Supplementary Material

A Background on Samplers

This is the more formal definition of a sampler: Let V and W be two random variables, we define a sampler as a function ϕ that maps an input $v \in \mathcal{V}$ to a density function on a space $\mathcal{W}$ i.e. $\phi : \mathcal{V} \to \{\psi \mid \psi \text{ density on } \mathcal{W}\}$. The two most common samplers in the context of PD and PFI are the marginal and the conditional sampler: the marginal sampler ϕ_{marg} maps every input $v \in \mathcal{V}$ to the density of W i.e. for all $v \in \mathcal{V}: \ \phi_{marg}(v) = \psi_W$; the conditional sampler ϕ_{cond} maps every input $v \in \mathcal{V}$ with $\psi_V(v) > 0$ to the conditional density of W i.e. for all $v \in \mathcal{V}: \ \phi_{cond}(v) = \psi_{W|V=v} = \frac{\psi_{W,V=v}}{\psi_{V=v}}$. As such, samples from $\phi_{marg}(v)$ follow $P(W)$, and samples from $\phi_{cond}(v)$ follow $P(W \mid V = v)$.

Like all model-agnostic interpretation techniques, both PD and PFI are based on sampling data and evaluating the model on these data [52]. Dependent on how we sample, we obtain different versions of PD and PFI and their results must be interpreted in a different way [31,30,18,42,2]. The two most common theoretical samplers in PD and PFI research are the marginal and the conditional sampler. The choice of the sampler should depend on the modeler's objective and the structure of the data. Under certain conditions, the marginal sampler allows to estimate causal effects [46]. However, for correlated input features the marginal sampler may create unrealistic data outside the training distribution, which is problematic if the goal is to draw inference about the DGP; under such conditions, the conditional sampler may be a better choice [19]. Samplers, especially conditional samplers, are generally not readily available, but must be learned with techniques such as conditional subgroups [30] or conditional density estimators [49,48,53,54,55,50]. The learning process of the sampler may introduce another source of uncertainty that we do not consider in this work; we discuss this limitation in Section 5.

B Bias and Variance of PD

The expected squared difference between model-PD and DGP-PD can be decomposed into bias and variance.

Proof.

$$\begin{aligned}
\mathbb{E}_F[(PD_{\hat{f}}(x) - PD_f(x))^2] &= \mathbb{E}_F[PD_{\hat{f}}(x)^2] + \mathbb{E}_F[PD_f(x)^2] \\
&- 2\mathbb{E}_F[PD_{\hat{f}}(x)PD_f(x)] \\
&= \mathbb{V}_F[PD_{\hat{f}}(x)] + \mathbb{E}_F[PD_{\hat{f}}(x)]^2 \\
&+ PD_f(x)^2 - 2\mathbb{E}_F[PD_{\hat{f}}(x)PD_f(x)] \\
&= \underbrace{(PD_f(x) - \mathbb{E}_F[PD_{\hat{f}}(x)])^2}_{\text{Bias}} + \underbrace{\mathbb{V}_F[PD_{\hat{f}}(x)]}_{\text{Variance}}
\end{aligned}$$

C Bias and Variance of PFI

The expected squared difference between model-PFI and DGP-PFI can be decomposed into bias and variance.

Proof.

$$\begin{aligned}
\mathbb{E}_F[(PFI_{\hat{f}} - PFI_f)^2] &= \mathbb{E}_F[PFI_{\hat{f}}^2] + \mathbb{E}_F[PFI_f^2] \\
&- 2\mathbb{E}_F[PFI_{\hat{f}} PFI_f] \\
&= \mathbb{V}_F[PFI_{\hat{f}}] + \mathbb{E}_F[PFI_{\hat{f}}]^2 \\
&+ PFI_f^2 - 2\mathbb{E}_F[PFI_{\hat{f}} PFI_f] \\
&= (PFI_f - \mathbb{E}_F[PFI_{\hat{f}}])^2 + \mathbb{V}_F[PFI_{\hat{f}}] \\
&= Bias_F^2[PFI_{\hat{f}}] + \mathbb{V}_F[PFI_{\hat{f}}]
\end{aligned}$$

D Model-PD Unbiasedness Regarding Theoretical PD

Proof. By the law of large numbers, the Monte Carlo integration converges with $r \to \infty$ to the true integral. Assuming we have a fixed x, r identically distributed random draws $\tilde{X}_C^{(1,x)}, \dots, \tilde{X}_C^{(r,x)} \sim \phi(x)$ and a model $\hat{f}$, the estimate is:

$$\begin{aligned}
\mathbb{E}_{\tilde{X}_C}[\widehat{PD}_{\hat{f}}(x)] &= \mathbb{E}_{\tilde{X}_C^{(1,x)}, \dots, \tilde{X}_C^{(r,x)}} \left[\frac{1}{r} \sum_{i=1}^{r} \hat{f}(x, \tilde{X}_C^{(i,x)}) \right] \\
&= \frac{1}{r} r \mathbb{E}_{\tilde{X}_C}[\hat{f}(x, \tilde{X}_C)] \\
&= PD_{\hat{f}}(x)
\end{aligned}$$

and therefore unbiased for the interval, i.e. the theoretical PD of the model.

