Total Variation Floodgate for Variable Importance Inference in Classification

Wenshuo Wang Lucas Janson[1] **Lihua Lei**[2] **Aaditya Ramdas**[3]

Abstract

Inferring variable importance is the key goal of many scientific studies, where researchers seek to learn the effect of a feature X on the outcome Y in the presence of confounding variables Z. Focusing on classification problems, we define the expected total variation (ETV), which is an intuitive and deterministic measure of variable importance that does not rely on any model assumption. We then introduce algorithms for statistical inference on the ETV under design-based/model-X assumptions. We name our method Total Variation Floodgate in reference to its shared high-level structure with the Floodgate method of Zhang & Janson (2020). The algorithms we introduce can leverage any user-specified regression function and produce asymptotic lower confidence bounds for the ETV. We show the effectiveness of our algorithms with simulations and a case study in conjoint analysis on the US general election.

1. Introduction

1.1. Motivation

In many scientific studies, researchers would like to understand the effect of a feature X on a response variable Y, while controlling for potential confounding features Z. While this question is sometimes simplified to a hypothesis testing problem of "does X affect Y at all in the presence of Z", it is more desirable to follow up with "if so, by how much"; that is, we wish to provide a quantitative variable importance measure (VIM). In traditional statistical frameworks, such a follow-up question is addressed by postulating a parametric model of $\mathcal{L}(Y \mid X, Z)$ and looking at the inferred parameters. However, such parametric models are often limited in their capacity to capture complex relationships. This paper aims at defining a VIM for classification problems and conducting inference on it. As a design objective, this (population-level) VIM must be model-free, in that it does not rely on an underlying model assumptions. We would also like the VIM to be intuitive and easy to interpret.

1.2. Our Contribution

The main contributions of this work are listed below.

1. We propose the expected total variation (ETV) as a VIM that is well-defined for any type of variables (X, Y, Z). In this paper, we focus on categorical response variables Y, for which VIMs with rigorous statistical guarantees were rarely discussed in the literature and ETV has both an intuitive model-free interpretation and sound statistical properties.

2. We introduce algorithms that provide lower confidence bounds on the ETV without imposing any assumptions on the distribution $\mathcal{L}(Y \mid X, Z)$, but instead make the design-based/model-X assumption that we can sample from $\mathcal{L}(X \mid Z)$, which we discuss in Secion 2.2. We accompany our algorithms with hyperparameter choice recommendations and show that they work well in our extensive simulation results. Again, the same idea also works for continuous and binary Y.

3. We demonstrate the effectiveness of our algorithms in a real conjoint data analysis study.

In the remainder of this section, we will discuss related work and introduce notation. The mathematical definition of ETV will be given in Section 2. We will discuss the properties of ETV in Section 2.1. Section 2.2 is devoted to our main algorithm to conduct inference on ETV. We then study the algorithm parameters in Section 2.3 to facilitate practical applications and discuss a generalization of ETV in Section 2.4. Section 3 includes simulations on synthetic data to support the effectiveness of our proposed method. We then apply our method to a conjoint analysis example on political candidate preferences in Secton 4.

[1] Department of Statistics, Harvard University, Cambridge, MA, USA [2] Stanford Graduate School of Business, Stanford University, Stanford, CA, USA [3] Departments of Statistics and Machine Learning, Carnegie Mellon University, Pittsburgh, PA, USA. Correspondence to: Wenshuo Wang <wenshuowang1997@gmail.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

1.3. Related Work

The canonical VIM is defined through parametric models. When Y is categorical, the textbook approach is to parameterize $\mathcal{L}(Y \mid X, Z)$ with a generalized linear model (Agresti, 2015), and there has been work on parameter inference in high-dimensional sparse models (Van de Geer et al., 2014; Belloni et al., 2016). However, generalized linear models have limited capacity in capturing non-linear effects, and such parameter-based VIMs crucially rely on the model being well specified and become ill-defined when the model is misspecified.

A more contemporary line of work utilizes machine learning methods to capture variable importance (Fisher et al., 2019; Watson & Wright, 2021; Molnar et al., 2023) in a model-free manner. While these VIMs are well-defined without parametric assumptions, they are associated with a trained machine learning model, which depends on the model choice and the data itself. We aim to define a VIM whose population-level definition depends on the underlying distribution but does not depend on the model and data.

Another existing approach (Castro et al., 2009; Williamson & Feng, 2020; Ning et al., 2022) borrows ideas from the game theory literature and considers VIMs based on the Shapley value (Shapley, 1953). Shapley-value based VIMs capture the variable's predictive power, and are often positive even for statistically null variables (that is, X where $X \perp\!\!\!\perp Y \mid Z$). Thus, while these VIMs have attractive predictive interpretations, they lack a causal interpretation.

Azadkia & Chatterjee (2021) introduced a model-free VIM based on cumulative distributions functions for non-categorical Y; Huang et al. (2020) generalized it to categorical Y, but it relies on a user-specified kernel function. These VIMs have the appealing property that they are 0 if and only if $Y \perp\!\!\!\perp X \mid Z$ and 1 if and only if Y is a deterministic measurable function of (X, Z). Both papers considered consistent estimators of the VIMs and not lower confidence bounds like we do here. Zhang & Janson (2020) proposed a model-free VIM called the minimum mean squared error (mMSE) for non-categorical Y and provided a lower confidence bound. Zhang & Janson (2020, Section 3.1) extended their inference to a VIM called the mean absolute conditional mean (MACM) gap, which is defined for binary Y. Similarly, Williamson et al. (2021) defined a class of VIMs based on a variable's additional predictiveness and constructed estimators and confidence intervals. These VIMs hinge on a predetermined family of predictors, and while the VIMs could work for any type of responses, Williamson et al. (2021) also only considered inference in cases of non-categorical and binary responses where the conditional mean is a meaningful quantity.

To the best of our knowledge, our work is the first to propose a VIM for general categorical responses that is model-free, natural and easy to interpret, and provide a lower confidence bound for it.

Finally, there are two methods in the literature (Zhang & Janson, 2020; Näf et al., 2022) that have close connections with our proposed method. As it requires first introducing our algorithm to properly discuss comparisons with these methods, we will review these works and their relationship to ours in Section 2.5.

2. VIM Inference in Classification

For random variables or vectors W_1 and W_2, $\mathcal{L}(W_1)$ means the distribution of W_1 and $\mathcal{L}(W_1 \,|\, W_2)$ means the conditional distribution of W_1 given W_2. $\mathrm{TV}(\mathcal{L}_1, \mathcal{L}_2)$ means the total variation distance between distributions $\mathcal{L}_1$ and $\mathcal{L}_2$. Unless otherwise specified, vectors are column vectors. Φ denotes the cumulative distribution function of the standard Gaussian distribution $\mathcal{N}(0, 1)$.

2.1. The Expected Total Variation Distance

Let (X, Y, Z) be a random vector with three components. We would like to quantify the effect size of X; that is, the strength of the conditional dependence between X and Y given Z. We propose to use the expected total variation distance (ETV) between $\mathcal{L}(Y \mid X, Z)$ and $\mathcal{L}(Y \mid Z)$:

$$\mathrm{ETV}(X, Y, Z) := \frac{E[\mathrm{TV}(\mathcal{L}(Y \mid X, Z), \mathcal{L}(Y \mid Z))]}{1 - 1/|\mathcal{Y}|}, \quad (1)$$

as the VIM, where $|\mathcal{Y}|$ is the support size of Y (if $|\mathcal{Y}| = \infty$, we simply normalize by 1) and the expecation is taken over (X, Z). The normalizing factor is to ensure the value of ETV is in $[0, 1]$, as stated below.

Lemma 2.1 (Range of ETV). *If* $\mathbb{P}(Y \in \mathcal{Y}) = 1$, *then* $0 \leq \mathrm{ETV}(X, Y, Z) \leq 1$, *where the right equality is achieved when* X*, conditional on* Z*, deterministically determines* Y *and* $\mathbb{P}(Y = y \mid Z) = 1/|\mathcal{Y}|$ *for all* $y \in \mathcal{Y}$.

If X is continuous and $\mathbb{P}(Y = y \mid Z) = 1/|\mathcal{Y}|$ for all $y \in \mathcal{Y}$, then there always exists $\mathcal{L}(X \mid Z)$ such that Y is a deterministic function of (X, Z). Thus, for any support $\mathcal{Y}$, there exists (X, Y, Z) such that $\mathrm{ETV}(X, Y, Z)$ is equal to 1 for finite $|\mathcal{Y}|$, or gets arbitrarily close to 1 for $|\mathcal{Y}| = \infty$.

The ETV has several desirable properties that distinguish it from other VIMs in the literature:

1. ETV attains its minimum value zero if and only if $X \perp\!\!\!\perp Y \mid Z$. This means ETV captures the aggregation of all possible effects of X on Y conditional on Z, including both linear and non-linear effects.

2. ETV has a very simple and intuitive form and does not

depend on any pre-specified model or kernel function (that is, it is model-free).

3. ETV is particularly suitable to categorical Y because it does not change under one-to-one mapping of Y.

ETV's simple form should already make it very easy to interpret conceptually. The key component is the total variation distance, which is also the Wasserstein distance with 0-1 loss and half the the L_1 distance between probability density/mass functions. To visually demonstrate the ETV, we consider an example where Y could be one of 5 categories $s_1, \dots, s_5$. In Figure 1, the red bars represent the probability of Y being from each category conditional on (X, Z) taking on some particular value (x, z), the blue bars represent the probability of Y taking each value conditional on $Z = z$, and the yellow bars represent the difference. The expected sum of the yellow bars, when averaged over (X, Z), is equal to $2(1 - 1/|\mathcal{Y}|)$ ETV.

