# Predicting Preference Reversals via Gaussian Process Uncertainty Aversion

**Rikiya Takahashi**
IBM Research - Tokyo
Rikiya.Takahashi@gmail.com

**Tetsuro Morimura**
IBM Research - Tokyo
tetsuro@jp.ibm.com

## Abstract

Modeling of a product or service's attractiveness as a function of its own attributes (e.g., price and quality) is one of the foundations in econometric forecasts, which have been provided with an assumption that each human rationally has a consistent preference order among his choice decisions. Yet the preference orders by real humans become irrationally reversed, when the choice set of available options is manipulated. In order to accurately predict choice decisions involving preference reversals, which existing econometric methods have failed to incorporate, the authors introduce a new cognitive choice model whose parameters are efficiently fitted with a global convex optimization algorithm. The proposed model captures each human as a Bayesian decision maker facing a mental conflict between objective evaluation samples and a subjective prior, where the underlying objective evaluation function is rationally independent from contexts while the subjective prior is irrationally determined by each choice set. As the key idea to analytically handle the irrationality and to yield the convex optimization, the Bayesian decision mechanism is implemented as a closed-form Gaussian process regression using similarities among the available options in each context. By explaining the irrational decisions as a consequence of averting uncertainty, the proposed model outperformed the existing econometric models in predicting the irrational choice decisions recorded in real-world datasets.

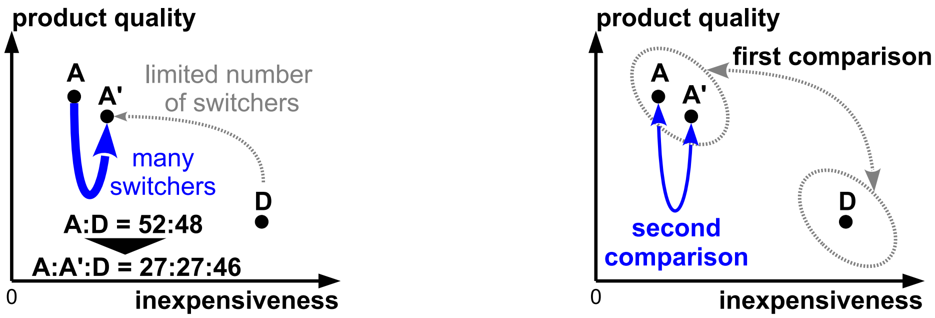
## 1 INTRODUCTION

Accurately predicting which option a human prefers to the other alternatives based on their attributes is one of the central interests in social science. Random utility maximization approaches have been adopted in lots of econometric applications, such as product design (Brownstonea et al., 2000), demand forecasting (Train and Winston, 2007; Frischknecht et al., 2010), and various marketing problems as summarized in (Chandukala et al., 2007). These approaches have assumed independence among every option's attractiveness, based on the rationale that attributes of unchosen options are irrelevant to the chosen option's benefit. If this rationale holds probabilistically, each decision maker must have his own random utility function that satisfies probabilistic Independence from Irrelevant Alternatives (IIA) (Luce, 1959). Random utility function quantifies the attractiveness of an option as random noise plus a function of only its own attributes, such as costs and benefits (e.g., Gaussian-distributed probit (Louviere, 1988) or Gumbel-distributed logit (McFadden, 1980)). By applying advanced functional approximators (e.g., (Chu and Ghahramani, 2005)), machine learners have played essential roles in estimating random utility functions.

### 1.1 Context Effects as Irrational Behavior

In the real world, however, experimental psychologists have clarified the dependence of an option's attractiveness on the other options, which is referred to as 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). The actual choice by real human is affected by the set of available options, which is called the choice set. As we show in Figure 1, attractiveness for each of multiple options is observed to be not independent but *correlated* with one another. One possible reason of such correlation is humans' limited abilities in processing perceived information, where grouping of

Appearing in Proceedings of the $18^{th}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38.



Change of shares by introducing option A$'$

Interpretation as grouping of options

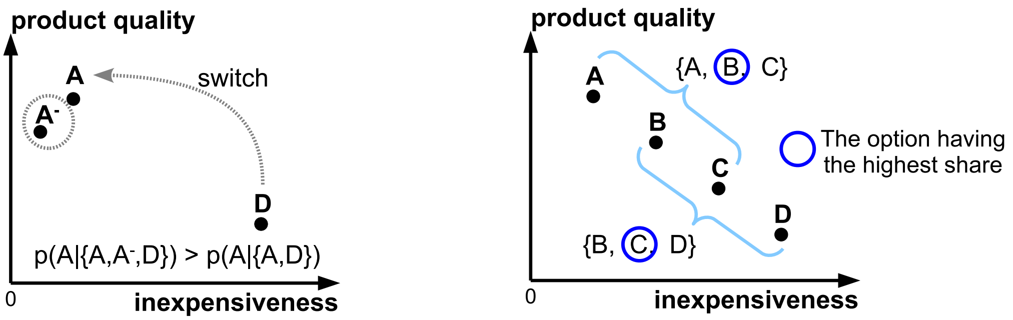
Figure 1: The similarity effect (Tversky, 1972) as a violation of probabilistic IIA. When option A$'$ is closer to option A than to option D, introduction of option A$'$ grabs market share more from option A than from option D. Then option D often has the highest share among the three options, even when option A was more preferred to option D in the two-option context. As shown in the right figure, one plausible mechanism behind the similarity effect is mental grouping of *available* options. Choosers of option A or A$'$ are supposed to experience two-step evaluation processes, whose first step is comparison of option D with a set {A,A$'$}, and whose second step is choice of one option from the set {A,A$'$}. Such hierarchical evaluation can be regarded as an outcome of correlation among the random utilities, where a pair of similar options is assigned high correlation.

options into a smaller number of categories can reduce the high mental workload of independently evaluating every option. An interesting phenomena caused by the correlation is called the similarity effect, with which an option grabs market share more from similar ones than from dissimilar ones. In the literature, various models of the correlated random utilities have been proposed, such as full-covariance probit (Hausman and Wise, 1978; Louviere, 1988), nested logit (Williams, 1977), mixed multinomial logit (McFadden and Train, 2000; Brownstonea et al., 2000; Glasgow, 2001; Karabatsos and Walkerbbook, 2012), and generalized nested logit (Wen and Koppelman, 2001). These predictive models are designed to accurately fit the structure of correlation, and have provided successful forecasts in practice.

