Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Decomposed Characteristic Function-Based Goodness-of-Fit Test for Functional Semi-Parametric Regression Models

Mengru You

Ymr0826@outlook.com

Xiangtan University

Research Article

Keywords: Functional data analysis, Goodness-of-fit test, Semi-parametric regression, Characteristic function, Random projection, Functional principal component analysis, U-statistic, Model specification testing

Posted Date: November 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11047979/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A DECOMPOSED CHARACTERISTIC FUNCTION-BASED GOODNESS-OF-FIT TEST FOR FUNCTIONAL SEMI-PARAMETRIC REGRESSION MODELS

MENGRU YOU

School of Business, Xiangtan University, Beierhuan, Yuhu District, Xiangtan 411105, Hunan, China
E-mail: *ymr0826@outlook.com*

Abstract. Functional semi-parametric regression models are widely used to analyze curve-valued predictors, yet model misspecification can severely bias inference. Existing goodness-of-fit tests either target fully parametric models or ignore the hybrid structure of semi-parametric function-on-scalar regressions. This paper proposes a suite of test statistics that decompose model checking into three components: a functional-part statistic, a parametric-part statistic, and a full covariate statistic, all built upon weighted L^2 distances between empirical characteristic functions. The functional covariate is first compressed via functional principal component analysis and then combined with scalar covariates; the resulting statistics are formulated as U-statistics with analytically convenient kernels derived from spherical stable or Laplace densities. We rigorously study the asymptotic behavior under both correct specification and a broad class of alternatives. A key theoretical contribution is the proof that, under the null hypothesis, the estimation effect introduced by replacing unobservable true errors with estimated residuals is asymptotically negligible when the permutation procedure is used for calibration; under alternatives the effect is of the same order as the statistic itself and is rigorously quantified. We also provide jackknife variance estimators that are consistent under alternatives, enabling the construction of critical regions without relying on resampling when $\Delta > 0$. Monte Carlo simulations demonstrate that the functional-part statistic excels at detecting omitted nonlinearities in the functional component, the parametric-part statistic is sensitive to misspecification of the scalar covariates, and the full statistic captures interactive misspecifications, thereby enabling directional diagnosis of model inadequacy. The tests control size accurately even with heavy-tailed errors and maintain high power across a range of sample sizes and dimensions. A real-data application using near-infrared absorbance spectra of meat samples demonstrates how the decomposed statistics can diagnose different sources of model inadequacy in predicting fat content. The decomposed testing framework provides a practical tool for diagnosing model misspecification in functional semi-parametric regression and can be extended to other complex functional models.

Keywords: Functional data analysis, Goodness-of-fit test, Semi-parametric regression, Characteristic function, Random projection, Functional principal component analysis, U-statistic, Model specification testing

1. Introduction

Functional data, where each observation is a curve or function, are now routinely collected in diverse fields such as chemometrics, meteorology, and biomedical engineering [Ramsay and Silverman, 2005, Horváth and Kokoszka, 2012]. Among the many models

2020 *Mathematics Subject Classification.* Primary 62G10, 62J05, 62R10; Secondary 62G20.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

proposed, semi-parametric functional regression models have gained popularity because they combine the interpretability of a parametric component (e.g., linear effects of scalar covariates) with the flexibility of a non-parametric functional component [Aneiros-Pérez and Vieu, 2006, Shin, 2009, Novo and Vieu, 2024]. The general semi-parametric regression framework, including penalized spline and mixed-model representations of smooth terms, is developed comprehensively by Ruppert et al. [2003], while nonparametric approaches for functional data are covered by Ferraty and Vieu [2006]. In such models, the response Y is related to a functional predictor $X(t)$ and a vector of scalar covariates Z through the semi-parametric functional regression model (1), where $\beta(\cdot)$ is an unknown functional coefficient, γ a parameter vector, and ε a random error that is independent of (X, Z) under the correct specification. This model extends the classical partial linear model [Heckman, 1986, Speckman, 1988, Ruppert et al., 2003] to functional data, where the nonparametric component is the inner product $\langle X, \beta \rangle$ rather than a univariate smooth function.

