

One Permutation Is All You Need: Fast, Reliable Variable Importance and Model Stress-Testing

Albert Dorador
 trustalgorithm.dev@gmail.com

December 24, 2025

Abstract

Reliable estimation of feature contributions in machine learning models is essential for trust, transparency and regulatory compliance, especially when models are proprietary or otherwise operate as “black boxes”. While permutation-based methods are a standard tool for this task, classical implementations rely on repeated random permutations, introducing computational overhead and stochastic instability. In this paper, we show that by replacing multiple random permutations with a single, deterministic, and optimal permutation, we achieve a method that retains the core principles of permutation-based importance while being non-random, faster, and more stable. We validate this approach across nearly 200 scenarios, including real-world household finance and credit risk applications, demonstrating improved bias-variance tradeoffs and accuracy in challenging regimes such as small sample sizes, high dimensionality, and low signal-to-noise ratios. Finally, we introduce Systemic Variable Importance, a natural extension designed for model stress-testing that explicitly accounts for feature correlations. This framework provides a transparent way to quantify how shocks or perturbations propagate through correlated inputs, revealing dependencies that standard variable importance measures miss. Two real-world case studies demonstrate how this metric can be used to audit models for hidden reliance on protected attributes (e.g., gender or race), enabling regulators and practitioners to assess fairness and systemic risk in a principled and computationally efficient manner.

1 Introduction

In machine learning applications, it is generally agreed nowadays that estimating variable importance (VI) provides valuable insight, and for that reason it has become a cornerstone of explainable AI (XAI), as well as modern, responsible machine learning practice more broadly. However, what does not seem to be as well-established is what VI actually means.

There are, at least, three different notions of VI, each one trying to answer a different question:

- Population VI: “how important is this feature in the true population model?”
- Model class VI: “how important or replaceable is this feature in a model like this?”
- Model instance VI: “how much does this specific fitted model rely on this feature?”

All notions are valid ways to interpret VI and they try to answer fundamentally different questions, which may naturally merit fundamentally different answers.

The first notion, while very intuitive and appealing, presents an important complication: when we speak of “the” importance of a variable, we often implicitly assume the uniqueness of the population model, but this has been strongly challenged in the machine learning literature: the

same phenomenon may have different plausible explanations (Rashomon effect [Breiman, 2001b]). One attempt to resolve the tension derived from the lack of a unique population model has been proposed in Fisher et al. [2019], where the authors suggest considering a set of high-performing models instead of committing to just one model. While principled, this approach, requiring the highly non-trivial task of first estimating the Rashomon set of models for a particular problem, lacks widespread adoption. In practice, approaches that instead suggest using a flexible, non-parametric class of models to estimate “the” population model are more common, like Random Forests [Breiman, 2001a] or GUIDE trees [Loh, 2012, Loh and Zhou, 2021]. These methods, in turn, can be used for the different (although related) goal of variable selection, as a pre-processing step in a predictive machine learning pipeline.

The second notion ties VI to a concrete class of models, e.g. the class of Random Forests with 500 decision trees. In this case, we want to know how much the class of models at hand relies on a particular feature given the set of features at our disposal. Features with more predictive power and weaker association with the rest of features will score higher. Several methods to estimate this type of VI exist, the quintessential one being LOCO (leave-one-covariate-out) [Lei et al., 2018], which estimates the class reliance on each feature (conditional on the rest) by tracking the difference in predictive performance between the model fitted with all features and a model leaving one of the features out. This is a computationally intensive procedure that can be prohibitively expensive for some model classes and/or in high-dimensional prediction. In linear models, there is no need to carry out this costly procedure, as statistical theory already provides ways to estimate the conditional importance of variables efficiently and reliably under the standard Gauss-Markov assumptions. For more general frameworks, there exist more efficient alternatives to LOCO that accomplish similar goals and perform well in practice, such as estimating ghost variables [Delicado and Peña, 2023] with a fast estimator [Dorador, 2025b,c].

Given a trained machine learning model (a “master” model), the third notion of VI refers to the influence that each covariate has within that specific fitted model. Because we have a fixed model (i.e. a concrete member of a certain model class), methods that estimate VI in this context are not concerned with the ability of the model (class) to potentially replace any given covariate with the rest, as that would yield a *different* model (i.e. another member of the given model class). Therefore, estimating the *direct* importance of a variable in this context *requires* ignoring association among covariates. In other words, ignoring the correlation among covariates is not a bug, but a feature. In that sense, the criticism that classical permutation methods [Breiman, 2001a] have received [Strobl et al., 2008, Hooker et al., 2021, Delicado and Peña, 2023] for ignoring feature interdependence when estimating *direct* VI is undeserved.

Indeed, classical VI methods, which randomly permute multiple times the values of each variable x_j to create a new variable x'_j and compare the resulting loss in prediction accuracy, are some of the most popular VI methods among practitioners, for several reasons:

- Conceptual simplicity and intuitive appeal
- Efficiency: only one predictive model has to be fitted (the one using all covariates)
- Comparability: the comparison is made between two models with the same number of variables, so any differences in prediction accuracy cannot be attributed to a difference in dimensionality
- Model-agnosticism: this VI method can be applied to *any* predictive model [Fisher et al., 2019]

- Theoretical guarantees: convergence results via Law of Large Numbers or Central Limit Theorem [Föge et al., 2024]

However, this does not mean that Breiman-style permutation does not have room for improvement. There are at least four aspects of traditional permutation methods for direct VI estimation that can still be improved:

- Randomness: their non-deterministic nature introduces stochastic instability, making scores dependent on the random seed. In high-stakes regulatory and auditing environments, this ambiguity compromises reproducibility and auditability
- Estimator variance: as a direct consequence of the previous shortcoming, this type of VI estimator can have substantial variance, which can only be reduced – although never fully eliminated – by increasing the number of per-feature permutations B , which is naturally expensive
- Improvable efficiency: while vastly more efficient than most other VI methods, the fact that each feature must be permuted B times introduces inefficiency. There are variants that suggest considering all $n!$ permutations for each feature, even [Fisher et al., 2019].
- Evaluation metric: the canonical implementation of this type of method considers the drop in predictive performance e.g. increase in test mean squared error (MSE) resulting from replacing x_j by x'_j when evaluating the original fitted model. While intuitive, it might not be the best choice of metric under weak fitted models, as said feature replacement could in fact *improve* predictive performance and hence yield *negative* importance scores, which complicates standardization and interpretation in this context, which is different than e.g. variable selection.

In this paper, we propose a Breiman-inspired VI method that corrects the above shortcomings while preserving (or improving) its qualities. We also present a natural extension to model stress-testing that does account for feature correlations. As such, this work both challenges common practice and expands the theoretical understanding of permutation-based importance measures.

Lastly, it is worth noting that permutation-based methods are not the only types of methods concerned with the third notion of VI that we have discussed. For example, methods based on Shapley values [Shapley, 1953, Lundberg and Lee, 2017] have gained traction in recent years (especially as local, rather than global, model-agnostic explanation tools), although they have their own limitations too: they are computationally expensive and, in order to mitigate that drawback, often assume independence among features, which is unrealistic in many scenarios and can compromise the reliability of importance scores. Note that assuming feature independence is not the same as ignoring feature dependence, which is what classical permutation-based methods (correctly) do.

The rest of this paper is structured as follows: Section 2 provides rigorous theoretical justification for the replacement of multiple random permutations by a single, deterministic, and optimal one. Section 3 starts by formally defining our proposed measures of direct variable importance (DVI), and then discusses the remaining methodological aspects of the paper, such as our choice of deterministic permutation, evaluation metric and empirical simulation setup, including our correlated feature design. Section 4 presents the main empirical results of our proposed DVI method relative to classical baselines. Section 5 transitions from direct VI to systemic VI (which we formally define), showing how our proposed direct VI method can be extended to this new context by accounting for feature interdependence, highlighting with two real-world case studies its practical relevance in model fairness evaluation and stress-testing. Section 6 concludes.

2 One Optimal Permutation is Worth a Thousand Random Ones

Traditional permutation-based VI [Breiman, 2001a, Fisher et al., 2019] samples B independent permutations uniformly at random for each feature, computes a measure of model-specific variable importance each time, and then averages the results across the B runs, where B in practice is typically at least ten [Alonso and Carbó, 2022], and sometimes more (e.g. 50 or 100).

While this method has enjoyed remarkable popularity for the reasons we described previously, its inherent randomness introduces non-zero estimator variance for finite B . Formally, this type of estimator of VI is defined as

$$\hat{I}_B = \frac{1}{B} \sum_{b=1}^B g(\pi_b) \quad (1)$$

where $g(\pi_b)$ is the importance functional for a single permutation π_b (e.g. MSE increase per feature following the permutation of its values). Thus, under i.i.d. permutation sampling, the variance of this estimator is

$$\text{Var}(\hat{I}_B) = \frac{\text{Var}(g(\pi))}{B} \quad (2)$$

and hence positive for any finite B . On the other hand, the variance of a deterministic permutation $\tilde{I} = g(\pi^*)$ is 0 (given the dataset).

In addition, it is clear that not all possible random permutations contribute equally to the estimation of the VI vector: minimally perturbing ones (e.g. only swapping two out of n feature values) contribute less than another one that significantly displaces every single feature value from its original position. Thus, in general, some of the B permutations are partially wasted. In particular, it is conceivable that, in general, a comparable effect could be achieved with a smaller amount of more carefully-chosen permutations.

Taking these limitations into account, we wish to replace B random permutations by a single, deterministic and optimal permutation. By optimal we mean that it perturbs all ranks of a given feature as much as possible and as evenly as possible. This is a reasonable optimality criterion, as we wish to find a permutation (hence a marginal distribution-preserving mechanism) that removes as much signal from the feature at hand as possible overall, while making sure we do not leave any local structure barely untouched.

Given a sample size of n observations per feature, we can achieve this goal by shifting the ranks of the feature at hand by $\lfloor n/2 \rfloor$ (modulo n). Let us formalize this idea. We will prove via a max-min argument that this choice of permutation indeed fulfills the optimality criterion that we have intuitively defined before and which we will rigorously define next.

Proposition 1 (Max-min optimality of $\pi_{\lfloor n/2 \rfloor}$). *Let $n \geq 2$. For any integers $a, b \in \{0, \dots, n-1\}$ define the circular displacement function as*

$$d(a, b) := \min(|a - b|, n - |a - b|) \quad (3)$$

For any permutation π of $\{0, \dots, n-1\}$, define

$$m(\pi) := \min_j d(j, \pi(j)) \quad (4)$$

Then, for every π we have $m(\pi) \leq \lfloor n/2 \rfloor$.

Moreover, the cyclic shift permutation

$$\pi_k(j) = (j + k) \bmod n \quad (5)$$

with $k = \lfloor n/2 \rfloor$ satisfies $m(\pi_k) = \lfloor n/2 \rfloor$. Hence π_k attains the maximal possible value of the objective function, and is therefore optimal.

Proof. Observe that for any $a, b \in \{0, \dots, n-1\}$ we have

$$d(a, b) \leq \lfloor n/2 \rfloor \quad (6)$$

because the maximum possible value of $\min(t, n-t)$ for integer $t \in \{0, \dots, n-1\}$ is $\lfloor n/2 \rfloor$. Consequently, for any permutation π ,

$$m(\pi) = \min_j d(j, \pi(j)) \leq \lfloor n/2 \rfloor \quad (7)$$

Hence, $\lfloor n/2 \rfloor$ is an upper bound on the objective. We will show that choosing π_k with $k = \lfloor n/2 \rfloor$ attains said upper bound.

Let $k = \lfloor n/2 \rfloor$ and consider the permutation (a cyclic shift) $\pi_k(j) = (j+k) \bmod n$. Note that for each j ,

$$|\pi_k(j) - j| = k \quad \text{or} \quad |\pi_k(j) - j| = n - k \quad (8)$$

Thus, the circular displacement is

$$d(j, \pi_k(j)) = \min(k, n-k) \quad (9)$$

but by definition of $k = \lfloor n/2 \rfloor$ we have $\min(k, n-k) = k$. Therefore,

$$d(j, \pi_k(j)) = k \quad \text{for every } j \quad (10)$$

and hence

$$m(\pi) = \min_j d(j, \pi(j)) = k = \lfloor n/2 \rfloor \quad (11)$$

□

Note that the function $d(a, b) = \min(|a-b|, n-|a-b|)$ is the *circular* displacement: it measures distance around the n -cycle. For example, if $n = 5$ and $j = 3$ then the ordinary absolute difference is $|\pi_2(3) - 3| = |0 - 3| = 3$, but the circular displacement is $d(3, \pi_2(3)) = \min(3, 5-3) = 2$. Thus the quantity $\lfloor n/2 \rfloor$ is indeed the maximal possible circular displacement for every n , and the proof above does not require n to be even. Indeed, the cyclic shift π_k with $k = \lfloor n/2 \rfloor$ not only maximizes the minimum displacement but makes every circular displacement equal to the maximum possible.

In addition, other natural candidate permutations like full rank reversal ($\pi(j) = n-1-j$) also produce large displacements on average, but they are not equal: some middle ranks map to themselves (for odd n) or near themselves (for even n), which nullifies the permutation effect for those values. Thus, full rank reversal maximizes other criteria (e.g. certain forms of average linear displacement) but does not achieve the evenness property that we seek.

Lastly, note that rank shifting can handle categorical features exactly like continuous ones, as all that is needed is to have an ordered sequence of distinct ranks, which is (deterministically) guaranteed even if the values themselves are neither numeric nor distinct.

Next we will show that under appropriate conditions a single permutation recovers ground-truth variable importance scores better in terms of MSE than traditional VI methods based on B i.i.d. random permutations, whether in finite samples or asymptotically.

Proposition 2 (Component-wise MSE dominance). *Given a dataset and a fitted model, fix a feature index $j \in \{1, \dots, p\}$. Let $\mu_j \in \mathbb{R}$ denote the ground-truth importance for feature j . Let $\hat{I}_{B,j}$ be the Breiman Monte Carlo estimator for feature j obtained by averaging B i.i.d. random permutations, and let \tilde{I}_j denote the deterministic single-permutation estimator (e.g. obtained with our rank-shift permutation).*

Write $g_j(\pi)$ for the scalar importance computed for feature j under permutation π , and let

$$\sigma_j^2 := \text{Var}_\Pi(g_j(\Pi))$$

be the variance of the importance under a uniformly random permutation Π . Let expectations be taken over the random permutations (and any randomness in the Monte Carlo draw). Define

$$\text{MSE}(\hat{I}_{B,j}) := \mathbb{E}[(\hat{I}_{B,j} - \mu_j)^2], \quad \text{MSE}(\tilde{I}_j) := \mathbb{E}[(\tilde{I}_j - \mu_j)^2].$$

Then

$$\text{MSE}(\tilde{I}_j) < \text{MSE}(\hat{I}_{B,j})$$

if and only if

$$(\tilde{I}_j - \mu_j)^2 < (\mathbb{E}[\hat{I}_{B,j}] - \mu_j)^2 + \frac{\sigma_j^2}{B} \quad (12)$$

which implies

$$|\tilde{I}_j - \mu_j| - |\mathbb{E}[\hat{I}_{B,j}] - \mu_j| < \frac{\sigma_j}{\sqrt{B}}. \quad (13)$$

Proof. First note that expectations are taken with respect to the permutation distribution. Thus, $\mathbb{E}[\tilde{I}_j] = \tilde{I}_j$ trivially. Now, write the bias-variance decompositions

$$\text{MSE}(\hat{I}_{B,j}) = (\mathbb{E}[\hat{I}_{B,j}] - \mu_j)^2 + \text{Var}(\hat{I}_{B,j}), \quad \text{MSE}(\tilde{I}_j) = (\tilde{I}_j - \mu_j)^2 + \text{Var}(\tilde{I}_j).$$

Because $\hat{I}_{B,j}$ is an average of B i.i.d. draws of $g_j(\Pi)$ we have $\text{Var}(\hat{I}_{B,j}) = \sigma_j^2/B$. The deterministic estimator has zero sampling variance so $\text{Var}(\tilde{I}_j) = 0$. Thus the inequality $\text{MSE}(\tilde{I}_j) < \text{MSE}(\hat{I}_{B,j})$ is equivalent to

$$(\tilde{I}_j - \mu_j)^2 < (\mathbb{E}[\hat{I}_{B,j}] - \mu_j)^2 + \frac{\sigma_j^2}{B}.$$

Taking square roots and re-arranging gives the final stated form; the displayed inequality (13) follows from the elementary bound

$$\sqrt{a^2 + c^2} \leq |a| + |c| \quad (14)$$

applied to $a = \mathbb{E}[\hat{I}_{B,j}] - \mu_j$ and $c = \sigma_j/\sqrt{B}$, which holds with strict inequality if neither a nor c are zero. Note c is never zero, and if $a = 0$, then the bound (14) is no longer necessary and (13) immediately holds. \square

Corollary 1 (Vector (global) MSE dominance). *If one defines the vector-valued estimators $\hat{I}_B, \tilde{I} \in \mathbb{R}^p$ and measures MSE in ℓ_2 ,*

$$\text{MSE}(\hat{I}_B) := \mathbb{E}[\|\hat{I}_B - \mu\|_2^2], \quad \text{MSE}(\tilde{I}) := \mathbb{E}[\|\tilde{I} - \mu\|_2^2],$$

and if the permutation-based variances σ_j^2 are available for each component, then

$$\text{MSE}(\tilde{I}) < \text{MSE}(\hat{I}_B)$$

if and only if the component-wise inequalities of the proposition combine to give

$$\sum_{j=1}^p (\tilde{I}_j - \mu_j)^2 < \sum_{j=1}^p \left((\mathbb{E}[\hat{I}_{B,j}] - \mu_j)^2 + \frac{\sigma_j^2}{B} \right).$$

Proof. The result follows directly from the definition of ℓ_2 norm and Proposition 2. \square

Proposition 2 formalizes the intuitive notion we described previously: the single deterministic permutation wins in MSE for a given feature whenever its *excess absolute bias* relative to the Monte Carlo estimator is smaller than the Monte Carlo estimator’s standard error σ_j/\sqrt{B} .

A few remarks are in order.