Unlike the similarity effect, the attraction and the compromise effects cannot be explained solely with correlation. In the cases of these two context effects, Figure 2 illustrates how reversals of preference orders occur when the choice sets are manipulated. For instance of abuse in marketing, such manipulations force parts of consumers to select not what they want but what firms intend to sell for high profits (e.g., an online shopping case in (Kivetz et al., 2004)). In statistical viewpoints, these context effects imply nonexistence of the consistent mean utility function, and hence strikingly shake the foundation of many economic implications that have implicitly adopted consistent mean utility functions. The non-existence of mean utility functions may also cause unpleasant feelings for many machine learners, because most of functional approximators as their main expertise fail to predict the actual choices, as long as they view choice decision as a result of scoring based on consistent mean plus random noise.

## 1.2 Discrete Choice Models as Related Work

Despite the necessity of quantitatively predicting the total impacts by all of the context effects, the literature of choice modeling provides neither generalization capability for choice sets that do not appear in training data, nor tractable learning algorithms. A promising direction to handle the preference reversals is to model the correlation as a function of option attributes, and/or to introduce relative features (e.g., price difference) in forming the mean utility function. Many of existing correlation models except the structured probit (Yai, 1997; Dotson et al., 2009) and mixed multinomial logit, however, formalize the correlation not as a function of option attributes, but as a constant parameter that cannot be generalized for the contexts that do not occur in training data. Custom engineering of the relative features adopted in the Proportional Difference Model (González-Vallejo, 2002) and a marketing study (Kivetz et al., 2004) requires ad hoc selection of variable pairs, though the desirable selection criteria have not yet been clarified. Multialternative Decision Field Theory (MDFT) (Roe et al., 2001) is able to explain the three context effects, and experimentally exhibits high predictive powers (Scheibehenne et al., 2009; Berkowitsch et al., 2014) thanks to the combination of correlation and relative features. Unfortunately, optimizing the parameters of MDFT is a hard problem (Chandukala et al., 2007), and applying machine learning algorithms for MDFT is not promising. Another Bayesian modeling is proposed in (Shenoy and Yu, 2013), while this work neither provides effective learning algorithms.



Attraction effect

Compromise effect

Figure 2: Preference reversals with the attraction (Huber et al., 1982) and the compromise (Simonson, 1989) effects. Options (A,B,C,D) are on a two-dimensional Pareto efficient frontier. For a market in which options A and D exist, introduction of option $A^-$, which is absolutely inferior to option A, shifts portions of shares from option D to option A while option $A^-$ is never chosen. In each choice set containing three options, the moderate option obtains the highest share. Preference order for options A and D, and that for options B and C are reversed with these context effects.

### 1.3 Bayesianism to Represent Irrationality

We propose a new choice model that is able to predict the preference reversals for choice sets that are not contained in training data, and whose parameters are stably fitted with global convex optimization. The core component required in predicting the preference reversals is a systematic feedback mechanism from correlation or variance into the expected utility. To directly link the mean with correlation, we borrow an idea of regarding each human as a Bayesian *shrinkage* estimator of utility (Natenzon, 2010), while this prior work does not model the mean and correlation as functions of option attributes. Our Bayesianism captures each human as a decision maker facing a mental conflict between a subjective prior that essentially generates the irrationality, and objective evaluation samples whose values are rationally determined by independent application of one consistent evaluation function for each option's own attributes. One core trick to generate the preference reversals is dependence of the subjective prior on the limited number of available options in each choice set, and this set-dependent prior naturally embodies the required feedback from correlation into mean. Because the expected attractiveness of an option dissimilar to the others is strongly shrunk into the prior mean, our approach generates the attraction and compromise effects as a result of uncertainty aversion by each decision maker.

In order to yield tractable predictions and the convex optimization, we further add one key assumption that the Bayesian decision making process involves a Gaussian Process Regression (GPR), in addition to the Gumbel-distributed random noises. The training samples in this GPR are the objective evaluations as results of applying the consistent evaluation function for the attributes of every option, and hence the total attractivenesses of all of the options are *simultaneously* evaluated through simple matrix operations. The set-dependent subjective prior is formalized with similarities among the available options. Our GPR's tricky embedding of the evaluation function in *labels*, instead of providing constant label data as in ordinary GPRs, leads a unique optimization objective that still consists of simple matrix multiplications and whose convexity is guaranteed.

The remainder of this paper is organized as follows. Section 2 introduces our cognitive GPR mechanism and the uncertainty-aversion interpretation to explain the context effects. Then Section 3 addresses the optimization formula to fit the parameters. The experimental prediction results using real-world datasets appear in Section 4, and Section 5 concludes the paper.

## 2 PRODUCING CONTEXT EFFECTS VIA GAUSSIAN PROCESS REGRESSION

The prediction formulas of our model are provided in this section. We define the choice prediction task in Section 2.1. Then we capture the irrational choice as a conflict between the objective evaluations and a subjective prior introduced in Section 2.2. Section 2.3 clarifies how the proposed model yields the context effects, while Section 2.4 discusses a cognitive interpretation of our model.

