Comparing Model-agnostic Feature Selection Methods through Relative Efficiency

Chenghui Zheng[1], Garvesh Raskutti[1]

[1] *University of Wisconsin - Madison*

Supplementary Material

The online supplementary materials provide an additional efficiency ratio comparison in additive model in Section 1. Mean and variance for test statistics under different models in Section 2, technical proofs in the main paper in Section 3, and additional simulation and real data analysis results in Section 4.

S1 Additional additive model example

Example 1. Let $Y = cos(X_1) + sin(X_2) + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$. We have $X_j \sim N(0,1)$, X_j's are independent, and we have $\tilde{X}_j = X_j - E(X_j|X_{-j}) = X_j$.

For $cos(X_1)$, we have $\tilde{f}_1(X_1) = cos(X_1) - E(X_1|X_2)$, and then the $\text{Corr}^2(\tilde{X}_1, \tilde{f}_1(X_1)) = 0$ due to the independence assumption and the symmetry of cosine function. Thus, the even function with symmetric distribution of X violates the condition (A) in Theorem 4, and LOCO estimator is

more efficient than GCM estimator.

For $sin(X_2)$, we have $\tilde{f}_2(X_2) = sin(X_2)$. By Taylor series expansion, let's approximate $sin(X_2) \approx X_2 - \frac{X_2^3}{3!} + \frac{X_2^5}{5!}$. The numerator in RHS of the inequality in condition (A) is bounded by

$$E(\tilde{X_2}^2 \tilde{f_2}^2)\frac{E(\tilde{f_2}^2)}{E(\tilde{X_2}^2)} + \sigma^2 E(\tilde{f_2}^2) \leq E[X_2^2(3X_2^{p_j})^2]E[(3X_2)^{2*p_j}] + \sigma^2 E[(3X_2^{p_j})^2]$$
$$\approx 81E(X_2^{2+2p_j})E(X_2^{2p_j}) + 9\sigma^2 E(X_2^{2p_j}). \tag{S1.1}$$

Let $p_j = 5$ based on the approximation chosen above, and plug into Eq.(3.1) and (3.2) derived from Example 3.2 in the main context. We see that the upper bound (S1.1) of RHS of the numerator in condition (A) is less than the lower bound of Corr^2 times LHS of denominator in condition (A) (ie: $E(X_j^{4p_j})\frac{E^2(X_j^{p_j+1})}{E(X_j^{2p_j})} + 4\sigma^2 E^2(X_j^{p_j+1})$). Thus, GCM estimator is more efficient than LOCO estimator up to the Taylor approximation of the sine function.

S2 Mean and variance for test statistics under different models

In the following sections, we compute all the expectation and variance of test statistics for different models. We will focus on the expression of

population parameters since the expectation and variance of the plug-in estimator can directly be derived, i.e., $E(\hat{\psi}_{0,j}^{(\cdot)}) = E(\psi_{0,j}^{(\cdot)})$ and $Var(\hat{\psi}_{0,j}^{(\cdot)}) = \frac{1}{n}Var(\psi_{0,j}^{(\cdot)})$.

S2.1 Linear Model

Recall linear model setup: $Y = X_{-j}\beta_{-j} + \beta_j X_j + \epsilon$ where ϵ are i.i.d mean zero with finite variance σ^2, and $\epsilon \perp X$.

GCM test statistics For GCM hypothesis testing of each feature j, we have the product of residuals from regression as:

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

where $h_{0,j}(X_{-j}) = E(X_j|X_{-j}), g_{0,j}(X_{-j}) = E(Y|X_{-j}) = X_{-j}\beta_{-j} + \beta_j E(X_j|X_{-j})$, and $\tilde{X}_j := X_j - E(X_j|X_{-j})$.

$$\begin{aligned}
E(\psi_{0,j}^{(gcm)}) &= E[(X_j - E(X_j|X_{-j}))(Y - E(Y|X_{-j}))] \\
&= E[(X_j - E(X_j|X_{-j}))(X_{-j}\beta_{-j} + \beta_j X_j + \epsilon - X_{-j}\beta_{-j} - \beta_j E(X_j|X_{-j}))] \\
&= E\big\{[X_j - E(X_j|X_{-j})][\beta_j(X_j - E(X_j|X_{-j})) + \epsilon]\big\} \\
&= \beta_j E(\tilde{X}_j^2);
\end{aligned}$$

$$
\begin{aligned}
Var(\psi_{0,j}^{(gcm)}) &= E(\psi_{0,j}^{(gcm)^2}) - (E(\psi_{0,j}^{(gcm)}))^2 \\
&= E\big\{[X_j - E(X_j|X_{-j})]^2[\beta_j(X_j - E(X_j|X_{-j})) + \epsilon]^2\big\} - \beta_j^2 Var^2(X_j|X_{-j}) \\
&= \beta_j^2 Var(\tilde{X_j}^2) + \sigma^2 E(\tilde{X}_j^2).
\end{aligned}
$$

LOCO test statistics In general,

$$
\begin{aligned}
E(\psi_{0,j}^{(loco)}) &= E[Y - g_{0,j}(X_{-j})]^2 - E[Y - f_0(X)]^2 \\
&= E[(f_0(X) + \epsilon - g_{0,j}(X_{-j}))^2 - (f_0(X) + \epsilon - f_0(X))^2] \\
&= E[(f_0(X) - g_{0,j}(X_{-j}))^2].
\end{aligned}
$$

Under the setting of multivariate linear regression model, we have

$$
\begin{aligned}
f_0(X) &= E(Y|X) = X_{-j}\beta_{-j} + \beta_j X_j \\
g_{0,j}(X_{-j}) &= E(Y|X_{-j}) = X_{-j}\beta_{-j} + \beta_j E(X_j|X_{-j})
\end{aligned}
$$

Then, $E[\psi_{0,j}^{(loco)}] = E[(f_0(X) - g_{0,j}(X_{-j}))^2] = \beta_j^2 E(X_j - E(X_j|X_{-j}))^2 = \beta_j^2 E(\tilde{X}_j^2)$;

$$
\begin{aligned}
Var(\psi_{0,j}^{(loco)}) &= Var[(f_0(X) - g_{0,j}(X_{-j}))^2 + 2\epsilon(f_0(X) - g_{0,j}(X_{-j})))] \\
&= E[(f_0(X) - g_{0,j}(X_{-j}))^2 + 2\epsilon(f_0(X) - g_{0,j}(X_{-j})))]^2 - \beta_j^4 E^2(\tilde{X}_j^2) \\
&= \beta_j^2\big\{\beta_j^2 E[(X_j - E(X_j|X_{-j}))^4] + 4\sigma^2 Var(X_j|X_{-j}) - \beta_j^2 E^2(\tilde{X}_j^2)\big\} \\
&= \beta_j^4 Var(\tilde{X_j}^2) + 4\sigma^2\beta_j^2 E(\tilde{X}_j^2).
\end{aligned}
$$

Dropout test statistics For Dropout estimator, let's recall $X^{(j)}$ is the j^{th} column replaced by its marginal mean μ_j. In general,

$$\begin{aligned} E(\psi_{0,j}^{(dr)}) &= V(f_0, P_{0,-j}) - V(f_0, P_0) \\ &= E[Y - f_0(X^{(j)})]^2 - E_0[Y - f_0(X)]^2 \\ &= E[(f_0(X) - f_0(X^{(j)}))^2 + 2\epsilon(f_0(X) - f_0(X^{(j)}))] \\ &= E[(f_0(X) - f_0(X^{(j)}))^2]; \end{aligned}$$

Under the setting of multivariate linear regression model, we have

$$\begin{aligned} f_0(X) &= E(Y|X) = X_{-j}\beta_{-j} + \beta_j X_j, \\ f_0(X^{(j)}) &= X_{-j}\beta_{-j} + \beta_j \mu_j. \end{aligned}$$