First, the effect of increasing B as a stabilizing mechanism has only sub-linear effect and exhibits diminishing marginal returns: the cost of each additional permutation stays constant (permuting and evaluating p times the fitted model) while the additional gain decreases.

Second, note that this MSE dominance result does not require the single, deterministic permutation to be optimal in any sense. However, choosing a permutation that is in some sense optimal (like the optimal rank shift discussed earlier) makes the result more likely to hold in practice for a given dataset.

Third, in practice, given a dataset and a fitted model, one can decide whether the deterministic estimator appears preferable for the target budget B by checking whether $\|\tilde{I} - \hat{I}_B\|_2^2 < \|\hat{\sigma}^2\|_2^2/B$, where for each $j = 1, \dots, p$, $\hat{\sigma}_j^2 := \sum_b (g(\pi_b) - \hat{I}_B)^2/(B-1)$. This comparison is conservative in the sense that it implicitly gives the Monte Carlo method a head-start by supposing it is arbitrarily close in expectation to the ground-truth vector μ .

Lastly, it is possible to state and prove analogous results that hold only asymptotically as the sample size n grows. In such case, the single-permutation method may have asymptotically lower MSE than the Monte Carlo method provided that the absolute bias of the former decays faster than that of the latter as n increases. If σ decays with n , it should also be accounted for in that comparison.

3 Methods

In this section we formally define our notion of direct variable importance, and then describe more specific design choices for our proposed VI method, as well as our empirical simulation setup. As in Breiman-style permutation importance, the direct importance of a variable within a model is understood as the expected immediate change in prediction after the signal in a feature is (largely) removed by permuting its values. We formalize this notion next.

Definition 1 (Direct Variable Importance). *Given the evaluation function of a trained model M denoted by $f_M : \mathbb{R}^{n \times p} \rightarrow \mathbb{R}^{n \times q}$ ($q = 1$ in standard regression, or $q = C$ in C -class classification), and a dataset $D = (X, y)$ where $X \in \mathcal{S}_X \subseteq \mathbb{R}^{n \times p}$ and $y \in \mathcal{S}_y$, the estimated direct importance of the k -th variable x_k is defined as*

$$d_k := \frac{1}{nq} \sum_{i=1}^{nq} d(f_{M_i}(X) - f_{M_i}(X_{k'})) \quad (15)$$

where $d(a, b)$ is a suitable metric (e.g. $d(a, b) = |a - b|$) that can be applied element-wise, and $X_{k'}$ denotes X replacing x_k by its permuted representation x'_k . Lastly, final direct scores are normalized to sum to one.

We deliberately did not require having access to a test set where the feature permutations and subsequent model evaluations are carried out: access to training data is enough, as the model parameters are estimated using the training set and fixed thereafter – in model instance VI, they are all that matter to estimate direct variable importance for that specific trained model. Common practice often requires using a test set, but this belief stems from the confusion surrounding the

definitions of VI (for instance, whether the model overfits is irrelevant to estimate its reliance on a given feature). Nonetheless, we follow the convention to use a test set to estimate VI in all our experiments so that our results cannot possibly be attributed to a non-standard choice of evaluation set.

3.1 Algorithm Design Choices

3.1.1 Choice of Deterministic Permutation

A crucial design aspect is the choice of deterministic permutation, as it has a direct impact on the VI scores. Following our theoretical discussion in Section 2, we consider two types of permutations: a rank shift of $\lfloor n/2 \rfloor \bmod n$ (optimal) and an index shift of $\lfloor n/2 \rfloor \bmod n$ too (sub-optimal).

Indeed, the rank-shift version is optimal as Proposition 1 showed, but it requires sorting a length- n array, which takes $\mathcal{O}(n \log n)$ time, while an approximate solution based on indices instead of ranks takes only $\mathcal{O}(n)$ time.

In our extensive simulations, the difference in accuracy between the optimal and sub-optimal approaches tends to be small, as is their difference in compute time – even with 10,000 samples. We show that either singular permutation scheme is at least as accurate as traditional random permutation schemes (and often more accurate in challenging regimes), while being at least an order of magnitude faster than the classical 10-repetition permutation-based methods (and competitive with single random permutation VI scoring).

3.1.2 Evaluation Metric

We consider different scoring metrics to measure the impact that permuting a feature has on prediction: mean absolute error (MAE, the default), MSE and Root MSE (RMSE). Formally, in the case of MAE or MSE, for each feature we compute $(nq)^{-1} \sum_{i=1}^{nq} d(\hat{y}_i - \hat{y}'_i)$, where $d(x) = |x|$ with MAE and $d(x) = x^2$ with MSE, while \hat{y}_i and \hat{y}'_i are the master model’s predictions for the i -th observation before and after permuting a given feature, respectively (in C -class classification, they are length- C vectors). Importance scores are then normalized to sum to one. In the case of RMSE, the formula we stated is adjusted by simply taking its square root.

In order to isolate the effect of replacing multiple random permutations in traditional VI methods by a single, deterministic permutation in the VI method we propose, we will use the MSE of the prediction differences as scoring metric (instead of our MAE default), just like the classical baseline does in regression. In classification settings, we still compute the MSE of the difference in class probabilities, but, for comparability purposes, we deviate from the default scoring metric used by the baseline (drop in classification accuracy) and replace it by the change in (negative) Brier score, which is the MSE of predicted class probabilities versus their observed class.

Using the default scoring metrics instead (MAE for our methods, MSE / classification accuracy for the baseline) would in fact benefit our proposed methods compared to the classical baseline (see Appendix A), which out-of-the-box show significantly better ground-truth score recovery than scikit-learn’s implementation of Breiman-style permutation across nearly 200 scenarios.

Unlike other evaluation metrics that are often used (e.g. difference in MSE or classification accuracy before and after permuting a feature) the MAE/MSE/RMSE of the prediction differences ensures non-negative scores that can easily be normalized to a convenient, context-independent scale (e.g. the closed interval $[0,1]$) and remain maximally granular regardless of the master task (regression or classification).

3.1.3 Optional Feature Prescreening

Our method optionally supports a prescreening step that detects features whose perturbation has no measurable effect on the model’s predictions. While beneficial in settings with many noise variables that can be identified and discarded going forward, this step involves evaluating a single, deterministic perturbation per feature – conceptually similar to a lightweight, structured permutation. To avoid any ambiguity in the empirical results section and to ensure a completely “pure” one-permutation framework, we disabled prescreening in all experiments comparing our direct importance method to Breiman’s classical permutation importance. This guarantees that both methods operate on the full set of features, without any prior filtering, and that our reported advantages do not stem from auxiliary preprocessing steps. The resulting comparison is therefore strictly conservative and isolates the core mechanistic differences between the two approaches.

3.2 Empirical Simulation Setup

3.2.1 Scenarios Considered

We consider 192 individual scenarios in total, with varying sample size n , dimensionality p , noise level σ_ε (Gaussian standard deviation), feature correlation ρ and choice of ground-truth linear/nonlinear regression or classification model, each repeated 50 times. Concretely, we consider all combinations that result from the parameter values below, both for regression and binary classification:

- $n \in \{100, 1000, 10000\}$
- $p \in \{10, 100\}$
- $\sigma_\varepsilon \in \{0.1, 5\}$
- $\rho \in \{0, 0.3\}$, see Section 3.2.3 for more details

For the linear response case, we fix a random seed (123) and randomly sample from a multivariate Gaussian distribution with covariate matrix C (the identity matrix in the $\rho = 0$ case), and define our response variable as a randomly chosen (seeded) linear function of the design matrix obtained with each combination of the above parameters. We set half of the p covariates as noise features (unrelated with the response).

For the nonlinear response case we follow the process described in the next subsection to generate correlated uniformly-distributed features, and use a highly nonlinear function known as “Friedman function” that is often used for benchmarking purposes. This function is defined as

$$y \equiv f(x_1, x_2, x_3, x_4, x_5) = 10 \sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5$$

for $x_j \stackrel{i.i.d.}{\sim} \text{Unif}[0, 1]$ for all j , although in addition we also consider the case of correlated covariates (see Section 3.2.3). Since the standard Friedman function is defined for exactly five covariates, the remaining number of covariates are set as noise features, which gives the opportunity of testing a much sparser regression problem in the $p = 100$ regime.

In the case of classification, the response variable is simply binarized as follows: $y_{bin} = 1$ if $y > \text{median}(y)$ and 0 otherwise.

Thus, the data-generating process is conceptually consistent across linear/nonlinear and regression/classification scenarios, with the response variable adjusted accordingly to fit each scenario.

Lastly, being able to establish the “ground-truth” VI scores requires using master models that inherently provide this information. For that reason, we consider generalized linear models with identity or logit link (respectively, for regression or classification) and with or without ℓ_1 regularization to check if the VI methods considered can estimate well zero or small scores too, respectively.

3.2.2 VI Methods Considered

We test both variants of our proposed method: the one using the optimal rank-shift permutation, and the one using the simpler index-shift permutation, which acts as a cheaper approximation to the optimal permutation scheme. As baseline we consider the classical Breiman-style permutation importance methods implemented in standard Python libraries like scikit-learn, both using one random repetition (the fastest possible setup) and the much more common default of ten [Alonso and Carbó, 2022]. We use the same MSE-based scoring metric in all cases so that differences in results cannot be attributed to differences in scoring metrics. However, for completeness, we include in Appendix A the results obtained using default scoring metrics, which are even more favorable to our proposed methods.

3.2.3 Correlated Features Design

In order to evaluate the performance of our proposed method relative to Breiman’s under correlated features, we consider a block covariance matrix with stronger correlation among informative features ($\rho = 0.3$) to reflect their shared underlying factor structure, weaker correlation among noise features ($\rho = 0.3/2 = 0.15$), and zero cross-correlation between the informative and uninformative feature blocks. This type of covariance design is common in the literature, see e.g. Loh [2012].

Therefore, we intentionally included (rather mild) correlations among noise features, which can slightly affect ordinary least squares (OLS) coefficient estimates and, consequently, the performance of VI methods. This choice creates challenging scenarios that act as a stress-test for VI approaches: while Breiman’s methods exhibit substantially reduced accuracy under these slightly hardened conditions, our direct VI methods consistently recover the true variable importance. By incorporating correlated noise features, we not only reflect a plausible structure encountered in real datasets, but also provide a less idealized evaluation framework in which the robustness of our methods is clearly demonstrated.

Although the correlations we introduced are fairly moderate, they already prove challenging enough for the traditional Breiman-style baseline and thus testing more extreme correlations does not appear necessary and could compromise numerical stability for no additional gain in insight.

To generate correlated uniform features on $[0, 1]$, we first construct correlated Gaussian variables and then apply the probability integral transform. Let $C \in \mathbb{R}^{p \times p}$ denote the desired covariance matrix. We obtain its Cholesky factorization $C = LL^\top$, where L is lower triangular. If $X \in \mathbb{R}^{n \times p}$ has independent standard Gaussian entries, $X_{ij} \sim \mathcal{N}(0, 1)$, then $Z = XL^\top$ has rows $Z_i \sim \mathcal{N}(0, C)$. This follows from

$$\text{Var}(Z) = \text{Var}(XL^\top) = L \text{Var}(X) L^\top = LIL^\top = C,$$

showing that the linear transformation L^\top injects the desired correlation structure into otherwise independent Gaussian features.

To obtain correlated uniform variables, we apply the standard normal CDF componentwise,

$$U_{ij} = \Phi(Z_{ij}),$$

which yields $U_{ij} \sim \text{Unif}(0, 1)$ by the probability integral transform. Importantly, Φ is strictly increasing, and monotone transformations preserve all rank-based dependence measures (e.g. Spearman or Kendall correlations). Thus, the uniform variables U retain the dependence structure implied by C while matching the marginal distributional assumptions of the Friedman benchmark function.

3.2.4 Ensuring Breiman’s VI Scores are Non-Negative

As we mentioned previously, classical permutation-based methods measure how much an external performance metric (e.g., mean squared error or classification accuracy) deteriorates when a feature is permuted; we refer to this as a *performance-drop* score. While appropriate when the goal is to quantify contribution to a chosen metric, performance-drop scores can be unstable and even negative in finite samples: permuting a feature may, by chance or due to model weakness, *improve* the metric. Such negative values are hard to interpret and complicate normalization to a chosen context-independent scale (e.g. [0,1]).

Due to its widespread use, our chosen Breiman-style implementation of permutation-based VI is scikit-learn’s implementation. This implementation considers a widespread literal interpretation of Breiman’s method, and hence the returned values may be negative, reflecting spurious score improvements rather than meaningful negative importance. Because our ground-truth importance vectors are non-negative and normalized to sum to one, a compatible representation of Breiman-style importance scores requires transforming these raw effects into the same scale.

Although a sign-agnostic measure of how much predictions change when a feature is perturbed would be more principled (and motivates our own method), our aim here is to remain faithful to the widely used Breiman-style permutation framework. We therefore adopt only the minimally necessary modification to ensure meaningful downstream comparisons – such as maximum discrepancy with ground-truth importances (which could otherwise exceed 1). Specifically, and in line with common practice, we simply set negative permutation effects to zero before normalization. This minor post-processing step is conservative and fully consistent with Breiman’s intent: it is reasonable to assume the goal of permutation importance was not to reflect cases where random perturbations *improve* predictive performance.

To ensure a fair and conservative comparison, we applied the following minimal post-processing: all negative values were set to zero, and the resulting vector was normalized to sum to one. Formally, letting \mathcal{B}_j denote the raw permutation effect for feature j à la Breiman, we define

$$\tilde{\mathcal{B}}_j := \frac{\max\{\mathcal{B}_j, 0\}}{\sum_{k=1}^p \max\{\mathcal{B}_k, 0\}}.$$

This transformation preserves all positive information, removes negative artifacts incompatible with normalized importance scores, and is strictly favorable to Breiman’s method: features that receive negative permutation scores are treated as having zero importance rather than as detractors. This yields well-behaved importance profiles in $[0, 1]^p$ and prevents pathological maximum-discrepancy values exceeding 1, allowing meaningful comparison with ground truth, including bounded metrics.

3.2.5 Result Aggregation Across Scenarios

To aggregate results across scenarios while accounting for both within-scenario variability and between-scenario differences, we compute a combined standard error using a well-established variance-decomposition approach. For each scenario we first obtain its mean estimate and its empirical variance from repeated simulations (50 times each scenario). The total variance across all scenarios is then defined as

$$\text{Var}_{\text{combined}} := \underbrace{\mathbb{E}[\text{Var}_{\text{within},i}]}_{\text{average within-scenario variance}} + \underbrace{\text{Var}(\mu_i)}_{\text{between-scenario variance}}$$

where μ_i is the estimate for scenario i . The first term captures random fluctuations within each scenario, while the second captures systematic differences in the scenario means. The square root

of this total variance gives the combined standard error, and reported error bands use ± 2 times this value. This approach ensures that the final uncertainty quantification reflects both sources of variability, producing a single coherent measure of precision for the aggregated performance metrics.

4 Results

4.1 Regression

4.1.1 Linear Response

Table 1: Linear regression regime. Global results across 48 scenarios with varying sample size, dimensionality, noise level, feature correlation and choice of ground-truth regression model

Metric	Direct-Opt	Direct-Approx	Breiman (B=1)	Breiman (B=10)
Ground-truth cor	0.953 ± 0.008	0.941 ± 0.012	0.900 ± 0.037	0.910 ± 0.035
Max score diff	0.082 ± 0.019	0.093 ± 0.020	0.095 ± 0.024	0.091 ± 0.024
Time (ms)	24.9 ± 13.7	19.4 ± 10.7	17.1 ± 5.3	98.4 ± 32.4

Table 1 shows the global results for linear regression across 48 scenarios. They show that, under a linear regression regime (linear response estimated by OLS) the proposed direct variable importance approach is both more accurate and robust than classical Breiman-style permutation methods: both the method with optimal rank shift and its approximation (using a simpler index shift) display higher correlation with ground-truth scores on average, and beyond the corresponding two-standard-error bands in the case of the former. Maximum score discrepancies across features are on average comparable or slightly lower with our proposed methods as well, although in this case their advantage over classical baselines is not as significant. As Figures 3, 4, 5, and 6 illustrate, our proposed methods particularly shine (relative to the classical methods) under challenging circumstances: small sample size n , high-dimensionality ($n \approx p$), low signal-to-noise, or correlated-feature scenarios. This shows that while in idealized conditions both our methods and the classical baselines are similarly accurate, in less ideal scenarios the variability injected by random permutations tends to compromise performance, on average.

Despite their superior accuracy – especially in challenging conditions – our proposed methods remain efficient (on average, competitive with single-permutation Breiman’s method).

Overall, our direct variable importance deterministic methods are capable of recovering true variable importance patterns more accurately and, above all, more robustly across scenarios, without incurring significant computational cost.

4.1.2 Nonlinear Response

For the Friedman benchmarks we add i.i.d. Gaussian observation noise with standard deviation $\sigma_\varepsilon \in \{0.1, 5\}$.

As Table 2 illustrates, the results under nonlinear regression are very similar to those under the linear setup. While maximum score differences increase for all methods, our proposed method using a single, deterministic and optimal permutation remains the most accurate and robust on all accounts, while being significantly faster than the common 10-repetition Breiman-style permutation method.

Table 2: Nonlinear regression regime. Global results across 48 scenarios with varying sample size, dimensionality, noise level, feature correlation and choice of ground-truth regression model

Metric	Direct-Opt	Direct-Approx	Breiman (B=1)	Breiman (B=10)
Ground-truth cor	0.953 ± 0.007	0.941 ± 0.008	0.899 ± 0.036	0.910 ± 0.034
Max score diff	0.154 ± 0.019	0.176 ± 0.021	0.173 ± 0.021	0.168 ± 0.023
Time (ms)	25.3 ± 13.9	19.7 ± 11.0	17.0 ± 5.1	98.2 ± 31.8

Figures 7, 8, 9, and 10 show that individual per-scenario performance again mirrors the linear regime: our proposed single-permutation methods and the classical baselines are similarly accurate in idealized conditions, while the gap widens significantly in favor of our methods in more challenging (yet realistic) conditions.

4.2 Classification

4.2.1 Linear Response

We next evaluate all methods in a binary classification setting, keeping the same linear data-generating process as in the regression experiments and binarizing the response as described previously. A logistic regression model (with class balancing) serves as the master model, and importance scores are computed exactly as in the regression setting.