### 2.1 Stochastic Choice Containing Logit as a Special Case

Assume that a decision maker in a context indexed by $i \in \mathbb{N}$ is shown $m[i]$ available options (e.g., products or services) denoted as a choice set $\mathcal{M}_i = \{1, \ldots, m[i]\}$, and he stochastically chooses one option $y_i \in \mathcal{M}_i$. The characteristics of the decision maker (e.g., demographics and queries to retrieve the choice set) is given as a vector $\boldsymbol{q}_i \in \mathbb{R}^{d_Q}$, and each option $j \in \mathcal{M}_i$ is associated with a vector $\boldsymbol{r}_{ij} \in \mathbb{R}^{d_R}$ to represent its attributes (e.g., price and quality). For notational convenience, we define an input vector $\boldsymbol{x}_{ij}$ by a concatenation $\boldsymbol{x}_{ij} \triangleq (\boldsymbol{q}_i^\top, \boldsymbol{r}_{ij}^\top)^\top \in \mathbb{R}^{d_X}$ such that $d_X \equiv d_Q + d_R$, and $\boldsymbol{X}_i \triangleq (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{im[i]})^\top \in \mathbb{R}^{m[i] \times d_X}$.

We aim to probabilistically predict the choice $y_i$ given the matrix $\boldsymbol{X}_i$ for any context $i$. Let us define the final evaluation function $u : \mathbb{R}^{d_X} \times \mathbb{R}^{m \times d_X} \to \mathbb{R}$ such that $u(\boldsymbol{x}, \boldsymbol{X})$ embodies the total attractiveness of an option whose attributes are represented as input vector $\boldsymbol{x}$, while attributes of the available options are represented by matrix $\boldsymbol{X}$. For notational convenience, we also define $u_i(\boldsymbol{x}) \triangleq u(\boldsymbol{x}, \boldsymbol{X}_i)$ for every context $i$. The final choice $y_i$ is given as $y_i = \arg\max_j (u_i(\boldsymbol{x}_{ij}) + \varepsilon_{ij})$, where $\varepsilon_{ij} \in \mathbb{R}$ is an independent and identically distributed noise variable for option $j$.

Because it is natural to regard an option's attractiveness as depending on its own attributes and characteristics of each decision maker, we also assume that every decision maker has a vector of objective evaluations $\boldsymbol{v}_i \triangleq (v_{i1}, \ldots, v_{im[i]})^\top \in \mathbb{R}^{m[i]}$ computed as

$$v_{ij} = b + \boldsymbol{w}_\phi^\top \boldsymbol{\phi}(\boldsymbol{x}_{ij}), \tag{1}$$

where $b$ is a bias term, $\boldsymbol{\phi} : \mathbb{R}^{d_X \to d_\phi}$ is a mapping function into a $d_\phi$-dimensional feature space, and $\boldsymbol{w}_\phi \in \mathbb{R}^{d_\phi}$ is a vector of absolute importance. Evaluating options only with Eq. (1) is rational, because every option's attractiveness is independent from each other's and manipulation of choice sets never changes the preference orders. In matrix representation, $\boldsymbol{v}_i = b\boldsymbol{1}_{m[i]} + \boldsymbol{\Phi}_i \boldsymbol{w}_\phi$ where $\boldsymbol{\Phi}_i \triangleq (\boldsymbol{\phi}(\boldsymbol{x}_{i1}), \ldots, \boldsymbol{\phi}(\boldsymbol{x}_{im[i]}))^\top \in \mathbb{R}^{m[i] \times d_\phi}$, and $\boldsymbol{1}_m$ is the $m$-dimensional vector whose elements are all one. When the mapping function $\boldsymbol{\phi}$ is non-linear, priority of each option attribute is heterogeneous among the decision makers.

An important aspect in our modeling is that the final evaluation $u_i(\boldsymbol{x}_{ij})$ and the objective evaluation $v_{ij}$ are not always identical. For analytical tractability, we assume the noise $\varepsilon_{ij}$ to obey a type-I extreme value distribution whose probability density function is $p(\varepsilon_{ij}) = \exp(-\varepsilon_{ij} - \exp(-\varepsilon_{ij}))$. Before proceeding into the novel parts of our approach, let us confirm the special case when $\forall j \in \{1, \ldots, m[i]\}$ $u_i(\boldsymbol{x}_{ij}) \equiv v_{ij}$. Then $u_i(\boldsymbol{x}) \equiv u(\boldsymbol{x}) \triangleq b + \boldsymbol{w}_\phi^\top \boldsymbol{\phi}(\boldsymbol{x})$ yields the logit model (McFadden, 1980), whose probability of choosing option $j$ is

$$P(j|\mathcal{M}_i, \boldsymbol{X}_i, \boldsymbol{w}_\phi, \boldsymbol{\phi}) = \frac{\exp(\boldsymbol{w}_\phi^\top \boldsymbol{\phi}(\boldsymbol{x}_{ij}))}{\sum_{j' \in \mathcal{M}_i} \exp(\boldsymbol{w}_\phi^\top \boldsymbol{\phi}(\boldsymbol{x}_{ij'}))} \tag{2}$$

and whose formalism is identical to the well-known multinomial logistic regression. The bias term $b$ is canceled and has no effect at least in the logit model (2), whereas it is essential for generating the preference reversals as we show in the next sections.

## 2.2 Gaussian Process Regression to Shrink the Evaluations

We consider generalized cases in which the final evaluations are different from the objective evaluations. Each decision maker is assumed to be a Bayesian who *estimates* the evaluation function $u_i(\cdot)$, by regarding the vector of objective evaluations $\boldsymbol{v}_i$ as just a *sample* in fitting instead of fully relying on their values. The sample $\boldsymbol{v}_i$ and the final evaluation $u_i(\boldsymbol{x}^*)$ for any test input $\boldsymbol{x}^*$ are perceived as noisy observations around the latent values $\boldsymbol{\mu}_i \in \mathbb{R}^{m[i]}$ and $\mu_i(\boldsymbol{x}^*)$, respectively. This observation process in context $i$ is modeled as

$$\left(\boldsymbol{v}_i^\top, u_i(\boldsymbol{x}^*)\right)^\top \sim \mathcal{N}\left(\left(\boldsymbol{\mu}_i^\top, \mu_i(\boldsymbol{x}^*)\right)^\top, \sigma^2 \boldsymbol{I}_{m[i]+1}\right), \tag{3}$$

where $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the multivariate Gaussian distribution whose mean is $\boldsymbol{\mu}$ and whose variance-covariance matrix is $\boldsymbol{\Sigma}$, $\sigma^2$ is a noise level, and $\boldsymbol{I}_m$ is the $m$-dimensional identity matrix.