E Model-PD Unbiasedness Regarding DGP-PD

Proof. Unbiasedness of the model $\hat{f}$ implies unbiasedness of the model-PD.

$$\begin{aligned}
\mathbb{E}_F[PD_{\hat{f}}(x)] &\overset{Def}{=} \int_F \int_{\tilde{x}_c \in \tilde{\mathcal{X}}_C} \phi(x)(\tilde{x}_c) \hat{f}(x, \tilde{x}_c) \, d\tilde{x}_c \, dP(F) \\
&\overset{Fub}{=} \int_{\tilde{x}_c \in \tilde{\mathcal{X}}_C} \int_F \phi(x)(\tilde{x}_c) \hat{f}(x, \tilde{x}_c) \, dP(F) \, d\tilde{x}_c \\
&\overset{const.}{=} \int_{\tilde{x}_c \in \tilde{\mathcal{X}}_C} \phi(x)(\tilde{x}_c) \int_F \hat{f}(x, \tilde{x}_c) \, dP(F) \, d\tilde{x}_c \\
&\overset{unbiased}{=} \int_{\tilde{x}_c \in \tilde{\mathcal{X}}_C} \phi(x)(\tilde{x}_c) f(x, \tilde{x}_c) \, d\tilde{x}_c \\
&\overset{def}{=} PD_f(x)
\end{aligned}$$

Fubini's theorem requires that $\int_{F,\tilde{X}_C} \mid \phi(x)(\tilde{X}_c)\hat{f}(\tilde{X}_c) \mid d\mathbb{P}_{F,X_C} < \infty$. One sufficient condition for this is when the model predictions have an upper bound $c :\mid \hat{f}(x) \mid < c < \infty$.

F Model-PFI Regarding theoretical PFI

Proof. As a function of random variables, the loss L itself is a random variable. We assume that the loss $L^{(i)}$ of observation i is a sample from the distribution of losses: $L^{(i)} \sim L$ and, similarly for the loss: $\tilde{L}^{(k,i)} \sim \tilde{L}$, where $L^{(i)} = L(y^{(i)}, \hat{f}(x^{(i)}))$ and $\tilde{L}^{(k,i)} = L(y^{(i)}, \hat{f}(\tilde{x}_S^{(k,i)}, x_C^{(i)}))$.

The expectation of our estimator is:

$$
\begin{aligned}
\mathbb{E}_{\tilde{X}_S X_S X_C Y}[\widehat{PFI}_{\hat{f}}] &= \mathbb{E}_{\tilde{X}_S X_S X_C Y}\left[\frac{1}{n_2}\sum_{i=1}^{n_2}(\frac{1}{r}\sum_{k=1}^{r}(\tilde{L}^{(k,i)} - L^{(i)}))\right] \\
&= \frac{1}{n_2} n_2 \mathbb{E}_{\tilde{X}_S X_S X_C Y}[((\frac{1}{r} r\tilde{L}) - L)] \\
&= \mathbb{E}_{\tilde{X}_S X_C Y}[\tilde{L}] - \mathbb{E}_{X_S X_C Y}[L] \\
&= PFI_{\hat{f}}
\end{aligned}
$$

In expectation, we retrieve the theoretical PFI of the model.

G PFI Biases for L2

In this proof, we use the conditional sampler ϕ_{cond} for both, the DGP-PFI and the model-PFI. Moreover, we assume that L is the squared loss $L(y, \hat{f}) = (y - \hat{f}(x))^2$ and that $\mathbb{E}[Y \mid X]$ can be described by f with some additive, irreducible, error ϵ with $\mathbb{E}(\epsilon) = 0$ and $\mathbb{V}(\epsilon) = \sigma^2$. To further examine the bias for the PFI, we apply the Bias-Variance Decomposition additionally on the loss itself: In addition, we use that $\mathbb{E}_{XY}[Y] = \mathbb{E}_X[f(X)]$, $\mathbb{V}_Y[Y] = \sigma^2$ and $\mathbb{E}[A^2] = \mathbb{V}[A] + \mathbb{E}[A]^2$. We first derive the bias-variance decomposition of (i) permuted loss and (ii) original loss and therefrom derive the expected PFI.

