorithms, the combination of the acceleration by SMO and locally-adaptive multi-scale basis

functions is expected to work much beneficially in predicting probability density functions of low- or middle-dimensional variables, including the unidimensional travel time in the next Chapter 4. In contrast, for extremely high-dimensional datasets, the k-nearest neighbor heuristic does not work well due to the curse of dimensionality (e.g., (Radovanović et al., 2010)), and we need to combine effective non-linear dimensionality reduction techniques, such as the random-forest-based features (Botsch and Nossek, 2008; Vens and Costa, 2011). The inconsistency between clustering and density estimation, which is addressed in Section 3.3.2, implies that incorporating the automatic model selection algorithm, such as our empirical-Bayes method, does not completely solve the problem of the curse of dimensionality, because most of the model selection algorithms rely on the generalization capability of the fitted probability density functions. Rather we need another model selection criteria specialized for clustering. Fortunately, solving the curse of dimensionality is not the main issues in the remaining chapters of this dissertation, while it is still beneficial for solving more general machine learning and decision making problems.

Because we treat large-scale datasets from Chapter 4, let us address how to choose the number of initial kernels m. When we have datasets and the output variables are multivariate, we must admit some randomness in the results. We should try to obtain stable results by taking more initial kernels than the true number of clusters. The essence of the convex clustering exists in the process to choose the relevant kernels from the given set of the kernels. While c-means algorithm directly samples the c initial kernels, the convex clustering with $m > c$ can choose the initial kernels more accurately. Hence we recommend to set $c < m \ll n$ when handling huge data. Actually, when n is huge, the probability with which the m subset includes the relevant kernels is high. Fortunately, in the travel-time prediction problem in Chapter 4, this randomness problem does not occur thanks to a customized quantile-based method.

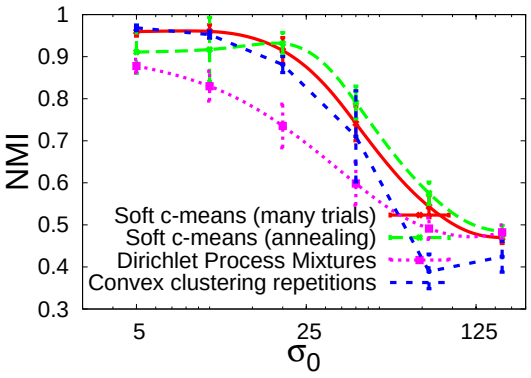
When we focus on clustering problems apart from the main aim in this dissertation, specializing the accelerated and empirical-Bayes density estimation algorithms for detecting non-spherical or time-varying clusters is a considerable direction. This chapter addressed only the case when every cluster is spherical, because our main aim in this dissertation is not the clustering but the acceleration of nonparametric density estimation. Many of the high-dimensional clustering problems, however, need to handle non-spherical clusters that are at least able to be detected by kernel k-means Dhillon et al. (2004) or spectral clustering algorithms (Zelnik-Manor and Perona, 2005). A temporal extension of such clustering problems is the stable fitting of hidden Markov models Beal et al. (2002); Siddiqi (2007); Gael et al. (2008). Since some clusters can have the same emission distributions while have different temporal dynamics Siddiqi (2007), many local optima exist in fitting a hidden Markov model and identifying the true clusters is more challenging than the standard clustering problem. The value of the globally-optimal clustering algorithm would be more prominent in these challenging tasks.

3.7 Summary

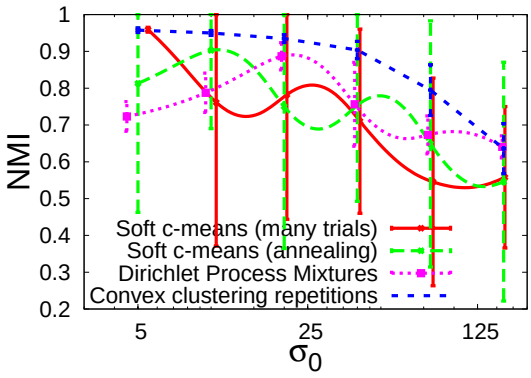
To realize an effective and globally-optimal algorithm to fit nonparametric mixture models, this chapter introduced an accelerated algorithm for the exemplar-based convex clustering, which is an equivalent problem with the nonparametric conditional density estimation. The proposed algorithm consists of the fast pruning of the irrelevant kernels with exact conditions, and the element-wise Newton-Raphson updating for optimizing the mixture weights of the relevant kernels. The algorithm is further accelerated when we incorporate locally-adaptive bandwidths, which provide a multivariate extension of the multi-scale basis density functions and which improve the predictive capabilities for data involving either of much dense regions, much sparse regions, or outliers. In addition to the acceleration

as the basis enhancement, extensions for handling high-dimensional data and enabling an empirical-Bayes model selection have been discussed in the latter part of this chapter. The essence to handle the high-dimensional data is the repetition of the convex clustering algorithms, which consist of gradual decreases of the bandwidth parameters. Unlike the existing maximum-likelihood or Bayesian clustering algorithms, the proposed algorithm does not depend on random initializations and yields stable estimates that are the requirements in real applications.

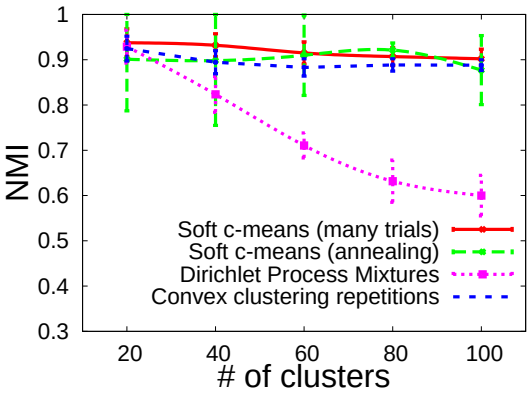
While this chapter has been spent for the basic algorithmic enhancement using only the Gaussian-distributed artificial datasets, in the next Chapter 4, we apply the accelerated convex clustering algorithm for fitting many instances of travel-time distributions, which do not obey Gaussian distributions in real vehicle traffic. Unlike the simple k-nearest neighbor methods used in this chapter, a custom design of the basis density functions involving multiple characteristic scales plays the essential role in predicting the fat tails, that exist in the real traffic data.



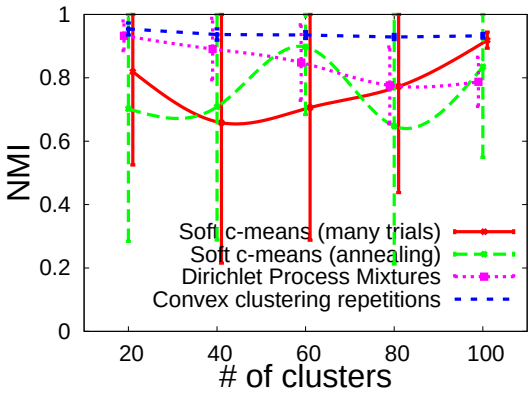
$(c, \nu) = (50, 5)$ and varying σ_0



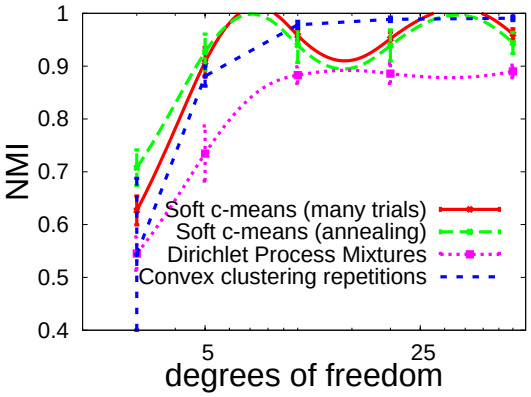
$(c, \nu) = (50, 5)$ and varying σ_0



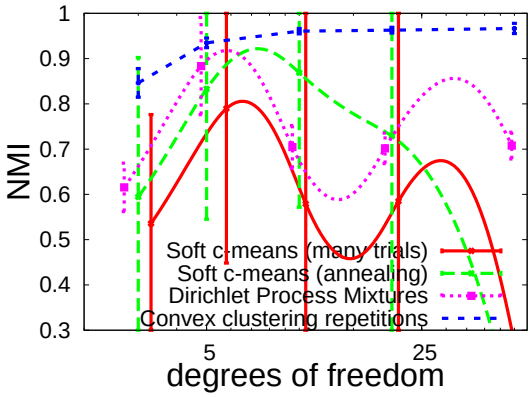
$(\sigma_0, \nu) = (20, 5)$ and varying c



$(\sigma_0, c) = (20, 50)$ and varying ν



$d = 20$



$d = 500$

Figure 3.10: Comparisons of the detectability of the hidden clusters in the higher-dimensional datasets. Compared to other c-means type algorithms, the convex clustering with refitting exhibited stable performances though the average performances were slightly inferior to the soft c-means, when $d = 20$. The main reason of the bad performance in $d = 20$ and $\sigma_0 > 80$ was the empirical-Bayes model selection, where parsimonious models were chosen to regularize the fitted density. Yet in other settings, the automatic selection with the empirical-Bayes methods achieved higher scores in detecting the hidden clusters. In the high-dimensional data ($d = 500$), the proposed algorithm was prominently stable and almost outperformed all of the other algorithms in all of the settings.

Chapter 4

Nonparametric Vehicle-Traffic Prediction

Vehicle traffic is a complex phenomena caused by lots of interacting vehicle drivers through a road network, and risks about the travel time for a specific route are the key factors for accurate prediction of entire traffic and rational decision-making in route choice. When we have a probability density function of travel-time for every link of the road network, optimal routes with respect to a wide range of risk measures are found with stochastic optimization involving dynamic programming (e.g., (Miller-Hooks and Mahmassani, 2000; Osogami, 2011)). As we already discussed in Section 2.2, many of the travel-time prediction algorithms provide only the expectation by following the Ordinary Least Squares principle, and cannot adequately model the actual distributions that possess fat tails stemming from a small portion of extreme delays in traffic jams. We realize more accurate modeling of such fat-tail distributions by using the accelerated nonparametric density estimators, whose algorithmic details have been introduced in Chapter 3. Because we need to fit the distributions for all of the links, to efficiently interpolate among the lots of the links is a key requirement in solving our specific problem. With regarding the membership to each link as an explanatory variable for predicting the distribution, we perform a conditional density estimation of travel time, whose training data consist of the graph structure of the road network and travel-time samples contained probe-car datasets acquired with Global Positioning Systems (GPS).

As well as the incorporation of the fat tails, the proposed nonparametric density estimator emulates the positive feedback among the drivers through the road network. To represent the geometrical positive feedback that yields a similarity of distributions for close links, we design a nonparametric density estimator called the spatial estimator, which interpolates certain probability density functions among multiple links. These probability density functions to be interpolated are called the basis density functions, which are assigned for parts of the links having several or more samples of travel time. By using the accelerated convex clustering algorithm, we fit the basis density functions as mult-scale mixtures of gamma or log-normal distributions, for yielding the fat tails. The interpolation among the basis density functions improves the predictive accuracy particularly for the links having a limited number of travel-time samples. The weights in the interpolation are provided by a combination of similarity metric between any two links and absolute importance of each link. To emulate the propagation of traffic in the road-network graph, we exploit a diffusion kernel (Kondor and Lafferty, 2002) to set the similarity for pairs of links that are not directly connected. As an additional idea to realize the computational feasibility for the real road network having the millions of links, we sparsify the diffusion kernel by approximating the matrix exponential as a power of the matrix. The absolute importance of every link is optimized also with the accelerated convex clustering algorithm, and hence the final estimate adopts a nested structure of nonparametric mixture models, where the nonparametric gamma or

log-normal mixtures are further mixed with another nonparametrics provided by the sparse diffusion kernel. The fast convex optimization and the sparse matrix multiplication make the nonparametric density estimation applicable for large probe-car datasets.

Considering the dependence of the traffic on time-zones, we give the final estimate by further interpolating multiple spatial estimators to one another in the time domain, where one spatial estimator is applied for the data of one specific timezone. Experimental results using real probe-car datasets show advantages of the new nonparametric estimator, over parametric regression methods in the spatial domain, and clarify benefits of the temporal interpolation.

Parts of this work were first published in (Takahashi et al., 2012) in the proceedings of the 12th SIAM International Conference on Data Mining (SDM 2012), which is published by the Society of Industrial and Applied Mathematics (SIAM). We borrowed the formulations, figures, and implications from this publication, based on the permission for authors[1].

Section 4.1 addresses the characteristics of our datasets and introduces the problem of density estimation with a sparse nonparametric method. Sections from 4.2 to 4.4 mainly describe about the spatial estimators, where Section 4.2 introduces the fitting of the basis density functions, Section 4.3 handles the approximation of the diffusion kernel on a link connectivity graph, and Section 4.4 shows the final optimization of the importance weights. The temporal interpolation of the multiple spatial estimators is introduced in Section 4.5. Section 4.6 experimentally shows our method's significant advantages of the predictive accuracy over several existing regression methods. Section 4.7 discusses the possible extensions or applications of the travel-time density estimation for realistic traffic simulation. Section 4.8 summarizes this chapter.

4.1 Road Network and Travel-Time Samples

This section describes the characteristics of our datasets, defines the density estimation problem to be solved, and introduces the model of travel-time distributions. Section 4.1.1 explains how we prepared the datasets from the original probe-car trajectories measured with GPS. For the density estimation problem defined in Section 4.1.2, a nonparametric model of the relative travel-time distribution is introduced in Section 4.1.3.

4.1.1 Properties of our Real Traffic Datasets

Our road network consists of the major roads of the Greater Tokyo Area in Japan. The road network is represented as a digraph $G = (V, E)$ whose node $v \in V$ represents an intersection and whose edge $(u, v) = e \in E$ represents a link from an intersection u to another intersection v. The numbers of intersections and links are $|V| = 1,183,358$ and $|E| = 3,290,523$, respectively. For every link $e \in E$, we defined a standard travel-time $\tau_e^{(0)}$ by dividing its length by its legal speed limit. The standard travel-time is not the actual observation but a reference value in measuring the relative magnitude of the actual travel-time.

The original probe-car dataset includes 58,584 trajectories of taxis, where each trajectory is given as a sequence of two-dimensional GPS coordinates whose sampling intervals were 30 seconds. For each sequence of the GPS coordinates, we applied a map matching algorithm based on (Newson and Krumm, 2009), for obtaining a sequence of links and travel-time spent at each link. The total number of travel-time samples is $3,144,669$, and the number of links having at least one travel-time sample

[1] See the SIAM consent to publish `https://www.siam.org/students/siuro/consent.pdf`. Copyright © by SIAM. Unauthorized reproduction of this chapter is prohibited.

Table 4.1: The numbers, N, of travel-time samples, and the numbers, $|E_+|$, of links that have at least one sample for each timezone.

HOUR	N	$\lvert E_+\rvert$	HOUR	N	$\lvert E_+\rvert$	HOUR	N	$\lvert E_+\rvert$
0:00-	273,168	69,126	8:00-	149,906	47,400	16:00-	134,056	37,794
1:00-	185,567	53,018	9:00-	154,597	47,067	17:00-	174,748	43,074
2:00-	109,662	38,994	10:00-	131,383	42,445	18:00-	196,978	45,676
3:00-	49,821	25,620	11:00-	111,664	37,080	19:00-	162,816	41,468
4:00-	22,501	15,484	12:00-	129,148	41,569	20:00-	149,438	42,592
5:00-	24,433	16,189	13:00-	133,987	40,083	21:00-	169,125	47,856
6:00-	23,868	16,579	14:00-	128,288	37,594	22:00-	169,956	49,328
7:00-	62,753	30,025	15:00-	130,971	36,980	23:00-	165,835	47,297

is $187,872$. Every link-dependent travel-time observation was classified into one of the 24 hourly datasets indexed as $h \in \{0, 1, \ldots, 23\}$, where the dataset $h = 0$ contains samples from 12 a.m. to 1 a.m., the dataset $h = 1$ contains those from 1 a.m. to 2 a.m., and so on.

We perform the spatial estimation independently for every of the 24 datasets. For easier understanding of the spatial estimation algorithms, we omit the notation of the index h, until we introduce the temporal interpolation in Section 4.5. Remember that every variable related with the travel-time samples actually depends on the index h, even when it is omitted.

Here are our notations to describe the training data. An edge $e \in E$ is assigned $n[e](\geq 0)$ travel-time samples $\tau_e^{(1)}, \tau_e^{(2)}, \ldots, \tau_e^{(n[e])}$ [2]. We define a set of relative travel times $\mathcal{Y}[e] \triangleq \{y_e^{(1)}, \ldots, y_e^{(n[e])}\}$, where $y_e^{(j)} \triangleq \tau_e^{(j)}/\tau_e^{(0)}$. $E_+ \triangleq \{e; n[e] \geq 1\}$ is the set of edges having at least one travel-time sample, and $N \triangleq \sum_{e \in E} n[e]$ is the sum of the numbers of travel-time samples.

Table 4.1 shows the sizes of our datasets for each timezone. We have relatively many samples at midnight because people who miss the last trains tend to take taxis. Another peak at 6 p.m. is partly caused by people who finish work.

Figure 4.1 shows the total number of travel-time samples observed in 24 hours for each link. While we have a total of $3,144,669$ travel-time samples, many links in suburban or rural regions lack observations of their travel times. Hence efficient methods to interpolate the probability density functions are crucial in the accurate fitting of many distributions.

4.1.2 Conditional Density Estimation of the Relative Travel Time

Here is the problem to be solved. Let Y_e be a random variable to represent the relative travel time spent by a car driving on a link $e \in E$. Our aim is to model and estimate the probability density function of Y_e as $f_e(y)$ for every link $e \in E$, given the set of training samples $\mathcal{D} = \{\mathcal{Y}[e]; e \in E_+\}$. After we fit the function $f_e(y)$, the distribution of actual travel time $X_e \triangleq \tau_e^{(0)} Y_e$ is easily computed by rescaling the density function $f_e(y)$. The main reason to focus on Y_e instead of X_e is the accuracy in interpolation. Because even the connected links have different values of the length and the speed limit, it is natural to assume that similar links have similar distributions of not the absolute travel-time but the driving velocity.

Before proceeding to mathematical modeling, we should study examples of histograms of the relative travel-time. As in Figure 4.2, we confirmed that the modes of the density function $f_e(y)$ are usually between 0 and 2. Some links have widely variable distributions depending on timezones.

[2] $\tau_{h,e}^{(1)}, \tau_{h,e}^{(2)}, \ldots, \tau_{h,e}^{(n[h,e])}$ is the precise notation when the index h is not omitted.

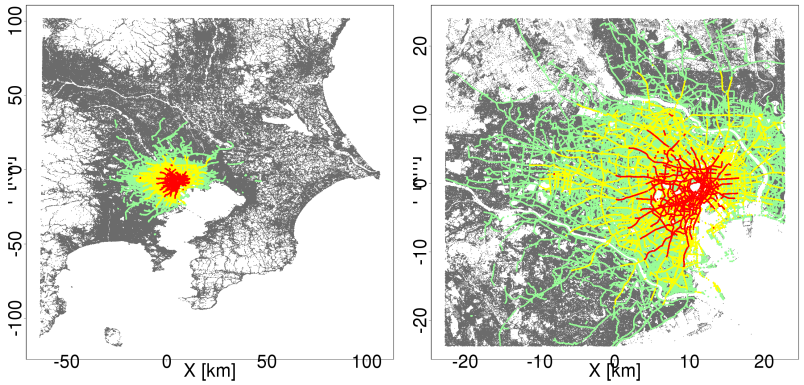


Figure 4.1: Heatmaps based on the total number of travel-time samples in 24 hours for each link. The green, yellow or red points are located on the links that have at least 1, 10, or 100 travel-time samples, respectively. The left figure shows the entire map while the right figure shows a magnified view for the center of Tokyo.

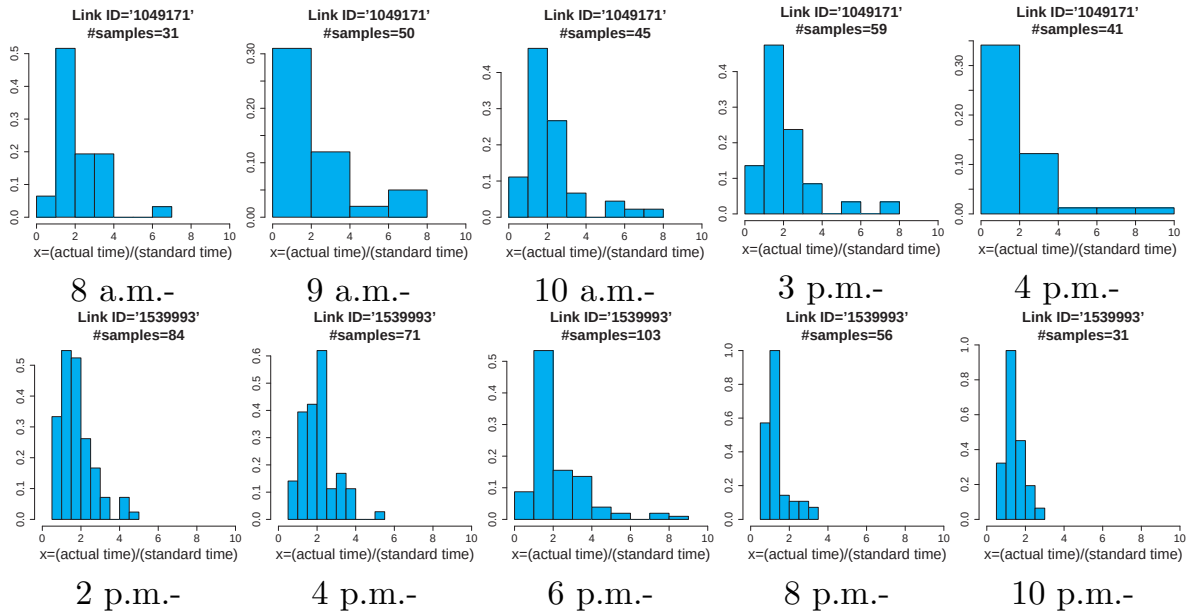


Figure 4.2: Examples of the histograms about the relative travel-time, where the number of bins was automatically determined using Scott's rule (Scott, 1992), and every vertical axis represents the ratio between the respective bin's and the total numbers of samples. For some links that have sufficient samples of the travel time, we are able to roughly infer the shapes of probability density functions without interpolation methods. The modes of the distributions are usually from 0 to 2. The distributions assigned for the link "1049171" do not widely vary among the timezones. In contrast, those assigned for the link "1539993" are more variable depending on the timezones, where the distribution around 6 p.m. has a fat tail.

4.1.3 The Nonparametric Formalism for the Conditional Densities

Let us index all of the edges by $e_1, e_2, \ldots, e_{|E|}$. Let $\pi[1], \ldots, \pi[m]$ be m indices of edges that are associated with basis density functions, and let us define a set of edges $E_\Phi \triangleq \{e_{\pi[1]}, \ldots, e_{\pi[m]}\}$. We

model the distribution of the relative travel-time with a nonparametric form

$$f_e(y) = \frac{\lambda_0\varphi_0(y)+\sum_{i=1}^{m}\lambda_i K_E(e,e_{\pi[i]})\varphi_i(y)}{\lambda_0+\sum_{i=1}^{m}\lambda_i K_E(e,e_{\pi[i]})}, \tag{4.1}$$

where $\Phi \triangleq \{\varphi_0,\varphi_1,\ldots,\varphi_m\}$ is a set of basis density functions, $K_E(e,e_{\pi[i]})$ is a similarity function between the edges e and $e_{\pi[i]}$, $\boldsymbol{\lambda} \triangleq (\lambda_0,\lambda_1,\ldots,\lambda_m)^\top$ is a vector of link importance, and $(\cdot)^\top$ denotes the transpose. A basis density function $\varphi_i(\cdot)$ needs to satisfy $\int_0^\infty \varphi_i(y)dy \equiv 1$, and the link similarity function $K_E(\cdot,\cdot)$ must be non-negative. Eq. (4.1) resembles the nonparametric Nadaraya-Watson kernel regression (Nadaraya, 1964; Watson, 1964) except that we add link-independent weight and basis density function, λ_0 and $\varphi_0(\cdot)$. The terms λ_0 and $\varphi_0(\cdot)$ are introduced for defining $f_e(y)$ even when $\forall i \in \{1,\ldots,m\}, K_E(e,e_{\pi[i]}) \equiv 0$.

The following three sections introduce our approaches to setting or estimating the parameters $(\Phi, E_\Phi, \boldsymbol{\lambda})$. The formulations of the basis density functions and their estimates are discussed in Section 4.2. Then Section 4.3 introduces our method to compute the kernel function $K_E(\cdot,\cdot)$ for each pair of links. Finally, Section 4.4 describes the algorithm to optimize the vector $\boldsymbol{\lambda}$ given Φ and E_Φ. As we show in Section 4.4 and thanks to the property of the convex clustering formulations, the optimum of the vector $\boldsymbol{\lambda}$ becomes sparse.

4.2 Fitting of the Basis Density Functions

The link-dependent basis density functions are fitted for the links that have at least two travel-time samples. Based on the number of available travel-time samples, let us define sets of links $E_1 \triangleq \{e; n[e]=1\}$ and $E_{2+} \triangleq \{e; n[e]\geq 2\}$. Because the link-dependent basis density functions $\varphi_1,\ldots,\varphi_m$ should be fitted for the links having multiple travel time samples, we set $E_\Phi = E_{2+}$ and $m = |E_{2+}|$. Since the link-independent basis function φ_0 is crucial for the links in E_1, we fit the function φ_0 using the samples assigned for links in E_1. The set of relative travel-time samples for fitting the function φ_i is denoted by $\mathcal{Y}_i$, where

$$\mathcal{Y}_i = \begin{cases} \bigcup_{e\in E_1}\mathcal{Y}[e] & \text{if } i=0 \\ \mathcal{Y}[e_{\pi[i]}] & \text{otherwise} \end{cases}.$$

Because the value of the relative travel-time is non-negative, Section 4.2.1 introduces parametric gamma and log-normal distributions to model the basis density functions. Despite their simplicity and stability in the fitting, we show that parametric basis density functions are not sufficiently expressive to model the travel-time distributions. For more flexible modeling, we introduce nonparametric mixtures of gamma or log-normal distributions whose bandwidths involve multiple scales. The global optimality in fitting such nonparametric mixtures is guaranteed with our convex clustering formalism. Details in fitting the nonparametric basis density functions are described in Section 4.2.2.

4.2.1 Parametric Density Functions

Let us first consider instances of probability density functions, which call the parametric density functions and for which we easily obtain the global optimum of the parameters. Since the maximum likelihood estimation of an exponential family is done with convex optimization, we first consider a usage of gamma distributions, which belong to the exponential family. Let $Gam(\cdot;\alpha,\mu)$ be the probability density function of a gamma distribution whose shape parameter is α and whose mean is μ.

The basis density function φ_i for $i \in \{0, 1, \ldots, m\}$ is given as

$$\varphi_i(y) = Gam(y; \alpha_i, \mu_i) \triangleq \frac{\alpha_i^{\alpha_i} y^{\alpha_i - 1} \exp(-\alpha_i \frac{y}{\mu_i})}{\Gamma(\alpha_i)\mu_i^{\alpha_i}}.$$

The maximum likelihood estimate of the basis density function φ_i is given by a convex optimization based on the sufficient statistics

$$\mu_i = \frac{1}{|\mathcal{Y}_i|} \sum_{y \in \mathcal{Y}_i} y \quad \text{and} \quad \nu_i = \frac{1}{|\mathcal{Y}_i|} \sum_{y \in \mathcal{Y}_i} \log y. \tag{4.2}$$

While the value of μ_i is simply given as the empirical average, we obtain the shape parameter α_i by solving a non-linear unidimensional equation

$$\log(\alpha_i) - \Psi(\alpha_i) = \log \mu_i - \nu_i, \tag{4.3}$$

where $\Psi(\cdot)$ is the digamma function such that $\Psi(x) \triangleq \frac{\partial \log \Gamma(x)}{\partial x}$. The solution of (4.3) is easily obtained by Newton-Raphson methods.

Second we consider a log-normal distribution as another model of the parametric basis density function. Let $\mathcal{LN}(; \nu, \sigma^2)$ be the probability density function of a log-normal distribution where

$$\varphi_i(y) = \mathcal{LN}(y; \nu_i, \sigma_i^2) \triangleq \frac{1}{y\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(\log y - \nu_i)^2}{2\sigma_i^2}\right).$$

The location parameter ν_i is given with (4.2), while the scale parameter σ_i^2 is given as

$$\sigma_i^2 = \frac{1}{|\mathcal{Y}_i| - 1} \sum_{y \in \mathcal{Y}_i} (\log y - \nu_i)^2. \tag{4.4}$$

Eq. (4.4) is an unbiased estimator of the variance for the logarithm of the relative travel-time.

4.2.2 Nonparametric Density Functions

To more flexibly model the basis density functions than the aforementioned parametric approaches, we introduce a nonparametric form

$$\varphi_i(y) = \sum_{\ell=1}^{L} \theta_{i\ell} \psi_\ell(y), \tag{4.5}$$

where each basis function φ_i is a linear interpolation of common (link-independent) fundamental density functions $\psi_1, \ldots, \psi_L$. For the mixture weights, we define a vector $\boldsymbol{\theta}_i \triangleq (\theta_{i1}, \ldots, \theta_{iL})^\top$. While one could adopt link-dependent fundamental density functions possibly denoted by $\psi_{i1}, \ldots, \psi_{iL}$, the common fundamental density functions in Equation (4.5) can stabilize and accelerate the fitting. As another merit, the common fundamental density functions make the predictive distribution in (4.1) at most L mixtures of parametric distributions, which are more compact than those computed with the link-dependent fundamental functions.