Correct model specification is a prerequisite for valid inference. Misspecifying the functional form, omitting important nonlinear terms or interactions, or using an incorrect link function can seriously bias estimates and degrade predictions. Therefore, developing goodness-of-fit tests specifically tailored to functional semi-parametric regression is of paramount importance.

Model checking for parametric regression has been approached through three complementary strategies: distance-based comparisons between parametric and nonparametric fits [Härdle and Mammen, 1993], conditional-moment tests employing nonparametric kernel estimation [Zheng, 1996], and marked empirical-process diagnostics [Stute, 1997]. A comprehensive review of these and subsequent developments is given by González-Manteiga and Crujeiras [2013]. The curse of dimensionality severely limits the applicability of smoothing-based tests when the covariate dimension is high or moderate. Escanciano [2006] addressed this problem by introducing a projection approach: instead of checking the full conditional moment restriction $E[\varepsilon \mid X] = 0$ directly, he showed that the null is equivalent to $E[\varepsilon \cdot 1(b'X \leq u)] = 0$ for all directions b on the unit sphere, and constructed a Cramér–von Mises statistic based on the residual marked projection process (RMPP). This projection principle avoids the choice of bandwidth or integrating measure and remains powerful in high-dimensional settings. Xu and Zhou [2021] extended this approach to single-index models with unspecified link functions, constructing an omnibus goodness-of-fit test that is consistent against any global alternative and detects local alternatives at the parametric rate. For functional data, Cuesta-Albertos et al. [2019], Chen et al. [2020] have advanced tests that employ empirical processes or random projections for functional linear models, and García-Portugués et al. [2014] proposed a goodness-of-fit test based on integrated projections for the functional linear model with scalar response. García-Portugués et al. [2020] later extended this approach to functional response, providing a unified framework for both scalar and functional predictors and responses. However, these methods are designed for fully parametric functional linear models and do not address the semi-parametric setting with scalar covariates. In the semi-parametric setting, Kloodt et al. [2021] proposed specification tests for transformation models, but their framework does not directly address the functional covariate case. Ferraccioli et al. [2024] proposed nonparametric tests for the linear component of semiparametric regression models based on random sign-flipping of transformed residuals, providing a useful benchmark for inference in partially linear models. Perhaps more importantly, almost all existing tests produce a single global decision (reject or not) and provide no information about the *source* of the misfit. In semi-parametric regression, misspecification can originate from the functional part, the parametric part, or an interaction between the two, and knowing which component is faulty is essential for model revision.

Distance-based measures [Székely and Rizzo, 2013] and characteristic function-based approaches [Chen et al., 2019, 2024] offer flexible toolkits for goodness-of-fit testing because they can be converted into U-statistics with analytically simple kernels. The recent work by Chen et al. [2024] introduced equivalence tests for symmetry, homogeneity, and independence using weighted L^2 distances between characteristic functions. Characteristic function-based approaches have also proven effective in the two-sample setting for functional data. Krzyśko and Smaga [2021] constructed weighted L^2 distances between characteristic functions of random vectors obtained via basis function representation, demonstrating excellent finite-sample properties through permutation tests. Their work illustrates that the characteristic function distance, combined with an appropriate dimensionality reduction of the functional data, yields a powerful and computationally tractable inferential tool. Building on this methodology, the present paper proposes a *decomposed* goodness-of-fit test for functional semi-parametric regression. Instead of computing a single statistic, we construct three statistics that examine different facets of the model:

- T_{func} uses the functional principal component scores $\boldsymbol{\xi}$ (extracted from $X(t)$) as the covariate, thereby being sensitive to misspecification of the functional component;
- T_{param} uses the scalar covariates Z, targeting the parametric part;
- T_{full} uses the full covariate vector $(\boldsymbol{\xi}^\top, Z^\top)^\top$, detecting joint misspecifications.