When decision makers have limited information and do not fully rely on the sample $\boldsymbol{v}_i$, we imagine that an alternative stochastic process supports their subjective decision making. Our key assumption is that this subjective process is implemented as a Gaussian process prior formed solely with each choice set $\mathcal{M}_i$ and the similarity among the $m[i]$ options. Let $K : \mathbb{R}^{d_X} \times \mathbb{R}^{d_X} \to \mathbb{R}$ be a covariance function to measure the similarity between two options. The actual instance of the covariance function $K(\cdot, \cdot)$ is later introduced in Section 3.2. Then let $\boldsymbol{K}(\boldsymbol{X}_i) \in \mathbb{R}^{m[i] \times m[i]}$ be a variance-covariance matrix whose $(j, j')$ element is $K(\boldsymbol{x}_{ij}, \boldsymbol{x}_{ij'})$ and $\boldsymbol{k}_i(\boldsymbol{x}^*) \triangleq (K(\boldsymbol{x}^*, \boldsymbol{x}_{i1}), \ldots K(\boldsymbol{x}^*, \boldsymbol{x}_{im[i]}))^\top \in \mathbb{R}^{m[i]}$. The subjective Gaussian process prior is

$$\begin{pmatrix} \boldsymbol{\mu}_i \\ \mu_i(\boldsymbol{x}^*) \end{pmatrix} \sim \mathcal{N}\left(\boldsymbol{0}_{m[i]+1}, \sigma^2 \begin{pmatrix} \boldsymbol{K}(\boldsymbol{X}_i) & \boldsymbol{k}_i(\boldsymbol{x}^*) \\ \boldsymbol{k}_i(\boldsymbol{x}^*)^\top & K(\boldsymbol{x}^*, \boldsymbol{x}^*) \end{pmatrix}\right), \tag{4}$$

where $\boldsymbol{0}_m$ denotes the $m$-dimensional zero vector. Using Eqs. (3) and (4), each decision maker obtains a posterior of the final evaluation function as

$$u_i(\boldsymbol{x}^*)|\boldsymbol{v}_i \sim \mathcal{N}\Big(\boldsymbol{k}_i(\boldsymbol{x}^*)^\top \left(\boldsymbol{I}_{m[i]} + \boldsymbol{K}(\boldsymbol{X}_i)\right)^{-1} \boldsymbol{v}_i, \\ \sigma^2 \Big[1 + K(\boldsymbol{x}^*, \boldsymbol{x}^*) - \|\boldsymbol{k}_i(\boldsymbol{x}^*)\|^2_{\boldsymbol{I}_{m[i]} + \boldsymbol{K}(\boldsymbol{X}_i)}\Big]\Big), \tag{5}$$

where $\|\boldsymbol{\beta}\|^2_{\boldsymbol{\Sigma}} \triangleq \boldsymbol{\beta}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\beta}$.

For analytical tractability, every decision maker is assumed to adopt the posterior mean as the final evaluation, while full-Bayesian approaches to sample from

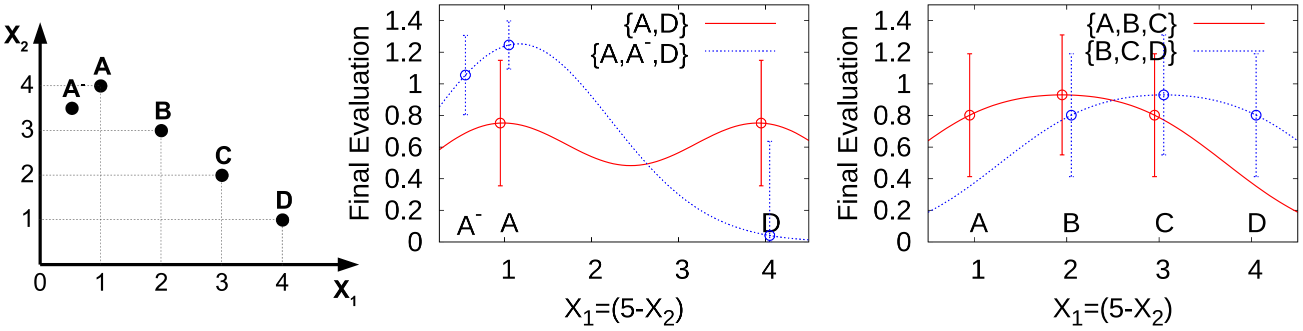


Figure 3: Preference reversals caused by utility shrinkage into the prior mean. For each option in the left-most figure and for each choice set, the mean and variance of the final evaluation function are provided in the right two figures, where $K(\boldsymbol{x},\boldsymbol{x}') = 3\exp(-\frac{1}{4}\|\boldsymbol{x}-\boldsymbol{x}'\|^2)$, $b=1$, $\boldsymbol{w}_\phi = \boldsymbol{0}_{d_\phi}$, and $\sigma = 0.3$. While all of the options have the same objective evaluation $b=1$, the perceptual shrinkage discounts the final evaluations of the options dissimilar to the others. As clarified in the center figure, evaluations of extreme options involve high posterior variance, as well as the strong shrinkage.