We stably fit gamma or log-normal distributions for the fundamental density functions $\psi_1, \ldots, \psi_L$, using a quantile method, which corresponds to a nearest-neighbor method for uni-dimensional samples. Let us sort all of the N travel-time samples in $\{y; y \in \mathcal{Y}[e], e \in E_+\}$ in the ascending order,

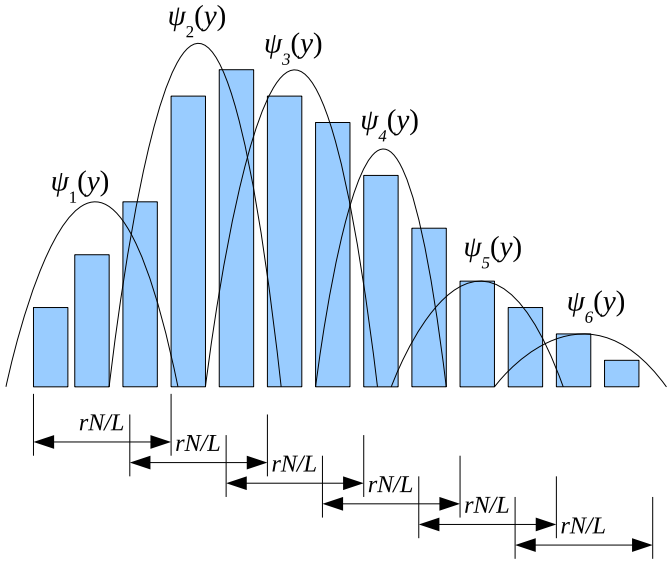


Figure 4.3: Fitting the fundamental density functions $\psi_1, \ldots, \psi_L$ ($L = 6$) for N marginal samples of the relative travel-time. Given a bandwidth hyperparameter r, we prepare a set of sliding window whose length $c_{r,N,L}$ is the rounded integer of $\frac{rN}{L}$. For each of the L windows located at 0%-, $\frac{100}{L-1}$%-, ..., $\frac{100(L-2)}{L-1}$%-, and 100%-tile of the $(N - c_{r,N,L} + 1)$ windows, we separately fit gamma or log-normal distributions of the relative travel-time. Because the resulting fundamental density functions possess larger variances in the tails of the mixture distribution, this quantile-based method is a practical implementation of the multi-scale nonparametric mixtures.

as $y_{s[1]}, y_{s[2]}, \ldots, y_{s[N]}$. Given a bandwidth hyperparameter r, let $c_{r,N,L}$ be the rounded integer of $\frac{rN}{L}$. We obtain $(N - c_{r,N,L} + 1)$ subsets of the samples $\mathcal{Y}_{s[t]} = \{y_{s[t]}, y_{s[t+1]}, \ldots, y_{s[t+c_{r,N,L}-1]}\}$ for $t \in \{1, 2, \ldots, (N - c_{r,N,L} + 1)\}$. Then the training set $\mathcal{Y}_\ell$ to fit the function ψ_ℓ is chosen from $\{\mathcal{Y}_{s[t]}\}$, based on the equal-interval splitting of the index $s[1], \ldots, s[c_{r,N,L}]$. With the same manner adopted in(4.2), (4.3) and (4.4), we compute the values

$$\begin{aligned}
\mu_\ell &= \frac{1}{|\mathcal{Y}_\ell|} \sum_{y \in \mathcal{Y}_\ell} y, \nu_\ell = \frac{1}{|\mathcal{Y}_\ell|} \sum_{y \in \mathcal{Y}_\ell} \log y, \\
\alpha_\ell &= \arg_\alpha \left[\log(\alpha) - \Psi(\alpha) = \log \mu_\ell - \nu_\ell \right], \text{ and} \\
\sigma_\ell^2 &= \frac{1}{|\mathcal{Y}_\ell| - 1} \sum_{y \in \mathcal{Y}_\ell} (\log y - \nu_\ell)^2 .
\end{aligned}$$

The values (α_ℓ, μ_ℓ) are the parameters of the density function when $\psi_\ell(y) \triangleq Gam(y; \alpha_\ell, \mu_\ell)$, and the values $(\nu_\ell, \sigma_\ell^2)$ are those when $\psi_\ell(y) \triangleq \mathcal{LN}(y; \nu_\ell, \sigma_\ell^2)$. Figure 4.3 illustrates the quantile-based estimation of the fundamental density functions $\psi_1, \ldots, \psi_L$.

Let us comment on why we adopted the equal-interval splitting in fitting fundamental density functions. Existing nonparametric methods that form the probability density function by using all of the total N samples of travel-time are inapplicable for our problem due to its high computational costs. One considerable way to reduce the number of basis functions without losing the high expressive powers of the nonparametric approaches is to randomly choose L ($\ll N$) instances of basis density functions. The quantile method is a deterministic alternative to such random choice, where the

influence of random-sampling noises is removed. Because mutually-exclusive choice of the training samples for each fundamental density function results in over-fitting, we improve the generalization capability by introducing the bandwidth-scaling hyperparameter $r > 1$. Consequently, the class of functions expressed by a linear combination of the selected fundamental density functions is still broad. In the later experiments, we set $L = 100$ that achieves both high explanation capabilities and reasonable computational costs in optimization.

After the fitting of the fundamental density functions, for each $i \in \{0, 1, \ldots, m\}$, we obtain the maximum-likelihood estimate of the vector $\boldsymbol{\theta}_i$, by solving an optimization problem

$$\max_{\boldsymbol{\theta}_i} \sum_{y \in \mathcal{Y}_i} \log \left[\sum_{\ell=1}^{L} \theta_{i\ell} \psi_\ell(y) \right]. \tag{4.6}$$

As we already discussed in Chapter 3, Optimization (4.6) is a convex clustering problem for which we are able to attain the sparse and global optimum of the vector $\boldsymbol{\theta}_i$. We solve Optimization (4.6) by using the accelerated convex clustering algorithm also introduced in Chapter 3. Remember that the accelerated algorithm requires a metric to evaluate the similarity between two of the fundamental density functions $\psi_1, \ldots, \psi_L$. Based on the fact that the output variable y is unidimensional in our problem, we measure the similarity based on the first moment of each fundamental density function.

Since we can compute the global optima in fitting both the fundamental density functions and the mixture weights, all of the nonparametric basis density functions are stably computed. The bandwidth-scaling hyperparameter r is determined with a validation method whose details are explained in Section 4.6.

4.3 Sparse Diffusion Kernel on a Link Connectivity Graph

Based on the original digraph of the road network, we design an undirected weighted graph which we call the link connectivity graph, whose adjacency matrix is given in Section 4.3.1. The link connectivity graph incorporates only the direct connections between links, while the basis density functions could be interpolated to one another even among indirectly connected links. For such intrinsic interpolations among the indirectly connected links, Section 4.3.2 introduces a sparse diffusion kernel whose computation is more efficient than the standard diffusion kernel.

4.3.1 Link Adjacency Matrix

We design the link connectivity graph based on a consideration about the flows of traffic and direct connections between links. Let $\boldsymbol{x}_v \in \mathbb{R}^2$ be the location of an intersection $v \in V$ and $\boldsymbol{\Delta}(e) \triangleq \boldsymbol{x}_v - \boldsymbol{x}_u$ for $e = (u, v)$ such that $u, v \in V$. The link connectivity graph G_E is an undirected and weighted graph, whose nodes are links and whose edges are connections between the links. Let us define an $|E|$-by-$|E|$ matrix $\boldsymbol{A} = (a_{ij}; e_i = (u_i, v_i), e_j = (u_j, v_j))$ as

$$a_{ij} = \begin{cases} \frac{1}{2} + \frac{\boldsymbol{\Delta}(e_i)^\top \boldsymbol{\Delta}(e_j)}{2\|\boldsymbol{\Delta}(e_i)\|\|\boldsymbol{\Delta}(e_j)\|} & \text{if } u_i = v_j \vee v_i = u_j \\ 0 & \text{otherwise} \end{cases}, \tag{4.7}$$

where $\vee$ denotes the logical "or" operator. Two links are connected if and only if the head of one link is the tail of another link, because we consider movement of vehicles from one link to another link. While we could connect two links that share heads or tails, physical influences between such two links are not so strong as the relationships we consider.

In addition, Equation (4.7) incorporates the cosine similarity between the directional vectors of the links. When the directions of the two links are the same, the weight is one. Conversely, the weight between the links having opposite directions to each other is set zero. Such design of weights is based on the physical observation that a lane of road are often crowded while its opposite lane is not.

4.3.2 Sparse Approximation of the Diffusion Kernel

Diffusion kernel on an undirected graph (Kondor and Lafferty, 2002) is a widely-used metric to compute the similarity between two unconnected nodes of the graph. Let $\boldsymbol{D}$ be a diagonal matrix such that $\boldsymbol{D} = diag\left(\sum_{j=1}^{|E|} a_{1j}, \ldots, \sum_{j=1}^{|E|} a_{|E|j}\right)$. We compute the negative of normalized graph Laplacian matrix as $\boldsymbol{H} = \boldsymbol{D}^{-1/2}\left(\boldsymbol{A} - \boldsymbol{D}\right)\boldsymbol{D}^{-1/2}$. The diffusion kernel matrix is given as a matrix exponential

$$\exp(\beta \boldsymbol{H}) \equiv \lim_{p \to \infty} \left(\boldsymbol{I} + \frac{\beta}{p}\boldsymbol{H}\right)^p,$$

where $\boldsymbol{I}$ is the identity matrix, β (> 0) is a diffusion hyperparameter, and the (i, j) element of the matrix $\exp(\beta \boldsymbol{H})$ represents the similarity between the edges e_i and e_j.

The diffusion kernel matrix is interpreted as a transition-probability matrix of a continuous-time Markov chain. For every row vector in the negative Laplacian matrix $\boldsymbol{H}$, the sum of its elements is zero. Hence $\boldsymbol{H}$ is regarded as the generator matrix of a continuous-time Markov chain, whose diffusion time is represented as the hyperparameter β.

Unfortunately, evaluating the exact value of the matrix exponential $\exp(\beta \boldsymbol{H})$ is computationally infeasible. Though the generator matrix $\boldsymbol{H}$ is sparse, the matrix $\exp(\beta \boldsymbol{H})$ becomes dense in general. Since our generator matrix $\boldsymbol{H}$ is extremely large as $3,290,523$-by-$3,290,523$, accurately computing and storing all of the elements of $\exp(\beta \boldsymbol{H})$ is unrealistic in our limited computing environment.

To approximately compute the diffusion kernel matrix within the feasible computational cost and memory, we instead calculate a power of the matrix $\boldsymbol{H}$. Assume that the traffic in a link only affects those in its close links. Then the hyperparameter β should be set by a small value and we are able to use an approximate kernel matrix

$$\boldsymbol{K}(\beta, p) = \left(\boldsymbol{I} + \frac{\beta}{p}\boldsymbol{H}\right)^p = \sum_{q=0}^{p} \frac{p!\beta^q}{q!(p-q)!p^q}\boldsymbol{H}^q, \tag{4.8}$$

where p is a resolution hyperparameter in discretizing the time in the Markov chain. The $(\pi[i], j)$ element of the matrix $\boldsymbol{K}(\beta, p)$ gives the link similarity $K_E(e_{\pi[i]}, e_j)$ in (4.1).

Due to the limited memory in computation, we adopt $p = 8$ and choose β from $\{1, 2, 3, 4, 5\}$. Since the sum of each row vector of $\left(\boldsymbol{I} + \frac{\beta}{p}\boldsymbol{H}\right)$ is one, $\left(\boldsymbol{I} + \frac{\beta}{p}\boldsymbol{H}\right)$ is a transition-probability matrix of a discrete-time Markov chain, and $\boldsymbol{K}(\beta, p)$ is also a transition-probability matrix whose duration time is p. Because the amount of the travel-time samples is limited, we calculate the $(\pi[i], j)$ element of the kernel matrix $\boldsymbol{K}(\beta, p)$ only for each pair $(\pi[i], j)$ satisfying $e_{\pi[i]} \in E_+$ and $e_j \in E$. We save the computational cost by calculating $\boldsymbol{c}^\top \boldsymbol{H}^q$ instead of fully computing $\boldsymbol{H}^q$, where $\boldsymbol{c}$ is an $|E|$-dimensional sparse vector whose ith element is one if $e_i \in E_+$ or zero otherwise. After caching the values of $\boldsymbol{c}^\top \boldsymbol{H}, \boldsymbol{c}^\top \boldsymbol{H}^2, \ldots, \boldsymbol{c}^\top \boldsymbol{H}^p$, we compute each of the link similarities depending on the value of the hyperparameter β.

4.4 Optimization of the Absolute Importances of Links

This section describes the optimization algorithm for the vector of link importance $\boldsymbol{\lambda}$ in (4.1), after the basis functions are fitted and the approximate diffusion kernel matrix is computed. By following the principle of the Kullback-Leibler Importance Estimation Procedure (KLIEP; Sugiyama et al. (2008)), Section 4.4.1 introduces the objective in optimizing the link importances. Considering the fact that the original KLIEP cannot be directly applied due to its computational inefficiency for our datasets, Section 4.4.2 discusses the application of the accelerated convex clustering algorithm for our optimization problem.

4.4.1 KLIEP for Travel-Time Distributions

Our statistical problem is a conditional density estimation of the relative travel-time y, conditional on each link e. When we count a pair (e, y) as one sample in the estimation, the optimization of the link importances is formalized as

$$\begin{aligned} &\max_{\boldsymbol{\lambda}} \sum_{e \in E_+} \sum_{y \in \mathcal{Y}[e]} \log \left[\lambda_0 \varphi_0(y) + \sum_{i=1}^{m} \lambda_i K_E(e, e_{\pi[i]}) \varphi_i(y) \right] \\ &\text{subject to } \sum_{e \in E_+} \sum_{y \in \mathcal{Y}[e]} \left[\lambda_0 + \sum_{i=1}^{m} \lambda_i K_E(e, e_{\pi[i]}) \right] = N. \end{aligned} \tag{4.9}$$

In addition to Optimization (4.9), we evaluate the outcomes by solving anther optimization problem

$$\begin{aligned} &\max_{\boldsymbol{\lambda}} \sum_{e \in E_+} \frac{1}{n[e]} \sum_{y \in \mathcal{Y}[e]} \log \left[\lambda_0 \varphi_0(y) + \sum_{i=1}^{m} \lambda_i K_E(e, e_{\pi[i]}) \varphi_i(y) \right] \\ &\text{subject to } \sum_{e \in E_+} \frac{1}{n[e]} \sum_{y \in \mathcal{Y}[e]} \left[\lambda_0 + \sum_{i=1}^{m} \lambda_i K_E(e, e_{\pi[i]}) \right] = |E_+|. \end{aligned} \tag{4.10}$$

In Optimization (4.10), each travel-time sample is reweighed so that the total number of samples belonging to the same link becomes one. The difference between the outcomes by (4.9) and (4.10) is evaluated in the later experiments.

For convenience, in this chapter, the optimization objectives (4.9) and (4.10) are called the Type 1 and Type 2 objectives, respectively. Whether to adopt the Type 1 or Type 2 objective depends on what kind of the characteristics in vehicle traffic we want to accurately model. Links having many travel-time samples are actually favored by many drivers. Hence the Type 1 objective is appropriate for predicting typical traffics, because it heavily weighs the predictive capability to the links that are well often chosen by drivers. On the other side, the data sparseness especially in rural areas is not due to the nature of the traffic, but due to our limited capability in collecting the automobile samples. Hence the Type 2 objective that equally weighs every link is reasonable when we want to rigorously evaluate the generalization capability even for the rural links. In the later experiments, instead of selecting either of the Type 1 or 2, we test two types of the prediction tasks where predictive capabilities of our nonparametric estimators are compared with the existing regression methods on both of the objectives. Measure to evaluate the predictive performances depends on the choice of the objective, as we show in Section 4.6.

Both of Optimizations (4.9) and (4.10) yield the sparse optima of the vector $\boldsymbol{\lambda}$. While one can compute the optimum simply with gradient ascending, such naïve optimization is known to run very slowly (Sugiyama et al., 2008). Since the vector $\boldsymbol{\lambda}$ is about 30,000-dimensional in our problem, we need an accelerated algorithm in solving the optimization problem (4.9).

We here complement the reason why we adopt a two-step procedure for optimizing the link-importance weights. Equation (4.1) consists of a linear combination of the basis density functions that are the linear combinations of the fundamental density functions. While we could consider a single optimization where Equation (4.1) is replaced with another equation that is solely a linear combination of the fundamental density functions, such expansion makes the computational cost very high. The bi-linear modeling using the sampling of the L fundamental density functions is an outcome of considering the computational feasibility. Fortunately, the models fitted with the bi-linear principle still successfully predict the real vehicle traffic with high accuracy, as we show later in Section 4.6.

4.4.2 Transformation of the KLIEP into Convex Clustering

To accelerate the optimization, we transform (4.9) or (4.10) into an optimization in convex clustering. Let $K_E(e, e_{\pi[0]}) \equiv 1$ for convenience. For the Type 1 objective, we introduce a variable transformation

$$\phi_i \triangleq \frac{\lambda_i}{N} \sum_{e \in E_+} \sum_{y \in \mathcal{Y}[e]} K_E(e, e_{\pi[i]}),$$

and a normalized density function

$$\kappa_i(e, y) \triangleq \frac{N \cdot K_E(e, e_{\pi[i]}) \varphi_i(y)}{\sum_{e' \in E_+} n[e'] K_E(e', e_{\pi[i]})}.$$

For the Type 2 objective, we set

$$\phi_i \triangleq \frac{\lambda_i}{|E_+|} \sum_{e \in E_+} \sum_{y \in \mathcal{Y}[e]} K_E(e, e_{\pi[i]}) \text{ and } \kappa_i(e, y) \triangleq \frac{|E_+| \cdot K_E(e, e_{\pi[i]}) \varphi_i(y)}{\sum_{e' \in E_+} K_E(e', e_{\pi[i]})}.$$

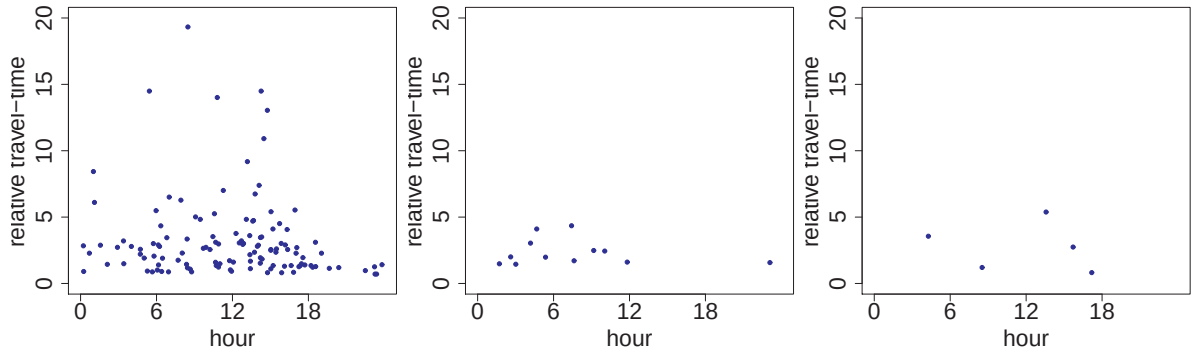
We are able to easily compute the values of $\kappa_i(e, y)$ for $e \in E_+$ and $y \in \mathcal{Y}[e]$, when the basis density functions $\varphi_0, \varphi_1, \ldots, \varphi_m$ and the link similarity function $K_E(\cdot, \cdot)$ are given. Then the optimization problems (4.9) and (4.10) are respectively modified into

$$\text{Type 1:} \quad \max_{\boldsymbol{\phi}} \sum_{e \in E_+} \sum_{y \in \mathcal{Y}[e]} \log \left[\sum_{i=0}^{m} \phi_i \kappa_i(e, y) \right] \text{ subject to } \sum_{i=0}^{m} \phi_i = 1, \text{ and} \tag{4.11}$$

$$\text{Type 2:} \quad \max_{\boldsymbol{\phi}} \sum_{e \in E_+} \frac{1}{n[e]} \sum_{y \in \mathcal{Y}[e]} \log \left[\sum_{i=0}^{m} \phi_i \kappa_i(e, y) \right] \text{ subject to } \sum_{i=0}^{m} \phi_i = 1, \tag{4.12}$$

where $\boldsymbol{\phi} \triangleq (\phi_0, \phi_1, \ldots, \phi_m)^\top$. Since Optimizations (4.11) and (4.12) have the equivalent forms to those in the convex clustering, we utilize the accelerated convex clustering algorithm also in optimizing the link-importance weights.

Note that the SMO in the accelerated convex clustering algorithm requires the information about which functions $\{\kappa_{i'}(e, y)\}$ are regarded as the neighbors of the function $\kappa_i(e, y)$. Here we introduce a sparse vector $\boldsymbol{w}_i \in \mathbb{R}^{|E|}$ such that $\boldsymbol{w}_i^\top = \Big(\kappa_i(e_1, y_{e_1}^{(1)}), \ldots, \kappa_i(e_1, y_{e_1}^{(n[e_1])}),$



link ID="1153023" link ID="1418272" link ID="1530669"

Figure 4.4: Time-dependent travel-time samples for some of the links. The horizontal axis represents time-stamp and the vertical axis represents the relative travel-time. While samples on the link "1153023" are covered for almost all of the timezones, those on the link "1418272" are lacking in the afternoon. In addition, we can confirm outliers resulting in heavy-tails of the travel-time distributions, in the link "1153023".

$\ldots, \kappa_i(e_{|E|}, y^{(1)}_{e_{|E|}}), \ldots, \kappa_i(e_{|E|}, y^{(n[e_{|E|}])}_{e_{|E|}})\Big)$. Then we compute the neighborhood indices by computing the inner product $\boldsymbol{w}_i^\top \boldsymbol{w}_{i'}$. The nearest neighbor computations are not required for all of the pairs $\{(i, i')\}$, because we eliminate the pairs whose inner products are zero by caching the auxiliary variables to indicate which elements of each vector $\boldsymbol{w}_i$ are positive.

4.5 Incorporating the Dependence on Each Timezone

Our spatio-temporal estimator interpolates the multiple spatial estimators to one another in the time domain. To efficiently incorporate the timezone-dependent nature of the travel-time distributions, we first take care of the amount of samples for each timezone. If drivers select a link during daylight times but do not select such link at night, travel-time distributions become significantly heterogeneous among the 24 timezones, and samples of the travel-time are observed only for parts of the timezones. Figure 4.4 shows examples of the amount about the travel-time samples in different timezones, where the shortage of available samples suggests the inefficiency of fitting the distributions independently for each link. The missing travel-time should rather be inferred by exploiting the neighboring links within the same timezone, while the degree of the spatial interpolation should depend on each timezone.

Section 4.5.1 shows how we model the spatio-temporal travel-time distributions, with introducing a timezone-dependent spatial interpolation. We first fit the spatial estimators independently for all of the timezones $h \in \{0, \ldots, 23\}$, and then interpolate them to one another with a temporal similarity function between two time-stamps. Section 4.5.2 addresses a truncated von Mises kernel as the concrete form of the temporal similarity. Though we have datasets observed in only one day, it is natural to assume that the characteristics of traffic are cyclic, because traffic properties of the same hours are similar even among different days. The truncated von Mises kernel allows us for reflecting such cyclic nature of the human behavior, and its sparseness reduces the computational costs in interpolation.

4.5.1 The Spatio-Temporal Distribution

Our spatio-temporal model of the travel-time distribution is given as follows. Assume that we have already performed the spatial estimation for every timezone $h \in \{0, \ldots, 23\}$. Based on the actual dependence on the timezone index h, we re-notate each spatial model as

$$f_{h,e}(y) = \frac{\lambda_{h0}\varphi_{h0}(y) + \sum_{i=1}^{m_h} \lambda_{hi} K_{h,E}(e, e_{\pi_h[i]})\varphi_{hi}(y)}{\lambda_{h0} + \sum_{i=1}^{m_h} \lambda_{hi} K_{h,E}(e, e_{\pi_h[i]})},$$

where each variable $(\cdot)_{h\cdot}$ corresponds to the variable $(\cdot)$ in which the index h was omitted in Sections from 4.1 to 4.4. The probability density function of the relative travel-time y conditional on link e and time-stamp t is given as

$$f_e(y|t) = \frac{\sum_{h=0}^{H-1} K_T(t, t_h) f_{h,e}(y)}{\sum_{h=0}^{H-1} K_T(t, t_h)},$$

where t_h is the time-stamp associated with each timezone h, $K_T(\cdot, \cdot)$ is a similarity function between two time-stamps, and $H = 24$. Since one dataset indexed with h belongs to an hour $[h, h+1)$, we set $t_0, t_1, \ldots, t_{23}$ as the intermediate time-stamps 0:30, 1:30, ..., and 23:30, respectively.

4.5.2 Truncated Von Mises Kernel to Represent a Cycle

Let $\rho(t) \in [0, 2\pi)$ be the angle to represent a time-stamp t during one day. For example, $\rho(0:00) = 0$, $\rho(1:00) = \frac{1}{12}\pi$, and $\rho(23:00) = \frac{23}{12}\pi$, respectively. A natural way to incorporate the circularity of the human behaviors is to adopt von Mises kernel (Evans et al., 2000) function

$$K_T(t, t_h) \propto \exp\left[\xi \cos\left(\rho(t) - \rho(t_h)\right)\right] \tag{4.13}$$

where ξ is the concentration hyperparameter to determine the degree of interpolation. Since we prefer a sparse version of (4.13) for efficient computation, we introduce a truncated similarity function as

$$K_T(t, t_h) = \left[\exp\left[\xi \cos\left(\rho(t) - \rho(t_h)\right) - \xi\right] - \varepsilon\right]_+, \tag{4.14}$$

where $[\cdot]_+ \triangleq \max\{\cdot, 0\}$ and ε is a threshold hyperparameter fixed as $\varepsilon = 0.01$ in this chapter. Note that the value of the kernel is standardized as $K_T(t_h, t_h) = 1 - \varepsilon$ in (4.14), for easily setting the magnitude of the threshold hyperparameter ε. The operator $[\cdot]_+$ makes the similarity exactly zero when two time-stamps are distant, and reduces the computational costs in interpolation. We determine the concentration hyperparameter ξ with cross-validation or hold-out methods.

Because of the nature of the cosine function, Equation (4.14) incorporates cyclic patterns of human behavior. While we only deal with the diurnal patterns, one can easily incorporate longer-term cycles such as the dependence on each day of week and month, if having multi-day probe-car datasets. When we incorporate the longer-term cycles, the new similarity function would be an additive or a multiplicative combination of equations derived from (4.14), with multiple concentration hyperparameters.

4.6 Experimental Evaluations

The main aims in this chapter's experiments are to compare our methods with parametric regression methods, and to assess what types of the basis density functions make the predictions highly accurate. Section 4.6.1 provides the settings in the experiments and explains how we measure each of the performances. Since it is not trivial how to apply the existing regression methods for our datasets, we

describe the procedures of the reference regression methods, which adopt the Ordinary Least Squares principle, in Section 4.6.2. The main results are provided in Section 4.6.3 where our estimator using the nonparametric basis density functions yielded the best performances. We also confirmed certain improvements by introducing the temporal interpolation, especially when we rigorously evaluate the generalization capability of each model. In Section 4.6.4, we visualize which links possess travel-time distributions that are significantly different from simple parametric distributions.

4.6.1 Settings and Performance Metrics

For each hourly dataset, we perform 10-fold likelihood cross-validation where the entire dataset is randomly split into an 80% training dataset and a 20% test dataset in each of the 10 trials. In choosing the hyperparameters, we maximized the average log-likelihood for the validation subset that is 20% of the training dataset, based on a fitted model using the remaining 80% subset of the training dataset[3]. After the best-score hyperparameters are chosen, the final model is fitted for the entire training dataset and the performance for the test dataset is evaluated.

The main task is to compare our nonparametric estimators with the existing regression methods independently for each hourly dataset. In the main task, a model fitted with the training dataset whose hour is h is applied for the test dataset of the same hour h. The additional task is to evaluate the improvements with the temporal interpolation introduced in Section 4.5. In this additional task, we fit the spatio-temporal model using all of the training datasets $h \in \{0, 1, \ldots, 23\}$, and then evaluate the predictive performance of the fitted model for each hourly test dataset.

Here are our notations to discriminate each dataset, where the timezone index h is omitted. Using the same manner in the training dataset $\mathcal{D} = \{\mathcal{Y}[e]; e \in E_+\}$, we denote the test dataset by $\mathcal{D}^* = \{\mathcal{Y}^*[e]; e \in E_+^*\}$. In the hyperparameter optimization, the entire training dataset $\mathcal{D}$ is randomly split into the training subset $\widetilde{\mathcal{D}} = \{\widetilde{\mathcal{Y}}[e]; e \in \widetilde{E}_+\}$ and the validation subset $\widetilde{\mathcal{D}}^* = \{\widetilde{\mathcal{Y}}^*[e]; e \in \widetilde{E}_+^*\}$. The number of travel-time samples for a link e is represented with $n^*[e]$ in the test dataset, $\widetilde{n}[e]$ in the training subset, or $\widetilde{n}^*[e]$ in the validation subset. We also define $N^* \triangleq \sum_{e \in E} n^*[e]$, $\widetilde{N} \triangleq \sum_{e \in E} \widetilde{n}[e]$, and $\widetilde{N}^* \triangleq \sum_{e \in E} \widetilde{n}^*[e]$.

We adopt two types of cross-validation criteria that match with the Type 1 and Type 2 optimization objectives introduced in Section 4.4.1. Let $\widetilde{f}_e(y)$ be a density model fitted for the training subset $\widetilde{\mathcal{D}}$. We first choose the hyperparameters such as the diffusion parameter β and the bandwidth-scaling parameter r, by maximizing the validation log-likelihood per sample

$$\text{Type 1:} \quad \mathcal{L}(\widetilde{\mathcal{D}}^* | \widetilde{f}) = \frac{1}{\widetilde{N}^*} \sum_{e \in \widetilde{E}_+^*} \sum_{y \in \widetilde{\mathcal{Y}}^*[e]} \log \widetilde{f}_e(y), \text{ or}$$

$$\text{Type 2:} \quad \mathcal{L}(\widetilde{\mathcal{D}}^* | \widetilde{f}) = \frac{1}{|\widetilde{E}_+^*|} \sum_{e \in \widetilde{E}_+^*} \frac{1}{\widetilde{n}^*[e]} \sum_{y \in \widetilde{\mathcal{Y}}^*[e]} \log \widetilde{f}_e(y).$$

The bandwidth hyperparameter r was chosen from $\{1, 1.5, 2, 2.5, 3\}$, and the diffusion hyperparameter β was chosen from $\{1, 2, 3, 4, 5\}$.

In the full spatio-temporal model, we simplified the hyperparameter selection procedure because the cross-validation of the three hyperparameters (β, r, ξ) is computationally intensive. First, for each hourly dataset, we fixed the values of (β, r) with the same ones as in models without the temporal

[3]Ones who favor rigorous hyperparameter optimization would use multiple-fold cross-validation. Yet in our preliminary experiments, one trial was sufficient to choose appropriate hyperparameters because of the large amounts of the samples.

interpolation. Then we only search for the best value of ξ from $\{1, 3, 10, 30, 100, 300\}$. Since $\xi = 100$ provided the best performance for most of the datasets, we show the predictive performances when we used the concentration hyperparameter $\xi = 100$.