For the permuted loss (i):

$$
\begin{aligned}
\mathbb{E}_{F\tilde{X}_S XY}[\tilde{L}] &= \mathbb{E}_{F\tilde{X}_S XY}[(Y - \tilde{\hat{f}})^2] \\
&= \mathbb{E}_{\tilde{X}_S XY}[Y^2 - 2Y\mathbb{E}_F[\tilde{\hat{f}}] + \mathbb{E}_F[\tilde{\hat{f}}^2]] \\
&= \mathbb{E}_{\tilde{X}_S XY}[Y^2 - 2Y\mathbb{E}_F[\tilde{\hat{f}}] + \mathbb{E}_F[\tilde{\hat{f}}]^2 + \mathbb{V}_F[\tilde{\hat{f}}]] \\
&= \mathbb{V}_Y[Y] + \mathbb{E}_{\tilde{X}_S X}[f^2 - 2f\mathbb{E}_F[\tilde{\hat{f}}] + \mathbb{E}_F[\tilde{\tilde{\hat{f}}}]^2 + \mathbb{V}_F[\tilde{\hat{f}}]] \\
&= \underbrace{\sigma^2}_{\text{Data Var}} + \mathbb{E}_{\tilde{X}_S X}\underbrace{\left[(f - \mathbb{E}_F[\tilde{\hat{f}}])^2\right]}_{\text{Bias}^2} + \mathbb{E}_{\tilde{X}_S X}\underbrace{[\mathbb{V}_F[\tilde{\hat{f}}]]}_{\text{Variance}}
\end{aligned}
$$

For the original loss (ii):

$$\begin{aligned}
\mathbb{E}_{FXY}[L] &= \mathbb{E}_{FXY}[(Y-\hat{f})^2] \\
&= \mathbb{E}_{XY}[Y^2 - 2Y\mathbb{E}_F[\hat{f}] + \mathbb{E}_F[\hat{f}^2]] \\
&= \mathbb{E}_{XY}[Y^2 - 2Y\mathbb{E}_F[\hat{f}] + \mathbb{E}_F[\hat{f}]^2 + \mathbb{V}_F[\hat{f}]] \\
&= \mathbb{V}_Y[Y] + \mathbb{E}_X[f^2 - 2f\mathbb{E}_F[\hat{f}] + \mathbb{E}_F[\hat{f}]^2 + \mathbb{V}_F[\hat{f}]] \\
&= \underbrace{\sigma^2}_{\text{Data Var}} + \mathbb{E}_X\underbrace{\left[(f-\mathbb{E}_F[\hat{f}])^2\right]}_{\text{Bias}^2} + \mathbb{E}_X\underbrace{[\mathbb{V}_F(\hat{f})]}_{\text{Variance}}
\end{aligned}$$

The expected PFI for feature X_S then is:

$$\begin{aligned}
\mathbb{E}_F[PFI_{\hat{f}}] &= \mathbb{E}_{F\tilde{X}_SXY}[\tilde{L}] - \mathbb{E}_{FXY}[L] \\
&\overset{(i)+(ii)}{=} \sigma^2 + \mathbb{E}_{\tilde{X}_SX}\left[(f-\mathbb{E}_F[\tilde{\hat{f}}])^2\right] + \mathbb{E}_{\tilde{X}_SX}[\mathbb{V}_F(\tilde{\hat{f}})] \\
&- (\sigma^2 + \mathbb{E}_X\left[(f-\mathbb{E}_F[\hat{f}])^2\right] + \mathbb{E}_X[\mathbb{V}_F(\hat{f})]) \\
&= \mathbb{E}_{\tilde{X}_SX}\left[(f-\mathbb{E}_F[\tilde{\hat{f}}])^2\right] - \mathbb{E}_X\left[(f-\mathbb{E}_F[\hat{f}])^2\right] \\
&+ \mathbb{E}_{\tilde{X}_SX}[\mathbb{V}_F[\tilde{\hat{f}}]] - \mathbb{E}_X[\mathbb{V}_F[\hat{f}]]
\end{aligned}$$

We can derive the same L2 decomposition for the DGP-PFI by replacing $\hat{f}$ with f in the equation above. This yields $PFI_f = \mathbb{E}_{\tilde{X}_SX}[(f(X) - f(\tilde{X}_S, X_C))^2]$, since $\mathbb{V}_F[f] = \mathbb{V}_F[\tilde{f}] = 0$ and $\mathbb{E}_F[f] = f$ and $\mathbb{E}_F[\tilde{f}] = \tilde{f}$.

The bias of the model-PFI compared to the DGP-PFI is:

$$\begin{aligned}
\mathbb{E}_F[PFI_{\hat{f}}] - PFI_f &= \underbrace{\mathbb{E}_{\tilde{X}_SX}[(f-\mathbb{E}_F[\tilde{\hat{f}}])^2 - (f-\tilde{f})^2]}_{\text{Permutation Loss Bias}} && (11) \\
&- \underbrace{\mathbb{E}_X\left[(f-\mathbb{E}_F[\hat{f}])^2]\right]}_{\text{(Learner Bias)}^2} + \underbrace{\mathbb{E}_{\tilde{X}_SX}[\mathbb{V}_F[\hat{f}]] - \mathbb{E}_X[\mathbb{V}_F[\hat{f}]]}_{\text{Variance Inflation}} && (12) \\
&\overset{unbiased}{=} \underbrace{\mathbb{E}_{\tilde{X}_SX}[\mathbb{V}_F[\hat{f}]] - \mathbb{E}_X[\mathbb{V}_F[\hat{f}]]}_{\text{Variance Inflation}} && (13) \\
&\overset{\tilde{X}_S\sim X_S|X_C}{=} 0 && (14)
\end{aligned}$$