the posterior should be considered in the future. Let $\boldsymbol{u}_i^* \triangleq (u_{i1}^*, \ldots, u_{im[i]}^*)^\top \in \mathbb{R}^{m[i]}$ be the posterior mean for the options in choice set $\mathcal{M}_i$. The probability of selecting option $j$ is $\exp(u_{ij}^*)/[\sum_{j'=1}^{m[i]} \exp(u_{ij'}^*)]$. By substituting $\boldsymbol{x}^* = \boldsymbol{x}_{ij}$ into Eq. (5) and by using Eq. (1), we get

$$\boldsymbol{u}_i^* = \boldsymbol{K}(\boldsymbol{X}_i)\left(\boldsymbol{I}_{m[i]} + \boldsymbol{K}(\boldsymbol{X}_i)\right)^{-1}\left(b\boldsymbol{1}_{m[i]} + \boldsymbol{\Phi}_i\boldsymbol{w}_\phi\right). \tag{6}$$

The logit model such that $\boldsymbol{u}_i^* \equiv \boldsymbol{v}_i$ is a special case when the prior is non-informative as $|\boldsymbol{K}(\boldsymbol{X}_i)| \to \infty$. The noise level $\sigma^2$ may matter in fitting the parameters, while it is ignored in predictions as long as we merely use the posterior mean.

### 2.3 Attraction and Compromise Effects as Uncertainty Aversion

In the posterior mean (6), the attraction and compromise effects appear as aversion toward options that do not resemble the others. As we illustrate in Figure 3, Eq. (5) makes the final evaluation of an extreme option strongly shrunk into the prior mean 0, and makes the posterior variance high. The reason of the strong shrinkage is high uncertainty, which is caused by absence of reliable samples for a decision maker to interpolate an evaluation with one another. The shrinkage makes the decision makers uncertainty-averse, as long as the score of the objective evaluations are positive. While shrinking negative objective evaluations makes decision makers uncertainty-seeking, such opposite behaviors are not implied statistically, because of the observed avoidance of extreme options.

We would like to stress that what the decision makers are avoiding are not risks but uncertainties. Someone would say that choosing extreme options looks risky, because higher variance leads to more avoidance. Yet posterior variance does not correspond to risks involving explicit losses such as drop of a stock price, but merely reflects the lack of training samples or prior knowledge. A more plausible description is that outcomes by choosing extreme options are felt to be more uncertain than those by choosing moderate options.

### 2.4 Cognitive Interpretations

For deeper understanding, let us address possible interpretations of the proposed model from the viewpoints of bounded rationality (Rieskamp et al., 2006; Manzini and Mariotti, 2009) and brain structure. Our assumptions may look strange, because each decision maker possesses double personality involving both the rational objective evaluations and the irrational subjective prior. Parts of these decision makers discard the merits of the rationality that the objective evaluations provide.

By expressing understanding of the high uncertainty in the real world, we partly defend humans exhibiting the context effects, whose vulnerability to intentional manipulations of the choice sets looks irrational for economists. The decision makers exploit available information as maximally as possible, may possess "wisdom of unknown," and are different from the true naïve who considers nothing. Humans sometimes neither know which option attributes are *absolutely* important for them, nor rely on their poor utility functions. Another clue they can exploit is merely the *relative* similarity among the existing options they know, i.e., the current choice set. The vulnerability to manipulation is an inevitable compensation in discounting the dubious objective evaluations. Remember that our model can also produce rational choices satisfying the probabilistic IIA, when non-informative prior is placed.

Our model is an implementation of the dual process theory (Barrett et al., 2004), which captures decision making as a conflict between two fictitious characters, System 1 and System 2, that yield quick intuitions and time-consuming deliberations, respectively. Such a conflicting dual personality is a radical simplification of the interaction among multiple brain tissues, where System 1 and System 2 mimic older and newer regions such as the limbic system and the cerebral neocortex, respectively.

# 3 OPTIMIZATION AND CHOICE OF THE PARAMETERS

This section discusses the efficient fitting and choice of the model parameters. Section 3.1 derives the convex optimization objective in fitting the objective evaluation function. Then Section 3.2 discusses the choices of the mapping function $\boldsymbol{\phi}(\cdot)$ and the covariance function $K(\cdot,\cdot)$, for the objective evaluation function, and the similarity between options, respectively.

## 3.1 Convex Optimization of the Objective Evaluation Function

Our training data consist of $n$ contexts and choices $(\boldsymbol{X}_i, y_i)_{i=1}^n$. For each context $i$, let us denote its log-likelihood term by $\ell(\boldsymbol{u}_i^*, y_i) \triangleq u_{iy_i}^* - \log(\sum_{j'=1}^{m[i]} \exp(u_{ij'}^*))$ and its vector of choice probabilities by $\boldsymbol{g}(\boldsymbol{u}_i^*) \triangleq (\exp(u_{i1}^*), \ldots, \exp(u_{im[i]}^*))^\top / [\sum_{j=1}^{m[i]} \exp(u_{ij}^*)]$. Based on Eq. (6), let $\boldsymbol{H}_i \triangleq \boldsymbol{K}(\boldsymbol{X}_i)(\boldsymbol{I}_{m[i]} + \boldsymbol{K}(\boldsymbol{X}_i))^{-1}$ be a constant matrix during the updates. With placing a Gaussian prior in order to avoid over-fitting of the vector $\boldsymbol{w}_\phi$, we perform a MAP estimation by solving an optimization problem

$$\max_{b,\boldsymbol{w}} \sum_{i=1}^{n} \ell(b\boldsymbol{H}_i \boldsymbol{1}_{m[i]} + \boldsymbol{H}_i \boldsymbol{\Phi}_i \boldsymbol{w}) - \frac{c}{2}\|\boldsymbol{w}_\phi\|^2, \qquad (7)$$

where $\|\cdot\|$ is the $L_2$-norm and $c$ is a regularization hyperparameter. Thanks to the log-concavity of the likelihood of multinomial logistic regression, Optimization (7) is convex with respect to the parameters $(b, \boldsymbol{w})$, whose global optimum is attained with Newton-Raphson methods. This convexity is a nice property of the proposed GPR model, which quickly predicts the context effects solely by multiplying the similarity matrix for the vector of the objective evaluations.