Using the optimal hyperparameters selected, we fit the final model $f_e(y)$ using the entire training dataset $\mathcal{D}$. The test-set log-likelihood per sample is given as

$$\text{Type 1:} \quad \mathcal{L}(\mathcal{D}^*|f) = \frac{1}{N^*} \sum_{e \in E_+^*} \sum_{y \in \mathcal{Y}^*[e]} \log f_e(y), \text{ or} \tag{4.15}$$

$$\text{Type 2:} \quad \mathcal{L}(\mathcal{D}^*|f) = \frac{1}{|E_+^*|} \sum_{e \in E_+^*} \frac{1}{n^*[e]} \sum_{y \in \mathcal{Y}^*[e]} \log f_e(y). \tag{4.16}$$

Our performance score is the average of Equation (4.15) or Equation (4.16) among the random 10 folds, and a high score is the evidence of good performance.

We regard our performance measurement as an approach of evaluating the generalization capability for the links that lack the observations of the GPS trace. In the cross-validation, links having only a few travel-time samples are often eliminated from the training subset. Hence we perform an out-of-sample prediction for the link that has no sample but is close to some links having travel-time samples. As a result, the generalization capabilities are more rigorously measured when we adopt the Type 2 objective than when we adopt the Type 1 objective. It remains an open problem to evaluate for the link whose neighboring links never have travel-time samples.

4.6.2 Parametric Models as Reference Methods

As the reference regression methods, we selected the k-nearest neighbor regression and the Nadaraya-Watson kernel regression based on the Ordinary Least Squares principle. In both methods we regressed the logarithms of the relative travel-time, and fit log-normal distributions of travel-time. While we also tried Gaussian process regression, its costly matrix inversion and determinant evaluation were impractical for our datasets.

We implemented a k-nearest neighbor regression as the simplest predictor, whose spatial structure is not based on the link connectivity but based on the Euclidean distance between the two-dimensional locations. Let us define the location of a link $e = (u, v)$ as $\boldsymbol{x}(e) = \frac{1}{2}(\boldsymbol{x}_u + \boldsymbol{x}_v)$, where $\boldsymbol{x}_u, \boldsymbol{x}_v \in \mathbb{R}^2$ are introduced in Section 4.3.1. Let $\mathcal{E}_k(e)$ be a set of k-nearest-neighbor links calibrated with the Euclidean distance from the location $\boldsymbol{x}(e)$. The travel-time distribution for the link e is determined with the set of samples $\{y \in \mathcal{Y}_{e'}; e' \in \mathcal{E}_k(e)\}$.

The relative travel time is modeled by a log-normal distribution $f_e(y) = \mathcal{LN}(y; \nu_e, \sigma_e^2)$ such that

$$\nu_e = \begin{cases} \dfrac{\nu_0 + \sum_{e' \in \mathcal{E}_k(e)} \sum_{y \in \mathcal{Y}_{e'}} \log y}{1 + \sum_{e' \in \mathcal{E}_k(e)} n[e']} & \text{for Type 1} \\ \dfrac{\nu_0 + \sum_{e' \in \mathcal{E}_k(e)} \frac{1}{n[e']} \sum_{y \in \mathcal{Y}_{e'}} \log y}{1 + |\mathcal{E}_k(e)|} & \text{for Type 2} \end{cases} \text{and}$$

$$\sigma_e^2 = \begin{cases} \dfrac{\sigma_0^2 + \sum_{e' \in \mathcal{E}_k(e)} \sum_{y \in \mathcal{Y}_{e'}} (\log y - \nu_e)^2}{1 + \sum_{e' \in \mathcal{E}_k(e)} n[e']} & \text{for Type 1} \\ \dfrac{\sigma_0^2 + \sum_{e' \in \mathcal{E}_k(e)} \frac{1}{n[e']} \sum_{y \in \mathcal{Y}_{e'}} (\log y - \nu_e)^2}{1 + |\mathcal{E}_k(e)|} & \text{for Type 2} \end{cases},$$

where the stabilization parameters ν_0 and σ_0^2 are given in (4.2) and (4.4). The hyperparameter k was chosen from $\{1, 2, \ldots, 10\}$, and $k = 7$ was optimal in many cases.

Another reference method is the Nadaraya-Watson kernel regression that incorporates the link connectivity graph as the spatial structure but assumes the relative travel-time to be log-normal distributed. We give

$$\nu_e = \begin{cases} \dfrac{\nu_0 + \sum_{e' \in E_+} K_E(e,e') \sum_{y \in \mathcal{Y}[e']} \log y}{1 + \sum_{e' \in \mathcal{E}_k(e)} K_E(e,e') n[e']} & \text{for Type 1} \\ \dfrac{\nu_0 + \sum_{e' \in E_+} K_E(e,e') \frac{1}{n[e']} \sum_{y \in \mathcal{Y}[e']} \log y}{1 + \sum_{e' \in \mathcal{E}_k(e)} K_E(e,e')} & \text{for Type 2} \end{cases} \quad \text{and}$$

$$\sigma_e^2 = \begin{cases} \dfrac{\sigma_0^2 + \sum_{e' \in E_+} K_E(e,e') \sum_{y \in \mathcal{Y}[e']} (\log y - \nu_e)^2}{1 + \sum_{e' \in \mathcal{E}_k(e)} K_E(e,e') n[e']} & \text{for Type 1} \\ \dfrac{\sigma_0^2 + \sum_{e' \in E_+} K_E(e,e') \frac{1}{n[e']} \sum_{y \in \mathcal{Y}[e']} (\log y - \nu_e)^2}{1 + \sum_{e' \in \mathcal{E}_k(e)} K_E(e,e')} & \text{for Type 2} \end{cases},$$

where the link similarity function $K_E(\cdot,\cdot)$ is based on the approximate diffusion kernel whose diffusion hyperparameter β is chosen from $\{1, 2, \ldots, 5\}$.

4.6.3 Evaluation Results

First, we mention the comparisons between our nonparametric estimators and the existing parametric regression methods, without the temporal interpolation. Figures 4.5 and 4.6 show the Type 1 and Type 2 scores based on the 10-fold likelihood cross-validation, where our nonparametric estimators outperformed other methods for all or many of the hourly datasets on the Type 1 or Type 2 objective, respectively. The necessity of tuning the importance weight for each link was confirmed with the worst performances provided by the Nadaraya-Watson regression having uniform weights. In the parametric basis density functions, models with log-normal distributions slightly outperformed those with gamma distributions. Since the difference between the mixture of gamma distributions and that of log-normal distributions was small, we think that both of the nonparametric estimators are sufficiently flexible to model the relative travel-time for every link.

On the Type 2 objective that rigorously evaluates the generalization capabilities in the rural regions, performances of the nonparametric estimators become relatively worse during the night time. Since our training samples are more limited at night than in daylight time, we regard the over-fitting of the nonparametric method as the main reason of the worse performances. Yet the nonparametric estimators are totally the most outperforming methods for our large-scale datasets, and the nonparametric nature of our methods is still evidenced to be beneficial on severe valuation criteria.

We should also report an insight in the hyperparameter optimization for the spatial estimators. In many cases, the setting $(r, \beta) = (1.5, 3)$ or $\beta = 5$ was optimal for the nonparametric or parametric basis density functions, respectively. The optimal interior value, as the best trade-off point between the smoothness and the flexibility, was found only with the nonparametric approaches.

Second, we show the improvements provided with the temporal interpolation for models using the nonparametric basis density functions. Since our nonparametric basis density functions were shown to provide the good performances, our next aim is to evaluate the full spatio-temporal model that extends the nonparametric spatial estimators. For the Type 1 and Type 2 objectives, Figures 4.7 and 4.8 respectively show the comparisons among the models, depending on whether or not the temporal interpolation is incorporated.

In both of the Type 1 and Type 2 objectives, the temporal interpolation improves the generalization capability of our nonparametric estimators, while improvements in the Type 2 objective are higher than

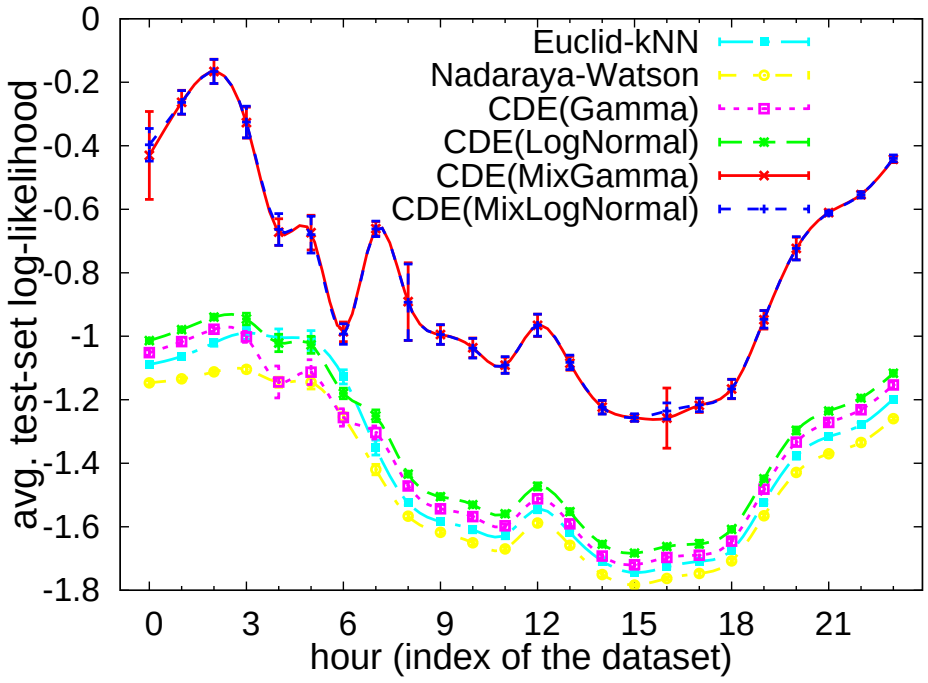


Figure 4.5: Type 1 average test-set log-likelihood for each hourly dataset based on the 10-fold cross-validation. The horizontal axis denotes the hour of a day, and the lengths of error bars are two standard deviations. The method "Euclid-kNN" is the k-nearest neighbor regression using the Euclidean distance, "Nadaraya-Watson" is the Nadaraya-Watson regression using the diffusion kernel, and methods named "CDE($\cdot$)" are our conditional density estimators. Methods using the nonparametric basis density functions, which are given as mixtures of gamma or log-normal distributions, achieved the highest prediction performances for all of the 24 datasets.

those in the Type 1 objective. The temporal interpolation is effective particularly when we rigorously evalute the generalization capabilities, and when we predict the travel-time in the rural regions.

4.6.4 Visualization of the Special Links

We consider that a link is special if the fitted travel-time distribution is quite different from the simple parametric distribution. For the fitted probability density function $f_e(y)$, let $g_e(y)$ be its approximation by a single parametric distribution, where the approximation is done with a moment matching. To compute the distance between the distributions $f_e(y)$ and $g_e(y)$, we use the Cauchy-Schwarz (CS) divergence measure (Príncipe, 2010) defined as

$$CS(f,g|e) = -\log \frac{\int_y f_e(y) g_e(y) dy}{\sqrt{\int_y f_e^2(y) dy \int_y g_e^2(y) dy}},$$

which is the cosine similarity between the two probability density functions. The CS divergence is zero if and only if $f_e(y)$ and $g_e(y)$ are the same, and becomes large when $f_e(y)$ and $g_e(y)$ are dissimilar.

We calculate the CS divergence in case of the mixture of gamma distributions. Since every link

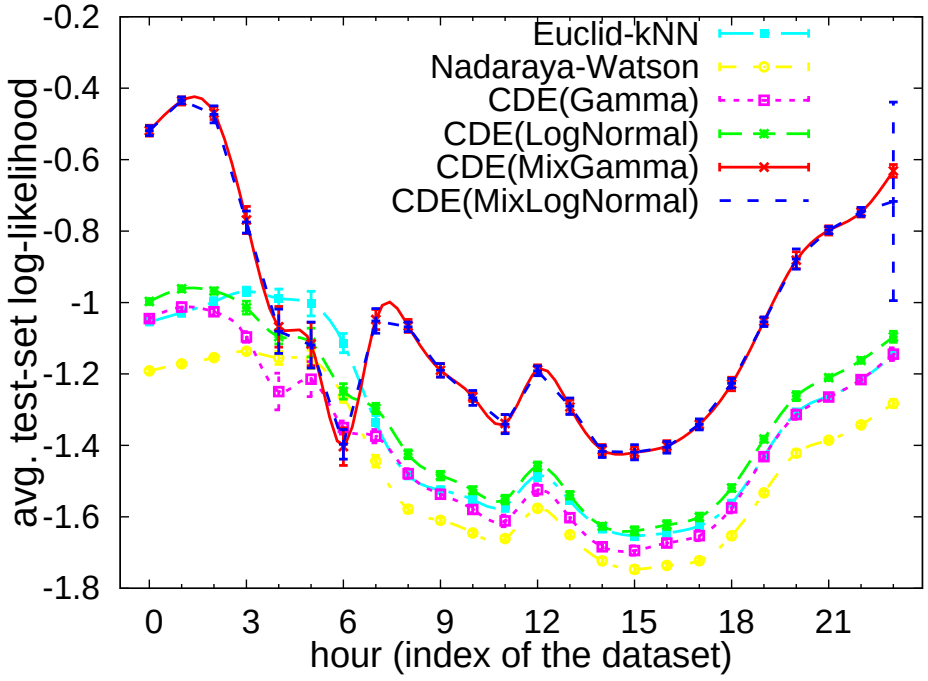


Figure 4.6: Type 2 average test-set log-likelihood for each hourly dataset based on the 10-fold cross-validation, where meanings of each model, axis, and error bar are the same as in Figure 4.5. While models with the nonparametric basis density functions become worse in nighttime, they are still most outperforming in many of the datasets during daylight timezones.

has at most L instances of gamma distributions, we rewrite the distribution as

$$f_e(y) = \sum_{\ell=1}^{L} w_{e\ell} Gam(y; \alpha_\ell, \mu_\ell)$$

where the coefficients $w_{e1}, \ldots, w_{eL}$ are computed with (4.1) and (4.5) in such a way that $\sum_{\ell=1}^{L} w_{e\ell} = 1$. By matching the first and second moments, we set the density function $g_e(y) = Gam(y; \widetilde{\alpha}_e, \widetilde{\mu}_e)$ with

$$\widetilde{\mu}_e = \sum_{\ell=1}^{L} w_{e\ell}\mu_\ell \text{ and } \widetilde{\alpha}_e = \left[\sum_{\ell=1}^{L} w_{e\ell} \frac{\alpha_\ell + 1}{\alpha_\ell} \left(\frac{\mu_\ell}{\widetilde{\mu}_e}\right)^2 - 1\right]^{-1}.$$

The CS divergence is calculated as $CS(f, g|e) = \frac{1}{2}\log \overline{f_e^2} - \log \overline{f_e g_e} + \frac{1}{2}\log \overline{g_e^2}$, where

$$\overline{f_e^2} = \sum_{\ell=1}^{L}\sum_{t=1}^{L} \frac{w_{e\ell} w_{et} \alpha_\ell^{\alpha_\ell} \alpha_t^{\alpha_t}}{\Gamma(\alpha_\ell)\Gamma(\alpha_t)\mu_\ell^{\alpha_\ell}\mu_t^{\alpha_t}} \frac{\Gamma(\alpha_\ell + \alpha_t - 1)}{\left(\frac{\alpha_\ell}{\mu_\ell} + \frac{\alpha_t}{\mu_t}\right)^{\alpha_\ell + \alpha_t - 1}},$$

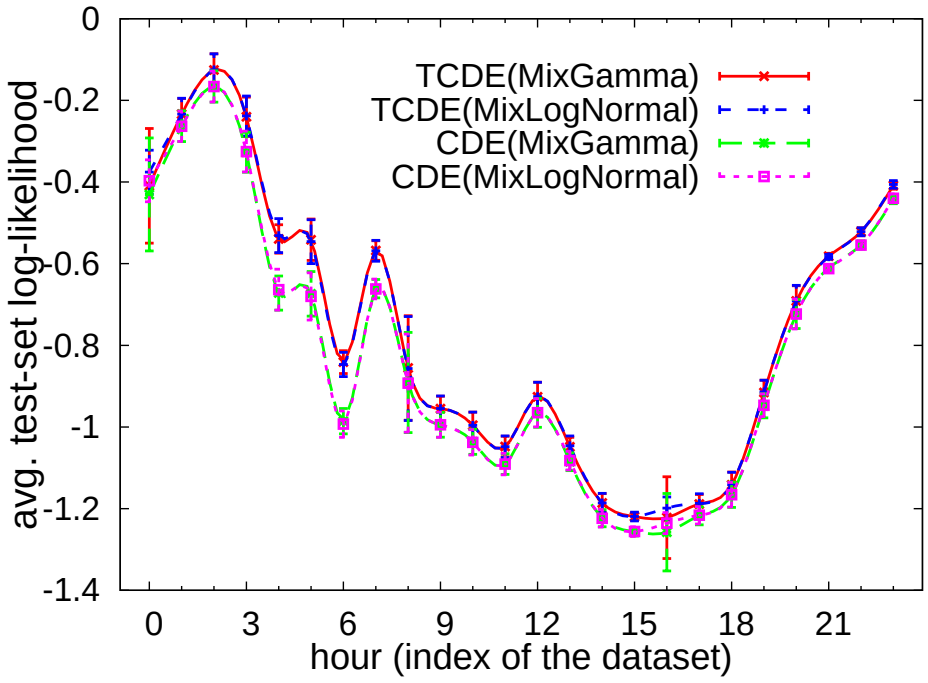


Figure 4.7: Performance improvements with temporal interpolation, when we adopt Type 1 average test-set log-likelihood with 10-fold cross-validation. Both "TCDE($\cdot$)" and "CDE($\cdot$)" use nonparametric basis density functions, while the temporal interpolation is incorporated only in "TCDE($\cdot$)". "CDE($\cdot$)" is the same model as that in Figure 4.5. In the temporal interpolation, we choose the concentration hyperparameter $\xi = 100$ based on the held-out method. For the 21 datasets such that $h \notin \{0, 8, 16\}$, models with the temporal interpolation outperformed those without the temporal interpolation. For the datasets $h \in \{0, 8, 16\}$, models with the temporal interpolation yielded statistically indistinguishable performances with those without the temporal interpolation.

$$\overline{f_e g_e} = \frac{\widetilde{\alpha}_e^{\widetilde{\alpha}_e}}{\Gamma(\widetilde{\alpha}_e)\widetilde{\mu}_e^{\widetilde{\alpha}_e}} \sum_{\ell=1}^{L} \frac{w_{e\ell}\alpha_\ell^{\alpha_\ell}}{\Gamma(\alpha_\ell)\mu_\ell^{\alpha_\ell}} \frac{\Gamma(\alpha_\ell + \widetilde{\alpha}_e - 1)}{\left(\frac{\alpha_\ell}{\mu_\ell} + \frac{\widetilde{\alpha}_e}{\widetilde{\mu}_e}\right)^{\alpha_\ell + \widetilde{\alpha}_e - 1}},$$

$$\text{and} \quad \overline{g_e^2} = \frac{\Gamma(2\widetilde{\alpha}_e - 1)\widetilde{\alpha}_e}{2^{2\widetilde{\alpha}_e - 1}\Gamma^2(\widetilde{\alpha}_e)\widetilde{\mu}_e}.$$

Figure 4.9 plots the special links whose travel-time distributions are quite different from their approximates. The special links tend to be located in the urban regions, because many travel-time samples are needed for fitting the probability density functions having complex shapes. The special links, however, are not concentrated in specific areas inside the urban region, because risks of encountering extreme delays also exist outside the center of Tokyo. In terms of the difference among the hours, the special links in midnight are more dispersed than those in daylight hours, probably because many samples are available (see Table 4.1) and persons who missed the last trains often ride taxis for a long time.

Let us finally comment on one future scenario to exploit the visualization of the special links. In the risk-sensitive routing algorithms, the values of the risk measures such as Iterated Conditional Tail

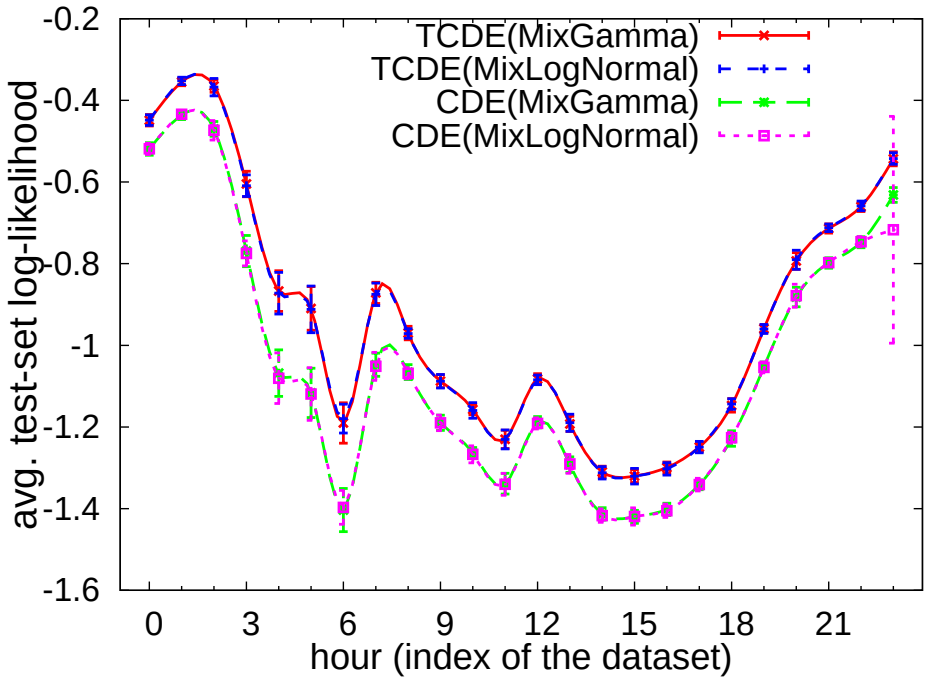
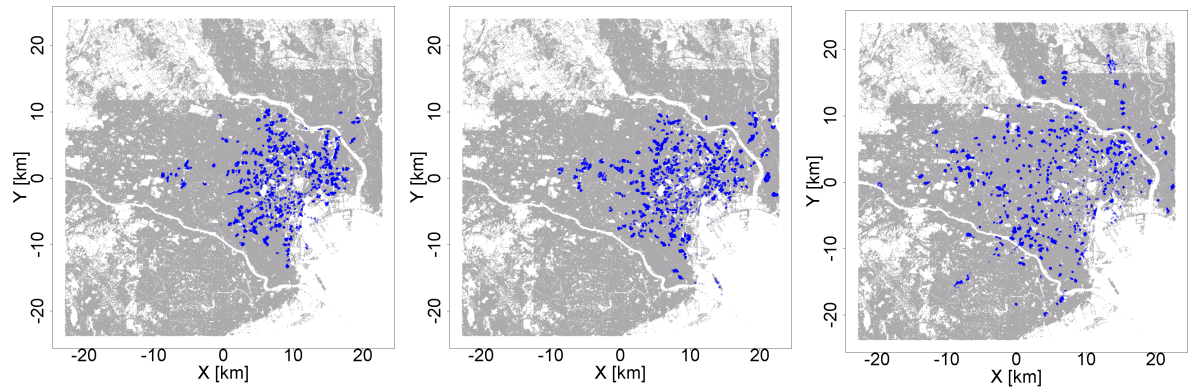


Figure 4.8: Performance improvements with temporal interpolation, when we adopt Type 2 average test-set log-likelihood with 10-fold cross-validation. Both "TCDE(·)" and "CDE(·)" correspond to the same models as those of Figure 4.7, except that Type 2 objectives are adopted in the training. Models with the temporal interpolation outperformed those without the temporal interpolation, for all of the hourly datasets. In addition, on the Type 2 objectives, the additional improvements provided by the temporal interpolation are larger than those for the Type 1 objectives.

Expectation strongly depend on the particular quantiles of the travel-time distributions. Thus, when the distribution has multiple modes or a heavy tail, small variability in tuning each driver's risk-sensitivity parameter strongly affects the computed values of the risk measures and would dramatically change the chosen routes. Visualizing the complexity of the distribution is beneficial in assessing the details of the route choice.

4.7 Discussion

Let us discuss the possible directions based on the high experimental performances achieved by the proposed models. As long as we have large-scale probe-car datasets, we expect that the remaining margin of additional performance improvement is not so high, even when we further improve the nonparametric density estimation algorithms. In contrast, travel-time prediction by using only limited amount of datasets would require characteristically different approaches. When we model the traffic of countries that do not have a good infrastructure of GPS, we instead need to use counting records about the numbers of vehicles staying for each link. When we regard the numbers of vehicles as training data, we are able to perform nonparametric density estimation about the number of vehicles, as well as regression modeling from the number of vehicles into the travel time. Such new travel-time prediction methods would become keys to simulate the real traffic in many developing countries, while



12 p.m.- 6 p.m.- 12 a.m.-

Figure 4.9: Examples of the special links whose travel-time distributions are significantly different from their exponential-family approximations, using the magnified view for the center of Tokyo. The blue-colored points specify the locations of the links having top-1% highest CS divergence scores, in terms of the dissimilarity between the nonparametric travel-time distribution and its parametric approximation. The special links exist in broad regions, while those in midnight are more broadly scattered than those in daylight hours.

our estimation algorithms are specialized for the Greater Tokyo Area that has a mature infrastructure of automobile transportation.

When we adopt the current high-accuracy estimates of the travel-time distributions, the remaining component we need to develop is a new data-mining algorithm to perform whole traffic simulation. One of the essential tasks to make the simulation realistic is the fitting of the risk-sensitivity parameters for every of the individual vehicle drivers. Like the choice modeling we discussed in Section 1.2, we need to formalize a descriptive route choice problem whose available options are lots of the traveling routes associated with the fitted travel-time distributions. How to model the bounded rationality of each vehicle driver, by comparing the normative choice with that of the descriptive models, is an important study to clarify the microscopic mechanism of each driver to cause the complex macroscopic behavior of the entire traffic system. In addition, since the reality of the found insights depends on the accuracy of map matching, preparing multiple datasets by several map matching algorithms would provide more robust information, even when the same trajectories of probe-cars are used.

4.8 Summary

This chapter introduced an application of our nonparametric mixture modeling for transportation decision making, which exploits travel-time distributions conditional on both the link of a road network and timezone. In order to fit the distributions for huge number of links and to reflect the positive feedback in a road network, we designed a nonparametric density estimator of the relative travel time, which is called the spatial estimator specific to one timezone. Then the full spatio-temporal estimates are provided by interpolating multiple spatial estimators to one another in the time domain. Every spatial estimator interpolates the nonparametric basis density functions among close links, where each basis density function is a link-dependent mixture of link-independent fundamental density functions modeled as gamma or log-normal distributions. The quantile-based heuristics to set the fundamental

density functions makes the resulting basis density functions involve multiple scales of the relative travel time. The fitting of the mixture weights for such nonparametric basis density functions are efficiently done with the accelerated convex clustering algorithm. In contrast, the interpolation among the multiple basis density functions, for reflecting the positive feedback, is done with the combination of a sparse diffusion kernel to emulate the traffic propagation, and each link's global importance to be optimized also with the accelerated convex clustering algorithm. Because of the flexibility and the global optimality in the fitting, our nonparametric estimator outperformed the parametric regression methods for all of the datasets classified with timezones. The temporal interpolation using the truncated von Mises kernels were shown to improve the predictive performances, particularly when we rigorously evaluate the generalization capability for the links located in the rural regions.

This study for transportation decision making has proved the benefits of our nonparametric mixture modeling to predict the fat tails as a result of the positive feedback. In the next Chapter 5, we proceed into another social-science study that involves the nonparametric mixture modeling of the long-range dependence, by taking the marketing decision making as a practical application example.

Chapter 5

Nonparametric Consumer-Response Prediction

Modeling how marketing actions in various channels influence or cause consumer purchase decisions is crucial for marketing decision-making. Marketing campaigns stimulate consumer awareness, interest and help drive interactions such as the browsing of product web pages, ultimately impacting an individual's purchase decision. The delaying impacts of the marketing campaigns, which accompany the changes of each individual consumer's mental state, are statistically observed as long-range dependences between the marketing-stimulus events and purchase events. In addition, some successful campaigns stimulate word-of-mouth and social trends by causing positive feedbacks among consumers, and such collective behavior of consumers result in concurrent and correlated responses over a short term.

To model both of the long-range dependence among the events each individual consumer experiences, and the positive feedback among interacting consumers, we propose a new continuous-time predictive model for time-dependent response rates of each consumer. The proposed model follows the literature of using the Inhomogeneous Poisson Processes (IPPs; Chatfield and Goodhardt (1975); Schmittlein et al. (1987); Wagner and Taudes (1986)), and our philosophy of the nonparametric multi-scale mixtures is applied for modeling the total influence from all of the past events the same individual has experienced, which we refer to as the individual factor. For both of the computational efficiency and high predictive accuracy, the individual factor is regressed with staircase functions that correspond to nonparametric mixtures of step functions involving multiple activation periods. An illustration of such staircase functions is provided in Figure 5.1, where multiple lengths of the sliding windows handle the nature of time decay in human memory. As in Poisson Network (Rajaram et al., 2005) and Piecewise-Constant Conditional Intensity Model (Gunawardana et al., 2011), we formalize a Poisson regression problem whose total log-likelihood is given as a sum of finite numbers of log-likelihood terms. The piecewise-constant features, which give the staircase functions in predictive formulas, make the likelihood maximization efficient, as well as efficiently approximating the time-decaying human memories.

The individual factor is also regarded as an outcome of the interaction among the multiple memories. As we discussed in Chapter 2, time decays in human memories obey power laws and hyperbolic discounting with respect to elapsed time (Wixted, 1990; Rubin and Wenzel, 1996; Wixted and Ebbesen, 1997). Such power laws are the compound outcomes of the interactions among the short-, mid-, and long-term memories of each human, because an important event that has survived in the short-term memory is transferred into the mid-term memory, in which the survival probability gets higher than in

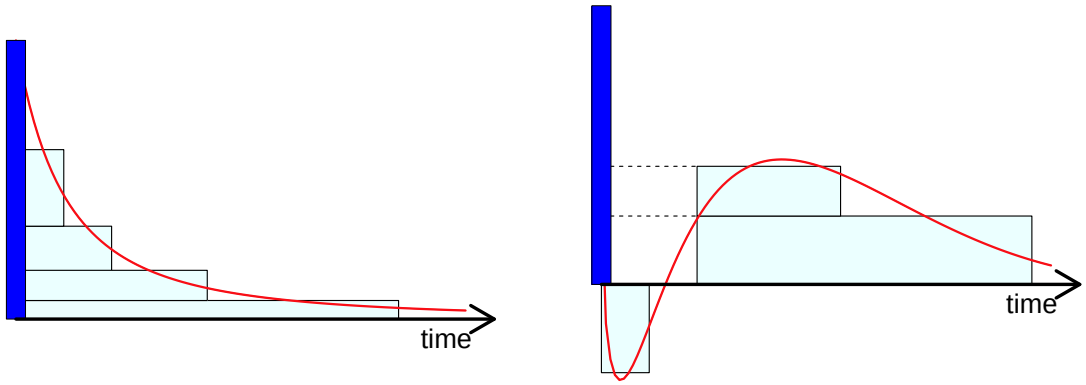


Figure 5.1: Examples of time-varying log-intensities approximated with staircase functions. Each blue-colored bar represents one spike by a stimulus or response event. Each red-colored curve approximated with a stack of rectangles is a dynamic log-intensity of response. Impacts of a marketing action typically decay over time like the curve on the left, while the spike by a purchase could give the curve on the right. Because a consumer who satisfied immediate demands waits for a while until the next purchase, the short-term impact can be negative while the mid- or long-term impacts become positive thanks to customer satisfaction.

the short-term memory. By nonparametrically fitting many types of heterogeneous staircase functions whose shapes depend on the type of each event, we are able to precisely predict how each of the marketing advertising communications stimulates short-term and mid-term memories of consumers. Note that long-term memory could be incorporated in principle, but is actually represented as the bias term due to the limited length of real time-series datasets.