The permutation loss bias and the squared learner bias are zero due to the unbiasedness assumption, i.e. $\mathbb{E}_F[\hat{f}] = f$. The variance inflation term is zero if $\tilde{X}_S \sim X_S \mid X_C$, which is here the case due to conditional sampling.

H conditional DGP-PFI minus model-PFI for L2

In this proof, we use the conditional sampler ϕ_{cond} for both, the DGP-PFI and the model-PFI.

$$\begin{aligned}
PFI_f - PFI_{\hat{f}} &= \mathbb{E}_{\tilde{X_S} X_C Y}[(Y-f)^2] - \mathbb{E}_{X_S X_C Y}[(Y-f)^2] \\
&\quad - \Big(\mathbb{E}_{\tilde{X_S} X_C Y}[(Y-\hat{f})^2] - \mathbb{E}_{X_S X_C Y}[(Y-\hat{f})^2]\Big) \\
&= \underbrace{\Big(\mathbb{E}_{X_S X_C Y}[(Y-\hat{f})^2] - \mathbb{E}_{X_S X_C Y}[(Y-f)^2]\Big)}_{\text{T1:=}} \\
&\quad + \underbrace{\Big(\mathbb{E}_{\tilde{X_S} X_C Y}[(Y-f))^2] - \mathbb{E}_{\tilde{X_S} X_C Y}[(Y-\hat{f})^2]\Big)}_{\text{T2:=}}
\end{aligned}$$

We know that for any $g : X \to Y$ holds:

$$\mathbb{E}_{X,Y}[(Y-g)^2] = \mathbb{E}_X[\mathbb{V}_{Y|X}[Y]] + \mathbb{E}_X[(\mathbb{E}_{Y|X}[Y] - g)^2]$$

Since $f = \mathbb{E}_{Y|X_S,X_C}[Y]$ we can conclude for our first term T1 that:

$$\begin{aligned}
\text{T1} &= \mathbb{E}_{X_S X_C}[\mathbb{V}_{Y|X_S,X_C}[Y]] + \mathbb{E}_{X_S X_C}[(f-\hat{f})^2] \\
&\quad - \Big(\mathbb{E}_{X_S X_C}[\mathbb{V}_{Y|X_S,X_C}[Y]] + \underbrace{\mathbb{E}_{X_S X_C}[(f-f)^2]}_{=0}\Big) \\
&= \mathbb{E}_{X_S X_C}[(f-\hat{f})^2]
\end{aligned}$$

We apply the same strategy to T2. Moreover, $Y \perp\!\!\!\perp \tilde{X}_S \mid X_C$.

$$\begin{aligned}
\text{T2} &= \mathbb{E}_{\tilde{X_S} X_C}[\mathbb{V}_{Y|\tilde{X}_S,X_C}[Y]] + \mathbb{E}_{\tilde{X_S} X_C}[(\mathbb{E}_{Y|\tilde{X}_S,X_C}[Y] - f)^2] \\
&\quad - \Big(\mathbb{E}_{\tilde{X_S} X_C}[\mathbb{V}_{Y|\tilde{X}_S,X_C}[Y]] + \mathbb{E}_{\tilde{X_S} X_C}[(\mathbb{E}_{Y|\tilde{X}_S,X_C}[Y] - \hat{f})^2]\Big) \\
&= \mathbb{E}_{\tilde{X_S} X_C}[(\mathbb{E}_{Y|X_C}[Y] - f)^2] - \mathbb{E}_{\tilde{X_S} X_C}[(\mathbb{E}_{Y|X_C}[Y] - \hat{f})^2]
\end{aligned}$$

If we now set together the two terms again and use in the first step that $P(X_S, X_C) = P(\tilde{X}_S, X_C)$, we obtain:

