Causal Forests versus Penalized Splines for Heterogeneous Treatment Effects

Ivan Korolev* Abiodun Musbaudeen† Nency Dhameja‡

September 19, 2026

Abstract

This paper compares causal forests with varying-coefficient penalized spline estimators for heterogeneous treatment effects. We conduct simulations with randomized treatment assignment that vary heterogeneity, predictor dimension, and discrete nuisance structure. Performance is measured by out-of-sample mean squared error. Forests perform best in Wager–Athey designs with sharp treatment-effect transitions. Penalized splines dominate when heterogeneity is smooth, when the baseline mean and treatment effect differ in structure, and in mixed designs with continuous, binary, and group-level covariates. As group cardinality rises, forest performance deteriorates, while `mgcv` random-effect smooths handle group structure effectively and remain easy to implement.

Keywords: causal forests, penalized splines, heterogeneous treatment effects, varying coefficients, semiparametric estimation, simulations

JEL Codes: C14, C15, C21, C52

*Department of Economics, Binghamton University, ikorolev@binghamton.edu
†Department of Economics, Binghamton University, amusbau1@binghamton.edu
‡Department of Economics, Binghamton University, ndhamej1@binghamton.edu

1 Introduction

Empirical researchers increasingly use causal forests and related machine learning methods to estimate heterogeneous treatment effects (Wager and Athey, 2018; Athey et al., 2019). At the same time, econometrics has long offered structured alternatives, including linear interaction models, varying-coefficient estimators, and penalized spline procedures (Hastie and Tibshirani, 1993; Fan and Zhang, 1999; Li and Racine, 2007; Chen, 2007; Wood, 2011, 2017). These methods are often presented as belonging to different methodological worlds, yet they are trying to recover closely related objects. In a partially linear varying-coefficient setup,

$$Y = m(X) + D\tau(X) + \varepsilon, \qquad E[\varepsilon \mid X, D] = 0,$$

the common target is the treatment effect function $\tau(X)$. Here, Y is the outcome of interest, D is the treatment dummy, X denotes pre-treatment covariates, and ε is the error term. We assume randomized treatment assignment, so conventional nonparametric and semiparametric techniques can be used directly.

This paper asks a practical question: when do causal forests improve on penalized spline estimators of treatment effect heterogeneity, and when do they not? We compare them in simulation designs that vary the shape of heterogeneity, the dimensionality of predictors, and the presence of discrete nuisance structure. Causal forests perform best when treatment effects exhibit sharp localized transitions, as in the Wager–Athey logit-style design. Spline-based varying-coefficient methods tend to dominate when heterogeneity is smooth, when the baseline mean and treatment effect have different structures, and in settings with binary and categorical variables, which are common in applied work.

A second question concerns implementability. Traditional nonparametric estimators are often viewed as impractical for applied work because they require choices of bandwidths, basis functions, sieve dimensions, and interaction structure. Recent developments in statistical software substantially lower this barrier. In our comparison, `mgcv` allows the researcher

to write varying-coefficient GAM specifications directly, with smoothing parameters and effective degrees of freedom selected by penalized likelihood criteria such as REML or fREML (Wood, 2011, 2017, 2025). Causal forests are implemented using `grf` (Tibshirani et al., 2026). Thus, the R environment (R Core Team, 2026) provides mature, well-documented implementations of both estimator classes, making their use feasible for applied researchers.

2 Estimands

We use the potential outcomes framework:

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0),$$

where $Y_i(d)$, $d = 0, 1$, denotes the potential outcome in state d (Rubin, 1974). The observed outcome can be written as:

$$Y_i = m(X_i) + D_i \tau(X_i) + \varepsilon_i,$$

where $m(X_i) = E[Y_i(0) \mid X_i]$ is the baseline conditional mean of the untreated potential outcome, $\tau(X_i) = E[Y_i(1) - Y_i(0) \mid X_i]$ is the conditional average treatment effect (CATE), and ε_i collects deviations from the conditional mean representation. We compare the ability of different methods to recover the CATE function $\tau(X)$.