Table 3: Classification (linear response). Global results across 48 scenarios with varying sample size, dimensionality, noise level, feature correlation and choice of ground-truth classifier

Metric	Direct-Opt	Direct-Approx	Breiman (B=1)	Breiman (B=10)
Ground-truth cor	0.962 ± 0.010	0.947 ± 0.018	0.868 ± 0.052	0.882 ± 0.051
Max score diff	0.066 ± 0.019	0.067 ± 0.018	0.088 ± 0.026	0.083 ± 0.029
Time (ms)	27.1 ± 14.7	21.2 ± 11.5	26.2 ± 8.0	189.1 ± 60.4

Table 3 reports global performance aggregated across all linear-response classification scenarios. The proposed *Direct-Opt* and *Direct-Approx* methods remain highly accurate, achieving ground-truth correlations of 0.962 and 0.947, respectively, substantially outperforming Breiman’s permutation importance with $B = 1$ (0.868) and even $B = 10$ (0.882). Maximum score differences show a similar although less pronounced pattern, with our methods remaining below 0.07 on average, compared to 0.09 for Breiman with $B = 1$.

Runtime comparisons also remain favorable: for $B = 1$, Breiman’s speed is comparable to our methods, whereas for $B = 10$ it becomes an order of magnitude slower. Overall, these results confirm that the proposed direct importance estimators extend naturally to classification, preserving both accuracy and computational efficiency while avoiding the instability that many permutation-based methods (including Breiman’s) exhibit even in simple linear setups.

Figures 11, 12, 13, and 14 illustrate that in challenging conditions the performance of the classical baselines can deteriorate substantially, reaching low correlation values with ground-truth scores (below 0.40) while our optimal method maintains correlations around 0.90 in those same circumstances.

4.3 Nonlinear Response

Table 4: Classification (nonlinear response). Global results across 48 scenarios with varying sample size, dimensionality, noise level, feature correlation and choice of ground-truth classifier

Metric	Direct-Opt	Direct-Approx	Breiman (B=1)	Breiman (B=10)
Ground-truth cor	0.953 ± 0.009	0.942 ± 0.013	0.863 ± 0.047	0.879 ± 0.045
Max score diff	0.153 ± 0.021	0.162 ± 0.020	0.173 ± 0.022	0.168 ± 0.023
Time (ms)	27.0 ± 14.6	20.9 ± 11.4	26.0 ± 7.8	185.1 ± 57.3

Table 4 shows that the results with a logistic regression master model used to forecast a nonlinear (binarized) response are similar to the linear response case. However, as we observed in the nonlinear regression case as well, maximum score differences are inflated, but our proposed methods remain the most accurate in this respect too.

In terms estimation efficiency, our methods remain competitive with 1-repetition Breiman importance, while being an order of magnitude faster compared with the more common 10-repetition setup.

See Figures 15, 16, 17, and 18 for individual results per scenario.

4.4 Global Result Conclusions

4.4.1 Accuracy

Our proposed methods are on average at least as accurate as Breiman-style VI methods across nearly 200 scenarios, while substantially more so in the more challenging scenarios tested, confirming a superior bias-variance tradeoff. The difference in mean accuracy is smallest in linear regression problems (0.953 vs. 0.900), while it is largest in binary classification (0.962 vs. 0.868). Our methods are remarkably more stable across scenarios as well – even compared with 10-repetition Breiman scoring.

4.4.2 Runtime

For small datasets (low n and p), our direct variable importance calculation – even the variant using optimal rank shifts – is up to $9\times$ faster than scikit-learn’s permutation importance implementation with $B = 1$. This suggests that scikit-learn’s implementation includes overhead that only amortizes for larger datasets. Indeed, our experiments indicate that such overhead may become advantageous for $n \gtrsim 10000$ and moderate $p \in [11, 99]$, say $p \gtrsim 50$. Regardless, our method remains competitive with even the fastest possible evaluation of Breiman’s method (the rather uncommon $B = 1$) across all tested regimes (including $n = 10000$ with $p = 100$), and its current speed reflects a conservative baseline, as further low-level optimization is possible.

Crucially, for the standard $B = 10$ ensemble (the typical benchmark for stable scores), our method is typically at least an order of magnitude faster across scenarios.

5 From Variable Importance to Model Stress-Testing

Beyond efficiently producing deterministic, reliable VI scores, our framework naturally extends to *model stress-testing*.

In this context, stress-testing a model – or more formally, a (model, data) pair – refers to evaluating the impact that the degradation of each feature would have on model performance. Data degradation can happen in practice due to several reasons, and some of the most common include covariate shifts, sensor failure, human error, and adversarial attacks.

Although superficially the problem at hand might appear identical to estimating variable importance (and it is indeed related), it is fundamentally different, because in order to determine the true risk that a compromised feature poses to a given fitted model we must, in most cases, account for network effects and contagion risk: the degradation of a feature rarely takes place in isolation; in highly interconnected systems (like financial markets or macro-economies), closely related features will likely be impacted to some extent too, creating contagion risk which may increase the *systemic* importance of the original feature of interest.

A simple analogy will suffice to illustrate the difference between direct and systemic VI: the systemic economic importance of a component often exceeds its direct contribution. Consider Spain: it accounts for 8.7% of the European Union’s GDP (Eurostat, 2023 figures) so its direct economic importance within the EU could reasonably be estimated as that number; however, it is not hard to imagine that its systemic importance might be substantially higher due to significant trade flows and financial integration with larger European economies like France, Italy or Germany, as well as its role as a key gateway between Europe, North Africa, and Latin America. Therefore, estimating the systemic economic importance of Spain within the EU requires taking into account the strength of its ties with the rest of member states, which is something that direct VI measures (e.g. Breiman-style permutation VI) cannot do by definition.

Unlike classical sensitivity analyses that treat input changes in isolation (e.g. “x decreases by 10%”), systemic importance quantifies the model’s vulnerability to correlated covariate corruption or covariate-shift scenarios in which a disturbance to one predictor co-occurs with predictable changes in its correlated neighbors.

Observe that model class VI methods like LOCO or ghost variables, despite accounting for feature correlation, cannot estimate systemic importance: in fact, by discounting the importance of features that are more correlated with the rest of covariates, they do the opposite. These methods are not incorrect, they are simply designed to estimate a different quantity (and thereby answer a different question).

By allowing our permutation-based framework to account for feature correlations we seamlessly switch from direct to systemic VI estimation, where more strongly correlated features have their importance scores boosted, not deflated.

We define systemic importance as the expected disruption to model predictions when a single feature is perturbed and that perturbation is allowed to *propagate* to other covariates according to their empirical correlations. The resulting systemic importance vector is non-negative and normalized to sum to one, and admits a direct interpretation as the fraction of total predictive disruption attributable to each feature when correlated propagation is accounted for. We formalize this notion next.

Definition 2 (Systemic Variable Importance). *With the same setup as in Definition 1, the estimated systemic importance score s_k of the k -th variable x_k is the model prediction disruption that occurs when x_k is permuted, and this perturbation is propagated to all other covariates.*

1. Propagation Rule: When the j -th feature, \mathbf{x}_j , is perturbed (permuted) to \mathbf{x}'_j , the permuted value of any other feature \mathbf{x}_k ($k \neq j$) in the permuted data matrix X' is simultaneously updated according to the following rule:

$$\mathbf{x}'_k \leftarrow \mathbf{x}'_k + \text{cor}(\mathbf{x}_k, \mathbf{x}_j) \times (\mathbf{x}'_j - \mathbf{x}_j) \quad (16)$$

where $\text{cor}(\mathbf{x}_k, \mathbf{x}_j)$ is the empirical correlation between features k and j (e.g. Spearman's ρ) and $(\mathbf{x}'_j - \mathbf{x}_j)$ is the vector of changes induced by the permutation of x_j . In order to disregard spurious correlations, the propagation only occurs if $|\text{cor}(\mathbf{x}_k, \mathbf{x}_j)|$ exceeds a data-driven tolerance threshold.

2. Systemic Score: The raw systemic importance score s_k^{raw} is the resulting expected change in prediction when only x_k is perturbed and the full propagation rule is applied. The final normalized systemic importance score s_k is calculated as:

$$s_k := \frac{s_k^{\text{raw}}}{\sum_{j=1}^p s_j^{\text{raw}}} \quad (17)$$

Decomposition for Interpretation: For analytical purposes, the normalized systemic importance score s_k can be decomposed into its direct importance score (d_k) and its indirect importance score (i_k):

$$s_k = d_k + i_k \quad (18)$$

The direct score d_k is the normalized performance degradation from the standard, isolated permutation of x_k (i.e., when no propagation occurs). The indirect score i_k is the difference that captures the network amplification or dampening of x_k 's total importance (relative to the rest of covariates), reflecting its interconnectedness with the feature set.

Put simply, the systemic importance of a feature is equal to its direct importance plus an indirect component which is dynamically computed by propagating the perturbations of all other covariates proportionally to their correlation with the feature of interest, so that correlated covariates amplify or dampen the original (direct) perturbation.

Because our systemic importance measure relies on correlations, all covariates must either be numeric or have a numeric encoding in the master model: ordinal encoding (the ideal case if there is a meaningful order), one-hot encoding or target-encoding. In that case, our suggested rank-based correlation default metric (Spearman) is still appropriate for the purposes of detecting non-parametric feature association. We emphasize that systemic importance reflects (and is bounded by) model behavior, rather than causal structure.

Observe that with this simple framework there is no need to simulate dynamic systems, impose any kind of parametric model or assume a particular multivariate feature distribution. As such, it does not intend to necessarily replace much more complex tools specifically designed to model and analyze shocks in dynamic systems like Dynamic Stochastic General Equilibrium (DSGE) or Structural Equation Models (SEM). Instead, it is intended to be a fast and transparent post-hoc tool built on top of classical permutation scoring, which can be applied to any predictive model (all it takes is knowledge of ifs evaluation function, no need to know any internals) to provide a more holistic view of the importance of each feature in a model taking into account their correlation as a first-order approximation to their dependence structure.

5.1 Statistical Calibration of the Propagation Threshold

To ensure that systemic variable importance (SVI) captures structural dependencies rather than numerical artifacts, we implement a non-parametric calibration mechanism to establish a minimum correlation threshold, τ . In any finite dataset, features may exhibit non-zero empirical correlations due to sampling error, even when their true population-level associations are zero.

5.1.1 Hypothesis Testing

We define the global null hypothesis of independence, H_0 , such that the joint distribution of any two features X_i and X_j factors into the product of their marginals: $f(x_i, x_j) = f(x_i)f(x_j)$. To

maintain the integrity of the forensic audit, the propagation of perturbations is restricted to feature pairs whose association is statistically significant relative to the permutation distribution. We seek to learn a threshold τ that provides a family-wise safeguard, ensuring that the probability of any spurious correlations exceeding this limit is controlled at a global significance level α :

$$\mathbb{P} \left(\max_{1 \leq i < j \leq p} |\hat{\rho}_{ij}| > \tau \mid H_0 \right) \leq \alpha \quad (19)$$

5.1.2 Non-parametric Estimation

In practice, we estimate τ in-sample using an efficient permutation-based approach. Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be the training matrix. We generate a null representation by independently permuting the row indices of each column, creating a synthetic dataset $\tilde{\mathbf{X}}$. This procedure preserves the marginal distributions $f(x_k)$ for all $k \in \{1, \dots, p\}$ while stochastically breaking all inter-feature dependencies.

Under H_0 , the $M = p(p - 1)/2$ pairwise correlations in the upper triangular matrix are exchangeable. We therefore define τ as the $(1 - \alpha)$ quantile of this pooled empirical distribution:

$$\tau = Q \left(\{|\hat{\rho}(\tilde{X}_i, \tilde{X}_j)|\}_{1 \leq i < j \leq p}; 1 - \alpha \right) \quad (20)$$

where $Q(S; p)$ is the empirical quantile function defined as the infimum of the set of values whose empirical cumulative distribution F_S is at least p :

$$Q(S; p) = \inf \{x \in \mathbb{R} : F_S(x) \geq p\} \quad (21)$$

Note that utilizing the inverted cumulative distribution function ensures that τ is an observed value from the permutation set, which provides a conservative family-wise safeguard that strictly respects the empirical noise floor.

By pooling all pairwise entries from the permuted matrix, τ provides a robust estimate of the global noise floor while accounting for the multiple comparisons inherent in auditing the full feature space. This approach provides a computationally efficient estimator for strong Family-Wise Error Rate (FWER) control (i.e., control under any configuration of the true dependencies). Correlations exceeding τ are treated as systemic proxies, whereas those below τ are dismissed as sampling artifacts. To ensure the audit remains deterministic and reproducible, the internal permutation seed is fixed.

Now, although our initial motivating example involved covariate shifts, sensor failure or the collapse of a major European economy, model-stress testing has another, equally relevant application: evaluating model fairness. This application is particularly salient for financial regulation and model governance, where evaluating model fairness against protected attributes is mandatory. We will assess the systemic impact of sensitive attributes in the two real-world case studies that we introduce next.

5.2 Case Study A: Debt-to-Income Ratio Estimation

The Boston HMDA dataset is a well-known dataset with 2380 observations and 12 features collected in the 1990s that derived from the federal Home Mortgage Disclosure Act (HMDA) data, used extensively in econometrics to study mortgage lending patterns, especially racial bias, by analyzing applicant features (e.g. debt ratios, job, age, or race) against loan outcomes (approval/denial). This dataset, available in the author's GitHub repository (<https://github.com/adc-trust-ai/>)

`trust-free`), is often used to explore if lenders treat minorities differently, finding persistent disparities even after controlling for financial factors [Munnell et al., 1996, Bostic, 1997].

In our case, however, we will use this dataset to test, on the one hand, our proposed direct VI method in terms of the usual accuracy metrics, and on the other hand, the systemic VI method that we propose, with a view to model fairness.

We will estimate the debt-to-income ratio of mortgage applicants in the Boston HMDA dataset, based on 12 personal attributes, spanning 5 numeric and 7 binary. In the process, we take the chance to dispel any lingering doubts about the applicability of our proposed methods to non-continuous features by testing an extreme continuity violation: binary features.

The response variable `dir` has the following empirical order statistics: $\min = 0$, $Q1 = 0.28$, $\text{median} = 0.33$, $Q3 = 0.37$, and $\max = 3$. It is therefore heavily right-skewed, which typically makes estimation by standard regression methods more challenging – not an issue for our purposes.

Since we wish to be able to define ground-truth VI scores to serve as benchmark in our experiments, we choose unregularized and ℓ_1 -regularized linear models to estimate the debt-to-income ratio of mortgage applicants in the Boston HMDA dataset. To demonstrate that our proposed method, just like the classical baseline, can be applied to *any* forecasting model as long as we have access to its prediction function, our choice of ℓ_1 -regularized linear model is TRUST [Dorador, 2025d], a custom algorithm that fits relaxed Lasso models [Meinshausen, 2007] in the leaves of an axis-aligned regression tree, recently implemented in the `trust-free` Python library. To preserve ground-truth scores, the depth of the TRUST trees is purposely kept at zero so that only a relaxed Lasso model is fitted (to the entire training set). A black-box benchmark – Random Forest (RF) [Breiman, 2001a] – is also included to showcase that our VI methods can be applied to black-box models as well – and that is arguably their most useful application, in fact. Although it might be debatable whether the VI scores that RF automatically provides in most implementations should be considered “ground-truth” for that model or just an estimate based on somewhat arbitrary criteria, we confirm that our proposed methods achieve a similar or slightly higher correlation with those scores compared to the classical baseline methods (around 0.98 in all cases).

All master models are run with default parameters, which in the case of RF involves an ensemble of 100 trees as implemented in standard Python libraries like scikit-learn.

We randomize the order of the rows in the original dataset and perform 10-fold cross-validation to obtain repeated measures of master model accuracy as well as direct VI method accuracy.

For consistency with our 10-fold cross-validation, we learn the spurious correlation threshold using the training set while we estimate SVI on the 10% held-out test set.

5.2.1 Direct VI

In this dataset, as Table 5 shows, the predictive performance of a depth-0 TRUST tree is essentially identical to that of a plain OLS solver, but fewer nonzero coefficients are estimated, aiding interpretability. Meanwhile, a black-box model like RF only manages to achieve similar accuracy but employed 100 trees with over 125,000 logic rules in total (although many of those trees were possibly redundant [Dorador, 2025a]).

Tables 6 and 7 show that our proposed methods (using MSE scoring as in all experiments) have a very similar performance to the classical baselines but are more stable and 14 to 105 times faster when compared to the $B = 1$ and $B = 10$ classical variants, respectively.

Tables 15 and 16 in Appendix A show the results employing default scoring functions for our methods and the classical baseline (but in the case of the latter it is the same scoring function in the case of regression). Those results highlight that correlation with ground truth is rather stable

Table 5: Forecasting out-of-fold average master model results with two standard error bands in the Boston HMDA dataset

Metric	TRUST	OLS	RF
R^2	0.48 ± 0.08	0.49 ± 0.10	0.46 ± 0.08
MAE	0.048 ± 0.002	0.048 ± 0.002	0.048 ± 0.002
Nonzero coef.	10.4 ± 0.1	12.0 ± 0.0	125802.4 ± 99.9

Table 6: Results of different VI methods under a sparse linear regression master estimator (TRUST) – Boston HMDA dataset

Metric	Direct-Opt	Direct-Approx	Breiman (B=1)	Breiman (B=10)
Ground-truth cor	0.944 ± 0.005	0.944 ± 0.005	0.943 ± 0.007	0.944 ± 0.007
Max score diff	0.473 ± 0.025	0.473 ± 0.024	0.472 ± 0.030	0.471 ± 0.028
Mean score diff	0.079 ± 0.004	0.079 ± 0.004	0.079 ± 0.005	0.079 ± 0.005
Time (ms)	1.63 ± 0.04	1.45 ± 0.03	5.06 ± 0.08	33.84 ± 0.43

across scoring functions, but that is not the case for a more extreme metric like the maximum score difference.