All three statistics are built from the empirical characteristic function of the residuals and the corresponding covariate, and they share the same U-statistic structure with a spherical kernel.

Our theoretical contribution is twofold. First, we rigorously establish the asymptotic normality of each statistic under fixed alternatives, and we prove that under the null hypothesis the estimation effect from replacing the unobservable true errors by estimated residuals is asymptotically negligible when calibration is performed via permutation; under alternatives the effect is of order $O_p(1)$ and is explicitly quantified. Second, we provide jackknife variance estimators that are consistent under alternatives, enabling the construction of critical regions without relying on resampling when $\Delta > 0$ (though for the degenerate null distribution of the true-error statistic we employ a permutation procedure).

Extensive Monte Carlo simulations demonstrate the directional diagnostic power of the three statistics. When the functional component is misspecified (e.g., omitted nonlinear functional term), T_{func} quickly attains high power while T_{param} remains near the nominal level. Conversely, when the parametric part is incorrect (e.g., omitted quadratic effect in Z), T_{param} is highly effective and T_{func} is insensitive. In cases of interactive or global misspecification, T_{full} exhibits the highest power, and the simultaneous elevation of T_{func} and T_{param} points to both components being affected. These patterns are illustrated using a Tecator near-infrared dataset where we diagnose the specification of a model predicting fat content from absorbance spectra and chemical covariates.

This decomposition is the key novelty that distinguishes our test from existing omnibus tests, which produce only a single global decision and provide no information about the source of the misfit. While distance-based measures [Székely and Rizzo, 2013] and characteristic function-based approaches [Chen et al., 2019, 2024] offer flexible toolkits for goodness-of-fit testing, the present paper moves beyond a single omnibus decision by introducing a decomposed test structure that enables directional diagnosis of model misspecification, together with a rigorous treatment of the residual estimation effect that arises in regression diagnostics but not in the two-sample problem.

The rest of the paper is organized as follows. Section 2 introduces the model and the three decomposed test statistics. Section 3 develops the asymptotic theory, including the

effect of residual estimation. Section 4 describes the practical implementation via permutation tests. Section 5 reports the Monte Carlo simulation results. Section 6 presents the Tecator data analysis. Section 7 concludes with a discussion and directions for future research.

2. Model and test statistics

2.1. **Functional semi-parametric regression.** For each observation $i = 1, \dots, n$, we observe a real-valued response Y_i, a functional covariate $X_i(t)$, $t \in \mathcal{T}$, and a q-dimensional vector of scalar covariates Z_i. The semi-parametric functional regression model is

$$Y_i = \langle X_i, \beta \rangle + Z_i^\top \gamma + \varepsilon_i, \tag{1}$$

where $\beta \in L^2(\mathcal{T})$ is an unknown functional coefficient, $\gamma \in \mathbb{R}^q$ is a parameter vector, and ε_i is an error term. Under the correct specification the errors are independent of (X_i, Z_i).

The null hypothesis of a correctly specified model can be phrased as

$$H_0 : \varepsilon \perp (X, Z),$$

whereas the alternative H_1 allows any form of dependence between ε and (X, Z). Misspecification may arise from an inadequate functional component (e.g. omitted nonlinearity in $X(t)$), from a poorly modelled parametric part (e.g. missing quadratic terms in Z), or from interactions between $X(t)$ and Z.