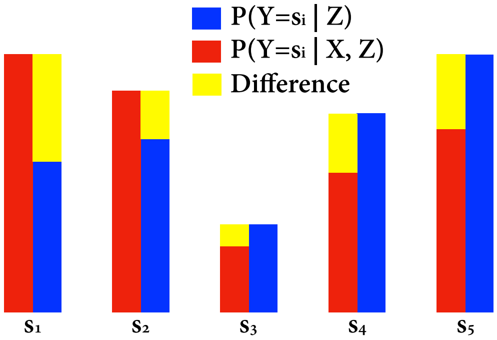

Figure 1. ETV illustration. The expected sum of the yellow bars, when averaged over (X, Z), is equal to $2(1 - 1/|\mathcal{Y}|)$ ETV.

We wish to note that ETV cannot be interpreted causally without certain standard causal assumptions. One case where those assumptions are met is the standard randomized experiments, where X is the treatment variable.

Finally, we briefly discuss the interpretation of ETV in the context of sensitivity analysis. When $Y \perp\!\!\!\perp X \mid Z$ does not hold, we can hypothesize the existence of a confounding variable U, such that $Y \perp\!\!\!\perp X \mid Z, U$. We can define $B(X, Z, U) \geq 1$ as the (almost sure) supremum of

$$\max\left\{\frac{p(X \mid U, Z)}{p(X \mid Z)}, \frac{p(X \mid Z)}{p(X \mid U, Z)}\right\}.$$

Therefore, B measures the minimal confounding effect that can explain away the conditional non-independence. We can show that $B(X, Z, U) \geq 1 + 2(1 - 1/|\mathcal{Y}|)\,\mathrm{ETV}(X, Y, Z)$ (see Appendix B). This interpretation is very relevant for low-signal problems such as the genome-wide association studies (GWAS), where most signals are weak and one would want to know the sensitivity of the conditional non-independence.

2.2. Lower Confidence Bound for the ETV

Having introduced ETV as a VIM, we now focus on designing algorithms to do inference on it. We recognize that (1) measures the distinction between the distributions

$$\mathcal{L}(X, Y, Z) \text{ and } \mathcal{L}(Y \mid Z) \times \mathcal{L}(X \mid Z) \times \mathcal{L}(Z), \tag{2}$$

where $\mathcal{L}(Y \mid Z) \times \mathcal{L}(X \mid Z) \times \mathcal{L}(Z)$ represents the joint distribution of (X, Y, Z) when $Z \sim \mathcal{L}(Z)$, $X \mid Z \sim \mathcal{L}(X \mid Z)$, $Y \mid Z \sim \mathcal{L}(Y \mid Z)$, and $X \perp\!\!\!\perp Y \mid Z$. We can see this by

$$\begin{aligned}
&2E[\mathrm{TV}(\mathcal{L}(Y \mid X, Z), \mathcal{L}(Y \mid Z))]\\
&= \int_z \left(\int_x \left(\int_y |p(y|x,z) - p(y|z)|\, dy\right) p(x|z)\, dx\right) p(z)\, dz\\
&= \int_z \int_x \int_y |p(y,x,z) - p(y|z)p(x|z)p(z)|\, dy\, dx\, dz\\
&= 2\,\mathrm{TV}(\mathcal{L}(X, Y, Z), \mathcal{L}(Y \mid Z) \times \mathcal{L}(X \mid Z) \times \mathcal{L}(Z)).
\end{aligned}$$

In fact, the ETV measure (1) exactly corresponds to the optimal error rate when classifying samples into the two populations. We state a general result in Theorem 2.2 below.

Theorem 2.2. *Let π_0 and π_1 be two distributions supported on the same continuous or discrete space Ω. Let π be the distribution of (ω, A), where $A \sim \mathrm{Bern}(a)$ and $\omega \mid A \sim \pi_A$. Then for any $f : \Omega \to [0, 1]$,*

$$1 - E_\pi\left[\frac{1}{a} I(A = 1)(1 - f(\omega)) + \frac{1}{1-a} I(A = 0) f(\omega)\right] \leq \mathrm{TV}(\pi_1, \pi_0),$$

where the equality could be achieved by an optimal f.

Armed with this observation, we can convert the task of inferring ETV into inferring the classification error rate of samples from the two distributions in (2). Note that this binary classification has nothing to do with $|\mathcal{Y}|$, the support size of Y, and all results in Sections 2.1 to 2.4 are agnostic to $|\mathcal{Y}|$. Using the available samples from $\mathcal{L}(X, Y, Z)$, the design-based/model-X approach (Rubin, 1974; Holland, 1986; Janson, 2017; Candès et al., 2018) allows us to obtain samples from $\mathcal{L}(Y \mid Z) \times \mathcal{L}(X \mid Z) \times \mathcal{L}(Z)$ by utilizing our ability to sample from $\mathcal{L}(X \mid Z)$. In randomized experiments (Rubin, 1974; Holland, 1986), $\mathcal{L}(X \mid Z)$ is known by design, such as the example of conjoint analysis in Section 4. Even for some observational data sets, $\mathcal{L}(X \mid Z)$ can be estimated accurately from unlabeled samples of (X, Z) without Y. We acknowledge that this is a limitation of our algorithms, while we wish to point out that inference on conditional inference is in general an impossible task without extra assumptions; specifically, when Z is continuous, Shah & Peters (2020) showed that a universally valid conditional independence test must be trivial.

Assuming[1] it is possible to sample from $\mathcal{L}(X \mid Z)$, we have the following simple corollary to Theorem 2.2.

Corollary 2.3. *Let $(X^{(0)}, Y, Z) \sim p(z)p(x|z)p(y|x,z) \equiv p(x,y,z)$ be the original data. Draw $X^{(1)}, \dots, X^{(J)} \mid X^{(0)}, Y, Z \overset{\text{i.i.d.}}{\sim} p(x|Z)$. Then for any classifier $f : (\mathcal{X}, \mathcal{Y}, \mathcal{Z}) \to [0,1]$, we have*

$$1 - E\left[1 - f(X^{(0)}, Y, Z) + \frac{1}{J}\sum_{j=1}^{J} f(X^{(j)}, Y, Z)\right] \leq \frac{1}{2}\int \Big|p(y|x,z) - p(y|z)\Big| p(x|z)p(z)\, dx\, dy\, dz\,.$$

In light of Corollary 2.3, we can design an algorithm that produces a lower confidence bound on (1) using the central limit theorem. The algorithm works by first choosing a classification function f as in Corollary 2.3, which can be understood as a labeling function, then using the sample mean and variance of classification error to produce a lower confidence bound for the real classification accuracy rate, which (by the corollary) is itself a lower bound on the optimal classification accuracy rate and a rescaled ETV.

Algorithm 1 Total variation floodgate.

```
Input: An i.i.d. data set $(X_i, Y_i, Z_i)_{i=1}^n$, conditional distribution $\mathcal{L}(X \mid Z)$ and a classifier $f : (\mathcal{X}, \mathcal{Y}, \mathcal{Z}) \to [0,1]$, number of resamples $J$, $Y$'s support size $|\mathcal{Y}|$, confidence level $\alpha \in (0,1)$
Output: a lower confidence bound for $\mathrm{ETV}(X, Y, Z)$
for $i = 1$ to $n$ do
    Draw $X_i^{(1)}, \dots, X_i^{(J)} \overset{\text{i.i.d.}}{\sim} \mathcal{L}(X \mid Z = Z_i)$.
    Set $(Y_i^{(j)}, Z_i^{(j)}) = (Y_i, Z_i)$, $E_i = 1$ and $E_i^{(j)} = 0$, $1 \leq j \leq J$.
    $f_i \leftarrow f(X_i, Y_i, Z_i)$.
    $f_i^{(j)} \leftarrow f(X_i^{(j)}, Y_i^{(j)}, Z_i^{(j)})$.
    $L_i \leftarrow (|f_i - 1| + (1/J)\sum_{j=1}^{J} |f_i^{(j)} - 0|)$.
end for
$\bar{L} \leftarrow \sum_{i=1}^n L_i/n$, $\bar{L^2} \leftarrow \sum_{i=1}^n L_i^2/n$.
Return $L_n^\alpha(f) = \frac{\max(0, 1 - \bar{L} - z_\alpha \frac{\sqrt{\bar{L^2} - \bar{L}^2}}{\sqrt{n}})}{(1 - 1/|\mathcal{Y}|)}$, where $z_\alpha$ satisfies $1 - \Phi(z_\alpha) = \alpha$.
```

This idea of producing a lower confidence bound on a lower bound on the quantity of interest is metaphorically termed "floodgate" in Zhang & Janson (2020), hence the name of Algorithm 1. Note that there is an oracle f that provides the best lower confidence bound, in the sense given in Theorem 2.4; thus, the coverage of Algorithm 1 can be tight.

