Statistica Sinica Preprint No: SS-2025-0337	
Title	Comparing Model-agnostic Feature Selection Methods Through Relative Efficiency
Manuscript ID	SS-2025-0337
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0337
Complete List of Authors	Chenghui Zheng and Garvesh Raskutti
Corresponding Authors	Chenghui Zheng
E-mails	chenghui.zheng@wisc.edu
Notice: Accepted author version.	

Comparing Model-agnostic Feature Selection Methods through Relative Efficiency

Chenghui Zheng[1], Garvesh Raskutti[1]

[1] *University of Wisconsin - Madison*

Abstract: Feature selection and importance estimation in a model-agnostic setting is an ongoing challenge of significant interest. Wrapper methods are commonly used because they are typically model-agnostic. In this paper, we develop a general comparison framework for model-agnostic feature selection methods based on relative efficiency, using *relative variability* σ/μ to account for different statistics having different means. In particular we focus on state-of-the-art feature selection methods, the Generalized Covariance Measure (GCM) and Leave-One-Covariate-Out (LOCO) estimation. In particular, we present a theoretical comparison under three model settings: linear models, non-linear additive models, and single index models that mimic a single-layer neural network. We complement this with simulations and real data examples for the above models and mis-specified models. Our theoretical results, along with empirical findings, demonstrate that GCM-related methods generally out-perform LOCO under suitable regularity conditions defined by a suitably defined correlation quantity which quantifies the asymptotic relative efficiency of these approaches. Our simulations and real data analysis include widely used machine learning methods such as neural networks and gradient boosting trees.

Key words and phrases: conditional independence; feature selection; generalized covariance measure; leave-one-covariate-out; model-agnostic methods; variable importance.

1. Introduction

Feature selection is one of the fundamental challenges in statistics and machine learning. While embedded methods such as Lasso have been widely used under parametric assump-

tions, there is a growing trend toward using "black-box" prediction methods such as random forests, neural networks, and others. In such settings, wrapper methods for feature selection which are typically *model-agnostic* have become increasingly popular.

A natural question to consider is how we provide a theoretical comparison for model-agnostic feature selection methods. In this paper, we use relative efficiency based on the *relative variability* σ/μ statistic to account for the fact that different estimators have different means. In particular we focus on two state-of-the-art wrapper methods: Leave-One-Covariate-Out (LOCO) (Lei et al. (2018)) and the Generalized Covariance Measure (GCM) (Shah and Peters (2020)). LOCO may be viewed as a generalization of ANOVA-based methods that measure the change in prediction error when a feature is removed (Fisher (1992); Duda et al. (2000)). GCM, on the other hand, generalizes traditional correlation-based tests (Spearman (1904); Pearson and Galton (1997); Székely et al. (2007)) to conditional correlation, and can capture general functional relationships. While there has been prior work analyzing each of these methods and establishing asymptotic normality (Shah and Peters (2020); Williamson et al. (2022)), there is no quantitative comparison of their performance in feature selection that we are aware of.

Concretely, consider a prediction model:

$$E[Y \mid X_1, X_2, \ldots, X_p] = f(X_1, X_2, \ldots, X_p)$$

for some unknown function f, where Y is the response and $(X_1, X_2, \ldots, X_p)$ are the covariates. The goal of feature selection is to identify which covariates significantly affect Y. This enhances model interpretability, reduces overfitting, and can improve predictive performance.

In this paper, we compare GCM and LOCO with respect to statistical inference and

practical applications under three statistical models: linear models, non-linear additive models and single-index models. We also include other model-agnostic variable importance methods such as: the *dropout method*, which replaces a feature with its marginal mean to avoid retraining; *Lazy-VI* (Gao et al. (2022)), which approximates the reduced model for efficiency while maintaining inferential guarantees.

The remainder of the paper is organized as follows. Section 2 introduces variable importance tests and asymptotic relative efficiency test for the comparison between feature selection methods. Section 3 compares LOCO, and GCM under linear, additive, and single index models. Section 4 extends the comparison to dropout and Lazy-VI. Sections 5 and 6 present simulation studies and real data experiments, respectively. We conclude in Section 7.

1.1 Related work

Feature selection techniques fall into three main categories: filter-based, wrapper-based, and embedded methods. *Filter methods* involve calculating a relevance score for each feature independently of any model and rank features accordingly. They are computationally efficient but due to their simplicity often don't capture conditional relationships. Examples include mutual information (Hoque et al. (2014)), information gain, Pearson correlation (Pearson and Galton (1997); Yu and Liu (2003)), the Chi-squared test (Witten et al. (2005)), and the Fisher score (Duda et al. (2000)). *Wrapper methods* involve evaluating subsets of features by training models and measuring performance, often using criteria like AIC, BIC, or Cp. Common algorithms include forward selection, backward elimination (Kittler (1978)), and recursive feature elimination (Guyon et al. (2002)). While they account for feature interactions, they are computationally expensive and prone to overfitting due to repeated training.

Embedded methods perform feature selection during model training. Examples include Lasso (Ma and Huang (2008)), elastic net (Zou and Hastie (2005)), decision trees, random forests (Sandri and Zuccolotto (2006)), and naive Bayes (Cortizo and Giraldez (2006)). Embedded methods are generally more efficient than wrappers but are tied to specific model classes.

Wrapper-style, model-agnostic feature selection has grown in popularity due to their flexibility and the increased use of so-called black-box methods. Among global model-agnostic approaches, permutation importance (Breiman (2001)) is the most widely used. It disrupts a feature's relationship with the response by permuting values and measuring the effect on predictive accuracy (Strobl et al. (2008); Fisher et al. (2019); Mentch and Hooker (2016); Altmann et al. (2010)). However, these methods introduce randomness, which can lead to instability. The knockoff framework (Barber and Candès (2015); Candès et al. (2018)) uses conditional randomization to control false discovery by generating synthetic "null" copies of features. While powerful, this approach relies on the assumption of feature exchangeability, which is difficult to verify in practice. Furthermore, the theoretical results focus on false discovery rate (FDR) rather than the mean and variance which is the focus of this paper. Replacing rather than permuting a feature, e.g., the dropout method, offers computational efficiency but may underfit (Chang et al. (2018)). LOCO avoids this issue by removing a feature and retraining, and has been generalized for conformal inference (Lei et al. (2018)) and bootstrapping in high-dimensional settings (Rinaldo et al. (2019)).

Conditional Independence Tests Feature selection directly relates to *conditional independence testing*. In particular, any feature X_j is informative to Y if and only if:

$$Y \not\perp X_j | X_1, X_2, ..., X_{j-1}, X_{j+1}, X_{j+2}, ..., X_p$$

or Y is not independent of X_j given all other variables/features. Classical tools like partial correlations rely on strong parametric assumptions. More flexible tests include Conditional randomization (Candès et al. (2018), Holdout randomization (Tansey et al. (2022)), Kernel-based (Zhang et al. (2012)) and Mutual information-based (Fleuret (2004)). However, many of these methods are computationally expensive and rely on estimating conditional distributions. Generalized Covariance Measure (GCM) (Shah and Peters (2020)), which tests for vanishing normalized covariance between residuals, is a regression-based test that requires only mild assumptions and works well in high dimensions and nonlinear settings.

This paper focuses on analyzing wrapper-based feature selection using LOCO and GCM. All methods yield population-level variable importance scores and support hypothesis testing. We provide theoretical and empirical comparisons across various settings.

1.2 Our contributions

This paper includes the following main contributions:

- Provides an asymptotic relative efficiency comparison for GCM and LOCO in 3 different model settings: linear models, additive models and single index models. In particular we prove that GCM is asymptotically more efficient compared to LOCO for the linear model case, provide conditions under which GCM is more asymptotically efficient than LOCO for non-linear models, and single-index models. We further provide a quantity related to an adjusted correlation under which GCM has superior relative efficiency to LOCO.

- We also provide a comparison of asymptotic variance between LOCO and more computationally efficient dropout and lazy fine-tuning based estimator (Gao et al. (2022)).

- Finally, we conduct a comparison of these different approaches along with other state-of-the-art approaches over a series of simulated and real data with different algorithms including neural networks and tree-based algorithms. The simulation results support the theory and show that GCM is often more accurate although computationally more intensive than other methods.

2. Feature selection method comparison

2.1 Asymptotic relative efficiency and relative variability

We begin by discussing how we provide a theoretical comparison for feature selection methods. Since model-agnostic feature selection methods are often difficult to characterize and there are a number of different comparison metric choices, this is a non-trivial issue. In our simulations and real data section, we provide many different comparison metrics.

For the purpose of this paper, we focus on the natural quantity *relative efficiency* for comparing two methods. The next natural question is which statistic we use for the efficiency comparison. When the means of two feature importance estimators are the same, efficiency simply equates to the variance. However, for feature importance statistics across different methods, the variance alone may provide a misleading notion of statistical efficiency, since different methods can produce importance scores on substantially different scales. A statistic with a larger variance may nevertheless exhibit substantially greater discriminatory power if its expected importance under the alternative is correspondingly larger.

Classical statistics addresses this issue through scale-invariant measures such as the coefficient of variation, which normalizes variability by the magnitude of the mean (Vangel (1996)). In the feature selection setting however, the relevant reference point is the expecta-

tion under the null hypothesis H_0 rather than the absolute mean. This motivates measuring variability relative to the separation between the null and alternative expectations rather than relative to the mean itself. Specifically, let $\mu_0 = \mathrm{E}[T \mid H_0]$ and $\mu_1 = \mathrm{E}[T \mid H_1]$. We define the *relative variability* of a feature importance statistic as

$$R(T) = \frac{\sqrt{\mathrm{Var}(T \mid H_1)}}{|\mu_1 - \mu_0|}.$$

Equivalently, its reciprocal

$$S(T) = \frac{|\mu_1 - \mu_0|}{\sqrt{\mathrm{Var}(T \mid H_1)}}$$

is a standardized signal-to-noise ratio measuring the separation between the null and alternative distributions.

This criterion is closely related to classical standardized effect sizes (Cohen (1988)) and Pitman efficacy, which characterizes the local asymptotic power of statistical procedures through the ratio of the derivative of the mean to the asymptotic standard deviation (Lehmann and Romano (2005); van der Vaart (2000)). Consequently, comparing feature importance statistics through $R(T)$ provides a scale-invariant notion of statistical efficiency that directly reflects their ability to distinguish null from informative features.

Definition 1 (Relative variability)**.** Let T be a feature importance statistic with

$$\mu_0 = \mathrm{E}[T \mid H_0], \qquad \mu_1 = \mathrm{E}[T \mid H_1].$$

We define the *relative variability* of T as

$$\mathrm{RV}(T) = \frac{\sqrt{\mathrm{Var}(T \mid H_1)}}{|\mu_1 - \mu_0|}.$$

This quantity measures the stochastic variability of the statistic relative to its discriminative signal, namely the separation between its null and alternative expectations.