$$
\begin{aligned}
\text{T1+T2} &= \mathbb{E}_{X_S X_C}[(f-\hat{f})^2] + \mathbb{E}_{X_S X_C}[(\mathbb{E}_{Y|X_C}[Y] - f)^2] \\
&- \mathbb{E}_{X_S X_C}[(\mathbb{E}_{Y|X_C}[Y] - \hat{f})^2] \\
&= \mathbb{E}_{X_S X_C}\Big[f^2 - 2f\hat{f} + \hat{f}^2 + \mathbb{E}_{Y|X_C}[Y]^2 - 2\mathbb{E}_{Y|X_C}[Y]f + f^2 \\
&- \mathbb{E}_{Y|X_C}[Y]^2 + 2\mathbb{E}_{Y|X_C}[Y]\hat{f} - \hat{f}^2\Big] \\
&= 2\mathbb{E}_{X_S X_C}\Big[(f^2 - \mathbb{E}_{Y|X_C}[Y]f) - (f\hat{f} - \mathbb{E}_{Y|X_C}[Y]\hat{f})\Big] \\
&= 2\mathbb{E}_{X_C}\Big[\mathbb{E}_{X_S|X_C}\big[(f^2 - \mathbb{E}_{Y|X_C}[Y]f) - (f\hat{f} - \mathbb{E}_{Y|X_C}[Y]\hat{f})\big]\Big] \\
&\overset{*}{=} 2\mathbb{E}_{X_C}\Big[(\mathbb{E}_{X_S|X_C}[f^2] - \mathbb{E}_{Y|X_C}[Y]\mathbb{E}_{X_S|X_C}[f]) \\
&- (\mathbb{E}_{X_S|X_C}[f\hat{f}] - \mathbb{E}_{Y|X_C}[Y]\mathbb{E}_{X_S|X_C}[\hat{f}])\Big] \\
&\overset{**}{=} 2\mathbb{E}_{X_C}\Big[(\mathbb{E}_{X_S|X_C}[f^2] - \mathbb{E}_{X_S|X_C}[f]^2) \\
&- (\mathbb{E}_{X_S|X_C}[f\hat{f}] - \mathbb{E}_{X_S|X_C}[\hat{f}]\mathbb{E}_{X_S|X_C}[f])\Big] \\
&= 2\mathbb{E}_{X_C}\big[\mathbb{V}_{X_S|X_C}[f] - Cov_{X_S|X_C}[f,\hat{f}]\big]
\end{aligned}
$$

At *, we use the fact that the random variable $\mathbb{E}_{Y|X_C}[Y]$ is measurable by the σ-Algebra generated from X_C, and we are inclined to pull it out of the expectation. In **, we use that from $f = \mathbb{E}_{Y|X_S,X_C}[Y]$ follows $\mathbb{E}_{X_S|X_C}[f] = \mathbb{E}_{Y|X_C}[Y]$.

I CI simulation results

J Theoretical background of model-based uncertainty

[32] leverage the kernel of GPs to analytically calculate the model-based uncertainty contained in the PD function. Let $\hat{f}$ be a GP and $\hat{\boldsymbol{m}}(x) = \left(\hat{m}(x, x_C^{(i)})\right)_{i=1,\ldots,n_2}$ its estimated posterior mean and $\hat{\boldsymbol{K}}(x) = \left(\hat{k}\big((x, x_C^{(i)}), (x, x_C^{(j)})\big)\right)_{i,j=1,\ldots,n_2}$ its estimated posterior covariance on the test set D_{n_2} for fixed feature values $x \in X_S$. The PD estimate $\widehat{PD}$ of $\hat{f}$ can be seen as a random variable. Thus, the PD for the posterior mean function is given by the expected value of $\widehat{PD}$:

$$
\mathbb{E}_{\hat{f}}\left[\widehat{PD}(x)\right] = \mathbb{E}_{\hat{f}}\left[\frac{1}{n_2}\sum_{i=1}^{n_2}\hat{f}(x, x_C^{(i)})\right] = \frac{1}{n_2}\sum_{i=1}^{n_2}\hat{m}(x, x_C^{(i)}). \tag{15}
$$

The variance of the PD is estimated accordingly and can be calculated straightforwardly by leveraging the posterior covariance of the GP:

$$
\mathbb{V}_{\hat{f}}\left[\widehat{PD}(x)\right] = \mathbb{V}_{\hat{f}}\left[\frac{1}{n_2}\sum_{i=1}^{n_2}\hat{f}(x, x_C^{(i)})\right] = \frac{1}{{n_2}^2}\mathbf{1}^\top\hat{\boldsymbol{K}}(x)\mathbf{1}. \tag{16}
$$