Then, $E[\psi_{0,j}^{(dr)}] = E[(f_0(X) - f_0(X^{(j)}))^2] = \beta_j^2 E((X_j - \mu_j)^2) = \beta_j^2 E(\hat{X}_j^2)$, where $\hat{X}_j := X_j - E(X_j)$;

$$\begin{aligned} Var(\psi_{0,j}^{(dr)}) &= Var[(f_0(X) - f_0(X^{(j)}))^2 + 2\epsilon(f_0(X) - f_0(X^{(j)}))] \\ &= E[(f_0(X) - f_0(X^{(j)}))^2 + 2\epsilon(f_0(X) - f_0(X^{(j)}))]^2 - \beta_j^4 E^2(\hat{X}_j^2) \\ &= \beta_j^2 \{\beta_j^2 E[(X_j - \mu_j)^4] + 4\sigma^2 E(\hat{X}_j^2) - \beta_j^2 E^2(\hat{X}_j^2)\} \\ &= \beta_j^4 Var(\hat{X}_j{}^2) + 4\sigma^2 \beta_j^2 E(\hat{X}_j^2). \end{aligned}$$

S2.2 Non-linear Additive Model

In this section, we compute all the expressions of test statistics under non-linear regression system. Recall the non-linear additive model $Y = f_0(X) + \epsilon = f_{0,-j}(X_{-j}) + f_{0,j}(X_j) + \epsilon$ where ϵ are i.i.d mean zero with finite variance σ^2, and $\epsilon \perp X$. Let $h_{0,j}(X_{-j}) = E(X_j|X_{-j}), g_{0,j}(X_{-j}) = E(Y|X_{-j})$, and $\tilde{f}_{0,j}(X_j) := f_{0,j}(X_j) - E(f_{0,j}(X_j)|X_{-j})$.

GCM test statistics

$$\begin{aligned}
E(\psi_{0,j}^{(gcm)}) &= E[(X_j - E(X_j|X_{-j}))(Y - E(Y|X_{-j}))] \\
&= E[(X_j - E(X_j|X_{-j}))(f_{0,-j}(X_{-j}) + f_{0,j}(X_j) + \epsilon - f_{0,-j}(X_{-j}) \\
&\qquad\qquad - E(f_{0,j}(X_j)|X_{-j}))] \\
&= E\big\{[X_j - E(X_j|X_{-j})][f_{0,j}(X_j) - E(f_{0,j}(X_j)|X_{-j}) + \epsilon]\big\} \\
&= E(\tilde{X}_j \tilde{f}_{0,j}(X_j));
\end{aligned}$$

$$\begin{aligned}
Var(\psi_{0,j}^{(gcm)}) &= E(\psi_{0,j}^{(gcm)^2}) - (E(\psi_{0,j}^{(gcm)}))^2 \\
&= E\big\{[X_j - E(X_j|X_{-j})]^2[f_{0,j}(X_j) - E(f_{0,j}(X_j)|X_{-j}) + \epsilon]^2\big\} \\
&\qquad - \Big\{E\big\{[X_j - E(X_j|X_{-j})][f_{0,j}(X_j) - E(f_{0,j}(X_j)|X_{-j})]\big\}\Big\}^2 \\
&= Var(\tilde{X}_j \tilde{f}_{0,j}(X_j)) + \sigma^2 E(\tilde{X}_j^2).
\end{aligned}$$

LOCO test statistics Let $f_{0,j}(X) = E(Y|X) = f_{0,-j}(X_{-j}) + f_{0,j}(X_j)$, and $g_{0,j}(X_{-j}) = E(Y|X_{-j}) = f_{0,-j}(X_{-j}) + E(f_{0,j}(X_j)|X_{-j})$. Then, $E(\psi_{0,j}^{(loco)}) = E(f_{0,j}(X_j) - E(f_{0,j}(X_j)|X_{-j}))^2 = E(\tilde{f}_{0,j}^2(X_j))$. By the same computation process in linear system, we have $Var(\psi_{0,j}^{(loco)}) = Var[(f_0(X) - g_{0,j}(X_{-j}))^2 + 2\epsilon(f_0(X) - g_{0,j}(X_{-j}))] = Var(\tilde{f}_{0,j}^2(X_j)) + 4\sigma^2 E(\tilde{f}_{0,j}^2(X_j))$.

Dropout test statistics

$$\begin{aligned} E(\psi_{0,j}^{(dr)}) &= E[(f_0(X) - f_0(X^{(j)}))^2] \\ &= E[(f_{0,j}(X_j) - f_{0,j}(\mu_j))^2] \\ &= E[(f_{0,j}(X_j) - f_{0,j}(\mu_j) - E(f_{0,j}(X_j)) + E(f_{0,j}(X_j))^2] \\ &= Var(f_{0,j}(X_j)) + (E(f_{0,j}(X_j)) - f_{0,j}(\mu_j))^2; \end{aligned}$$

$$\begin{aligned} Var(\psi_{0,j}^{(dr)}) &= Var[(Y - f_0(X^{(j)}))^2 - (Y - f_0(X))^2] \\ &= Var[(f_0(X) - f_0(X^{(j)}))^2 + 2\epsilon(f_0(X) - f_0(X^{(j)}))] \\ &= E[(f_{0,j}(X_j) - f_{0,j}(\mu_j))^4] + 4\sigma^2[Var(f_{0,j}(X_j)) + (E(f_{0,j}(X_j)) - f_{0,j}(\mu_j))^2] \\ &\qquad - [Var(f_{0,j}(X_j)) + (E(f_{0,j}(X_j)) - f_{0,j}(\mu_j))^2]^2. \end{aligned}$$

Expansion of 4^{th} moment:

$$\begin{aligned}E[(f_{0,j}(X_j) - f_{0,j}(\mu_j))^4] &= E\{[f_{0,j}(X_j) - E(f_{0,j}(X_j)) + E(f_{0,j}(X_j)) - f_{0,j}(\mu_j)]^4\} \\ &= E[(f_{0,j}(X_j) - E(f_{0,j}(X_j)))^4] \\ &\quad + 6Var(f_{0,j}(X_j))(E(f_{0,j}(X_j)) - f_{0,j}(\mu_j))^2 + (E(f_{0,j}(X_j)) - f_{0,j}(\mu_j))^4.\end{aligned}$$

Thus,

$$\begin{aligned}Var(\psi_{0,j}^{(dr)}) &= E[(f_{0,j}(X_j) - E(f_{0,j}(X_j)))^4] - Var^2(f_{0,j}(X_j)) + 4\sigma^2 Var(f_{0,j}(X_j)) \\ &\quad + 4(E(f_{0,j}(X_j)) - f_{0,j}(\mu_j))^2(Var(f_{0,j}(X_j)) + \sigma^2) \\ &= Var(\hat{f}_{0,j}^2(X_j)) + 4\sigma^2 E(\hat{f}_{0,j}^2(X_j)) + 4(E(f_{0,j}(X_j)) - f_{0,j}(\mu_j))^2(Var(f_{0,j}(X_j)) + \sigma^2)\end{aligned}$$

where $\hat{f}_{0,j}(X_j) := f_j(X_j) - E(f_j(X_j))$.

S2.3 Single Index Model

In this section, we will compute all the expressions of test statistics in single index model which is a semi-parametric extension of linear model. Recall the single index model $Y = \eta(X^T\beta) + \epsilon$ where β 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 $\epsilon \perp X$.