In the context of feature selection considered in this paper, the null hypothesis corresponds to zero importance, so that

$$\mu_0 = \mathrm{E}[T \mid H_0] = 0.$$

In this case, the relative variability simplifies to

$$\mathrm{RV}(T) = \frac{\sqrt{\mathrm{Var}(T \mid H_1)}}{|\mathrm{E}[T \mid H_1]|}.$$

Thus, the relative variability is simply the ratio of the standard deviation to the expected importance under the alternative. This naturally leads to the definition of asymptotic relative efficiency when comparing two estimators T_1 and T_2.

Definition 2 (Asymptotic relative efficiency)**.** Let $T_1^{(n)}, T_2^{(n)}$ be two estimators, then the *asymptotic relative efficiency* (ARE) e_{T_1,T_2} based on the relative variability $R(T)$:

$$e_{T_1,T_2} := \lim_{n\to\infty} \frac{\mathrm{Var}(T_2^{(n)})/\mathrm{E}^2[T_2^{(n)}]}{\mathrm{Var}(T_1^{(n)})/\mathrm{E}^2[T_1^{(n)}]}.$$

Next we define the two estimators we focus on in this paper, GCM and LOCO.

2.2 Generalized Covariance Measure

GCM is a regression-based conditional independence test. Let's review the setup in Shah and Peters (2020) and connect it to variable selection. Let $(U, V, W) \in \mathbb{R}^{d_U} \times \mathbb{R}^{d_V} \times \mathbb{R}^{d_W}$ with joint distribution P_0, where $d_U = d_V = 1$ and $d_W \geq 1$. Define:

$$U = f_0(W) + \xi \quad ; \quad V = g_0(W) + \epsilon$$

where $f_0(W) = E(U|W)$, $g_0(W) = E(V|W)$, $\xi = U - f_0(W)$, and $\epsilon = V - g_0(W)$. The generalized covariance between residuals $E[Cov(U, V|W)]$ is defined as:

$$\psi^{(gcm)}_{0,(U,V)} := E[\xi * \epsilon]$$

Given observations $\{(U_i, V_i, W_i)\}_{i=1}^n$, let $\hat{f}$ and $\hat{g}$ be estimates of conditional expectations f_0 and g_0, respectively. The empirical GCM estimator is a normalized version of:

$$\hat{\psi}^{(gcm)}_{(U,V)} = \frac{1}{n}\sum_{i=1}^{n}\{U_i - \hat{f}(W_i)\}\{V_i - \hat{g}(W_i)\},$$

which estimates a population residual covariance that is zero under $U \perp V \mid W$.

Theorem 1 (Re-formulated from Shah and Peters (2020))**.** *Under suitable assumptions defined in Shah and Peters (2020), the GCM estimator* $\hat{\psi}^{(gcm)}_{(U,V)}$ *has the following result:*

$$\sqrt{n}\big[\hat{\psi}^{(gcm)}_{(U,V)} - \psi^{(gcm)}_{(U,V)}\big] \xrightarrow{d} N(0, \sigma^2_{(U,V)}),$$

where $\sigma^2_{(U,V)} = Var(\xi\epsilon)$.

To see how we apply to the feature selection setting, let's assume we have samples $Z_i = (X_i, Y_i)$, $i \in [n]$ drawn from data matrices $\boldsymbol{X}^{(n)} \in \mathbb{R}^{n \times p}$ and $\boldsymbol{Y}^{(n)} \in \mathbb{R}^n$. Those observations are drawn from a data-generating distribution P_0. X (resp. X_i) denotes the multivariate random variable containing all features, Y (resp. Y_i) denotes the response. For feature index $j \in [p]$ where $[p] := \{1, ..., p\}$, let X_j (resp. X_{ij}) denote the j^{th} feature in X and $X_{-j} \in \mathbb{R}^{p-1}$ (resp. $X_{i,-j}$) denote the features in X with the j^{th} variable removed. In particular, let (U, V, W) be (X_j, Y, X_{-j}) applied to the GCM, for all $j \in [p]$. For notation simplicity we use $\psi_j^{(gcm)}$ to represent $\psi_{(X_j,Y)}^{(gcm)}$, then we can test the hypotheses

$$H_0 : \psi_j^{(gcm)} = 0 \text{ versus } H_a : \psi_j^{(gcm)} \neq 0$$

We expect the p-value to be small if Y is not conditional independent of X_j given X_{-j}.

GCM limitation One of the well-known limitations of GCM is that a covariance of 0 does not necessarily imply independence which is especially relevant for non-linear relationship. Consider the simple example

$$Y = X^2 + \epsilon$$

where X is a symmetric distribution (e.g.normal) and ϵ is zero-mean independent noise. It follows that $\text{Cov}(X^2, Y) = 0$ and the GCM statistic satisfies $\psi_j^{(gcm)} = 0$ when clearly X and Y are dependent. This extends to any scenario where the functional relationship is an even function and the random variable has a symmetric distribution. This is a clear limitation of GCM that will be observed in both the theoretical and empirical results.

2.3 Leave-one-covariate-out

Before outlining variable importance for feature selection, we review the model-agnostic population framework of Williamson et al. (2022), where each feature importance score is measured by the change in predictive accuracy after omitting or replacing the feature.

Let's denote $s \subset \{1, ..., p\}$ as an index set of the covaraite subgroup of interest. Then, $X_{-s} \in \mathbb{R}^{p-|s|}$ under population $P_{0,-s}$ is defined as the features in X with index in s features removed. Let $V(f, P_0)$ be a measure of the predictive power of a prediction function f of the model, where $f \in \mathcal{F}$, $\mathcal{F}$ is a rich function class. A natural candidate prediction function would be a population maximizer f_0 over the class $\mathcal{F}$:

$$f_0 \in \text{argmax}_{f \in \mathcal{F}} V(f, P_0).$$

Similarly, define prediction function $f_{0,-s}$ be the maximizer over function class $\mathcal{F}_{-s}$ where $\mathcal{F}_{-s}$ is a collection of function $f \in \mathcal{F}$ whose evaluation ignores entries of input X with index in s. In the population setting, for the subset of features X_s, importance is estimated by the difference in the predictive power by excluding s features from X. Thus, the feature importance score of the subgroup X_s is defined as:

$$\psi_{0,s} := V(f_0, P_0) - V(f_{0,-s}, P_{0,-s}).$$

Throughout this paper we use a negative mean-squared error predictive power:

$$V(f_0, P_0) = -\text{E}[(Y - f_0(X_1, X_2, ..., X_p))^2].$$

For feature selection, we focus on LOCO (Lei et al. (2018)), which evaluates one feature at a time by setting $s = \{j\}$ and comparing the full model using X with the reduced model using X_{-j}. A feature is considered important if removing it significantly worsens predictive performance. Define the respective error term as

$$\epsilon := Y - f_0(X) \quad ; \quad \epsilon_{-j} := Y - f_{0,-j}(X_{-j})$$

where $f_0(X) = \mathrm{E}[Y|X]$ and $f_{0,-j}(X_{-j}) = \mathrm{E}[Y|X_{-j}]$. We can adapt Theorem from Williamson et al. (2022) to LOCO estimator:

Theorem 2 (Re-formulated from Williamson et al. (2022))**.** *Under suitable assumptions defined in Williamson et al. (2022), the LOCO estimator* $\hat{\psi}_j^{(loco)} := V(f_n, P_n) - V(f_{n,-j}, P_{n,-j})$ *has the following result:*

$$\sqrt{n}\big[\hat{\psi}_j^{(loco)} - \psi_j^{(loco)}\big] \xrightarrow{d} N(0, \sigma_j^2)$$

where $\widehat{\psi}_j^{(loco)} = \frac{1}{n}\sum_{i=1}^{n}\{[Y_i - f_n(X_{ij})]^2 - [Y_i - f_{n,-j}(X_{i,-j})]^2\}$, *and* $\sigma_j^2 = Var(\epsilon^2 - \epsilon_{-j}^2)$.

Thus, for all $j \in [p]$, we can test the hypotheses

$$H_0 : \psi_{0,j}^{(loco)} = 0 \text{ versus } H_a : \psi_{0,j}^{(loco)} \neq 0.$$

We expect the corresponding p-value to be small if X_j provides additional information for predicting Y not already captured by X_{-j}.

The focus of this paper on the ARE comparison

$$e_{\hat{\psi}_j^{(gcm)}, \hat{\psi}_j^{(loco)}} := \lim_{n\to\infty} \frac{\mathrm{Var}(\widehat{\psi}_j^{(loco)})/\mathrm{E}^2[\widehat{\psi}_j^{(loco)}]}{\mathrm{Var}(\widehat{\psi}_j^{(gcm)})/\mathrm{E}^2[\widehat{\psi}_j^{(gcm)}]}.$$

for these two model-agnostic methods GCM and LOCO. The ARE comparison would apply to any model-agnostic feature importance estimator where the asymptotic mean and variance can be computed. In the simulations and real data comparison, we consider a broader set of methods.

3. Main results: ARE comparison

In this section, we compare and evaluate the asymptotic relative efficiency of GCM and LOCO for linear, non-linear additive and single index model. We focus on these models because these models allow us to provide theoretical guarantees and they provide a reasonable class of data-generating models. We consider mis-specified models in the simulations section.

Throughout we define

$$\tilde{X}_j := X_j - E[X_j|X_{-j}]$$

which denotes the residual based on predicting X_j from X_{-j}. Note that if X_j is independent of X_{-j}, then $\tilde{X}_j = X_j - E[X_j]$.

3.1 Linear Model

Let's consider $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$ with joint distribution P_0 satisfying the linear model

$$Y = X_{-j}^T \beta_{-j} + \beta_j X_j + \epsilon$$

where $\beta_j \in \mathbb{R}$ and $\beta_{-j} \in \mathbb{R}^{p-1}$ are regression coefficients, and ϵ is a random noise term with zero mean and finite variance σ^2.

To select the most important features in the linear model, we are perform individual

hypothesis tests: for all $j \in [p]$, $H_0 : \beta_j = 0$ which is equivalent to the test: X_j is dependent of the target Y given the rest features X_{-j}, and LOCO test: if X_j provide additional model prediction power when the rest features are already included.