Theorem 2.4 (Validity of Algorithm 1)**.** *For any given f and $\alpha \in (0,1)$, $\lim_{n\to\infty} \mathbb{P}(\mathrm{ETV} \geq L_n^\alpha(f)) \geq 1 - \alpha$, where $L_n^\alpha(f)$ is defined in Algorithm 1. Additionally, $\lim_{n\to\infty} \mathbb{P}(\mathrm{ETV} \geq L_n^\alpha(f_{\text{oracle}})) = 1 - \alpha$, where*

$$\begin{aligned} f_{\text{oracle}}(x,y,z) &= \mathbb{I}(p(y \mid x,z) > p(y \mid z)) \\ &= \mathbb{I}\left(\frac{p(y \mid x,z)}{p(y \mid x,z) + p(y \mid z)} > 0.5\right). \end{aligned} \quad (3)$$

The proof of Theorem 2.4 follows directly from Corollary 2.3 and the central limit theorem once we note that L_i's are bounded and independent and identically distributed.

A natural question that the reader may have is whether we could provide an *upper* confidence bound for the ETV. We present Theorem 2.5, which states that in some sense the answer is no: a generic confidence upper bound on the ETV must simply cover the theoretical upper bound even under the most ideal scenario: there is no Z variable, X and Y are independent, X's distribution is known, and Y's distribution is uniform.

Theorem 2.5. *Let $(X_i, Y_i)_{i=1}^n$ be i.i.d. samples from $\mathcal{L}$, where the marginal distribution of Y_i is $\mathrm{Unif}(\{1, \dots, K\})$. Let $C_{\mathcal{L}_X}$ be an algorithm tailored for the marginal distribution of X_i that takes $(X_i, Y_i)_{i=1}^n$ as input and produces a confidence upper bound, such that $\mathbb{P}_{\mathcal{L}}(C_{\mathcal{L}_X}(X_{1:n}, Y_{1:n}) \geq \mathrm{ETV}(X, Y)) \geq 1 - \alpha$, $\alpha \in (0,1)$, for any $\mathcal{L}$ that respects the marginal distributions $\mathcal{L}_X$, where $\mathrm{ETV}(X, Y)$ is (1) with an empty Z. Then,*

$$\mathbb{P}(C_{\mathcal{L}_X}(X_{1:n}, Y_{1:n}) \geq 1) \geq 1 - \alpha \quad (4)$$

when $(X_i, Y_i) \overset{\text{i.i.d.}}{\sim} \mathcal{L}_X \times \mathrm{Unif}(\{1, \dots, K\})$, where $\mathcal{L}_X$ is a continuous distribution and 1 is the theoretical ETV upper bound given by Lemma 2.1.

To provide some intuition on Theorem 2.5, we can understand the hardness of producing an upper bound by thinking about the general problem of upper bounding the total variance distance between $\mathcal{L}_1$ and $\mathcal{L}_2$. By writing $\mathrm{TV}(\mathcal{L}_1, \mathcal{L}_2) = \sup_A |\mathcal{L}_1(A) - \mathcal{L}_2(A)|$ (where the supremum is over measurable sets A), we can easily obtain a lower bound for TV by fixing a nontrivial set A, and it is then straightforward to empirically estimate the lower bound $|\mathcal{L}_1(A) - \mathcal{L}_2(A)|$. However, in order to upper bound or estimate the actual TV, one would need to be able to consistently estimate the set $\arg\max_A |\mathcal{L}_1(A) - \mathcal{L}_2(A)|$.

[1] This assumption can be relaxed if one is willing to assume certain parametric models for $\mathcal{L}(X \mid Z)$ (Zhang & Janson, 2020, Section 3.2), and the floodgate approach for other VIMs has been extended to be doubly robust, but such an extension is beyond the scope of this paper.

For the ETV, this translates to consistently estimating the optimal classifier f given by (3), which requires to impose conditions on $\mathcal{L}(Y \mid X, Z)$. On the other hand, we do not need such conditions to produce a lower confidence bound.

Moving back to Algorithm 1, the function f in practice would have to be trained on a separate dataset to maintain validity, which is not fully utilizing the whole dataset. Next, we show how to apply cross-validation in a way that every data point is used for inference.

Data Splitting, Cross Validation and Derandomization. To avoid excluding any data in the inference step, we use the idea of cross-validated floodgate. The idea borrows results from central limit theorems for cross-validation (Austern & Zhou, 2020; Bayle et al., 2020) and ensures the validity of Algorithm 2, a cross-validated version of Algorithm 1.

Algorithm 2 Cross-validated total variation floodgate.

```
Input: An i.i.d. data set $(X_i, Y_i, Z_i)_{i=1}^n$, conditional distribution $\mathcal{L}(X \mid Z)$ and a classifier training rule $f$, number of resamples $J$, number of CV folds $k$, $Y$'s support size $|\mathcal{Y}|$, confidence level $\alpha \in (0,1)$
Output: a lower confidence bound for $\mathrm{ETV}(X, Y, Z)$
Randomly partition the data into $k$ folds $B_1^c, \dots, B_k^c$ with sizes differing by at most one.
for $r = 1$ to $k$ do
    Train a classifier $f_{B_r}$ with data $B_r$, plug in $f_{B_r}$ and data $(X_i, Y_i, Z_i)_{i \in B_r^c}$ to Algorithm 1, and record the sample mean and sample variance of the $L$ vector as $\hat{\mu}_r$ and $\hat{\sigma}_r^2$.
end for
$\hat{\mu} \leftarrow \sum_{r=1}^k \hat{\mu}_r / k$, $\hat{\sigma}^2 \leftarrow \sum_{r=1}^k \hat{\sigma}_r^2 / k$.
Return $L_n^\alpha(f) = \max(0, (1 - \hat{\mu} - z_\alpha \hat{\sigma}/\sqrt{n})/(1 - 1/|\mathcal{Y}|))$, where $z_\alpha$ satisfies $1 - \Phi(z_\alpha) = \alpha$.
```

Theorem 2.6 (Validity of Algorithm 2). *For any given f and $\alpha \in (0,1)$, let*

$$h_n((x, x^{(1:J)}, y, z); B_1) = |f_{B_1}(x, y, z) - 1| + \frac{1}{J}\sum_{j=1}^J |f_{B_1}(x^{(j)}, y, z)|,$$

$$\bar{h}_n((x, x^{(1:J)}, y, z)) = E_{B_1}[h_n((x, x^{(1:J)}, y, z); B_1)],$$

$$\sigma_n = \sqrt{\mathrm{Var}(\bar{h}_n((X, X^{(1:J)}, Y, Z)))}$$

where the subscript B_1 means taking expectation over B_1, which contains the $n(1 - 1/k)$ training samples for f_{B_1}. Assume

(a) $\left(\bar{h}_n((X, X^{(1:J)}, Y, Z)) - E[\bar{h}_n((X, X^{(1:J)}, Y, Z))]\right)/\sigma_n^2$ is uniformly integrable;

(b) and the asymptotic linearity condition (2.2) in Bayle et al. (2020) holds in probability:

$$\frac{1}{\sigma_n\sqrt{n}}\sum_{r=1}^k\sum_{i\in B_r^c}\Big((h_n(X_i, X_i^{1:J}, Y_i, Z_i); B_r) - E[h_n(X_i, X_i^{1:J}, Y_i, Z_i); B_r) \mid B_r] - \Big(\bar{h}_n((X, X^{(1:J)}, Y, Z)) - E[\bar{h}_n((X, X^{(1:J)}, Y, Z))]\Big)\Big) \xrightarrow{p} 0.$$

Then, $\lim_{n\to\infty}\mathbb{P}(\mathrm{ETV} \geq L_n^\alpha(f)) \geq 1 - \alpha$. *Additionally,* $\lim_{n\to\infty}\mathbb{P}(\mathrm{ETV} \geq L_n^\alpha(f_{\mathrm{oracle}})) = 1 - \alpha$, *where f_{oracle} is given by* (3).

Assumption (a) holds if $\bar{h}_n((X, X^{(1:J)}, Y, Z))$ does not converge to a degenerate distribution. Section 3 in Bayle et al. (2020) discussed some sufficient conditions of assumption (b). Notably, when the number of cross-validation folds $k = O(1)$, then a sufficient condition for (b) is

$$\frac{E[\mathrm{Var}[h_n((X, X^{(1:J)}, Y, Z); B_1) \mid (X, X^{(1:J)}, Y, Z)]]}{\mathrm{Var}(\bar{h}_n((X, X^{(1:J)}, Y, Z)))} \to 0 \text{ in probability.}$$

Assuming the denominator converges to a positive constant, this condition says that the out-of-sample loss is asymptotically stable over the random training sample. Because h_n is bounded, the conditions of Theorem 2.6 hold if there exists f_* such that $f_{B_1}(x, y, z) \to f_*(x, y, z)$ in probability, uniformly over (x, y, z).

2.3. Classification Function

In this section, we discuss how to train the function f in Algorithms 1 and 2.