In modeling the positive feedback among consumers, a crucial problem specific to our datasets is how to adequately estimate the influences of the unobserved word-of-mouth, by only handling sequences of observable event spikes. One straight formulation for incorporating word-of-mouth is to model hidden social graphs with Ising models (e.g., (Ravikumar et al., 2010)), graphical Granger models (e.g., (Lozano et al., 2009)), Poisson network or continuous-time Bayesian networks (Nodelman et al., 2002). These graph-based approaches, however, suffer from the lack of training data due to usual rareness of word-of-mouth as a consequence of low coupling strengths for most pairs of consumers. In our preliminary experiments, graph-based approach to detect existences of edges resulted in the pruning of most edges and underestimation of word-of-mouth effects. In addition, graph-based algorithms essentially have quadratic computational costs to population of consumers, and hence they are computationally inapplicable for lots of consumers.

For the stability in estimation, we regard the positive feedback among the interacting consumers as calibrated by aggregating the frequencies of the response events. As we illustrate in Figure 5.2, the aggregation of the response frequencies is done for each group of the consumers that are expected to mutually interact with one another. We refer to each of the aggregated frequencies as a collective factor, which is formalized as another multi-scale mixture to reflect the time-decaying nature of the positive feedback. Here one additional challenge is how to automatically detect such mutually-interacting groups over which the response frequencies are summed. Our key idea in this automatic grouping is the clustering of *residuals* in time-series regression. Using the initial estimate of the Poisson regression without the collective factor, we compute a time-series of regression residual that is the difference between the expected and actual frequencies of response. We finally fit group-dependent Poisson regression models whose feature variables contain both the individual factors and the aggregate response

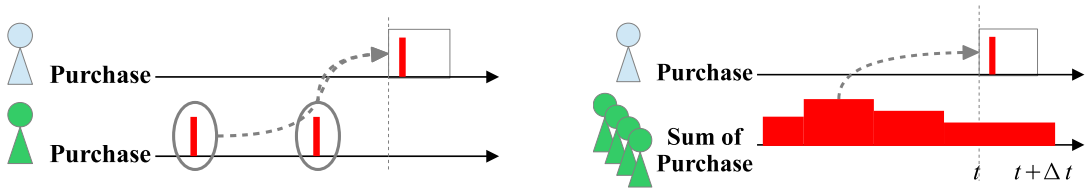


Figure 5.2: Stabilizing the fitting by aggregating the frequencies of response events. In the left figure, the limited number of spikes per one consumer makes the estimation of interaction unreliable. In contrast, the aggregation of the response frequencies among many consumers are robust indicators of social trend, and the detection of the interaction becomes more resistant to the over-fitting. The aggregation should be done within a group of consumers who are interacting to one another.

frequencies each detected group supplies. Because the centroid of each cluster represents a hidden but common trend that is unpredictable merely from the same individual's experiences, our collective-factor modeling handles the mutual interaction more robustly than the graph-based approaches using pairs of individual consumers.

High predictive accuracy of the proposed approach is empirically validated by using real-world data provided from an online retailer in Europe. The main outcomes of this work were first published in (Takahashi et al., 2013), from which we borrowed the formulations, figures, and implications based on the permission for authors[1]. Section 5.1 introduces the piecewise-constant Poisson regression without the collective factor, and Section 5.2 extends the model to contain the collective factor. Section 5.3 gives experimental results on real-world datasets provided from an online retailer, and these results become the basis of the generalizing discussion in Section 5.4. Section 5.5 summarizes this chapter.

5.1 Poisson Regression with the Individual Factor

This section introduces basic concepts and notation of our continuous-time regression problem for predicting future response events from the past events. Section 5.1.1 identifies the several sub-tasks required for continuous-time probabilistic event prediction. In Section 5.1.2, we introduce an algorithm for computing vector time-series of mental states for each consumer. In Section 5.1.3, we introduce a piecewise Poisson log-likelihood when all consumers share the same parameters, and discuss a Maximum A Posteriori estimation of the parameters.

5.1.1 Continuous-Time Response Regression

We first introduce basic notations for event sequences. There are K_R types of response events such as purchase orders and browsing of web pages, and K_S types of marketing stimulus events such as e-mail promotions. K_S is usually the number of marketing-communication channels. We define each event e with a triplet $(k[e], t[e], v[e])$, where $k[e] \in \{1, 2, \ldots, K_R + K_S\}$ is an index to represent the type of the event, $t[e] \in \mathbb{R}$ is a continuous time-stamp, and $v[e] \in \mathbb{R}$ is the amount of stimulus or response. Event e is a response event when $1 \leq k[e] \leq K_R$, or is a stimulus event when $K_R+1 \leq k[e] \leq K_R+K_S$. For every consumer indexed as $i \in \{1, \ldots, n\}$, we have a sequence of events $(e_{i1}, e_{i2}, \ldots, e_{iE_i})$ whose length is E_i. Since only the purchase events involve real-valued response amounts, $v[e]$ is the sales revenue if e is a purchase event, or $v[e]$ is 0,1-valued otherwise (e.g. web-browsing).

[1] See the IEEE copyright and consent form `http://www.ieee.org/documents/ieeecopyrightform.pdf`.

As in Figure 2.2, using all or parts of the stimulus and response events before time t, we wish to probabilistically predict whether a response event of type $k \in \{1, \ldots, K_R\}$ occurs in right half-open interval $[t, t+\Delta t)$. Let $\boldsymbol{x}_i(t) \in \mathbb{R}^{d_X}$ be a vector to represent the mental state of consumer i at time t, which is called the state vector and which is determined with the events before time t. A random variable $Y_{ik}(t, t+\Delta t)$ represents the number of type-k responses by consumer i in period $[t, t+\Delta t)$, and consumer i's response probability at time t is assumed to be a function of the state vector $\boldsymbol{x}_i(t)$ for every response-type k. When Δt is sufficiently small, the value of $Y_{ik}(t, t+\Delta t)$ is either one or zero and all of the elements in the state vector $\boldsymbol{x}_i(t)$ are constant over the period $[t, t+\Delta t)$. Hence we model the time-varying occurrence probability of type-k response as

$$P\left(Y_{ik}(t, t+\Delta t) \geq 1\right) = \Delta t \exp\left(b_k + \boldsymbol{w}_k^\top \boldsymbol{\Phi}\left(\boldsymbol{x}_i(t)\right)\right), \tag{5.1}$$

where $\boldsymbol{\Phi} : \mathbb{R}^{d_X} \rightarrow \mathbb{R}^{d_\Phi}$ is a high-dimensional mapping function applied to the state vector $\boldsymbol{x}_i(t)$, b_k is a bias term, and $\boldsymbol{w}_k \in \mathbb{R}^{d_\Phi}$ is a vector of regression coefficients. The mapping function $\boldsymbol{\Phi}$ incorporates non-linear relationship between the state vector and the logarithm of the response probability. For convenience, we define a function of time t as

$$z_{ik}(t) = \lim_{\Delta t \rightarrow 0} \log \frac{P\left(Y_{ik}(t, t+\Delta t) \geq 1\right)}{\Delta t}, \tag{5.2}$$

which is called the log-intensity function. The log-intensity function $z_{ik}(t)$ is the logarithm of the time-varying intensity in our inhomogeneous Poisson processes.

We will now introduce the "data mining" tasks required to realize the stochastic response-event prediction. We first need to design an efficient algorithm to compute the state vector $\boldsymbol{x}_i(t)$ adequate for realizing high predictive accuracy. We then need to estimate the vector of coefficients $\boldsymbol{w}_k$ for every response-type k, while introducing an effective mapping function $\boldsymbol{\Phi}$. The training data is a set of observed sequences $\{(e_{i1}, e_{i2}, \ldots, e_{iE_i})\}_{i=1}^n$. Once all of the parameters in (5.1) as well as the algorithm to compute the state vectors are specified, the impacts of any marketing-action sequence can be forecasted by iterating the computation of the state vectors and simulation of the next response for every consumer.

5.1.2 Variable-Interval State Vectors

Since humans have limited memory and tend to forget inessential information, the intensity of any given spike necessarily decays over time. The state vector $\boldsymbol{x}_i(t)$ needs to be designed to effectively model the time-decaying intensity, where each event-type k can have a different magnitude of intensity and decay-time.

Modeling the intensity with piecewise-constant features corresponds to an approximation with staircase functions, as shown in Figure 5.1. For a type-k' spike to consumer i, let $f_{ik'}^{(k)}(\tau) \in \mathbb{R}$ be its impact to the future type-k response, after time τ has passed since the time of the type-k' spike. We model $f_{ik'}^{(k)}(\tau)$ as

$$f_{ik'}^{(k)}(\tau) = \sum_{h \in \mathcal{H}} \beta_{ik'h}^{(k)} I(\tau < h),$$

where $I(\cdot)$ denotes the indicator function, $\mathcal{H} = \{h_1, \cdots, h_{|\mathcal{H}|}\}$ is a set of allowed window lengths whose multiple characteristics scales are prepared for incorporating the long-range dependence, and $\beta_{ik'h}^{(k)} \in \mathbb{R}$ is a weight for a pair of type k' and window length h. Note that, given the maximum

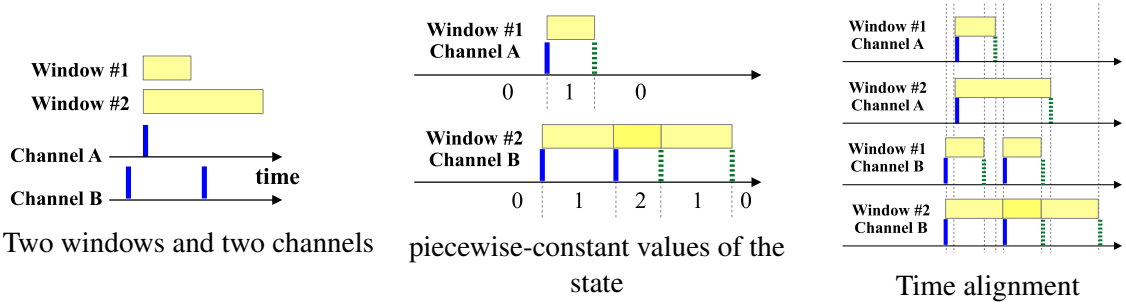


Figure 5.3: The variable-interval time-series of state vectors. A blue-colored bar represents an event spike and each dashed and green-colored bar represents a "terminator" of a stimulus associated with one window. In the left-most figure, we shown an example in which two types of windows #1 and #2 are applied to each of the two stimulus channels A and B. As the center figure clarifies, each element of the state vector is incremented or decremented at the time of a stimulus or a terminator, respectively. The right-most figure shows a time-series of four-dimensional state vectors given by the time alignment, when the two windows and the two channels are all combined. The state-vector time-series is piecewise constant with varying time intervals. At each change point, at least one element of the state vector varies with a jump.

window-length $h_{max} \triangleq \max\{h \in \mathcal{H}\}$, the intensity at time t with (5.3) is solely computed with the events in period $[t-h_{max}, t)$.

When the impacts by multiple events are additive, the log-intensity function becomes

$$\begin{aligned} z_{ik}(t) &= const. + \sum_{e:t[e]<t} f^{(k)}_{ik[e]}(t - t[e]) \\ &= const. + \sum_{e:t[e]<t} \sum_{k'=1}^{K_R+K_S} \sum_{h\in\mathcal{H}} \beta^{(k)}_{ik'h} I(t[e] \geq t-h). \end{aligned} \tag{5.3}$$

Here the state vector $\boldsymbol{x}_i(t)$ is defined as the sum of the indicator functions over pairs of type k' and length h, so that Equations (5.2) and (5.3) are identical if $\boldsymbol{\Phi}(\boldsymbol{x}) = \boldsymbol{x}$. The set of weights $\{\beta^{(k)}_{ik'h}\}$ corresponds to the vector $\boldsymbol{w}_k$. The idea behind the state vector computation is illustrated in Figure 5.3, where we primarily calculate a variable-interval sequence of each element of the state vector separately, by picking up only the relevant spikes. We then take time-alignment among all of the elements and generate a sequence of the state vector, capturing those time-stamps in which at least one element changes. The overall procedure is summarized in Algorithm 4.

Let us make a few remarks on the structure of our model when the mapping function $\boldsymbol{\Phi}$ is non-linear. A degressive mapping function for each element of the state vector would represent diminishing returns of stimuli within the same type. Higher-order (e.g., quadratic) terms would represent synergy or cannibalization among different types of stimuli.

5.1.3 Piecewise-Constant Poisson Regression

In each variable interval identified by Algorithm 4, we show the likelihood function of the associated Poisson regression. Let $[t_{i0}, t_{i1}), [t_{i1}, t_{i2}), \ldots, [t_{i(T_i-1)}, t_{iT_i})$ be the sequence of right half-open intervals of length T_i, such that values of $\boldsymbol{x}_i(t)$ do not vary within $t \in [t_{i(j-1)}, t_{ij})$ but do change at $t = t_{ij}$.

Algorithm 4 Computing state-vectors for consumer i

input Set of event-type and window-length pairs $\mathcal{R} = \{k, h\}$ with index $\pi[k, h]$, event sequence $\boldsymbol{e}_i \triangleq (e_{i1}, \ldots, e_{iE_i})$

output State-vector sequence $\{[t_{i(j-1)}, t_{ij}), \boldsymbol{x}_{ij}\}_{j=1}^{T_i}$

function $\{[t_{i(j-1)}, t_{ij}), \boldsymbol{x}_{ij}\}_{j=1}^{T_i} = GenStateVec(\mathcal{R}, \boldsymbol{e}_i)$
  for $k = 1$ **to** $K_R + K_S$ **do**
    $\boldsymbol{e}_k := ()$
    for $j = 1$ **to** E_i **do**
      if $k[e_{ij}] = k$ **then** $\boldsymbol{e}_k := (\boldsymbol{e}_k, e_{ij})$
    end for
  end for
  for all $(k, h) \in \mathcal{R}$ **do**
    $(t'_{kh1}, \ldots, t'_{khT_{kh}}, x_{kh1}, \ldots, x_{kh(T_{kh}-1)})$:=*UniS*$(h, \boldsymbol{e}_k)$
    $u[k, h] := 1$
  end for
  $t_{i0} := -\infty$, $\boldsymbol{x}_{i0} := \mathbf{0}_d$, $j := 1$
  repeat
    $\boldsymbol{x}_{ij} := \boldsymbol{x}_{i(j-1)}$
    $t_{ij} := \min\{t'_{khu[k,h]}; \forall(h, k)\}$ s.t $t_{ij} > t_{i(j-1)}$
    for all $(k, h) \in \mathcal{R}$ **do**
      if $t_{ij} = t'_{khu}$ **then**
        ($\pi[k, h]$-th element of $\boldsymbol{x}_{ij}$) := $x'_{khu[k,h]}$
        if $u[k, h] < T_{kh}$ **then** $u[k, h] := u[k, h] + 1$
      end if
    end for
    $j := j + 1$
  until $t_{ij} = \max\{t'_{khu[k,h]}; \forall(h, k)\}$
end function
function $(t'_1, \ldots, t'_{2E}, x_1, \ldots, x_{2E-1}) = UniS(h, \boldsymbol{e})$
  Sort $(t[e_1], v[e_1]), (t[e_1] + h, -v[e_1]), \ldots, (t[e_E], v[e_E]),$
  $(t[e_E] + h, -v[e_E])$ as $(t'_1, v'_1), \ldots, (t'_{2E}, v'_{2E})$ with the ascending order of time
  $x_0 := 0$
  for $j = 1$ **to** $2E$ **do** $x_j := x_{j-1} + v'_j$
end function

Consider the case when $t \in [t_{i(j-1)}, t_{ij})$. Since the state vector is constant, $\boldsymbol{x}_i(t) \equiv \boldsymbol{x}_i(t_{ij})$, the value of the log-intensity is also constant, $z_{ijk} \triangleq \boldsymbol{w}_k^\top \boldsymbol{\Phi}(\boldsymbol{x}_i(t_{ij}))$. Let y_{ijk} be the observed frequency of type-k response, and $\tau_{ij} \triangleq t_{ij} - t_{i(j-1)}$. The discrete-time likelihood for consumer i's responses is

$$(\Delta t \exp(z_{ijk}))^{y_{ijk}} \left[1 - \Delta t \exp(z_{ij})\right]^{\left(\frac{\tau_{ij}}{\Delta t} - y_{ijk}\right)},$$

and the continuous-time ($\Delta t \to 0$) limit of its logarithm is

$$const. + y_{ijk} z_{ijk} - \tau_{ij} \exp(z_{ijk}). \tag{5.4}$$

Eq. (5.4) is nothing but the unnormalized log-likelihood of a Poisson distribution whose outcome is y_{ijk} and whose mean is $\tau_{ij} \exp(z_{ijk})$.

Based on (5.4), the predictive log-likelihood of the entire type-k response sequences is given by summing a finite number of terms. For a time-series set of the variable-interval state vectors and response frequencies $\mathcal{D}_k \triangleq \{(\tau_{i1}, \boldsymbol{x}_{i1}, y_{i1k}), \ldots, (\tau_{iT_i}, \boldsymbol{x}_{iT_i}, y_{iT_ik})\}_{i=1}^n$, the data log-likelihood is

$$\mathcal{L}(\mathcal{D}_k|\Theta_k) = \sum_{i=1}^{n} \sum_{j=1}^{T_i} \ell\left(y_{ijk}; b_k + \boldsymbol{w}_k^\top \boldsymbol{\Phi}(\boldsymbol{x}_{ij}), \tau_{ij}\right), \tag{5.5}$$

where $\ell(y; z, \tau) \triangleq yz - \tau \exp(z)$ and $\Theta_k \triangleq \{b_k, \boldsymbol{w}_k, \boldsymbol{\Phi}\}$ is the set of parameters and mapping function.

The parameters $(b_k, \boldsymbol{w}_k)$ are optimized with an L_1 regularization. Since Eq. (5.5) is concave with respect to the variables $(b_k, \boldsymbol{w}_k)$, we are able to perform a global optimization

$$\max_{b_k, \boldsymbol{w}_k} \left[\mathcal{L}(\mathcal{D}_k|\Theta_k) - nC_0\|\boldsymbol{w}_k\|_1\right], \tag{5.6}$$

where $\|\cdot\|_1$ denotes the L_1-norm and C_0 is a regularization hyperparameter. The L_1 penalty term encourages sparse estimates as in the Lasso regression (Tibshirani, 1994). We apply coordinate-wise updating for solving Optimization (5.6). In each iteration the objective is approximated with the sum of a quadratic function and the L_1 norm, and then the update is done paying attention to the indifferentiable points in the L_1 norm.

5.2 Introducing the Collective Factor

This section extends the model to contain the collective factor for modeling word-of-mouth and trend following, and addresses the estimation of mutually-interacting groups. Section 5.2.1 shows the limitation of aggregate-level prediction when only the individual-factor is incorporated, and evidences the existence of inter-consumer dependence in our dataset. The high dispersion of the actual responses, which is addressed in Section 5.2.1, suggests the benefits of detecting correlated groups, which is done with the clustering of the regression residuals as in Section 5.2.2. Section 5.2.3 introduces our final estimator to handle both the individual and collective factors, with exploiting the detected groups. Design of mapping functions and collective state is briefly provided in Section 5.2.4.

5.2.1 Over-Dispersion in Aggregate-Level Prediction

Using the estimates $(b_k = \widehat{b}_k, \boldsymbol{w}_k = \widehat{\boldsymbol{w}}_k)$ given by solving Optimization (5.6), we are able to predict the sum of the response frequencies for all of the consumers. Assume that each consumer responds independently from one another[2]. Since sum of Poisson random variables also obeys a Poisson distribution, the total response frequency from all consumers in period $[t, t+\Delta t)$ obeys a Poisson distribution whose mean is

$$\Delta t \sum_{i=1}^{n} \exp(z_{ik}(t)) \equiv \Delta t \sum_{i=1}^{n} \exp\left(\widehat{b}_k + \widehat{\boldsymbol{w}}_k^\top \boldsymbol{\Phi}(\boldsymbol{x}_{it})\right). \tag{5.7}$$

The minimum time-granularity of our dataset, whose details are provided in Section 5.3, is 1 day. Hence we are able to evaluate the validity of the assumption of the independence among different consumers, by using daily total responses over all consumers. Figure 5.4 shows an example of the actual and predicted frequencies of the total purchase orders, where we fit the parameters $(\widehat{b}_k, \widehat{\boldsymbol{w}}_k)$

[2]In the real world, responses among many consumers are still correlated even when word-of-mouth does not exist, because firms tend to send the same campaigns to similar customers.

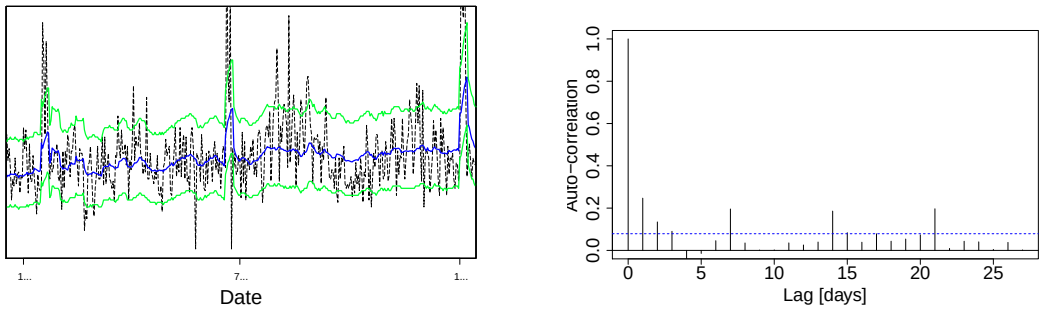


Figure 5.4: High dispersion from the sum of individual predictions. In the left figure, the vertical axes represent the daily frequencies of the purchase orders, where we omitted the absolute values for confidentiality reasons. The black dashed and blue solid lines specify the actual and the predicted values, respectively. The green solid lines represent the 0.5%- and 99.5%-tiles when Poisson distributions are assumed. While the predicted mean is reasonable with its somewhat conservative estimate, since the actual frequency is out of the confidence intervals in many days, rejection of the independence assumption is naturally implied. The right figure represents the autocorrelation of the regression residuals with respect to the days to represent the lags. The significant autocorrelation evidences the predictability of the residuals, which could be provided with mutual-interaction information among consumers.

using 10,000 consumer samples. In Figure 5.4, the actual frequency is over-dispersed against the confidence intervals. Because the standard deviation of a Poisson distribution is the square root of its mean, models whose aggregate-level prediction obeys a Poisson distribution much underestimate the variances of the total responses, particularly when the number of consumers is large. In addition, the statistically-significant autocorrelation of the regression residuals suggests the predictability of such residuals with autoregressive models, while the events for the same individual cannot more improve the predictability.

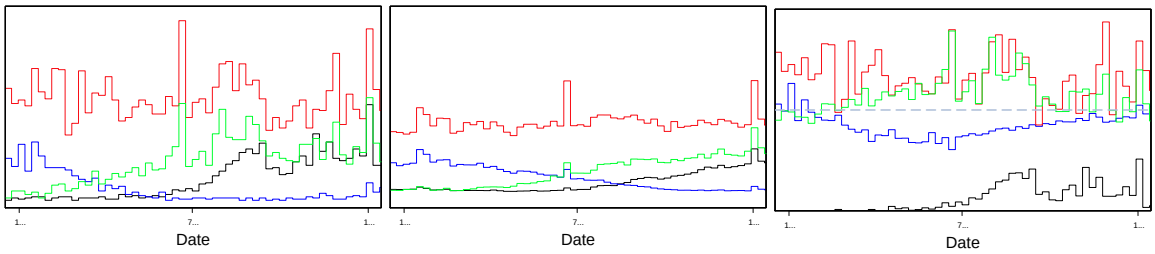
5.2.2 Clustering of the Residual Time-Series

The remaining autocorrelation in the aggregate level suggests a feedback between the total past responses summed over many consumers and the future response by each individual consumer. While simply incorporating the total responses as explanatory variables would improve the predictability, we intend to retrieve richer information by exploiting the large number of consumer samples. The residual in each individual's response prediction cannot be more regressed within the same individual's past events but could be regressed with other consumers' responses.

Let $\widetilde{\tau}$ be a unit time-length in aligning the entire period. We fix $\widetilde{\tau}$ to be 1 week in this paper, and introduce common time periods $[\widetilde{t}_0, \widetilde{t}_1), \ldots, [\widetilde{t}_{H-1}, \widetilde{t}_H)$ where $\widetilde{t}_j - \widetilde{t}_{j-1} \equiv \widetilde{\tau}$. Let $\widetilde{y}_{ijk}$ be the total frequency of type-k responses in period $[\widetilde{t}_{j-1}, \widetilde{t}_j)$ by consumer i. Then we define a residual variable

$$r_{ijk} \triangleq \widetilde{y}_{ijk} - \int_{\widetilde{t}_{(j-1)}}^{\widetilde{t}_j} \exp\left(\widehat{b}_k + \widehat{\boldsymbol{w}}_k^\top \boldsymbol{\Phi}\left(\boldsymbol{x}_{it}\right)\right) dt, \tag{5.8}$$

where the integral in (5.8) is actually given with a finite number of terms stemming from the piecewise-constant features. We define an H-dimensional residual time-series vector for each consumer i, as $\boldsymbol{r}_{ik} \triangleq (r_{i1k}, \ldots, r_{iHk})^\top$.



actual frequency predicted frequency residual

Figure 5.5: Time-series associated with 4 example clusters, whose members are inferred to share the common trends. In the left-most and the center figures, the vertical axes represent the weekly-smoothed actual and predicted values of the total response frequencies summed over the members belonging to each cluster. The right-most figure shows the residual which is the difference between the actual and predicted frequencies, where the center dashed line represents the zero frequency. Because persistent biases are confirmed for the residuals of some clusters, cluster-dependent autoregressive modeling is implied to improve the prediction capability.

We assume that consumers having similar residuals follow common social trends, and hence clustering of the residual vectors $\{\boldsymbol{r}_{ik}\}_{i=1}^{n}$ provides groups of mutually-interacting consumers. For m mutually-exclusive sets of consumers $\{\mathcal{S}_c \subset \{1,\ldots,n\}\}_{c=1}^{m}$, we expect significant mutual interaction existing between any pair of consumers in the same group $\mathcal{S}_c$.

While any clustering algorithm may be applied, we adopt m-medians clustering (Jain and Dubes, 1981) for robustly obtaining the groups against the outlying peaks. Figure 5.5 provides an actual example of the clusters for the same 10,000 consumers in Figure 5.4, where the actual, predicted, and residual time-series are plotted with the 1-week time granularity.

5.2.3 Heterogeneous Model with Mutual Interaction

The detected clusters in Section 5.2.2 are exploited in the final Poisson regression with heterogeneous parametrization. Our log-intensity function incorporating the collective factor is given as

$$z_{ik}(t) = b_{c[i]k} + \boldsymbol{w}_{c[i]k}^{\top} \boldsymbol{\Phi}\left(\boldsymbol{x}_i(t)\right) + \sum_{c'=1}^{m} \boldsymbol{\theta}_{c[i]c'k}^{\top} \boldsymbol{\Psi}\left(\boldsymbol{z}_{c'}(t)\right),$$

where $c[i]$ represents the cluster index for consumer i, $\boldsymbol{z}_c(t) \in \mathbb{R}^{d_Z}$ represents cluster c's collective state modeled with the total response frequencies, $\boldsymbol{\Psi} : \mathbb{R}^{d_Z} \to \mathbb{R}^{d_\Psi}$ is another high-dimensional mapping function applied to the collective state vector $\boldsymbol{z}_c(t)$, b_{ck} is a bias term for cluster c, $\boldsymbol{w}_{ck} \in \mathbb{R}^{d_\Phi}$ is a vector of individual factors and $\boldsymbol{\theta}_{cc'k}$ is a vector of coupling strength between the clusters c and c'.

Because the minimum time granularity is $\tau_0 = 1$ [day], we introduce a daily sequence of common time-periods $[t_0^*, t_1^*), \ldots, [t_{T^*-1}^*, t_{T^*}^*)$, and rewrite the data log-likelihood as

$$\mathcal{L}\left(\mathcal{D}_k | \Theta_k^*\right) = \sum_{i=1}^{n} \sum_{j=1}^{T^*} \ell\left(y_{ijk}^*; b_{c[i]k} + \boldsymbol{w}_{c[i]k}^{\top} \boldsymbol{\Phi}\left(\boldsymbol{x}_{ij}^*\right) + \sum_{c'=1}^{m} \boldsymbol{\theta}_{c[i]c'k}^{\top} \boldsymbol{\Psi}\left(\boldsymbol{z}_{c'j}\right), \tau_0\right), \tag{5.9}$$

where y_{ijk}^* is the frequency of type-k response by consumer i in period $[t_{j-1}^*, t_j^*)$, $\boldsymbol{x}_{ij}^* \equiv \boldsymbol{x}_i(t_{j-1}^*)$, $\boldsymbol{z}_{cj} \equiv \boldsymbol{z}_c(t_{j-1}^*)$, and $\Theta_k^* \triangleq \left(\{\{b_{ck}, \boldsymbol{w}_{ck}, \{\boldsymbol{\theta}_{cc'k}\}_{c'=1}^{m}\}_{c=1}^{m}, \boldsymbol{\Phi}, \boldsymbol{\Psi}\right)$. In order to attain the final estimate,

we maximize another penalized objective

$$\mathcal{L}\left(\mathcal{D}_k|\Theta_k^*\right) - \frac{n}{m}\sum_{c=1}^{m}\left(C_1\|\boldsymbol{w}_{ck} - \widehat{\boldsymbol{w}}_k\|_1 + C_2\sum_{c'=1}^{m}\|\boldsymbol{\theta}_{cc'k}\|_1\right), \tag{5.10}$$

with respect to the parameters $\left(\{\{b_{ck}, \boldsymbol{w}_{ck}, \{\boldsymbol{\theta}_{cc'k}\}_{c'=1}^m\}_{c=1}^m,\right)$, given the regularization hyperparameters (C_1, C_2). The objective (5.10) involves a multi-task learning aspect because sharing of the coefficients among clusters are encouraged. Maximization of (5.10) is a convex optimization problem and hence we are able to obtain the global optimum.