Theorem 3. *For a linear model, where* $Y = X_{-j}^T \beta_{-j} + \beta_j X_j + \epsilon$ *and* ϵ *is random noise with finite variance* σ^2*. For all* $j \in [p]$*, we have*

$$\frac{sd(\hat{\psi}_j^{(gcm)})}{E(\hat{\psi}_j^{(gcm)})} = \frac{\frac{1}{\sqrt{n}}\beta_j \sqrt{Var(\tilde{X}_j^2) + \frac{1}{\beta_j^2}\sigma^2 E(\tilde{X}_j^2)}}{\beta_j E(\tilde{X}_j^2)},$$
$$\frac{sd(\hat{\psi}_j^{(loco)})}{E(\hat{\psi}_j^{(loco)})} = \frac{\frac{1}{\sqrt{n}}\beta_j^2 \sqrt{Var(\tilde{X}_j^2) + \frac{4}{\beta_j^2}\sigma^2 E(\tilde{X}_j^2)}}{\beta_j^2 E(\tilde{X}_j^2)},$$

and thus $e_{\hat{\psi}_j^{(gcm)}, \hat{\psi}_j^{(loco)}} > 1$.

The expressions for the ratio of standard deviation and expectation for each estimator is instructive in that they are equivalent once dividing numerator and denominator by β_j aside from an additional factor of 4 in the standard deviation for the LOCO estimator. This displays a clear increase in efficiency for the GCM estimator. As far as we are aware this is the first direct comparison between the two, even for a linear model. And note that there are no assumptions of dependence/correlation so the result holds regardless of the dependence/correlation structure.

The key quantities that determine how much the efficiency is larger for GCM compared to LOCO is depends on $\frac{\sigma^2}{\beta_j^2}$ as well as $E(\tilde{X}_j^2)$ and $Var(\tilde{X}_j^2)$ where $\tilde{X}_j = X_j - E[X_j|X_{-j}]$. In particular if $\frac{\sigma^2}{\beta_j^2}E(\tilde{X}_j^2)$ is large relative to $Var(\tilde{X}_j^2)$ then the gap in efficiency between GCM and LOCO increases.

Below is a simple two-variable example illustrating the efficiency ratio.

Example 3.1. Suppose $Y = \beta_0+\beta_1 X_1+\beta_2 X_2+\epsilon$ where $X_i \sim N(0,\sigma^2), i = 1, 2, Cov(X_1, X_2) = \rho$, and $\epsilon \sim N(0,\sigma_\epsilon^2)$ is independent of features. Also, known that $X_1|X_2 \sim N(\rho X_2, (1-\rho^2)\sigma^2)$, then the asymptotic relative efficiency of LOCO and GCM estimators for the testing of X_1 is

$$e_{\hat{\psi}_1^{(gcm)},\hat{\psi}_1^{(loco)}} = \frac{4\frac{\sigma_\epsilon^2}{\beta_1^2} + 2(1-\rho^2)\sigma^2}{\frac{\sigma_\epsilon^2}{\beta_1^2} + 2(1-\rho^2)\sigma^2} > 1.$$

Specifically, if $\frac{\sigma_\epsilon^2}{\beta_1^2} \to 0$, then $e_{\hat{\psi}_1^{(gcm)},\hat{\psi}_1^{(loco)}} \to 1$.

3.2 Non-linear Additive Model

Although the linear model is instructive, the primary focus of this paper is on non-linear additive regression settings. Let

$$Y = \sum_{j=1}^{p} f_{0,j}(X_j) + \epsilon$$

where $f_{0,j}$'s are non-linear and measurable, and there are no interaction terms among features, ϵ are i.i.d mean zero with finite variance σ^2, and $\epsilon \perp X$. For each $j \in [p]$, both GCM and LOCO are performing the following hypothesis test:

$$H_0 : \psi_j^{(\cdot)} = 0 \quad ; \quad H_a : \psi_j^{(\cdot)} \neq 0.$$

For regressing X_j and Y on X_{-j} respectively for GCM, let $g_0, h_0 \in \mathcal{F}$ for function class $\mathcal{F}$,

$$X_j = h_{0,j}(X_{-j}) + \xi_{-j} \quad ; \quad Y = g_{0,j}(X_{-j}) + \epsilon_{-j}$$

where ϵ_{-j}, ξ_{-j} are i.i.d mean zero with finite variance $\sigma_\epsilon^2, \sigma_\xi^2$ respectively, $\epsilon_{-j}, \xi_{-j} \perp X$. Let $h_{0,j}(X_{-j}) = E(X_j|X_{-j}), g_{0,j}(X_{-j}) = E(Y|X_{-j})$, recall $\tilde{X}_j := X_j - E(X_j|X_{-j})$, and $\tilde{f}_{0,j}(X_j) := f_{0,j}(X_j) - E(f_{0,j}(X_j)|X_{-j})$. We imposing the following assumption on $\text{Corr}(\tilde{X}_j, \tilde{f}_{0,j}(X_j))$:

Assumption (A). $\text{Corr}^2(\tilde{X}_j, \tilde{f}_{0,j}(X_j)) > \frac{E(\tilde{X}_j^2 \tilde{f}_{0,j}^2(X_j)) \frac{E(\tilde{f}_{0,j}^2(X_j))}{E(\tilde{X}_j^2)} + \sigma^2 E(\tilde{f}_{0,j}^2(X_j))}{E(\tilde{f}_{0,j}^4(X_j)) + 4\sigma^2 E(\tilde{f}_{0,j}^2(X_j))}$

Assumption (A) is an adjusted correlation condition that ensures the efficiency of GCM under ARE comparison. Note that Assumption (A) fails for the example where if X_j is drawn from a symmetric distribution and $f_{0,j}(.)$ is an even function since Corr $(\tilde{X}_j, \tilde{f}_{0,j}(X_j)) = 0$. The assumption therefore ensures that the correlation between $\tilde{X}_j$ and $\tilde{f}_{0,j}(X_j)$ is sufficiently far away from 0. Also note that in the case of the linear model where $f_{0,j}(X_j) = \beta_j X_j$ Assumption (A) is always satisfied which yields the linear model result from earlier.

Theorem 4. *For an additive regression model estimated with mean squared error loss, $Y = f_0(X) + \epsilon = f_{0,-j}(X_{-j}) + f_{0,j}(X_j) + \epsilon$, where ϵ are i.i.d ransom noise with zero mean and finite variance σ^2. For all $j \in [p]$, we have*

$$\frac{sd(\hat{\psi}_j^{(gcm)})}{E(\hat{\psi}_j^{(gcm)})} = \frac{\frac{1}{\sqrt{n}}\sqrt{Var(\tilde{X}_j \tilde{f}_{0,j}(X_j)) + \sigma^2 E(\tilde{X}_j^2)}}{E(\tilde{X}_j \tilde{f}_{0,j}(X_j))},$$
$$\frac{sd(\hat{\psi}_j^{(loco)})}{E(\hat{\psi}_j^{(loco)})} = \frac{\frac{1}{\sqrt{n}}\sqrt{Var(\tilde{f}_{0,j}^2(X_j)) + 4\sigma^2 E(\tilde{f}_{0,j}^2(X_j))}}{E(\tilde{f}_{0,j}^2(X_j))}.$$

Therefore under Assumption (A), $e_{\hat{\psi}_j^{(gcm)}, \hat{\psi}_j^{(loco)}} > 1$.

Next, we present a nonlinear additive regression functions as concrete example to illustrate that the GCM estimator is (approximately) asymptotically more efficient than the LOCO estimator when Assumption (A) in Theorem 4 is satisfied, and vice versa. The corresponding simulation study will also be shown in Section 5, and another further additive

model comparison example can be found in Supplementary Material S1.

Example 3.2. Let $Y = \sum_{j=1}^{q} f_j(X_j) + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$, and $f_j(X_j) = X_j^{p_j}$ where p_j is odd, and suppose $X_j \sim N(0,1)$, and X_j's are independent. Then $\tilde{X}_j = X_j - E(X_j|X_{-j}) = X_j$. For notation simplicity, we use $\tilde{f}_j$ to represent $\tilde{f}_j(X_j)$, then $\tilde{f}_j = X_j^p - E(X_j^p|X_{-j}) = X_j^{p_j}$.

By the property of the moment for standard normal that $E(X^{2t}) = \frac{(2t)!}{2^t t!}$ and the sterling approximation $n! \approx \sqrt{2\pi n}(\frac{n}{e})^n$, we have condition (A) in Theorem 4 as:

$$\begin{aligned} &E(\tilde{f}_j{}^4)\mathrm{Corr}^2(\tilde{X}_j, \tilde{f}_j) + 4\sigma^2\mathrm{Corr}^2(\tilde{X}_j, \tilde{f}_j)E(\tilde{f}_j{}^2) \\ &\qquad = E(X_j^{4p_j})\frac{E^2(X_j^{p_j+1})}{E(X_j^{2p_j})} + 4\sigma^2 E^2(X_j^{p_j+1}) \\ &\qquad \approx \frac{2^{5p_j+1}p_j^{p_j}(p_j+1)^{p_j+1}}{e^{2p_j+1}} + \frac{2^3(p_j+1)^{p_j+1}}{e^{p_j+1}}\sigma^2, \end{aligned} \tag{3.1}$$

$$\begin{aligned} E(\tilde{X}_j{}^2\tilde{f}_j{}^2)\frac{E(\tilde{f}_j{}^2)}{E(\tilde{X}_j{}^2)} + \sigma^2 E(\tilde{f}_j{}^2) &= E(X_j^{2+2p_j})E(X_j^{2p_j}) + \sigma^2 E(X_j^{2p_j}) \\ &\approx \frac{2^{2p_j+2}(p_j+1)^{p_j+1}p_j^{p_j}}{e^{2p_j+1}} + \frac{2^{p_j+\frac{1}{2}}p_j^{p_j}}{e^{p_j}}\sigma^2. \end{aligned} \tag{3.2}$$

Thus, (3.1) $>$ (3.2), and $e_{\hat{\psi}_j^{(gcm)},\hat{\psi}_j^{(loco)}} = \left(\frac{sd(\hat{\psi}_j^{(loco)})/E(\hat{\psi}_j^{(loco)})}{sd(\hat{\psi}_j^{(gcm)})/E(\hat{\psi}_j^{(gcm)})}\right)^2 > 1$ up to the Sterling approximation.

Note that in practice, higher-order interaction terms are natural. While our theory does not extend to this setting, we provide a simulation comparison which includes interaction terms in Section 5.

