# Gaussian Process Conjoint Analysis for Adaptive Marginal Effect Estimation

**Yehu Chen, Jacob Montgomery, Roman Garnett**
Washington University in St Louis
chenyehu,jacob.montgomery,garnett@wustl.edu

## Abstract

Choice-based conjoint analysis is an essential tool for learning the marginal effects of multidimensional explanatory features on preferences. However, existing marginal effect models rely on either non-parametric estimators that generalize poorly to individualized effects, or linear latent utility that completely ignores possible high-order interactions. We introduce Gaussian process conjoint analysis (GPCA) for learning marginal effects from observed choices as the first-order derivatives of the unknown systems. We also propose Gaussian mixture approximation for the predictive distributions of marginal effects that facilitates downstream tasks such as adaptive experimentation. Through both synthetic and real data, we show GPCA achieves more precise estimation of marginal effects and higher efficiency of effect estimation using adaptive experimentation.

## 1 Introduction

Understanding the relationship between targeted outcomes and features in survey experiments is fundamental in many disciplines such as social science [1–3], human-computer interaction [4, 5] and marketing research [6–8]. These associations are often captured by marginal effect, defined as the change in predicted outcomes resulting from changes in features. Depending on the type of attributes, marginal effects could either be computed as the discrete change in outcomes for categorical attributes or infinitesimal margins for continuous attributes. In survey experiments, marginal effects are often learned using the choice-based conjoint experiments which present a series of profile pairs at varying attribute values so as to compare the difference in averaged outcomes [6]. For example, researchers alternate background characteristics to study bias towards immigrants, and system designers change interface setups to improve click-through rates of their new web interface.

However, learning marginal effects from conjoint analysis encounters several challenges. First, effects of a single attribute may be heterogeneous when interacting with other attributes. To learn possible heterogeneous effects caused by high-order interactions, existing methods usually rely on stacking multiple attributes in a difference-in-different style that makes estimation of interaction effects involving more attributes extremely complicated [1, 2]. Alternatively, marginal effects may also be captured by the first-order derivatives of attributes w.r.t to the preference outcomes through a latent utility function. However, previous work typically depend on linear models such as support vector machine to learn partial utilities that overlooks possible interactions in the feature space [8–11].

Second, the multi-dimensional nature in conjoint experiments may lead to small-sample biases in effect estimation, as common randomization design would inevitably split sample sizes on each level of attributes. Hence, adaptive experimentation may be needed for acquiring next pairs of profiles when querying of unknown preferences is expensive. By utilizing prior responses and maintaining a belief model of the system, adaptive experimentation could balance between exploiting attributes that are more crucial to the preference and exploring attributes that the model is uncertain about.

Preprint. Under review.

In this work, we study the problem of marginal effect estimation in choice-based conjoint analysis and propose Gaussian process conjoint analysis (GPCA) that automatically learns high-order interactions by using the preference learning framework. We derive marginal effects using first-order derivatives of Gaussian process learned from observed preferences, and approximate the distributions of marginal effects via Gaussian mixture models. By building a predictive model of the latent system, GPCA could also facilitate adaptive experimentation such as Bayesian active learning by disagreement to accelerate effect estimation. As shown in the simulated experiments, GPCA is able to achieve more precise estimation of marginal effects than other non-parametric and parametric methods. Finally, we apply GPCA to two real-world online experiments: learning citizens' preferences across presidential candidates and examining attitudes toward immigrants.

## 2 Related work

**Conjoint analysis.** Originally introduced as a marketing tool [6, 7], conjoint analysis has been used for learning multi-dimensional treatment effects using non-parametric estimators in quantitative research [1, 2] or eliciting user preferences in recommendation system via parametric utility functions [10, 12]. Hainmueller et al. [1] proposed a difference-in-difference interaction effect estimator for eliciting preferences from multi-dimensional choices in survey experiments, where the inner and outer differences come from the target and interacted attributes. Subsequently, Egami and Imai [2] proposed a new effect estimator in factorial experiments that does not depend on the choice of baseline conditions and generalizes better for higher-order interaction effects. However, these work focus on discrete attributes and have to categorize continuous attributes into distinct subgroups that are subject to categorization. Alternatively, Chapelle and Harchaoui [10] introduced a generalized logistic approach by learning a parametric latent utility and explaining observed preferences via a softmax function. Similar utility-based methods include support vector machines [9, 8, 11], Gaussian processes [13, 12, 14–16] and decision trees [17]. However, these preference learning methods emphasize learning the most preferred recommendations through latent utilities of low interpretability, rather than estimation of marginal effects that explains the relation between attributes and outcomes.

**Marginal effects.** Marginal effects was studied in economics for measuring the responsiveness of economic variables by the concept of elasticity [18], for instance, how the percentage of demand quantity falls due to percentage of change in price. Hence, marginal effects are often used for understanding transformed features in regression models [19] or examining heterogeneous association between feature and outcomes [20]. Another stream of work focus on using marginal effects for machine learning model interpretability. Silva Filho et al. [21] provided a feature importance method for interpreting classification models based on marginal local effects. Merz et al. [22] proposed a marginal attribution method by conditioning on quantiles for analyzing global gradients in deep neural network. Scholbeck et al. [17] introduced forward marginal effects that unify and mixed-type features as a general model-agnostic interpretation method for general non-linear machine learning models. However, marginal effects in preference learning has not been investigated in these literature.

**Adaptive experiment.** Often framed as a sequential decision making or active learning problem [23], adaptive experimentation utilizes already collected responses for informing experiment setup or data acquisition in next iterations to maximize the usefulness of limited data. Adaptive experiment has been adopted by domain scientists to accelerate scientific discovery. For instance, Bayesian optimization via adaptive sample selection were successfully applied in material science for discovering new materials [24] and clinical trials for finding maximum tolerated dose [25, 26]. Meanwhile, active search was introduced for iterative design of virtual screening trials in chemoinformatics [27]. In machine learning, Chen et al. [28] studied the pairwise ranking problem in crowd-sourcing setup with online learning. Bıyık et al. [29] proposed an active preference-based learning based on information gain for reward functions in robotics. However, previous adaptive designs in quantitative research have been mainly focused on treatment selection in bandit settings [30–32], with limited attention to marginal effect estimation particularly within the GP preference learning framework.