## 3.2 Selection of the Mapping and the Covariance Functions

In the actual implementation, we adopt a Radial-Basis-Function (RBF) for each of the mapping function $\boldsymbol{\phi}$ and the covariance function $K(\cdot,\cdot)$.

Like the diminishing returns observed in real economy, attractiveness of an option is usually non-linear to the increase of each option attribute. One of the simplest way to universally incorporate such non-linearity is to adopt nonparametrics using RBF kernels. We adopt a mapping function $\boldsymbol{\phi}$ whose inner product becomes an isotropic RBF. After standardizing every element of the input vectors $((\boldsymbol{x}_{ij})_{j=1}^{m[i]})_{i=1}^n$ to have the unit variance, the mapping function $\boldsymbol{\phi}$ is chosen to realize $\boldsymbol{\phi}(\boldsymbol{x})^\top \boldsymbol{\phi}(\boldsymbol{x}') = \exp(-\frac{\gamma}{2}\|\boldsymbol{x} - \boldsymbol{x}'\|^2)$, where $\gamma$ is a bandwidth hyperparameter.

For computational efficiency, we explicitly introduce a finite-dimensional mapping function instead of handling only the inner products as in many kernel machines. An isotropic RBF kernel is efficiently approximated with random Fourier features (Rahimi and Recht, 2008). Let us distribute $L < n$ vectors $(\boldsymbol{\beta}_\ell \in \mathbb{R}^{d_X})_{\ell=1}^L$ such that $\forall \ell\ \boldsymbol{\beta}_\ell \sim \mathcal{N}(\boldsymbol{0}_{d_X}, \boldsymbol{I}_{d_X})$. Every inner product between two mapping functions designed as

$$\begin{aligned}\mathbb{R}^{2L} \ni \boldsymbol{\phi}(\boldsymbol{x}) = L^{-1/2}(&\cos(\sqrt{\gamma}\boldsymbol{\beta}_1^\top \boldsymbol{x}), \ldots, \cos(\sqrt{\gamma}\boldsymbol{\beta}_L^\top \boldsymbol{x}),\\ &\sin(\sqrt{\gamma}\boldsymbol{\beta}_1^\top \boldsymbol{x}), \ldots, \sin(\sqrt{\gamma}\boldsymbol{\beta}_L^\top \boldsymbol{x}))^\top\end{aligned}$$

converges into an RBF kernel when $L \to \infty$. The proof is provided in (Rahimi and Recht, 2008), and we adopt $L = 100$.

As the covariance function $K(\cdot,\cdot)$, we adopt another isotropic RBF kernel parametrized as

$$K(\boldsymbol{x}, \boldsymbol{x}') = \frac{1}{\eta} \exp\left(-\frac{\xi}{2}\|\boldsymbol{x}' - \boldsymbol{x}\|^2\right), \qquad (8)$$

where $\eta$ is a scaling hyperparameter and $\xi$ is another bandwidth hyperparameter. The scaling hyperparameter $\eta$ represents the strength of the subjective prior, where $\eta \to 0$ corresponds to the non-informative prior that yields rational decisions. The bandwidth hyperparameter $\xi$ determines a universal resolution in perceiving the similarity among multiple options.

Because every input vector $\boldsymbol{x}_{ij}$ contains a decision maker's characteristics vector $\boldsymbol{q}_i$, the similarity criteria (8) is heterogeneous among the decision makers.

# 4 EXPERIMENTAL EVALUATIONS

We compared the proposed model with existing discrete choice models, by validating their predictability for real-world datasets. For reproducibility in the future, we aimed to use public datasets while most of

Table 1: Two datasets exhibiting the compromise effect. Each subject in the PC and SP datasets is required to choose one option from a given choice set. Every subject is randomly assigned to one of the prepared choice sets. The highest shares of moderate options evidence the compromise effect.

Attributes of each portable PC (PC)

| | A | B | C | D | E |
|---|---|---|---|---|---|
| CPU [MHz] | 250 | 300 | 350 | 400 | 450 |
| Mem. [MB] | 192 | 160 | 128 | 96 | 64 |

Result of context-dependent choice

| Choice Set | #subjects |
|---|---|
| {A, B, C} | 36:176:144 |
| {B, C, D} | 56:177:115 |
| {C, D, E} | 94:181:109 |

Attributes of each speaker (SP)

| | A | B | C | D | E |
|---|---|---|---|---|---|
| Power [Watt] | 50 | 75 | 100 | 125 | 150 |
| Price [USD] | 100 | 130 | 160 | 190 | 220 |

Result of context-dependent choice

| Choice Set | #subjects |
|---|---|
| {A, B, C} | 45:135:145 |
| {B, C, D} | 58:137:111 |
| {C, D, E} | 95:155: 91 |

the prior work adopted proprietary and undisclosed datasets (e.g., (Keane and Wasi, 2012)). Given the limitation of public data, we simulated the real shares described in a marketing paper (Kivetz et al., 2004). Section 4.1 addresses the experimental settings including the properties of the datasets and reference models. The prediction results are summarized in Section 4.2.

### 4.1 Datasets and Settings

The tables attached in (Kivetz et al., 2004), which specify the number of decision-making subjects and shares of options, allowed us for reproducing two datasets PC ($n = 1,088$) and SP ($n = 972$) about the choice of a personal computer and a speaker, respectively. Every option in these datasets has ($d_R = 2$)-dimensional attributes, while absence of every subject's characteristics, i.e., $d_Q = 0$, imposed us to assume homogeneity among the subjects. Here the compromise effect matters as shown in Table 1.

We also tested a larger ($n = 10,719$) SwissMetro (SM) dataset (Antonini et al., 2007). In this dataset, every subject, whose characteristics and route in travel are represented by ($d_Q = 23$)-dimensional vector $\boldsymbol{q}_i$, is asked of choosing one transportation method either from a choice set {train, car, SwissMetro} or another choice set {train, SwissMetro} where numerical attributes of each transportation vary among the contexts. Each vector of the option attributes $\boldsymbol{r}_i$ is ($d_R = 7$)-dimensional and reflects cost, travel time, headway, seat type, and the three dummy variables to specify the type of transportation.