3.3 Single Index Model

Next, we will generalize to a semi-parametric extension of the linear model: single index models which has connections to a shallow neural networks. For single index models with a

scalar response variable $Y \in \mathbb{R}$ and a predictor vector $X \in \mathbb{R}^p$, the model is:

$$Y = \eta(X^T\beta) + \epsilon$$

where $\beta = (\beta_1, ..., \beta_p)^T$ is unknown parameter vector, η is an unknown link function, and ϵ are i.i.d errors with zero mean and finite variance σ^2, and $\epsilon \perp X$. Then, for each $j \in [p]$, GCM and LOCO perform the same individual hypothesis testing as discussed previously. One way to achieve the relative variability of these two estimators in single index model is through Taylor expansion of $\eta(X^T\beta)$ around $E(X^T\beta|X_{-j})$. For notation simplicity, we use $\eta' = \eta'(E(X^T\beta|X_{-j}))$.

$$E(Y|X) = \eta(X^T\beta) = \eta(E(X^T\beta|X_{-j})) + \eta' \cdot (X^T\beta - E(X^T\beta|X_{-j})) + \Delta$$

where $\Delta = \sum_{s=2}^{\infty} \frac{1}{s!}\eta^{(s)}(E(X^T\beta|X_{-j}))(X^T\beta - E(X^T\beta|X_{-j}))^s$. Let's introduce the notation for order of approximation used in the following result: $f(n) = o(1)$ means if for all $\epsilon > 0$, there exists $M > 0$, such that $|f(n)| \leq \epsilon$ for all $n > M$.

Assumption (B1). η is fixed, monotone differentiable function with bounded derivatives $\eta^{(s)}(X^T\beta) \leq O(a^{-X^T\beta})$, $a > 1$, for all $s \geq 2$.

Assumption (B2). X has a sub-Gaussian distribution; that is, there exists $\sigma^2 > 0$ such that for all vectors $v \in \mathbb{R}^p$ with $\|v\|_2 = 1$ and all $t \in \mathbb{R}$, with $T = v^T X$, we have $P(|T - E(T)| > t) \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right)$.

Assumption (B3). $X^T\beta$ has a density that is bounded from above by $\bar{q}$.

Assumption (B4). For any $\delta > 0$, there exists a constant C, such that $E(X^T\beta|X_{-j}) \geq$

$C\max\{\sqrt{\frac{1}{\delta}}, \sqrt{K}\}$, for some constant $K \geq 3$ and tail probability δ.

Theorem 5. *Under the assumptions (B1) to (B4), consider a single index model estimated with mean squared error loss, $Y = \eta(X^T\beta) + \epsilon$, where ϵ is random noise with finite variance σ^2. For all $j \in [p]$, we have*

$$\frac{sd(\hat{\psi}_j^{(gcm)})}{E(\hat{\psi}_j^{(gcm)})} = \frac{\frac{1}{\sqrt{n}}\sqrt{\eta'^2\beta_j^2 Var(\tilde{X}_j^2) + \sigma^2 E(\tilde{X}_j^2) + o(1)}}{\beta_j \eta' E(\tilde{X}_j^2) + o(1)};$$
$$\frac{sd(\hat{\psi}_j^{(loco)})}{E(\hat{\psi}_j^{(loco)})} = \frac{\frac{1}{\sqrt{n}}\sqrt{\eta'^4\beta_j^4 Var(\tilde{X}_j^2) + 4\sigma^2\eta'^2\beta_j^2 E(\tilde{X}_j^2) + o(1)}}{\beta_j^2 \eta'^2 E(\tilde{X}_j^2) + o(1)}.$$

For any $\epsilon' > 0, \delta > 0$, with probability at least $1 - \delta$, we have $e_{\hat{\psi}_j^{(gcm)}, \hat{\psi}_j^{(loco)}} > 1 - \epsilon'$.

Remark 1.

- *The asymptotic efficiency proof technique here is similar to Theorem 3 except that we additionally employ a linear approximation via Taylor series expansion of $\eta(X^T\beta)$ around $E(X^T\beta|X_{-j})$, so assumption (B1) and (B4) ensure that first order approximation of η is sufficient and higher order derivatives terms are negligible, such that $\Delta, E(\Delta|X_{-j}) = o(1)$.*
- *(B2) and (B3) are relaxed assumptions for the guarantee of the convergence of the least-square estimator of (η, β) in monotone single index models introduced in Theorem 4.1 in Balabdaoui et al. (2019). Those assumptions will also be used to bound the polynomial terms in Taylor series expansion.*

The single index model is useful given its connection to single-layer neural networks if we have an unknown and fixed link function, or a generalized linear model if the link function is

known. Next, we illustrate a binary classification example to compare the relative efficiency of GCM and LOCO hypothesis testing where we continue using the same linear predictor set up as in Example 3.1 but with the sigmoid link function.

Example 3.3. Suppose $Y = \eta(\beta_0 + \beta_1 X_1 + \beta_2 X_2) + \epsilon$ where η is sigmoid function (ie: $\eta(x) = \frac{1}{1+e^{-z}}$), $X_i \sim N(0, \sigma^2), i = 1, 2$, $Cov(X_1, X_2) = \rho$, and $\epsilon \sim N(0, \sigma_\epsilon^2)$ is independent of matrix X.

The sigmoid function η is an increasing function with recursive derivatives (Minai and Williams (1993)) $\eta^{(s)}(x) = \sum_{k=0}^{s-1}(-1)^k \left\langle {s-1 \atop k} \right\rangle \eta(x)^{k+1}(1-\eta(x))^{s-k}$, where $\left\langle {s-1 \atop k} \right\rangle$ are Eulerian numbers. Thus, $\eta^{(s)}(x) \leq e^{-|x|}C_s = O(e^{-|x|}) = o(1)$ for $s \geq 2$ when $|x| \to \infty$, so the higher order derivatives in the remainder term of linear Taylor series expansion is negligible. The polynomial term in the remainder decays fast because of normal random variable X_j. For notation simplicity, let $\eta' = \eta'(E(\beta_0 + \beta_1 X_1 + \beta_2 X_2 | X_2)) = \eta'(E(\beta_0 + \beta_1 \rho X_2 + \beta_2 X_2)) = \eta\left(\beta_0 + (\beta_1\rho + \beta_2)X_2\right) \cdot \left(1 - \eta\left(\beta_0 + (\beta_1\rho + \beta_2)X_2\right)\right)$. Thus, the relative efficiency of GCM and LOCO test statistics for X_1 is

$$e_{\hat{\psi}_{0,1}^{(gcm)}, \hat{\psi}_{0,1}^{(loco)}} = \left(\frac{sd(\hat{\psi}_{0,1}^{(loco)})}{\beta_1 \eta' sd(\hat{\psi}_{0,1}^{(gcm)})}\right)^2 = \frac{\frac{4}{\beta_1^2 \eta'^2}\sigma_\epsilon^2(1-\rho^2)\sigma^2 + 2(1-\rho^2)^2\sigma^4}{\frac{1}{\beta_1^2 \eta'^2}\sigma_\epsilon^2(1-\rho^2)\sigma^2 + 2(1-\rho^2)^2\sigma^4}\Big(1 + o(1)\Big).$$

Thus, for any $\epsilon' > 0$, there exists a sufficiently large η', such that we have $e_{\hat{\psi}_{0,1}^{(gcm)}, \hat{\psi}_{0,1}^{(loco)}} > 1 - \epsilon'$.

4. Extension to comparison with Dropout and Lazy-VI

The previous section compares the asymptotic relative efficiency for GCM and LOCO. In this section, we extend the comparison to estimators from dropout and Lazy Variable Importance (Lazy-VI) (Gao et al. (2022)) approaches which are used as computationally

faster approaches that attempt to estimate similar parameters to LOCO. Let's first review the setup of dropout and Lazy-VI.

4.1 Dropout

The dropout approach is a counterpart to LOCO which avoids retraining a reduced model. Instead, the j^{th} feature is replaced by its corresponding marginal mean and then reinserted into the pre-trained full model f_0. Let's define the dropout features as $X^{(j)}$ (resp. $X_i^{(j)}$), where $X^{(j)} = (X_1, ..., X_{j-1}, \mu_j, X_{j+1}, ..., X_p)$. The plug-in estimator for the dropout is

$$\hat{\psi}_j^{(dr)} := V(f_n, P_n) - V(f_n, \tilde{P}_{n,-j})$$

where $\tilde{P}_{n,-j}$ is the empirical distributions of $X^{(j)}$. The dropout estimator for feature j is :

$$\widehat{\psi}_j^{(dr)} = \frac{1}{n}\sum_{i=1}^{n}\{[Y_i - f_n(X_{ij})]^2 - [Y_i - f_n(X_i^{(j)})]^2\}.$$

4.2 Lazy-VI

The prediction functions in GCM, LOCO, and dropout tests can be estimated using user-selected learners. Lazy-VI Gao et al. (2022) is a computationally efficient approximation to LOCO, primarily developed for neural networks because it relies on gradient-based updates, and an ensemble-based algorithm like random forest does not apply. Instead of retraining a reduced model for each feature, Lazy-VI fits the full neural network once and approximates the reduced model by solving a linearized ridge regression around the full-model parameters. Thus, Lazy-VI reduces the computational cost of LOCO while retaining asymptotic guarantees under neural network setup. We briefly outline its setup below.

Let a neural network function class $\left\{h_\theta(x) : \mathbb{R}^p \mapsto \mathbb{R} \mid \theta \in \mathbb{R}^M\right\}$ be parameterized by a vector θ, such that the full model estimator is

$$\theta_f = \arg\min_{\theta \in \mathbb{R}^M} \frac{1}{n} \sum_{i=1}^{n} \left[Y_i - h_\theta\left(X_i\right)\right]^2.$$

However, unlike LOCO, which need to retrain the reduced model, Lazy-VI approximates the reduced-model parameter by $\theta_f + \Delta\theta_j(\lambda, n)$, where

$$\Delta\theta_j(\lambda, n) = \arg\min_{\omega \in \mathbb{R}^M} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[Y_i - h_{\theta_f}(X_i^{(j)}) \; - \omega^\top \nabla_\theta h_\theta(X_i^{(j)}) \Big|_{\theta=\theta_f} \right]^2 + \lambda \|\omega\|_2^2 \right\}$$

and $\lambda > 0$ is the penalty parameter. Hence, the variable importance score under Lazy-VI is

$$\psi_j^{(lazy)} = V\left(h_{\theta_f}, P_0\right) - V\left(h_{\theta_f + \Delta\theta_j(\lambda,n)}, P_{0,-j}\right).$$

Theorem 6. *(Re-formulated from Gao et al. (2022)) Under the regularity conditions required in Theorem 4.4 of Gao et al. (2022), the Lazy-VI estimator under the mean squared error predictive measure*

$$\widehat{\psi}_j^{(lazy)} = \frac{1}{n} \sum_{i=1}^{n} \left\{ \left[Y_i - h_{\theta_f}\left(X_i\right)\right]^2 - \left[Y_i - h_{\theta_f + \Delta\theta_j(\lambda,n)}(X_i^{(j)})\right]^2 \right\}$$

has the following result:

$$\sqrt{n}[\hat{\psi}_j^{(lazy)} - \psi_j^{(lazy)}] \xrightarrow{d} N(0, \sigma_j^2), \text{ where } \sigma_j^2 = Var(\epsilon^2 - \epsilon_{-j}^2).$$