By looking at the ultimate goal of f, which is to predict whether X is a resample from $p(x|z)$ or the original sample, a naive approach is to train f by regressing E on (X, Y, Z) using samples

$$(E_i^{(j)}, (X_i^{(j)}, Y_i, Z_i)), i = 1, \dots, n, j = 0, \dots, J,$$

where $X_i^{(0)} = X_i$ and $E_i^{(j)} = \mathbb{I}(j = 0)$. However, this approach is ignoring important structural information. From the proof of Corollary 2.3, the oracle f that minimizes the expected error rate is the one given in (3), which motivates the following choice of f in practice

$$f(x, y, z; p_{\theta_1}, p_{\theta_2}, c) = \begin{cases} 1, & \hat{p}_i > 0.5 + c, \\ 0, & \hat{p}_i < 0.5 - c \\ 0.5, & |\hat{p}_i - 0.5| \leq c, \end{cases} \quad (5)$$

where

$$\hat{p}_i = \frac{p_{\hat{\theta}_1}(y \mid x, z)}{p_{\hat{\theta}_1}(y \mid x, z) + p_{\hat{\theta}_2}(y \mid z)},$$

$\hat{\theta}_1$ and $\hat{\theta}_2$ are parameter estimates of working models $p_{\theta_1}(y \mid x, z)$ and $p_{\theta_2}(y \mid z)$ and c sets a buffer to account for the estimation error and gives an extra degree of freedom. The working models can be from any model family, including simple generalized linear models and more sophisticated machine learning models. The accuracy of the working models directly impacts the performance of our algorithms. While an agnostic model may already have adequate performance, as we show in Section 3.1, practitioners are encouraged to incorporate domain knowledge into building these working models. The logic behind such f is that we classify the sample as 0 or 1 depending on which has the higher estimated likelihood, but when the two likelihoods are close and we are not sure, we set it to 0.5 and essentially discard the sample. The parameter c controls our comfort level of confidence. We will show the empirical effect of c in Section 3.2.

2.4. Generalization to Hierarchical Responses

In some cases, the response Y may have several levels, arranged in a hierarchy. For example, a wolf is also a type of dog, which is also an animal. We can choose to relabel wolf to dog or animal to reflect the relevant level of granularity. It is then straightforward to apply Algorithms 1 and 2 to the relabeled data. We wish to raise a subtle yet crucial point that one cannot simply drop certain labels. For example, if one only cares about a feature X's ability to distinguish $Y = \mathrm{A}$ from $Y = \mathrm{B}$, one might be tempted to simply drop all samples where $Y \notin \{\mathrm{A}, \mathrm{B}\}$. However, doing so would require one to be able to sample from $\mathcal{L}(X \mid Z, Y \in \{\mathrm{A}, \mathrm{B}\})$ to apply Algorithms 1 or 2, which is a different assumption from being able to sample from $\mathcal{L}(X \mid Z)$.

If all values of Y have the same number of levels, then we can define an overall VIM by weighting all levels. Let $Y = (Y_1, \ldots, Y_K)$ have K hierarchy levels, where for any possible values Y and $\tilde{Y}$, if $Y_k \neq \tilde{Y}_k$, then $Y_{k'} \neq \tilde{Y}_{k'}$ for all $k' > k$. For instance, we can let $K = 3$ and Y_1, Y_2, Y_3 be the taxonomic ranks of family, genus and species (a genus consists of many species, etc.), so if two labels disagree at the genus level, they must necessarily also disagree on the species. Next, we define a VIM at each level $k > 1$:

$$\begin{aligned}\mathrm{HETV}_k(X, Y, Z) = (1 - 1/|\mathcal{Y}_{1:k}|)\,\mathrm{ETV}(X, Y_{1:k}, Z) \\ - (1 - 1/|\mathcal{Y}_{1:k-1}|)\,\mathrm{ETV}(X, Y_{1:(k-1)}, Z),\end{aligned}$$

where we add back the normalizing constant to ensure $\mathrm{HETV}_k(X, Y, Z) \geq 0$. A sufficient condition of $\mathrm{HETV}_k(X, Y, Z) = 0$ is

$$\begin{aligned}&\mathcal{L}(Y_k \mid X, Y_{1:(k-1)}, Z) = \mathcal{L}(Y_k \mid Y_{1:(k-1)}, Z), \\ &\quad \text{equivalently } \mathcal{L}(X \mid Y_{1:k}, Z) = \mathcal{L}(X \mid Y_{1:(k-1)}, Z),\end{aligned}$$

so we can interpret $\mathrm{HETV}_k(X, Y, Z)$ as an ETV-based VIM of X at hierarchy level k. Finally, we define an overall VIM of X by aggregating $\mathrm{ETV}(X, Y_1, Z)$ and HETV at all others levels with a user-specified weight vector w:

$$\begin{aligned}\mathrm{HETV}_w(X, Y, Z) = w_1(1 - 1/|\mathcal{Y}_1|)\,\mathrm{ETV}(X, Y_1, Z) \\ + \sum_{k=2}^{K} w_k\,\mathrm{HETV}_k(X, Y, Z).\end{aligned}$$

Here, $w_{1:K}$ is a sequence of non-increasing weights, so that the coefficient for each ETV is non-negative. We can then modify Algorithm 1 to support HETV, as below.

Algorithm 3 Hierarchically weighted TV floodgate.

```
Input: An i.i.d. data set $(X_i, Y_i, Z_i)_{i=1}^n$, conditional distribution $\mathcal{L}(X \mid Z)$ and classifiers $f_k : (\mathcal{X}, \mathcal{Y}_{1:k}, \mathcal{Z}) \to [0, 1]$, number of resamples $J$, confidence level $\alpha \in (0, 1)$
Output: a lower confidence bound for $\mathrm{ETV}_w(X, Y, Z)$
for $i = 1$ to $n$ do
  Draw $X_i^{(1)}, \ldots, X_i^{(J)} \overset{\text{i.i.d.}}{\sim} \mathcal{L}(X \mid Z = Z_i)$.
  Set $(Y_i^{(j)}, Z_i^{(j)}) = (Y_i, Z_i)$, $E_i = 1$ and $E_i^{(j)} = 0$, $1 \leq j \leq J$.
  for $k = 1$ to $K$ do
    $f_{i,k} \leftarrow f_k(X_i, Y_{1:k,i}, Z_i)$.
    $f_{i,k}^{(j)} \leftarrow f_k(X_i^{(j)}, Y_{1:k,i}^{(j)}, Z_i^{(j)})$.
    $L_{i,k} \leftarrow (|f_{i,k} - 1| + (1/J) \sum_{j=1}^{J} |f_{i,k}^{(j)} - 0|)$.
  end for
  $L_i \leftarrow w_1 L_{i,1} + \sum_{k=2}^{K} w_k (L_{i,k} - L_{i,k-1})$.
end for
$\bar{L} \leftarrow \sum_{i=1}^{n} L_i/n$, $\bar{L^2} \leftarrow \sum_{i=1}^{n} L_i^2/n$.
Return $L_n^\alpha(f) = \max(0, 1 - \bar{L} - z_\alpha \sqrt{\bar{L^2} - \bar{L}^2}/\sqrt{n})$, where $z_\alpha$ satisfies $1 - \Phi(z_\alpha) = \alpha$.
```

In the same way as done above, we could also modify Algorithm 2 to work for HETV.

2.5. Relationship with Literature

Having introduced the definition of ETV and algorithms for inference, we pause to discuss two recent related works.

Connection to the MACM Gap in Zhang & Janson (2020) For the specific case where $Y \in \{1, -1\}$, Zhang & Janson (2020) defined the MACM gap, which has exactly twice the value of ETV. Their inference (Zhang & Janson, 2020, Algorithm 3) is equivalent to our Algorithm 1 with

$$f(X,Y,Z) = \begin{cases} \mathbb{I}(\mu(X,Z) \geq E[\mu(X,Z) \mid Z]), \text{ if } Y=1, \\ \mathbb{I}(\mu(X,Z) \leq E[\mu(X,Z) \mid Z]), \text{ if } Y=-1. \end{cases} \tag{6}$$

The details are deferred to Appendix C.

Connection to $\hat{\lambda}^{\rho}_{\text{bayes}}$ in Näf et al. (2022) Näf et al. (2022) studied lower confidence bounds for $\text{TV}(P,Q)$ based on i.i.d. samples from P and Q. One of their proposed estimators, $\hat{\lambda}^{\rho}_{\text{bayes}}$ in Näf et al. (2022, Proposition 3), is based on the same classification idea as Algorithm 1. Specifically, Näf et al. (2022) also utilized the relationship between the classification accuracy and the total variation distance, and $\hat{\lambda}^{\rho}_{\text{bayes}}$ is constructed based on this fact for a fixed classification function $\rho_t(x) = \mathbb{I}(\rho(x) > t)$ with $t = 0.5$. Similar to our discussion around the parameter c in Section 2.3, Näf et al. (2022) showed that there may exist better choices for t in $\rho_t(x)$ than the natural $t = 0.5$, depending on prior knowledge of P and Q. While we propose to use cross-validation to choose c, Näf et al. (2022) went on to consider estimators very different from $\hat{\lambda}^{\rho}_{\text{bayes}}$. While our method shares the same construction idea as $\hat{\lambda}^{\rho}_{\text{bayes}}$ in Näf et al. (2022), the key difference between the two works is the problem setting. Näf et al. (2022) studied two-sample testing, where the samples are naturally labeled with auxiliary information; our work is centered around the ETV, which is a novel VIM defined through a sample-labeling mechanism based on $\mathcal{L}(X \mid Z)$. Thus, while the algorithmic ideas are similar, our motivations and applications are quite different.

3. Simulations

The code to implement ETV floodgate and replicate all experiments is available at https://github.com/wenshuow/etv_floodgate. Results are obtained from a number of independent experiments. Each experiment has a brief runtime, consistently below 10 minutes on a single CPU. Given that standard errors are under 0.01 across all experiments, the error bars on the plots are deemed negligible and have been excluded.