3 Simulation Design

We consider four data-generating processes (DGPs) and two predictor regimes. Treatment is randomly assigned throughout, so the heterogeneity comparison is not confounded by propensity score modeling. We begin with the design studied by Wager and Athey (2018); Athey et al. (2019): a low-dimensional setting with $p = 2$ continuous predictors, each distributed uniformly on $[0, 1]$, and treatment heterogeneity generated from logistic components with a relatively sharp transition. We then consider a smooth nonlinear design in which

heterogeneity is generated by sine, cosine, and polynomial terms. In these first two designs, we follow the Wager–Athey convention and impose $m(X) = -0.5\tau(X)$.

This restriction need not hold in practice. The baseline mean $m(X)$ and the treatment effect function $\tau(X)$ may have different shapes. We therefore add two designs that vary one component at a time. In one design, $m(X)$ is bilinear in X_1 and X_2 while $\tau(X)$ remains nonlinear. In the other, $\tau(X)$ is bilinear while $m(X)$ remains at the smooth nonlinear baseline.

The restriction $m(X) = -0.5\tau(X)$ is a feature of the benchmark design, not a maintained assumption of either method. In the two-step `grf` implementation, the marginal outcome and propensity score are first estimated as nuisance functions; a causal forest then estimates $\tau(X)$ using the centered outcome and treatment (Athey et al., 2019). This separation allows the baseline mean and treatment effect to differ in complexity. Dandl et al. (2024) compare this procedure with model-based forests that estimate outcome and treatment-effect components jointly, and find that joint approaches with appropriate centering can perform on par with the two-step approach. Penalized splines likewise allow separate mean-side and treatment-side specifications and penalties. Thus, both method families can in principle accommodate different structures in $m(X)$ and $\tau(X)$; our asymmetric designs ask which adapts more effectively in finite samples.

The first predictor regime is a low-dimensional $p = 2$ design with continuous regressors only. The second is a mixed $p = 6$ design in which X_1, X_2, X_3 are continuous, X_4 is a group variable with either 4 or 20 levels, and X_5, X_6 are binary covariates. We use sample sizes $n = 800$ and $n = 1600$ to match Athey et al. (2019). The idiosyncratic disturbance is drawn i.i.d. as $\varepsilon_i \sim N(0, 1)$. The DGPs are presented in Table 1.

Table 1: Headline simulation designs

Design	$p = 2$	$p = 6$ mixed
Wager–Athey	$\tau(X) = \zeta(X_1)\zeta(X_2)$ $m(X) = -0.5\,\tau(X)$	$\tau(X) = \zeta(X_1)\zeta(X_2)$ $m(X) = -0.5\,\tau(X) + X_3 + \text{FE}(X_4)$ $+ \text{FE}(X_5) + \text{FE}(X_6)$
Smooth interacted	$\tau(X) = 0.5 + \sin(2\pi X_1)$ $+ 2(X_2 - 0.5)^2$ $+ 0.8\cos(2\pi X_1) \times (X_2 - 0.5)$ $m(X) = -0.5\,\tau(X)$	$\tau(X) = 0.5 + \sin(2\pi X_1)$ $+ 2(X_2 - 0.5)^2$ $+ 0.8\cos(2\pi X_1) \times (X_2 - 0.5)$ $m(X) = -0.5\,\tau(X) + X_3 + \text{FE}(X_4)$ $+ \text{FE}(X_5) + \text{FE}(X_6)$
Bilinear m, nonlinear τ	$\tau(X) = 0.5 + \sin(2\pi X_1)$ $+ 2(X_2 - 0.5)^2$ $+ 0.8\cos(2\pi X_1) \times (X_2 - 0.5)$ $m(X) = 3X_1 + 2X_2 + 2X_1X_2$	$\tau(X) = 0.5 + \sin(2\pi X_1)$ $+ 2(X_2 - 0.5)^2$ $+ 0.8\cos(2\pi X_1) \times (X_2 - 0.5)$ $m(X) = 3X_1 + 2X_2 + X_3 + 2X_1X_2$ $+ \text{FE}(X_4) + \text{FE}(X_5) + \text{FE}(X_6)$
Bilinear τ, nonlinear m	$\tau(X) = 2X_1 + X_2 + 1.5X_1X_2$ $m(X) = -0.5[0.5 + \sin(2\pi X_1)$ $+ 2(X_2 - 0.5)^2$ $+ 0.8\cos(2\pi X_1) \times (X_2 - 0.5)]$	$\tau(X) = 2X_1 + X_2 + 1.5X_1X_2$ $m(X) = -0.5[0.5 + \sin(2\pi X_1)$ $+ 2(X_2 - 0.5)^2$ $+ 0.8\cos(2\pi X_1) \times (X_2 - 0.5)]$ $+ X_3 + \text{FE}(X_4)$ $+ \text{FE}(X_5) + \text{FE}(X_6)$