LOCO, dropout and Lazy-VI are asymptotically unbiased estimators used to estimate

the ability of X_j in predicting Y in terms of the variance only. Then, we can extend the asymptotic relative efficiency comparison to those methods in the following theorem:

Theorem 7. *Under the regularity conditions required in Theorem 4.4 of Gao et al. (2022), together with respective assumptions in Theorem 3 Theorem 4, and Theorem 5, then we have the corresponding asymptotic variance relationship:*

$$\lim_{n\to\infty} Var(\hat{\psi}_j^{(loco)}) = \lim_{n\to\infty} Var(\hat{\psi}_j^{(lazy)}) \leq \lim_{n\to\infty} Var(\hat{\psi}_j^{(dr)}).$$

5. Simulations

In this section, we validate our theoretical results by comparing GCM, LOCO, dropout and Lazy-VI. We also provide a comparison to other state-of-the-art feature selection methods such as Lasso (Ma and Huang (2008)), Knockoff (Barber and Candès (2015); Candès et al. (2018)), and permutation testing (Breiman (2001)). We assess their efficiency and accuracy for the linear model, single index model, and various non-linear models both with and without interaction terms. Specifically, the following simulation settings are used:

(a) Linear model:

$Y \sim 0.01X_1 + 0.1X_2 + X_3 + X_4 + 1.5X_5 + 2X_6 + 3X_7 + 4X_8 + \epsilon$, where $X \sim N(0, \Sigma_{20\times 20})$, $\mathrm{Corr}(X_3, X_4) = 0.5$, $\epsilon \sim N(0, 0.1^2)$.

(b) Non-linear additive model :

$Y \sim 2X_1^2 + 2\cos(4X_2) + \sin(X_3) + \exp(X_4/3) + 3X_5 + X_6^3 + 5X_7 + \max(0, X_8) + \epsilon$, where $X \sim N(0, \Sigma_{20\times 20}), \epsilon \sim N(0, 0.1^2)$;

(c) Non-additive model with interaction :

$Y \sim 2\sin(X_1) + \log(|X_2| + 1) + 3X_1X_2 + \cos(X_3 + X_4) + X_5^3 + X_6X_7X_8 + \epsilon$, where $X \sim N(0, \Sigma_{20\times 20}), \epsilon \sim N(0, 0.1^2)$;

(d) Single Index Model :

$Y \sim f(X\beta) + \epsilon$, where f is sigmoid function, $\beta = (6, 5, 4, 3, 2, 1, 0.5, 0.1, 0, ..., 0)^T \in \mathbb{R}^{20}, X \sim N(0, \Sigma_{20\times 20})$, $\text{Corr}(X_1, X_2) = 0.5, \epsilon \sim N(0, 0.1^2)$.

5.1 Variable selection comparison across methods

In this section, the key theoretical results will be validated by the following simulations. For all models (a)-(d), the response variable depends only on the first 8 of $p = 20$ features, and we use $n = 1000$ observations in each of the 100 simulations. We discuss the feature selection methods through canonical implementations designated for each model (see Supplementary Material S4 for more details). However, the variable importance selection methods we investigate here are not restricted to regression problems. We consider both algorithms suggested by the corresponding models, and the same method across different models. Specifically, we choose a widely used tree-based algorithm: gradient boosted decision tree with default setting (`gbm`, Ridgeway et al. (2024) package in `R`). The average p-value for each feature performed by GCM, LOCO, dropout and permutation variable selection is presented in Fig. 1. What's more, the performance comparison is also extended with other popular methods which do not provide p-values, e.g. Lasso and Knockoff with target FDR = 0.2 across all models, is reported in Fig. 3 and the dashed line is the empirical threshold from random coin-flipping. (`glmnet`, Friedman et al. (2023); `knockoff`, Barber et al. (2022) packages in `R`).

In Fig. 1, plot (a) shows that all approaches have similar performance in the linear model

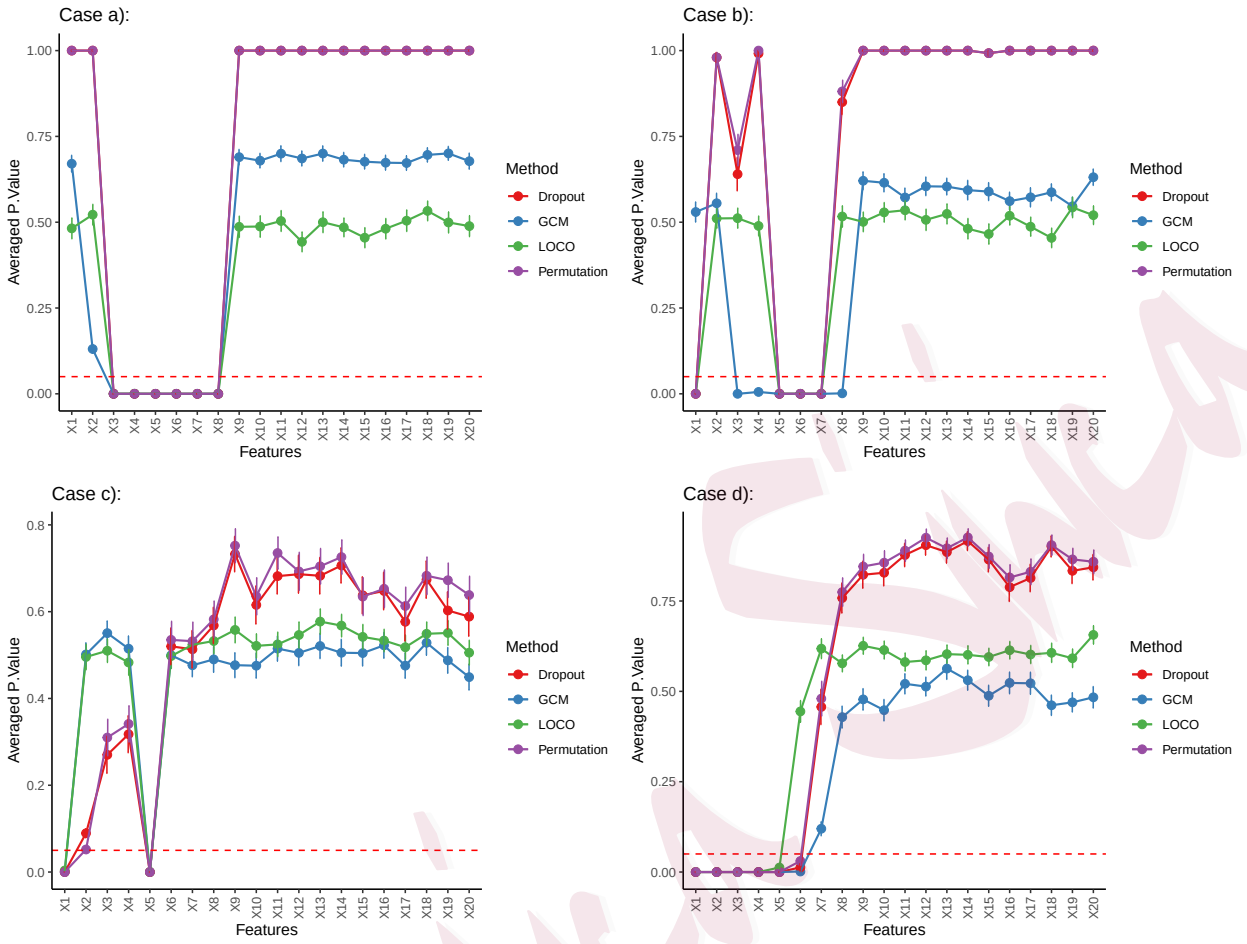

Figure 1: Averaged p-value and standard deviation of each feature for model (a) - (d) for GCM, LOCO, dropout and permutation univariate selection on dataset using gradient boosted decision tree. Red dashed line: targeted type-I error rate ($\alpha = 0.05$).

case. The performance of lasso is slightly better than others with the highest averaged accuracy and F1 score as shown in Fig. 3. For the additive non-linear model (b) in Fig. 1, LOCO requires approximately half the computational time of GCM, but GCM can identify more significant variables than LOCO as long as the ratio checking (see Fig. 2) LHS over RHS in assumption (A) in Theorem 4 holds. The first two ratios are 0 because the quadratic and cosine functions are even functions and yield the corresponding $\text{Corr}(\tilde{X}_j, \tilde{f}_{0,j}(X_j)) = 0, j = 1, 2,$

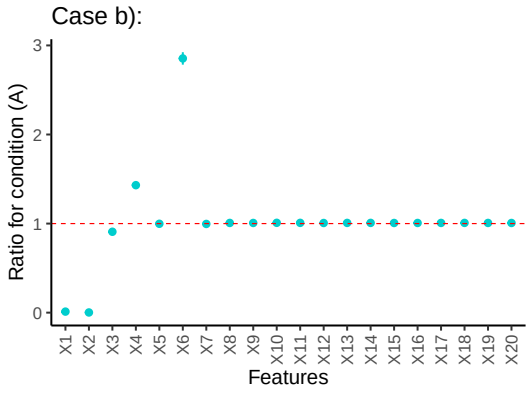

Figure 2: Averaged ratio with standard deviation for verifying theoretical condition (A) (LHS/RHS) in Theorem 4 for case b).

which violates assumption (A). What's more, GCM outperforms knockoff and lasso methods in the non-linear model with high accuracy, F1 score and low FDR, since the underlying model is not sparse as shown in Fig. 3. The simulation with the non-additive nonlinear model (c) examines the necessity of assuming the absence of joint effects in the application of all variable selection approaches even under "black-box" machine learning algorithm, since all of the approaches barely identify the significant features which yields low recall. For single index model (d), as illustrated in Fig. 1, all four methods effectively identify variables with relatively large weights but fail to detect those with smaller contributions. The GCM measure is the most computationally demanding method, as it requires training $2p$ models. In contrast, the LOCO approach trains $p + 1$ models, while the dropout, permutation and lasso method requires less computational effort with only a single model needs to be trained. The knockoff filter method maintains single-model computation efficiency but requires extra time for knockoff feature generation. Overall, the GCM approach is more stable and effective compared to all other feature selection methods. Additional simulation results for models trained with kernel SVM, presented in Supplementary Material S4, further validate