3.1. Effect of the Classification Function

In this section, we consider the model

$$Y \mid X \sim \text{Bern}(\Phi(X^\top \beta)), X \sim \mathcal{N}(0, \Sigma), \tag{7}$$

where X is a p-dimensional column vector and we provide lower confidence bound for $\text{ETV}(X_j, Y, X_{\text{-}j})$ for each j. We choose $\Sigma_{ij} = \rho^{|i-j|}$. We set $p = 4$ or 10, $\beta = (0,1,2,3)$ for $p = 4$ and $\beta = (0,0,0,0,1,2,3,4,5,6)$ for $p = 10$, $n = 100p$, and apply 10-fold cross validation in Algorithm 2. We use three types of classification functions as in (5). For the oracle model, p_{θ_1} and p_{θ_2} are set to the true models. For the logistic or tree models, p_{θ_1} and p_{θ_2} are logistic or tree models, and $\hat{\theta}_1$ and $\hat{\theta}_2$ are trained on cross-validated data. We find that the oracle gives the highest lower confidence bound (as expected), and the logistic model is a close second. Even the generic random forest model performs reasonably well.

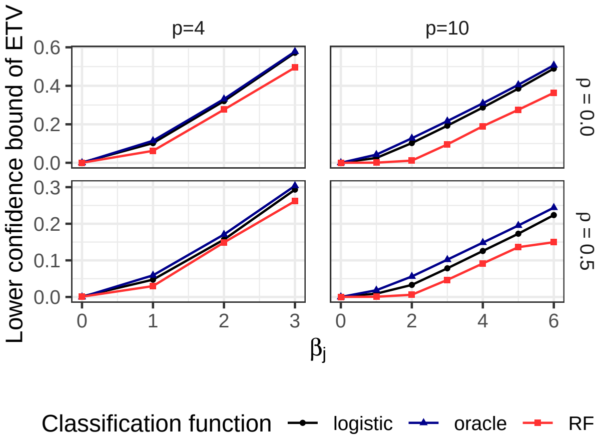

Figure 2. Lower confidence bound of ETV with different classification functions, averaged over 1536 independent experiments. In the $p = 10$ case where there are multiple β_j's equal to zero, we plot one case where $j = 1$.

3.2. Effect of Threshold c

In this section, we demonstrate the effect of c in (5). We consider a model M_k of the form

$$Y \mid X \sim \text{Bern}\left(\Phi\left(\beta_k X_k Z_k + \sum_{j \neq k} \beta_j X_j\right)\right), \tag{8}$$
$$X \sim \mathcal{N}(0, \Sigma), Z \sim \text{Bern}(0.5), X \perp\!\!\!\perp Z.$$

Here, X is a p-dimensional column vector and we provide lower confidence bound for $\text{ETV}(X_k, Y, (X_{\text{-}k}, Z))$ under M_k, where we rotate the value of k. We choose $\Sigma_{ij} = \rho^{|i-j|}$. We set $p = 10$, $\beta = (0,0,0,1,2,3,4,5,6,7)$ and $n = 220$. We focus on two classification functions as in (5) and the results are reported in Figure 3. For the "logistic" model, we use logistic models for p_{θ_1} and p_{θ_2}, and $\hat{\theta}_1$ and $\hat{\theta}_2$ are trained on cross-validated data; for the "logistic_int" model, we add interactions between Z and other X_j's into the models. We explore the following methods to choose c: "Naive" means setting $c = 0$; "CV" means using 10-fold cross validation to choose c. Note that this cross validation is different from one we use to train f in Algorithm 2. We further compare our methods with an oracle method described below. Consider the rth fold in Algorithm 2, where we have trained classifier f_{B_r} and the evaluation set B_r^c. We use $\hat{\mu}(f^c_{B_r}, D)$ and $\hat{\sigma}^2(f^c_{B_r}, D)$ to denote the sample

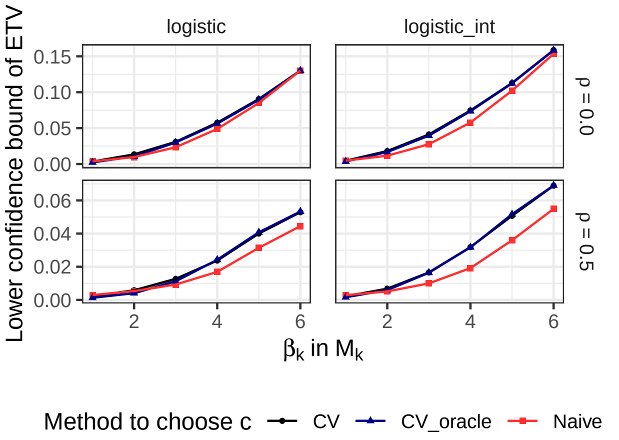

Figure 3. Lower confidence bound of ETV based on 1536 experiments. "logistic" and "logistic_int" denote the classification function. The $\beta_k = 0$ case is with $k = 1$.

mean and variance of applying Algorithm 1 with dataset D and $f^c_{B_r}$, where $f^c_{B_r}$ is combining c with f_{B_r} as in (5). "CV_oracle" means setting

$$c_{\text{oracle}} = \arg\max_c \frac{1}{K} \sum_{s=1}^{S} \Big(1 - \hat{\mu}(f^c_{B_r}, D_s) \\ - z_\alpha \hat{\sigma}(f^c_{B_r}, D_s) / \sqrt{|B_r \cup B^c_r|} \Big),$$

where $D_1, \ldots, D_S$ are S (taken to be 10 in the experiments) independent regenerations of the dataset B^c_r with the true distribution. The results are summarized in Figure 3. We can see that "Oracle" outperforms "CV" and "Naive", matching intuition. "CV" outperforms "Naive" and is quite close to the oracle method.

4. Application in Conjoint Analysis

4.1. Conjoint Analysis

Conjoint analysis (Luce & Tukey, 1964) is a survey-based statistical technique, where respondents are given a number of profiles with different attributes are asked to pick a favourite or rank them. A popular VIM used by social scientists is the average marginal component effect (AMCE), and there has been work on constructing confidence intervals on the AMCE (Hainmueller et al., 2014; Ono & Burden, 2019). The AMCE, as its name suggests, considers only the marginal effect and may fail to capture some interactions. Ham et al. (2022) introduced a hypothesis testing procedure for the null hypothesis $Y \perp\!\!\!\perp X \mid Z$ in the conjoint analysis context, but they did not propose a VIM. We will bridge this gap by using the ETV as the VIM in conjoint analysis and construct lower confidence bounds for it. The code is available at `https://github.com/wenshuow/etv_floodgate`.

4.2. US General Election Data

In this section, we analyze the election data in Ono & Burden (2019), which is under the CC0 license. In the experiment, each respondent is given two hypothetical political candidate profiles and asked to pick the one that they prefer. Each data point can thus be written in the form

$$(Y, X^0, X^1, Z^0, Z^1, Z^R),$$

where (X^k, Z^k) are the attributes of Candidate k with X being the attribute of interest, Z^R is the attribute of the respondent, and $Y \in \{0, 1\}$ is the choice of the respondent. We use Z to denote the collection of (Z^1, Z^2, Z^R). We focus on the presidential election data with $n = 7190$ observations. In each observation, there are 13 attributes of two political candidates and 11 attributes of the respondent, so X^0, X^1 are scalars, Z^0, Z^1 are 12-dimensional and Z^R is 11-dimensional. Here, each candidate's attributes are uniformly and independently randomized, with a few hard constraints; for example, a candidate with a high-skill profession must have at least two years of college experience. More details on the data can be found in Appendix D.1.

4.3. Inference for ETV

We choose X^0 and X^1 to be the party affiliations of the candidates, which take value from {**D**emocratic, **R**epublican}. We can see that while one would expect $X^{0,1}$ to play an important role in the respondent's choice Y, its marginal effect would be close to zero (assuming there is no party affiliation bias in the respondents). To use the AMCE, we would have to re-define $X^{0,1}$ as whether that candidate has the same party affiliation as the respondent. The ETV, on the other hand, can be employed directly. This issue could be more severe for other features that are not as straightforward to correct. For instance, the original analysis in Ono & Burden (2019) based on the AMCE suggested that gender is a statistically significant factor for congressional political candidates, while the analysis Ham et al. (2022) suggested that gender does matter for congressional candidates through interactions with other factors, including the respondent's party affiliation.

Before presenting our data analysis, we pause to consider what are reasonable values for the ETV. We have shown in Lemma 2.1 that in the case of binary response, the upper bound of ETV is 1. In our specific case, we should expect an even lower upper bound. Suppose we have the following ideal data generating distribution, where

$$\mathbb{P}(\text{candidate party affiliation is independent}) = q \in [0, 1],$$
$$X = (X^0, X^1) \mid Z \sim \text{Unif}\{(\text{D}, \text{D}), (\text{D}, \text{R}), (\text{R}, \text{D}), (\text{R}, \text{R})\},$$

and $\mathcal{L}(Y \mid X, Z)$ is given by Table 1.

Table 1. Ideal distribution by case.