## 3 Backgrounds

**Notations.** Formally, let $\mathbf{x} \subseteq \mathbb{R}^d$ denote all $d$-dimensional attributes of the full profile, and $\mathbf{x}_l$ represents the $l$th attribute and $\mathbf{x}_{-l}$ represents the remaining attributes other than the $l$th. Furthermore, for pairwise comparison, let $y_{ij} \in \{0,1\}$ denote whether the left-side profile $\mathbf{x}^{(i)}$ is preferred to the right-side $\mathbf{x}^{(j)}$, where $y_{ij} = 1$ if $\mathbf{x}^{(i)} \succ \mathbf{x}^{(j)}$ and $y_{ij} = 0$ otherwise. Here we focus on choice-based conjoint analysis with pairwise comparison, as multiple choices could be easily transformed into multiple comparison of pairs. For instance, $\mathbf{x}^{(i)}$ is mostly preferred amongst $\{\mathbf{x}^{(i)}, \mathbf{x}^{(j)}, \mathbf{x}^{(k)}\}$ is equivalent to $\mathbf{x}^{(i)} \succ \mathbf{x}^{(j)}$ and $\mathbf{x}^{(i)} \succ \mathbf{x}^{(k)}$. Our notation could also account for score-based conjoint experiments, where $\mathbf{x}^{(i)} \succ \mathbf{x}^{(j)}$ could indicate $\mathbf{x}^{(i)}$ having higher score than $\mathbf{x}^{(j)}$. Furthermore, suppose all revealed preferences are collected into $\mathcal{D} = \{(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}), y_{ij}\}$.

**Marginal effects of discrete attributes.** In conjoint analysis with factorial design, attributes usually take discrete values of different levels $\mathbf{x}_l = 1, \ldots, C_l$. For a target distribution of profiles $\mathcal{P}$, the marginal effects $\pi_l(a,b)$ of attribute $\mathbf{x}_l$ from level $a$ to $b$ ($1 \le a < b \le C_l$) are captured by the average marginal component effect (AMCE), defined as the difference in expected preferential outcomes averaged over all the possible values of the remaining attributes $\mathbf{x}_{-l}$ over $\mathcal{P}$:

$$\pi_l(a,b) = \mathbb{E}_{\mathbf{x}^{(i)}_{-l}, \mathbf{x}^{(j)} \sim \mathcal{P}}[y_{ij} \mid \mathbf{x}^{(i)}_l = b] - \mathbb{E}_{\mathbf{x}^{(i)}_{-l}, \mathbf{x}^{(j)} \sim \mathcal{P}}[y_{ij} \mid \mathbf{x}^{(i)}_l = a] \tag{1}$$

Intuitively, $\pi_l(a,b)$ represents the increase in the probability of one profile being preferred if the $l$th attribute were $b$ instead of $a$ for profile distribution $\mathcal{P}$. With the conditionally independent assumption, $\pi_l(a,b)$ can be estimated straight-forwardly using a difference-in-mean approach:

$$\hat{\pi}_l(a,b) = \frac{\sum_{(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) \in \mathcal{D}} y_{ij} \mathbb{I}[\mathbf{x}^{(i)}_l = b]}{\sum_{(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) \in \mathcal{D}} \mathbb{I}[\mathbf{x}^{(i)}_l = b]} - \frac{\sum_{(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) \in \mathcal{D}} y_{ij} \mathbb{I}[\mathbf{x}^{(i)}_l = a]}{\sum_{(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) \in \mathcal{D}} \mathbb{I}[\mathbf{x}^{(i)}_l = a]} \tag{2}$$

However, this difference-in-mean approach for estimating marginal effects suffers from two issues. First, generalization of this estimator for heterogeneous effect resulting from either background characteristics or high-level interactions could get more complicated as calculation of multiple differences is required. For instance, for obtaining interaction effects of between $\mathbf{x}^{(i)}_l$ and $\mathbf{x}^{(i)}_m$ from level $c$ to $d$ in $\mathbf{x}^{(i)}_m$, one needs to compute $[\hat{\pi}_l(a,b)\big|_{\mathbf{x}^{(i)}_m = c} - \hat{\pi}_l(a,b)\big|_{\mathbf{x}^{(i)}_m = c}] - [\hat{\pi}_l(a,b)\big|_{\mathbf{x}^{(i)}_m = d} - \hat{\pi}_l(a,b)\big|_{\mathbf{x}^{(i)}_m = d}]$ [1]. Second, in practice, continuous attributes are rarely repeated and thus often need to be discretized into multiple levels; otherwise, each level $\mathbf{x}^{(i)}_l = a$ would have very few observations. However, this discretization is subject to the chosen cutoff points and may lead to an oversimplification of the system, threatening the internal validity of marginal effect estimation.

# 4 Gaussian process conjoint analysis

We now introduce Gaussian process conjoint analysis (GPCA) for estimating marginal effects in conjoint analysis of mixed-type attributes. We then derive marginal effects in GPCA and propose the use of Gaussian mixture model for effectively approximating their distributions.

## 4.1 Preference learning with Gaussian process

Conjoint analysis can also be framed as a preference learning problem with a latent utility function $u(\mathbf{x})$ that takes mixed-type attributes. The preferential relation between $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(j)}$ is then determined by comparing their utilities $u(\mathbf{x}^{(i)})$ and $u(\mathbf{x}^{(j)})$. Through a sigmoid probabilistic model $\sigma(\cdot)$, the probability of observed preference $p(\mathbf{x}^{(i)} \succ \mathbf{x}^{(j)}) = \sigma\big(u(\mathbf{x}^{(i)}) - u(\mathbf{x}^{(j)})\big)$ could also allow possible labeling error. Gaussian process (GP) preference learning places a GP prior on latent utility $u(\mathbf{x}) \sim \mathcal{GP}(0, K)$ with RBF kernel $K(x,x') = \exp(-\|x - x'\|^2/2)$, and uses a cumulative standard normal function for observation model $p\big(\mathbf{x}^{(i)} \succ \mathbf{x}^{(j)} \mid u(\mathbf{x}^{(i)}), u(\mathbf{x}^{(j)})\big) = \Phi\big(u(\mathbf{x}^{(i)}) - u(\mathbf{x}^{(j)})\big)$. Although the posterior of $u(\mathbf{x})$ is no longer analytical for GP classification, it could be approximated using standard methods such as Laplace approximation and expectation propagation [33, 12].

Furthermore, the inferred latent utility posterior could also be used for prediction. For any new pair of profiles $(\mathbf{x}^{(i)}_*, \mathbf{x}^{(j)}_*)$, suppose their corresponding utility vector has been approximated by a bivariate normal $\mathbf{u}_* = [u(\mathbf{x}^{(i)}_*), u(\mathbf{x}^{(j)}_*)]^T \sim \mathcal{N}(\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*)$. Let $\boldsymbol{\mu}_* = [\mu^{(i)}_*, \mu^{(j)}_*]^T$ and $\sigma^2_* =$

$1 + [1, -1]\mathbf{\Sigma}_*[1, -1]^T$, then the predictive probability has the following closed-form:

$$p(\mathbf{x}_*^{(i)} \succ \mathbf{x}_*^{(j)}) = \int \Phi\big(u(\mathbf{x}_*^{(i)}) - u(\mathbf{x}_*^{(j)})\big) p(\mathbf{u} \mid \mathcal{D}) d\mathbf{u} = \Phi(\frac{\mu_*^{(i)} - \mu_*^{(j)}}{\sigma_*}) \tag{3}$$

Sometimes the predictive probability in Eq. (3) are directly defined on pairs of profiles $(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$ using preference kernel. As the difference of two Gaussians remains Gaussian, a GP on $u(\mathbf{x}^{(i)})$ will also induce a GP on $u(\mathbf{x}^{(i)}) - u(\mathbf{x}^{(j)})$ but with a preference kernel $K_{\text{pref}}\big((\mathbf{x}_1^{(i)}, \mathbf{x}_1^{(j)}), (\mathbf{x}_2^{(i)}, \mathbf{x}_2^{(j)})\big) = K(\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)}) - K(\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(j)}) - K(\mathbf{x}_2^{(i)}, \mathbf{x}_1^{(j)}) + K(\mathbf{x}_1^{(j)}, \mathbf{x}_2^{(j)})$. We adopted this preference kernel in our implementation of GPCA.