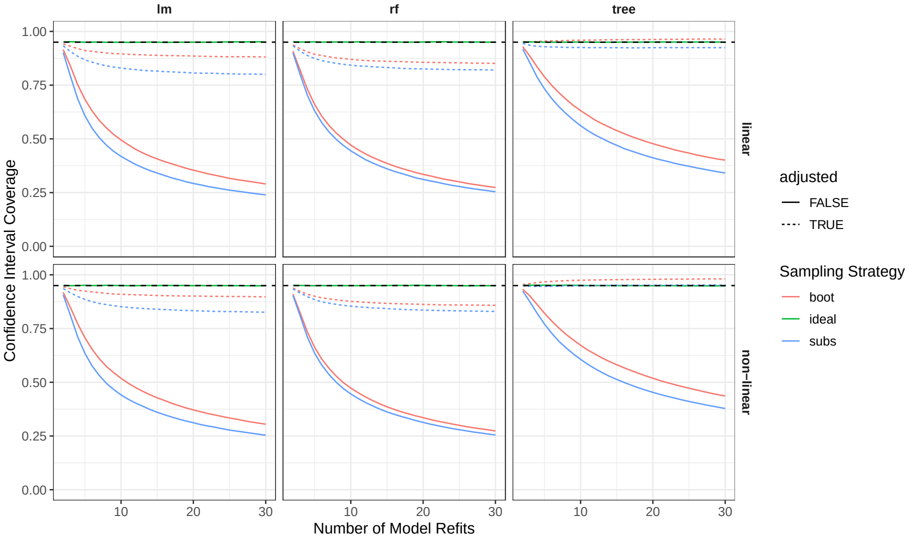

Figure I.7: CI coverage for PD with n=100.

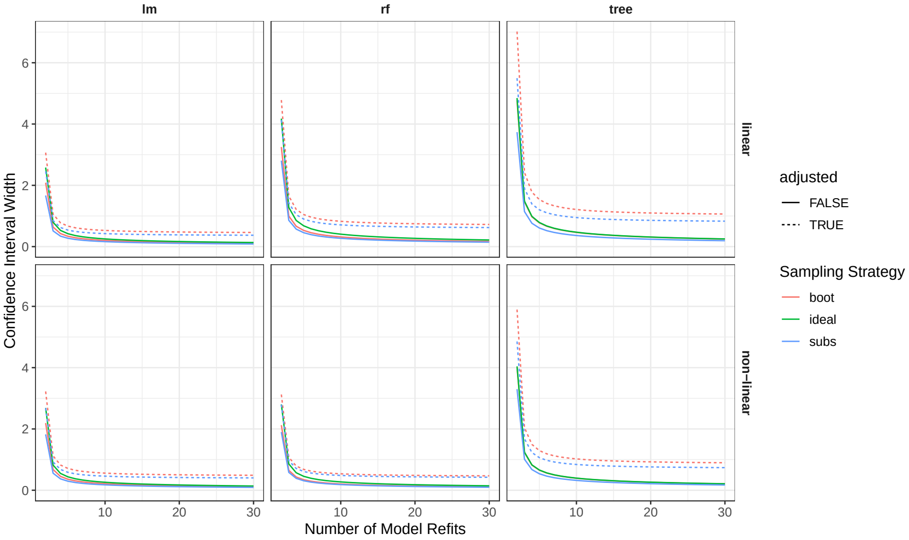

Figure I.8: CI width for PD with n=100.

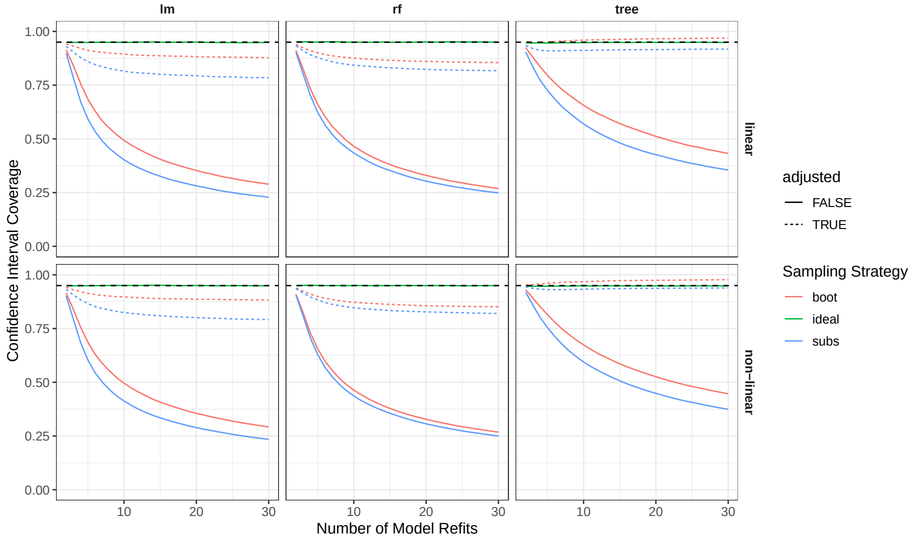

Figure I.9: CI coverage for PD with n=1,000.