Suppose under the conditions in Theorem 5 it is sufficient to approximate η via linear Taylor series expansion and the remainder term Δ involv-

ing higher order terms in the expansion is negligible. That is to say that $\Delta, E(\Delta|X_{-j}) = o(1)$, and we will prove it later in Appendix S3.3. Thus, the test statistics derivation procedure is similar to linear model section. Specifically, let's recall the Taylor expansion of $\eta(X^T\beta)$ around $E(X^T\beta|X_{-j})$, and notation $\eta' = \eta'(E(X^T\beta|X_{-j}))$.

$$
\begin{aligned}
f_0(X) = E(Y|X) &= \eta(E(X^T\beta|X_{-j})) + \eta' \cdot (X^T\beta - E(X^T\beta|X_{-j})) + \Delta \\
&= \eta(E(X^T\beta|X_{-j})) + \eta' \cdot \beta_j(X_j - E(X_j|X_{-j})) + \Delta \\
&\text{where } \Delta = \sum_{s=2}^{\infty} \frac{1}{s!}\eta^{(s)}(E(X^T\beta|X_{-j}))(X^T\beta - E(X^T\beta|X_{-j}))^s.
\end{aligned}
$$

GCM test statistics

$$
\begin{aligned}
E(\psi_{0,j}^{(gcm)}) &= E[(X_j - E(X_j|X_{-j}))(Y - E(Y|X_{-j}))] \\
&= E[(X_j - E(X_j|X_{-j}))(\eta(X^T\beta) + \epsilon - \eta(E(X^T\beta|X_{-j}))) - E(\Delta|X_{-j})] \\
&= E[(X_j - E(X_j|X_{-j}))(\eta'\beta_j(X_j - E(X_j|X_{-j})) + \Delta - E(\Delta|X_{-j})] \\
&= \beta_j E(\tilde{X}_j^2)\eta' + E[(X_j - E(X_j|X_{-j}))o(1)] \\
&= \beta_j E(\tilde{X}_j^2)\eta' + o(1);
\end{aligned}
$$

$$
\begin{aligned}
Var(\psi_{0,j}^{(gcm)}) &= E(\psi_{0,j}^{(gcm)^2}) - (E(\psi_{0,j}^{(gcm)}))^2 \\
&= E[\tilde{X}_j^2(\eta(X^T\beta) + \epsilon - \eta(E(X^T\beta|X_{-j}) - E(\Delta|X_{-j}]))^2] - (\beta_j E(\tilde{X}_j^2)\eta' + o(1))^2 \\
&= E[\tilde{X}_j^2(\eta'\beta_j\tilde{X}_j + \epsilon + o(1))^2] - \beta_j^2 E^2(\tilde{X}_j^2)\eta'^2 + o(1) \\
&= \eta'^2\beta_j^2 E(\tilde{X}_j^4) + [\sigma^2 + o(1)]E(\tilde{X}_j^2) + 2o(1)\eta'\beta_j E(\tilde{X}_j^3) - \beta_j^2 E^2(\tilde{X}_j^2)\eta'^2 + o(1) \\
&= \eta'^2\beta_j^2 Var(\tilde{X}_j^2) + \sigma^2 E(\tilde{X}_j^2) + o(1).
\end{aligned}
$$

LOCO test statistics

$$
\begin{aligned}
E[\psi_{0,j}^{(loco)}] &= E[(E(Y|X) - E(Y|X_{-j}))^2] \\
&= E[(\eta'\beta_j(X_j - E(X_j|X_{-j})) + \Delta - E(\Delta|X_{-j}))^2] \\
&= \beta_j^2 E(\tilde{X}_j^2)\eta'^2 + o(1);
\end{aligned}
$$

$$
\begin{aligned}
Var[\psi_{0,j}^{(loco)}] &= Var[(E(Y|X) - E(Y|X_{-j}))^2 + 2\epsilon(E(Y|X) - E(Y|X_{-j})))] \\
&= \beta_j^4 E(\tilde{X}_j^4)\eta'^4 + 4\sigma^2\beta_j^2 E(\tilde{X}_j^2)\eta'^2 - \beta_j^4 E^2(\tilde{X}_j^2)\eta'^4 + o(1) \\
&= \beta_j^4 Var(\tilde{X}_j^2)\eta'^4 + 4\sigma^2\beta_j^2 E(\tilde{X}_j^2)\eta'^2 + o(1).
\end{aligned}
$$

Dropout test statistics Under the single index model, let's make estimation of $f_0(X^{(j)})$ also via first order Taylor series expansion around $E(X^T\beta|X_{-j})$:

$$f_0(X^{(j)}) = \eta(\beta_j\mu_j + X_{-j}^T\beta_{-j}) = \eta(E(X^T\beta|X_{-j})) + \eta' \cdot \beta_j(\mu_j - E(X_j|X_{-j})) + \Delta,$$

$$\begin{aligned}
E[\psi_{0,j}^{(dr)}] &= E[(f_0(X) - f_0(X^{(j)}))^2] \\
&= \beta_j^2\eta'^2 E((X_j - \mu_j)^2) + o(1) \\
&= \beta_j^2\eta'^2 E(\hat{X}_j^2) + o(1); \\
Var(\psi_{0,j}^{(dr)}) &= Var[(f_0(X) - f_0(X^{(j)}))^2 + 2\epsilon(f_0(X) - f_0(X^{(j)}))] \\
&= \beta_j^2\eta'^2\left\{\beta_j^2\eta'^2 E[(X_j - \mu_j)^4] + 4\sigma^2 E(\hat{X}_j^2) - \beta_j^2\eta'^2 E^2(\hat{X}_j^2)\right\} + o(1) \\
&= \beta_j^4 Var(\hat{X}_j^{\,2})\eta'^4 + 4\sigma^2\beta_j^2 E(\hat{X}_j^2)\eta'^2 + o(1).
\end{aligned}$$

S3 Proofs of Theorems

S3.1 Proof of Theorem 3

Since the estimators of GCM and LOCO have different means: one is the product of residuals between models, and the other is the difference of MSE between models. Thus, we use the so-called *relative variability* σ/μ to compare the asymptotic relative efficiency.

$$
\begin{aligned}
e_{\hat{\psi}_{0,j}^{(gcm)},\hat{\psi}_{0,j}^{(loco)}} &= \left(\frac{sd(\hat{\psi}_{0,j}^{(loco)})/E(\hat{\psi}_{0,j}^{(loco)})}{sd(\hat{\psi}_{0,j}^{(gcm)})/E(\hat{\psi}_{0,j}^{(gcm)})}\right)^2 \\
&= \left(\frac{sd(\hat{\psi}_{0,j}^{(loco)})}{\beta_j sd(\hat{\psi}_{0,j}^{(gcm)})}\right)^2 \\
&= \frac{\beta_j^2\big\{\beta_j^2 E[(X_j - E(X_j|X_{-j}))^4] + 4\sigma^2 Var(X_j|X_{-j}) - \beta_j^2 Var^2(X_j|X_{-j})\big\}}{\beta_j^2\{\beta_j^2 E[X_j - E(X_j|X_{-j})]^4 + \sigma^2 Var(X_j|X_{-j}) - \beta_j^2 Var^2(X_j|X_{-j})\}} \\
&= \frac{\beta_j^4 Var(\tilde{X_j}^2) + 4\sigma^2\beta_j^2 E(\tilde{X}_j^2)}{\beta_j^4 Var(\tilde{X_j}^2) + \sigma^2\beta_j^2 E(\tilde{X}_j^2)} > 1.
\end{aligned}
$$