Lastly, to further test the stability of the traditional VI methods on real datasets we execute them ten times and check the top 5 features that they report each time. Table 19 in Appendix B shows that the classical baseline employing only one repetition yields very unstable rankings for the top 5 features: every single run yields a different ranking. Employing 10 repetitions reduces this instability to some degree although it is still notable: 7 distinct rankings were produced in 10 runs. In stark contrast, our deterministic framework is designed to yield the exact same scores (and, hence, rankings) every single time it is run. In addition, only our proposed methods (either with optimal or sub-optimal permutation) recovered the ground-truth top 5 components.

5.2.2 Systemic VI

We will test two scenarios: one where we use all of our features to estimate the debt-to-income ratio of mortgage applicants via unregularized linear regression (plain OLS), and another scenario where we force a sparse linear regressor (TRUST) to not use the sensitive attribute `black` (1 if the applicant is African American, 0 otherwise). The purpose of this setup is to show that even if a model does not directly use a sensitive feature, the model might still rely on the sensitive attribute via proxy variables, which is something that traditional Breiman-style VI methods cannot possibly detect.

- With a plain OLS master, the systemic relative importance of feature `black` in the fitted model is estimated to be 0.44%, which can be decomposed as the sum of its direct (0.09%) and indirect (0.35%) relative importance scores. Thus, its systemic importance is about 5 times larger than its direct importance (which is what classical Breiman-style methods would only be able to estimate). Using MAE instead of MSE, we obtain the following importance decomposition: $2.31\% = 1.21\% + 1.10\%$, which in this case implies a smaller multiplier of 2.
- With a sparse linear regressor master, the systemic relative importance of feature `black` in the fitted model is estimated to be 0.16%, which is due entirely to its indirect component, as

Table 7: Results of different VI methods under an unregularized master linear model – Boston HMDA dataset

Metric	Direct-Opt	Direct-Approx	Breiman (B=1)	Breiman (B=10)
Ground-truth cor	1.000 ± 0.000	1.000 ± 0.000	0.999 ± 0.000	0.999 ± 0.000
Max score diff	0.075 ± 0.010	0.076 ± 0.010	0.074 ± 0.017	0.073 ± 0.016
Mean score diff	0.013 ± 0.002	0.013 ± 0.002	0.013 ± 0.003	0.013 ± 0.002
Time (ms)	0.48 ± 0.01	0.39 ± 0.01	5.84 ± 0.11	42.06 ± 0.37

the direct importance of the feature in the model is exactly 0. Using MAE instead of MSE, we obtain a larger systemic importance score of 1.41%, again entirely due to network effects.

Hence, even if the sensitive attribute `black` is not used directly by the model, our systemic importance method identified that the model still relies on it indirectly through correlated covariates beyond any spurious correlation doubts (tolerance quantile: 0.99), such as `lvr` (ratio of size of loan to assessed value of property), `ccs` (consumer credit score) or `deny` (1 if the applicant had a mortgage denied, 0 otherwise). Our method quantified this indirect reliance on the protected attribute to be between 0.16% and 1.41% of the total predictive reliance of the model, despite not being present in the model as such. This reliance on a protected attribute via proxy variables would be an undetected model risk and potential regulatory non-compliance when assessed using traditional VI methods.

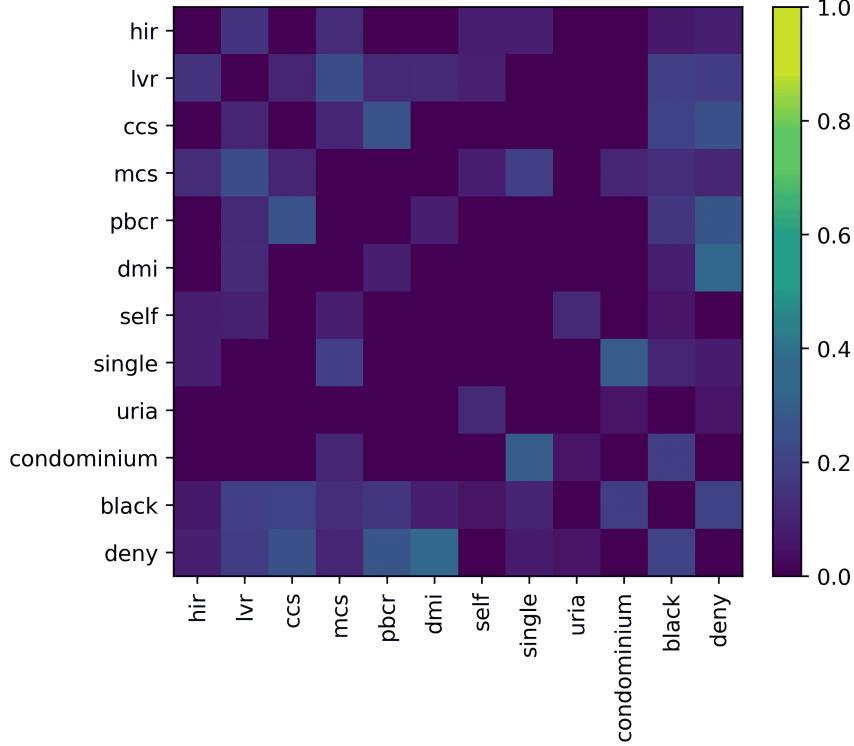


Figure 1: Feature interconnections in the Boston HMDA dataset (tolerance quantile: 0.99)

Observe that the protected attribute `black` shows Spearman correlations of about 0.2 or higher

with 4 of the 12 features included in the dataset: `deny` (whether the mortgage application was denied), `css` (consumer credit score from 1 to 6, where a lower value represents a better score), `lvr` (loan-to-value ratio), and `condominium` (whether the applicant lives in a condominium as opposed to a detached single-family home). This means that the model can “reconstruct” a sizable proportion of the racial information using those proxies, even when the race variable is explicitly excluded from the model.

More generally, taking into account all features having Spearman correlation with `black` exceeding the 0.99 quantile from a non-parametric null distribution to define its significant proxies, we note that 90% of the total ground-truth score contributions in a model that excludes the race variable is nonetheless influenced by race.

For completeness, we also tested fitting the sparse linear regression master with all covariates and the results are similar to the OLS case. The profile comparison report made with the `trust-free` library shows that the direct counterfactual effect on prediction of `black` indeed typically is about 1% (see Figure 19 in Appendix D).

5.3 Case Study B: Credit Risk Forecasting

The Statlog (German Credit Data) dataset is a classic machine learning benchmark with 1000 records and 20 covariates, used to classify loan applicants as high or low credit risks, featuring mixed numerical (7) / categorical (13) features like credit history, duration, amount, housing, job, and demographics. This dataset is available on the UCI Repository, as well as in the author’s GitHub repository (<https://github.com/adc-trust-ai/trust-free>) with more convenient decoded feature names.

Like we did in the regression case study of the previous section, we will use this dataset to test, on the one hand, our proposed direct VI method in terms of the usual accuracy metrics, and on the other hand, the systemic VI method that we propose, with a view to model fairness.

As it is standard with this benchmark dataset, we will train a classifier to learn to classify loan applicants as either low (`class = 0`) or high (`class = 1`) risk based on their personal attributes. In our records, only 30% are classified as high risk, so there is a noticeable class imbalance. This imbalance, together with the misclassification cost matrix that accompanies this dataset whereby a false negative is 5 times as costly as a false positive, suggests reporting the PR-AUC curve as the main performance metric relative to which compare master models, with the ROC-AUC as secondary metric. A custom metric such as $A = (5FN + FP + 0.001)^{-1}$ that reflects the exact misclassification cost asymmetry would also be a reasonable choice.

Again, being able to define ground-truth VI scores requires using a model class in which they are undisputedly identifiable, such as linear models (in log-odds). Hence we consider ℓ_1 -regularized logistic regression (with a strong penalty so that most coefficients are zeroed-out) and unregularized logistic regression (so that all feature coefficients are kept nonzero). We also consider RF as a black-box benchmark to illustrate one of the main use cases of VI methods. As in the previous regression case study, we confirm that our proposed methods achieve a similar correlation with the RF-native VI scores compared to the classical baseline methods (0.50-52 in all cases).

The rest of the experimental setup is identical to the previous case study.

5.3.1 Direct VI

In this case, as Table 8 shows, the performance of the strongly-regularized logistic regression model is surprisingly competitive given that it only used 10% of the available features, on average. As in

the previous case study, the RF model only managed to achieve a performance comparable to the (unregularized) linear baseline but employed 100 trees with nearly 12,000 logic rules in total.

Table 8: Forecasting out-of-fold average master model results with two standard error bands in the German credit dataset

Metric	ℓ_1 -Logistic	Logistic	RF
$PR - AUC$	0.51 ± 0.07	0.61 ± 0.07	0.63 ± 0.07
$ROC - AUC$	0.73 ± 0.03	0.79 ± 0.03	0.79 ± 0.03
Nonzero coef.	2.2 ± 0.4	20 ± 0.0	18222.4 ± 78.2

Tables 9 and 10 show that our proposed methods (using MSE / negative Brier scoring as in all experiments) have comparable or better performance with respect to classical baselines while being more stable and 8 to 52 times faster in relation to the $B = 1$ and $B = 10$ classical variants, respectively.

Tables 17 and 18 in Appendix A show the results employing default scoring functions for our methods and the classical baseline. In that case, results favor very significantly our proposed methods compared to baseline.

Table 9: Results of different VI methods under a sparse logistic regression master estimator – German credit dataset

Metric	Direct-Opt	Direct-Approx	Breiman (B=1)	Breiman (B=10)
Ground-truth cor	0.996 ± 0.002	0.997 ± 0.002	0.999 ± 0.001	0.998 ± 0.001
Max score diff	0.072 ± 0.025	0.069 ± 0.024	0.035 ± 0.020	0.044 ± 0.021
Mean score diff	0.007 ± 0.003	0.007 ± 0.002	0.004 ± 0.002	0.004 ± 0.002
Time (ms)	0.82 ± 0.02	0.67 ± 0.01	5.27 ± 0.09	34.94 ± 0.31

Table 10: Results of different VI methods under an unregularized logistic regression master model – German credit dataset

Metric	Direct-Opt	Direct-Approx	Breiman (B=1)	Breiman (B=10)
Ground-truth cor	0.875 ± 0.015	0.900 ± 0.013	0.806 ± 0.036	0.773 ± 0.077
Max score diff	0.255 ± 0.031	0.200 ± 0.025	0.196 ± 0.084	0.196 ± 0.031
Mean score diff	0.029 ± 0.002	0.025 ± 0.002	0.035 ± 0.004	0.036 ± 0.003
Time (ms)	0.76 ± 0.02	0.67 ± 0.02	5.16 ± 0.07	34.77 ± 0.30

Similarly to the previous case study, Table 20 in Appendix B shows that across 10 independent runs on the German Credit dataset, both our optimal and approximate VI methods recovered the exact ground-truth top-5 feature set in all runs. On the other hand, both classical baselines employing one or even ten repetitions yielded very unstable rankings for the top 5 features, producing 9 and 10 distinct top 5 rankings, respectively.

5.3.2 Systemic VI

Similar to the previous case study in the context of regression, we will test two scenarios: one where we use all of our features, and another scenario where we guarantee the sparse estimator

does not use the sensitive attribute, in this case `Sex-Marital_status`. The purpose of this setup is to show that even if a model does not directly use a sensitive feature, the model might still rely on the sensitive attribute via proxy variables, which is something that traditional Breiman-style VI methods cannot possibly detect, as they ignore feature correlations.

- With an unregularized logistic regression master, the systemic relative importance of feature `Sex-Marital_status` in the fitted model is estimated to be 3.24%, which can be decomposed as the sum of its direct (2.09%) and indirect (1.15%) relative importance scores. Thus, its systemic importance is about 50% larger than its direct importance. Using MAE instead of MSE, we obtain the following importance decomposition: $4.94\% = 4.20\% + 0.74\%$.
- With a sparse logistic regression master that excludes the protected feature, the systemic relative importance of said feature in the fitted model is estimated to be exactly 0 as well. Using MAE instead of MSE, the estimated systemic importance is still 0, showing effective safeguarding by exclusion, in this case.

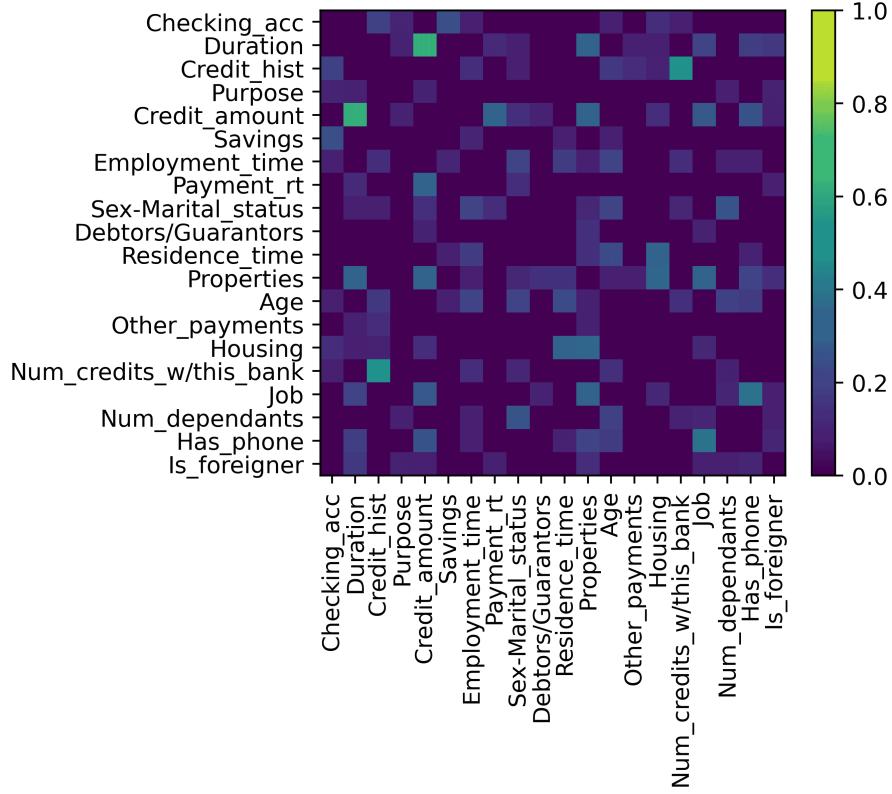


Figure 2: Feature interconnections in the German credit dataset (tolerance quantile: 0.99)

Hence, in this case study, our proposed systemic importance extension to our single-permutation VI method allowed us to audit the sparse fitted model and confirm that it indeed does not rely to any meaningful extent on a protected attribute like sex or marital status.

6 Conclusion

As machine learning models become increasingly central to high-stakes decision-making in finance and macroeconomic policy, the tools used to interpret and stress-test them must meet rigorous standards of stability, efficiency, and reliability. In this paper, we challenged the long-standing convention that permutation-based VI requires repeated random sampling to be effective.

We demonstrated that a single, deterministic, and optimal permutation is sufficient to recover meaningful variable importance scores and, both in theory and practice, offers clear advantages over traditional stochastic baselines. By removing the noise inherent in random sampling, our Direct Variable Importance (DVI) method yields scores that are strictly deterministic and reproducible – an essential property for model auditing, validation, and regulatory compliance. At the same time, our DVI method achieves substantial computational savings, with speedups of up to two orders of magnitude relative to standard permutation ensembles, making it scalable to high-dimensional models and intensive stress-testing pipelines. Across nearly 200 scenarios, our DVI method consistently matches or exceeds classical methods in ground-truth recovery, offering particularly strong gains in challenging regimes such as small sample sizes, high dimensionality, and low signal-to-noise ratios with correlated features.

Beyond contributing a faster, more reliable alternative to classical permutation-based variable importance methods, we introduced the concept of Systemic Variable Importance (SVI). In interconnected systems – whether they be financial markets prone to contagion or credit scoring models relying on proxy variables – evaluating features in isolation is insufficient and can be misleading. By explicitly integrating feature correlations into the permutation mechanism, our SVI method complements direct importance measures by providing a novel lens through which to quantify a model’s vulnerability to propagated shocks in a simple, non-causal, and transparent way. Our case studies on the Boston HMDA and German Credit datasets highlighted the practical power of this metric, revealing how models may indirectly rely on protected characteristics through correlation networks, a risk that traditional variable importance methods fail to detect.

In summary, this work introduces a unified, model-agnostic framework for both explaining model behavior and assessing its vulnerabilities. By shifting variable importance from randomized estimation toward deterministic stress-testing, we offer practitioners and regulators a tool that is faster, more transparent, and better aligned with the systemic nature of model risk.

Future work will explore computationally efficient ways to extend our SVI framework beyond pairwise correlations to capture richer dependence structures, further strengthening its applicability to real-world model stress-testing and regulatory audits.

References

- A. Alonso and J.M. Carbó. Accuracy of explanations of machine learning models for credit decisions. *Banco de España, Documentos de Trabajo N.^o 2222*, 2022.
- R. W. Bostic. The Role of Race in Mortgage Lending: Revisiting the Boston Fed Study. *Federal Reserve System. Board of Governors. Finance and Economics Discussion Series, Division of Research and Statistics - Division of Monetary Affairs*, 97(2), 1997.
- L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001a.
- L. Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16:199–231, 2001b.

- P. Delicado and D. Peña. Understanding complex predictive models with ghost variables. *TEST*, 32:107–145, 2023.
- A. Dorador. Theoretical and empirical advances in forest pruning. *Conference on Parsimony and Learning, Proceedings of Machine Learning Research (PMLR)*, 280:615–651, 2025a.
- A. Dorador. The normal equations are normally faster: Benchmarking ordinary least squares solvers for $n \geq p$. *SSRN preprint:5811842*, 2025b.
- A. Dorador. The normal equations strike back: The $n < p$ case. *SSRN preprint:5884203*, 2025c.
- A. Dorador. TRUST: Transparent, Robust and Ultra-Sparse Trees. *arXiv preprint:2506.15791*, 2025d.
- A. Fisher, C. Rudin, and F. Dominici. 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*, 20:1–81, 2019.
- N. Föge, L. Schmid, M. Ditzhaus, and M. Pauly. A central limit theorem for the permutation importance measure. *arXiv preprint:2412.13020*, 2024.
- G. Hooker, L. Mentch, and S. Zhou. Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. *Statistics and Computing*, 31, 2021.
- J. Lei, M. G’Sell, A. Rinaldo, R. Tibshirani, and L. Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.
- W.-Y. Loh. Variable selection for classification and regression in large p , small n problems. *Probability Approximations and Beyond, Lecture Notes in Statistics-Proceedings*, 205:133–157, 2012.
- W.-Y. Loh and P. Zhou. Variable importance scores. *Journal of Data Science*, 19(4):569–592, 2021.
- S. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017)*, pages 4768–4777, 2017.
- N. Meinshausen. Relaxed lasso. *Computational Statistics and Data Analysis*, 52:374–393, 2007.
- A. Munnell, G. Tootell, L. Browne, and J. McEneaney. Mortgage lending in boston: Interpreting hmida data. *American Economic Review*, pages 25–53, 1996.
- L. S. Shapley. *A Value for n-Person Games. Contributions to the Theory of Games*. Princeton University Press, 1953.
- C. Strobl, A-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 9(307), 2008.