Case	$\mathcal{L}(Y \mid X, Z)$
respondent is independent or two candidates have same party affiliation	$\text{Bern}(0.5)$
candidates' party affiliations differ and candidate 1 is same as respondent	$\text{Bern}(p)$
candidates' party affiliations differ and candidate 0 is same as respondent	$\text{Bern}(1-p)$

In this case, $\text{ETV}(X, Y, Z) = (1-q)|p-0.5|$. In the election data, $q \approx 0.27$, so even if $p = 1$, which means a respondent deterministically prefers the candidate from the same party, $\text{ETV}(X, Y, Z)$ is merely around 0.365, far from the general upper bound of 1. Simulations show that we are able to produce floodgate estimates close to the actual ETV with Algorithm 2. The derivation and supporting simulations are included in Appendix D.2.

Returning to the real data analysis, we apply Algorithm 2 with $k = 10$ and $J = 100$. The classifier family f is chosen to be the model-based f in equation (5), where the models are HierNet (Bien et al., 2013), following Ham et al. (2022, Section 3.3). We summarize our analysis in Figure 4. Each violin plot summaries 40 independent runs, with each run (or one fold of cross validation for "CV") taking less than 20 minutes on a single CPU. Here, we include both the floodgate lower bound and the floodgate estimate (that is, manually setting the confidence interval width to zero). We use the "Naive" and "CV" methods to choose c as in Section 3.2. We can see that activating c in f boosts performance, and we obtain an ETV estimate of around 0.1 and an ETV lower bound of around 0.08. Note that 0.1 is quite a high VIM, translating to around $p = 0.63$ in the model given in Table 1, though we do not assume that model. Further details are deferred to Appendix D.3. Our analysis shows that there is a strong presence of co-partisanship in the US, where voters prefer candidates of the same party, even after controlling for other candidate attributes. Co-partisanship is a well-documented phenomenon in the US (Campbell et al., 1980) and our finding corroborates existing research within the literature; for instance, in a different experiment, Peterson (2017) showed that even when the respondent is given the highest level of additional information, co-partisanship still increases the probability that a respondent selects a candidate by 0.29.

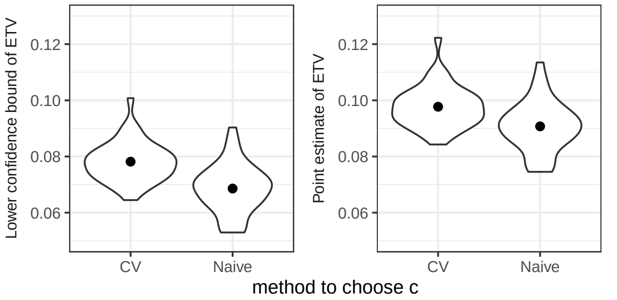

Figure 4. Conjoint analysis result of the US general election data. The black dots denote the mean of the violin plots.

Impact Statement

This paper's goal is to make classification methods more trustworthy and interpretable in a rigorous manner. The positive consequences include improved human ability to understand and interpret classifiers, and we do not foresee any negative consequences that require highlighting here.

Acknowledgements

L.J. was partially supported by a CAREER grant from the National Science Foundation (Grant #DMS2045981).

References

Agresti, A. *Foundations of linear and generalized linear models*. John Wiley & Sons, 2015.

Austern, M. and Zhou, W. Asymptotics of cross-validation. *arXiv preprint arXiv:2001.11111*, 2020.

Azadkia, M. and Chatterjee, S. A simple measure of conditional dependence. *The Annals of Statistics*, 49(6): 3070–3102, 2021.

Barber, R. F. Is distribution-free inference possible for binary regression? *Electronic Journal of Statistics*, 14 (2):3487 – 3524, 2020. doi: 10.1214/20-EJS1749. URL https://doi.org/10.1214/20-EJS1749.

Bayle, P., Bayle, A., Janson, L., and Mackey, L. Cross-validation confidence intervals for test error. In *Advances in Neural Information Processing Systems*, volume 33, pp. 16339–16350, 2020.

Belloni, A., Chernozhukov, V., and Wei, Y. Post-selection inference for generalized linear models with many controls. *Journal of Business & Economic Statistics*, 34(4): 606–619, 2016.

Bien, J., Taylor, J., and Tibshirani, R. A lasso for hierarchical interactions. *Annals of Statistics*, 41(3):1111, 2013.

Campbell, A., Converse, P., Miller, W., and Stokes, D. *The American Voter*. University of Chicago Press, 1980.

Candès, E., Fan, Y., Janson, L., and Lv, J. Panning for gold: Model-X knockoffs for high-dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B*, 80(3):551–577, 2018.

Castro, J., Gómez, D., and Tejada, J. Polynomial calculation of the shapley value based on sampling. *Computers & Operations Research*, 36(5):1726–1730, 2009.

Fisher, A., Rudin, C., and Dominici, F. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.*, 20(177):1–81, 2019.

Hainmueller, J., Hopkins, D. J., and Yamamoto, T. Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments. *Political analysis*, 22(1):1–30, 2014.

Ham, D. W., Imai, K., and Janson, L. Using machine learning to test causal hypotheses in conjoint analysis. *arXiv preprint arXiv:2201.08343*, 2022.

Holland, P. W. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

Huang, Z., Deb, N., and Sen, B. Kernel partial correlation coefficient–a measure of conditional dependence. *Journal of Machine Learning Research*, 23(216):1–58, 2020.

Janson, L. B. *A model-free approach to high-dimensional inference*. PhD thesis, Stanford University, 2017.

Luce, R. D. and Tukey, J. W. Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of mathematical psychology*, 1(1):1–27, 1964.

Molnar, C., König, G., Bischl, B., and Casalicchio, G. Model-agnostic feature importance and effects with dependent features: a conditional subgroup approach. *Data Mining and Knowledge Discovery*, pp. 1–39, 2023.

Ning, Y., Ong, M. E. H., Chakraborty, B., Goldstein, B. A., Ting, D. S. W., Vaughan, R., and Liu, N. Shapley variable importance cloud for interpretable machine learning. *Patterns*, 3(4):100452, 2022.

Näf, J., Michel, L., and Meinshausen, N. High probability lower bounds for the total variation distance. *arXiv:2005.06006*, 2022.

Ono, Y. and Burden, B. C. The contingent effects of candidate sex on voter choice. *Political Behavior*, 41:583–607, 2019.

Peterson, E. The role of the information environment in partisan voting. *The Journal of Politics*, 79(4):1191–1204, 2017.

Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

Shah, R. D. and Peters, J. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514 – 1538, 2020. doi: 10.1214/19-AOS1857. URL https://doi.org/10.1214/19-AOS1857.

Shapley, L. S. A value for n-person games. 1953.

Van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. On asymptotically optimal confidence regions and tests for high-dimensional models. 2014.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Watson, D. S. and Wright, M. N. Testing conditional independence in supervised learning algorithms. *Machine Learning*, 110(8):2107–2129, 2021.

Williamson, B. and Feng, J. Efficient nonparametric statistical inference on population feature importance using shapley values. In *International Conference on Machine Learning*, pp. 10282–10291. PMLR, 2020.

Williamson, B. D., Gilbert, P. B., Simon, N. R., and Carone, M. A general framework for inference on algorithm-agnostic variable importance. *Journal of the American Statistical Association*, pp. 1–14, 2021.

Zhang, L. and Janson, L. Floodgate: Inference for model-free variable importance. *arXiv preprint arXiv:2007.01283*, 2020.

A. Proofs

Proof of Lemma 2.1. Let $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ be spaces X, Y and Z live in; let p, q, r and s denote the densities of $\mathcal{L}(Z)$, $\mathcal{L}(X \mid Z)$, $\mathcal{L}(Y \mid X, Z)$ and $\mathcal{L}(Y \mid Z)$. We only prove the case where $|\mathcal{Y}| < \infty$, while the case $|\mathcal{Y}| = \infty$ can be treated similarly.

When $|\mathcal{Y}| < \infty$, We scale (1) as

$$\begin{aligned}
2(1 - 1/|\mathcal{Y}|)\,\mathrm{ETV}(X, Y, Z) &= \sum_{y\in\mathcal{Y}} \int_{x\in\mathcal{X}} \int_{z\in\mathcal{Z}} |r(y \mid x, z) - s(y \mid z)| q(x \mid z)\, dx\, p(z)\, dz \\
&= \sum_{y\in\mathcal{Y}} \int_{z\in\mathcal{Z}} E_{X|Z=z}\left[|r(y \mid X, z) - s(y \mid z)|\right] p(z)\, dz \\
&\le \sum_{y\in\mathcal{Y}} \int_{z\in\mathcal{Z}} 2s(y \mid z)(1 - s(y \mid z))p(z)\, dz \;\text{ (Lemma A.1)} \\
&= 2\int_{z\in\mathcal{Z}} \left[\sum_{y\in\mathcal{Y}} s(y \mid z)(1 - s(y \mid z))\right] p(z)\, dz \\
&\le 2\int_{z\in\mathcal{Z}} (1 - 1/|\mathcal{Y}|)p(z)\, dz = 2(1 - 1/|\mathcal{Y}|) \text{ (Lemma A.2)}.
\end{aligned}$$

Here, we are using two simple lemmas of which the proofs are omitted. The upper bound is achieved when X, conditional on Z, deterministically determines Y and $s(y \mid Z) = 1/|\mathcal{Y}|$ almost surely for all y. □

Lemma A.1. *If $X \in [0, 1]$, $E[X] = \mu$ and $\mathbb{P}(X = \mu) = p$, then $E[|X - \mu|] \le 2(1 - p)\mu(1 - \mu)$, where the equality is achieved when $X \mid (X \ne \mu) \sim \mathrm{Bern}(\mu)$.*

Lemma A.2. *Let $0 \le a_i \le 1$, $\sum_{i=1}^{n} a_i = 1$, then*

$$\sum_{i=1}^{n} a_i(1 - a_i) \le 1 - 1/n.$$

The equality is achieved when $a_i = 1/n$ for all i.