Notes: $p = 2$: $X_1, X_2 \overset{i.i.d.}{\sim} U[0,1]$. Mixed $p = 6$: $X_1, X_2, X_3 \overset{i.i.d.}{\sim} U[0,1]$, X_4 has 4 or 20 groups, $X_5 \sim$ Bernoulli(0.5), and $X_6 \sim$ Bernoulli(0.3). Discrete covariates enter through fixed effects: X_4 effects are centered draws from $N(0, 2^2)$; X_5 and X_6 effects are centered draws from $N(0,1)$. The 20-group mixture variant uses centered, scale-matched three-component normal-mixture draws for X_4. Fixed-effect draws are common to training and test samples within replication. $D \sim$ Bernoulli(0.5), $\varepsilon_i \sim N(0,1)$, and test samples have 1000 observations. Wager–Athey: $\zeta(u) = 1 + (1 + \exp(-20(u - 1/3)))^{-1}$.

We focus on randomized treatment assignment so that the simulations isolate how well each method recovers treatment-effect heterogeneity, rather than differences in how the methods adjust for confounding. In observational settings, a causal interpretation would additionally require unconfounded treatment assignment given observed pre-treatment covariates and sufficient overlap in treatment probabilities. A GAM-based approach would also require a separate adjustment step, such as propensity-score adjustment, orthogonalization, or the construction of doubly robust scores, before estimating treatment-effect heterogeneity.

This focus on moderate-dimensional settings is consistent with recent applied practice: Rehill (2025) reviews 133 causal forest applications and finds that researchers often use causal forests with relatively low-dimensional covariate sets and the standard `grf` implementation.

For each design and each method, we draw 1,000 fresh test covariates in every replication

and compute the replication-specific CATE mean squared error (MSE) as

$$MSE_{met}^{(r)} = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \left(\hat{\tau}_{met}^{(r)}(X_{i,test}^{(r)}) - \tau^{(r)}(X_{i,test}^{(r)}) \right)^2,$$

where met indexes the method, r indexes the Monte Carlo replication, and $n_{test} = 1{,}000$. The main table entries average this quantity across $R = 250$ simulation replications:

$$\overline{MSE}_{met} = \frac{1}{R} \sum_{r=1}^{R} MSE_{met}^{(r)}.$$

For readability, all MSE entries in the tables are reported as $100 \times \overline{MSE}_{met}$.

4 Methods Compared

We compare two families of methods: varying-coefficient penalized spline estimators and causal forests. The spline family is motivated by the varying-coefficient tradition of Hastie and Tibshirani (1993) and the broader nonparametric econometrics literature (Li and Racine, 2007). Its theoretical foundations include the sieve and penalized sieve framework of Chen (2007), Chen and Pouzo (2012), and Chen and Pouzo (2015), with practical implementations developed in the generalized additive models (GAM) literature (Wood, 2011, 2017).

The spline specifications are dimension-specific. For $p = 2$, *Pairwise GAM* uses smooth main effects and a tensor-product smooth in (X_1, X_2); in the asymmetric designs, *Oracle GAM* uses the true mean-side and treatment-side structure. For $p = 6$, *Pairwise GAM* uses smooth and pairwise tensor terms among the continuous covariates and additive discrete terms; *Full GAM* additionally interacts treatment with the discrete covariates; and *Oracle GAM* uses the true mean-side and treatment-side structure. In the 20-group designs, X_4 enters through a random-effect smooth that treats group coefficients as penalized deviations from zero; Appendix A reports unrestricted fixed-effect versions.

Causal forests instead grow many honest treatment-effect trees on subsamples of the

data (Wager and Athey, 2018; Athey et al., 2019). The fitted forest acts as an adaptive kernel: observations falling in the same leaves as the target point receive higher weight in a locally weighted, orthogonalized estimating equation. Regularization comes from honesty, subsampling, averaging across trees, and tuning of forest parameters. This is attractive when heterogeneity involves thresholds or irregular interactions that are difficult to specify ex ante.