A Results with Default Scoring Metrics

Table 11: Linear regression regime. Global results across 48 scenarios with varying sample size, dimensionality, noise level, feature correlation and choice of ground-truth regression model

Metric	Direct-Opt	Direct-Approx	Breiman (B=1)	Breiman (B=10)
Ground-truth cor	0.996 ± 0.002	0.992 ± 0.004	0.900 ± 0.037	0.910 ± 0.035
Max score diff	0.007 ± 0.002	0.011 ± 0.003	0.095 ± 0.024	0.091 ± 0.024
Time (ms)	25.0 ± 13.7	19.3 ± 10.7	17.9 ± 5.4	104.8 ± 33.1

Table 12: Nonlinear regression regime. Global results across 48 scenarios with varying sample size, dimensionality, noise level, feature correlation and choice of ground-truth regression model

Metric	Direct-Opt	Direct-Approx	Breiman (B=1)	Breiman (B=10)
Ground-truth cor	0.997 ± 0.001	0.993 ± 0.003	0.899 ± 0.036	0.910 ± 0.034
Max score diff	0.013 ± 0.004	0.015 ± 0.005	0.173 ± 0.021	0.168 ± 0.023
Time (ms)	24.1 ± 13.2	18.7 ± 10.3	17.3 ± 5.2	101.9 ± 32.0

Table 13: Classification (linear response). Global results across 48 scenarios with varying sample size, dimensionality, noise level, feature correlation and choice of ground-truth classifier

Metric	Direct-Opt	Direct-Approx	Breiman (B=1)	Breiman (B=10)
Ground-truth cor	0.991 ± 0.004	0.982 ± 0.008	0.838 ± 0.070	0.867 ± 0.071
Max score diff	0.017 ± 0.006	0.019 ± 0.005	0.079 ± 0.021	0.076 ± 0.026
Time (ms)	26.7 ± 14.4	20.8 ± 11.3	20.1 ± 6.2	126.6 ± 41.6

Table 14: Classification (nonlinear response). Global results across 48 scenarios with varying sample size, dimensionality, noise level, feature correlation and choice of ground-truth classifier

Metric	Direct-Opt	Direct-Approx	Breiman (B=1)	Breiman (B=10)
Ground-truth cor	0.995 ± 0.002	0.987 ± 0.009	0.798 ± 0.070	0.837 ± 0.064
Max score diff	0.018 ± 0.004	0.020 ± 0.005	0.165 ± 0.033	0.156 ± 0.029
Time (ms)	26.4 ± 14.3	20.8 ± 11.3	20.0 ± 6.2	125.6 ± 41.6

Table 15: Results of different VI methods under a sparse linear regression master estimator (TRUST) with default scoring metrics – Boston HMDA dataset

Metric	Direct-Opt	Direct-Approx	Breiman (B=1)	Breiman (B=10)
Ground-truth cor	0.952 ± 0.005	0.956 ± 0.004	0.943 ± 0.011	0.944 ± 0.011
Max score diff	0.225 ± 0.016	0.219 ± 0.015	0.472 ± 0.047	0.471 ± 0.045
Mean score diff	0.045 ± 0.003	0.044 ± 0.002	0.079 ± 0.008	0.079 ± 0.007
Time (ms)	1.60 ± 0.03	1.45 ± 0.03	5.06 ± 0.11	33.84 ± 0.91

Table 16: Results of different VI methods under an unregularized master linear model with default scoring metrics – Boston HMDA dataset

Metric	Direct-Opt	Direct-Approx	Breiman (B=1)	Breiman (B=10)
Ground-truth cor	0.997 ± 0.000	0.998 ± 0.000	0.999 ± 0.000	0.999 ± 0.000
Max score diff	0.185 ± 0.012	0.190 ± 0.015	0.074 ± 0.026	0.073 ± 0.025
Mean score diff	0.031 ± 0.002	0.032 ± 0.002	0.013 ± 0.004	0.013 ± 0.004
Time (ms)	0.49 ± 0.03	0.39 ± 0.01	5.80 ± 0.15	42.03 ± 0.99

Table 17: Results of different VI methods under a sparse logistic regression master estimator with default scoring metrics – German credit dataset

Metric	Direct-Opt	Direct-Approx	Breiman (B=1)	Breiman (B=10)
Ground-truth cor	1.000 ± 0.000	1.000 ± 0.000	0.996 ± 0.002	0.996 ± 0.002
Max score diff	0.022 ± 0.010	0.013 ± 0.008	0.077 ± 0.028	0.077 ± 0.028
Mean score diff	0.002 ± 0.001	0.001 ± 0.001	0.008 ± 0.003	0.008 ± 0.003
Time (ms)	0.82 ± 0.04	0.67 ± 0.01	5.07 ± 0.10	32.56 ± 0.40

Table 18: Results of different VI methods under an unregularized logistic regression master model – German credit dataset

Metric	Direct-Opt	Direct-Approx	Breiman (B=1)	Breiman (B=10)
Ground-truth cor	0.930 ± 0.014	0.935 ± 0.017	0.682 ± 0.082	0.658 ± 0.148
Max score diff	0.060 ± 0.009	0.045 ± 0.005	0.180 ± 0.073	0.191 ± 0.096
Mean score diff	0.012 ± 0.001	0.010 ± 0.001	0.042 ± 0.011	0.042 ± 0.013
Time (ms)	0.75 ± 0.01	0.66 ± 0.01	4.95 ± 0.08	32.48 ± 0.22

B Stability Across Repeated Independent Runs

Table 19: Top 5 features by VI method, across 10 independent runs. Ground truth: [0, 1, 11, 6, 3] – Boston HMDA dataset.

Top 5 indices	Direct-Opt	Direct-Approx	Breiman (B=1)	Breiman (B=10)
[0, 11, 3, 1, 6]	10	10	0	0
[0, 11, 2, 9, 3]	0	0	0	3
[0, 11, 3, 9, 2]	0	0	0	2
[0, 11, 9, 2, 1]	0	0	0	1
[0, 11, 9, 3, 2]	0	0	0	1
[0, 11, 1, 2, 9]	0	0	1	0
[0, 11, 2, 9, 1]	0	0	1	0
[0, 11, 1, 2, 7]	0	0	1	0
[0, 11, 2, 3, 9]	0	0	1	0
[0, 3, 11, 9, 1]	0	0	0	1
[0, 3, 11, 1, 9]	0	0	0	1
[0, 2, 3, 9, 11]	0	0	0	1
[0, 1, 11, 9, 3]	0	0	1	0
[0, 3, 1, 9, 7]	0	0	1	0
[0, 3, 11, 9, 2]	0	0	1	0
[0, 3, 11, 9, 7]	0	0	1	0
[0, 1, 11, 4, 7]	0	0	1	0
[0, 3, 2, 11, 9]	0	0	1	0

Table 20: Top 5 features across 10 independent runs. Ground truth: [0, 3, 2, 1, 5] – German credit dataset.

Top 5 indices	Direct-Opt	Direct-Approx	Breiman (B=1)	Breiman (B=10)
[0, 3, 2, 1, 5]	10	10	0	0
[0, 5, 7, 3, 2]	0	0	2	0
[0, 19, 5, 18, 3]	0	0	1	0
[0, 2, 3, 5, 13]	0	0	1	0
[0, 7, 5, 15, 18]	0	0	1	0
[0, 5, 3, 2, 19]	0	0	1	0
[0, 7, 9, 4, 18]	0	0	1	0
[0, 3, 7, 13, 18]	0	0	0	1
[0, 3, 7, 2, 18]	0	0	0	1
[0, 2, 7, 3, 13]	0	0	0	1
[0, 2, 7, 5, 3]	0	0	0	1
[0, 2, 13, 3, 7]	0	0	0	1
[0, 7, 5, 3, 13]	0	0	0	1
[0, 7, 5, 2, 13]	0	0	0	1
[0, 7, 2, 5, 3]	0	0	0	1
[0, 7, 3, 2, 5]	0	0	0	1
[0, 5, 7, 2, 13]	0	0	0	1
[3, 2, 0, 5, 12]	0	0	1	0
[3, 13, 19, 0, 18]	0	0	1	0
[3, 0, 19, 13, 7]	0	0	1	0

C Additional Figures and Plots

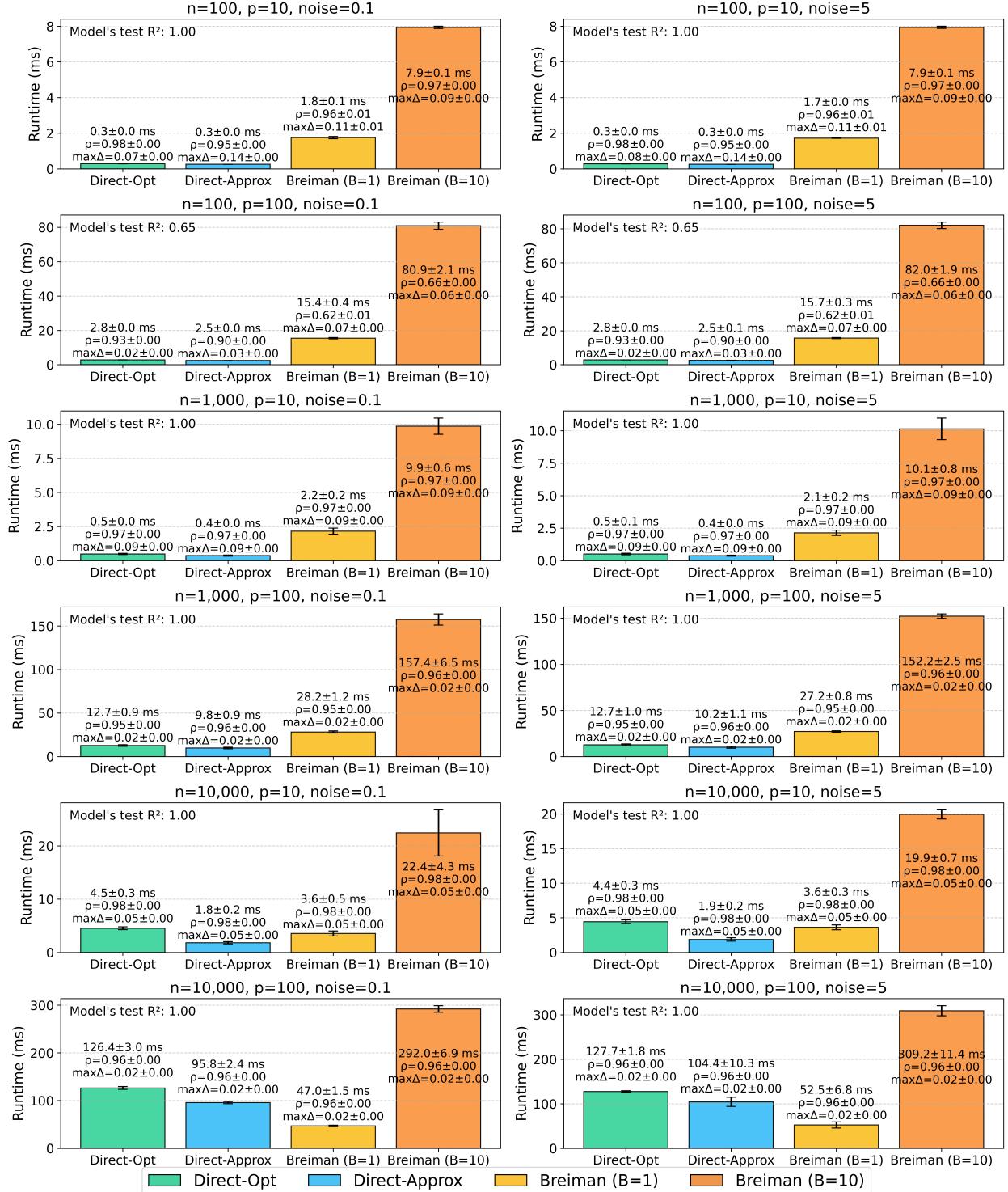


Figure 3: Performance of different variable importance methods in linear regression with uncorrelated features and linear response under varying sample size, dimensionality and noise

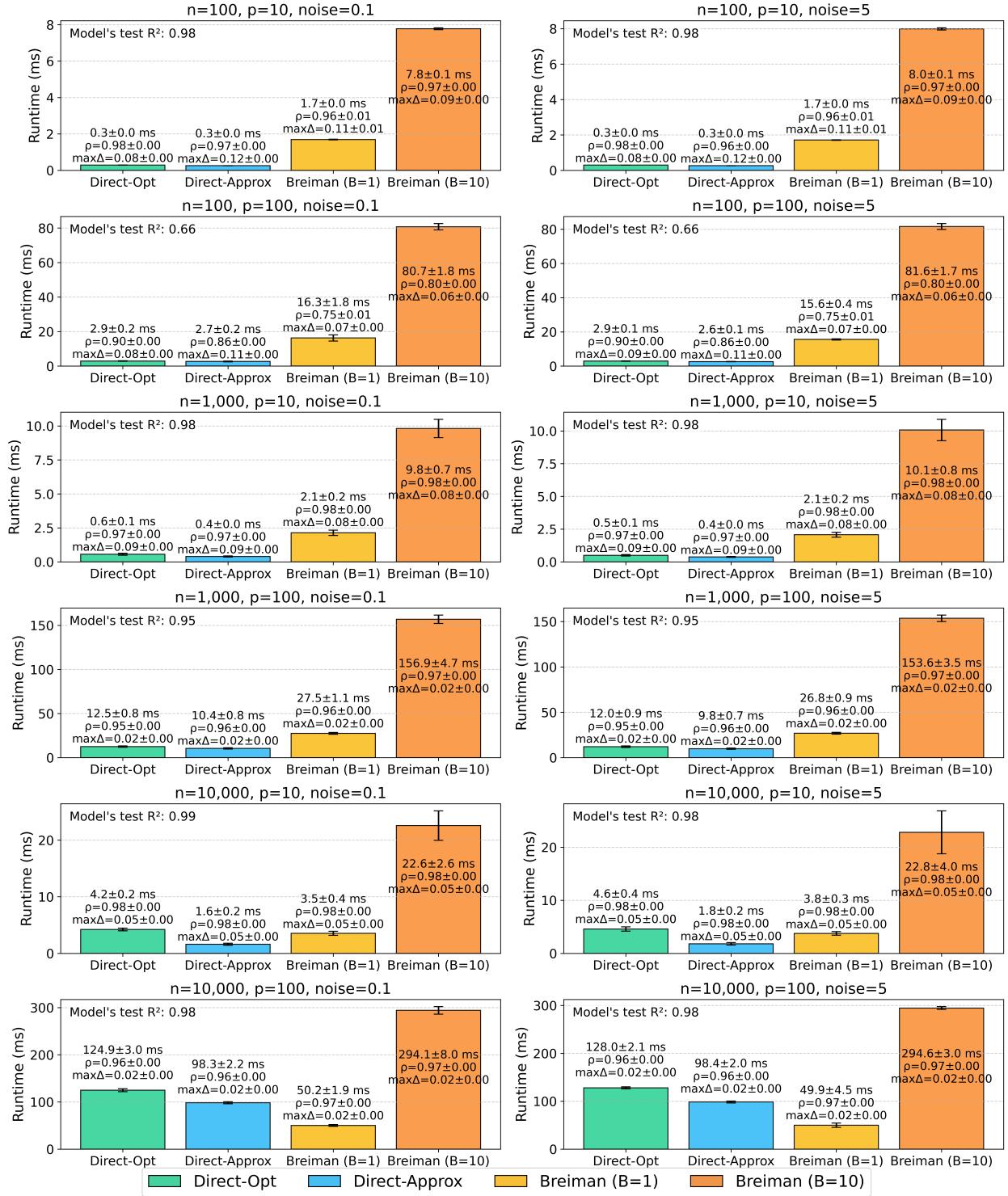


Figure 4: Performance of different variable importance methods in ℓ_1 -linear regression with uncorrelated features and linear response under varying sample size, dimensionality and noise