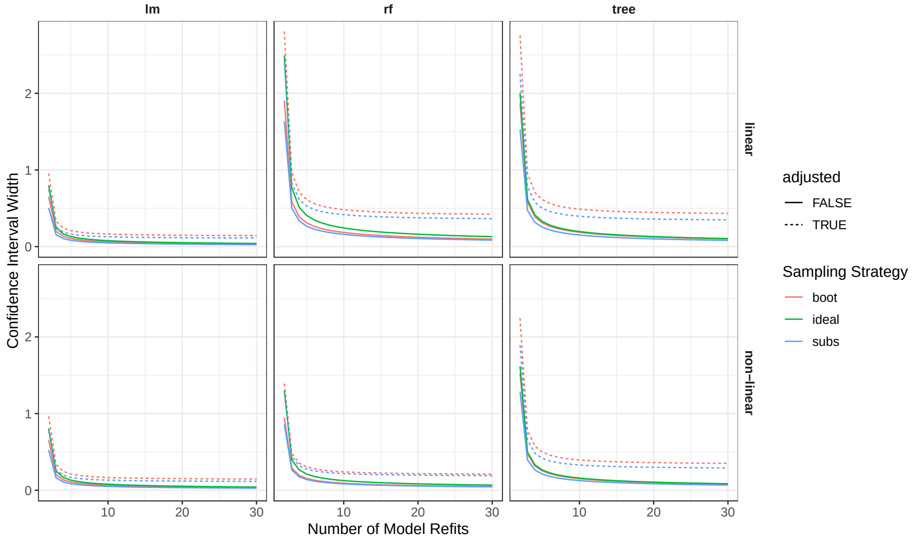

Figure I.10: CI width for PD with n=1,000.

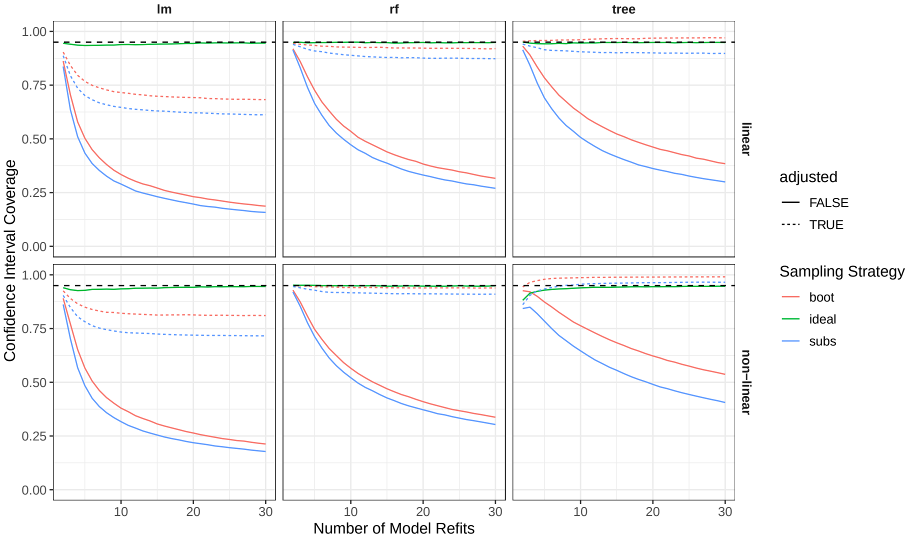

Figure I.11: CI coverage for PFI with n=100.

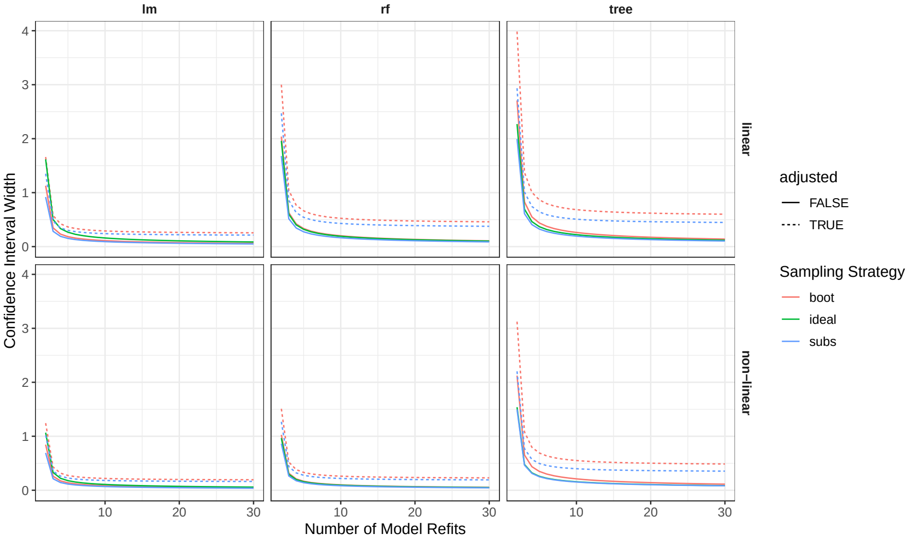

Figure I.12: CI width for PFI with n=100.