### 4.2 Marginal effects in GPCA

We follow the definition of AMCE in Eq. (1) but adapted to our GPCA framework. We exploit the affine property of Gaussian processes to derive marginal effects of mixed-type attributes using first-order gradients, where discrete attributes can be converted to continuous attributes with additional dummy variables. Our discussion will focus on marginal effects of profile pairs $(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$ on both sides. Specifically, the gradient $\pi\big((\mathbf{x}^{(i)}, \mathbf{x}^{(j)})\big)$ in probability of target profile $\mathbf{x}^{(i)}$ being preferred to opponent profile $\mathbf{x}^{(j)}$ can be derived as:

$$\pi\big((\mathbf{x}^{(i)}, \mathbf{x}^{(j)})\big) = \frac{\partial}{\partial(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})}[p(\mathbf{x}^{(i)} \succ \mathbf{x}^{(j)})] \qquad \text{definition of AMCE} \tag{4}$$

$$= \frac{\partial}{\partial(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})}\mathbb{E}_{u|\mathcal{D}}\Big[\Phi\big(u(\mathbf{x}^{(i)}) - u(\mathbf{x}^{(j)})\big)\Big] \qquad \text{averaged by } u \mid \mathcal{D} \tag{5}$$

$$= \mathbb{E}_{u|\mathcal{D}}\Big[\phi\big(u(\mathbf{x}^{(i)}) - u(\mathbf{x}^{(j)})\big)\big(\nabla u(\mathbf{x}^{(i)}), -\nabla u(\mathbf{x}^{(j)})\big)\Big] \qquad \text{chain rule} \tag{6}$$

Note that in the second step we swapped the order of expectation and differentiation. Intuitively, the marginal effects of $(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$ on the outcome space can be computed as the expected gradient $\big(\nabla u(\mathbf{x}^{(i)}), -\nabla u(\mathbf{x}^{(j)})\big)$ in the latent utility space further weighted by the probability densities $\phi\big(u(\mathbf{x}^{(i)}) - u(\mathbf{x}^{(j)})\big)$ of a normal distribution at the latent utility distance $u(\mathbf{x}^{(i)}) - u(\mathbf{x}^{(j)})$. For the sake of notation, further denote the one-sided marginal effect $\phi\big(u(\mathbf{x}) - u(\mathbf{x}^{(j)})\big)\nabla u(\mathbf{x})$ as $\mathbf{g}(\mathbf{x}; \mathbf{x}^{(j)}, \mathcal{D})$ where $\mathcal{D}$ indicates the posterior of utility on $\mathcal{D}$. Since the normal pdf is symmetric, we could conveniently write the right-side gradient as $-\phi\big(u(\mathbf{x}^{(i)}) - u(\mathbf{x})\big)\nabla u(\mathbf{x}) = -\phi\big(u(\mathbf{x}) - u(\mathbf{x}^{(i)})\big)\nabla u(\mathbf{x})$ as $-g(\mathbf{x}; \mathbf{x}^{(i)}, \mathcal{D})$ and hence marginal effect as $\pi\big((\mathbf{x}^{(i)}, \mathbf{x}^{(j)})\big) = \big(\mathbf{g}(\mathbf{x}^{(i)}; \mathbf{x}^{(j)}, \mathcal{D}), -\mathbf{g}(\mathbf{x}^{(j)}; \mathbf{x}^{(i)}, \mathcal{D})\big)$. Lastly, $\pi\big((\mathbf{x}^{(i)}, \mathbf{x}^{(j)})\big)$ captures the concatenated multi-variate distribution of marginal effects for the entire profile vectors, and could be easily projected along any unit vector $\hat{\mathbf{e}}_l$ to obtain the component effects analogous to Eq. (1). Intuitively, component effects represent the attribute-specific effects on preferences, averaged over profile population:

$$\pi_l(\mathbf{x}_l^{(i)}) = \sum_{(\mathbf{x}_{-l}^{(i)}, \mathbf{x}^{(j)}) \sim \mathcal{P}} \langle \pi\big((\mathbf{x}^{(i)}, \mathbf{x}^{(j)})\big), \hat{\mathbf{e}}_l \rangle \tag{7}$$

### 4.3 Gaussian mixture approximation of marginal effects

As $\pi\big((\mathbf{x}^{(i)}, \mathbf{x}^{(j)})\big)$ involves taking weighted averages of utility gradient vector $\nabla u(\mathbf{x})$ over utility posterior $u \mid \mathcal{D}$, we propose the use of Gaussian mixture model (GMM) to approximate its distribution. As the gradient of a GP is still a GP, we can first write the joint distribution of utility $u(\cdot) \mid \mathcal{D}$ and utility gradient $\nabla u \mid \mathcal{D}$ under utility posterior $\mathcal{GP}\big(\mu_{u|\mathcal{D}}(\mathbf{x}), K_{u|\mathcal{D}}(\mathbf{x}, \mathbf{x}')\big)$ on $\mathcal{D}$ as:

$$\begin{bmatrix} u \mid \mathcal{D} \\ \nabla u \mid \mathcal{D} \end{bmatrix} \sim \mathcal{GP}\Big( \begin{bmatrix} \mu_{u|\mathcal{D}} \\ \nabla \mu_{u|\mathcal{D}} \end{bmatrix}, \begin{bmatrix} K_{u|\mathcal{D}} & \nabla K_{u|\mathcal{D}}{}^T \\ \nabla K_{u|\mathcal{D}} & \nabla^2 K_{u|\mathcal{D}} \end{bmatrix} \Big) \tag{8}$$

where $\nabla \mu_{u|\mathcal{D}} = \partial \mu_{u|\mathcal{D}}(\mathbf{x})/\partial \mathbf{x}$ is the first-order derivative of the posterior mean, $\nabla K_{u|\mathcal{D}} = \partial K_{u|\mathcal{D}}(\mathbf{x}, \mathbf{x}')/\partial \mathbf{x}$ is the first-order derivative of the posterior covariance and $\nabla^2 K_{u|\mathcal{D}} = \partial^2 K_{u|\mathcal{D}}(\mathbf{x}, \mathbf{x}')/\partial \mathbf{x} \partial \mathbf{x}'$ is its second-order mixed derivatives.