2.2. **Low-dimensional representation via FPCA.** Because $X_i(t)$ is infinite-dimensional, we extract a small number of features by functional principal component analysis (FPCA). Let $\widehat{\mu}(t)$ and $\widehat{\phi}_1(t), \widehat{\phi}_2(t), \dots$ be the estimated mean and eigenfunctions obtained from the sample $\{X_i\}$. For a chosen truncation level K, we compute the scores

$$\hat{\xi}_{ik} = \int_{\mathcal{T}} \big(X_i(t) - \widehat{\mu}(t)\big)\widehat{\phi}_k(t)\, dt, \qquad k = 1, \dots, K, \tag{2}$$

and form the K-dimensional vector $\hat{\boldsymbol{\xi}}_i = (\hat{\xi}_{i1}, \dots, \hat{\xi}_{iK})^\top$. We then model the functional predictor through these scores, plugging them into (1) as $\langle X_i, \beta \rangle \approx \hat{\boldsymbol{\xi}}_i^\top b$, where $b \in \mathbb{R}^K$ is a coefficient vector. This FPCA-based compression is a data-driven analogue of the basis function representation employed by Krzyśko and Smaga [2021] in the two-sample setting; instead of using fixed basis functions, we let the data determine the eigenfunctions that capture the dominant modes of variation. For notational simplicity we drop the hat and write $\boldsymbol{\xi}_i$ for the scores, but it is understood that all quantities are estimated.

2.3. **Characteristic function distance for independence.** We quantify the distance between the joint distribution of two random vectors $U \in \mathbb{R}$ and $V \in \mathbb{R}^d$ from the situation where they are independent. Let ϕ_U, ϕ_V, and $\phi_{U,V}$ denote the characteristic functions. A weighted L^2 distance is

$$\Delta_w(U, V) = \iint_{\mathbb{R}^{1+d}} \big|\phi_{U,V}(t, s) - \phi_U(t)\phi_V(s)\big|^2 \, w(t, s)\, dt\, ds, \tag{3}$$

where $w(t, s) = w_1(t)w_d(s) > 0$ is a product weight function. If w is chosen as the density of a spherically symmetric distribution, Parseval's identity transforms (3) into an expectation form that avoids numerical integration. Define $C_w(\cdot)$ to be the characteristic function (CF) of the spherical density w. Then Chen et al. [2019, 2024]

$$\begin{aligned}\Delta_w(U, V) = {} & \mathbb{E}\big[C_w(U - U_1)C_w(V - V_1)\big] + \mathbb{E}\big[C_w(U - U_1)\big]\,\mathbb{E}\big[C_w(V - V_1)\big] \\ & - 2\mathbb{E}\big[C_w(U - U_1)C_w(V - V_2)\big],\end{aligned} \tag{4}$$

where (U_1, V_1) and (U_2, V_2) are independent copies of (U, V).

Throughout the paper we employ two families of spherical kernels, both of which yield simple closed-form CFs:

- **Spherical stable kernel:** $C_\gamma(t) = \exp\bigl(-\|t\|^\gamma\bigr)$, $0 < \gamma \le 2$;
- **Generalized Laplace kernel:** $C_\gamma(t) = \bigl(1 + \|t\|^2\bigr)^{-\gamma}$, $\gamma > 0$.

The choice of the kernel function directly affects the power and sensitivity of the test. In the two-sample functional setting, Krzyśko and Smaga [2021] demonstrated that weight functions derived from density distributions (such as the spherical stable family) are generally more powerful than the classical energy statistic, and that different tail behaviours of the weight function can lead to markedly different power profiles. The parameter γ controls the tail behaviour of the weight function. Following Chen et al. [2019], we view γ as indexing a family of Bayesian priors on the frequency domain: smaller γ corresponds to a heavier-tailed prior, which places more emphasis on the characteristic function away from the origin and therefore detects smoothness properties of the underlying density rather than tail behaviour. Conversely, $\gamma = 2$ (the Gaussian kernel) corresponds to a medium-tailed prior that emphasizes the behaviour of the characteristic function near the origin, thereby detecting differences in the tail properties of the distribution. Chen et al. [2019] show that the Gaussian weight ($\gamma = 2$) is "seriously suboptimal" in high dimensions because it is dominated by non-Gaussian weight functions with heavier tails ($\gamma < 2$). In our regression diagnostic setting, however, the dimension of the covariate vector V is typically small (the number of FPCA scores plus scalar covariates), so the high-dimensional bias is negligible. Nevertheless, the choice of γ remains crucial for directional diagnosis: smaller γ amplifies the detection of high-frequency misspecification but increases leakage across components (Section 5.2), while larger γ suppresses leakage at the cost of reduced sensitivity to fine-scale structure. We therefore recommend a data-driven selection of γ (Section 4.4) and provide a systematic comparison of $\gamma \in \{0.5, 1, 2\}$ in the simulation study.