S3.2 Proof of Theorem 4

$$
\begin{aligned}
e_{\hat{\psi}_{0,j}^{(gcm)},\hat{\psi}_{0,j}^{(loco)}} &= \left(\frac{sd(\hat{\psi}_{0,j}^{(loco)})/E(\hat{\psi}_{0,j}^{(loco)})}{sd(\hat{\psi}_{0,j}^{(gcm)})/E(\hat{\psi}_{0,j}^{(gcm)})}\right)^2 \\
&= \left(\frac{sd(\hat{\psi}_{0,j}^{(loco)})}{\frac{E[\tilde{f}_{0,j}(X_j)^2]}{E[\tilde{X}_j\tilde{f}_{0,j}(X_j)]} sd(\hat{\psi}_{0,j}^{(gcm)})}\right)^2, \text{ replace } \tilde{f}_{0,j}(X_j) \text{ with notation } \tilde{f}_{0,j} \\
&= \frac{Var(\hat{\psi}_{0,j}^{(loco)})}{\frac{1}{Corr^2(\tilde{X}_j,\tilde{f}_{0,j})}\frac{E[\tilde{f}_{0,j}^2]}{E[\tilde{X}_j^2]} Var(\hat{\psi}_{0,j}^{(gcm)})}, \text{ replace } Corr^2(\tilde{X}_j,\tilde{f}_{0,j}) \text{ with notation } \rho^2_{\tilde{X}_j,\tilde{f}_{0,j}} \\
&= \frac{E(\tilde{f}_{0,j}^4) + 4\sigma^2 E(\tilde{f}_{0,j}^2) - E^2(\tilde{f}_{0,j}^2)}{\big\{E[\tilde{X}_j^2\tilde{f}_{0,j}^2] + \sigma^2 E(\tilde{X}_j^2) - E^2(\tilde{X}_j\tilde{f}_{0,j})\big\}\frac{1}{\rho^2_{\tilde{X}_j,\tilde{f}_{0,j}}}\frac{E(\tilde{f}_{0,j}^2)}{E(\tilde{X_j}^2)}} \\
&= \frac{E(\tilde{f}_{0,j}^4)\rho^2_{\tilde{X}_j,\tilde{f}_{0,j}} + 4\sigma^2\rho^2_{\tilde{X}_j,\tilde{f}_{0,j}} E(\tilde{f}_{0,j}^2) - \frac{E^2(\tilde{X}_j\tilde{f}_{0,j})}{E(\tilde{X}_j^2)} E^2(\tilde{f}_{0,j}^2)}{E[\tilde{X}_j^2\tilde{f}_{0,j}^2]\frac{E(\tilde{f}_{0,j}^2)}{E(\tilde{X}_j^2)} + \sigma^2 E(\tilde{f}_{0,j}^2) - \frac{E^2(\tilde{X}_j\tilde{f}_{0,j})}{E(\tilde{X}_j^2)} E^2(\tilde{f}_{0,j}^2)} > 1, \text{ if assumption (A) holds.}
\end{aligned}
$$

S3.3 Proof of Theorem 5

In this section, we first prove that the remainder term Δ in Taylor series expansion is negligible, so it is sufficient to apply linear approximation. Then we prove that the GCM estimator is asymptotically more efficient than LOCO.

By assumption (B1) that $\eta^{(s)} \leq O(a^{-X^T\beta})$,$a > 1$ for all $s \geq 2$, assumption (B2) that X has sub-Gaussian distribution with variance proxy σ^2, and assumption (B3) which makes β_j to be finite, we can bound the higher order derivatives and polynomial terms in Δ. For any $\delta > 0$, with probability at least $1 - \delta$:

$$\begin{aligned}
\Delta &= \sum_{s=2}^{\infty} \frac{1}{s!} g^{(s)}(E(X^T\beta|X_{-j}))(X^T\beta - E(X^T\beta|X_{-j}))^s \\
&\leq \sum_{s=2}^{\infty} \frac{O(a^{-E(X^T\beta|X_{-j})})\beta_j}{s!}(X_j - E(X_j|X_{-j}))^s \\
&\leq \sum_{s=2}^{\infty} \frac{\beta_j}{C' a^{E(X^T\beta|X_{-j})} s!} \sigma^s \left(\sqrt{2\log\left(\frac{2}{\delta}\right)}\right)^s, \text{ for some constant } C' \\
&\leq \frac{\beta_j}{C' a^{E(X^T\beta|X_{-j})}} e^{\sigma\sqrt{2\log(\frac{2}{\delta})}} \\
&= o(1)
\end{aligned}$$

The last inequality holds due to the exponential series that for any $s \geq 2, \sum_{s=2}^{\infty} \frac{t^s}{s!} = e^t - 1 - t \leq e^t$. Finally, the remainder term vanishes if there

exists a constant C, such that $E(X^T\beta|X_{-j}) \geq C\sqrt{\log(\frac{1}{\delta})}$, then $\Delta \leq \epsilon$ for any $\epsilon > 0$. Besides, we also have that

$$\begin{aligned}
E(\Delta|X_{-j}) &= E[\sum_{s=2}^{\infty} \frac{1}{s!} g^{(s)}(E(X^T\beta|X_{-j}))(X^T\beta - E(X^T\beta|X_{-j}))^s|X_{-j}] \\
&\leq \sum_{s=2}^{\infty} \frac{O(a^{-E(X^T\beta|X_{-j})})\beta_j}{s!} E[(X_j - E(X_j|X_{-j}))^s|X_{-j}] \\
&\leq \sum_{s=2}^{\infty} \frac{\beta_j}{C'a^{E(X^T\beta|X_{-j})}s!} (\sigma^2)^{\frac{s}{2}} s\Gamma(s/2), \text{ for some constant } C' \\
&\leq \sum_{s=2}^{\infty} \frac{\beta_j}{C'a^{E(X^T\beta|X_{-j})}s!} \sigma^s s(s/2)^{s/2} \\
&\leq \sum_{s=2}^{\infty} \frac{\beta_j}{C'a^{E(X^T\beta|X_{-j})}s!} \sigma^s K^s, \text{ for some constant } K \geq 3 \\
&\leq \frac{\beta_j}{C'a^{E(X^T\beta|X_{-j})}} e^{\sigma\sqrt{K}} \\
&= o(1)
\end{aligned}$$

Similarly, $E(\Delta|X_{-j})$ is negligible if there exists a constant C, such that $E(X^T\beta|X_{-j}) \geq C\sqrt{K}$, then $E(\Delta|X_{-j}) < \epsilon$. Thus, as stated in assumption B4), if there exists a constant C, such that $E(X^T\beta|X_{-j}) \geq C\max\{\sqrt{\frac{1}{\delta}}, \sqrt{K}\}$, where the upper bound $K \geq 3$, then $\Delta, E(\Delta|X_{-j}) = o(1)$. Next, with the estimator expressions derived in Supplementary Material S2.3, the same relative efficiency comparison technique, and monotonicity of η in assumption

(B1), we have that

$$
\begin{aligned}
e_{\hat{\psi}_{0,j}^{(gcm)},\hat{\psi}_{0,j}^{(loco)}} &= \left(\frac{sd(\hat{\psi}_{0,j}^{(loco)})/E(\hat{\psi}_{0,j}^{(loco)})}{sd(\hat{\psi}_{0,j}^{(gcm)})/E(\hat{\psi}_{0,j}^{(gcm)})}\right)^2 \\
&= \left(\frac{sd(\hat{\psi}_{0,j}^{(loco)})}{\beta_j\eta' sd(\hat{\psi}_{0,j}^{(gcm)})}\right)^2, \text{ recall notation } \eta' = \eta'(E(X^T\beta|X_{-j})) \\
&= \frac{\beta_j^2\eta'^2\big\{\beta_j^2\eta'^2 Var(\tilde{X}_j^2)) + 4\sigma^2 E(\tilde{X}_j^2)\big\} + o(1)}{\beta_j^2\eta'^2\{\beta_j^2\eta'^2 Var(\tilde{X}_j^2) + \sigma^2 E(\tilde{X}_j^2)\} + o(1)} \\
&= \frac{\beta_j^2\eta'^2\big\{\beta_j^2\eta'^2 Var(\tilde{X}_j^2)) + 4\sigma^2 E(\tilde{X}_j^2)\big\}}{\beta_j^2\eta'^2\{\beta_j^2\eta'^2 Var(\tilde{X}_j^2) + \sigma^2 E(\tilde{X}_j^2)\}}\Big(1 + o(1)\Big)
\end{aligned}
$$

Thus, if assumption (B1)-(B4) hold, then for any $\epsilon' > 0, \delta > 0$, with probability at least $1-\delta$, we have $e_{\hat{\psi}_{0,j}^{(gcm)},\hat{\psi}_{0,j}^{(loco)}} > 1-\epsilon'$.

Remark 1. *Note that the expressions in GCM and LOCO test statistics involve the estimate of η'. If η is fixed and known, then the result can be directly computed. If η is fixed but unknown, we can employ the Nadaraya-Watson kernel-weighted average estimation method proposed by Foster et al. (2013) (see details in section 2.2) where η' is estimated from the model $\tilde{y}_i = \eta'(\beta^T\tilde{x}_i) + \tilde{\epsilon}_i$ with $\tilde{x}_i = x_i + \frac{x_{i+1}-x_i}{2}$, $\tilde{y}_i = \frac{\eta(\beta_0 x_{i+1}) - \eta(\beta_0^T x_i)}{\beta_0 x_{i+1} - \beta_0^T x_i}$, and known $\beta_0^T X$ (in our case, it is $E(\beta^T X|X_{-j})$).*

S3.4 Proof of Theorem 7

Under the regularity conditions of neural network required in Theorem 4.4 of Gao et al. (2022), the asymptotic variance of the Lazy-VI estimator

$\hat{\psi}_{0,j}^{(lazy)}$ is the same as that of LOCO estimator $\hat{\psi}_{0,j}^{(loco)}$, which is $\frac{1}{n}Var(\epsilon^2 - \epsilon_{-j}^2)$, see reformulated Theorem 6 in main paper.