Proof of Theorem 2.2. Define $\ell(f) = E_\pi\left[\frac{1}{a}I(A = 1)(1 - f(\omega)) + \frac{1}{1-a}I(A = 0)f(\omega)\right]$. Then

$$\begin{aligned}
\ell(f) &= E_\pi\left[|f(\omega) - A|\left(\frac{1}{a}I(A = 1) + \frac{1}{1 - a}I(A = 0)\right)\right] \\
&= E[(1 - f(\omega))/a \mid A = 1]\mathbb{P}(A = 1) + E\left[f(\omega)/(1 - a) \mid A = 0\right]\mathbb{P}(A = 0) \\
&= \int (1 - f(\omega))\pi_1(\omega)\, d\omega + \int f(\omega)\pi_0(\omega)\, d\omega \\
&= 1 + \int f(\omega)(\pi_0(\omega) - \pi_1(\omega))\, d\omega\,.
\end{aligned}$$

The minimum of $\ell(f)$ is attained when

$$f(\omega) = f^*(\omega) = I(\pi_0(w) < \pi_1(\omega)).$$

It is not hard to see that $1 - \ell(f^*) = \mathrm{TV}(\pi_1, \pi_0)$. □

Proof of Corollary 2.3. Let $K \sim \mathrm{Unif}\{0, 1, \ldots, J\}$ independent of $(X^{(0:J)}, Y, Z)$. Then we apply Theorem 2.2 with $(X^{(K)}, Y, Z)$ as ω and $I(K = 0)$ as E to get

$$\begin{aligned}
1 - E\left[(J + 1)I(E = 1)(1 - f(X^{(K)}, Y, Z)) + \frac{J + 1}{J}I(E = 0)f(X^{(K)}, Y, Z)\right]& \\
\le \frac{1}{2}\int |p_{y|x,z}(y|x, z) - p_{y|z}(y|z)|p_{x|z}(x|z)p_z(z)\, dx\, dy\, dz\,.&
\end{aligned}$$

Evaluate the left hand side by integrating out K and we prove the claim. □

Proof of Theorem 2.5. The proof technique of this theorem is a generalization of the strategy used in the proof of Barber (2020, Lemma 1), which is itself a generalization of the construction used in the proof of Vovk et al. (2005, Proposition 5.1).

We fix $\mathcal{L}_X$ in the proof, so we omit the subscript $\mathcal{L}_X$ of C. We partition the sample space of X into NK equal-probability Borel sets $B_{1:NK}$, $N > n$, which is possible because $\mathcal{L}_X$ is a continuous distribution.

We are going to define data generating distributions $D_0, \ldots, D_5$ for $(X_i, Y_i)_{i=1}^n$, where D_0 is the distribution we care about, and we construct $D_{1:5}$ in a way such that $\mathrm{TV}(D_{i-1}, D_i)$ is small for $i = 1, \ldots, 5$. Our goal is to show (4) holds for D_5, so that it also has to hold for D_0. We use $\mathcal{L}(B)$ to denote the distribution $\mathcal{L}$ restricted to the set B.

- D_0: sample $(X_i, Y_i) \overset{\text{i.i.d.}}{\sim} \mathcal{L} = \mathcal{L}_X \times \text{Unif}(\{1, \ldots, K\})$;
- D_1: randomly sample n sets $\tilde{B}_{1:n}$ with replacement from $B_{1:NK}$; sample $Y_i \overset{\text{i.i.d.}}{\sim} \text{Unif}(\{1, \ldots, K\})$ and $X_i \mid \tilde{B}_{1:n} \sim \mathcal{L}_X(\tilde{B}_i)$ independently;
- D_2: randomly sample n sets $\tilde{B}_{1:n}$ without replacement from $B_{1:NK}$; sample $Y_i \overset{\text{i.i.d.}}{\sim} \text{Unif}(\{1, \ldots, K\})$ and $X_i \mid \tilde{B}_{1:n} \sim \mathcal{L}_x(\tilde{B}_i)$ independently;
- D_3: randomly permutate $B_{1:NK}$ to be $(\tilde{B}_{k,m})_{1\le k\le K, 1\le m\le N}$; sample $Y_i \overset{\text{i.i.d.}}{\sim} \text{Unif}(\{1, \ldots, K\})$, sample $I_i \overset{\text{i.i.d.}}{\sim} \text{Unif}(\{1, \ldots, N\})$ but resample until all the I_i's are distinct, and then sample $X_i \mid \tilde{B} \sim \mathcal{L}_x(\tilde{B}_{Y_i, I_i})$ independently;
- D_4: randomly permutate $B_{1:NK}$ to be $(\tilde{B}_{k,m})_{1\le k\le K, 1\le m\le N}$; sample $Y_i \overset{\text{i.i.d.}}{\sim} \text{Unif}(\{1, \ldots, K\})$, $I_i \overset{\text{i.i.d.}}{\sim} \text{Unif}(1 : N)$ and $X_i \mid \tilde{B} \sim \mathcal{L}_x(\tilde{B}_{Y_i, I_i})$ independently;
- D_5: randomly permutate $B_{1:NK}$ to be $(\tilde{B}_{k,m})_{1\le k\le K, 1\le m\le N}$; sample $Y_i \overset{\text{i.i.d.}}{\sim} \text{Unif}(\{1, \ldots, K\})$ and $X_i \mid Y_i, \tilde{B} \sim \mathcal{L}_X\left(\cup_{m=1}^N \tilde{B}_{Y_i, m}\right)$ independently;

By assumption, because D_5 is an i.i.d. data generating distribution for (X_i, Y_i) conditional on $\tilde{B}$ that respects the marginal distributions of X and Y,

$$\mathbb{P}_{D_5}(C(X_{1:n}, Y_{1:n}) \ge 1 \mid \tilde{B}) \ge 1 - \alpha,$$

where 1 is the attained ETV upper bound per the calculation in the proof of Lemma 2.1. After marginalizing out $\tilde{B}$, we have $\mathbb{P}_{D_5}(C(X_{1:n}, Y_{1:n}) \ge 1) \ge 1 - \alpha$.

We then notice that D_4 and D_5 are actually the same data generating distribution, so (4) holds under D_4 as well.

Now we examine the difference between D_3 and D_4. The probability of not having to resample is $N!/(N^n(N-n)!)$, so the total variation distance between D_3 and D_4 is upper bounded by $\epsilon(n, N) = 1 - N!/(N^n(N-n)!)$. Thus,

$$\mathbb{P}_{D_3}(C(X_{1:n}, Y_{1:n}) \ge 1) \ge 1 - \alpha - \varepsilon(n, N). \tag{9}$$

Next, we notice that D_2 and D_3 are also the same. This is because they both essentially use n random samples without replacement from $B_{1:NK}$. Therefore, (9) also holds for D_2.

Similarly, we can observe that the total variation distance between D_1 and D_2 is upper bounded by one minus the probability of all sampled sets $B_{1:n}$ in D_1 are distinct. This gives us the upper bound of $\varepsilon(n, NK)$. As a result, we get

$$\mathbb{P}_{D_1}(C(X_{1:n}, Y_{1:n}) \ge 1) \ge 1 - \alpha - \varepsilon(n, N) - \varepsilon(n, NK). \tag{10}$$

Finally, there is no difference between D_1 and D_0, so (10) also holds for D_0. Since $\epsilon(n, N) \to 0$ as $N \to \infty$, the fact that (10) holds for D_0 for any N means that (4) holds for D_0, as desired. □

B. ETV and Sensitivity Analysis

Let $B(X, Z, U)$ be the almost sure supremum of

$$\max\left\{\frac{p(X \mid U, Z)}{p(X \mid Z)}, \frac{p(X \mid Z)}{p(X \mid U, Z)}\right\}.$$

For notational convenience, we write $B(X, Z, U)$ as B. Then

$$\begin{aligned}
p(y \mid z, x) &= \int p(y \mid z, x, u) p(u \mid z, x)\, du \\
&= \int p(y \mid z, u) \frac{p(x \mid u, z) p(u \mid z)}{p(x \mid z)}\, du \\
&= \int \underbrace{\frac{p(x \mid u, z)}{p(x \mid z)}}_{\in[1/B, B]} \underbrace{p(y \mid z, u) p(u \mid z)}_{\text{integrates to } p(y \mid z)}\, du \in [p(y \mid z)/B, Bp(y \mid z)].
\end{aligned}$$

Then

$$\begin{aligned}
2(1 - 1/|\mathcal{Y}|)\,\mathrm{ETV}(X, Y, Z) &= \int |p(y \mid x, z) - p(y \mid z)|\, dy \cdot p(x \mid z)\, dx \cdot p(z)\, dz \\
&\leq \int \max(B - 1, 1 - 1/B) p(y \mid z)\, dy \cdot p(x \mid z)\, dx \cdot p(z)\, dz \\
&= \max(B - 1, 1 - 1/B) = B - 1.
\end{aligned}$$

Thus, $B \geq 1 + 2(1 - 1/|\mathcal{Y}|)\,\mathrm{ETV}(X, Y, Z)$.