5.2.4 Feature Designs

Here are notes on designing the collective states and the mapping functions. For each of the collective states $\{\boldsymbol{z}_c(t)\}_{c=1}^m$, we contained the past 1-day, 2-day, 4-day, and 1-week response frequencies summed over group c by following our philosophy of multi-scale basis functions, and hence $d_Z = 4m$. We utilize a sub-linear mapping function $\Phi_u : \mathbb{R} \to \mathbb{R}$ as $\boldsymbol{\Phi}(x_1, \ldots, x_d) = (\Phi_1(x_1), \ldots, \Phi_d(x_d))^\top$ and $\boldsymbol{\Psi}(z_1, \ldots, z_{d_Z}) = (\Phi_1(z_1), \ldots, \Phi_{d_Z}(z_d))^\top$. Following the prior art (Rajaram et al., 2005), we adopt $\Phi_u(x) = \log(1 + x/h(u))$ where $h(u)$ is the length of sliding window associated with the u-th feature.

When the input of the function Φ_u is a frequency of events, Φ_u handles the marginally diminishing return of the same stimuli. Due to the intensive overall computational cost, presently we omitted quadratic terms to represent synergy or cannibalization. Degressive functions performed well empirically, and we leave the development of an efficient algorithm for handling higher-order terms as future work.

5.3 Experimental Evaluations

We perform validation experiments with large real world datasets, which are supplied by an online retail company in Europe, and whose characteristics is introduced in Section 5.3.1. For intuitively evaluating the correlation between the predicted response intensity and the actual occurrence of the response events, we introduce an evaluation metric called the "Continuous-Time Area-Under-Curve" (CTAUC) in Section 5.3.2. The first experimental results in Section 5.3.3 evidence the benefits of containing many events to form the individual factor, and provide some example shapes of the fitted log-intensity functions, which are valuable outcomes for marketing implications. The second experiment in Section 5.3.4 clarifies the additional advantages yielded by the collective factor. In all of the experiments, we split the entire datasets into training and test datasets, using the middle date in the entire periods as the dividing point. Hyperparameters were chosen by a hold-out method to further separate the training datasets into two subsets. The regularization hyperparameter C_0 in (5.6) is chosen from $\{10^{-5}, 10^{-4}, \ldots, 10^2\}$, while we adopted a grid search to choose (C_1, C_2) in (5.10) for $C_1 \in \{1, 10, 10^2\}$ and $C_2 \in \{1, 10, 10^2, 10^3\}$.

5.3.1 Individual-Level Datasets

We use individual-level daily records of marketing-action and response events for 2 years from 2009 to 2011. We have 3 datasets, where each dataset contains about 2 million events, including about 900 thousands browsing events and about 60 thousands purchase events produced by 30,000 customers. There are $K_S = 10$ types of responses, including 1 type of purchase order and 9 types of browsing

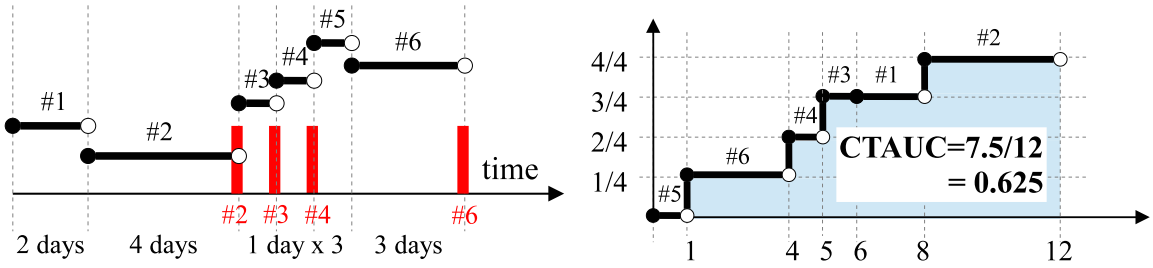


Figure 5.6: Principle in computing the CTAUC. The line segments show the predicted response intensity and the red-colored bars are actual response events. Each response event could occur in the terminal time of a right half-open interval, because occurrence of the response changes the following intensity. In the right figure, the CTROC curve is composed with the permutation of the periods based on the descending order of the response intensity.

events reaching a commerce web-cite. The 9 types in classifying the browsing are based on the referral links from various external pages. The number of marketing-action channels is $K_S = 5$, including e-mail promotions and other message materials.

For each of the $(10+5)$ event types, we prepared $|\mathcal{H}| = 9$ types of the multi-scale window lengths including 1 day, 2 days, 4 days, 1 week, 2 weeks, ..., and 32 weeks for the individual state vectors. Elements of the state vector corresponding to the purchase events contain both the frequency of purchase and the real-valued sales amount. Hence the input dimensionality is $d_X = 15 \times 9 + 9 = 144$. Due to the finite length of the entire periods in the datasets, we do not predict the responses in the first 32 weeks during which we cannot have the correct values of the states. Still the events in these first 32 weeks are incorporated in the states used for the prediction for time stamps after 32 weeks from the first date.

5.3.2 Performance Metrics

Here we employ two types of predictive performance metrics. The first is a Continuous-Time Area-Under-Curve (CTAUC), for calibrating the correlation between the predicted response intensities and actual occurrence of the response events. The second is an average log-likelihood which devides Eq. (5.5) by the total length $\sum_{i=1}^{n}(t_{iT_i} - t_{i0})$.

Remember that the state vectors and their corresponding response intensities are constant in certain time periods. We permute these periods with the descending order of the response intensities. We then compute what fraction of the actual response events are covered in the periods that are assigned high intensities, as the Continuous-Time Receiver-Operator-Characteristics (CTROC) curve. The CTAUC is the ratio between the area under the CTROC curve and the overall sum of the periods. We consequently compute the total CTAUC by assembling all of the time periods for all of the consumers, where the permutation is done only with the intensities and not with the ID of each consumer.

5.3.3 Basic Performances Achievable with the Individual Factor

We first validated the advantage attained by incorporating multiple types of events as covariates, and confirmed the detection of event-dependent time-decaying curves. Here the predictive accuracy on the final purchase order is evaluated without the collective factor, and hence the fitted model is a

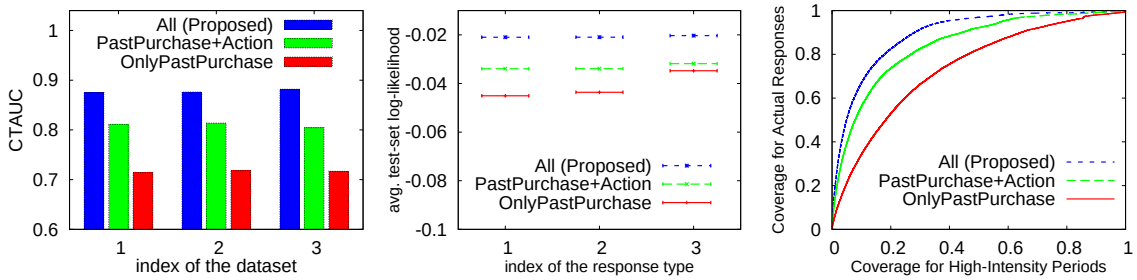


Figure 5.7: Comparing performances for the inclusion of covariates. The left-most and center figures show the CTAUC and average log-likelihood for each test dataset, respectively. The right-most figure shows the actual CTROC curves in the first dataset. For all of the 3 datasets, the model containing all types of events as the covariates is most outperforming, where over-80% of the actual responses are covered in the top-20% high-intensity periods.

variation of Piecewise-Constant Conditional Intensity Model (Gunawardana et al., 2011). Then we also inspected the fitted log-intensity functions to obtain insights on the nature of effects yielded by various types of marketing actions.

Figure 5.7 exhibits the results for three types of models whose covariates are controlled for comparison. The first model emulates the self-exciting IPPs that include auto-correlation only among the events of the same type, where we included only the past purchase events in the state of the first reference model. The second model emulates regression methods to directly predict the impacts of marketing actions, without considering intermediate phases. Here the state contains both the marketing-stimulus and purchase events, but does not contain browsing events. The third model is the proposed model containing all types of the events as its covariates. The results in Figure 5.7 evidence the significant advantage in predictive accuracy with handling many types of events in modeling.

Figure 5.8 exhibits several example "shapes" of the fitted log-intensity functions. Note we omitted the absolute values due to confidentiality considerations. Insights are obtained from these relative magnitudes as a function of the elapsed time and event type. As expected from the hyperbolic discounting theory, we are able recognize the long-term tails in the intensity. These fitted shapes evidence the high predictive capability of the proposed nonparametrics to fit such long-term tails, without directly imposing strict power-law parametrics.

As demonstrated in Figure 5.8, marketers must consider whether a marketing action stimulates the direct purchase that could later become self-exciting, or intermediate consumer activities that are equally, and sometimes more, self-exciting. Visualization of the time-varying intensity helps decision making in such a complex environment, and precisely-fitted intensity functions enable quantitative forecasting of outcomes in the presence of a large variety of marketing stimuli.

5.3.4 Gains Obtained with the Collective Factor

We next validated the additional advantage attained by incorporating the collective factor. Figure 5.9 exhibits the comparisons among the collective-factor models with $m \in \{1, 16, 64\}$ and the individual-factor only model. While the model with $m = 1$ incorporates only the total response frequency without clustering, the basic performance gain by adding the collective information is the most significant. Gains by adding residual clustering depend on the response-type labels, but we can confirm many-cluster models are outperforming for considerable portions of the labels, in terms of log-likelihood.

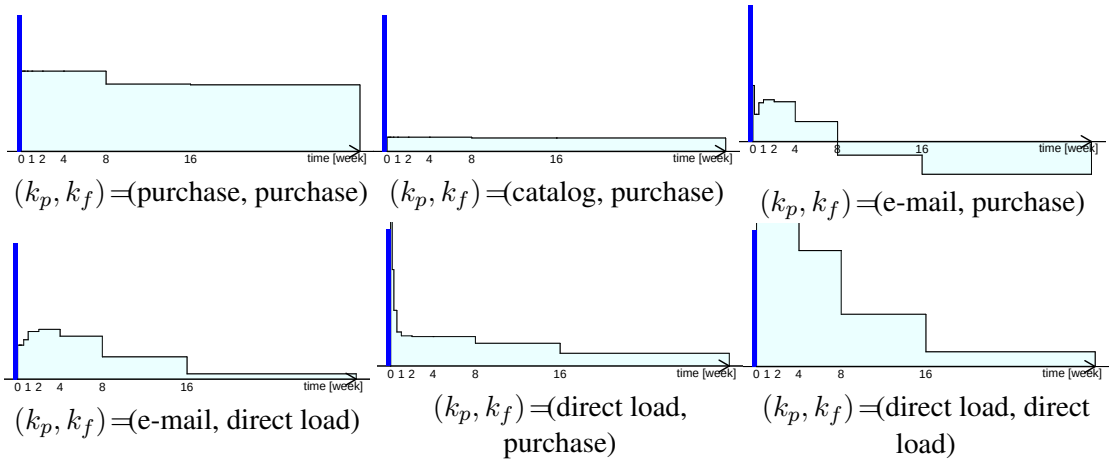


Figure 5.8: Examples of nonparametrically fitted log-intensity functions, when type-k_f response event is predicted from a past type-k_p event. It is seen that purchase events are strongly self-exciting, and influence by a purchase experience is larger than those by exogenous marketing stimuli. Sending a rich catalog of products does not impact in short-term, but yields certain outcomes in the longer-term. In contrast, an e-mail strongly stimulates purchase in short-term but its negative longer-term impact is striking. Graphs in the second row are for the intermediate responses named 'direct load'. These are browsing events without any link from other pages or e-mail. Even when the embedded web-links are ignored by customers, an e-mail has long-term positive impact to the direct load events, which in turn have positive impact on the final purchase and exhibit high autocorrelation. Hence, the long-term marginal return of an e-mail is seen to be the compound outcome of its negative direct impact and positive indirect impact involving a chain of self-exciting direct load events.

We infer that log-likelihood's more sensitivity to the predictability of rare word-of-mouth effects is the main reason of the relative advantages. While the numbers of clusters are fixed for easy comparisons, automatic choice of the number of clusters such as Dirichlet Process Mixtures (e.g., (Blei and Jordan, 2006)) would enhance the acquired gains. Choice from the variety of possible penalty terms in multi-task learning is also a considerable additional work.

5.4 Discussion

Given each of the validated significances about the explanatory variables used in the IPP regression, we discuss what are the essential approaches to make our continuous-time IPP model practical in many marketing decision-making problems. As shown in Section 5.3.3, accurate modeling of the individual factor was shown to significantly contribute for the performance improvements. Therefore, having richer structure of the state vector $\boldsymbol{x}_i(t)$ and the mapping function ϕ is a key to further improve the predictive capabilities. Because there would be more types of marketing stimuli and responses in the entire marketing operations, efficient design of the non-linear mapping function ϕ, which accompanies an efficient dimensionality reduction to handle lots of the event types, is a practical direction to improve the quality of the model. We consider that efficient implementation of such non-linear mapping functions would be provided by random forest (Breiman, 2001) or boosted regression trees (Freund and Schapire, 1997; Schapire and Singer, 1998; Friedman, 1999).

We regard time-dependent and robust modeling of the collective factor as other essential parts to

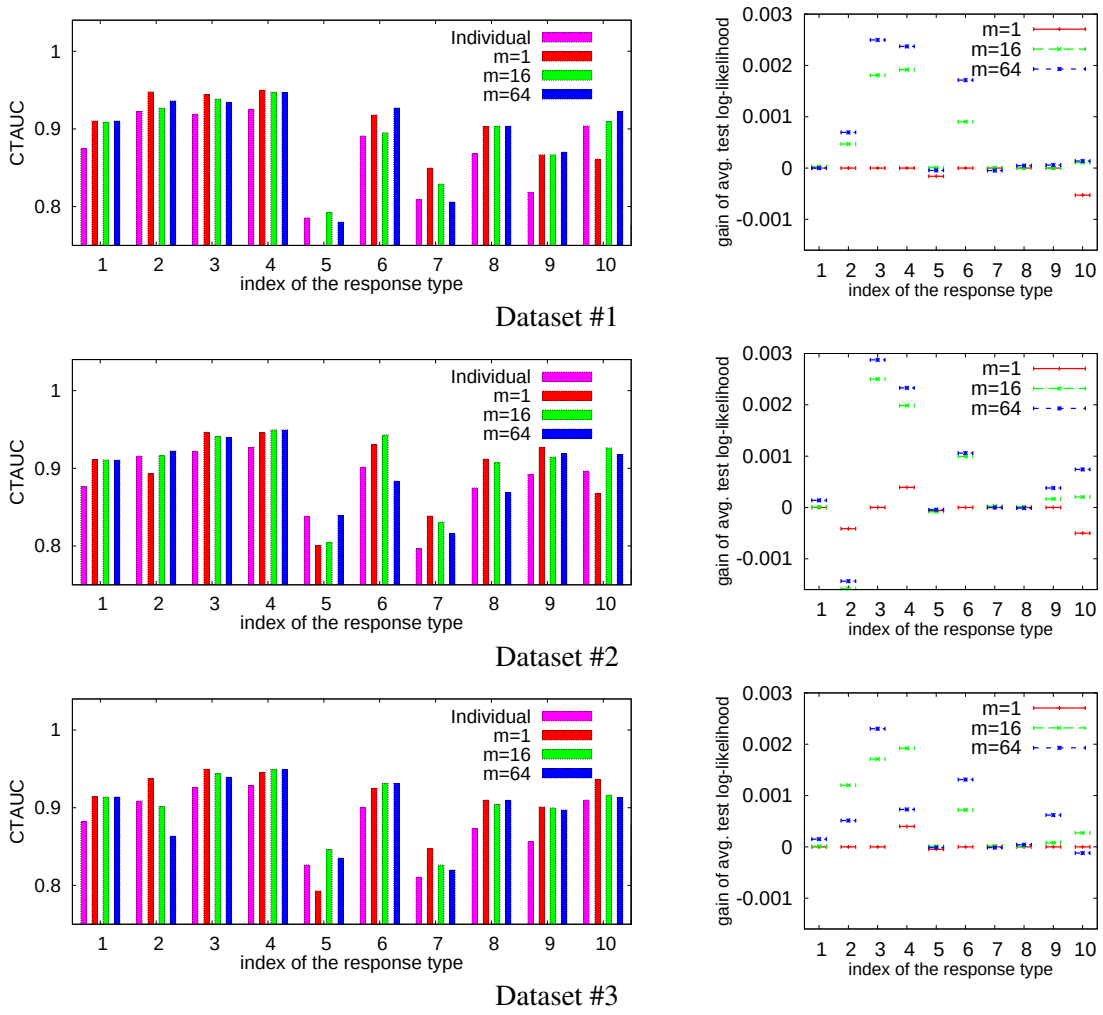


Figure 5.9: Label-dependent performances for each collective-factor model. Because the scales of log-likelihood are different among each response-type label, the relative gains of the test log-likelihood from the individual-factor model are shown. Simple feedback with $m = 1$ successfully works in many labels, while many-cluster models sometimes outperform others, particularly in terms of log-likelihood.

derive further useful implications. For more microscopic implications, we should let the structure of interactions be richer than that of the current model. The current cluster-based model assumes a block-diagonal adjacency matrix to represent the interaction among consumers, given the limitation of the amount of training data. Therefore, new estimation algorithms that are more resistant to the over-fitting will allow us for introducing more numbers of model parameters, that result in lower bias and higher generalization capabilities. One essential nature to be modeled is the non-stationary of the interactions among consumers, where the coupling strengths among consumers are low in most of the entire days, while suddenly get higher in a few days. A new parametrization of such non-stationary covariance structures, whose estimation is robust against the lack of training data, is the most crucial component for deriving feasible viral-marketing strategies.

5.5 Summary

This chapter introduced a novel continuous-time IPP model for marketing decision-making, in order to predict response spikes as compound outcomes of marketing actions and self-exiting consumer activities. The proposed model incorporates both of the individual and the collective factors, which are the total influence of the past events one consumer has experienced and the influence of the positive feedback among interacting consumers, respectively. The individual factor is implemented as an interpolation of the staircase functions that are nonparametric multi-scale mixtures of step functions, and provides clear psychological interpretation about how each of the marketing-communication events affects the future responses by each individual. The collective factor, for handling social word-of-mouth and trend-following effects, is formulated with the aggregation of the frequencies about response events within each mutually-interacting group of consumers. The detection of each group is automatically done with clustering of time-series regression residuals, and the aggregated frequencies are again processed with another nonparametric mixture also involving multiple scales. High predictive accuracy of the proposed approach is empirically confirmed using real-world datasets, and the fitted intensity function for each event-type is seen to provide actionable marketing insights.

As well as predicting the impacts of certain marketing-campaign strategies, we exploit the fitted parameters in the proposed model for rational marketing-decision making. As we have seen in this chapter, the staircase approximation of the time-decaying curves yields the finite-dimensional state vector of each consumer, and we are able to consider the *optimal* marketing decision making depending on each of these state vectors. The next Chapter 6 provides an approach of normative marketing decision making that is based on constrained Markov Decision Processes, when we are able to assume the correctness of the fitted parameters and the insensitivity of consumer behaviors against our decision making process to optimize the targeting rules. Given the experimental evidence that most of the predictive capabilities are provided by the individual factor, we focus on the interactions among multiple memories while accepting a compromise to ignore the influences of the positive feedback among consumers.

Chapter 6

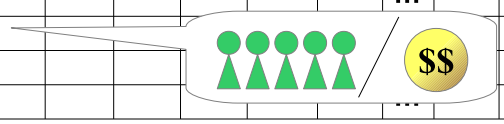
Normative Marketing-Mix Optimization Using the Nonparametric Forecasts

The nonparametric descriptive model introduced in Chapter 5 enables the forecasts of the total responses in the future, based on discrete event simulation whose input and output are sequences of marketing-stimulus and response events for each individual, respectively. Such forecasting process is embodied as sampling from marked point processes (Jacobsen, 2006), which involve the prediction of response amounts, as well as the per-time occurrence probability discussed in Chapter 5. As a real-world example of exploiting the nonparametric forecasts for normative and rational decision making, this chapter introduces a marketing-mix optimization algorithm. Marketing-mix optimization and budget allocation, which also accompany a finely detailed execution plan to target each of the many individual consumers, are ones of the essential requirements in today's complex marketing planning processes. A rational strategy to target individual consumers is realized as a state-based policy used in constrained Markov Decision Processes (cMDPs), across various channels and target segments, along with the timing of each marketing action. For feasibility of the executions in real operations, we define every segment of consumers, which we call the Micro Segment (MS), by discretizing the continuous state-vector space introduced in Chapter 5. Here the multi-scale nature of the sliding windows, which were used in defining the state vectors, again plays an essential role in deriving an efficient MS-based marketing-mix policy. With assuming the accuracy of the nonparametric descriptive model, we solve a multi-period optimization problem for marketers to obtain the maximum mid-term profit with leading each consumer's dynamic state into the desirable directions. One essential assumption in this application is that consumers are not adversarial, i.e., they do not act against our optimizing strategy but simply follow the input-output relationship predicted by the nonparametric models.

As we illustrate in Figure 6.1, the decision variables in our optimization problem consist of the target population for every triplet of a timing, a segment of consumers, and a marketing-communication channel. In each of the marketing-communication channels such as e-mail and tele-marketing channels, we are able to target every individual consumer with some execution costs. While the standard MDPs provide the optimal probability of each action for each segment, in the context of marketing decision making with budget constraints, we optimize the number of consumers (Abe et al., 2004; Tirenni et al., 2007b; Abe et al., 2009) targeted with each action, because the total marketing cost is a non-decreasing function of the number of targeted consumers.

One specific issue to generate executable policies is to properly incorporate a set of constraints imposed in real marketing operations. Targeting of many customers requires an appropriate planning of the marketing costs, in advance to the execution of contact operations. The discrepancy between

	2014/01/01			2014/01/08			⋯	2014/12/31		
	EM	DM	TM	EM	DM	TM	⋯	EM	DM	TM
Segment #1							⋯			
Segment #2										
⋯										
Segment #N							⋯			



EM: e-mail DM: direct mail TM: tele-marketing

Figure 6.1: Target populations as the decision variables in marketing decision making. Every cell in the table is conditional on a triplet of a timing, a segment of consumers, and a marketing-communication channel. Before the beginning date of the entire period (e.g., 1 year), we must plan the target population and its resulting budget amount for all of such triplets. The entire marketing cost is a function of such set of the target populations.

Max Budget	From 2014/01/01 To 2014/06/30			From 2014/07/01 To 2014/12/31		
	EM	DM	TM	EM	DM	TM
GOLD	$400,000		$150,000	$600,000		$200,000
SILVER						
NORMAL	$200,000			$200,000		

Max Budget	From 2014/01/01 To 2014/06/30			From 2014/07/01 To 2014/12/31		
	EM	DM	TM	EM	DM	TM
GOLD	$1,000,000.00					
SILVER	$500,000.00					
NORMAL	$200,000.00					

Figure 6.2: Structure of the complex budget constraints visualized as tables. For example, in the first half of 2014, the total e-mail and direct-mail budget spent for the `GOLD` and `SILVER` Strategic Segments (SSs) must not exceed $400,000. Another example of the budget constraint is imposed for the annual budget for the `GOLD` SS, whose maximum is $1M. The optimization algorithm must comply with all of these budget constraints involving complex combinations of the timings, segments, and channels.

the planning and execution dates necessitates a careful design of the optimization algorithm. Contact executions and budget planning are often done by different decision makers, where the targeting rules should be designed based on detailed profiles of consumers whilst their budget support must satisfy higher-level financial constraints over multiple executions. As shown in Figure 6.2, each of the required budget constraints is given as the maximum, or sometimes the minimum, of the costs summed across multiple time-stamps, segments, and channels whose combination often becomes complex due to the organizational structure of marketing departments. We must remember the fact that interest of most marketers is not limited to a single run of optimization but includes a What-If analysis, which runs the optimization processes numerous times with changing segmentation criteria and budget constraints.

We provide a framework to fairly evaluate how constraints and segmentation criteria affect the resulting policies and their economic impacts. As illustrated in Figure 6.3, the approximation of the original continuous state vector by the discrete MS motivates us to separate the process of optimizing the policy from that of simulating its impacts. To the best of our knowledge, the expected revenue gains in all of the cited literature heavily rely on the estimates of MDP parameters. Since the accuracy of the predicted revenue depends on segmentation, profit by some segmentation and its resulting policy cannot be directly compared with those by others. Profits should rather be forecast with a unified simulation system whose setting is independent from the segmentation and optimization processes. Such separation allows for both usage of finely detailed sales prediction and effective simplification of the marketing-mix policy with high interpretability. Figure 6.4 summarizes our unified framework that combines machine learning, cMDP-based optimization, and discrete event simulation, for both

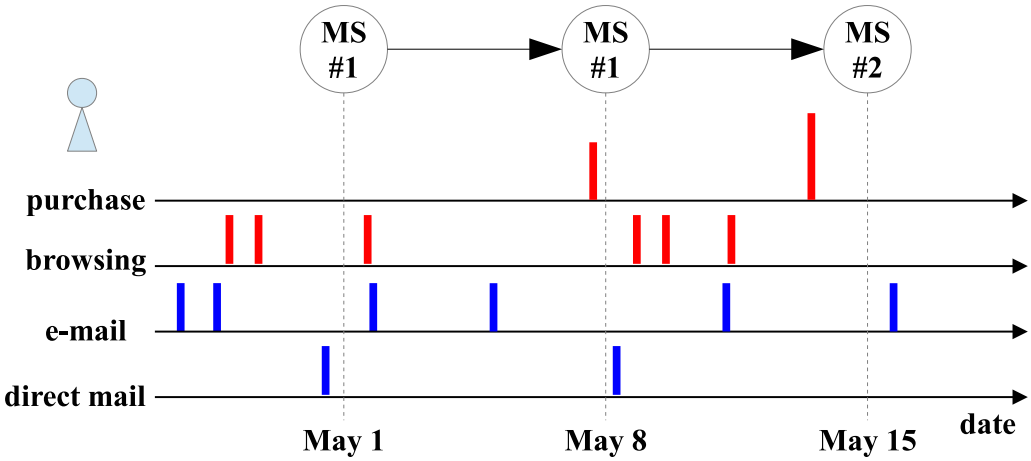


Figure 6.3: Marketing-stimulus (e-mail and direct-mail) and response (purchase and browsing) event spikes having continuous time-stamps, and an inferred sequence of Micro Segments (MSs) from these events. The true forecast of the individual-level sales should be based not on erroneous MS-level estimates, but on simulating every response event spike. In contrast, optimal targeting with an MS-based policy is justified for the purpose of operability and interpretability.

optimizing the policy and predicting its impacts. The final forecasts provided by a continuous-state simulation system, whose model parameters are precisely fitted with the machine learning algorithm, are more accurate than those by the approximately optimized objective based on the discrete MSs.

Our algorithmic contribution in solving the cMDPs is a new LP formulation to efficiently optimize the target population with satisfying budget constraints over multiple periods. In an MDP, each segment's population in the future is not deterministic but stochastic, where the dynamics is captured by an equality on the expected population for each segment in consecutive time points. The key ideas for efficiently optimizing such stochastic population is to introduce slack variables to represent the differences between the realized and expected populations, and to handle the state dynamics of MDP as probabilistic flow constraints. Then we solve a soft optimization problem that is completely linear in the target populations and the slack variables. This linearization enables the usage of very fast LP solvers and allows us for handling a large number of segments, channels, and many budget constraints in a unified interface.

Main formulations and experimental results in this chapter were borrowed from (Takahashi et al., 2014). Section 6.1 overviews the entire system and introduces the notations about our decision variables and the related variables behind the main optimization. The main solving of our LP problem is introduced in Section 6.2. In actual computation, before the solving of the LP, we need to statistically estimate the coefficients in the main objective and constraints. Section 6.3 summarizes the whole procedure for such parameter estimation, which consists of the fitting of the simulation model, discretization of the state space, and the computation of the coefficients associated with discrete MSs. Section 6.4 introduces a case-study using a real-world dataset provided by an online retailer in Europe, for clarifying the characteristic aspects of the optimized policies and the associated revenue gains. Section 6.5 discusses considerable extensions of our What-If analysis framework to more precisely emulating the real market, and Section 6.6 summarizes this chapter.

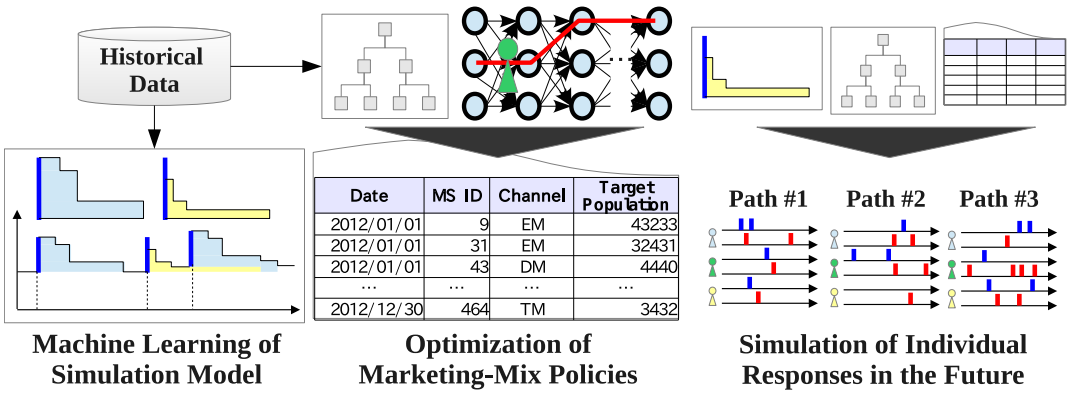


Figure 6.4: Flows of the input/output for/from each component in our What-If analysis framework. The event-based machine learning algorithms provide the nonparametric time-decaying forecasting models of the response-event spikes (left). Such nonparametric models involve multi-scale continuous state vectors, and we discretize the space of the continuous state vectors into discrete Micro Segments (MSs). By solving discrete-segment cMDPs, we obtain an MS-based marketing-mix policy consisting of the target population for each triplet of a timing, a segment, and a channel (center). The fitted simulation model, the micro-segmentation tree, and the optimized marketing-mix policy are finally inputted for the discrete event simulator that stochastically generates multiple paths of future sales, where each path consists of a sequence of response events for every individual consumer (right). We evaluate the mean and standard deviation of the total sales forecast among these paths, by summing the amounts of the response events over all of the consumers.