2.4. **Sample statistics as U-statistics.** Given a sample $\{(U_i, V_i)\}_{i=1}^n$, the quantity $\Delta_w(U, V)$ can be estimated by a combination of U-statistics. Define the symmetric kernels

$$\psi_1(u, \tilde{u}; v, \tilde{v}) = C_w(u - \tilde{u})\, C_w(v - \tilde{v}), \tag{5}$$

$$\psi_2(u, \tilde{u}) = C_w(u - \tilde{u}), \tag{6}$$

$$\psi_3(v, \tilde{v}) = C_w(v - \tilde{v}), \tag{7}$$

$$\begin{aligned}\psi_4(u, \tilde{u}; v, \tilde{v}, v^*) = \frac{1}{6}\Big[&C_w(u - \tilde{u})C_w(v - v^*) + C_w(u - u^*)C_w(v - \tilde{v}) \\ &+ C_w(u - \tilde{u})C_w(\tilde{v} - v^*) + C_w(\tilde{u} - u^*)C_w(v - \tilde{v}) \\ &+ C_w(u - u^*)C_w(\tilde{v} - v^*) + C_w(\tilde{u} - u^*)C_w(v - v^*)\Big],\end{aligned} \tag{8}$$

where ψ_4 is the symmetrized version of the kernel $C_w(u-\tilde{u})C_w(v-v^*)$. The corresponding U-statistics are

$$U_{n1} = \frac{1}{n(n-1)} \sum_{i \neq j} \psi_1(U_i, U_j; V_i, V_j), \tag{9}$$

$$U_{n2} = \frac{1}{n(n-1)} \sum_{i \neq j} \psi_2(U_i, U_j), \tag{10}$$

$$U_{n3} = \frac{1}{n(n-1)} \sum_{i \neq j} \psi_3(V_i, V_j), \tag{11}$$

$$U_{n4} = \frac{1}{n(n-1)(n-2)} \sum_{i \neq j \neq k} \psi_4(U_i, U_j; V_i, V_j, V_k). \tag{12}$$

An unbiased estimator of $\Delta_w(U, V)$ is then

$$T_n(U, V) = U_{n1} + U_{n2}U_{n3} - 2U_{n4}. \tag{13}$$

2.5. **Decomposed test statistics for the semi-parametric model.** In our setting we have residuals $\hat{\varepsilon}_i$ obtained from a preliminary fit of the postulated model, which is linear in the first K FPCA scores and the scalar covariates Z. These residuals serve as the U-variable in (13). For the V-variable we construct three different covariate vectors, each targeting a different source of model misspecification:

(1) **Functional component statistic:** $T_{\text{func}} = T_n\big(\hat{\varepsilon},\ V_{\text{func}}\big)$, where $V_{\text{func}} \in \mathbb{R}^{d_f}$ is a collection of features extracted from $X(t)$ that are suspected to capture misspecification of the functional part. In the simplest case V_{func} consists of the first K FPCA scores $\boldsymbol{\xi}$, but it may also be chosen as the scores of higher-order components that were not included in the fitted model, or random projections of $X(t)$.
(2) **Parametric component statistic:** $T_{\text{param}} = T_n\big(\hat{\varepsilon},\ Z\big)$, which uses the scalar covariates Z as the V-variable.
(3) **Full (joint) statistic:** $T_{\text{full}} = T_n\big(\hat{\varepsilon},\ \Xi\big)$, where $\Xi = (V_{\text{func}}^\top, Z^\top)^\top$ combines the features from the functional and parametric parts.