Each of the three datasets was randomly split into 20 folds of 80%-training and 20%-test sets, and we evaluated the average test-set log-likelihood with 20-fold cross-validation. By picking up the highest-probability option, we also evaluated the average test-set classification accuracy for easier understanding of the predictability.

We denote the proposed model by GPUA whose naming is the acronym of "Gaussian Process Uncertainty Aversion". As the first class of reference models, we fit linear and nonparametric logit models denoted as LinLogit and NpLogit, respectively. The LinLogit model assumes $\boldsymbol{\phi}(\boldsymbol{x}) \equiv \boldsymbol{x}$, while the NpLogit model adopts the same random Fourier features as GPUA.

As the second class of reference models, we prepared Mixed Multinomial Logit Models (MMLMs) providing

$$P(j|\mathcal{M}_i, \boldsymbol{X}_i, \Theta) = \sum_{t=1}^{T} \lambda_t \frac{\exp(\boldsymbol{w}_{\phi,t}^\top \boldsymbol{\phi}(\boldsymbol{x}_{ij}))}{\sum_{j'\in\mathcal{M}_i} \exp(\boldsymbol{w}_{\phi,t}^\top \boldsymbol{\phi}(\boldsymbol{x}_{ij'}))},$$

where $\boldsymbol{\lambda} \triangleq (\lambda_1, \ldots, \lambda_T)^\top$ is a vector of mixture weights such that $\sum_{t=1}^{T} \lambda_t \equiv 1$, $\boldsymbol{w}_{\phi,t}$ is a vector of coefficients assigned for the $t$th component, and $\Theta \triangleq \{\boldsymbol{\phi}, \boldsymbol{\lambda}, (\boldsymbol{w}_{\phi,t})_{t=1}^{T}\}$. Choice probabilities by any discrete choice model can be approximated arbitrarily well by an MMLM (McFadden and Train, 2000). Hence we regard MMLMs as good representatives of the state-of-the-art random utility models that can be fitted to real data, but cannot predict some irrational decisions (Rieskamp et al., 2006) and lack psychological interpretations (Stern and Richardson, 2005). Depending on the choice of the mapping function $\boldsymbol{\phi}$, we prepared both linear and nonparametric MMLMs denoted as LinMix and NpMix, respectively. In the fitting, we applied a variational Bayes method (e.g., Blei and Jordan (2006)) to obtain the posterior mean of the mixture vector $\boldsymbol{\lambda}$, and applied the MAP estimation for the set of vectors $\{\boldsymbol{w}_{\phi,t}\}_{t=1}^{T}$. We placed a symmetric Dirichlet distribution prior $\boldsymbol{\lambda} \sim Dir(1/T, \ldots, 1/T)$ and an isotropic Gaussian prior $p(\boldsymbol{w}_{\phi,t}) \propto \exp(-\frac{c}{2}\|\boldsymbol{w}_{\phi,t}\|^2)$.

The hyperparameters in every model were chosen with 3-fold likelihood cross-validation that further divides the 80% training data. The $L_2$ hyperparameter $c$ was chosen from $\{10^{-2}, 10^{-1}, 1, 10, 10^2\}$ in all of the three models. For parsimonious computations, two bandwidth hyperparameters are constrained to be the same,

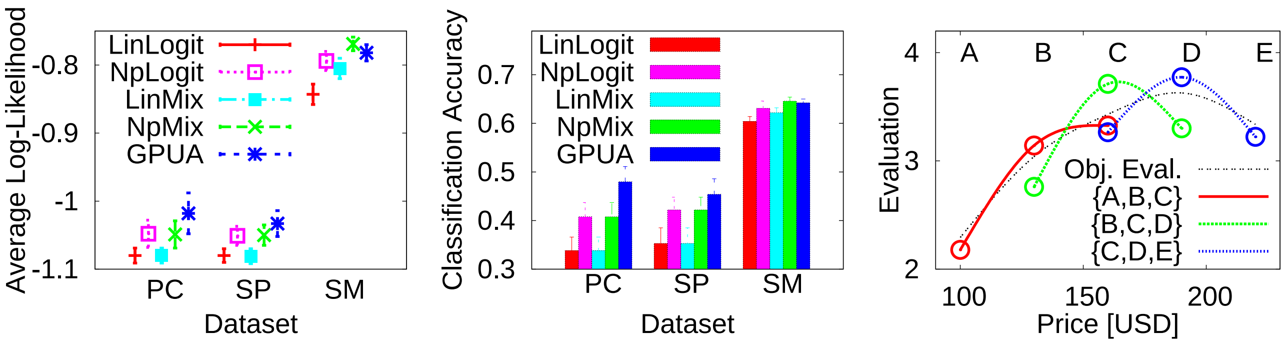


Figure 4: Prediction performances measured as average log-likelihood (left) and classification accuracy (center), and an estimated final evaluation function $u_i(\boldsymbol{x}^*)$ compared with the objective evaluation function $b+\boldsymbol{w}_\phi^\top \boldsymbol{\phi}(\boldsymbol{x}^*)$ (right). Each error bar represents one standard deviation among the 20 folds. We took an example of applying the `GPUA` model for the `SP` dataset in drawing the right-most figure. The highest evaluation for option C in the choice set $\{\mathrm{A}, \mathrm{B}, \mathrm{C}\}$ and those for the moderate options in other two choice sets are consistent with the shares of speakers in Table 1.

i.e., $\gamma \equiv \xi$. The bandwidth hyperparameter $\gamma (\equiv \xi)$ in each of the `NpLogit`, `NpMix`, and `GPUA` models was chosen from $\{10^{-1}, 10^{-1/2}, 1, 10^{1/2}, 10\}$ for the `PC` and `SP` datasets, and from $\{10^{-4}, 10^{-3}, \ldots, 1\}$ for the `SM` dataset. The strength of the subjective prior $\eta$ in `GPUA` was chosen from $\{10^{-3}, 10^{-2}, 10^{-1}, 1\}$. We tried to choose the mixture number $T$ in each MMLM from $\{16, 64\}$. Yet models with $T = 64$ performed almost equally to those with $T = 16$, and hence we provide only the $T\!=\!16$ cases for making the results simpler.