To assess the gains from modeling heterogeneity, we also report a constant-effect benchmark that assigns every test observation the treated-minus-control mean difference from the training sample. Performance relative to this benchmark quantifies how much each flexible method improves CATE estimation over imposing a common effect.

We fit splines with `mgcv::bam()` (Wood, 2025) (`select = TRUE`, `method = "fREML"`) and forests with `grf::causal_forest()` (Tibshirani et al., 2026) using 2,000 trees and cross-validated tuning. All computations are implemented in R (R Core Team, 2026). Exact implementation details, including model formulas and covariate encoding, are reported in Appendix B.

5 Results

Table 2 reports results for the $p = 2$ continuous-predictor designs. Causal forests perform best in the Wager–Athey logit-style setup, while the pairwise GAM dominates for smooth interacted heterogeneity, especially at $n = 1{,}600$. When $m(\cdot)$ and $\tau(\cdot)$ have different structures, the oracle spline estimator is best, as expected. More importantly, the feasible pairwise GAM adapts well: it nearly matches the oracle when $m(\cdot)$ is bilinear and $\tau(\cdot)$ is nonlinear, and remains well ahead of causal forests in the reverse case.

Tables 3 and 4 report mixed $p = 6$ results with 4 groups and with 20 mixture-distributed group effects. Forests again perform best for the Wager–Athey DGP with 4 groups, but spline-based methods dominate in the other scenarios. As group count rises, forest performance deteriorates. Because forests can split on group indicators, more groups expand the

effective predictor space. Spline estimators are less sensitive, partly due to the random-effect smooth `s(X4, bs = "re")` in `mgcv::bam()`. The oracle GAM is typically best but infeasible. Among feasible estimators, the practical pairwise GAM with flexible continuous covariates and additive regularized group effects is generally best. *Full GAM*, which also interacts treatment with discrete variables, generally trails *Pairwise GAM*, although it often narrows the gap at larger sample sizes and still usually outperforms forests outside the Wager–Athey DGP.

The constant-effect benchmark in Tables 2–4 makes the magnitude of these gains clear. Both forests and GAMs substantially outperform the benchmark in every design and sample size. The best feasible method reduces MSE by a factor of at least six in every case and by more than ten in most cases, with the reduction exceeding fiftyfold in the most favorable setting. These gains indicate that the flexible estimators recover substantial variation in $\tau(X)$ rather than merely estimating the average treatment effect accurately.

Table 2: Mean test-set MSEs for $p = 2$

Design	n	Constant	Forest	Pairwise GAM	Oracle GAM
Wager–Athey	800	99.34	**6.41**	8.48	–
	1600	99.07	**3.87**	5.27	–
Smooth interacted	800	55.43	8.85	**8.59**	–
	1600	55.16	6.08	**4.36**	–
Bilinear m, nonlinear τ	800	56.22	9.31	**4.84**	***4.04***
	1600	55.63	6.70	**2.43**	***2.04***
Bilinear τ, nonlinear m	800	91.30	7.62	**3.43**	***1.74***
	1600	90.92	5.08	**1.74**	***0.91***

Notes: Entries are 100× test-set MSEs for $\hat{\tau}(X)$, averaged over 1000 test points and 250 replications; max Monte Carlo SE = 0.24. Constant: treated-minus-control mean difference, applied to all test observations. Bold: best feasible flexible method, including ties at reported precision. Bold italic: overall-best Oracle GAM. Forest: tuned causal forest, no variable screening. Oracle GAM appears only in the two asymmetric designs and is infeasible by construction.

Table 3: Mean test-set MSEs for mixed $p = 6$ designs with 4 X_4 groups

Design	n	Constant	Forest	Pairwise GAM	Full GAM	Oracle GAM
Wager–Athey	800	101.35	**7.80**	9.31	12.01	8.29
	1600	100.13	**4.59**	5.77	7.14	5.28
Smooth interacted	800	57.41	10.29	**9.15**	11.81	***8.17***
	1600	56.27	7.47	**4.81**	6.17	***4.48***
Bilinear m, nonlinear τ	800	58.09	10.68	**5.51**	8.21	***3.85***
	1600	56.58	7.57	**3.03**	4.38	***2.11***
Bilinear τ, nonlinear m	800	93.44	10.37	**4.17**	6.79	***1.66***
	1600	91.97	6.47	**2.27**	3.60	***0.93***

Notes: Entries are 100× test-set MSEs for $\hat{\tau}(X)$, averaged over 1000 test points and 250 replications; max Monte Carlo SE = 0.53. Constant: treated-minus-control mean difference, applied to all test observations. Bold: best feasible flexible method, including ties at reported precision. Bold italic: overall-best Oracle GAM. Pairwise GAM: continuous smooths and pairwise tensors, discrete covariates additive on the mean side. Full GAM: also treatment interactions with discrete covariates. Oracle GAM: true mean and treatment structure. Forest: tuned causal forest, no variable screening.