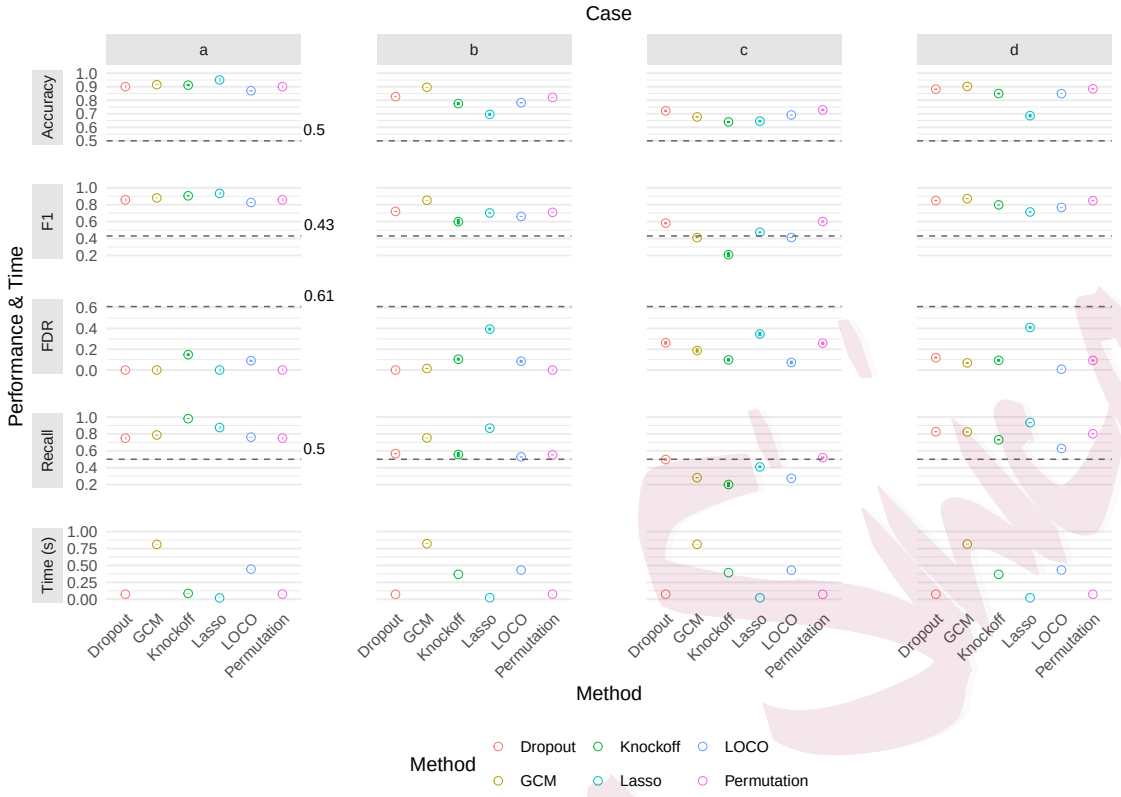

Figure 3: Average performance measure and time (seconds) with standard deviation across 100 simulations for lasso, knockoff (target FDR = 0.2), GCM, LOCO, dropout and permutation univariate selection on dataset using gradient boosted decision tree for (a) - (d). Dashed line: empirical random-guessing baseline.

the strength and robustness of GCM.

5.2 Model Mis-specification

To assess the robustness of the proposed comparisons beyond the correctly specified settings used in the theoretical results, we consider two mis-specified models, one with interactions and the other with heavy-random variable with heavy-tailed distributions (which violates sub-Gaussianity) with $n = 1000$ observations in each of the 50 simulations:

(e) Non-additive model with surrogate feature:

$$Y \sim \beta X_1 + 0.05\,(X_3 X_4^2 + X_4 X_3^2) + X_5 + \varepsilon, \quad \varepsilon \sim N(0,1), \text{ and } X_2 = \rho X_1 + \sqrt{1-\rho^2} Z,$$

where $Z \sim N(0,1), \rho \in \{0, 0.3, 0.5, 0.7, 0.8, 0.9\}$ and $X \sim N(0, I_{20\times 20})$.

(f) Nonlinear additive model with heavy-tailed covariates :

$$Y \sim \beta X_1 + 1.5\sin(X_2) + 2 * \mathbf{I}\{X_3 > 0\} + \tanh(X_4) + 0.6 X_5^3 + \epsilon, \text{where } \epsilon \sim N(0,1), X_j \overset{\text{i.i.d.}}{\sim}$$

$t_\nu, \nu \in \{3, 5, 10, 15, 30, \inf\}, j = 1, ..., 20.$

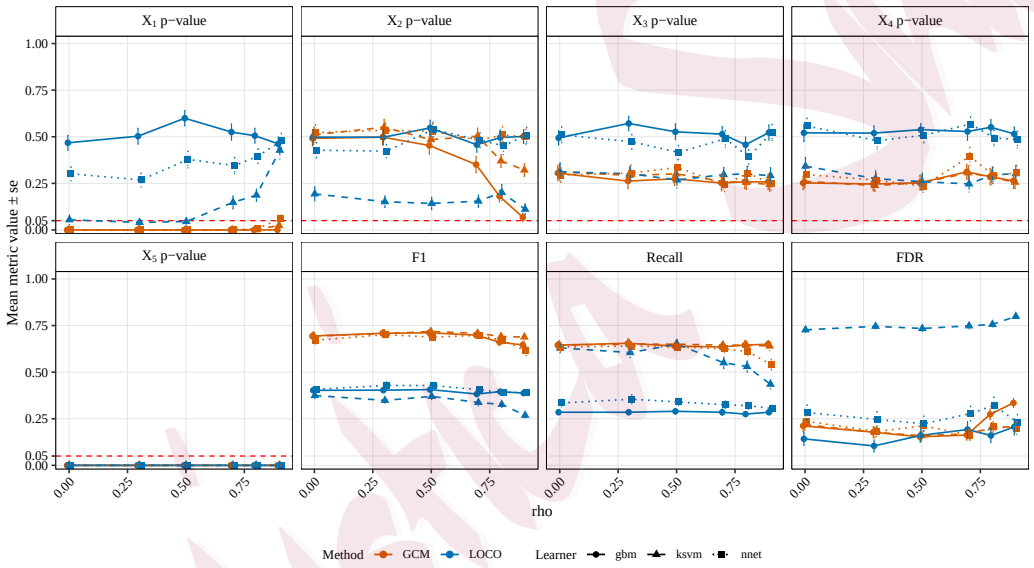

Figure 4: Performance of GCM and LOCO under t-distributed covariates with varying degrees of freedom. Panels show mean X_j p-values and performance metrics over 50 simulations, with error bars denoting standard errors. The blue dashed line marks $\alpha = 0.05$.

In model (e), X_2 is correlated with the significant variable X_1, but has no direct effect on the response. The model is intended to evaluate robustness of comparison between LOCO and GCM to model misspecification. For the learners, GBM is implemented with 50 trees and a learning rate of 0.05, neural networks are fitted using the default hyperparameters,

and KSVM is fitted with (C=0.1). From Fig.4, we see that across different learners, conclusions from the theoretical results remain consistent. GCM always identifies X_1 regardless of whether X_2 is weakly or strongly correlated with X_1, whereas LOCO is more sensitive to the correlation and produces more false discoveries. Neither method reliably detects the non-additive interaction term X_3 and X_4. Overall, GCM is more stable than LOCO, achieving higher F1 score and recall. Model (f) considers a nonlinear additive model with heavy-tailed

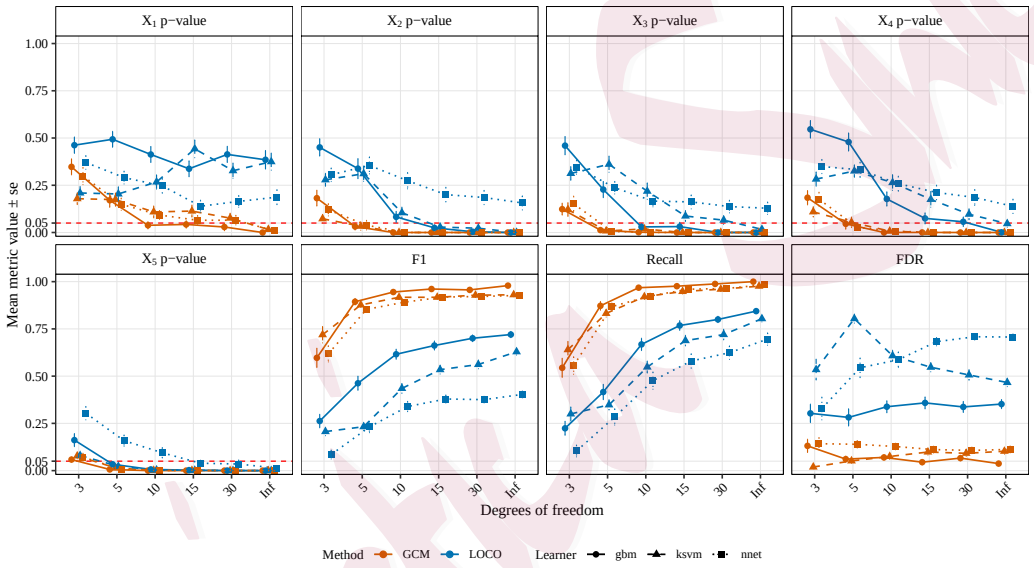

Figure 5: Performance of GCM and LOCO under t-distributed covariates with varying degrees of freedom. Panels show mean X_j p-values and performance metrics over 50 simulations, with error bars denoting standard errors. The blue dashed line marks $\alpha = 0.05$.

covariates. The covariates are generated independently from t-distributions. To make it comparable across different degrees of freedom, each covariate is standardized to have unit variance. The model evaluates the performance of GCM and LOCO generalizes beyond the model from beyond Gaussian or light-tailed covariate distributions. All three learners are implemented with default settings. Fig.5 shows that GCM is more stable than LOCO across

different degrees of freedom and machine learners. GCM consistently achieves higher F1 score and recall while maintaining a lower FDR. In contrast, LOCO performs worse under heavier tails, but gradually improves as the degrees of freedom increase, in other words, the covariate approaches Gaussian distribution, allowing it to better identify the significant features.

These findings indicate that the superior performance of GCM remains across the misspecified settings considered, suggesting greater empirical robustness beyond standard modeling assumptions.

5.3 Comparison with Lazy-VI in neural networks

In the next simulation, we extend the performance comparison with Lazy-VI in a wide neural network setup. For this comparison, we implement simulations in Python as the estimation procedure for a lazy training method is not easily transferable to R because it relies on a sufficiently large neural network function class and other designated assumptions. To generate synthetic data, we use the same setup of the neural network as that of Gao et al. (2022) proposed for evaluating the performance of Lazy-VI. We train a wide, fully connected two-layer neural network for all simulations, using Adam with learning rate 10^{-3} and the width of the hidden layer in the training network is 50.