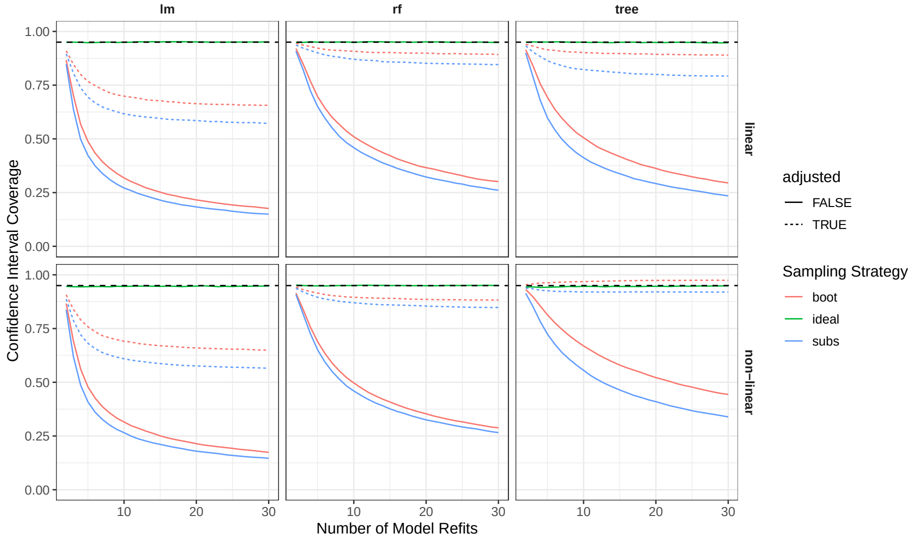

Figure I.13: CI coverage for PFI with n=1,000.

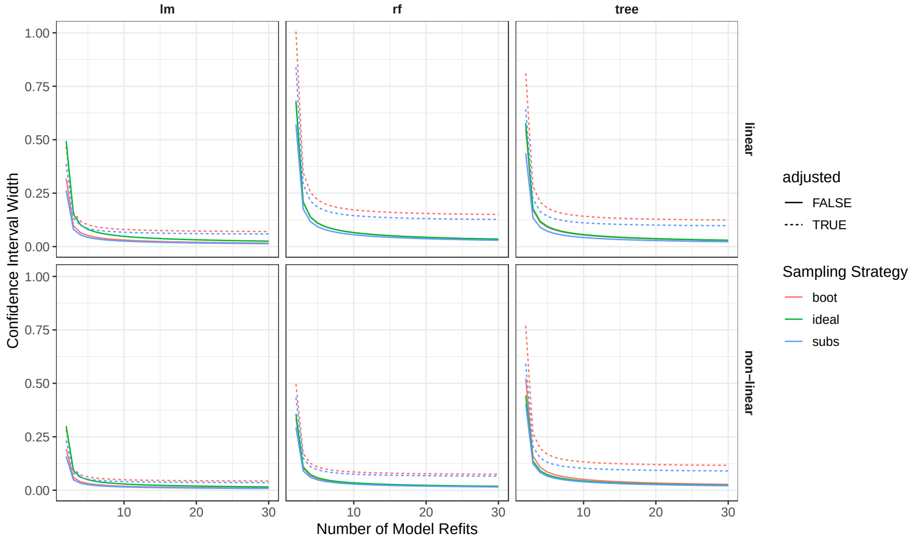

Figure I.14: CI width for PFI with n=1,000.

Since the n_2 predictors $\hat{f}(x, x_C^{(i)})$ of the GP follow a Gaussian distribution, their sum is also normally distributed. Hence, we can construct confidence bands for the mean estimate in Eq. (15) by using the variance estimate in Eq. (16) together with the respective $1-\alpha/2$ quantiles of the Gaussian distribution. This approach is applicable to any models (including non-GPs) that provide a fully specified covariance matrix between the predictions.

As Eq. (16) solely quantifies the variance w.r.t. the model given the observed data, the resulting confidence bands only capture model variance but not the variance induced by MC integration.

Supplementary References

48. Bashtannyk, D.M., Hyndman, R.J.: Bandwidth selection for kernel conditional density estimation. Computational Statistics & Data Analysis **36**(3), 279–298 (2001)
49. Bishop, C.M.: Mixture density networks. Tech. rep., Aston University (1994)
50. Hothorn, T., Zeileis, A.: Predictive distribution modeling using transformation forests. Journal of Computational and Graphical Statistics **30**(4), 1181–1196 (2021)
51. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2018), `https://www.R-project.org/`
52. Scholbeck, C.A., Molnar, C., Heumann, C., Bischl, B., Casalicchio, G.: Sampling, intervention, prediction, aggregation: A generalized framework for model-agnostic interpretations. Communications in Computer and Information Science p. 205–216 (2020). `https://doi.org/10.1007/978-3-030-43823-4_18`
53. Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. Advances in neural information processing systems **28** (2015)
54. Trippe, B.L., Turner, R.E.: Conditional density estimation with bayesian normalising flows. arXiv preprint arXiv:1802.04908 (2018)
55. Winkler, C., Worrall, D., Hoogeboom, E., Welling, M.: Learning likelihoods with conditional normalizing flows. arXiv preprint arXiv:1912.00042 (2019)