Table 4: Mean test-set MSEs for mixed $p = 6$ designs with 20 mixture X_4 groups

Design	n	Constant	Forest	Pairwise GAM	Full GAM	Oracle GAM
Wager–Athey	800	102.28	10.68	**9.56**	11.04	***8.65***
	1600	100.06	**5.47**	5.76	6.60	***5.34***
Smooth interacted	800	57.94	11.38	**9.63**	11.26	***8.93***
	1600	56.22	7.64	**4.80**	5.76	***4.49***
Bilinear m, nonlinear τ	800	58.88	15.39	**5.69**	7.27	***4.27***
	1600	56.47	9.35	**2.93**	3.76	***2.26***
Bilinear τ, nonlinear m	800	94.34	14.02	**4.50**	5.98	***1.83***
	1600	91.92	7.73	**2.29**	3.07	***0.97***

Notes: Entries as in Table 3; max Monte Carlo SE = 0.41. The 20 X_4 effects are centered, scale-matched three-component normal-mixture draws. GAMs encode X_4 with `s(X4, bs = "re")`.

6 Conclusion

We compare causal forests with varying-coefficient penalized spline estimators across a set of simulation designs. Our results indicate that causal forests perform best in settings where heterogeneity is naturally represented by adaptive partitions, especially when treatment effects exhibit sharp localized transitions, as in the Wager–Athey logit-style design with continuous predictors.

In many empirically relevant settings, however, modern penalized spline estimators perform very well and often outperform causal forests. This is especially true when treatment

heterogeneity is low-dimensional and smooth, when the baseline mean and treatment effect have different structures, and when the covariates mix continuous, binary, and categorical variables.

An added advantage of spline-based methods is that they allow researchers to specify different parts of the model separately. One can model group effects additively or interactively, include some variables linearly and others flexibly, and control the menu of interactions, which is attractive in empirical settings with time trends, unit or time fixed effects, and categorical variables such as gender or race. At the same time, `mgcv` allows researchers to write a single model formula while smoothing parameters are selected automatically.

Declaration of Generative AI and AI-Assisted Technologies in the Manuscript Preparation Process

During the preparation of this work, the authors used Codex (OpenAI) to assist with simulation code, including implementation of `mgcv::bam()` and `grf::causal_forest()`, and with language editing. All substantive decisions about the simulation designs, estimation methods, interpretation, and presentation of results were made by the authors. After using this tool, the authors reviewed and edited the code, output, and manuscript as needed and take full responsibility for the content of the published article.

Data Availability

No external data were used. The analysis is based on simulated data generated by the replication code. A full replication package, containing all simulation code, run manifests with the exact DGP formulas and seeds, and instructions for reproducing every table, is publicly available in a dedicated GitHub repository. All simulations were run in R version 4.6.0 (R Core Team, 2026) with `grf` version 2.6.1 (Tibshirani et al., 2026) and `mgcv` version 1.9-4 (Wood, 2025); the computational environment is fully documented in the package.

References

Athey, S., J. Tibshirani, and S. Wager (2019): "Generalized Random Forests," *The Annals of Statistics*, 47, 1148–1178.

Chen, X. (2007): "Large Sample Sieve Estimation of Semi-Nonparametric Models," in *Handbook of Econometrics*, Elsevier, vol. 6B, 5549–5632.

Chen, X. and D. Pouzo (2012): "Estimation of Nonparametric Conditional Moment Models With Possibly Nonsmooth Generalized Residuals," *Econometrica*, 80, 277–321.

——— (2015): "Sieve Wald and QLR Inferences on Semi/Nonparametric Conditional Moment Models," *Econometrica*, 83, 1013–1079.