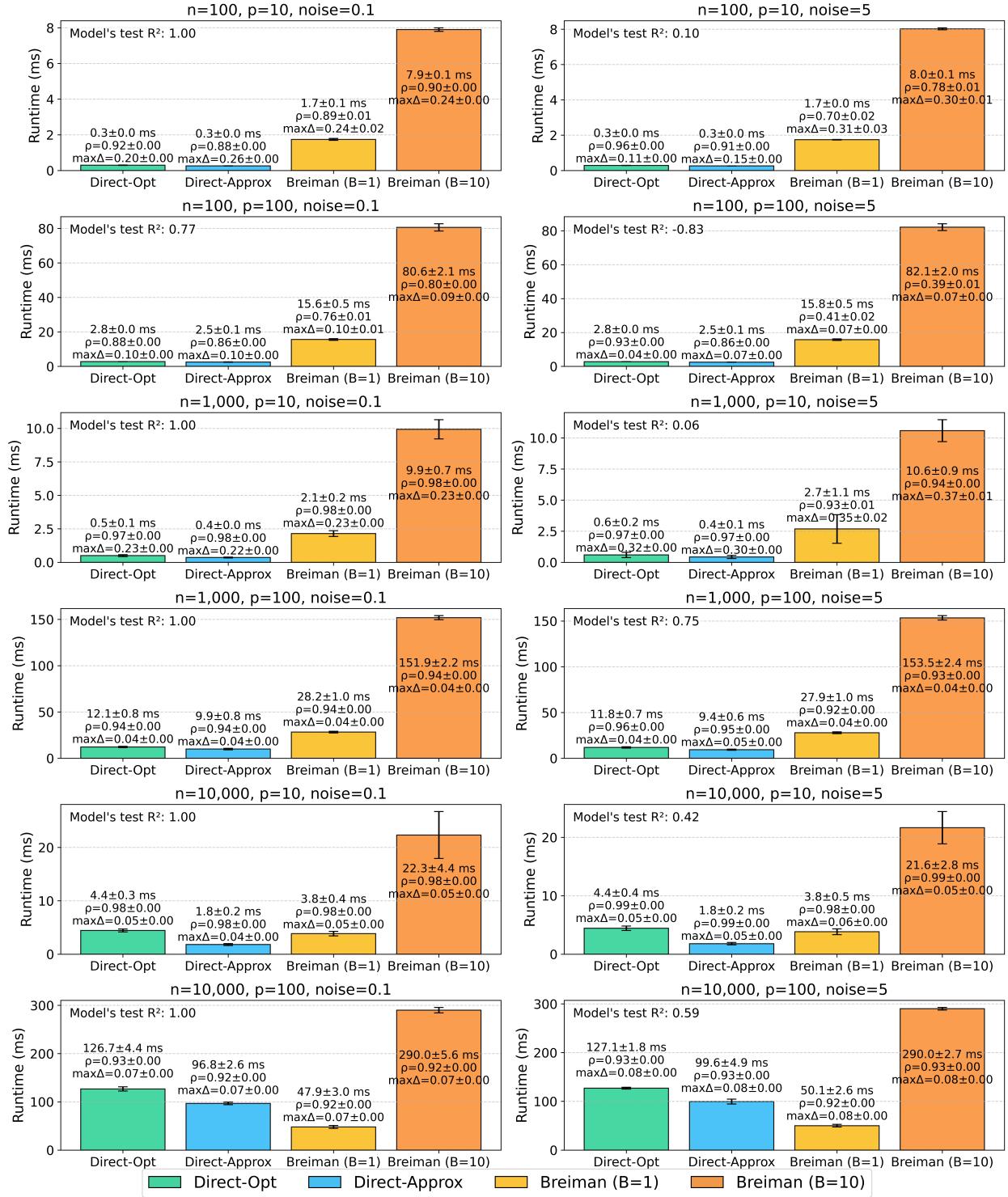


Figure 5: Performance of different variable importance methods in linear regression with correlated features and linear response under varying sample size, dimensionality and noise

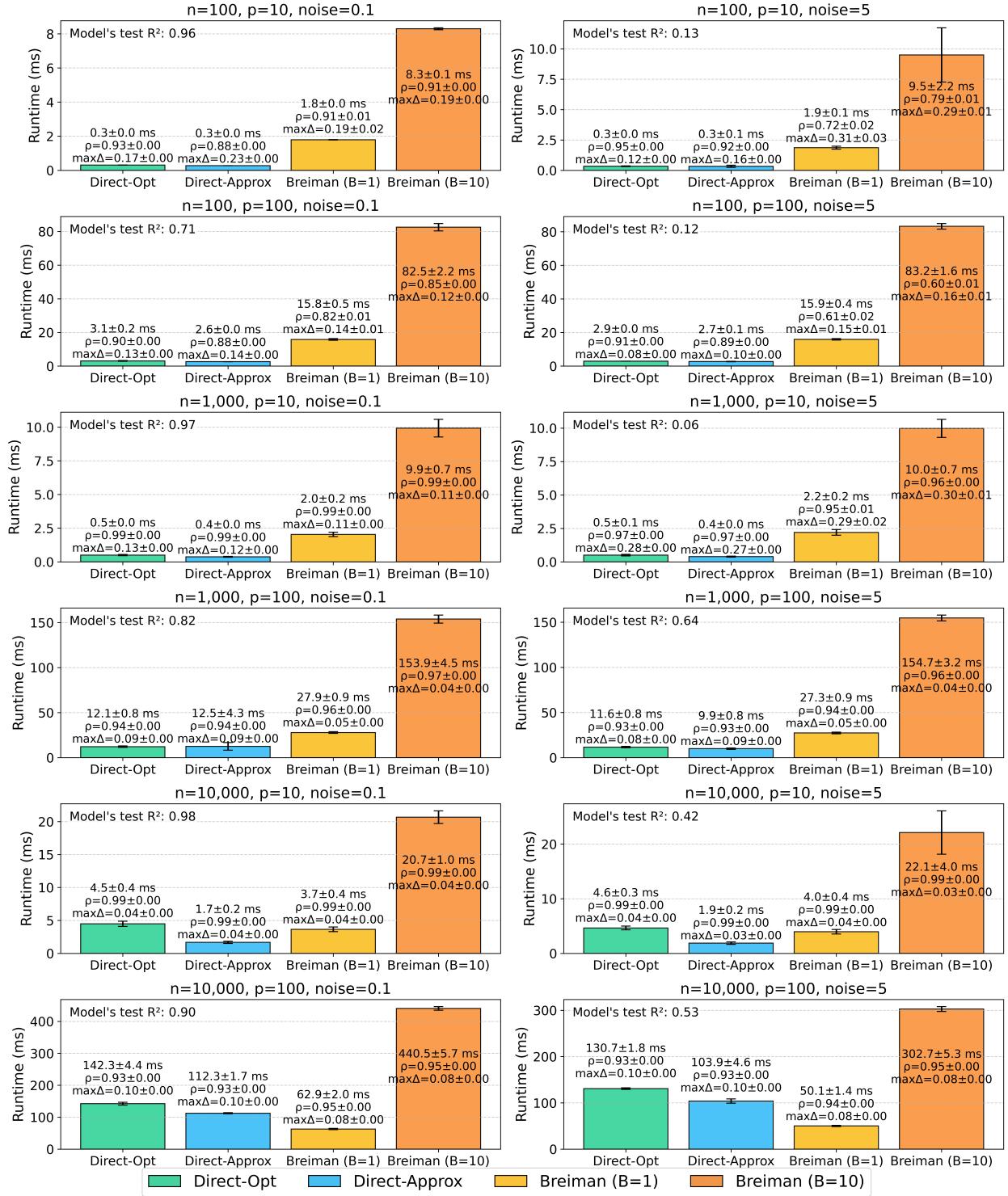


Figure 6: Performance of different variable importance methods in ℓ_1 -linear regression with correlated features and linear response under varying sample size, dimensionality and noise

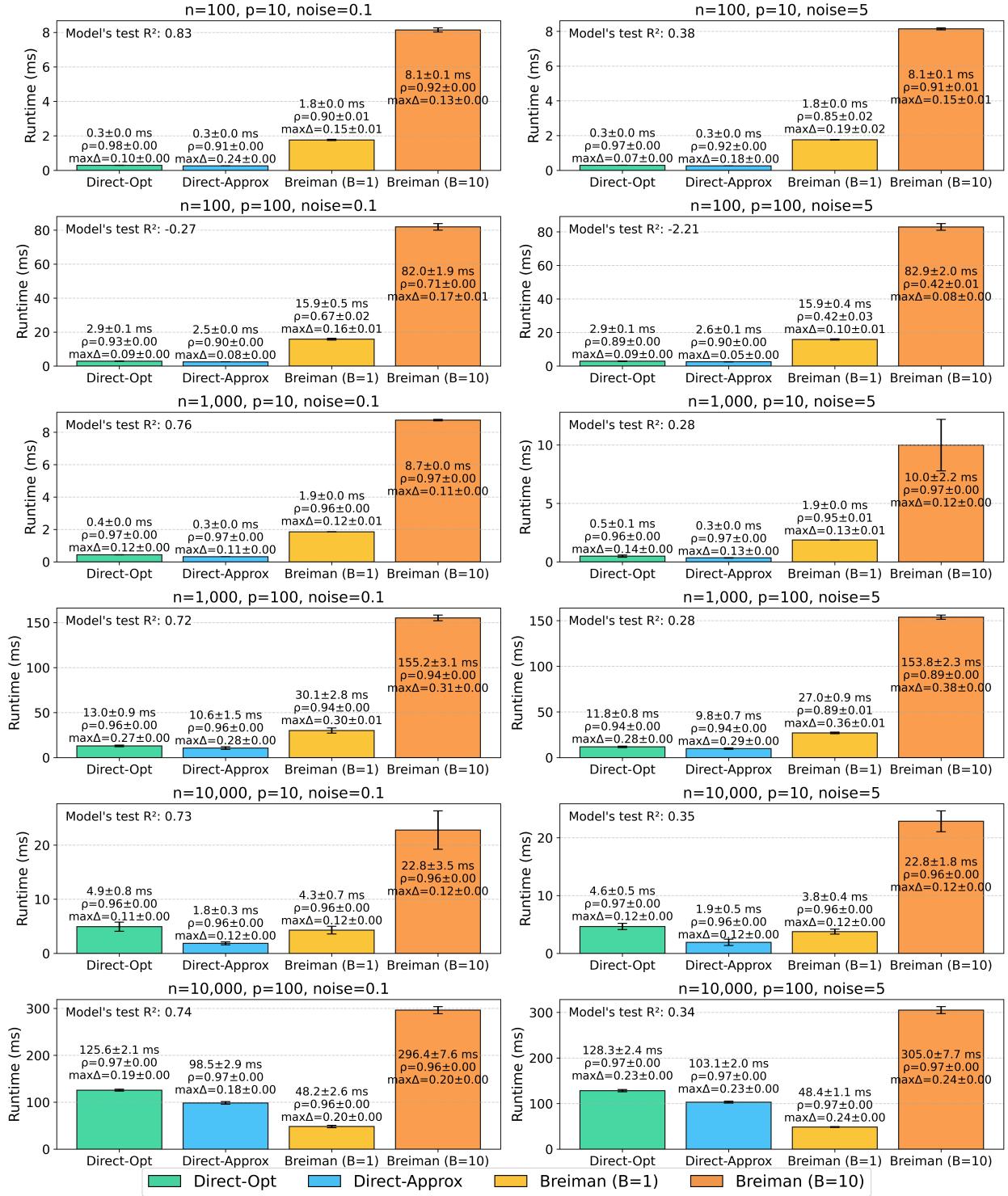


Figure 7: Performance of different variable importance methods in linear regression with uncorrelated features and nonlinear response under varying sample size, dimensionality and noise

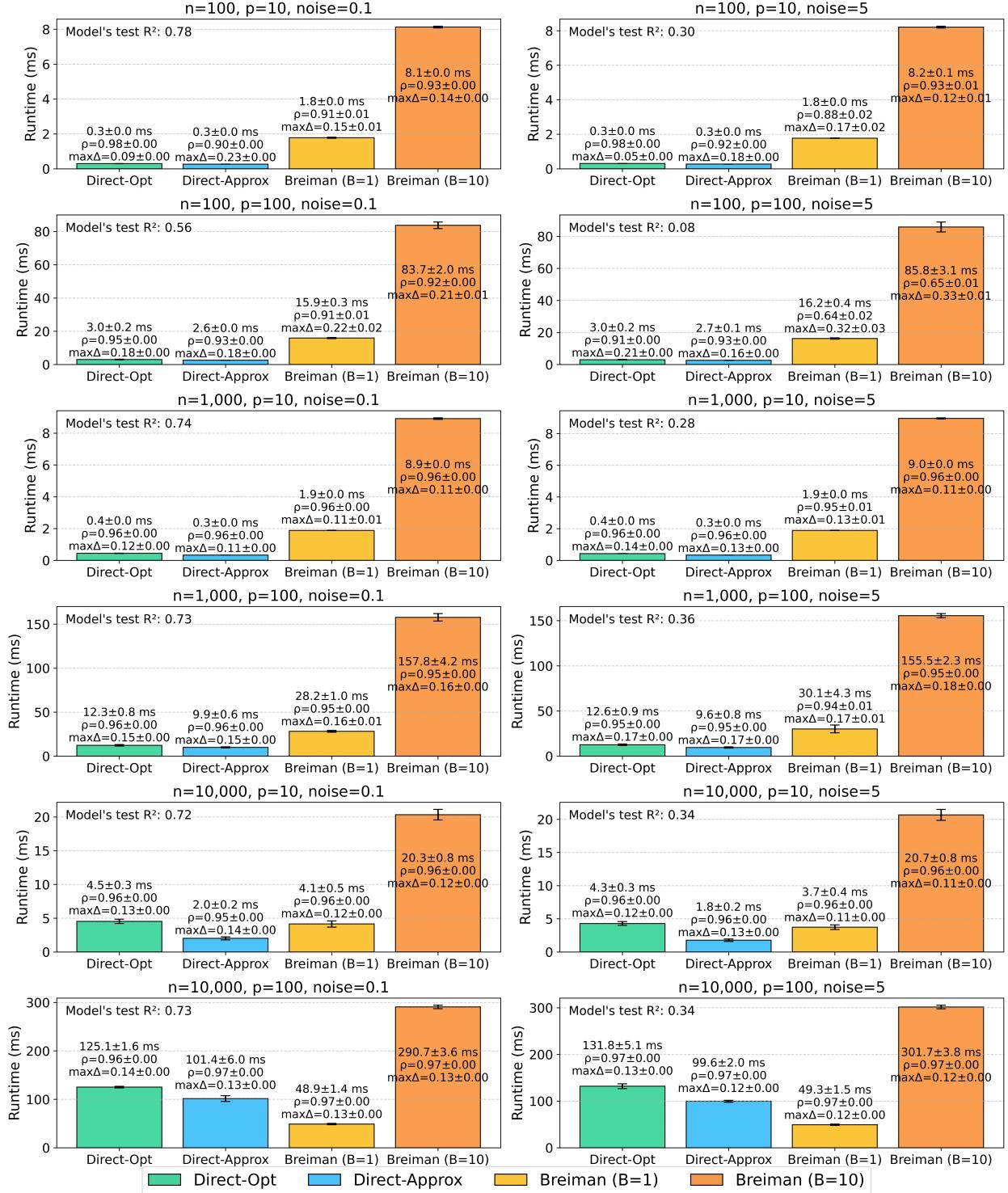


Figure 8: Performance of different variable importance methods in ℓ_1 -linear regression with uncorrelated features and nonlinear response under varying sample size, dimensionality and noise

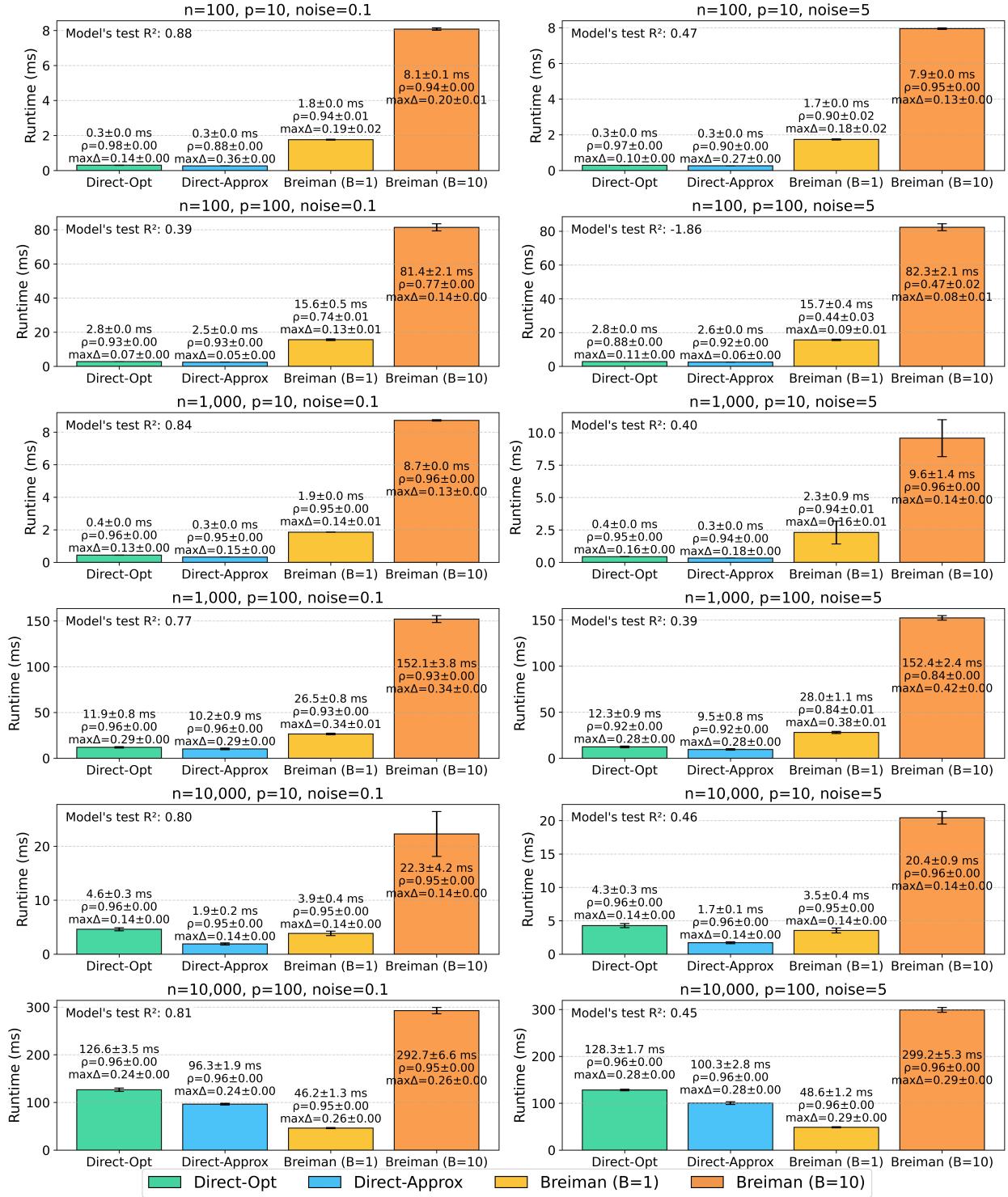


Figure 9: Performance of different variable importance methods in linear regression with correlated features and nonlinear response under varying sample size, dimensionality and noise