C. Comparison with the MACM Gap

Continuing equation (6), the R_i in Zhang & Janson (2020, Algorithm 3) is equivalent to $1 - L_i$ in Algorithm 1. Note that

$$R_i = \begin{cases} \mathbb{P}(U_i < 0 \mid Z_i) - \mathbb{I}(U_i < 0), & \text{if } Y = 1, \\ \mathbb{P}(U_i > 0 \mid Z_i) - \mathbb{I}(U_i > 0), & \text{if } Y = -1. \end{cases}$$

and

$$\begin{aligned}
1 - L_i &= f(X_i, Y_i, Z_i) - \frac{1}{J} \sum_{j=1}^{J} f(X_i^{(j)}, Y_i^{(j)}, Z_i^{(j)}) \\
&= f(X_i, Y_i, Z_i) - \hat{E}_{X \mid Z = Z_i}[f(X, Y_i, Z_i)] \\
&= \begin{cases} \mathbb{I}(U_i \geq 0) - \hat{\mathbb{P}}_{X \mid Z = Z_i}(U_i \geq 0), & \text{if } Y_i = 1, \\ \mathbb{I}(U_i \leq 0) - \hat{\mathbb{P}}_{X \mid Z = Z_i}(U_i \leq 0), & \text{if } Y_i = -1. \end{cases}
\end{aligned}$$

D. Conjoint Analysis Further Details

D.1. Additional Details on Data

In this section, we include some additional details on the data used in Section 4. Table 2 includes attributes of the candidate profiles. Table 3 includes attributes of the respondents.

D.2. ETV Upper Bound

We derive the ETV upper bound in Section 4.3. First, we notice that due to the symmetry of labeling, $P(Y = 0 \mid Z = z) = P(Y = 1 \mid Z = z) = 0.5$ for any z. If z is such that the respondent's party affiliation is independent, then $P(Y = 0 \mid X^0 = x_0, X^1 = x_1, Z = z) = P(Y = 1 \mid X^0 = x_0, X^1 = x_1, Z = z) = 0.5$ for any (x_0, x_1); otherwise, $P(Y = 0 \mid X^0 = x_0, X^1 = x_1, Z = z)$ takes value $0.5, 0.5, p, 1 - p$ for $(x_0, x_1) \in \{(\mathrm{D}, \mathrm{D}), (\mathrm{D}, \mathrm{R}), (\mathrm{R}, \mathrm{D}), (\mathrm{R}, \mathrm{R})\}$. Then the ETV is

$$\begin{aligned}
\mathrm{ETV} = q \times 0 + (1 - q) &\sum_{y \in \{0,1\}} \sum_{x_0 \in \{\mathrm{R}, \mathrm{D}\}} \sum_{x_1 \in \{\mathrm{R}, \mathrm{D}\}} \frac{1}{4} |P(Y = y \mid X^0 = x_0, X^1 = x_1, Z = z) - P(Y = y \mid Z = z)| \\
&= (1 - q) \sum_{y \in \{0,1\}} \frac{1}{4} (0 + 0 + |p - 0.5| + |1 - p - 0.5|) \\
&= (1 - q) \times 2 \times \frac{1}{4} \times |2p - 1| = (1 - q)|p - 0.5|.
\end{aligned}$$

Attributes	**Values**
Sex	Male, Female
Age	36, 44, 52, 60, 68, 76
Race/Ethnicity	White, Black, Hispanic, Asian American
Family	Single (never married), Single (divorced), Married (no child), Married (two children)
Experience in public office	12 years, 8 years, 4 years, No experience
Salient personal characteristics	Provides strong leadership, Really cares about people like you, Honest, Knowledgeable, Compassionate, Intelligent
Party affiliation	Democrat Party, Republican Party
Policy area of expertise	Foreign policy, Public safety (crime), Economic policy, Health care, Education, Environmental issues
Position on national security	Wants to cut military budget and keep U.S. out of war, Wants to maintain strong defense and increase U.S. influence
Position on immigrants	Favors giving citizenship or guest worker status to undocumented immigrants, Opposes giving citizenship or guest worker status to undocumented immigrants
Position on abortion	Abortion is a private matter (pro-choice), Abortion is not a private matter (pro-life), No opinion (neutral)
Position on government deficit	Wants to reduce the deficit through tax increase, Wants to reduce the deficit through spending cuts, Does not want to reduce the deficit now
Favorability rating among public	34%, 43%, 52%, 61%, 70%

Table 2. Types of attributes varied in candidate profiles (Table 1 in Ono & Burden (2019)).

To test how well our algorithm does in this ideal setting, We regenerate synthetic Y according to Table 1 and apply Algorithm 2. In Figure 5, we plot the average floodgate estimate of ETV from 40 independent experiments (but they share the same synthetic response) and the true value of ETV, which is the theoretical upper bound $(1-q)|p-0.5|$. We can see that in moderate to high signal regimes, the floodgate estimate is close to the true value.

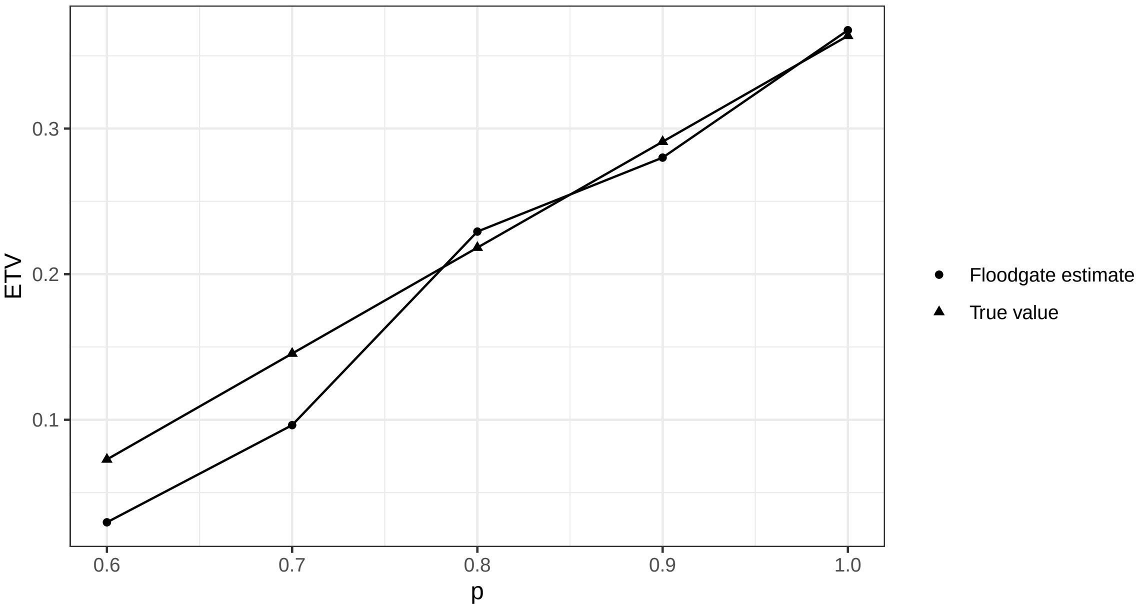

Figure 5. ETV floodgate estimate vs true value in conjoint analysis with synthetic responses.

Attributes	**Values**
Sex	Male, Female
Education level	BA degree, No BA degree
Age group	18-29, 30-50, 51-65, 66 or older
Age	Age in years
Social class	Lower class, Middle class, Upper class
Region	South, Nonsouth
Race/Ethnicity	White, Black, Hispanic, Other
Partisanship	Democrat Party, Republican Party, Independent
Thought on Hillary Clinton	Dislike, Like, Neutral
Interest in politics	Not at all interested, Not very interested, Somewhat interested, Very interested
Political ideology	Conservative or liberal levels (7 levels)

Table 3. Types of attributes recorded in respondents.

D.3. Analysis Details

In the experiments in Section 4.3, f is chosen to be

$$f(x,y,z;p_{\theta_1},p_{\theta_2},c)=\begin{cases}1, & \hat{p}_i>0.5+c,\\ 0, & \hat{p}_i<0.5-c\\ 0.5, & |\hat{p}_i-0.5|\le c,\end{cases}$$

where $p_{\theta_1}(y \mid x,z)$ is a HierNet model with a fixed penalty parameter, where interactions between politician's gender and party affliation are added as a feature, and $p_{\theta_2}(y \mid z)$ is a HierNet model with the same penalty parameter. In the method "CV" to choose c, we further partition the training data B_r into $m=10$ folds $C^c_{r1},\dots,C^c_{rm}$. We then calculate the cross-validated loss

$$\text{loss}_r(c)=\frac{1}{m}\sum_{j=1}^{m}\text{loss of } f(p_{\hat{\theta}_1(C_{rj})},p_{\hat{\theta}_2(C_{rj})},c)\text{ on } B_r\setminus C_{rj},$$

where $\hat{\theta}(C)$ means $\hat{\theta}$ estimated on dataset C, choose the c_r that minimizes $\text{loss}_r(c)$, and let

$$f_{B_r}=f(p_{\hat{\theta}_1(B_r)},p_{\hat{\theta}_2(B_r)},c_r).$$