Next, let compare the asymptotic variance relationship of LOCO and dropout estimators. Under assumptions in Theorem 3, for $j \in [p]$, we have

$$\frac{Var(\hat{\psi}_{0,j}^{(dr)})}{Var(\hat{\psi}_{0,j}^{(loco)})} = \frac{\beta_j^4 Var(\hat{X}_j^{\,2}) + 4\sigma^2\beta_j^2 E(\hat{X}_j^2)}{\beta_j^4 Var(\tilde{X}_j^{\,2}) + 4\sigma^2\beta_j^2 E(\tilde{X}_j^2)} \geq 1.$$

The inequality holds since the unconditional variance of the dropout estimator is greater than or equal to the conditional variance of the LOCO estimator. This follows from the principle that unconditional moments are greater than or equal to their corresponding conditional moments which is implied by Jensen's inequality for convex functions, such as variance.

Under assumptions in Theorem 4 , for all $j \in [p]$, similarly we have

$$\frac{Var(\hat{\psi}_{0,j}^{(dr)})}{Var(\hat{\psi}_{0,j}^{(loco)})} = \frac{Var(\hat{f}_{0,j}^2(X_j)) + 4\sigma^2 E(\hat{f}_{0,j}^2(X_j)) + 4(E(f_{0,j}(X_j)) - f_{0,j}(\mu_j))^2(Var(f_{0,j}(X_j)) + \sigma^2)}{Var(\tilde{f}_{0,j}^2(X_j)) + 4\sigma^2 Var(\tilde{f}_{0,j}(X_j))} \geq 1.$$

The inequality holds due to the extra non-negative term and the replacement of the unconditional moments for the variance of dropout estimator.

Finally, under assumptions in Theorem 5, using the same Taylor series expansion approach as shown in Supplementary Material S3.3, we have $\Delta, E(\Delta|X_{-j}) = o(1)$. Then, with monotonicity of η in assumption (B1), we

have for all $j \in [p]$

$$\begin{aligned}\frac{Var(\hat{\psi}_{0,j}^{(dr)})}{Var(\hat{\psi}_{0,j}^{(loco)})} &= \frac{\beta_j^4 Var(\hat{X}_j^{\,2})\eta'^4 + 4\sigma^2\beta_j^2 E(\hat{X}_j^2)\eta'^2 + o(1)}{\beta_j^4 Var(\tilde{X}_j^2)\eta'^4 + 4\sigma^2\beta_j^2 E(\tilde{X}_j^2)\eta'^2 + o(1)} \\ &= \frac{\beta_j^4 Var(\hat{X}_j^{\,2})\eta'^4 + 4\sigma^2\beta_j^2 E(\hat{X}_j^2)\eta'^2}{\beta_j^4 Var(\tilde{X}_j^2)\eta'^4 + 4\sigma^2\beta_j^2 E(\tilde{X}_j^2)\eta'^2}\Big(1 + o(1)\Big)\end{aligned}$$

If assumption (B1)-(B4) in Theorem 5 hold, then for any $\epsilon' > 0, \delta > 0$, with probability at least $1 - \delta$, we have $e_{\hat{\psi}_{0,j}^{(loco)}, \hat{\psi}_{0,j}^{(dr)}} > 1 - \epsilon'$.

S4 Additional experiment results

In this section, we explore more extra simulation and real data analysis to prove the consistency of our theoretical result that GCM continues to outperform LOCO.

S4.1 Additional simulation

Besides making a comparison for different feature selection methods with model-agnostic algorithm, we initially use canonical implementations for each type of the model. Specifically, we employ linear regression for (a), generalized additive model for (b) and (c), and generalized partially linear single index model for (d) (`stats`, Bolar (2019) ; `gam`, Hastie and Narasimhan (2024); `gplsim`, Zu and Yu (2023) package respectively in `R`).

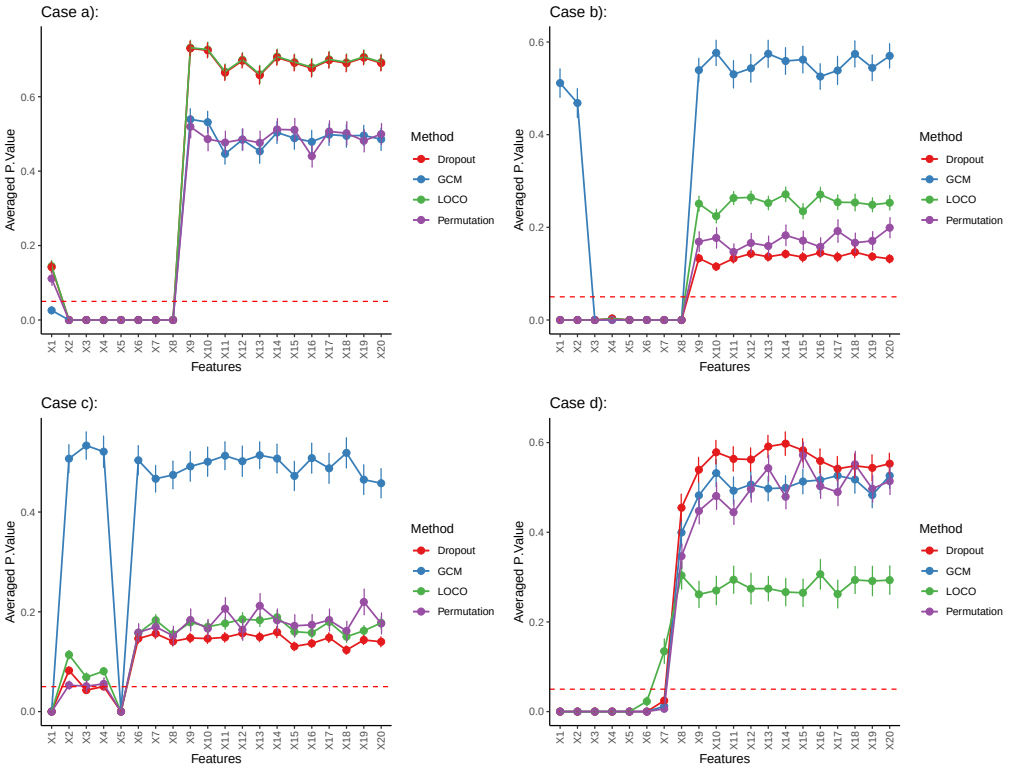

Figure 1: Averaged p-value with standard deviation of each feature for model (a) - (d) for GCM, LOCO, dropout and permutation univariate selection on dataset using linear regression for (a), generalized additive model for (b) and (c), and generalized partially linear single index model for (d). Red dashed line: targeted type-I error rate ($\alpha = 0.05$).

Fig. 1 shows the average p-value for each of the 20 features performed by GCM, LOCO, dropout and permutation variable selection.

In Fig. 1, plot (a) shows that all approaches work well in the linear model case. GCM demonstrates an advantage by accurately selecting variables with relatively small regression coefficients, highlighting its sensitivity

to subtle but important relationships in the linear model. The performance of dropout, permutation and LOCO are almost the same due to the independence structure in model (a), and the moderate correlation between X_3 and X_4 does not make a large difference in p-values. For plot (b), as discussed in Example 3.2 (main context) and Example 1 (Supplementary material S1), GCM successfully captures all other features except the first two in the non-linear additive model (b) because the quadratic and cosine functions are even functions, which violates assumption (A). None of the methods work well when the true model involves interaction terms as shown in plot (c). Overall, all methods have similar performance when the standard algorithm is selected for the true model, and GCM performs slightly better than LOCO as seen in plot (a) and (d).