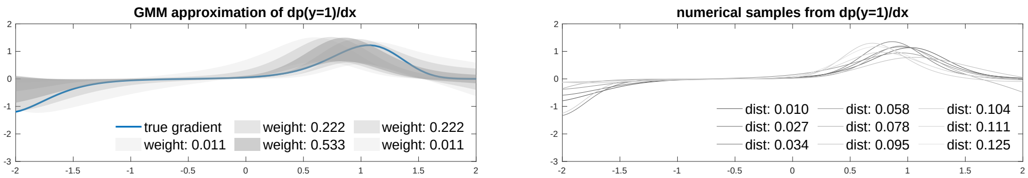



Figure 1: Visualization of the proposed GMM for approximating one-side marginal effect. Left figure shows our GMM approximation of the one-side marginal effect using 5 sampling points, and right figure shows 9 possible true effects obtained by numerical sampling. Darker colors indicate components with higher weights in the GMM and numerical samples closer to the one-side marginal effect posterior mode.

Although the joint distribution in Eq. (8) is Gaussian, the one-sided marginal effect $\mathbf{g}(\mathbf{x}; \mathbf{x}^{(j)}, \mathcal{D})$ is not because it involves the product of a multivariate Gaussian $\nabla u(\mathbf{x}) \mid \mathcal{D}$ and a non-linear transformation $\phi(\cdot)$ of an univariate Gaussian $u(\mathbf{x}) - u(\mathbf{x}^{(j)}) \mid \mathcal{D}$. Therefore, we use a Gaussian mixture model (GMM) to approximate $\mathbf{g}(\mathbf{x}; \mathbf{x}^{(j)}, \mathcal{D})$. Each component of the GMM is formed by scaling the multivariate Gaussian with the transformed values of quadrature points of the univariate Gaussian determined by Gauss-Hermite quadrature. Let $N$ be the number of points in the quadrature, $k_r$ be the roots of the physicists' version of the Hermite polynomial $H_N(k)$ and $\omega_r = \frac{2^{N-1}N!}{N^2[H_{N-1}(k_r)]^2}$ be the weights of each component [34]. We could then approximate $\mathbf{g}(\mathbf{x}; \mathbf{x}^{(j)}, \mathcal{D})$ as:

$$\mathbf{g}(\mathbf{x}; \mathbf{x}^{(j)}, \mathcal{D}) \approx \sum_{r=1}^{N} \omega_r \phi\big(\bar{f}_r(\mathbf{x})\big) \circ \mathcal{N}\Big(\nabla \mu_{u|\mathcal{D}}(\mathbf{x}), \nabla^2 K_{u|\mathcal{D}}(\mathbf{x}, \mathbf{x})\Big) \tag{9}$$

$$= \sum_{r=1}^{N} \omega_r \mathcal{N}\Big(\phi\big(\bar{f}_r(\mathbf{x})\big) \circ \nabla \mu_{u|\mathcal{D}}(\mathbf{x}), \phi\big(\bar{f}_r(\mathbf{x})\big)\phi\big(\bar{f}_r(\mathbf{x})\big)^T \circ \nabla^2 K_{u|\mathcal{D}}(\mathbf{x}, \mathbf{x})\Big) \tag{10}$$

where $\bar{f}_r(\mathbf{x}) = \sqrt{2}[\sigma^2_{u|\mathcal{D}}(\mathbf{x}) + \sigma^2_{u|\mathcal{D}}(\mathbf{x}^{(j)})]^{1/2} k_r + [\mu_{u|\mathcal{D}}(\mathbf{x}) - \mu_{u|\mathcal{D}}(\mathbf{x}^{(j)})]$ are locations of mixture components defined on the sample point $k_r$s, and $\circ$ denotes the Hadamard (element-wise) product. Figure 1 shows the visualization of the proposed GMM for approximating one-side marginal effect. The left-hand side shows our GMM approximation of the one-sided marginal effect using 5 sampling points, and the right-hand side shows 9 possible true effects obtained by numerical sampling. Darker colors indicate components with higher weights in the GMM and numerical samples closer to the one-side marginal effect posterior mode. We found in experiments with just $N = 10$ quadrature points, our GMM was able to effectively approximate the true distribution of $\mathbf{g}(\mathbf{x}; \mathbf{x}^{(j)}, \mathcal{D})$.

# 5 Adaptive experimentation in GPCA

We investigate the use of adaptive experimentation with GPCA to acquire the most informative pairs of profiles for estimating marginal effects. Informed by the posterior belief on the latent utility, adaptive experimentation may efficiently explore attributes whose marginal effects on preferences are less certain. To this end, we can determine the next pairs of profiles to compare by maximizing an *acquisition function* $(\mathbf{x}^{(i)}_*, \mathbf{x}^{(j)}_*) = \max_{(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) \sim \mathbb{P}} \alpha\big((\mathbf{x}^{(i)}, \mathbf{x}^{(j)}); \mathcal{D}\big)$. For simplicity, let $A = u(\mathbf{x}^{(i)}) - u(\mathbf{x}^{(j)})$ and $B = K_{u|\mathcal{D}}(\mathbf{x}^{(i)}, \mathbf{x}^{(i)}) + K_{u|\mathcal{D}}(\mathbf{x}^{(j)}, \mathbf{x}^{(j)})$. We consider the following policies:

1. Upper confident bound on predictive preference (UCB) maximizes the 95% confidence interval of preference prediction: $\alpha\big((\mathbf{x}^{(i)}, \mathbf{x}^{(j)}); \mathcal{D}\big) = \big|A + 1.96\sqrt{B}\big|$.
2. Differential entropy of the latent utility (DE-U) maximizes the log variance of utility posterior: $\alpha\big((\mathbf{x}^{(i)}, \mathbf{x}^{(j)}); \mathcal{D}\big) = \frac{1}{2}\log(2\pi B) + \frac{1}{2}$.
3. Differential entropy of the marginal effects (DE-ME) maximizes the log variance of marginal effects approximated using our proposed GMM in Eq. (9):

$$\alpha\big((\mathbf{x}^{(i)}, \mathbf{x}^{(j)}); \mathcal{D}\big) = \log\Big|\sum_{k\in\{i,j\}}\sum_{r=1}^{N} \omega_r \phi\big(\bar{f}_r(\mathbf{x}^{(k)})\big)\phi\big(\bar{f}_r(\mathbf{x}^{(k)})\big)^T \circ \nabla^2 K_{u|\mathcal{D}}(\mathbf{x}^{(k)}, \mathbf{x}^{(k)})\Big|.$$