### 4.2 Prediction Results

Figure 4 provides the prediction performances and illustrates how choice sets affect the posterior-mean evaluations. For all of the three datasets, the proposed `GPUA` model outperformed the logit models `LinLogit` and `NpLogit`, in terms of both average log-likelihood and classification accuracy. For the large `SM` dataset, the most outperforming was the nonparametric MMLM `NpMix` while the proposed model is ranked as the second best. Yet the poor performances of the MMLMs against the `PC` and `SP` datasets, for which predicting the compromise effect is essential, imply the limited applicability of MMLMs. In contrast, we expect higher generalization capability of the proposed model in many contexts causing irrationality.

The advantage of having a good psychological interpretation is further clarified when we compare the rightmost graph in Figure 4 with Table 1. Let us remember that preferences in the `SP` dataset are more complex than those in the the `PC` dataset, because the choice criteria is a combination of the compromise effect and prioritization of speaker power over cheapness. Unlike the `PC` dataset that allows for a heuristic to always pick up the moderate option as the highest-share one, the `SP` dataset is a real example that needs a finely-tuned quantitative model that evaluates the total attractiveness involving the complex combination. The proposed model accurately predicted which of the conflicting objective and subjective criteria matters, for every of the three choice sets.

## 5 Conclusion

We proposed a new discrete choice model to predict choice decisions involving preference reversals, with providing a Bayesian mechanism of mental conflict between objective evaluations and a subjective prior. Each decision maker is assumed to perform a GPR whose Bayesian shrinkage leads irrational decisions as a consequence of aversion toward uncertainty. The parameters of the proposed model are fitted with a convex optimization algorithm and high predictability of the proposed model is validated by using real-world datasets that exhibit the context effects.

In the future work, we will apply our methodology for larger-size and higher-dimensional choice tasks. Design or fitting of more detailed covariance functions in the subjective prior will yield interesting economic implications about what types of option attributes strongly lead irrational decisions. Such covariance structure would also clarify how humans perceive the values of options having complex attributes, such as shaping in product designs and tastes of foods. Another considerable application is improvement of information retrieval systems for economic decision making, such as travel planning and accommodation search.

### Acknowledgments

This research was supported by CREST, JST.

## References

G. Antonini, C. Gioia, E. Frejinger, and M. Thémans. Swissmetro: description of the data, 2007. http://biogeme.epfl.ch/swissmetro/examples.html.

L. Barrett, M. Tugade, and R. Engle. Individual differences in working memory capacity and dual-process theories of the mind. *Psychological Bulletin*, 130: 553–573, 2004.

N. A. J. Berkowitsch, B. Scheibehenne, and J. Rieskamp. Rigorously testing multialternative decision field theory against random utility models. *Journal of Experimental Psychology: General*, 143 (3):1331–1348, 2014.

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

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.

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.

W. Chu and Z. Ghahramani. Preference learning with Gaussian processes. In *Proceedings of the 22th International Conference on Machine Learning (ICML 2005)*, pages 137–144, 2005.

J. P. Dotson, P. Lenk, J. Brazell, T. Otter, S. N. Maceachern, and G. M. Allenby. A probit model with structured covariance for similarity effects and source of volume calculations, 2009. http://ssrn.com/abstract=1396232.

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

G. Glasgow. Mixed logit models for multiparty elections. *Political Analysis*, 9(2):116–136, 2001.

C. González-Vallejo. Making trade-offs: A probabilistic and context-sensitive model of choice behavior. *Psychological Review*, 109:137–154, 2002.

J. Hausman and D. Wise. A conditional probit model for qualitative choice: Discrete decisions recognizing interdependence and heterogenous preferences. *Econometrica*, 48(2):403–426, 1978.

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.

G. Karabatsos and S. G. Walkerbbook. Bayesian nonparametric mixed random utility models. *Computational Statistics and Data Analysis*, 56(6):1714–1722, 2012.

M. P. Keane and N. Wasi. Estimation of discrete choice models with many alternatives using random subsets of the full choice set: With an application to demand for frozen pizza. 2012.

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

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

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

P. Manzini and M. Mariotti. Consumer choice and revealed bounded rationality. *Economic Theory*, 41 (3):379–392, 2009.

D. McFadden and K. Train. Mixed MNL models for discrete response. *Journal of Applied Econometrics*, 15:447–470, 2000.

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

P. Natenzon. Random choice and learning. Working paper, 2010.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 20*, pages 1177–1184. Curran Associates, Inc., 2008.

J. Rieskamp, J. R. Busemeyer, and B. A. Mellers. Extending the bounds of rationality: Evidence and theories of preferential choice. *Journal of Economic Literature*, 44(3):631–661, 2006.

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

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.

P. Shenoy and A. J. Yu. A rational account of contextual effects in preference choice: What makes for a bargain? In *Proceedings of the Cognitive Science Society Conference*, 2013.

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

E. Stern and H. W. Richardson. Behavioural modeling of road users: Current research and future needs. *Transport Reviews*, 25(2):159–180, 2005.

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.

C.-H. Wen and F. Koppelman. The generalized nested logit model. *Transportation Research Part B*, 35: 627–641, 2001.

H. Williams. On the formulation of travel demand models and economic evaluation measures of user benefit. *Environment and Planning A*, 9(3):285–344, 1977.

T. Yai. Multinomial probit with structured covariance for route choice behavior. *Transportation Research Part B: Methodological*, 31(3):195–207, 1997.