Next, we also include another popular machine learning algorithm - kernel SVM - as a reference. For case (a) - (b) in Fig. 2, we use gaussian kernel with default setting in `R` since gaussian kernel is most widely used for non-linear models with complex decision boundaries.

We observe that GCM and LOCO show consistent results, whereas dropout and permutation methods, despite their computational efficiency, show a significant loss of type I error control in case (b), achieving high recall but low specificity. The gaussian kernel depends on all features si-

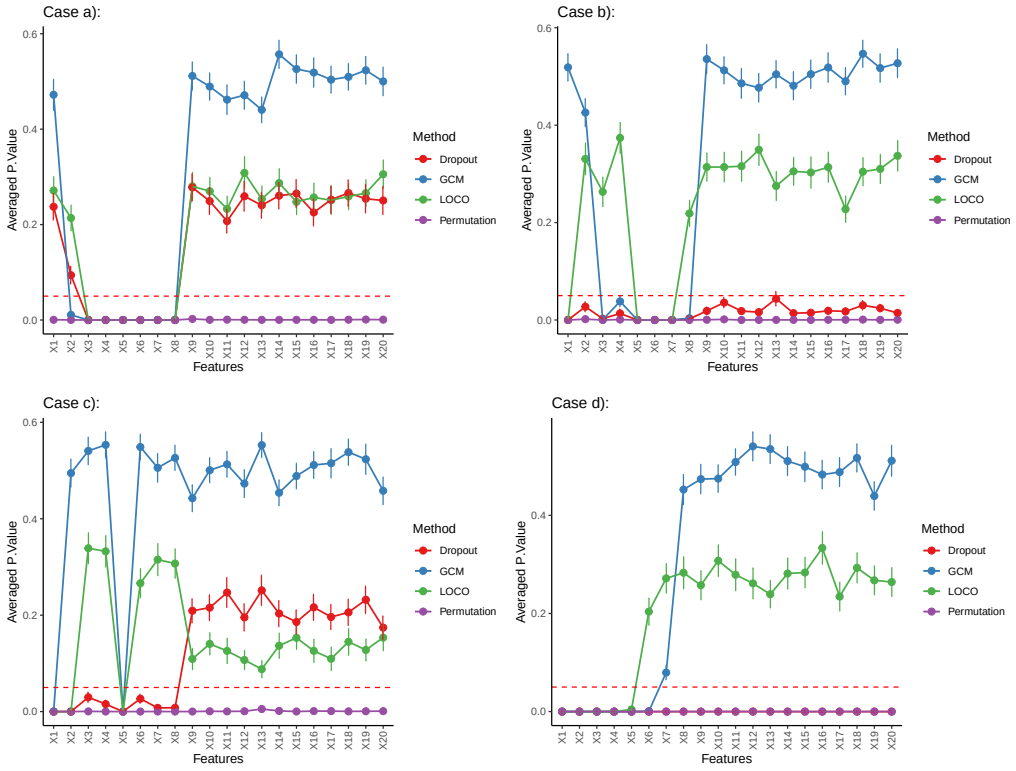

Figure 2: Averaged p-value and standard deviation of each feature for model (a) - (d) across 100 simulations for GCM, LOCO, dropout and permutation univariate selection on dataset with 1000 instances using kernel SVM. Red dashed line: targeted type-I error rate ($\alpha = 0.05$).

multaneously. Without retraining the model, any change to a feature can cause a significant difference in the learned model, making permutation and dropout methods less stable. LOCO requires approximately half the computational time of GCM, but GCM can identify more significant variables than LOCO. GCM outperforms knockoff and lasso methods in the non-

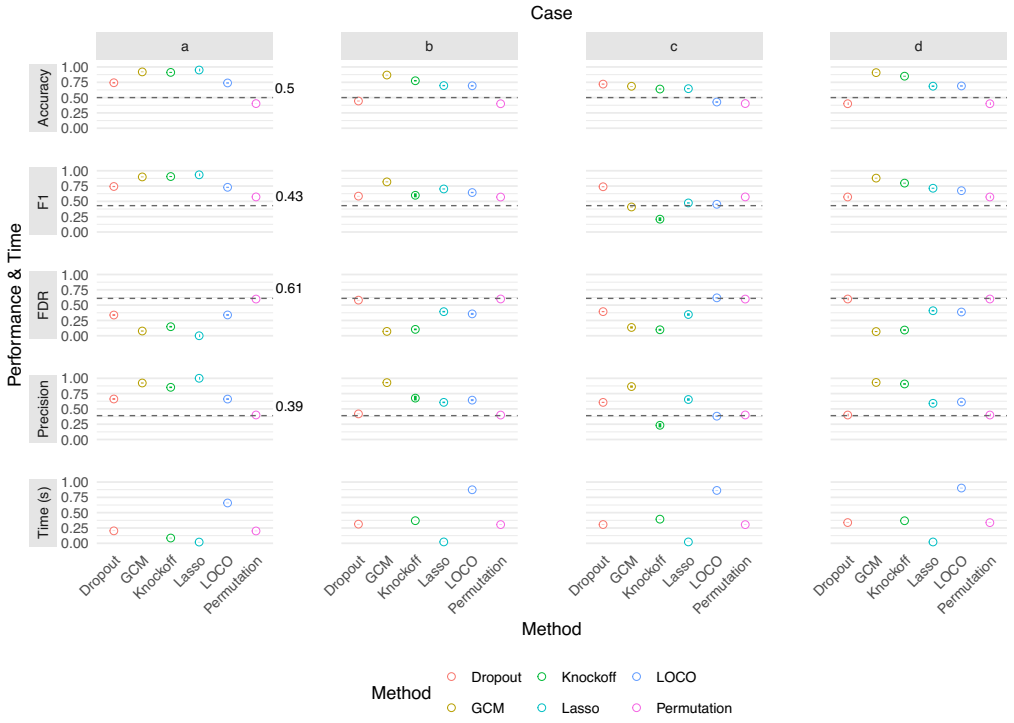

Figure 3: Averaged performance measure and time (seconds) with standard deviation across 100 simulations for lasso, knockoff (target FDR = 0.2), GCM, LOCO, dropout and permutation univariate selection on dataset with 1000 instances using kernel SVM for (a)-(d). Dashed line: empirical random-guessing baseline.

linear model when the underlying model is not sparse (see Fig. 3). In the high-dimensional regression setting, ksvm is implemented with the gaussian kernel. We use the default tuning parameters for setting (a), while for setting (b) we set the cost parameter to $C = 0.95$, and the setting a and b is the same as the main context Section 5.3. From Fig. 4, Knockoff filter method and lasso work well in the sparse models, and GCM still signifi-