In Fig. 6, we display the performance comparison among LOCO, GCM, Lazy-VI and dropout across 100 simulations. We find that GCM still has better performance than others in terms of high accuracy, F1 score and recall in model b) and d) with sigmoid link function albeit at the expense of computational time. Lazy-VI has proved to be fast and accurate in approximating the LOCO test. LOCO and Lazy-VI have similar performance which is con-

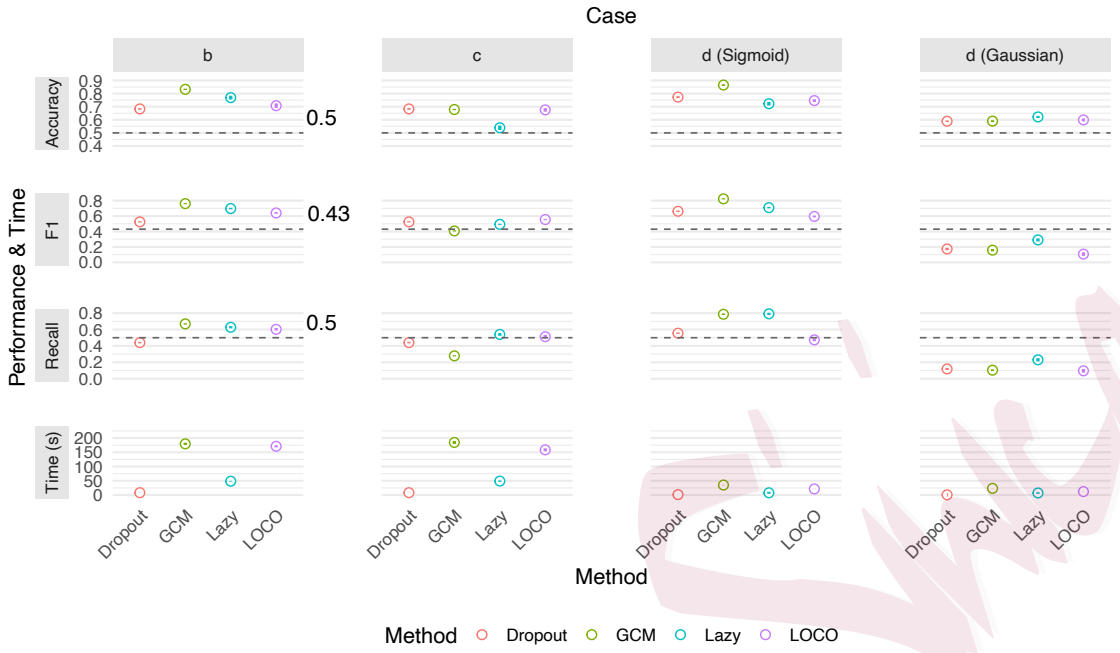

Figure 6: Average performance measure and time (seconds) with standard deviation across 100 simulations for GCM, LOCO, dropout and Lazy-VI univariate selection on dataset using large neural network for (b), (c), and (d) with sigmoid and gaussian link function. Dashed line: empirical random-guessing baseline.

sistent with our theory. For model (c), all methods perform poorly due to interaction terms, resulting in low statistical power. For the single index model (d), when the monotonicity assumption (B1) in Theorem 5 is violated, the gaussian link function (ie: e^{-x^2}), none of the feature selection methods work well.

5.4 High-dimensional Regression

We also check the behavior of all feature selection methods in high-dimensional setting. We use 500 covariates. Specifically, the first 40 variables are significant and are generated by repeating the same eight-significant-feature block in model (a) and (b) five times, while

the remaining 460 variables are noise variables. In setting (a), all covariates are independent except that the third and fourth active variables within each repeated block have correlation 0.5. In setting (b), all covariates are independent. The computational burden is more pronounced in high-dimensional setup since GCM requires retraining more models than other methods. Fig. 7 displays the average performance of each feature selection method over 50 simulations in the high-dimensional setting. All methods are implemented using gradient boosted trees with the default `gbm` settings in R, except that the tree depth is set to 4. The dashed line represents the empirical threshold obtained from fair coin-flipping sampling. Knockoff filter method works well in all sparse models with high F1 score and low FDR, lasso has the best performance in sparse linear model (a). The performance of GCM is slightly better than other wrapper-based selection with higher accuracy, recall and F1 score in case (a), and GCM performs as good as others in non-linear high-dimensional setting. Model training choices also influence feature selection. Further simulation results with kernel SVM also validate the stability and power of GCM (see Supplementary Material S4).

6. Real data analysis

In this section, we evaluate GCM and LOCO on two real datasets: Airbnb pricing data Inside Airbnb (2026) and social media addiction data Zaman (2026). We use 5-fold cross-validation and report the average MSE with its standard error across folds in tables. Stability selection Meinshausen and Bühlmann (2010) where if the feature is selected in 4 or 5 folds is used to select features. We use linear regression, gradient boosted decision tree and neural network prediction algorithms. Additional data analysis results are available in

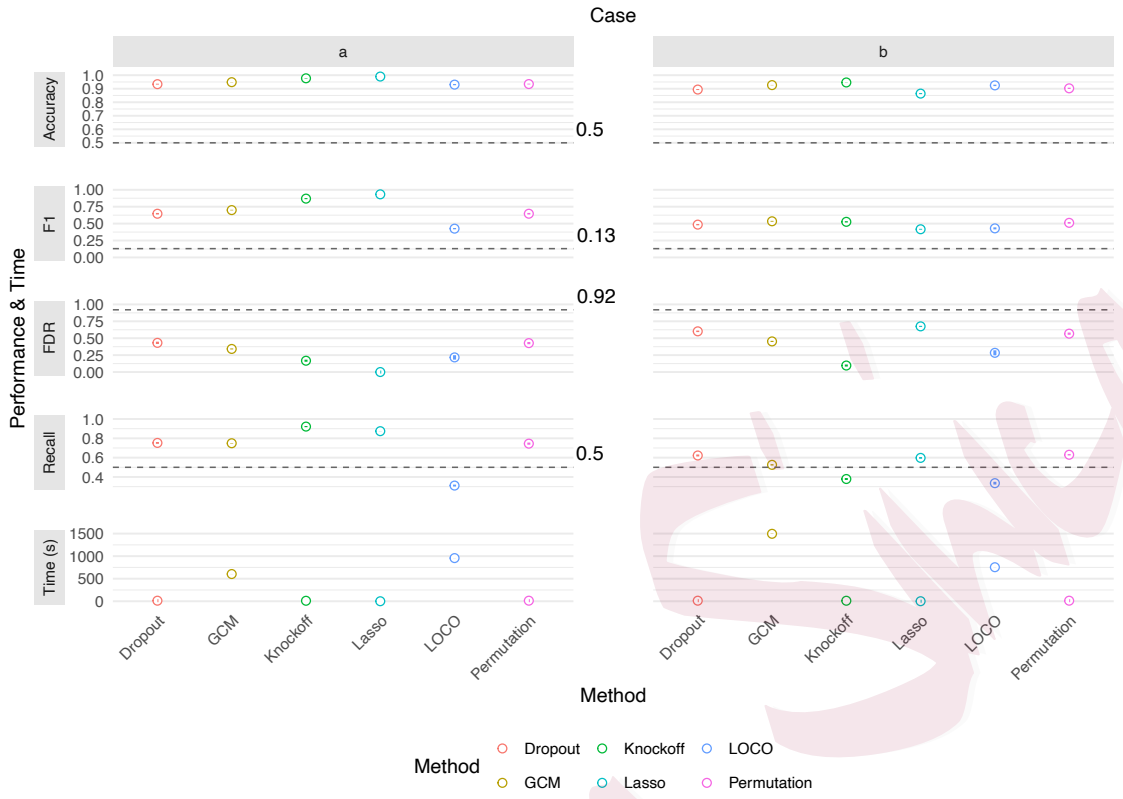

Figure 7: Average performance measure and time (seconds) with standard deviation across 50 simulations for lasso, knockoff (target FDR = 0.2), GCM, LOCO, dropout and permutation univariate selection on dataset using GBM for (a)-(d). Dashed line: empirical random-guessing baseline.

Supplementary Materials S4.

6.1 Airbnb listing price data

As Airbnb has transformed the short-term rental market, understanding the factors that influence listing prices has become increasingly important for hosts and travelers. The data consist of $1,863$ Airbnb listings from Quebec City, Canada, collected in June 2026 from the publicly available platform Inside Airbnb (2026). The response variable Y is

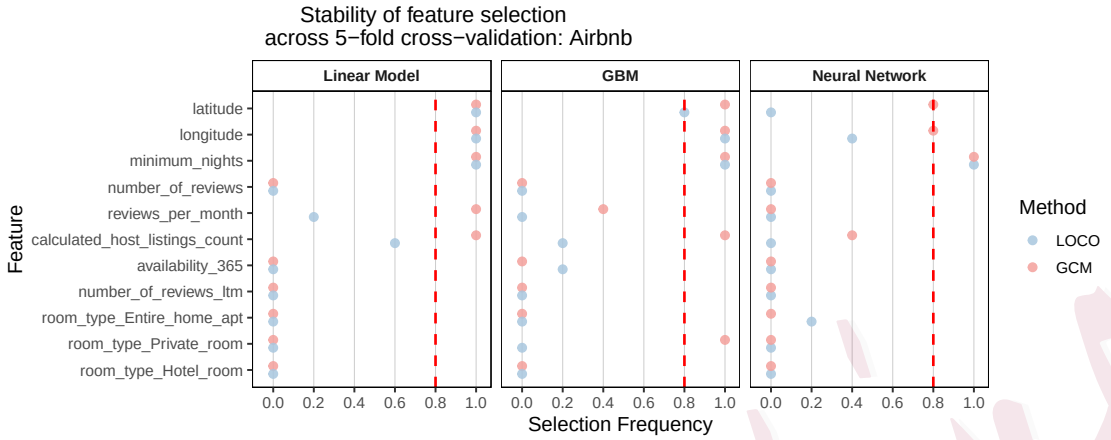

Figure 8: Selection rate of each feature selected by GCM and LOCO across 5-fold cross validation on Airbnb listing price data, using lm, GBM and NN as the base predictor respectively. Red dashed line: targeted selection frequency.

the log-transformed listing price, and the covariates include geographic location (`latitude` and `longitude`), listing characteristics (`minimum nights`, `room type`), host information (`calculated host listings count`), and booking activity measures (`number of reviews`, `reviews per month`, `number of reviews` in the last 12 months, and `availability` throughout the year). The categorical variable `room type` is embedded using one-hot encoding. For Airbnb listing price data, the GBM model uses the default setting in R. The neural network is a fully connected feed-forward model with hidden layers of 32 and 16 neurons, learning rate 10^{-3}, weight decay 10^{-3}, and batch size 64. From Table 1, we see that GCM consistently achieves the lower averaged MSE across all different machine learning learners. Fig.8 shows that GCM consistently identifying the significant variables in predicting listing price, such as `latitude`, `longitude`, and `minimum_nights`. Across all learners, GCM generally tends to select more variables than LOCO and yields more stable selections.