6.1 Framework for Normative What-If Analysis

The substance of the discrete event simulator is marked point processes that are extended from the nonparametric event-spike prediction models in Chapter 5. Each consumer has a real-valued and time-varying state vector that affects the occurrence probability and the amount of each response event. For optimization purpose we introduce a finite number of Micro Segments (MSs) by discretizing the state vectors with a dyadic decision-tree. A marketing-mix decision making problem, with respect to each MS, is introduced while its actual solving is deferred in Section 6.2.

For each consumer $i \in \{1, \ldots, m\}$, the simulator generates a sequence of response events given a sequence of marketing-stimulus events. As in Chapter 5, Every event is associated with a flag to represent either response or marketing-stimulus, a continuous time-stamp, and a response or stimulus amount such as the revenue in one purchase event. Every consumer i's response events are distributed solely based on her/his state vector $\boldsymbol{x}_i(t) \in \mathbb{R}^{dx}$. We forecast the future sales for each consumer i, by iterating the sampling of the response events and updating of the state vector $\boldsymbol{x}_i(t)$ based on adding the latest or removing the oldest events.

Using real data, we estimate a micro-segmentation function $\pi : \mathbb{R}^{dx} \to \mathcal{S}$ that deterministically converts any state vector into one MS $s \in \mathcal{S}$. Our segmentation is semi-automatic as follows. Marketers are accustomed to manually define a limited number of (typically from 3 to 20) segments of consumers, which we call the Strategic Segments (SSs). Each SS is usually defined with recency, frequency, or monetary values about the responses while more detailed segments than SS are desired for fully exploiting the benefits of MDP approaches. One aspect is to incorporate frequencies of past marketing stimuli, which are not handled in defining SSs usually, but which strongly affect the profitability of

one more action due to diminishing returns and forgetting of the old stimuli. We further segment every SS with the function involving past actions, where the substance of the function π is a dyadic binary tree that separates the d_X-dimensional state space into mutually-exclusive $|\mathcal{S}|$ regions. As we showed in Figure 6.2, marketers usually give the budget constraints using SS, while such constraints can always be converted into MS-based constraints. The manual definition for a limited number of SS provides easy understanding of budget plans, while the automatic estimation of many MSs yields profitable marketing-mix policy that carefully follows the finely-detailed states of consumers.

We need a dynamic budget allocation for each MS and channel, in order to efficiently generate the sequences of marketing-stimulus events input into the simulator. Let $t^{\mathrm{F}}_1, t^{\mathrm{F}}_2, \ldots, t^{\mathrm{F}}_H, t^{\mathrm{F}}_{H+1}$ be an increasing sequence of future time-stamps. At each time t^{F}_h such that $1 \leq h \leq H$, we need to choose a marketing-communication channel $a \in \mathcal{A}$ to target each of the m consumers, where the set $\mathcal{A}$ contains all of the available channels and an element $a_\emptyset \in \mathcal{A}$ representing no action. Let n_{hsa} be an integer to represent the target population in channel a for MS s at time t^{F}_h. For marketing budget planning in advance, at or before time t^{F}_1, we are required to optimize a set of the target populations $\{n_{hsa}; 1 \leq h \leq H, s \in \mathcal{S}, a \in \mathcal{A}\}$, where each MS s and channel a are regarded as a discrete state and an action of the MDP, respectively. When the target population exceeds the realized population of the associated MS during the simulation, we instead target the consumers belonging to other but similar MSs on the micro segmentation tree. Hence the entire amount of the planned costs are always consumed.

We maximize the multi-period sum of revenues minus marketing costs. We define an MDP with a set $\Theta \triangleq \{r_{hsa}, p_{s'|hsa}; 1 \leq h \leq H, s \in \mathcal{S}, a \in \mathcal{A}\}$, where r_{hsa} is the expected revenue acquired in period $[t^{\mathrm{F}}_h, t^{\mathrm{F}}_{h+1})$ when we target one consumer in MS s with channel a at time t^{F}_h, and $p_{s'|hsa}$ is the probability with which a consumer in MS s at time t^{F}_h targeted in channel a switches into MS s' at time t^{F}_{h+1}. After statistically fitting all of the elements in Θ with a procedure explained later in Section 6.3, we solve an optimization problem

$$\max_{\{n_{hsa}\}} \sum_{h=1}^{H} \gamma^{t_h - t_1} \sum_{a \in \mathcal{A}} \left[\sum_{s \in \mathcal{S}} n_{hsa} r_{hsa} - c_{ha} \left(\sum_{s \in \mathcal{S}} n_{hsa} \right) \right] \text{ subject to several constraints,} \tag{6.1}$$

where γ is a discounting factor provided by users and $c_{ha} : \mathbb{R} \to \mathbb{R}$ is a non-decreasing cost function to target consumers in channel a at time t^{F}_h. While the cost function in our case study is assumed to be time-independent, paper-based direct mails and telemarketing by human labors are considerable examples of the resources whose costs are time-dependent.

Optimization (6.1) is solved with financial constraints and observable state dynamics. Each of the constraints in (6.1) belongs to one of three types. The first type is a user-defined budget constraint. Because marketing organizations are complex, where different marketers are responsible for different segments or channels, we usually have multiple budget constraints denoted by a set $\boldsymbol{B}$. Each constraint $\beta \in \boldsymbol{B}$ is associated with a set of the triplets of timing, target-MS set, and channel denoted by $\mathcal{E}_\beta = \{(h, \widetilde{\mathcal{S}}, a)\}$. For each of the total cost $C_\beta \triangleq \sum_{(h,\widetilde{\mathcal{S}},a) \in \mathcal{E}_\beta} c_{ha} \left(\sum_{s \in \widetilde{\mathcal{S}}} n_{hsa} \right)$ associated with the constraint $\beta \in \boldsymbol{B}$, we have either its maximum or minimum formulated as either $C_\beta \leq C^{\max}_\beta$ or $C_\beta \geq C^{\min}_\beta$, respectively. The second type is a flow constraint to incorporate the transitions among segments and we provide its details later. The third one is the total-population constraint whose formula is $\forall h \in \{1, \ldots, H\}\ m \equiv \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} n_{hsa}$, based on the limitation that we use only one or no channel at each time.

When we expect synergy effects (Naik and Raman, 2003; Naik et al., 2005; Kolsarici and Vakratsas, 2011; Schultz et al., 2012) for Integrated Marketing Communication, it is theoretically possible to target in multiple channels at the same time with adding a combination of original channels into the

set $\mathcal{A}$. In practice, the better approach is to increase the number of periods H with narrower decision intervals, while only one channel is used for targeting in one period. If the number of MS is large, this practice avoids the combinatorial explosion while incorporates short-term synergy effects through segment transitions.

6.2 Linear Programming of the Target Populations

Let us introduce an LP algorithm to solve Optimization (6.1). The flow constraints on the segment transitions are given as linear inequalities, where we also pay attention to the difference between the expected and realized populations. The budget constraints are also linear inequalities, in which each cost function c_{ha} is a linear or piecewise-linear function of the target population. Both of these types of constraints yield additional slack variables to be optimized, and we consequently solve a high-dimensional LP to optimize both the target populations and the slack variables.

Our optimization problem corresponds to searching for a reasonable realization of random population in the future. Considering the population of each MS in the future to be stochastic, let N_{hsa} be a random variable to represent the target population for MS s with channel a at time t_h^{F}. Because the decision variable n_{hsa} can be regarded as one realization of the random variable N_{hsa}, we wish the realization n_{hsa} to provide high profits in a wide variety of possible scenarios caused by randomness. While n_{hsa} can theoretically be a function of the remaining budget at each time t_h^{F}, deterministically providing a constant value in advance yields a clearer and more practical budget planning for the marketer.

Handling the expected population in the future leads to certain equalities among the decision variables. We can constrain the expected populations in two consecutive periods by imposing $\left\langle \sum_{a\in\mathcal{A}} N_{(h+1)sa} \right\rangle = \left\langle \sum_{s'\in\mathcal{S}} \sum_{a\in\mathcal{A}} p_{s|hs'a} N_{hs'a} \right\rangle$, where $\langle\cdot\rangle$ represents the expectation operator for each random variable. We then get

$$\sum_{a\in\mathcal{A}} n_{(h+1)sa} = \varepsilon_{hs} + \sum_{s'\in\mathcal{S}} \sum_{a\in\mathcal{A}} p_{s|hs'a} n_{hs'a}, \tag{6.2}$$

by introducing a zero-mean random noise variable ε_{hs}.

To obtain the desired realization of the random target populations, we consider an alternative problem of maximizing the original objective minus the sum of the magnitudes of the noise variables. Because the noise variables must obey certain probabilistic distributions, larger values of the variance or absolute deviation of each noise variable ε_{hs} should be more penalized. For convenience of deriving an optimized solution completely linear in every decision variable, let us consider a penalty term which consists of the negation of the absolute deviation $|\varepsilon_{hs}|$ for every pair of $(h\in\{1,\ldots,H\}, s\in\mathcal{S})$, and which is related to the log-likelihood of Laplace distributions. We introduce a relaxation hyper-parameter $\eta_{hs} \geq 0$, which is the strength of the penalty term for the absolute deviations, and whose value is given a priori, and $\sigma_{hs} \triangleq |\varepsilon_{hs}|$. A modified maximization objective is introduced as

$$\max_{\{n_{hsa}\},\{\sigma_{hs}\}} \sum_{h=1}^{H} \gamma^{t_h - t_1} \sum_{a\in\mathcal{A}} \left[\sum_{s\in\mathcal{S}} n_{hsa} r_{hsa} - c_{ha}\left(\sum_{s\in\mathcal{S}} n_{hsa} \right) - \sum_{s\in\mathcal{S}} \eta_{hs}\sigma_{hs} \right], \tag{6.3}$$

where Optimization (6.3) looks similar to the Maximum A Posteriori estimation with L_1 regularization terms (Tibshirani, 1994), though the objective in (6.3) does not represent the penalized log-likelihood. Let $\wedge$ be the logical "and" operator and let $V_1 \geq \pm V_2$ mean $V_1 \geq V_2 \wedge V_1 \geq -V_2$. By converting (6.2)

into two inequalities

$$\sigma_{hs} \geq \pm \left[\sum_{a \in \mathcal{A}} n_{(h+1)sa} - \sum_{s' \in \mathcal{S}} \sum_{a \in \mathcal{A}} p_{s|hs'a} n_{hs'a} \right], \tag{6.4}$$

we can readily confirm that (6.3) and (6.4) are linear in the target population and the slack variable σ_{hs}, when we exclude the cost term $c_{ha}(\cdot)$. While one can assume Gaussian distributions on the noise variables with deriving a quadratic programming problem, the Laplace-distribution approximation yields the faster LP algorithm and robustness against the outliers of the realized populations.

Non-linear cost functions can be replaced by multiple linear inequalities with piecewise-linear approximations. For every channel $a \in \mathcal{A}$, we approximate its cost function by a piecewise-linear continuous and convex function $c_{ha}(n \geq 0) = \sum_{k=1}^{K_{ha}} I(n_{k-1} \leq n < n_k)(b_{hak} + \theta_{hak} n)$, where $n_0 \equiv 0, n_1, \ldots, n_{K_{ha}} \to \infty$ is a monotonically-increasing sequence of knots, $I(\cdot)$ is the boolean indicator function, K_{ha} is the number of supports, $b_{hak} \geq 0$ and $\theta_{hak} \geq 0$ are the bias and slope in the kth support, respectively. Because targeting no people incurs no cost, $\forall h \in \{1, \ldots, H\}\ \forall a \in \mathcal{A}\ b_{ha1} = 0$. Furthermore, the continuity and convexity assumptions provide a condition $\forall h \in \{1, \ldots, H\}\ \forall a \in \mathcal{A}\ \forall k \in \{1, \ldots, K_{ha} - 1\}\ \left(b_{hak} + \theta_{hak} x_k = b_{ha(k+1)} + \theta_{ha(k+1)} x_{k+1}\right) \wedge \left(\theta_{hak} < \theta_{ha(k+1)}\right)$. For the cost terms in the objective and the budget constraints, let us further introduce slack variables $\zeta_{ha} \triangleq c_{ha}\left(\sum_{s \in \mathcal{S}} n_{hsa}\right)$ and $\delta_{\beta z} \triangleq \sum_{z=(h,\widetilde{\mathcal{S}},a) \in \mathcal{E}_\beta} c_{ha}\left(\sum_{s \in \widetilde{\mathcal{S}}} n_{hsa}\right)$, respectively. Then each slack variable is constrained with a logical product of multiple linear inequalities stemming from the piecewise-linear approximations.

By integrating all of the aspects of linearizing the objective and constraints, an initial condition and total constraints for the population, we finally arrive at the LP problem we wish to solve:

$$\max_{\{n_{hsa}\},\{\sigma_{hs}\},\{\zeta_{ha}\},\{\delta_{\beta z}\}} \sum_{h=1}^{H} \left[\gamma^{t_h - t_1} \sum_{a \in \mathcal{A}} \left[\sum_{s \in \mathcal{S}} n_{hsa} r_{hsa} - \zeta_{ha} \right] - \sum_{s \in \mathcal{S}} \eta_{hs} \sigma_{hs} \right]$$

$$\text{subject to} \begin{cases} \forall a \in \mathcal{A}\ \forall h \in \{1, \ldots, H\} \forall k \in \{1, \ldots, K_{ha}\}\ \zeta_{ha} \geq b_{hak} + \theta_{hak} \sum_{s \in \mathcal{S}} n_{hsa} \\ \forall \beta \in \boldsymbol{B} \begin{cases} \sum_{z=(h,\widetilde{\mathcal{S}},a) \in \mathcal{E}_\beta} \delta_{\beta z} \leq C_\beta^{\max} & \text{if } \beta \text{ is a maximum constraint} \\ \sum_{z=(h,\widetilde{\mathcal{S}},a) \in \mathcal{E}_\beta} \delta_{\beta z} \geq C_\beta^{\min} & \text{if } \beta \text{ is a minimum constraint} \end{cases} \\ \forall \beta \in \boldsymbol{B}\ \forall z = (h, \widetilde{\mathcal{S}}, a) \in \mathcal{E}_\beta \forall a \in \{1, \ldots, K_{ha}\}\ \delta_{\beta z} \geq b_{hak} + \theta_{hak} \sum_{s \in \widetilde{\mathcal{S}}} n_{hsa} \\ \forall h \in \{1, \ldots, H-1\} \forall s \in \mathcal{S}\ \sigma_{hs} \geq \pm \left[\sum_{a \in \mathcal{A}} n_{(h+1)sa} - \sum_{s' \in \mathcal{S}} \sum_{a \in \mathcal{A}} p_{s|hs'a} n_{hs'a} \right] \\ \forall s \in \mathcal{S}\ n_{1s} = \sum_{a \in \mathcal{A}} n_{1sa} \\ \forall h \in \{1, \ldots, H\}\ m = \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} n_{hsa} \end{cases}, \tag{6.5}$$

where n_{1s} is the initial population of MS s deterministically given with historical data.

Each relaxation hyperparameter is set inversely proportional to the square root of the initial population of the associated MS, because it corresponds to the inverse of the absolute deviation of the target population, which is proportional to the square root of the variance. We set $\eta_{hs} = \rho m \gamma^{t_h - t_1} / \sqrt{1 + n_{1s}}$ using the entire population m, the initial population n_{1s}, and a global relaxation hyperparameter ρ.

Any additional constraints can be considered depending on the requirements in practice. One possible example is imposing autoregressive cost structures to affect the momentum of time-varying costs. Such autoregressive dependence is implemented as linear inequalities involving the slack variables in consecutive periods, i.e., ζ_{ha} and $\zeta_{(h+1)a}$.

6.3 Estimation of the Parameters in LP

Here we describe the steps in estimating the set of the parameters $\Theta \triangleq \{r_{hsa}, p_{s'|hsa}; 1 \leq h \leq H, s, s' \in \mathcal{S}, a \in \mathcal{A}\}$ from real data. We introduce marked point processes to stochastically generate the sequences of response events, whose occurrence obeys an inhomogeneous Poisson process that introduces time-dependent (inhomogeneous) rates in the Poisson processes, and whose amount conditional on its occurrence obeys a log-normal distribution. Both of the per-time probability and the amount distribution are modeled with each consumer's high-dimensional vector of state. The fitting of the parameters in the marked point processes is introduced in Section 6.3.1. Section 6.3.2 discusses the micro segmentation by discretizing the state space with a dyadic binary tree. Sections 6.3.3 and 6.3.4 discuss the computations of every parameter in the set Θ, by averaging the expected revenues that the marked point processes provide and by counting transition frequencies among micro segments, respectively. For representing the response amounts that are always non-negative, we denote the space of non-negative d-dimensional vectors by $\mathbb{R}^d_+$, and $\mathbb{R}_+ \triangleq \mathbb{R}^d_+$.

6.3.1 Fitting of the Simulation Models

For the primitive data containing all of the event sequences, we compute the sequences of piecewise-constant state vectors as we did in Chapter 5. Then we denote the training data by a set of the triplet series $\mathcal{D} = \{(\boldsymbol{x}_{i1}, \boldsymbol{y}_{i1}, t^{\mathrm{P}}_{i1}), \ldots, (\boldsymbol{x}_{iL_i}, \boldsymbol{y}_{iL_i}, t^{\mathrm{P}}_{iL_i})\}_{i=1}^m$, where L_i is the length of the sequence for consumer i, $\boldsymbol{x}_{ij} \in \mathbb{R}^{d_X}$ is a state vector, $\boldsymbol{y}_{ij} \in \mathbb{R}^{d_Y}$ is a response vector, and $t^{\mathrm{P}}_{ij} \in \mathbb{R}$ is the time at which the response amounts are observed. As we showed in Chapter 5. time-series of our state vectors is piecewise-constant in time, i.e., $\forall j \in \{2, \ldots, L_i\} \forall t \in [t^{\mathrm{P}}_{i(j-1)}, t^{\mathrm{P}}_{ij})\ \boldsymbol{x}_i(t) \equiv \boldsymbol{x}_{ij}$. The actual definition of the state vector is provided in the last part of this section, where we set the input dimensionality d_X high by exploiting a large variety of event types, in order to provide a good proxy of the hidden mental state of each consumer. In general we denote the response vector at time t and its uth element by $\boldsymbol{y}_i(t) \in \mathbb{R}^{d_Y}_+$ and $y_{iu}(t) \in \mathbb{R}_+$, respectively.

We model the generation of every consumer's response events with a marked point process, and we assume that the entire dataset for all of the consumers is generated with independent and parallel runs of these marked point processes. The inhomogeneous Poisson process for sampling the moments of response events is identical to Chapter 5's event-spike prediction model without the collective factor. The amount of each response represents a "mark" in our marked point processes. For the uth type of responses, we model its per-time occurrence probability as

$$\lim_{\Delta t \to 0} \frac{1}{\Delta t} \int_t^{t+\Delta t} I(y_{iu}(\widetilde{t}) \neq 0) d\widetilde{t} = \exp(b^{\text{timing}}_u + \left\langle \boldsymbol{w}^{\text{timing}}_u, \boldsymbol{\Phi}(\boldsymbol{x}_i(t)) \right\rangle), \tag{6.6}$$

where $\boldsymbol{\Phi} : \mathbb{R}^{d_X} \to \mathbb{R}^{d_\Phi}$ is a high-dimensional non-linear mapping function applied to the state vector, b^{timing}_u is a bias term, $\boldsymbol{w}^{\text{timing}}_u$ is a vector of coefficients, and $\langle \cdot, \cdot \rangle$ represents the inner product of two vectors. The parameters $(b^{\text{timing}}_u, \boldsymbol{w}^{\text{timing}}_u)$ are identical to the parameters $(b_k, \boldsymbol{w}_k)$ in Chapter 5, where both of the indexes u and k represent the type of the response events. The bias term represents the logarithm of the per-time response probability when $\boldsymbol{\Phi}(\boldsymbol{x}_i(t))$ is a zero vector, which usually corresponds to the case of no event history. Conditional on the occurrence of a response event, we model the distribution of the response amount as

$$\log y_{iu}(t)|_{y_{iu}(t) \neq 0} \sim \mathcal{N}(b^{\text{amount}}_u + \left\langle \boldsymbol{w}^{\text{amount}}_u, \boldsymbol{\Phi}(\boldsymbol{x}_i(t)) \right\rangle, \sigma^2_u), \tag{6.7}$$

where $\mathcal{N}\left(\mu, \sigma^2\right)$ represents a univariate Gaussian distribution whose mean is μ and whose variance is σ^2, b^{amount}_u is another bias term to represent the logarithm of the average response amount, and $\boldsymbol{w}^{\text{amount}}_u$

is another vector of coefficients. When the set of the parameters $\Xi \triangleq \{b_u^{\text{timing}}, \boldsymbol{w}_u^{\text{timing}}, b_u^{\text{amount}}, \boldsymbol{w}_u^{\text{amount}}, \sigma_u^2\}_{u=1}^{d_Y}$, the mapping function $\boldsymbol{\Phi}$, and the definition of the state vector are given, we are able to simulate all of the response events in the future, because Eqs. (6.6) and (6.7) specify how to sample the next response.

Here we summarize the optimization to fit the set of the parameters Ξ. For the fitting of the parameters $(b_u^{\text{timing}}, \boldsymbol{w}_u^{\text{timing}})$, we follow the procedure in Section 5.1.3. Let us denote the piecewise-constant and continuous-time Poisson log-likelihood function by $\ell(y; z, \tau)$. We solve the optimization problem

$$\max_{b_u^{\text{timing}}, \boldsymbol{w}_u^{\text{timing}}} \left[\sum_{i=1}^{m} \sum_{j=2}^{L_i} \ell(y_{iju}; b_u^{\text{timing}} + \left\langle \boldsymbol{w}_u^{\text{timing}}, \boldsymbol{\Phi}(\boldsymbol{x}_{ij}) \right\rangle, t_{ij}^{\text{P}} - t_{i(j-1)}^{\text{P}}) - \omega_u^{\text{timing}} \|\boldsymbol{w}_u^{\text{timing}}\|_1 \right], \quad (6.8)$$

where ω_u^{timing} is a regularization hyperparameter which controls the complexity and the number of non-zero coefficients in the fitted regression model, and whose optimal value is chosen with hold-out or cross-validation methods. Because each response amount obeys a log-normal distribution, for other parameters associated with response amounts, we solve the following penalized least-square problem

$$\max_{b_u^{\text{amount}}, \boldsymbol{w}_u^{\text{amount}}} \sum_{i=1}^{m} \sum_{j=1}^{L_i} I(y_{iju} \neq 0) \left[\log y_{iju} - (b_u^{\text{amount}} + \left\langle \boldsymbol{w}_u^{\text{amount}}, \boldsymbol{\Phi}(\boldsymbol{x}_{ij}) \right\rangle) \right]^2 + \omega_u^{\text{amount}} \|\boldsymbol{w}_u^{\text{amount}}\|_1,$$

with another regularization hyperparameter ω_u^{amount}. After fitting the parameters $(b_u^{\text{amount}}, \boldsymbol{w}_u^{\text{amount}})$, the corresponding noise level σ_u^2 is given as

$$\sigma_u^2 = \frac{\sum_{i=1}^{m} \sum_{j=1}^{L_i} I(y_{iju} \neq 0) \left[\log y_{iju} - (b_u^{\text{amount}} + \langle \boldsymbol{w}_u^{\text{amount}}, \boldsymbol{\Phi}(\boldsymbol{x}_{ij}) \rangle) \right]^2}{\sum_{i=1}^{m} \sum_{j=1}^{L_I} I(y_{iju} \neq 0)}.$$

We follow the principle of nonparametric multi-scale mixtures to provide the actual definition of the state vectors. Each element of the state vector is a frequency or total response amount for one type of events, using one specific sliding window chosen from 1 week, 2 weeks, , and 16 weeks. By incorporating all of the event types, we make the state vector a good proxy of the hidden mental state of each consumer. We use both of the frequencies and response amounts for purchase events, while we use only the frequency features for other types of events. In implementing an element-wise sub-linear function as the mapping function $\boldsymbol{\Phi}$, we adopt a definition $\boldsymbol{\Phi}(x_1, \ldots, x_{d_X}) = (x_1/(1 + x_1), \ldots, x_{d_X}/(1 + x_{d_X}))$ because of its higher predictive accuracy than that introduced in Section 5.2.4. While we are able to contain some seasonality variables (e.g., dummy variables to represent each month) in each state vector,in the later case study we did not include them due to the insufficient lengths of the training data.

For modeling the synergy effects depending on each pair of actions, the mapping function should contain quadratic terms of the state vector. Yet in the prior statistical experiments, using the quadratic mapping function did not improve the out-of-sample likelihood scores. We concluded that synergy effects among the marketing actions in our dataset are insignificant, while our methodology with the custom design of the mapping function allows for handling any types of higher-dimensional synergy effects, as well as the quadratic ones.

6.3.2 Tree-based Micro Segmentation

Here the fitting of micro-segmentation function $\pi : \mathbb{R}^{d_X} \to \mathcal{S}$ is addressed. As shown in Figure 6.5, we adopt a dyadic binary-tree structure, in which each branch uses only one variable of the state vector

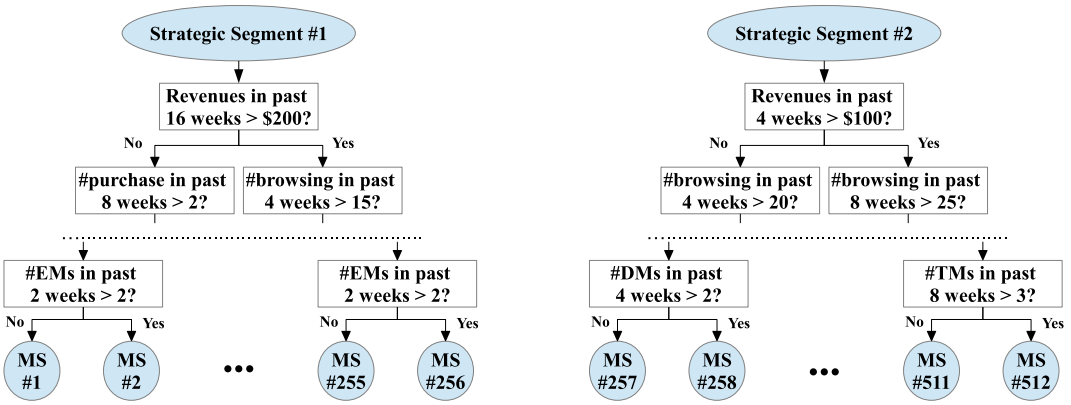


Figure 6.5: Examples of micro-segmentation using a dyadic binary tree, whose root node represents one Strategic Segment (SS) manually defined by users. Every node of the tree has a branching rule consisting of a single variable and its threshold, where many of the branching variables involve the number of specific events in specific window lengths (e.g., 4 weeks). Each of the leaves of a tree provides the definition of one Micro Segment (MS), and each consumer switches from one MS into another MS as time passes.

and the root node represents one SS. For each time stamp of events, let us compute a d_Y-dimensional vector $\widetilde{\boldsymbol{y}}_{ij} = (\sum_{j'=j}^{L_i} \widetilde{\gamma}^{t^{\mathrm{P}}_{ij'} - t^{\mathrm{P}}_{ij}} \boldsymbol{y}_{ij}) / (\sum_{j'=j}^{L_i} \gamma^{t^{\mathrm{P}}_{ij'} - t^{\mathrm{P}}_{ij}})$ using another discounting factor $\widetilde{\gamma}$. Like CART (Breiman et al., 1984) and C4.5 (Quinlan, 1996), we fit a regression tree to minimize the mean square errors of the discounted cumulative amounts of the responses $\{\{\widetilde{\boldsymbol{y}}_{ij}\}_{j=1}^{L_i}\}_{i=1}^{m}$. The choice of $\widetilde{\gamma}$ affects what types of the variables are used in each branch of the tree. In the case study, we adopted $\widetilde{\gamma} = 0.01$ or $\widetilde{\gamma} = 1$ to yield fast or slow switching among segments, respectively. Around the idea of performing micro segmentation with decision trees, there are a wide variety of relevant prior work (e.g., (Elsner et al., 2003; Abe et al., 2004; Tirenni et al., 2007b,a; Abe et al., 2009)).

6.3.3 Computing the Expected Revenues

By combining the parameters of the simulation model and the micro-segmentation function, we compute the set of the parameters Θ with manipulating the samples of state vectors. Let us synthesize a vector $\boldsymbol{x}_i^{+a}(t) \in \mathbb{R}^{d_X}$ by incrementing some elements of state vector where such elements represent frequencies of marketing stimuli in channel $a \neq a_\emptyset$. The vector $\boldsymbol{x}_i^{+a}(t)$ represents a new state of consumer i, if we immediately execute one more action in channel a. Also for the no-action element $a_\emptyset \in \mathcal{A}$, we denote $\boldsymbol{x}_i^{+a_\emptyset}(t) \triangleq \boldsymbol{x}_i(t)$ to unify notations. We further compute another vector $\boldsymbol{x}_i^{+a}(t, t')$ by replacing the seasonality elements in $\boldsymbol{x}_i^{+a}(t)$, which are functions of time t, by those of time t'.