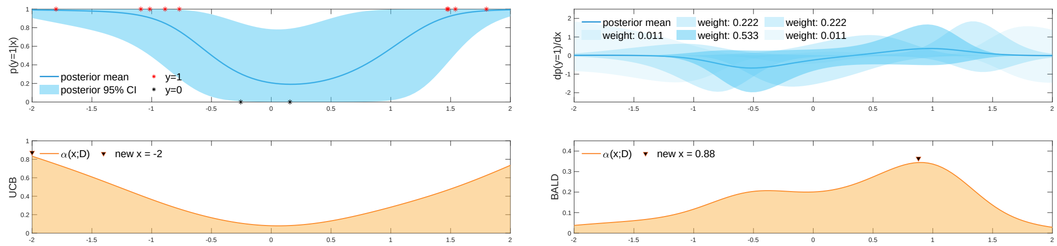


Figure 2: A 1-d example for illustrating acquisition functions of UCB and BALD. Upper panel shows the observed data with model posterior (left) and current marginal effect estimation (right). Lower panel shows acquisition value of UCB and BALD for selecting new profile, where the marginal effect variance at UCB's selection is low and that at BALD's selection is high. While UCB tends to exploit and optimize profile preference, BALD tends to explore and minimize model uncertainty.

4. Bayesian active learning by disagreement (BALD) aims to maximize the mutual information between the utility model and predictive preferences:

$$\alpha\big((\mathbf{x}^{(i)}, \mathbf{x}^{(j)}); \mathcal{D}\big) = \mathbf{I}(y_{ij}, u; \mathbf{x}^{(i)}, \mathbf{x}^{(j)}, \mathcal{D}).$$

With entropy function $h(p) = -p\log(p) - (1-p)\log(1-p)$ and constant $C = \sqrt{\pi \log(2)/2}$, the approximated mutual information is:

$$\alpha\big((\mathbf{x}^{(i)}, \mathbf{x}^{(j)}); \mathcal{D}\big) \approx h\Big(\Phi\big(\frac{A}{\sqrt{B+1}}\big)\Big) - \frac{C}{\sqrt{B+C^2}} \exp\big(-\frac{A^2}{2(B+C^2)}\big).$$

5. Random sampling (UNIFORM) simply selects pairs uniformly at random from $\mathcal{P}$.

While UCB emphasizes *exploiting* current belief to find the most preferred profile, DE-U, DE-ME and BALD focus on *exploring* the profile space by reducing model uncertainty on either latent utility, marginal effects or the predictive preferences. Figure 2 shows a 1-d example for illustrating acquisition functions of UCB and BALD. Upper panel shows the observed data with model posterior (left) and current marginal effect estimation (right). Lower panel shows acquisition value of UCB and BALD for selecting new profile, where the marginal effect variance at UCB's selection is low and that at BALD's selection is high. This demonstrates that while UCB tends to exploit and optimize profile preference, BALD tends to explore and minimize model uncertainty.

# 6 Experiments

We first evaluate the estimated marginal effects by GPCA using synthetic data when the functional relations are known and could be computed analytically, and then consider adaptive experimentation of GPCA with several active learning policies. We also apply GPCA to two real-world data.

**Data generating process**. Following the simulation specification in Chu and Ghahramani [12], we consider two generating processes with discrete (2DPLANE) and continuous (FRIEDMAN) attributes.[1] The 2DPLANE dataset has 5 discrete attributes where $x_1 \in \{-1, 1\}$ and $x_2, \ldots, x_5 \in \{-1, 0, 1\}$, with a piecewise linear utility $u(\mathbf{x}) = 1 + 2x_2 - x_3$ if $x_1 = -1$ and $u(\mathbf{x}) = 1 + x_4 - 2x_5$ if $x_1 = 1$. The FRIEDMAN dataset has 3 continuous attributes where $x_1, \ldots, x_3 \sim [0, 1]$ with a non-linear utility $u(\mathbf{x}) = 3\sin(\pi x_1 x_2) + 6(x_3 - 0.5)^2$. We randomly sample 1000 pairs of profiles in each dataset and set $y_{ij} = 1$ with probability of $\Phi\big(u(\mathbf{x}^{(i)} - u(\mathbf{x}^{(j)})\big)$ and $y_{ij} = 0$ otherwise.

## 6.1 Accuracy of marginal effect estimation

**Evaluation metrics and baselines.** We consider three metrics for evaluation of both marginal effects and component effects: (1) the RMSE of the estimated effects, (2) the correlation (COR) between the estimated effects and true effects, and (3) the log likelihood (LL) of the estimated effects. We also compare our proposed GMM approximation for marginal effects in GPCA to several baselines: (1)

[1] See `https://www.dcc.fc.up.pt/~ltorgo/Regression/DataSets.html` for details.

Table 1: Averaged performance and standard deviations of both marginal and component effects from our GP-GMM estimator and baselines on the 2DPLANE and FRIEDMAN datasets. Models that perform statistically significantly better than all the others in paired t-tests are indicated in bold, while methods performing comparably to best models are indicated in italics.

| DATASET | ESTIMATOR | Marginal effects | | | Component effects | | |
|---|---|---|---|---|---|---|---|
| | | RMSE ↓ | COR ↑ | LL ↑ | RMSE ↓ | COR ↑ | LL ↑ |
| 2DPLANE | DIM | 0.712±0.022 | 0.013±0.003 | −2.137±0.115 | 0.109±0.005 | 0.341±0.029 | 0.494±0.117 |
| | LM-GMM | 0.213±0.001 | 0.340±0.005 | −0.238±0.145 | 0.069±0.002 | 0.475±0.019 | −0.778±0.157 |
| | GP-MAP | 0.175±0.002 | 0.732±0.007 | −3.893±0.863 | 0.052±0.002 | *0.611±0.024* | 1.401±0.177 |
| | GP-GMM | **0.135±0.002** | **0.803±0.007** | **0.563±0.023** | **0.044±0.001** | **0.616±0.025** | **2.000±0.082** |
| FRIEDMAN | DIM | 0.910±0.008 | 0.024±0.005 | −9.658±0.392 | 0.150±0.010 | 0.944±0.017 | −1.824±0.480 |
| | LM-GMM | 0.845±0.010 | 0.328±0.007 | −1.001±0.271 | 0.078±0.005 | 0.980±0.005 | 0.503±0.245 |
| | GP-MAP | 0.510±0.008 | *0.830±0.006* | −3.869±0.530 | **0.042±0.003** | **0.995±0.001** | *1.680±0.045* |
| | GP-GMM | **0.478±0.008** | **0.847±0.005** | **−0.213±0.065** | **0.042±0.003** | **0.995±0.001** | **1.689±0.044** |

the non-parametric diff-in-mean estimator (DIM) [1], where the continuous attributes in FRIEDMAN are first discretized by splitting into equally-spanned intervals, (2) the standard preference learning method with linear utility (LM-GMM) [9, 10, 8, 11], and (3) an ablated GPCA method (GP-MAP) but with MAP estimation of marginal effects.