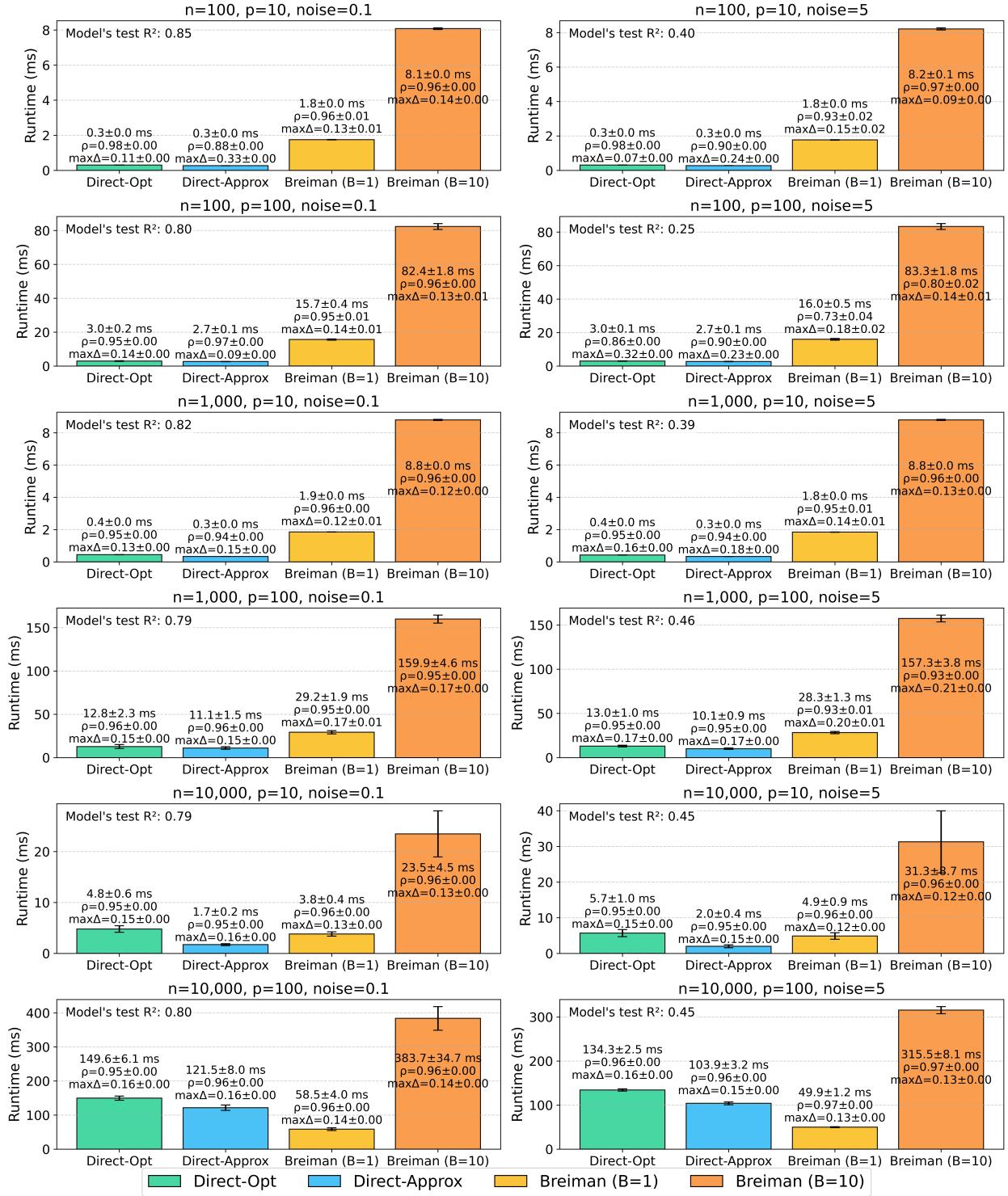


Figure 10: Performance of different variable importance methods in ℓ_1 -linear regression with correlated features and nonlinear response under varying sample size, dimensionality and noise

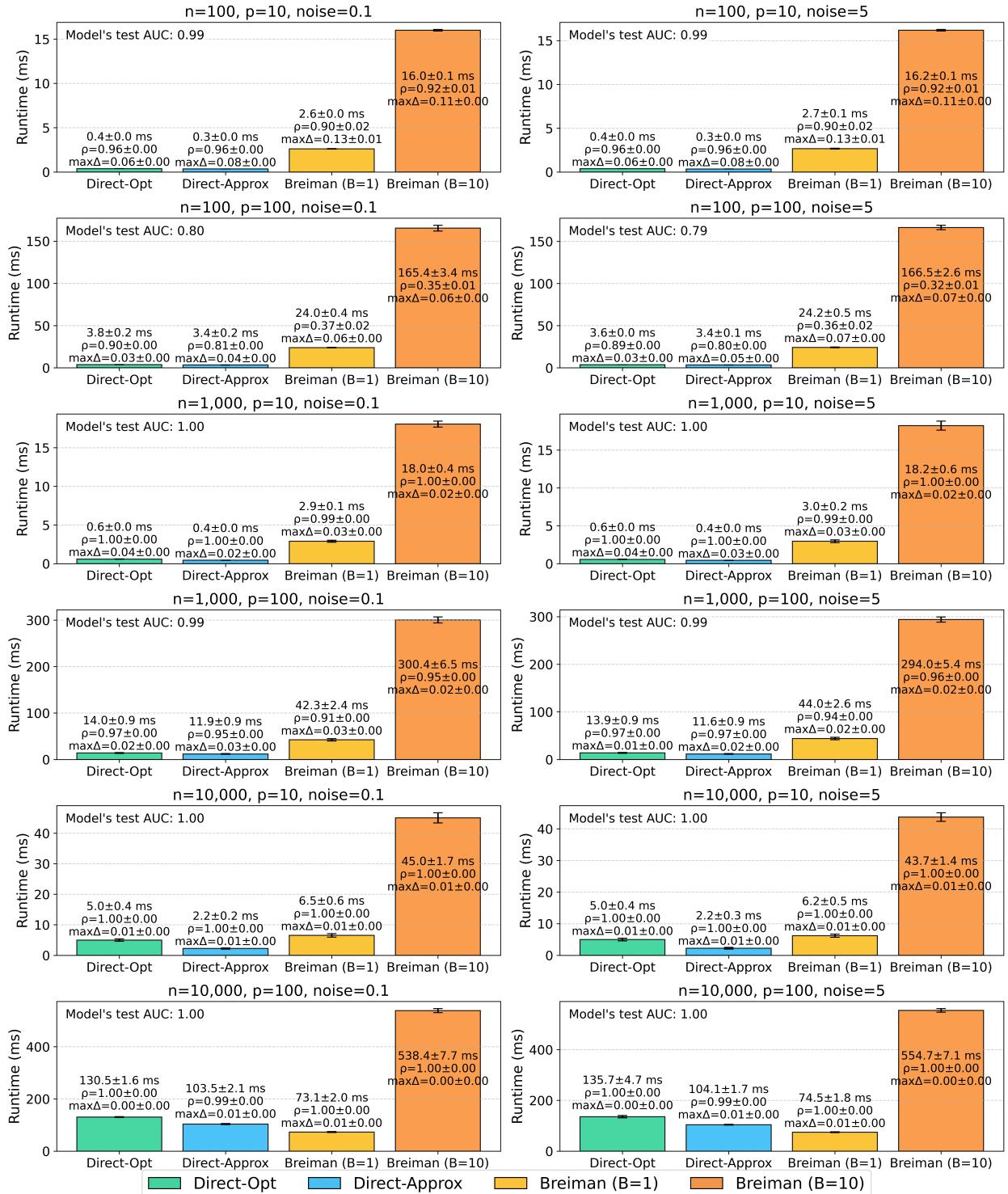


Figure 11: Performance of different variable importance methods in logistic regression with uncorrelated features and linear (binarized) response under varying sample size, dimensionality and noise

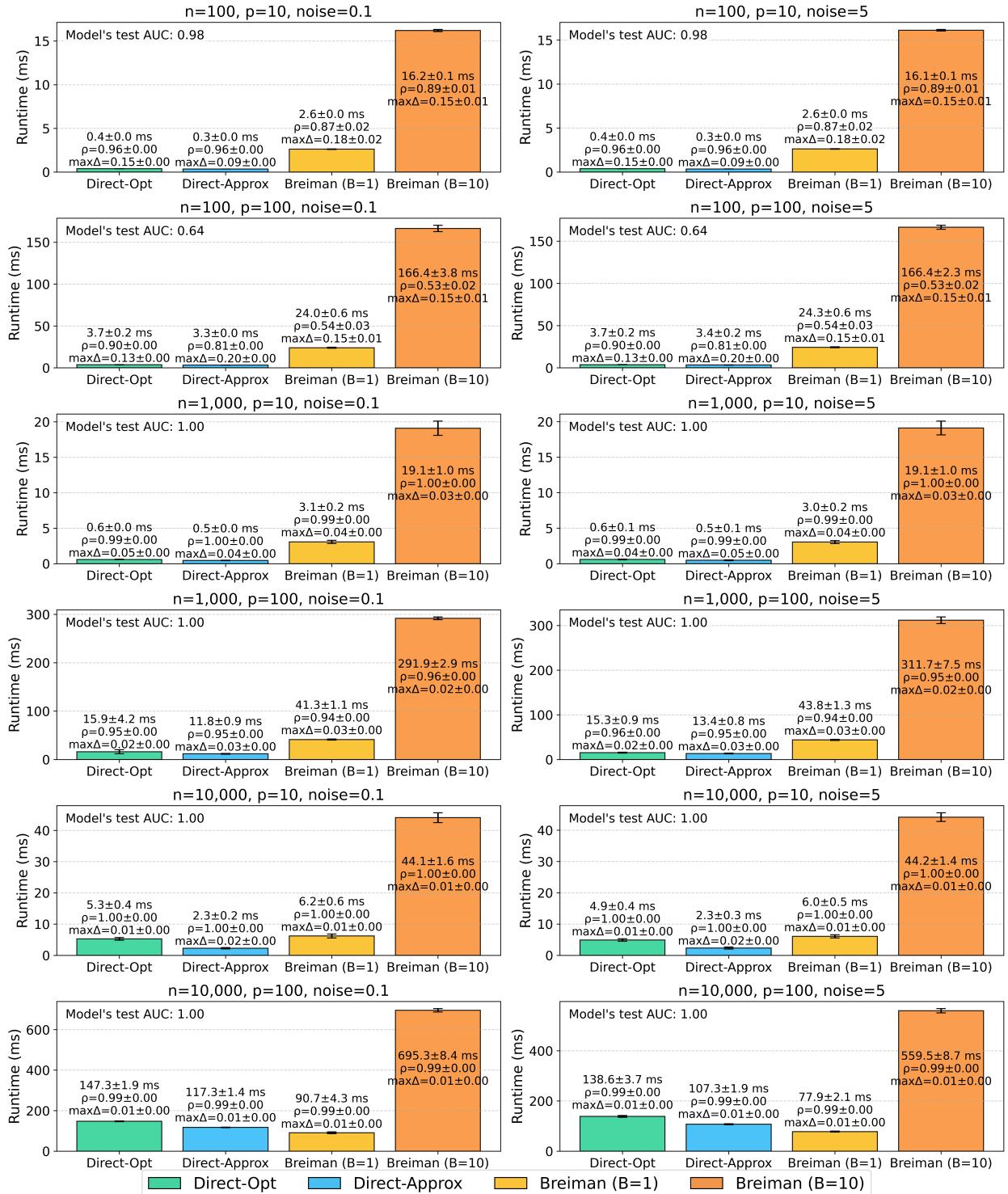


Figure 12: Performance of different variable importance methods in ℓ_1 -logistic regression with uncorrelated features and linear (binarized) response under varying sample size, dimensionality and noise

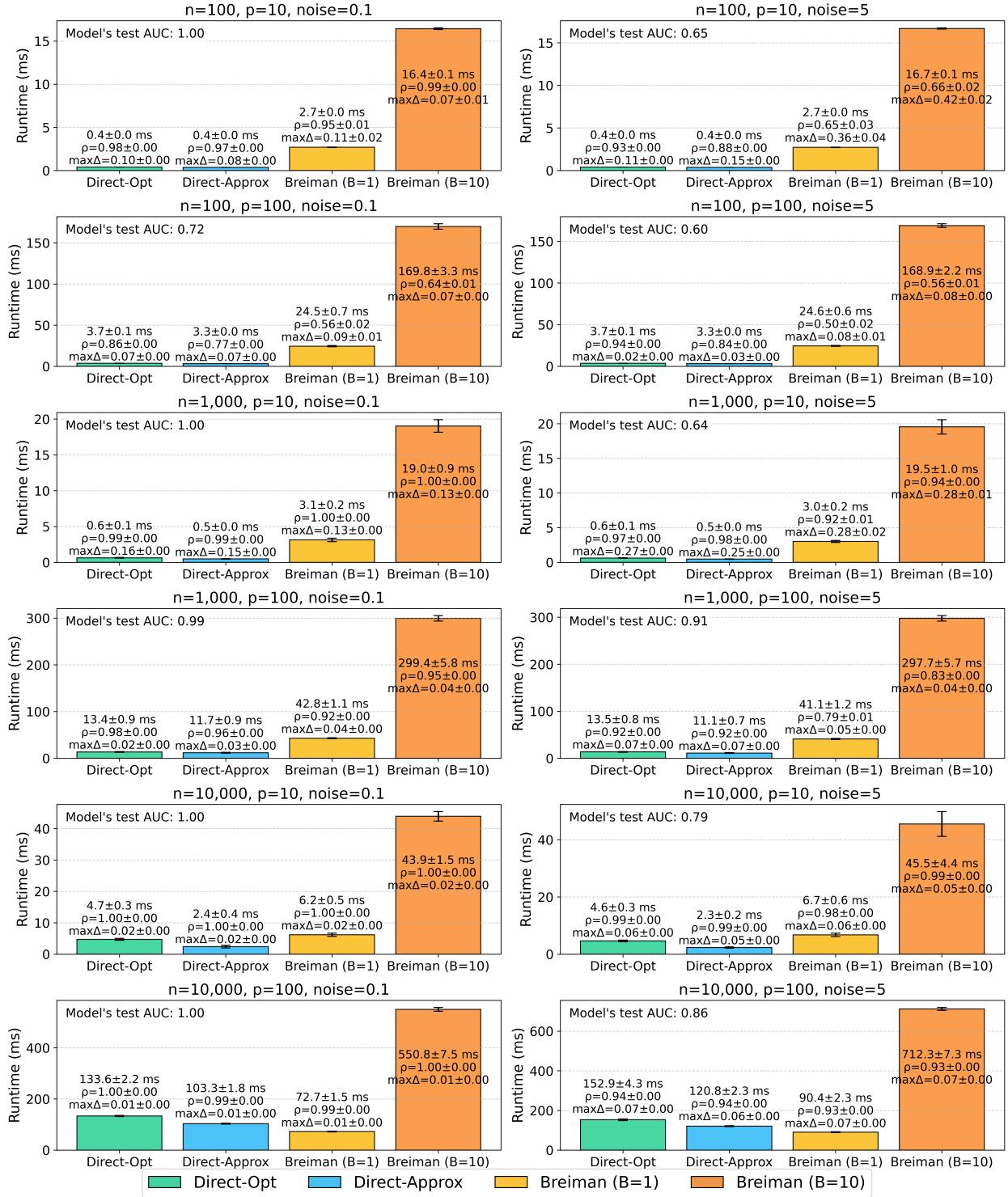


Figure 13: Performance of different variable importance methods in logistic regression with correlated features and linear (binarized) response under varying sample size, dimensionality and noise

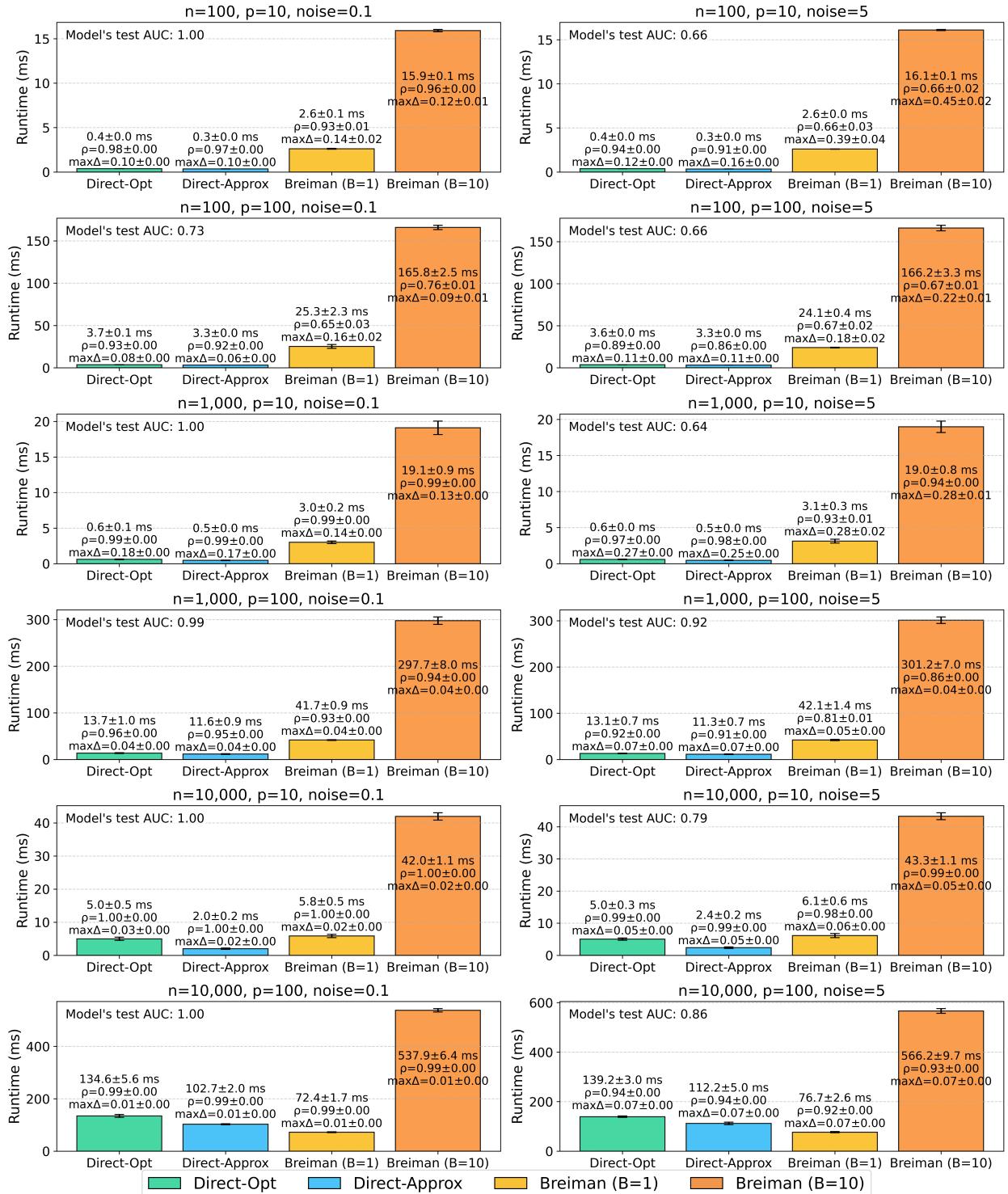


Figure 14: Performance of different variable importance methods in ℓ_1 -logistic regression with correlated features and linear (binarized) response under varying sample size, dimensionality and noise