Expected revenue is the product of expected frequency and per-event amount of responses. Let us denote by $g_u \in \{0, 1\}$ whether or not the uth type of response is a purchase event. By using the expectation of Poisson and log-normal distributions, we give a per-time expected revenue function of action a at time t, whose seasonality is modified with time t', as

$$R_{ia}(t, t') = \sum_{u=1}^{d_Y} g_u \exp\left(b_u^{\text{timing}} + \left\langle \boldsymbol{w}_u^{\text{timing}}, \boldsymbol{\Phi}(\boldsymbol{x}_i^{+a}(t, t')) \right\rangle + b_u^{\text{amount}} + \left\langle \boldsymbol{w}_u^{\text{amount}}, \boldsymbol{\Phi}(\boldsymbol{x}_i^{+a}(t, t')) \right\rangle + \frac{\sigma_u^2}{2} \right).$$

Assume that any interval time between the consecutive times of marketing decision in the future is sufficiently small, and hence the two vectors $\boldsymbol{x}_i(t_h^{\mathrm{F}})$ and $\boldsymbol{x}_i(t_{h+1}^{\mathrm{F}})$ are close. Then we can empirically

compute the expected-revenue parameter r_{hsa} as

$$r_{hsa} = (t^{\mathrm{F}}_{h+1} - t^{\mathrm{F}}_{h}) \frac{\sum_{i=1}^{m} \sum_{j=1}^{L_i - 1} I(\pi(\boldsymbol{x}^{+a}_{ij}) = s) \left(t^{\mathrm{P}}_{i(j+1)} - t^{\mathrm{P}}_{ij} \right) R_{ia}(t^{\mathrm{P}}_{ij}, t^{\mathrm{F}}_{h})}{\sum_{i=1}^{m} \sum_{j=1}^{L_i - 1} I(\pi(\boldsymbol{x}^{+a}_{ij}) = s) \left(t^{\mathrm{P}}_{i(j+1)} - t^{\mathrm{P}}_{ij} \right)}.$$

6.3.4 Computing the Transition Probabilities

While the transition probabilities we need depend on each channel and timing, we first compute a generator matrix $\mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|} \ni \boldsymbol{Q} = (Q_{ss'})$ of a channel-independent continuous-time Markov chain (Serfozo, 1979), whose transition-probability matrix for elapsed time τ is given as the matrix exponential $\exp(\boldsymbol{Q}\tau)$. Each element of the generator matrix is denoted as

$$Q_{ss'} = \begin{cases} -\lambda_s & \text{if} s = s' \\ \lambda_s q_{s'|s} & \text{otherwise} \end{cases},$$

where $\lambda_s > 0$ is the inverse of the expected duration in MS s while $q_{s'|s} \geq 0$ is the jump probability from MS s into MS s'.

The generator matrix $\boldsymbol{Q}$ is fitted with sequences of MSs. For each consumer i, by applying the function π for the state vector samples, we have a $\widetilde{L}_i (\leq L_i)$-length sequence of pairs of MS and duration time $(s_{i1}, \tau_{i1}), (s_{i2}, \tau_{i2}), \ldots, (s_{i\widetilde{L}_i}, \tau_{i\widetilde{L}_i})$. We can fit the parameter λ_s as the inverse of the average duration time for MS s. We also compute a transition frequency for every MS pair (s, s') such that $s \neq s'$, as

$$f_{ss'} = \sum_{i=1}^{m} \sum_{j=2}^{L_i} I(\pi(\boldsymbol{x}_{i(j-1)}) = s \wedge \pi(\boldsymbol{x}_{ij}) = s').$$

While simply taking the ratios among the the frequencies $\{f_{ss'}\}$ provides an estimate of the transition probabilities, we apply the modified Kneser-Ney smoothing (Chen and Goodman, 1998), which is the state-of-the-art statistical estimator of high-dimensional discrete-state Markov chains. For the details to compute $\{q_{s'|s}\}$ from $\{f_{ss'}\}$, refer (Chen and Goodman, 1998).

The action-dependent transition probabilities are finally given with mixing the action-independent transition probabilities with multiple source segments. Let us compute a ratio

$$\theta_{sz|ha} = \frac{\sum_{i=1}^{m} \sum_{j=1}^{\widetilde{L}_i} I(\pi(\boldsymbol{x}_i(t^{\mathrm{P}}_{ij})) = s \wedge \pi(\boldsymbol{x}^{+a}_i(t^{\mathrm{P}}_{ij}, t^{\mathrm{F}}_{h})) = z)}{\sum_{i=1}^{m} \sum_{j=1}^{\widetilde{L}_i} I(\pi(\boldsymbol{x}_{ij}(t^{\mathrm{P}}_{ij})) = s)}, \tag{6.9}$$

which represents the immediate switching probability from MS s into MS z when executing action a at time t^{F}_{h}. Using a transition-probability matrix $\boldsymbol{T}_h \triangleq \left\{ T_{ss'|h} \right\} = \exp(\boldsymbol{Q}(t^{\mathrm{F}}_{h+1} - t^{\mathrm{F}}_{h}))$, we obtain $p_{s'|hsa} = \sum_{z \in \mathcal{S}} \theta_{sz|ha} T_{zs'|h}$.

In practice, to avoid over-fitting and suppress the number of parameters, we adopt the same time intervals as $\forall h \in \{1, \ldots, H\}$ $(t^{\mathrm{F}}_{h+1} - t^{\mathrm{F}}_{h}) \equiv (t^{\mathrm{F}}_{2} - t^{\mathrm{F}}_{1})$, with ignoring the seasonality as $p_{s'|hsa} \equiv p_{s'|sa}$ by replacing the seasonality-adjusted vector $\boldsymbol{x}^{+a}_i(t^{\mathrm{P}}_{ij}, t^{\mathrm{F}}_{h})$ in (6.9) with $\boldsymbol{x}^{+a}_i(t^{\mathrm{P}}_{ij})$. Mainly for rigorous notations, here we discussed the general case in which we allow for variable intervals among the decision moments.

6.4 Experimental Evaluations and Implications

We demonstrate the values of our system using a common simulation model fitted with real data. From the same online retailer as that in Chapter 5, we received a smaller dataset containing 10,000-consumer and 2-year marketing-stimulus and response events. Time-stamp of every event is between 2009 and 2011, while we do not disclose the beginning date with confidentiality considerations. The dataset contains $d_Y = 2$ types of responses, browsing events for e-commerce sites and purchase events associated with a revenue amount. Because some browsing events lead later purchases, simulating both of the browsing and purchase events provides accurate forecasts of the mid-term revenues. We have 5 types of marketing-communication channels, including e-mails and direct mails. Each cost function c_{ha} does not depend on time t_h^{F} in this dataset and is linear to the number of consumers, while we do not disclose the underlying unit cost. In fitting the simulation model, the events in the first one-year period from 2009 to 2010 were chosen as the training data.

For relying on the results of What-If analysis, it is first essential to confirm the mid- or long-term predictability of the discrete event simulator, as well as the short-term predictability already validated in Chapter 5. Section 6.4.1 provides a qualitative evaluation of the accuracy of the simulation-based mid-term forecast. Section 6.4.2 provides a result of preliminary test to choose appropriate hyper-parameters in optimization. Using fixed values of these optimization hyperparameters, Section 6.4.3 demonstrates how the optimized policy allocates the entire budget for each of the triplet of timing, segment, and channel. Section 6.4.4 provides the main result of What-If analysis to compare multiple segmentation criteria and budget constraints for choosing the practical and best marketing-mix policy.

6.4.1 Validating the Mid-Term Predictability of the Simulator

Before the optimization and forecasting experiments, it is important to confirm mid-term predictability of the simulation model whose short-term predictability was already validated in Chapter 5. We simulated 100 paths of response events in the second one-year period from 2010 to 2011, where the inputs of the simulator were the actual marketing-stimulus events for all of the consumers. The difference between the simulated and actual revenue time-series in aggregate level is shown in Figure 6.6. The fitted model reasonably predicted the future with capturing parts of the temporal trends, while some of the transitive declines were failed to be predicted. While we assume the correctness of the fitted model in the following experiments, more accurate models fitted with lengthier time-series data will improve the reliabilities of the implications.

6.4.2 Choice of the Optimization Hyperparameters

Because our LP algorithm involves the hyperparameters to be specified, we tested the sensitivity of the maximal profit with these hyperparameters. Our goal is to maximize the mean among the simulated samples of the finite-horizon (annual) cumulative revenues minus costs. We decided not to discount this objective because the annual interest rate in the real economy is very low after the 2008 financial crisis (e.g., 0.28% for the 1-year Treasury bill in 2010 (of Governors of the Federal Reserve System, 2010)). Yet an optimized policy with a different discounting factor could yield a better result, due to the errors in approximating segment transitions. For the undiscounted true objective, Table 6.1 summarizes how combination of the discounting factor γ and the relaxation hyperparameter ρ in the optimization process affects the simulated revenue. While optimization with the larger discounting factor always ended in lower revenue, the relaxation hyperparameter ρ multiplied for the L_1-norm does not strongly affect the revenues.

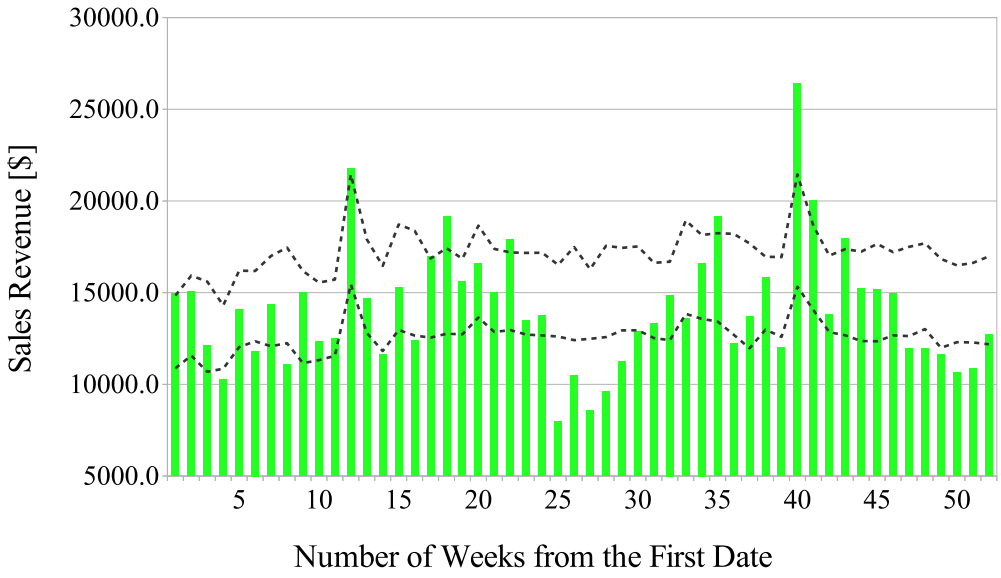


Figure 6.6: Mid-term predictability of the fitted simulation model. Each green-colored bar represents the actual weekly sales revenue, while the two dashed lines are the 2.5%- and 97.5%-tile estimates based on the simulated revenues for the actual marketing-stimulus events. While this mid-term forecasting is less dynamic than the actual, the simulator successfully predicted the two peaks located in the 12th and 40th weeks. Due to confidentiality considerations, all of the revenue values are multiplied with one secret scalar.

Based on the insensitivity to ρ, we adopted a fixed setting of $\gamma = 1$ and $\rho = 10^3$ for all of the remaining main experiments. Our test justified the usage of the same discounting factor in the optimization and simulation steps as $\gamma = 1$.

6.4.3 Properties of the Optimized Policy

To show a typical outcome of our system, we demonstrate profit gains using three Strategic Segments (SSs), Gold, Silver, and Normal, defined with the empirical percentiles of the latest 16-week revenue from each consumer. The upper 5%- and 25%-tiles of the 16-week revenues were chosen as the boundaries, where the Gold SS has the highest revenue. While we show the results for the sample 10,000 customers, expected revenues from a larger number of consumers can be inferred with scaling.

Typical outcomes of the optimization is clarified in Figure 6.7, as one path of revenue and marketing-cost allocation of a marketing-mix policy using a 4-depth micro-segmentation. The optimizations are done with bounding the annual budget to be from 75% to 125% of the actual total cost. The optimal spending is concentrated in later periods or the first week, with producing the revenue time-series that is quite different from the actual. Such characteristic results were given with our model assumption. While we can constantly stimulate the consumers with cheap e-mails, more costly direct mails should be concentrated on high-revenue consumers. The drastic increase of direct mails in the 19th week occurred after the population of the high-revenue consumers had reached certain critical mass. The change point of the revenue is around the 17th week, which is implied from each consumer's 16-week

Table 6.1: Sensitivity of the 20-path mean among the simulated revenues, for each triple of segmentation criteria (Fast or Slow Switching with depth=4 or 6), global relaxation hyperparameter (row) and annual interest rate (column) to set the discounting factor . In every segmentation, high discounting factors harm the profitabilities of the attained policies while relaxation hyperparameters do not.

		0%	5%	10%	20%	30%
FAST ($d=4$)	$\rho=10^0$	8.69×10^5	8.65×10^5	8.61×10^5	8.34×10^5	8.17×10^5
	$\rho=10^1$	8.68×10^5	8.68×10^5	8.68×10^5	8.32×10^5	8.06×10^5
	$\rho=10^2$	8.66×10^5	8.63×10^5	8.63×10^5	8.28×10^5	8.08×10^5
	$\rho=10^3$	8.63×10^5	8.66×10^5	8.66×10^5	8.24×10^5	8.06×10^5
	$\rho=10^4$	8.65×10^5	8.65×10^5	8.62×10^5	8.28×10^5	8.04×10^5
	$\rho=10^5$	8.67×10^5	8.68×10^5	8.63×10^5	8.26×10^5	8.06×10^5
SLOW ($d=6$)	$\rho=10^0$	9.21×10^5	9.17×10^5	9.14×10^5	8.96×10^5	8.58×10^5
	$\rho=10^1$	9.16×10^5	9.19×10^5	9.12×10^5	8.88×10^5	8.57×10^5
	$\rho=10^2$	9.13×10^5	9.10×10^5	9.10×10^5	8.89×10^5	8.57×10^5
	$\rho=10^3$	9.19×10^5	9.13×10^5	9.09×10^5	8.88×10^5	8.60×10^5
	$\rho=10^4$	9.20×10^5	9.18×10^5	9.13×10^5	8.95×10^5	8.58×10^5
	$\rho=10^5$	9.22×10^5	9.12×10^5	9.10×10^5	8.92×10^5	8.60×10^5

memory initialized with unoptimized stimulus events before the first week. The optimization algorithm assigned the stimuli not in these less profitable 16 weeks but in the later more profitable periods. Thus, we conclude that carefully modeling the memories of consumers, as we did in Chapter 5, is particularly important for practical marketing strategies.

6.4.4 What-If Analysis with Various Budget Constraints and Segmentation

In the final and main experiment we evaluated how the revenue forecasts depend on budget constraints and segmentation criteria. Our main aim is to assess what types of flexibility in allocating budgets are beneficial. We clarify the advantage of incorporating the past actions in defining each MS, over the conventional segmentation approaches that use only the past responses. Our MS trees are fitted to yield either fast or slow switching among segments, using short (e.g., 1-week) or lengthy (e.g., 16-week) sliding-window features in their branching. For each setting of the trees, we evaluate the outcome of choosing whether to involve the frequencies of the past actions. Then we also investigate how constraints manually added by marketers change the performances, using three types of the budget constraints in addition to the annual constraint used in Figure 6.7. The three constraints bound the total budget in every quarter, annual budget in every SS, and annual budget in every channel, respectively with the same 75% and 125% rules derived from the actual costs.

Figure 6.8 summarizes the results of our What-If analysis. Let us focus on the fact that solving cMDPs whose segmentations do not involve the action variables resulted in very low profits, even below the level of a current practice that would be based on this retailer's own scoring model whose details are unknown to us. In contrast, our approach with incorporating the action variables into segmentation is evidenced to yield more profits than both of the current practice and segmentations without actions, if we set the complexity of the micro-segmentation in an appropriate level. As a resulting implication for marketers, at least when we apply discrete-state MDPs for marketing-mix decision making, it is highly recommended to micro-segment consumers with the amounts of the past actions, in addition to the amounts of responses, for carefully searching for the optimal timing stemming from diminishing returns and forgetting.

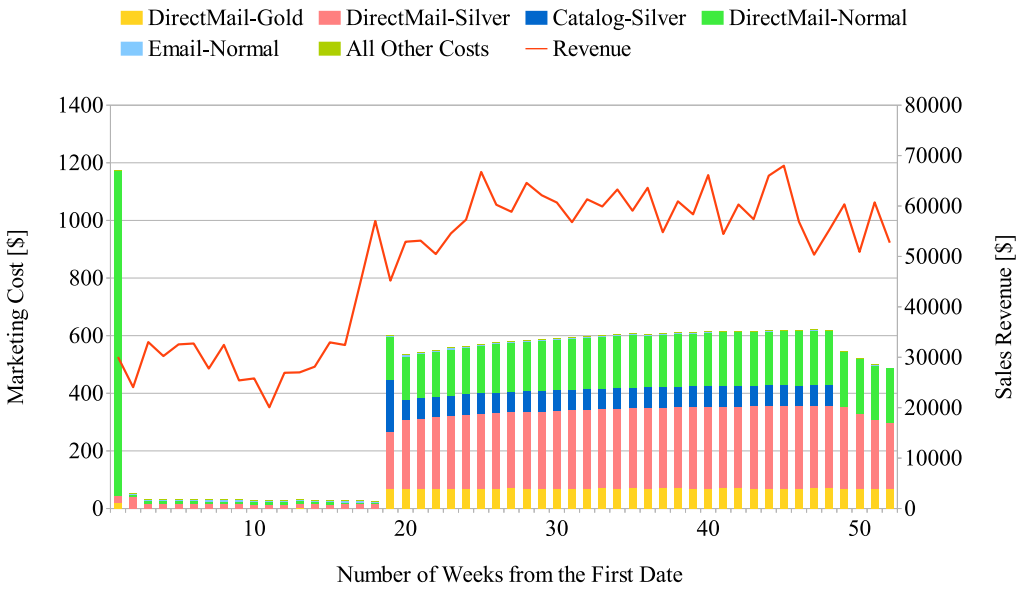


Figure 6.7: One realized path of revenues (the line) and cost allocation (each bar) for an optimized policy. All of the values are multiplied with the same secret scalar used in Figure 6.6. Cost items are classified with pairs of channel and strategic segment. To summarize the finely-detailed budget allocation, remaining items with smaller costs (e.g., DirectMail-Normal) are aggregated into one item, All Other Costs. The optimized revenue time-series is quite different from those in Figure 6.6, with an allocation concentrated in later periods or the first week.

Manually adding constraints usually suppresses the benefits of the automatic optimization algorithm, but is found to often produce better solutions particularly when simulation is required for evaluating the true objective. Marketers should be flexible in moving budgets from some segments into the other segments, because putting a cap on the budget for every SS more deprived the profit than bounding the total budgets in every quarter. In contrast, channel-specific constraints efficiently mitigated the approximation errors and yielded astonishing high revenues. Given these results, we believe that our approach of combining the approximate optimization and the full simulation can stimulate desirable interactions between the system and marketers, who consider simple but effective constraints instead of fully relying on a single optimization result.

6.5 Discussion

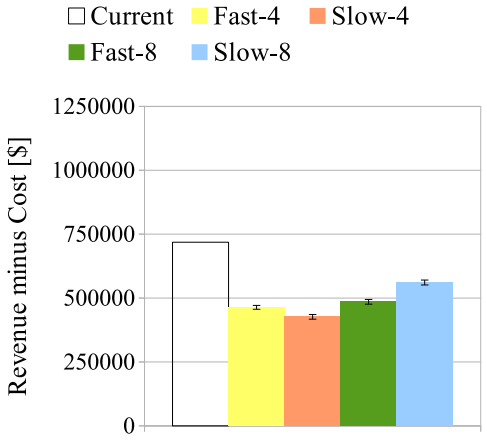
Let us discuss directions to further improve the applicability of our What-If analysis framework for more complicated marketing decision-making problems. A general direction we must consider is to narrow the gap between the approximate but computationally-efficient optimization and the accurate but time-consuming simulation. One essence to realize the ultimate consistency between the two different objectives is allowing the direct handling of the state vectors in the solving of cMDPs. Remember that the fortunate derivation of the efficient LP algorithm is heavily thanks to the mutual-exclusiveness of the Micro Segments. In contrast, when we directly solve the continuous-state cMDP problems based on the raw values of the continuous state vectors, we need a completely-new and

efficient optimization algorithm that does not assume the mutual-exclusivity of discrete segments. We would need to refer the literature of continuous-state reinforcement learning algorithms using piecewise-linear approximations. Converting the continuous state vectors into mutually-independent binary features is another considerable direction to be investigated.

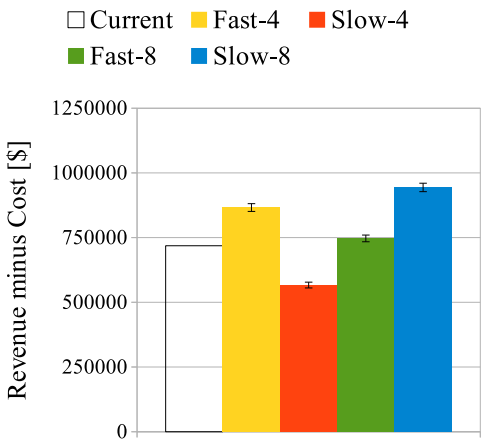
Other meaningful directions to enlarge the applicability of our framework is to increase the types of marketing actions the system is able to handle, by referring the actual details of real marketing operations. The first direction to be investigated is to incorporate mass advertising, in which we cannot control the exposure of an ad to each individual. A new optimization algorithm, which incorporates the constraints on the targeting of consumers with mass advertising, should be developed for Integrated Marketing Communications. The second direction is to incorporate the interactions among consumers whose predictive modeling has been already discussed in Chapter 6. The assumption of independence among the responses by all of the consumers does not hold in the real world, and the current discrete event simulation would be replaced by a more general agent-based simulation (e.g., (Libai et al., 2013)), introducing correlation among the response events. Rigorous risk evaluation for the marketing-mix policies would be enabled with such correlation modeling.

6.6 Summary

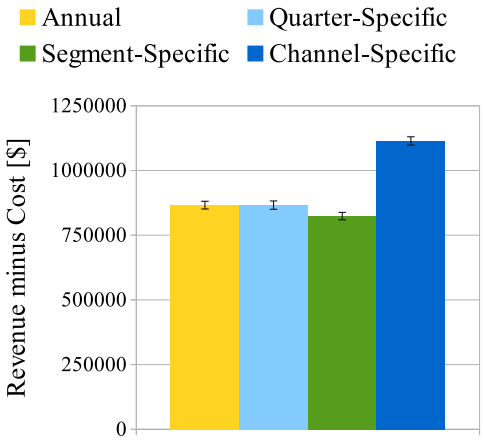
This chapter introduced a normative marketing decision-making algorithm that exploits the forecasts by our nonparametric descriptive model. The normative optimization problem is implemented as an optimization of target populations to form a time-dependent marketing-mix policy that allocates the marketing budgets across many periods, segments, and channels. We introduced a new LP algorithm to solve the optimization problem, which allows for inputting complex budget constraints required in real marketing operations. With simulating continuous-state marked point processes to generate each consumer's response events, we validated the implied revenue gains by our normative decision-making algorithm, based on an online retailing case study. By relying on this final experimental achievement, we believe the huge benefit of forecasting the behaviors of humans based on our philosophy of non-parametric multi-scale descriptive modeling, and the practical rationality of the normative decision making as an additional outcome derived from the philosophy.



Dependence on micro-segmentation without action variables



Dependence on micro-segmentation using action variables



Dependence on budget constraints

Figure 6.8: What-If analysis whose values are multiplied with the same secret scalar used in Figures 6.6 and 6.7. The left-most white bars in the top two figures show the actual value in the simulation period, while every other bar shows the 100-path mean and standard deviation (with an error bar) for an optimized policy. Segmentation names "Fast-d" and "Slow-d" represent d-depth micro-segmentation trees whose transitions from one segment into another segment are frequent or infrequent, respectively. The top-left and top-right figures provide comparisons among several micro-segmentation criteria, where action variables are not incorporated in the branching of the micro-segmentation in the left figure, while are incorporated in the right figure. The bottom figure shows the dependence of the realized profit on each of the set of budget constraints, when the Fast and depth=4 micro-segmentation involving action variables is commonly used. By comparing the top-right figure with the top-left one, segmentation using action variables is shown to make state-based policies significantly profitable. The highest performance when channel-specific budget constraints are imposed tells us that some constraints in LP are able to stabilize the optimization, by mitigating the errors introduced in approximating the true continuous-state cMDPs by a discrete-segment ones.

Chapter 7

Conclusion

This dissertation introduced efficient decision-making algorithms in the real world, based on the nonparametric descriptive models to predict the behaviors of interacting humans. The proposed nonparametric models interpolate multi-scale basis functions for handling either the fat tails produced by the positive feedback, or the long-range dependence yielded by the persistent interactions. By explicitly preparing the multi-scale basis functions in advance, we derived convex optimization algorithms to obtain the global optima of the model parameters in all of the technical contributions. The design of the multi-scale basis functions has been customized for each of the transportation and the marketing application examples. In the latter marketing application, the nonparametric mixture modeling contributed for the efficient computation of the time-varying state vectors, which consequently yielded a normative and optimal marketing-mix decision-making algorithm for marketers to obtain the maximum long-term profit.

In this concluding chapter, we start with reviewing the essences of our specific technical contributions. Section 7.1 discusses every specific contribution provided in each of the previous four chapters. Then we proceed into considerations about how our philosophy of the nonparametric multi-scale mixtures has been commonly applied in all of the technical contributions, or customized for implementing each specific contribution. Based on these considerations, Section 7.2 provides our main overview to conclude this dissertation in terms of the generalization capability of the nonparametric-mixture modeling in broader domains. In considering the future work we will next solve, we try to go back to the essential challenge introduced in Chapter 1, which is the accurate modeling and prediction of the bounded rationality and the positive feedback in the real world of interacting humans. Section 7.3 discusses how we should solve the remaining problems still by using the statistical descriptive modeling framework, but by focusing more on microsophic decision-making mechanisms each human possesses.

7.1 Our Contributions

In Chapter 3, as our first contribution, we introduced the acceleration of globally-optimal nonparametric density estimation, via deriving an efficient optimization algorithm applied in the convex clustering (Lashkari and Golland, 2008) or the Kullback-Leibler Importance Estimation Procedure (KLIEP; Sugiyama et al. (2008)). Such accelerated algorithm of nonparametric density estimation is a prerequisite for stably fitting many fat-tail distributions observed in large-scale datasets. Instead of the existing first-order EM algorithms, the accelerated algorithm adopted the Sequential Minimal Optimization approach that picks up a pair of two similar basis functions and quickly prunes the irrelevant

one of them. In each step to update the mixture weights for such two basis functions, exact evaluations of their first derivatives enabled quick judgement of the pruning of the irrelevant basis while an element-wise Newton-Raphson method produced efficient updating of the weights of relevant basis functions. The proposed algorithm is further accelerated when we incorporate locally-adaptive bandwidths, which embody a multivariate generalization of the multiple scales.

In Chapter 4, as our second contribution and an example of applying the first contribution for transportation decision making, we introduced a scalable nonparametric conditional density estimator of many travel-time distributions conditional on each of the huge number of links in a road network, and each of the one-hour timezones. The proposed estimation algorithm is designed on deliberations about how to realize all of the high accuracy in statistical prediction, scalability for large datasets, and physical interpretation to justify the interpolation mechanism in the proposed estimator, based on the positive feedback among many vehicle drivers. The proposed estimator begins with fitting of nonparametric basis density functions, which are modeled as mixtures of gamma or log-normal distributions having multiple characteristic scales. Then, for both generalizing for the links having limited numbers of travel-time samples and reflecting the positive feedback, the proposed estimator interpolates these nonparametric basis density functions among close links using a sparse diffusion kernel, which emulates the propagation of traffic and which is customized for guaranteeing the scalability. The accelerated convex clustering algorithm and the KLIEP are applied in optimizing both of the first mixture of gamma or log-normal distributions, and the second mixture of the nonparametric basis density functions. The full spatio-temporal estimates are provided by further interpolating these multiple spatial estimators among multiple timezones. Both normative and descriptive decisions in transportation, such as route recommendation and traffic simulation, are realized by inputting the fitted travel-time distributions into risk-sensitive stochastic routing algorithms.

As well as the modeling of fat tails, we addressed the modeling of long-range dependence in Chapter 5, by taking a marketing application example to predict the responses of each individual consumer. The philosophy of nonparametric multi-scale mixtures is here implemented as a new model of continuous-time Inhomogeneous Poisson Processes to distribute the spikes of response events by consumers, whose covariates are compound outcomes of marketing actions and self-exiting consumer activities. Here the nonparametric mixtures are implemented as staircase functions, which are the mixtures of multi-scale step functions and whose computational efficiency is essential for securing the scalability to large datasets. The staircase-functional nonparametrics was exploited for modeling both of the power-law forgetting curves caused by the interactions among each consumer's multiple memories, and those for the aggregated responses reflecting the word-of-mouth and social trends among mutually-interacting consumers. To detect each mutually-interacting group of consumers, we introduced an automatic clustering of time-series regression residuals, which is the difference between the actual responses and predicted ones by an initial Poisson regression without the mutual-interaction factor.

In Chapter 6, we addressed an economic decision making problem to optimizing the marketing-mix budget allocation, whose underlying forecasts are given by the nonparametric-mixture models reflecting the long-range dependence between marketing-stimulus and response events. Here the computationally-efficient staircase nonparametrics, which yielded the finite-dimensional state vectors, greatly helped us derive an efficient state-based marketing-mix policy based on the constrained Markov Decision Processes. Consideration about the budget constrains in real marketing operations led us to design an efficient Linear Programming algorithm to optimize a target population, for each triple of a timing, a target segment, and a marketing-communication channel. What-If analysis experiments, which test different segmentation criteria while adopt the common simulation model, enabled

us to compare many marketing-mix strategies within the common metric about the implied profit gains.

7.2 Overview

The common issue in all of the specific problems is the modeling of the fat tails and the long-range dependence observed in the real human activities. The philosophy of the efficient nonparametric multi-scale mixtures is implemented either as the globally-optimal sparse nonparametric density estimation for the fat-tail distributions, or the computationally-efficient time-varying explanatory variables in the Inhomogeneous Poisson Processes to model the long-range dependence. The Sequential Minimal Optimization technique to accelerate the convex clustering algorithm is regarded as a common and generally-useful technique to guarantee the applicability of sparse nonparametric density estimation for large-scale and real-world datasets. The high predictive accuracies about the lots of travel-time distributions in vehicle traffic modeling provides one experimental evidence of the efficiency of the proposed nonparametric philosophy, in addition to the purely algorithmic contribution. By focusing on the fact that we needed careful strategies to interpolate each basis density functions with one another, we regard the combination of the accelerated and global optimization of the mixture weights with domain-specific manual designs of basis functions, which should cover all of the required scales in each application, as the essential key to achieve high generalization capabilities in fitting lots of fat-tail and complex distributions. The success of the consumer-response prediction implies that the effective combination of the convex optimization and custom design of multi-scale basis density functions is applicable also for modeling the long-range dependence. Another outcome of the nonparametric mixture of finite-support multi-scale staircase functions is each of the efficiently-computed state vectors, which are explanatory variables in the Poisson regression. Such efficient representation of the state vectors while still incorporating the long-range dependence is the main enabler of the normative marketing-mix decision making algorithm, which is based on the constrained Markov Decision Processes and whose profit gains are experimentally evidenced with discrete event simulation.