All three statistics share the same algebraic form (13) and differ only in the choice of V. This decomposition separates the residual dependence into the functional and parametric subspaces, analogous to how partial regression diagnostics isolate the contribution of different covariate groups. Under the null hypothesis of correct specification, ε is independent of (X, Z) and consequently also of any measurable function of these variables; therefore $\Delta_w(\varepsilon, V) = 0$ for each choice. Large values of T_{func}, T_{param}, or T_{full} indicate that the corresponding component is misspecified. Moreover, by comparing the magnitudes (or the resulting p-values) of the three statistics the practitioner can gain insight into which part of the model is likely to be wrong, enabling a more targeted model revision.

Remark 2.1 (Independence versus conditional mean zero)**.** The null hypothesis $H_0 : \varepsilon \perp (X, Z)$ is strictly stronger than the conventional conditional-mean restriction $H_0^{\text{mean}} : \mathbb{E}[\varepsilon \mid X, Z] = 0$. A model may satisfy H_0^{mean} while exhibiting conditional heteroscedasticity or other forms of dependence (e.g. $\varepsilon = \sigma(X)\eta$ with $\eta \perp X$ and $\mathbb{E}[\eta] = 0$). Our test will reject such models because the characteristic-function distance (3) is sensitive to *any* departure from independence, not merely to mean dependence.

When is heteroscedasticity a misspecification? Whether conditional heteroscedasticity constitutes model inadequacy depends on the inferential goal. In the following settings it *is* a genuine concern:

- *Efficient estimation.* If the practitioner intends to use generalised least squares (GLS) or weighted least squares (WLS) to obtain asymptotically efficient estimators, a correct specification of $\mathrm{Var}(\varepsilon \mid X, Z)$ is required. Conditional heteroscedasticity renders the ordinary least-squares standard errors invalid, even when the mean function is correct.
- *Prediction intervals.* For constructing valid prediction intervals, one needs the entire conditional distribution of Y given (X, Z), not merely its mean. Heteroscedasticity widens or narrows the intervals in a covariate-dependent way that the homoscedastic model fails to capture.
- *Risk-sensitive applications.* In financial or medical contexts where tail risk (e.g. value-at-risk, interventional safety) is of interest, conditional variance is itself a target of inference. A model that predicts the mean correctly but ignores variance heterogeneity will systematically underestimate extreme risks.

In these scenarios the independence null H_0 serves as a comprehensive diagnostic: it detects not only mean misspecification but also second-order and higher-order departures that compromise the validity of downstream inference.

What if only the conditional mean matters? If the practitioner is concerned solely with the conditional mean and intends to report heteroscedasticity-robust (Eicker–Huber–White) standard errors, then conditional heteroscedasticity is not a misspecification in the substantive sense. In this case the independence-based test may be viewed as overly conservative. Two practical adjustments are available:

(1) *Studentized residuals.* Compute the nonparametric conditional variance estimate $\hat{\sigma}^2(X_i, Z_i)$ (e.g. by local polynomial regression or random-forest regression of $\hat{\varepsilon}_i^2$ on (X_i, Z_i)), and apply the test to the studentized residuals $\tilde{\varepsilon}_i = \hat{\varepsilon}_i / \hat{\sigma}(X_i, Z_i)$. Under H_0^{mean} with arbitrary conditional variance, $\tilde{\varepsilon}_i$ is approximately independent of (X_i, Z_i), so the test targets the mean restriction without rejecting variance dependence. Because the studentization introduces an additional estimation step, the permutation procedure must be applied to $\tilde{\varepsilon}_i$ rather than to $\hat{\varepsilon}_i$; the asymptotic validity follows from the same arguments as Theorem 3.7 provided $\hat{\sigma}(\cdot)$ is consistent.