**Results.** We repeat our simulation with 25 different random seeds using 300 Intel Xeon 2680 CPUs. Table 1 shows the averaged performance and standard deviations (STDs) of both marginal effects $\pi\big((\mathbf{x}^{(i)}, \mathbf{x}^{(j)})\big)$ and component effects $\pi_l(\mathbf{x}_l^{(i)})$ defined in Eq. (4 and 7) from our GP-GMM estimator and baselines on the 2DPLANE and FRIEDMAN datasets. Models that perform statistically significantly better than all the other models in paired t-tests are indicated in bold, while methods performing comparably to the best method are indicated in italics. Our proposed GP-GMM leads to more precise effect estimation with lower RMSE and higher COR/LL for both marginal and component effects. In addition, Table 2 shows the averaged accuracy and STDs of preference prediction from GPCA and baselines on both synthetic datasets. GPCA has the best prediction for capturing the underlying preferential relations in the system.

Table 2: Averaged accuracy and STDs of preference prediction from GPCA and baselines on both synthetic datasets. GPCA has the best prediction for capturing the underlying preferential relations in the system.

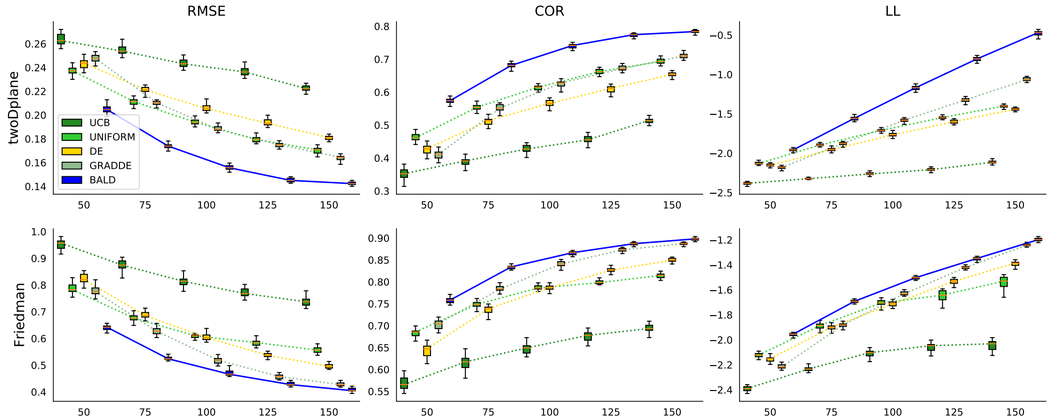
| DATASET | 2DPLANE | | | FRIEDMAN | | |
|---|---|---|---|---|---|---|
| | DIM | SVM | GPCA | DIM | SVM | GPCA |
| ACC | 0.696±0.006 | 0.824±0.003 | **0.986±0.002** | 0.785±0.006 | 0.795±0.005 | **0.956±0.002** |

### 6.2 Improved efficiency from adaptive experimentation

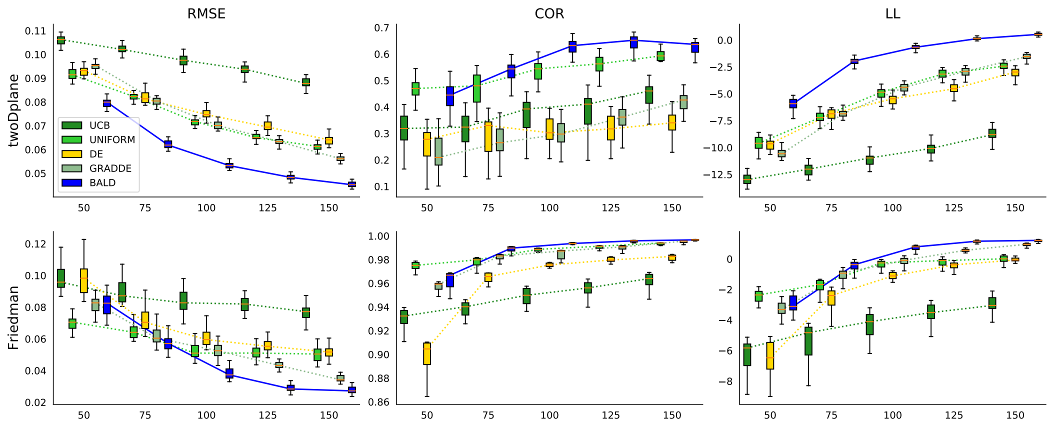
We then investigate adaptive experimentation in GPCA for increasing efficiency of effect estimation. We consider various policies: (1) UCB popular in multi-arm bandit setting [35], (2) DE-U and DE-ME for active learning based on differential entropy [36–38], (3) BALD in Bayesian active learning for model uncertainty reduction [39] and (4) UNIFORM design in non-parametric conjoint analysis [1, 2].

**Experimental details.** We initialize all the policies with the same 25 profile pairs from the 1000 candidate pairs, and update model posterior in GPCA once new preferences are revealed. Since the sampled profile distributions from each policy differ from each other due to their adaptive essence, we estimate the marginal and component effects w.r.t the same target profile distribution to ensure comparability. Specifically, we train our GPCA model on revealed preferences from profile pairs acquired so far and estimate both effects using GP-GMM at all the 1000 pairs.

**Results.** Figure 3 shows box plots of averaged RMSE, COR and LL and their STDs of marginal (top panel) and component (bottom panel) effects with adaptive experimentation under different acquisition policies. Sample size range from 50 to 150, and performance metrics are reported every other 25 acquisitions. Overall BALD (blue) outperforms the rest of policies including UNIFORM and UCB, indicating higher efficiency for effect estimation when the acquisition is designed to reduce




(a) Marginal effects




(b) Component effects

Figure 3: Box plots of averaged RMSE, COR and LL and their STDs of marginal (top panel) and component (bottom panel) effects with adaptive experimentation under different acquisition policies. Sample size range from 50 to 150, and performance metrics are reported every other 25 acquisitions. Overall BALD (blue) outperforms the rest of policies including UNIFORM and UCB, indicating higher efficiency for effect estimation when the acquisition is designed to reduce model uncertainty.

model uncertainty. Morever, UCB (forest green) has overall the worst performance in estimating both marginal and component effects as it solely reinforces current belief on the probability of preference.

**Preference prediction**. Besides estimation of marginal effects, we also examine the model quality of GPCA by evaluating the prediction accuracy of unrevealed preferences among the not acquired profile pairs. Figure 4 shows the averaged accuracy and STDs of preference prediction by various policies. With as few as $50$ data points, GPCA manages to predict at least $80\%$ of the unrevealed preference and $95\%$ when $150$ data points are adaptively acquired by BALD.