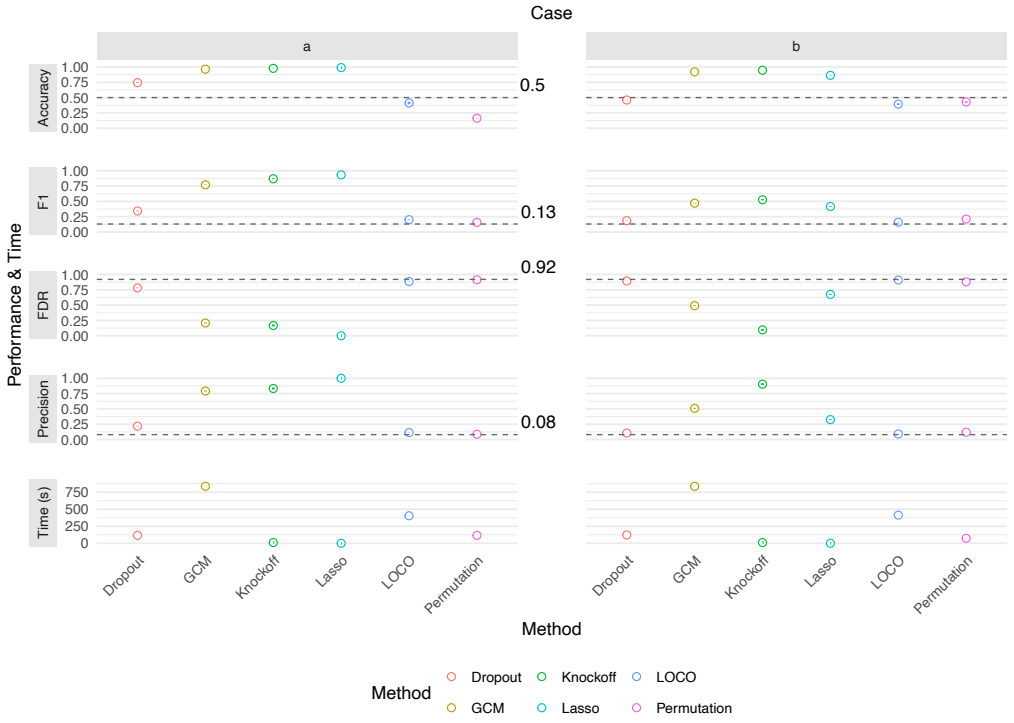

Figure 4: Averaged performance measure and time (seconds) with standard deviation across 50 simulations for lasso, knockoff (target FDR = 0.2), GCM, LOCO, dropout and permutation univariate selection on dataset with 1000 instances and 500 features using kernel SVM for (a) and (b). Dashed line: empirical random-guessing baseline.

cantly outshines other methods. Across different model training algorithms such as canonical implementations for each designated model, GBM, and kernel SVM, GCM consistently outperforms or performs as good as LOCO, indicating that GCM is a more reliable and stable approach.

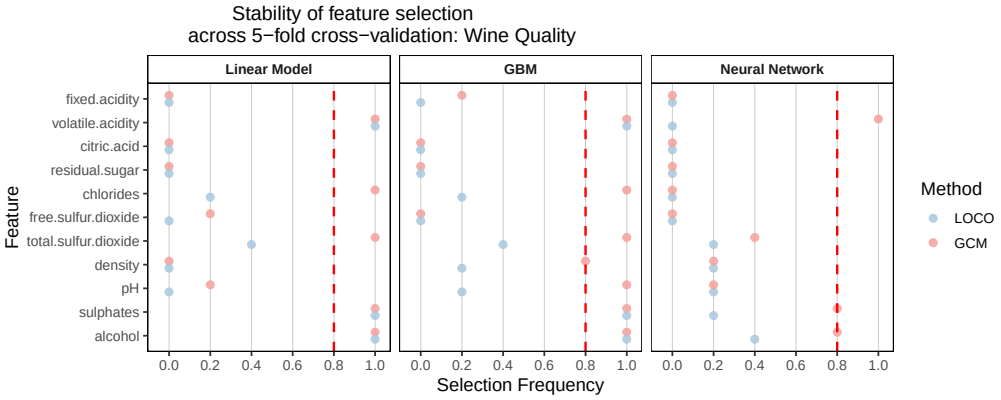

Figure 5: Selection rate of each feature selected by GCM and LOCO across 5-fold cross validation on Wine quality data, using lm, GBM and NN as the base predictor respectively. Red dashed line: targeted selection frequency.

S4.2 Wine quality data

With increasing interest in wine, quality certification has become essential, relying on sensory evaluations by human experts. We aim to predict wine quality based on objective analytical tests conducted during the certification process. The Wine Quality dataset, introduced by Cortez et al. (2009), contains $n = 1599$ samples of red wine, with $p = 11$ features related to physicochemical properties (e.g., `acidity`, `alcohol`). The target variable is the wine quality score, ranging from 0 to 10, which indicates the perceived quality of the wine. For Wine quality data, the GBM model uses the default setting in R. We use a fully connected neural network with hidden layer of width 50, trained with learning rate 5×10^{-4}, weight de-

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

	Models:		
Method	lm	GBM	NN
GCM	0.43 ± 0.015	0.41 ± 0.017	0.44 ± 0.026
LOCO	0.44 ± 0.015	0.42 ± 0.017	0.59 ± 0.066

cay 10^{-5}, and batch size 32. GCM constantly achieves lower MSE than LOCO across all three learners as shown in Table 1. From Fig. 5, we see that GCM shows more stable feature selection than LOCO. In particular, GCM consistently identifies key predictors such as alcohol, sulphates, and volatile acidity, which is the same as prior literature on wine quality analysis (Molnar et al. (2023)).

S4.3 Boston housing Data

Accurate prediction of house prices is crucial, requiring careful consideration of numerous factors that significantly influence property values. We aim to address pricing discrepancies by variable selection to predict real estate prices, providing insights for both buyers and sellers in the future. We analyze the Boston Housing dataset, originally introduced by Harrison and Rubinfeld (1978). This dataset contains $n = 506$ observations, $p = 13$ attributes of homes in Boston suburbs, along with a target variable, `medv`, representing the median value of owner-occupied homes (in thousands of

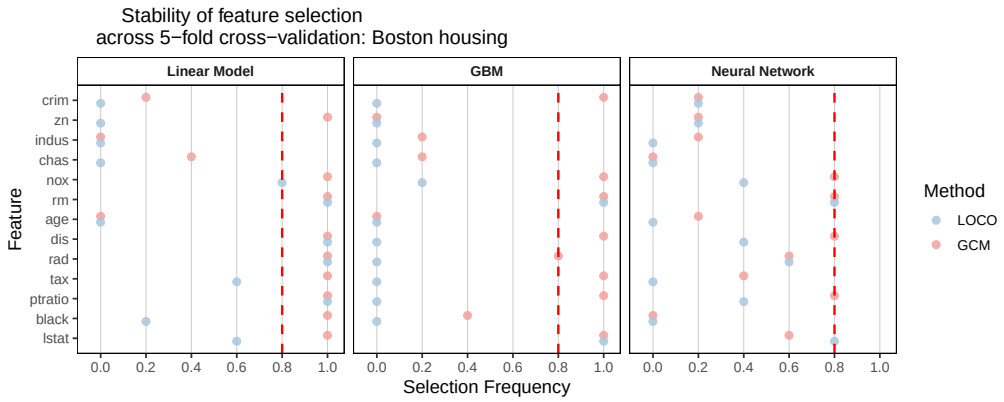

Figure 6: Selection rate of each feature selected by GCM and LOCO across 5-fold cross validation on Boston hosing data, using lm, GBM and NN as the base predictor respectively. Red dashed line: targeted selection frequency.

dollars). For Boston housing data, the GBM model uses the default setting in R. We use a fully connected neural network with hidden layer of width 32, trained with learning rate 5×10^{-4}, weight decay 10^{-3}, and batch size 32. From Fig. 6, we see that the most important features selected by GCM and LOCO differ depending on the machine learning methods employed, but the mean squared prediction error of the models selected by GCM is always lower than that of LOCO (Table 2). It can also be seen that the gradient boosted decision tree algorithm outperforms the others. Due to the robustness of gradient boosted decision tress and its ability to capture nonlinear relationships in the data, we now focus on the features selected by this model. Besides features selected by both LOCO and GCM, which

Table 2: Mean squared prediction error averaged across 5-fold cross-validation for GCM and LOCO on Boston housing data, using lm, GBM, NN respectively.

	Models:		
Method	lm	GBM	NN
GCM	25.79 ± 2.09	15.25 ± 1.79	21.21 ± 3.33
LOCO	32.28 ± 4.71	19.17 ± 2.37	23.13 ± 2.50

indicates the significance of property details (ie: average number of rooms per house, `rm`) and the proportion of population with lower socioeconomic status (ie: `lstat`),other moderately significant features selected by GCM but not LOCO are aligned with previous studies (Harrison and Rubinfeld (1978); Sanyal et al. (2022); Eze et al. (2023)). Those features indicate the significance of proximity to major employment centers (ie: `dis`), property tax rate (ie: `tax`), the education quality (ie: pupil-teacher ratio, `ptratio`), etc. The results align with intuitive understanding: factors related to comfort, education, and economic level are more important when middle-income families purchase a home.