Table 1: Mean squared prediction error averaged across 5-fold cross-validation for GCM and LOCO on Airbnb listing price data, using lm, GBM, NN respectively.

	Models:		
Method	lm	GBM	NN
GCM	0.3338 ± 0.0093	0.2507 ± 0.0134	0.2892 ± 0.0113
LOCO	0.3366 ± 0.0105	0.2960 ± 0.0103	0.3227 ± 0.0272

6.2 Social media addiction dataset

Whether social media use enhances or harms productivity has become an increasingly important question as digital platforms play a central role in everyday life. The dataset is publicly available on Kaggle (Zaman (2026)) and contains 4,999 observations after removing missing values. The response variable is `productivity score`, and the covariates include `age`, `daily screen time`, `social media hours`, `study hours`, `sleep hours`, `notifications per day`, `focus score`, and `addiction level`. For social media addiction data, the GBM model uses 500 trees and depth 4 in R. The neural network is a fully connected feed-forward model with hidden layers of 32 and 16 neurons, learning rate 10^{-3}, weight decay 10^{-3}, and batch size 128. Fig.9 shows that both GCM and LOCO select `social media hours`, `study hours`, `sleep hours`, and the number of `notifications per day` consistently across learners. Compared with LOCO, GCM exhibits more stable selection, with same or higher frequencies for important variables and lower frequencies for insignificant variables. Table 2 shows that the two methods have identical prediction error under the linear model, while GCM achieves slightly lower mean squared prediction error than LOCO under GBM and neural network.

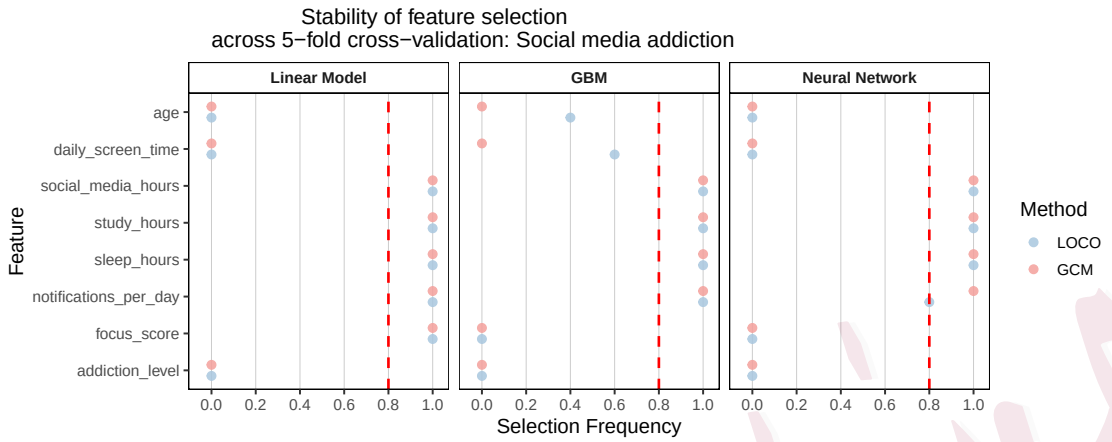

Figure 9: Selection rate of each feature selected by GCM and LOCO across 5-fold cross validation on impact of social media addiction on productivity data, using lm, GBM and NN as the base predictor respectively. Red dashed line: targeted selection frequency.

Table 2: Mean squared prediction error averaged across 5-fold cross-validation for GCM and LOCO on impact of social media addiction on productivity data, using lm, GBM, NN respectively.

	Models:		
Method	lm	GBM	NN
GCM	95.95 ± 2.56	91.23 ± 2.08	84.70 ± 1.95
LOCO	95.95 ± 2.56	91.85 ± 2.02	85.61 ± 2.31

7. Conclusion

A key result of this paper is that GCM demonstrates greater efficiency in variable selection compared to LOCO under suitable regular conditions for linear, additive and single index models. Application to real-data variable selection further confirms that GCM selects models with lower prediction MSE than LOCO and is more stable. Future work could develop a novel GCM-based test statistic that improves computational efficiency or strikes a compromise with Type II error, advancing hypothesis testing methodologies.

Supplementary materials

The online Supplementary Material contains an extra efficiency comparison example for additive model (Section S1). Mean and variance for test statistics under different models (Section S2), proofs of theorems (Section S3), and additional simulation and data analysis results (Section S4).

References

Altmann, A., L. Tološi, O. Sander, and T. Lengauer (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics 26*(10), 1340–1347.

Balabdaoui, F., C. Durot, and H. Jankowski (2019). Least squares estimation in the monotone single index model. *Bernoulli 25*(4B).

Barber, R. F., E. Candes, L. Janson, E. Patterson, and M. Sesia (2022). knockoff: The Knockoff Filter for Controlled Variable Selection.

Barber, R. F. and E. J. Candès (2015). Controlling the False Discovery Rate Via Knockoffs. *The Annals of Statistics 43*(5), 2055–2085.

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

Candès, E., Y. Fan, L. Janson, and J. Lv (2018). Panning for Gold: 'Model-X' Knockoffs for High Dimensional Controlled Variable Selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology 80*(3), 551–577.

Chang, C.-H., L. Rampasek, and A. Goldenberg (2018). Dropout Feature Ranking for Deep Learning Models.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2 ed.). Lawrence Erlbaum.

Cortizo, J. C. and I. Giraldez (2006). Multi Criteria Wrapper Improvements to Naive Bayes Learning. In *Intelligent Data Engineering and Automated Learning – IDEAL 2006*, Volume 4224, pp. 419–427.

Duda, R. O., P. Hart, and D. Stork (2000). Pattern classification/Duda RO, Hart PE, Stork DG–.

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

Fisher, R. A. (1992). Statistical Methods for Research Workers. In S. Kotz and N. L. Johnson (Eds.), *Breakthroughs in Statistics: Methodology and Distribution*, pp. 66–70. Springer.

Fleuret, F. (2004). Fast binary feature selection with conditional mutual information. *Journal of Machine learning research 5*(9).

Friedman, J., T. Hastie, R. Tibshirani, B. Narasimhan, K. Tay, N. Simon, J. Qian, and J. Yang (2023). glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models.

Gao, Y., A. Stevens, G. Raskutti, and R. Willett (2022). Lazy Estimation of Variable Importance for Large Neural Networks. *International Conference on Machine Learning*, 7122–7143.

Guyon, I., J. Weston, S. Barnhill, and V. Vapnik (2002). Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning 46*(1), 389–422.

Hoque, N., D. K. Bhattacharyya, and J. K. Kalita (2014). MIFS-ND: A mutual information-based feature selection method. *Expert Systems with Applications 41*(14), 6371–6385.

Inside Airbnb (2026). Inside airbnb: Get the data. `https://insideairbnb.com/get-the-data/`. Quebec City, Quebec, Canada listings dataset, June 2026.

Kittler, J. (1978). Feature set search algorithms. *Pattern recognition and signal processing*.

Lehmann, E. L. and J. P. Romano (2005). *Testing Statistical Hypotheses* (3 ed.). Springer.

Lei, J., M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman (2018). Distribution-Free Predictive Inference for Regression. *Journal of the American Statistical Association 113*(523), 1094–1111.

Ma, S. and J. Huang (2008). Penalized feature selection and classification in bioinformatics. *Briefings in Bioinformatics 9*(5), 392–403.

Meinshausen, N. and P. Bühlmann (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72*(4), 417–473.

Mentch, L. and G. Hooker (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *Journal of Machine Learning Research 17*(26), 1–41.

Minai, A. A. and R. D. Williams (1993). On the derivatives of the sigmoid. *Neural Networks 6*(6), 845–853.

Pearson, K. and F. Galton (1997). VII. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London 58*(347-352), 240–242.

Ridgeway, G., D. Edwards, B. Kriegler, S. Schroedl, H. Southworth, B. Greenwell, B. Boehmke, J. Cunningham, and GBM Developers (2024). gbm: Generalized Boosted Regression Models.

Rinaldo, A., L. Wasserman, and M. G'Sell (2019). Bootstrapping and sample splitting for high-dimensional, assumption-lean inference. *The Annals of Statistics 47*(6), 3438–3469.

Sandri, M. and P. Zuccolotto (2006). Variable Selection Using Random Forests. In S. Zani, A. Cerioli, M. Riani, and M. Vichi (Eds.), *Data Analysis, Classification and the Forward Search*, pp. 263–270. Springer.

Shah, R. D. and J. Peters (2020). The Hardness of Conditional Independence Testing and the Generalised Covariance Measure. *The Annals of Statistics 48*(3).

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology 15*(1), 72–101.

Strobl, C., A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis (2008). Conditional variable importance for random forests. *BMC Bioinformatics 9*(1), 307.

Székely, G. J., M. L. Rizzo, and N. K. Bakirov (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics 35*(6), 2769–2794.

Tansey, W., V. Veitch, H. Zhang, R. Rabadan, and D. M. Blei (2022). The Holdout Randomization Test for Feature Selection in Black Box Models. *Journal of Computational and Graphical Statistics*.

van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press.

Vangel, M. G. (1996). Confidence intervals for a normal coefficient of variation. *The American Statistician 50*(1), 21–26.

Williamson, B. D., P. B. Gilbert, N. R. Simon, and M. Carone (2022). A General Framework for Inference on Algorithm-Agnostic Variable Importance. *Journal of the American Statistical Association 118*(543), 1645–1658.

Witten, I. H., E. Frank, M. A. Hall, C. J. Pal, and M. Data (2005). *Practical machine learning tools and techniques*, Volume 2. Elsevier Amsterdam, The Netherlands. Issue: 4.

Yu, L. and H. Liu (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pp. 856–863.

Zaman, A. (2026). Social media addiction vs productivity dataset.

Zhang, K., J. Peters, D. Janzing, and B. Schoelkopf (2012). Kernel-based Conditional Independence Test and Application in Causal Discovery.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology 67*(2), 301–320.

Department of Statistics, University of Wisconsin - Madison, Madison, WI, USA

E-mail: chenghui.zheng@wisc.edu

Department of Statistics, University of Wisconsin - Madison, Madison, WI, USA

E-mail: raskutti@stat.wisc.edu