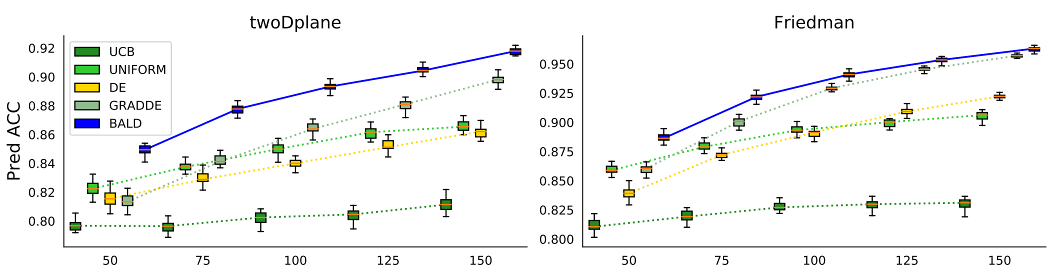



Figure 4: Averaged accuracy and STDs of preference prediction by various policies for simulated data.

### 6.3 Applications

**Data.** We apply GPCA to two real-world conjoint experiments: U.S. citizens' preferences across presidential candidates and attitudes toward immigrants containing 1733 and 6980 pairwise comparisons [1, 40]. Attributes in the candidate experiment include various aspects of candidates' personal background, demographics and issue positions, such as religion, education, profession, income and race, while attributes in the immigrant experiment include employment plans, job experience, language skills, country of origin, reasons for applying and so on.

Table 3: List of attributes with estimated component effects by GPCA and DIM used in the original studies, grouped by negative, null and positive effects.

| DATASET | GPCA \ DIM | NEG | NULL | POS |
|---|---|---|---|---|
| Candidate | NEG | Evangelical protestant, Mormon, car dealer,Age 68 | Jewish,Catholic,high school teacher, farmer,Income 210K,Black,Age 60 | — |
| | NULL | — | Mainline protestant,Lawyer,doctor,female, Income 54K,Hispanic,Asian American,Age 52 | Baptist college,Income 65K |
| | POS | — | Income 92K,5.1M,Caucasian, Native American,Age 45,75 | Military,community college, state university,Ivy League |
| Immigrant | NEG | India,China,will look for work, interview with employer,once as tourist | Broken English,Used interpreter,Germany, France,Mexico,Philippines,Poland,Iraq | — |
| | NULL | — | Mainline protestant,Lawyer,doctor,female, Income 54K,Hispanic,Asian American,Age 52 | — |
| | POS | — | Male,Somalia,financial analyst, waiter,child care provider | college degree,graduate degree,teacher,nurse,doctor computer programmer,research scientist,escape persecution |

**Results.** We run GPCA using all samples in both datasets. Table 3 shows the list of attributes with estimated component effects by GPCA and DIM used in original studies grouped by negative, null and positive effects. Overall, component effect estimation by GPCA is more reasonable. For example, in the candidate experiment GPCA found negative effects of Black candidates working as high school teachers or farmers on the probability of becoming U.S. presidents and positive effects of Caucasian candidates with 5.1M or more annual income, while DIM found no effects for any of these attributes. In the immigrant experiment, GPCA found negative effects of Iraqi applicants with broken English on the probability of immigration approval and positive effects of applicants working as financial analysts, while DIM found no effects. Figure 5 shows the averaged accuracy and STDs of preference prediction by various policies for real data with sample size varying from 100 to 800, where BALD has better prediction of unrevealed preferences than randomized policy.

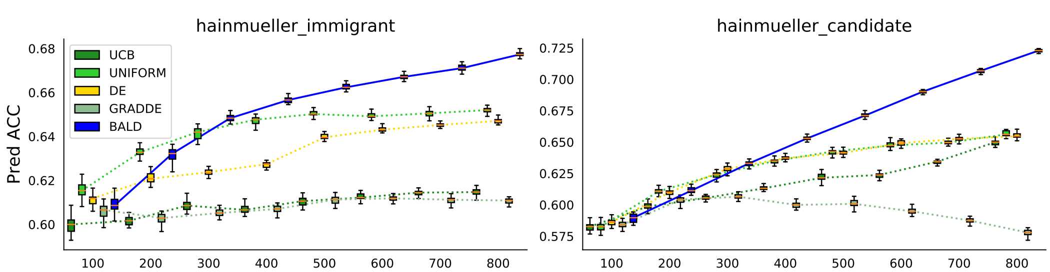



Figure 5: Averaged accuracy and STDs of preference prediction by various policies for real data, with sample size varying from 100 to 800. BALD has better prediction of unrevealed preferences than randomized policy.

## 7 Conclusion

We introduce GPCA, a Gaussian Process conjoint analysis model for estimating marginal effects in choice-based conjoint experiments. GPCA derives marginal effects as first-order derivatives and approximates their distributions using Gaussian mixtures, enhancing precision and efficiency in effect estimation aided by adaptive experimentation. GPCA has the potential of advancing causal inference in adaptive conjoint experiments. As distributional shifts are inevitable between adaptive acquired samples and uniformly randomized samples, directly interpreting marginal effects from adaptive samples in GPCA as causal effects may not be appropriate. Future research may explore methods such as inverse propensity weighting or doubly robust strategy for causal inference or feature interpretability in GPCA with adaptive experimentation.

## References


[1] Jens Hainmueller, Daniel J Hopkins, and Teppei Yamamoto. Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments. *Political Analysis*, 22(1):1–30, 2014.

[2] Naoki Egami and Kosuke Imai. Causal Interaction in Factorial Experiments: Application to Conjoint Analysis. *Journal of the American Statistical Association*, 2018.

[3] Thomas J Leeper, Sara B Hobolt, and James Tilley. Measuring Subgroup Preferences in Conjoint Experiments. *Political Analysis*, 28(2):207–221, 2020.

[4] Dana Naous and Christine Legner. Leveraging Market Research Techniques in IS: A Review and Framework of Conjoint Analysis Studies in the IS Discipline. In *Proceedings of the 38th International Conference on Information Systems (ICIS 2017)*, 2017.

[5] Eva-Maria Schomakers and Martina Ziefle. Privacy vs. Security: Trade-Offs in the Acceptance of Smart Technologies for Aging-in-Place. *International Journal of Human–Computer Interaction*, 39(5):1043–1058, 2023.

[6] Paul E Green and Vithala R Rao. Conjoint Measurement for Quantifying Judgmental Data. *Journal of Marketing Research*, 8(3):355–363, 1971.

[7] Paul E Green and Venkat Srinivasan. Conjoint Analysis in Marketing: New Developments with Implications for Research and Practice. *Journal of Marketing*, 54(4):3–19, 1990.

[8] Theodoros Evgeniou, Constantinos Boussios, and Giorgos Zacharia. Generalized Robust Conjoint Estimation. *Marketing Science*, 24(3):415–429, 2005.

[9] Sariel Har-Peled, Dan Roth, and Dav Zimak. Constraint Classification for Multiclass Classification and Ranking. *Advances in Neural Information Processing Systems*, 15, 2002.