S4.4 Crab age prediction data

Crab age prediction is useful for commercial farming because estimating age from physical attributes helps determine the optimal harvest time and improve profit. The data provided by Sidhu (2021) is available on Kaggle, and contain 3,893 observations with `Age` as the response variable. The orig-

Table 3: Mean squared prediction error averaged across 5-fold cross-validation for GCM and LOCO on Crab Age data, using lm, GBM, NN respectively.

	Models:		
Method	lm	GBM	NN
GCM	4.93 ± 0.16	4.73 ± 0.21	5.82 ±0.52
LOCO	5.03 ± 0.17	4.86±0.21	6.47 ± 0.57

inal predictors include `Sex`, `Length`, `Diameter`, `Height`, `Weight`, `Shucked Weight`, `Viscera Weight`, and `Shell Weight`. `Sex` is a categorical variable which has three levels: male, female, and indeterminate. We use two dummy variables to encode `Sex`, taking female as the reference level. For the Crab Age data, the GBM model uses the default setting in R. The neural network is a fully connected feed-forward model with hidden layer of 30 neurons, learning rate 5×10^{-4}, weight decay 10^{-3}, and batch size 128. The hyperparameter settings are kept consistent across LOCO and GCM. Similarly, GCM achieves lower averaged MSE than LOCO across all three learners as shown in Table 3. From Fig. 7, we see that Weight-related variables, such as `Shucked Weight` and `Shell Weight`, are selected most consistently, suggesting that body-mass measurements are significant predictors of crab age. Besides, GCM appears slightly more stable than LOCO from the selection stability perspective as it has fewer variables with intermediate selection frequencies than LOCO does.

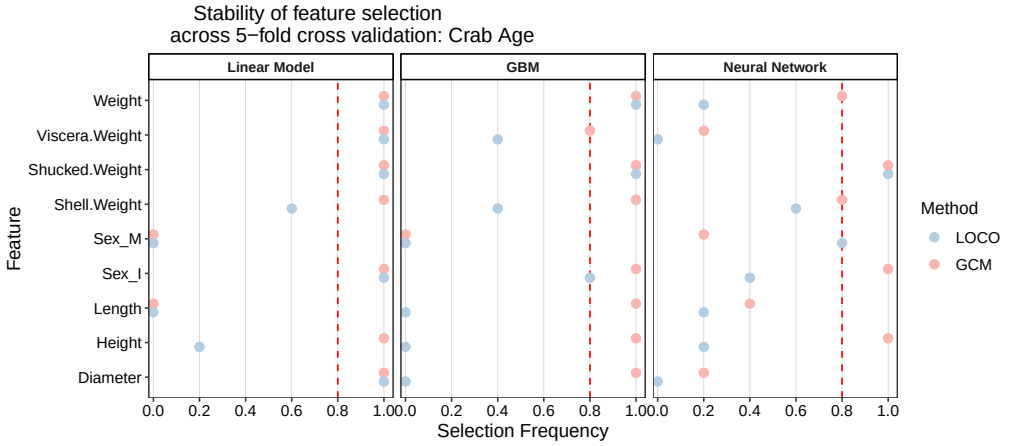

Figure 7: Selection rate of each feature selected by GCM and LOCO across 5-fold cross validation on Crab Age data, using lm, GBM and NN as the base predictor respectively. Red dashed line: targeted selection frequency.

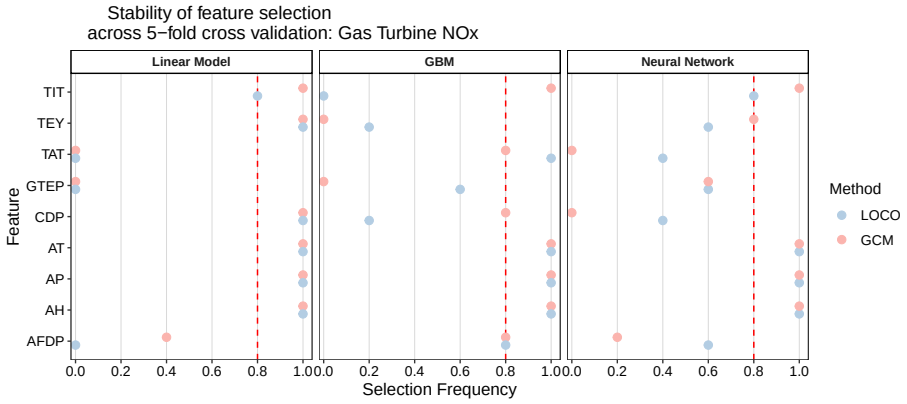

Figure 8: Selection rate of each feature selected by GCM and LOCO across 5-fold cross validation on gas turbine NO_x data, using lm, GBM and NN as the base predictor respectively. Red dashed line: targeted selection frequency.

S4.5 Gas turbine NO_x emission data

Table 4: Mean squared prediction error averaged across 5-fold cross-validation for GCM and LOCO on Gas Turbine NO_x data, using lm, GBM, NN respectively.

	Models:		
Method	lm	GBM	NN
GCM	83.64 ± 1.92	39.69 ± 0.93	38.61 ± 1.27
LOCO	84.61 ± 1.73	41.19 ± 0.73	38.91 ± 3.38

Gas turbines are significant sources of air pollution, with NO_x and CO as major pollutants. Accurate NO_x prediction can support emission monitoring and pollution control. The gas turbine dataset is publicly available in UCI (2013). In this study, we focus on emissions from 2013, which include 7,152 observations, and predict NO_x emissions concentration using nine input variables: ambient temperature, ambient pressure, ambient humidity, air filter difference pressure, gas turbine exhaust pressure, turbine inlet temperature, turbine after temperature, turbine energy yield, and compressor discharge pressure. For the Gas Turbine NO_x data, the GBM model uses the default setting in R. The neural network is a fully connected feed-forward model with hidden layer of 50 neurons, learning rate 10^{-3}, weight decay 10^{-5}, and batch size 32. These learner settings are kept fixed across LOCO and GCM for comparison. From Table 4, we see that GCM consistently achieves the lower averaged MSE across all different machine learning learners. From Fig. 8, we see that GCM is relatively more stable than LOCO , especially for GBM and neural network learners,

with fewer features having intermediate selection frequencies. Both methods frequently select `AH`, `AP`, and `AT`, while GCM tends to select additional operating variables such as `TIT`, `TEY`, and `CDP`. It is aligned with intuitive understanding since thermal NO_x production is sensitive to the operating temperature, load, and pressure conditions of the turbine.

Bibliography

Bolar, K. (2019). STAT: Interactive Document for Working with Basic Statistical Analysis.

Cortez, P., A. Cerdeira, F. Almeida, T. Matos, and J. Reis (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems 47*(4), 547–553.

Eze, E., S. Sujith, M. S. Sharif, and W. Elmedany (2023). A Comparative Study For Predicting House Price Based on Machine Learning. In *2023 4th International Conference on Data Analytics for Business and Industry (ICDABI)*, Bahrain, pp. 75–81. IEEE.

Foster, J. C., J. M. Taylor, and B. Nan (2013). Variable selection in monotone single-index models via the adaptive LASSO. *Statistics in medicine 32*(22), 3944–3954.

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.

Harrison, D. and D. L. Rubinfeld (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management 5*(1), 81–102.

Hastie, T. and B. Narasimhan (2024). gam: Generalized Additive Models.

Molnar, C., T. Freiesleben, G. König, J. Herbinger, T. Reisinger, G. Casalicchio, M. N. Wright, and B. Bischl (2023). Relating the Partial Dependence Plot and Permutation Feature Importance to the Data Generating Process. In *Explainable Artificial Intelligence*, Volume 1901, pp. 456–479. Cham: Springer Nature Switzerland.

Sanyal, S., S. Kumar Biswas, D. Das, M. Chakraborty, and B. Purkayastha (2022). Boston House Price Prediction Using Regression Models. In *2022 2nd International Conference on Intelligent Technologies (CONIT)*, Hubli, India, pp. 1–6. IEEE.

Sidhu, G. S. (2021). Crab age prediction. Kaggle.

UCI (2013). Gas Turbine CO and NOx Emission Data Set . UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5WC95.

Zu, T. and Y. Yu (2023). gplsim: Spline Estimation for GPLSIM.