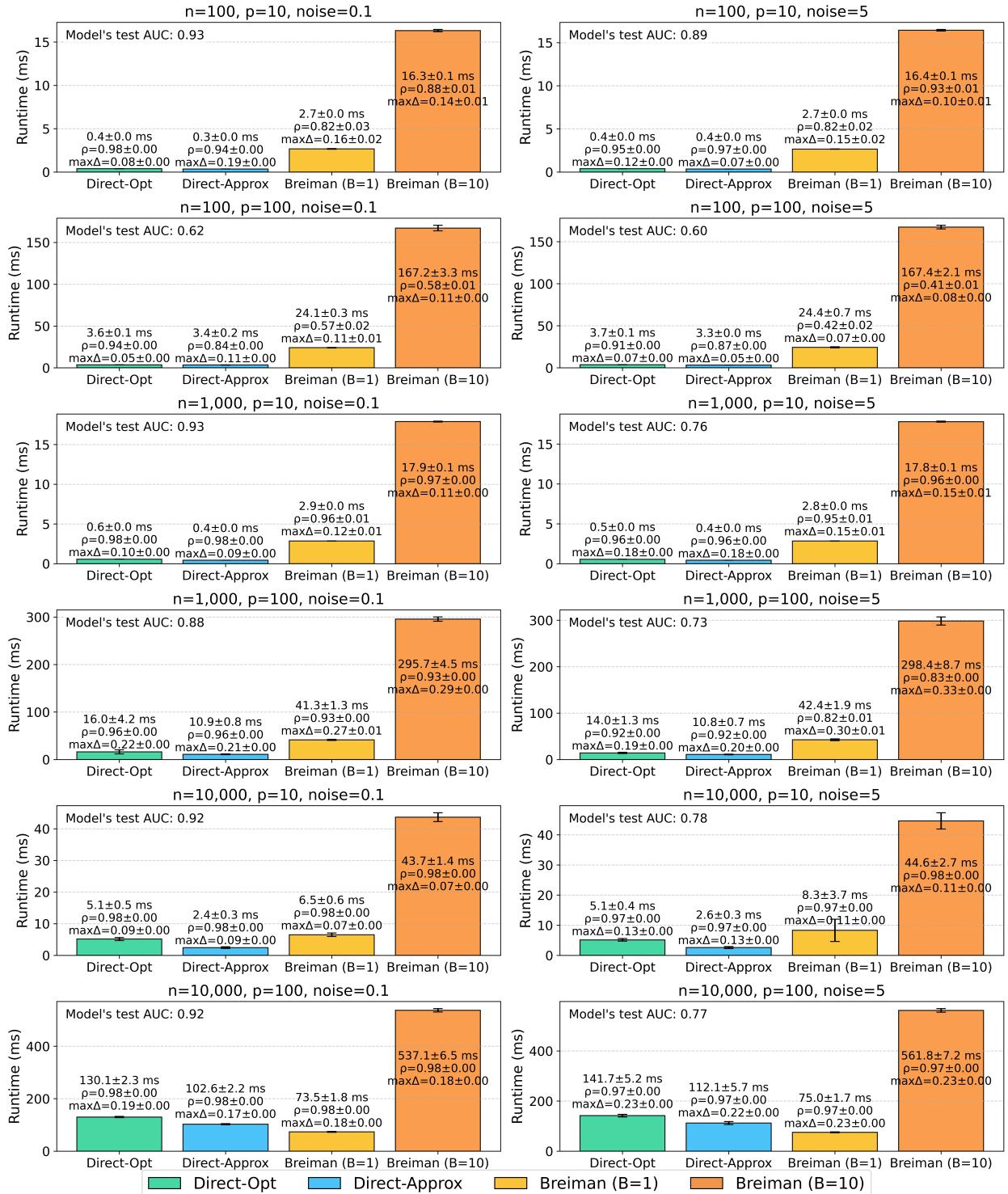


Figure 15: Performance of different variable importance methods in logistic regression with uncorrelated features and nonlinear (binarized) response under varying sample size, dimensionality and noise

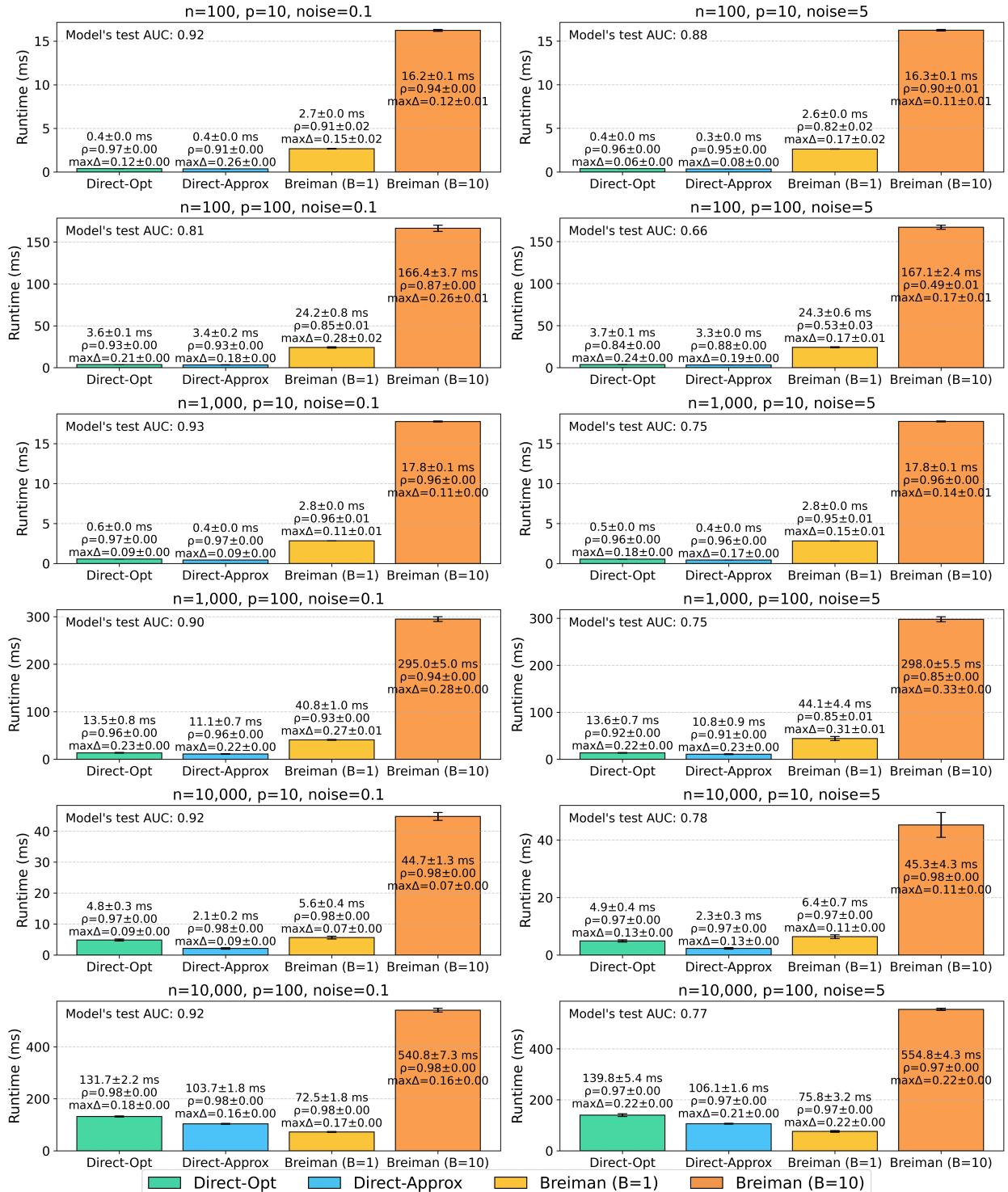


Figure 16: Performance of different variable importance methods in ℓ_1 -logistic regression with uncorrelated features and nonlinear (binarized) response under varying sample size, dimensionality and noise

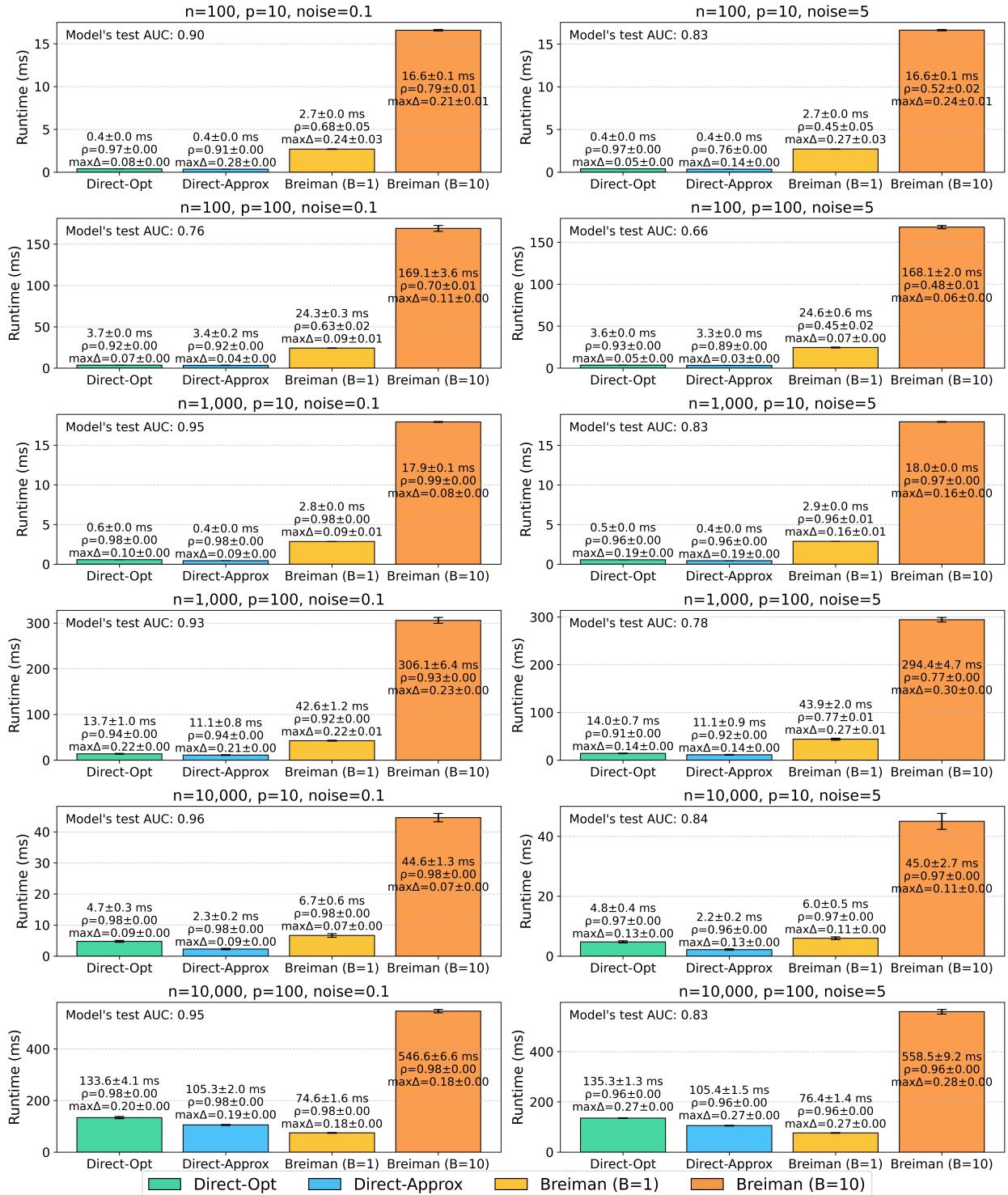


Figure 17: Performance of different variable importance methods in logistic regression with correlated features and nonlinear (binarized) response under varying sample size, dimensionality and noise

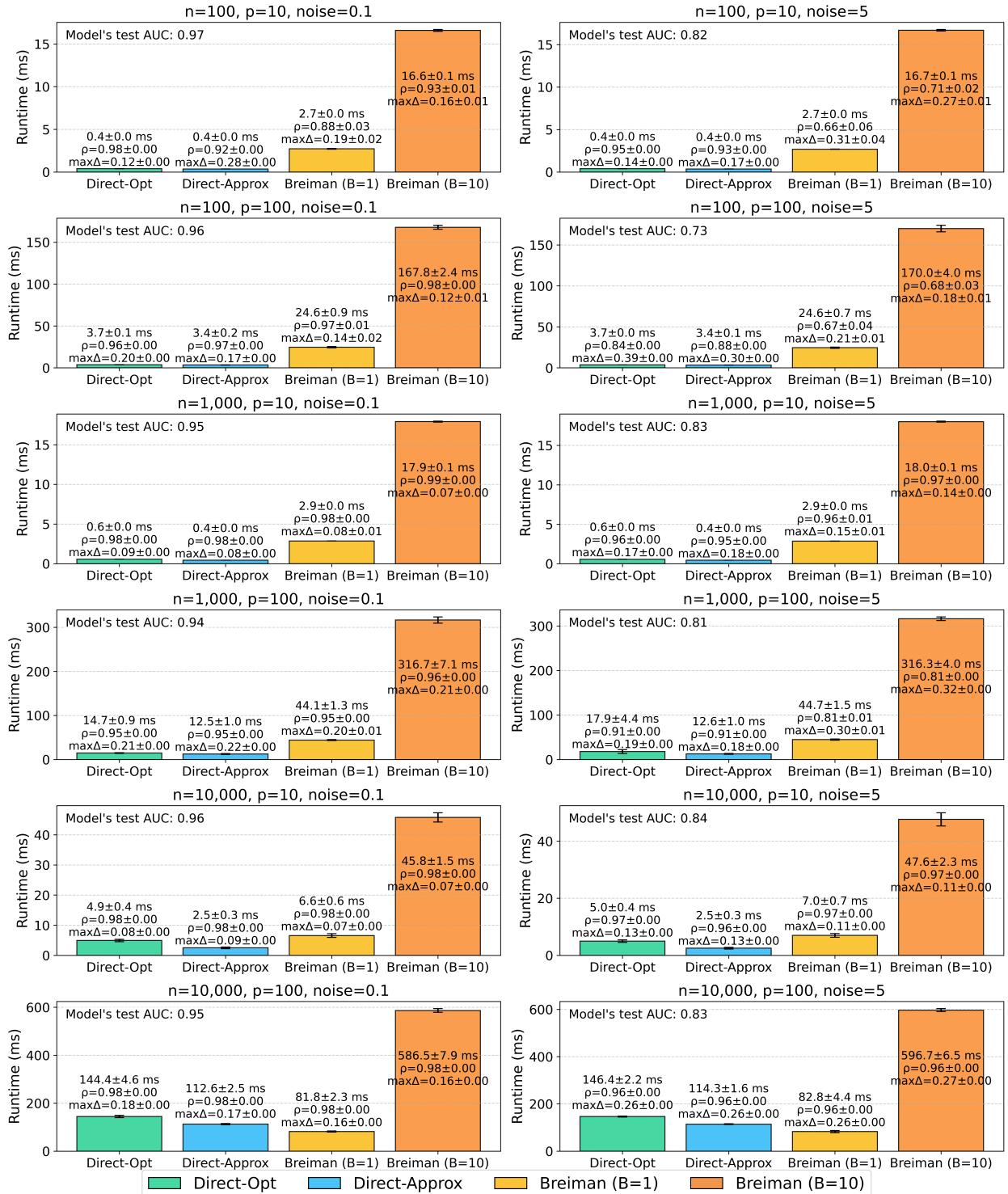


Figure 18: Performance of different variable importance methods in ℓ_1 -logistic regression with correlated features and nonlinear (binarized) response under varying sample size, dimensionality and noise

D Counterfactual Analysis With the trust-free Python Library

TRUST™ Profile Comparison Report

◆ Profile Similarity

We measure how close the feature patterns of the requested profiles are based on angular similarity.
Meaning: 0% = opposite profiles, 50% = uncorrelated, 100% = identical.

- Similarity: **----** (Moderate, 78%)
- Most similar in: hir (0.26 and 0.26)
- Most different in: black (1.00 vs. 0)

◆ Prediction Difference Analysis

In this section we compare head-to-head the two requested profiles and relate them to their predictions.
The table below shows how their differences in attributes affect their model predictions.

Profile comparison table:

	Profile 1	Profile 2	Effect
black	1.00	0	+0.00
11 other factors			0
Prediction	0.36	0.36	+0.00

Takeaway: The first profile's prediction (0.36) is slightly higher (+1%) than that of the second (0.36). This outcome is influenced by higher black (\uparrow), which increases the first prediction by 0.00.

Figure 19: Direct counterfactual effect on prediction of variable `black` for observation number 35