[10] Olivier Chapelle and Zaid Harchaoui. A Machine Learning Approach to Conjoint Analysis. *Advances in Neural Information Processing Systems*, 17, 2004.

[11] Sebastián Maldonado, Ricardo Montoya, and Richard Weber. Advanced conjoint analysis using feature selection via support vector machines. *European Journal of Operational Research*, 241 (2):564–574, 2015.

[12] Wei Chu and Zoubin Ghahramani. Preference learning with Gaussian processes. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 137–144, 2005.

[13] Johannes Fürnkranz and Eyke Hüllermeier. Pairwise preference learning and ranking. In *Machine Learning: ECML 2003: 14th European Conference on Machine Learning, Cavtat-Dubrovnik, Croatia, September 22-26, 2003. Proceedings 14*, pages 145–156. Springer, 2003.

[14] Brochu Eric, Nando Freitas, and Abhijeet Ghosh. Active Preference Learning with Discrete Choice Data. *Advances in Neural Information Processing Systems*, 20, 2007.

[15] Shengbo Guo, Scott Sanner, and Edwin V Bonilla. Gaussian Process Preference Elicitation. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010.

[16] Neil Houlsby, Ferenc Huszar, Zoubin Ghahramani, and Jose Hernández-lobato. Collaborative Gaussian Processes for Preference Learning. *Advances in Neural Information Processing Systems*, 25, 2012.

[17] Christian A Scholbeck, Giuseppe Casalicchio, Christoph Molnar, Bernd Bischl, and Christian Heumann. Marginal effects for non-linear prediction functions. *Data Mining and Knowledge Discovery*, pages 1–46, 2024.

[18] Roy George Douglas Allen. *Mathematical Analysis For Economists*. MacMillan and Co., Ltd., 1938.

[19] Daniel Lüdecke. ggeffects: Tidy Data Frames of Marginal Effects from Regression Models. *Journal of Open Source Software*, 3(26):772, 2018.

[20] Jens Hainmueller and Chad Hazlett. Kernel Regularized Least Squares: Reducing Misspecification Bias with a Flexible and Interpretable Machine Learning Approach. *Political Analysis*, 22 (2):143–168, 2014.

[21] Rogério Luiz Cardoso Silva Filho, Paulo Jorge Leitão Adeodato, and Kellyton dos Santos Brito. Interpreting Classification Models Using Feature Importance Based on Marginal Local Effects. In *Brazilian Conference on Intelligent Systems*, pages 484–497. Springer, 2021.

[22] Michael Merz, Ronald Richman, Andreas Tsanakas, and Mario V Wüthrich. Interpreting deep learning models with marginal attribution by conditioning on quantiles. *Data Mining and Knowledge Discovery*, 36(4):1335–1370, 2022.

[23] Burr Settles. *Active Learning Literature Survey*. University of Wisconsin-Madison Department of Computer Sciences, 2009.

[24] Turab Lookman, Prasanna V Balachandran, Dezhen Xue, and Ruihao Yuan. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *npj Computational Materials*, 5(1):21, 2019.

[25] Suyu Liu and Ying Yuan. Bayesian optimal interval designs for phase I clinical trials. *Journal of the Royal Statistical Society: Series C: Applied Statistics*, pages 507–523, 2015.

[26] Jakob Richter, Tim Friede, and Jörg Rahnenführer. Improving adaptive seamless designs through Bayesian optimization. *Biometrical Journal*, 64(5):948–963, 2022.

[27] Roman Garnett, Thomas Gärtner, Martin Vogt, and Jürgen Bajorath. Introducing the 'active search' method for iterative virtual screening. *Journal of Computer-Aided Molecular Design*, 29:305–314, 2015.

[28] Xi Chen, Paul N Bennett, Kevyn Collins-Thompson, and Eric Horvitz. Pairwise ranking aggregation in a crowdsourced setting. In *Proceedings of the sixth ACM International Conference on Web Search and Data Mining*, pages 193–202, 2013.

[29] Erdem Bıyık, Nicolas Huynh, Mykel J Kochenderfer, and Dorsa Sadigh. Active Preference-Based Gaussian Process Regression for Reward Learning. *arXiv preprint arXiv:2005.02575*, 2020.

[30] Maria Dimakopoulou, Zhengyuan Zhou, Susan Athey, and Guido Imbens. Estimation Considerations in Contextual Bandits. *arXiv preprint arXiv:1711.07077*, 2017.

[31] Molly Offer-Westort, Alexander Coppock, and Donald P Green. Adaptive Experimental Design: Prospects and Applications in Political Science. *American Journal of Political Science*, 65(4): 826–844, 2021.

[32] Aurélien Bibaut, Maria Dimakopoulou, Nathan Kallus, Antoine Chambaz, and Mark van Der Laan. Post-Contextual-Bandit Inference. *Advances in Neural Information Processing Systems*, 34:28548–28559, 2021.

[33] Carl Edward Rasmussen and Christopher K Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006.

[34] Milton Abramowitz and Irene A Stegun. *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*, volume 55. US Government Printing Office, 1948.

[35] Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, page 1015–1022, Madison, WI, USA, 2010. Omnipress. ISBN 9781605589077.

[36] Shujin Sun, Ping Zhong, Huaitie Xiao, and Runsheng Wang. Active Learning With Gaussian Process Classifier for Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 53(4):1746–1760, 2014.

[37] Jens Schreiter, Duy Nguyen-Tuong, Mona Eberts, Bastian Bischoff, Heiner Markert, and Marc Toussaint. Safe Exploration for Active Learning with Gaussian Processes. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part III 15*, pages 133–149. Springer, 2015.

[38] Xiaowei Yue, Yuchen Wen, Jeffrey H Hunt, and Jianjun Shi. Active Learning for Gaussian Process Considering Uncertainties With Application to Shape Control of Composite Fuselage. *IEEE Transactions on Automation Science and Engineering*, 18(1):36–46, 2020.

[39] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian Active Learning for Classification and Preference Learning. *arXiv preprint arXiv:1112.5745*, 2011.

[40] Jens Hainmueller and Daniel J Hopkins. The Hidden American Immigration Consensus: A Conjoint Analysis of Attitudes toward Immigrants. *American Journal of Political Science*, 59 (3):529–548, 2015.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims accurately reflect this paper's contribution and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of the work is discussed in Sec (7).

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: This paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All the information needed to reproduce the main experimental results of this paper is reported in Sec (6).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The data and code to reproduce the main experimental results of this paper are uploaded as supplementary materials.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All the training and test details to understand the results of this paper are reported in Sec (6).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All error bars about the statistical significance of experiments are reported in Sec (6).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All computer resources needed to reproduce the experiments are reported in Sec (6).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in this paper conforms the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have discussed the societal impacts of our work in Sec (7).

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All assets used in this paper are properly cited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.