Dandl, S., T. Hothorn, H. Seibold, E. Sverdrup, S. Wager, and A. Zeileis (2024): "What Makes Forest-Based Heterogeneous Treatment Effect Estimators Work?" *The Annals of Applied Statistics*, 18, 506–528.

Fan, J. and W. Zhang (1999): "Statistical Estimation in Varying Coefficient Models," *The Annals of Statistics*, 27, 1491–1518.

Hastie, T. and R. Tibshirani (1993): "Varying-Coefficient Models," *Journal of the Royal Statistical Society. Series B (Methodological)*, 55, 757–779.

Li, Q. and J. S. Racine (2007): *Nonparametric Econometrics: Theory and Practice*, Princeton University Press.

R Core Team (2026): *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Rehill, P. (2025): "How Do Applied Researchers Use the Causal Forest? A Methodological Review," *International Statistical Review*, 93, 288–316.

Rubin, D. B. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701.

Tibshirani, J., S. Athey, R. Friedberg, V. Hadad, D. Hirshberg, L. Miner, E. Sverdrup, S. Wager, and M. Wright (2026): *grf: Generalized Random Forests*, r package version 2.6.1.

Wager, S. and S. Athey (2018): "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests," *Journal of the American Statistical Association*, 113, 1228–1242.

WOOD, S. N. (2011): "Fast Stable Restricted Maximum Likelihood and Marginal Likelihood Estimation of Semiparametric Generalized Linear Models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73, 3–36.

——— (2017): *Generalized Additive Models: An Introduction with R*, Chapman and Hall/CRC, 2 ed.

——— (2025): *mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation*, r package version 1.9-4.

A Additional Simulation Results

This section compares the random-effect GAM estimator with the fixed-effect GAM estimator in the setup with 20 groups. Random-effect GAM treats group coefficients as penalized deviations from zero; fixed-effect GAM includes unrestricted group indicators. Tables A1 and A2 report the resulting MSEs. Two patterns stand out. First, when treatment is not interacted with the discrete covariates (*Pairwise GAM*), there is virtually no difference between the random-effect and fixed-effect approaches. Second, when treatment is interacted with the discrete covariates (*Full GAM*), the fixed-effect approach performs substantially worse.

Table A1: Mean test-set MSEs for mixed $p = 6$ designs with 20 normal X_4 groups and fixed-effect specifications

Design	n	Forest	Pairwise RE GAM	Full RE GAM	Oracle RE GAM	Pairwise FE GAM	Full FE GAM	Oracle FE GAM
Wager–Athey	800	10.67	**9.19**	10.66	8.34	**9.19**	21.24	***8.27***
	1600	**5.52**	5.83	6.64	5.37	5.81	11.31	***5.33***
Smooth interacted	800	11.70	10.00	11.63	9.31	**9.94**	21.93	***9.21***
	1600	7.67	**4.95**	5.94	4.63	**4.95**	10.56	***4.52***
Bilinear m, nonlinear τ	800	16.45	5.81	7.57	4.33	**5.77**	17.63	***4.20***
	1600	9.82	**3.05**	3.86	2.26	**3.05**	8.47	***2.11***
Bilinear τ, nonlinear m	800	13.76	**4.43**	6.04	***1.87***	4.48	16.16	***1.87***
	1600	7.78	**2.39**	3.12	0.91	**2.39**	7.79	***0.90***

Notes: Entries are 100× test-set MSEs, otherwise as in Table 3, averaged over 100 replications; max Monte Carlo SE = 0.61. RE: `s(X4, bs = "re")`; FE: unrestricted X_4 fixed effects. Bold: best feasible method, including ties at reported precision. Bold italic: overall-best Oracle GAM, including ties.

Table A2: Mean test-set MSEs for mixed $p = 6$ designs with 20 mixture X_4 groups and fixed-effect specifications

Design	n	Forest	Pairwise RE GAM	Full RE GAM	Oracle RE GAM	Pairwise FE GAM	Full FE GAM	Oracle FE GAM
Wager–Athey	800	10.11	9.21	10.72	8.38	**9.20**	21.18	***8.27***
	1600	**5.40**	5.83	6.60	5.39	5.81	11.34	***5.32***
Smooth interacted	800	11.68	**9.98**	11.63	9.38	**9.98**	21.93	***9.24***
	1600	7.61	**4.95**	5.93	4.62	4.99	10.57	***4.52***
Bilinear m, nonlinear τ	800	15.85	**5.74**	7.38	4.32	**5.74**	17.59	***4.20***
	1600	9.28	**3.04**	3.86	2.28	3.05	8.48	***2.11***
Bilinear τ, nonlinear m	800	13.81	**4.46**	5.94	***1.87***	**4.46**	16.16	***1.87***
	1600	7.68	**2.37**	3.12	***0.90***	2.40	7.85	***0.90***