In reviewing the connection between the successes in transportation and marketing applications, we are able to remember the fact that both fat-tail distributions and long-range dependence are commonly parametrized by power-law functions. Our nonparametric mixture modeling methodology is free from all of the three crucial limitations of high bias, local optimality, and non-scalability which were first discussed in Section 1.4. This significant computational advantages, over the existing parametric approach of directly modeling the power-law functions, are regarded as the main reasons of the experimental high performances by our nonparametric models in all of the example applications.

In contrast to the commonality of the nonparametric mixture philosophy, the required customizations in designing each basis function are caused by the limitation of available data, and physical interpretations or domain knowledge have been shown to be useful in such customizations. For example, while the main justification of the sparse diffusion kernel is the positive feedback that occurs as propagation of traffic on a road network, such sparse diffusion kernel also contributed for improving the predictive accuracies with alleviating the lack of training data. For another example, the aggregation of the frequencies of responses to account for unobservable word-of-mouth is an outcome by prioritizing the collective behavior of the entire market over the detectability of each inter-individual interaction. This marketing-specific customization was also inspired by limitation of available data, and further led the idea of clustering the consumers based on a time-series of the residuals, which is the difference between the actual responses and predicted ones by a basic Inhomogeneous Poisson regression. The effectiveness of elegantly solving the problems with constraints has also been observed

in the design of the normative marketing-mix optimization algorithm. Here the consideration about the budget constrains in real marketing operations led us to design an efficient Linear Programming algorithm to optimize a target population.

Based on all of the considerations in this section, we conclude that our philosophy of the nonparametric descriptive modeling leads effective and practical decision making strategies in interacting with other humans, as well as providing accurate forecasts. The source of this practicality is the nonparametric multi-scale nature in the descriptive modeling, which is supported by the stable and global optimization of the mixture weights and the broad coverage for all of the required scales to yield the fat tails or the long-range dependence in each application domain.

7.3 Future Work

In the future work, we primarily plan to directly model the bounded rationality of humans with focusing on more microscopic mechanisms. The nonparametric models in this dissertation have been specialized for predicting macroscopic observations, which are essentially the collective phenomena as the aggregation of many microscopic interactions among humans. While this macroscopic approach resulted in an efficient normative decision making, more interpretable approaches while keeping on the high predictive accuracy should be created for more useful implications. One possible way is to modify the nonparametric models as directly explaining the *mechanisms*, with which each human becomes often irrational. Such mechanical explanations would be a mixture of the optimizing models assuming the rationality of humans, and the statistical models merely focusing on the input-output relationship.

The other fundamental direction in extending our approach is the microscopic modeling of interactions, whose statistical estimation is highly accurate even when the amount of training data is limited. A key phenomenon that strongly impacts the way of our decision making and that we must predict with high accuracy is the bursty correlation among humans, such as financial crisis caused by fears of many investors and exploding word-of-mouth about exceptionally excellent products. In the real world, correlation among the behavior of each human is time-varying where strong interactions happen within very short periods despite the weak coupling strengths in most of the other days. The limited lengths of such strong-interaction periods make the statistical estimation of the time-varying correlation challenging, as we discussed in Chapter 5. Specialized designs of autoregressive predictive models, which have parsimonious parametrizations while keep on the sufficient explanatory powers about the variability of the correlation, must be provided for handling the burstness phenomena. Thus, efficient statistical modeling with careful considerations about the learnability to real data will remain as the core of our approaches in any possible direction.

References

N. Abe, N. K. Verma, C. Apté, and R. Schroko. Cross channel optimized marketing by reinforcement learning. In *Proceedings of the 10th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2004)*, pages 767–772, 2004.

N. Abe, P. Melville, C. Pendus, C. K. Reddy, D. L. Jensen, V. P. Thomas, J. J. Bennett, G. F. Anderson, B. R. Cooley, M. Kowalczyk, M. Domick, and T. Gardinier. Optimizing debt collections using constrained reinforcement learning. In *Proceedings of the 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2010)*, pages 75–84, 2009.

G. W. Ainslie. Impulse control in pigeons. *Journal of the Experimental Analysis of Behavior*, 21(3): 485–489, 1974.

G. M. Allenby, R. P. Leone, and L. Jen. A dynamic model of purchase timing with application to direct marketing. *Journal of the American Statistical Association*, 94(446):365–374, 1999.

E. Altman. *Constrained Markov Decision Processes*. Chapman and Hall/CRC, 1999.

M. R. Anderberg. *Cluster analysis for applications*. Academic Press, New York, 1973.

C. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics*, 2(6):1152–1174, 1974.

R. Axtell and G. McRae. A general mathematical theory of discounting. SFI Working Paper, 2007.

V. Bala and S. Goyal. Learning from neighbours. *Review of Economic Studies*, 65(3):595–621, 1998.

J. Barceló. *Fundamentals of Traffic Simulation*. Springer, 1st edition, 2010.

K. W. B.D. Frischknecht and P. Papalambros. On the suitability of econometric demand models in design for market systems. *Journal of Mechanical Design*, 132(12), 2010.

M. Beal, Z. Ghahramani, and C. Rasmussen. The infinite hidden Markov model. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 577–584. MIT Press, Cambridge, MA, 2002.

A. S. Benjamin, J. S. de Belle, B. Etnyre, T. A. Polk, M. Guadagnoli, and G. Stelmach. *Human Learning: Biology, Brain, and Neuroscience*. Elsevier Science, 2008.

J. Berger and E. Schwartz. What do people talk about? drivers of immediate and ongoing word-of-mouth. *Journal of Marketing Research*, 48(5):869–880, 2011.

D. P. Bertsekas. *Dynamic Programming and Optimal Control (2nd Edition)*. Athena Scientific, 2000.

S. Bikhchandani, D. Hirshleifer, and I. Welch. Information cascades, 2008.

C. M. Bishop and M. E. Tipping. Variational relevance vector machines. In *The 16th Conference on Uncertainty in Artificial Intelligence (UAI 2000)*, pages 46–53, 2000.

O. J. Blanchard and M. W. Watson. Bubbles, rational expectations and financial markets. *Crises in the Economic and Financial Structure*, pages 295–316, 1982.

O. J. Blanchard and M. W. Watson. Bubbles, rational expectations and financial markets. NBER Working Paper No. 945, 1983.

R. Blattberg and N. Gonnedes. A comparison of stable and student distribution as statistical models for stock prices. *Journal of Business*, 47(2), 1974.

D. M. Blei and M. I. Jordan. Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–144, 2006.

V. Borkar and R. Jain. Risk-constrained Markov decision processes. In *Proceedings of the 49th IEEE Conference on Decision and Control (CDC 2010)*, pages 2664–2669, Atlanta, GA, 2010.

M. Botsch and A. J. Nossek. Construction of interpretable radial basis function classifiers based on the random forest kernel. In *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN 2008)*, pages 220–227, Hong Kong, 2008.

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and regression trees*. Wadsworth and Brooks/Cole Advanced Books and Software, Monterey, CA, 1984.

D. Brownstonea, D. Bunch, and K. Train. Joint mixed logit models of stated and revealed preferences for alternative-fuel vehicles. *Transportation Research Part B: Methodological*, 34(5):315–338, 2000.

M. D. Buhmann. *Radial Basis Functions: Theory and Implementations*. Cambridge University Press, 2003.

J. R. Busemeyer and J. T. Townsend. Decision field theory: a dynamic cognition approach to decision making. *Psychological Review*, 100(3):432–459, 1993.

E. Camponogara and W. K. Jr. Distributed learning agents in urban traffic control. In *The 11th Portuguese Conference on Artificial Intelligence (EPIA 2003)*, pages 324–335, 2003.

H. M. Cannon, J. D. Leckenby, and A. Abernethy. Beyond effective frequency: Evaluating media schedules using frequency value planning. *Journal of Advertising Research*, 42(6):1–15, 2002.

M. Á. Carreira-Perpiñán. Gaussian mean shift is an EM algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(5):767–776, 2007.

S. R. Chandukala, J. Kim, T. Otter, P. E. Rossi, and G. M. Allenby. Choice models in marketing: Economic assumptions, challenges and trends. *Foundations and Trends in Marketing*, 2(2):97–184, 2007.

G. Chang and J. Feigenbaum. A Bayesian analysis of log-periodic precursors to financial crashes. *Quantitative Finance*, 6(1):15–36, 2006.

C. Chatfield and G. Goodhardt. Results concerning brand choice. *Journal of Marketing Research*, 12: 110–113, 1975.

C. F. Chen. Personality, safety attitudes and risky driving behaviors–evidence from young Taiwanese motorcyclists. *Accident Analysis and Prevention*, 41(5):963–968, 2009.

R. Chen and G. Blankenship. Dynamic programming equations for discounted constrained stochastic control. *IEEE Transactions on Automatic Control*, 49(5):699–709, 2004.

R. Chen and E. Feinberg. Non-randomized policies for constrained Markov decision process. *Mathematical Methods of Operations Research*, 66:165–179, 2007.

S. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Harvard Computer Science, 1998.

A. T. Ching. Consumer learning and heterogeneity: Dynamics of demand for prescription drugs after patent expiration. *International Journal of Industrial Organization*, 28(6):619–638, 2010.

Y.-L. Chow and M. Pavone. Stochastic optimal control with dynamic, time-consistent risk constraints. In *Proceedings of American Control Conference (ACC 2013)*, pages 390–395, Washington, DC, 2013.

R. Cont. Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, 1(2):223–236, 2001.

R. Cont. Long range dependence in financial markets. In E. Lutton and J. Vehel, editors, *Fractals in Engineering*, pages 159–180. Springer, 2005.

R. Crane and D. Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *PNAS*, 105(41):15649–15653, 2008.

I. Csiszár and P. C. Shields. Information theory and statistics: A tutorial. *Foundations and Trends in Communications and Information Theory*, 1(4):417–528, 2004.

W. Delaney and E. Vaccari. *Dynamic Models and Discrete Event Simulation*. Dekker INC, 1998.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

F. Deschatres and D. Sornette. The dynamics of book sales: Endogenous versus exogenous shocks in complex networks. *Physical Review E*, 72(1):016112, 2005.

I. S. Dhillon, Y. Guan, and B. Kulis. Kernel k-means, spectral clustering and normalized cuts. In *10th ACM KDD Conference*, pages 551–556, 2004.

E. W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1): 269–271, 1959.

D. V. Djonin and V. Krishnamurthy. MIMO transmission control in fading channels - a constrained Markov decision process formulation with monotone randomized policies. *IEEE Transactions on Signal Processing*, 55(10):5069–5083, 2007.

W. Dong and A. Pentland. A network analysis of road traffic with vehicle tracking data. In *Proceedings of the 2009 AAAI Spring Symposium*, pages 7–12, Menlo Park, CA, USA, 2009. The AAAI Press.

H. Ebbinghaus. *Memory: A contribution to experimental psychology*. Dover, New York, 1885.

E. Eberlein, U. Keller, and K. Prause. New insights into smile, mispricing and value at risk: the hyperbolic model. *Journal of Business*, 71(3):371–405, 1998.

R. Elsner, M. Krafft, and A. Huchzermeier. Optimizing Rhenania's mail-order business through Dynamic Multilevel Modeling (DMLM). *Interfaces*, 33(1):50–66, 2003.

R. Engle and J. Russell. Autoregressive conditional duration: A new model for irregularly spaced transaction data. *Econometrica*, 66(5):1127–1162, 1998.

M. Evans, N. Hastings, and B. Peacock. *Statistical Distributions*. Wiley, New York, 3rd edition, 2000.

J. A. Feigenbaum and P. Freund. Discrete scaling in stock markets before crashes. *International Journal of Modern Physics B*, 10(27):3737–3745, 1996.

T. S. Ferguson. Bayesian density estimation by mixtures of normal distributions. In *Recent Advances in Statistics*, pages 287–302, New York, 1983. Academic Press.

V. A. Filimonov and D. Sornette. Self-excited multifractal dynamics. *Europhysics Letters*, 94(4): 46003, 2011.

H. Föllmer and A. Schied. *Stochastic Finance: An Introduction in Discrete Time*. Walter de Gruyter, Berlin, Germany, 2nd edition, 2004.

E. W. Forgy. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*, 21(3):768–769, 1965.

M. Fosgerau and D. Fukuda. Valuing travel time variability: Characteristics of the travel time distribution on an urban road. *Transportation Research Part C*, 24:83–101, 2012.

Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315(5814): 972–976, 2007.

J. H. Friedman. Greedy function approximation: A gradient boosting machine. 1999.

D. Fudenberg. Word-of-mouth communication and social learning. *Quarterly Journal of Economics*, 109:93–125, 1995.

J. V. Gael, Y. Saatci, Y. Teh, and Z. Ghahramani. Beam sampling for the infinite hidden Markov model. In A. McCallum and S. Roweis, editors, *Proceedings of the 25th Annual International Conference on Machine Learning (ICML 2008)*, pages 1088–1095, Helsinki, Finland, 2008. Omnipress.

W. Goffman and V. A. Newill. Generalization of epidemic theory: An application to the transmission of ideas. *Nature*, 204(4955):225–228, 1964.

G. Gómez-Pérez, J. D. Martín-Guerrero, E. Soria-Olivas, E. Balaguer-Ballester, A. Palomares, and N. Casariego. Assigning discounts in a marketing campaign by using reinforcement learning and neural networks. *Expert Systems with Applications*, 36(4):8022–8031, 2009.

O. Grygorash, Y. Zhou, and Z. Jorgensen. Minimum spanning tree based clustering algorithms. In *Proceedings of the 18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI '06)*, pages 73–81, Washington, DC, USA, 2006. IEEE Computer Society.

A. Gunawardana, C. Meek, and P. Xu. A model for temporal dependencies in event streams. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 1962–1970. Curran Associates, Inc., 2011.

J. E. Gustafsson. A money-pump for acyclic intransitive preferences. *Dialectica*, 64(2):251–257, 2010.

B. Hammer and K. Gersmann. A note on the universal approximation capability of support vector machines. *Neural Process. Letters*, 17(1):43–53, 2003.

M. R. Hardy and J. L. Wirch. The iterated CTE: A dynamic risk measure. *North American Actuarial Journal*, 8(4):62–75, 2004.

A. Hawkes. Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society. Series B*, 33(3):438–443, 1971.

C. hsin Wu, J. ming Ho, and D. T. Lee. Travel-time prediction with support vector regression. *IEEE Transactions on Intelligent Transportation Systems*, 5:276–281, 2004.

J. Huber, J. W. Payne, and C. Puto. Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *Journal of Consumer Research*, 9:90–98, 1982.

T. Idé and S. Kato. Travel-time prediction using gaussian process regression: A trajectory-based approach. In *Proceedings of the Ninth SIAM International Conference on Data Mining (SDM 2009)*, pages 1185–1196, 2009.

T. Idé and M. Sugiyama. Trajectory regression on road networks. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI 2011)*, pages 203–208, 2011.

H. Ishwaran and L. F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.

M. Jacobsen. *Point Process Theory and Applications*. Birkhäuser, Boston, MA, 2006.

A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice-Hall, 1981.

D. Karlis. An EM type algorithm for maximum likelihood estimation of the normalinverse Gaussian distribution. *Statistics & Probability Letters*, 57(1):43–52, 2002.

R. Kivetz, O. Netzer, and V. S. Srinivasan. Alternative models for capturing the compromise effect. *Journal of Marketing Research*, 41(3):237–257, 2004.

C. Kolsarici and D. Vakratsas. The complexity of multi-media effects. Working Paper Series, Report No. 11-100, 2011.

R. I. Kondor and J. Lafferty. Diffusion kernels on graphs and other discrete input spaces. In C. Sammut and A. Hoffmann, editors, *Proceedings of the Nineteenth International Conference on Machine Learning (ICML 2002)*, pages 315–322. Morgan Kaufmann, 2002.

H. W. Kuhn and A. W. Tucker. Nonlinear programming. In *Proceedings of 2nd Berkeley Symposium*, pages 481–492, Berkeley, CA, 1951. University of California Press.

V. Kumar, S. Sriram, A. Luo, and P. Chintagunta. Assessing the effect of marketing investments in a business marketing context. *Marketing Science*, 30(5):924–940, 2011.

K. Kurihara, M. Welling, and Y. W. Teh. Collapsed variational Dirichlet process mixture models. In *Proceedings of the International Joint Conference on Artificial Intelligence*, volume 20, 2007.

D. Lashkari and P. Golland. Convex clustering with exemplar-based models. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 825–832. MIT Press, Cambridge, MA, 2008.

B. Libai, E. Muller, and R. Peres. Decomposing the value of word-of-mouth seeding programs: Acceleration versus expansion. *Journal of Marketing Research*, 50(2):161–176, 2013.

F. Liese and K.-J. Miescke. *Statistical Decision Theory: Estimation, Testing, and Selection*. Springer, 2008.

J. Louviere. Conjoint analysis modeling of stated preferences: A review of theory, methods, recent developments and external validity. *Journal of Transport Economics and Policy*, 22(1):93–119, 1988.

A. C. Lozano, N. Abe, Y. Liu, and S. Rosset. Grouped graphical granger modeling methods for temporal causal modeling. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 577–586, New York, NY, USA, 2009. ACM.

R. D. Luce. *Individual Choice Behavior: A Theoretical Analysis*. Wiley, New York, 1959.

R. D. Luce. The choice axiom after twenty years. *Journal of Mathematical Psychology*, 15(3):215–233, 1977.

J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297, 1967.

B. B. Mandelbrot. The variation of certain speculative prices. *Journal of Business*, 36(4):394–419, 1963.

A. A. Markov. The theory of algorithms. *Trudy Mat. Inst. Steklov.*, 42:3–375, 1954.

D. L. McFadden. Econometric models of probabilistic choice among products. *Journal of Business*, 53(3):13–29, 1980.

G. J. McLachlan and T. Krishnan. *The EM algorithm and extensions*. Wiley, New York, 1997.

E. Miller-Hooks and H. S. Mahmassani. Least expected time paths in stochastic, time-varying transportation networks. *Transportation Science*, 34(2):198–215, 2000.

B. Minasny and A. B. McBratney. The matérn function as a general model for soil variograms. *Geoderma*, 128(34):192–207, 2005.

T. Minka. Estimating a Dirichlet distribution. Technical report, Microsoft Research, 2003.

E. A. Nadaraya. On estimating regression. *Theory of Probability and its Applications*, 9(1):141–142, 1964.

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

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

J. Nash. Non-cooperative games. *The Annals of Mathematics*, 54(2):286–295, 1951.

P. Newson and J. Krumm. Hidden Markov map matching through noise and sparseness. In *Proceedings of the Seventeenth ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 336–343, Seattle, Washington, 2009. ACM.

U. Nodelman, C. R. Shelton, and D. Koller. Continuous time Bayesian networks. In *The 18th Conference on Uncertainty in Artificial Intelligence (UAI 2002)*, 2002.

R. B. Noland and J. W. Polak. Travel time variability: a review of theoretical and empirical issues. *Transport Reviews*, 22(1):39–54, 2002.

B. of Governors of the Federal Reserve System. Selected interest rates, 2010. URL `http://www.federalreserve.gov/releases/h15/2010.htm`.

T. Osogami. Iterated risk measures for risk-sensitive Markov decision processes with discounted cost. In *The Twenty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, 2011.

T. Osogami. Robustness and risk-sensitivity in Markov decision processes. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 233–241. Curran Associates, Inc., Lake Tahoe, NV, 2012.

S. Ossen and S. P. Hoogendoorn. Driver heterogeneity in car following and its impact on modeling traffic dynamics. *Journal of the Transportation Research Board*, 1999:95–103, 2007.

T. Otter, G. M. Allenby, and T. van Zandt. An integrated model of discrete choice and response time. *Journal of Marketing Research*, 45(5):593–607, 2008.

O. Papaspiliopoulos and G. O. Roberts. Retrospective MCMC for Dirichlet process hierarchical models. *Biometrika*, 95:169–186, 2008.

A. Piunovskiy. Dynamic programming in constrained Markov decision process. *Control and Cybernetics*, 35(3):646–660, 2006.

J. C. Platt. Fast training of support vector machines using sequential minimal optimization. In *Advances in kernel methods: support vector learning*, pages 185–208. MIT Press, Cambridge, MA, USA, 1999.

J. C. Príncipe. *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*. Springer, 2010.

J. Quinlan. Improved use of continuous attributes in c4.5. *Journal of Artificial Intelligence Research*, 4:77–90, 1996.

M. Radovanović, A. Nanopoulos, and M. Ivanović. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11:2487–2531, 2010.

S. Rajaram, T. Graepel, and R. Herbrich. Poisson-networks: A model for structured point processes. In *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics (AISTATS 2005)*, 2005.

K. Raman, M. Mantrala, S. Sridhar, and E. Tang. Optimal resource allocation with time-varying marketing effectiveness, margins and costs. *Journal of Interactive Marketing*, 40(1):43–52, 2012.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

P. Ravikumar, M. J. Wainwright, and J. D. Lafferty. High-dimensional ising model selection using ℓ_1-regularized logistic regression. *Annals of Statistics*, 38(3):1287–1319, 2010.

R. A. Redner and H. F. Walker. Mixture densities, maximum likelihood, and the EM algorithm. *SIAM Review*, 26(2):195–239, 1984.

H. Robbins. An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 1: Contributions to the Theory of Statistics, pages 157–163, 1956.

S. Robinson and J. W. Polak. Modelling urban link travel time with inductive loop detector data using the k-NN method. *Journal of the Transportation Research Record*, 1935:47–56, 2005.

R. M. Roe, J. R. Busemeyer, and J. T. Townsend. Multialternative decision field theory: A dynamic connectionist model of decision making. *Psychological Review*, 108(2):370–392, 2001.

B. M. Roehner and D. Sornette. "thermometers" of speculative frenzy. *The European Physical Journal B - Condensed Matter and Complex Systems*, 16(4):729–739, 2000.

S. Rogers and P. Langley. Personalized driving route recommendations. In *Proceedings of the American Association of Artificial Intelligence Workshop on Recommender Systems*, pages 96–100. Madison,WI, 1998.

K. Rose. Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. In *Proceedings of the IEEE*, volume 86, pages 2210–2239, 1998.

K. Ross and R. Varadarajan. Markov decision processes with sample path constraints: the communicating case. *Operations Research*, 37:780–790, 1989.

K. Ross and R. Varadarajan. Multichain Markov decision processes with a sample path constraint: a decomposition approach. *Mathematics of Operations Research*, 16:195–207, 1991.

D. C. Rubin and A. E. Wenzel. One hundred years of forgetting: A quantitative description of retention. *Psychological Review*, 103:734–760, 1996.

A. Ruszczynski. Risk averse dynamic programming for Markov decision process. *Mathematical Programming, Series B*, 125(2):235–261, 2010.

R. E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37(3):297–336, 1998.

B. Scheibehenne, J. R. J, and C. González-Vallejo. Cognitive models of choice: comparing decision field theory to the proportional difference model. *Cognitive Science*, 33(5):911–939, 2009.

D. C. Schmittlein, D. G. Morrison, and R. Colombo. Counting your customers: Who are they and what will they do next? *Management Science*, 33(1):1–24, 1987.

D. Schultz, M. Block, and K. Raman. Understanding consumer-created media synergy. *Special Issue on Cross-Media and Cross-Tool Effects, Journal of Marketing Communications*, 18(3):173–187, 2012.

D. W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley and Sons, Chichester, New York, 1992.

R. Serfozo. Technical note - an equivalence between continuous and discrete time Markov decision processes. *Operations Research*, 27(3):616–620, 1979.

G. R. Shorack and J. A. Wellner. *Empirical Processes with Applications to Statistics*. Wiley, New York, 1986.

S. M. Siddiqi. Fast state discovery for HMM model selection and learning. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics (AI-STATS 2007)*, 2007.

B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London, United Kingdom, 1986.

A. Simma and M. I. Jordan. Modeling events with cascades of Poisson processes. In *The 26th Conference on Uncertainty in Artificial Intelligence (UAI 2010)*, pages 546–555, 2010.

B. Simon. *Functional Integration and Quantum Physics*. Academic Press, 1979.

H. A. Simon. *Administrative Behavior*. Macmillan, New York, 1947.

I. Simonson. Choice based on reasons: The case of attraction and compromise effects. *Journal of Consumer Research*, 16:158–174, 1989.

K. A. Small and C. Winston. *The Demand for Transportation: Models and Applications*. Brookings Institution Press, Washington, DC, 1999.

K. A. Small and J. Yan. The value of "value pricing" of roads: Second-best pricing and product differentiation. *Journal of Urban Economics*, 49(2):310–336, 2001.

K. A. Small, C. M. Winston, and J. Yan. Uncovering the distribution of motorists' preferences for travel time and reliability. *Econometrica*, 73(4):1367–1382, 2005.

D. Sornette and G. Ouillon. Multifractal scaling of thermally activated rupture processes. *Phys. Rev. Lett.*, 94(3), 2005.

D. Sornette, A. Johansen, and J. P. Bouchaud. Stock market crashes, precursors and replicas. *J. Phys. I Finance*, 6:167–175, 1996.

M. L. Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, 1999.

M. Sugiyama, T. Suzuki, S. Nakajima, H. Kashima, P. von Bünau, and M. Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60 (4):699–746, 2008.

R. H. Thaler. Some empirical evidence on dynamic inconsistency. *Economic Letters*, 8(3):201–207, 1981.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1994.

M. E. Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1, 2001.

G. Tirenni, C. Kaiser, and A. Herrmann. Applying decision trees for value-based customer relations management: Predicting airline customers' future values. *Journal of Database Marketing and Customer Strategy Management*, 14:130–142, 2007a.

G. Tirenni, A. Labbi, C. Berrospi, A. Elisseeff, T. Bhose, K. Pauro, and S. Pöyhönen. The 2005 ISMS practice prize winner - Customer Equity and Lifetime Management (CELM) Finnair case study. *Marketing Science*, 26(4):553–565, 2007b.

E. Todorov. Linearly-solvable Markov decision problems. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 1369–1376. MIT Press, 2007.

E. Todorov. Eigenfunction approximation methods for linearly-solvable optimal control problems. In *Proceedings of IEEE International Symposium on Adaptive Dynamic Programming and Reinforcemenet Learning (ADPRL 2009)*, pages 161–168, 2009.

J. T. Townsend and F. G. Ashby. *Stochastic Modeling of Elemenatary Psychological Processes*. Cambridge University Press, Cambridge, 1983.

K. Train and C. Winston. Vehicle choice behavior and the declining market share of U.S. automakers. *International Economic Review*, 48(4):1469–1496, 2007.

A. Tversky. Elimination by aspects: A theory of choice. *Psychological Review*, 79:281–299, 1972.

N. Ueda and R. Nakano. Deterministic annealing EM algorithm. *Neural Networks*, 11(2):271–282, 1998.

P. Ulleberg and T. Rundmo. Personality, attitudes and risk perception as predictors of risky driving behavior among young drivers. *Safety Science*, 41(5):427–443, 2003.

V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

C. Vens and F. Costa. Random forest based feature induction. In *Proceedings of the 11th IEEE International Conference on Data Mining (ICDM 2011)*, volume 11, pages 744–753, Vancouver, BC, 2011.

E. T. Verhoef and K. A. Small. Product differentiation on roads: Constrained congestion pricing with heterogeneous users. *Journal of Transport Economics and Policy*, 38(1):127–156, 2004.

J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ, 1953.

U. Wagner and A. Taudes. A multivariate Polya model of brand choice and purchase incidence. *Marketing Science*, 5(3):219–244, 1986.

S. Walker. Sampling the Dirichlet mixture model with slices. *Communications in Statistics. Simulation and Computation*, 36:45–54, 2007.

M. Wardman. A review of british evidence of time and service quality valuations. *Transportation Research E: Logistics and Transportation Review*, 37(2-3):107–128, 2001.

G. S. Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 26(4): 359–372, 1964.

Y. Wen, T. Lee, and H. Cho. Hybrid models toward traffic detector data treatment and data fusion. In *Proceedings of Networking, Sensing and Control 2005*, pages 525–530, 2005a.

Y.-H. Wen, T.-T. Lee, and H.-J. Cho. Missing data treatment and data fusion toward travel time estimation for ATIS. *Journal of the Eastern Asia Society for Transportation Studies*, 6:2546–2560, 2005b.

J. T. Wixted. Analyzing the empirical course of forgetting. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 16:927–935, 1990.

J. T. Wixted and E. B. Ebbesen. Genuine power curves in forgetting: A quantitative analysis of individual subject forgetting functions. *Memory and Cognition*, 25:731–739, 1997.

L. Xu and M. I. Jordan. On convergence properties of the EM algorithm for Gaussian mixtures. *Neural Computation*, 8(1):129–151, 1995.

C. Yang, R. Duraiswami, N. A. Gumerov, and L. Davis. Improved fast gauss transform and efficient kernel density estimation. In *Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV'03)*, volume 1, pages 664–671, 2003.

J. Yu, G. Chang, H. Ho, and Y. Liu. Variation based online travel time prediction using clustered neural networks. In *Proceedings of the Eleventh International IEEE Conference on Intelligent Transportation Systems*, pages 85–90, 2008.

C. Zahn. Graph-theoretical methods for detecting and describing gestalt clusters. *IEEE Transactions on Computers*, C-20(1):68–86, 1971.

L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 1601–1608. MIT Press, Cambridge, MA, 2005. URL `http://www.vision.caltech.edu/lihi/Demos/SelfTuningClustering.html`.

G. Zhu and X. Wang. Study on route travel time prediction based on RBF neural network. In *2009 First International Workshop on Education Technology and Computer Science*, pages 1118–1122, 2009.

B. Zuckerman and D. Jefferson. *Human Population and the Environmental Crisis*. Jones & Bartlett Learning, 1996.

Publications

R. Takahashi. Sequential minimal optimization in adaptive-bandwidth convex clustering. In *Proceedings of the 11th SIAM International Conference on Data Mining (SDM 2011)*, pages 896–907, 2011.

R. Takahashi. Sequential minimal optimization in convex clustering repetitions. *Statistical Analysis and Data Mining*, 5(1):70–89, 2012.

R. Takahashi, T. Osogami, and T. Morimura. Large-scale nonparametric estimation of vehicle travel time distributions. In *Proceedings of the 12th SIAM International Conference on Data Mining (SDM 2012)*, pages 12–23, 2012.

R. Takahashi, H. Mizuta, N. Abe, R. L. Kennedy, V. J. Jeffs, R. Shah, and R. H. Crites. Collective response spike prediction for mutually interacting consumers. In *Proceedings of the 13th IEEE International Conference on Data Mining (ICDM 2013)*, volume 13, pages 727–736, Dallas, TX, 2013.

R. Takahashi, T. Yoshizumi, H. Mizuta, N. Abe, R. L. Kennedy, V. J. Jeffs, R. Shah, and R. H. Crites. Multi-period marketing-mix optimization with response spike forecasting. *IBM Journal of Research and Development*, 2014. to appear.