Notes: Entries are 100× test-set MSEs, otherwise as in Table 3, averaged over 100 replications; max Monte Carlo SE = 0.54. The 20 X_4 effects are centered, scale-matched normal-mixture draws. RE: `s(X4, bs = "re")`; FE: unrestricted X_4 fixed effects. Bold and bold italic as in Table A1.

Conceptually, FE GAM is less restrictive, which may reduce the risk of misspecification, but it also increases the number of parameters in the model. These effects appear negligible, or offsetting, in *Pairwise GAM*. In *Full GAM*, however, the added complexity dominates, leading to worse performance. In many applied settings, researchers include fixed effects additively rather than interacting treatment with the full set of fixed effects; in such cases, the random-effect and fixed-effect GAM approaches are likely to perform similarly.

B Implementation Details

For all GAM specifications, the estimation sample is converted to a data frame with outcome `y`, treatment `D`, and covariates `X1`, ..., `X6`. We fit:

```
mgcv::bam(formula = formula, data = model_df,
          method = "fREML", select = TRUE)
```

The estimated CATE is computed as the fitted difference between treatment and control predictions at the same covariates:

```
tau_hat <- predict(fit, newdata = transform(X, D = 1), type = "response")
   -
           predict(fit, newdata = transform(X, D = 0), type = "response")
```

For $p = 2$, Pairwise GAM uses:

```
y ~ s(X1) + s(X2) + ti(X1, X2) + D +
    s(X1, by = D) + s(X2, by = D) + ti(X1, X2, by = D)
```

In the asymmetric designs, Oracle GAM replaces the mean-side or treatment-side component with the true lower-dimensional structure:

```
# Bilinear m, nonlinear tau
y ~ X1 + X2 + I(X1 * X2) + D +
    s(X1, by = D) + s(X2, by = D) + ti(X1, X2, by = D)


# Nonlinear m, bilinear tau
```

```
y ~ s(X1) + s(X2) + ti(X1, X2) + D +
    D:X1 + D:X2 + D:I(X1 * X2)
```

For mixed $p = 6$, Pairwise GAM uses smooth and tensor terms for the continuous covariates, additive discrete terms, and treatment-varying smooths only for continuous covariates:

```
y ~ s(X1) + s(X2) + s(X3) +
    ti(X1, X2) + ti(X1, X3) + ti(X2, X3) +
    x4_mean + X5 + X6 + D +
    s(X1, by = D) + s(X2, by = D) + s(X3, by = D) +
    ti(X1, X2, by = D) + ti(X1, X3, by = D) +
    ti(X2, X3, by = D)
```

Full GAM additionally includes the discrete treatment interactions:

```
... + x4_treat + D:X5 + D:X6
```

For 4-group designs, `x4_mean = X4` and `x4_treat = D:X4`. For 20-group random-effect specifications, the corresponding terms are:

```
x4_mean = s(X4, bs = "re")
x4_treat = s(X4, by = D, bs = "re")
```

Appendix fixed-effect variants replace these terms with unrestricted factor versions of X_4 on the mean side and treatment side. Oracle GAM uses the true mean-side and treatment-side continuous structure while retaining the same discrete encoding.

The causal forest benchmark is fit through the project wrapper:

```
fit_causal_forest(data,
  num.trees = 2000,
  variable_selection = FALSE,
  screening.num.trees = 2000,
  tune.parameters = "all",
  seed = seed)
```

This calls `grf::causal_forest(X, Y, W)`. Before fitting, covariates are converted with `model.matrix(~. - 1, data = X)`; categorical variables therefore enter the forest as indi-

cator columns. Since `variable_selection = FALSE`, all generated columns are retained.

The complete runner scripts, estimator helpers, and per-run manifest files recording the exact DGP definitions, model formulas, seeds, and package versions are provided in the replication package (see the Data Availability statement).