(2) *Variance-stabilizing transformation.* If a parametric form of the conditional variance is suspected (e.g. $\sigma^2(X) = \exp(\delta^\top X)$), estimate δ and transform the response as $Y_i^* = Y_i / \hat{\sigma}(X_i, Z_i)$. The test is then applied to the transformed-model residuals. This approach is simpler than studentization but requires a correct variance specification.

Practical diagnostic. When the test rejects H_0, the practitioner should first determine whether the rejection is driven by mean misspecification or by higher-order dependence. A simple graphical diagnostic is to plot the absolute residuals $|\hat{\varepsilon}_i|$ against the fitted values $\hat{Y}_i = \langle X_i, \hat{\beta} \rangle + Z_i^\top \hat{\gamma}$ and against each covariate. A systematic pattern (e.g. a funnel shape in the $|\hat{\varepsilon}|$ vs. $\hat{Y}$ plot) indicates conditional heteroscedasticity, whereas a random scatter with isolated clusters suggests mean misspecification. In the Tecator analysis (Section 6), the absence of a funnel pattern in the residual plots supports the interpretation that the rejection is due to mean misspecification rather than to heteroscedasticity.

Theoretical underpinning. The stronger null is also what enables the permutation calibration: the exchangeability of residuals required by Theorem 3.7 holds under independence but not under conditional mean zero alone. Simulation scenario M6 (Table 7) confirms that the test does not spuriously reject under homoscedasticity ($\theta = 0$: empirical size 0.026), while it correctly detects heteroscedasticity when present ($\theta > 0$). Thus the

independence null broadens the diagnostic scope without sacrificing size control under standard conditions.

In the simulations of Section 5 we illustrate this directional diagnostic property by setting V_{func} to be the FPCA scores that were actually omitted from the fitted model (e.g. higher-order scores in scenario M5) or, when the functional part is correctly specified, to the full set of scores used in the fit (scenarios M1 and M3). In practice one may let V_{func} be the complete set of FPCA scores or a random projection of $X(t)$; the choice can be guided by domain knowledge about likely forms of misspecification. In the simulations, we let V_{func} be either the full set of FPCA scores retained in the fitted model (scenarios M1–M4) or the higher-order scores omitted from the fit (scenario M5), demonstrating the flexibility of the framework.

Remark 2.2 (Residual estimation)**.** The statistics above are computed with the estimated residuals $\hat{\varepsilon}_i$ rather than the true errors ε_i. In Section 3 we show that under the null hypothesis the estimation effect is asymptotically negligible and is correctly absorbed by the permutation procedure of Section 4.1, yielding an asymptotically valid test. Under alternatives the effect is of the same order as the statistic itself and is rigorously quantified in Theorem 3.3.

3. Asymptotic theory

3.1. **Assumptions.** Let $(X_i, Z_i, \varepsilon_i)$, $i = 1, \ldots, n$, be i.i.d. observations from the model (1). Denote by $\boldsymbol{\xi}_i$ the vector of FPCA scores extracted from X_i, and let V stand for any of the three covariate vectors (V_{func}, Z, Ξ) used in the tests. We impose the following conditions.

Assumption 1 (Moments and design)**.** $\mathbb{E}|\varepsilon|^4 < \infty$, the FPCA scores $\boldsymbol{\xi}$ have finite fourth moments, and the scalar covariates Z have finite fourth moments. Moreover, the functional predictor satisfies $\mathbb{E}[\|X\|_{L^2}^4] < \infty$.

Assumption 2 (Kernel)**.** The function $C_w(\cdot)$ is bounded, continuous, and satisfies $C_w(0) = 1$, $C_w(-t) = C_w(t)$. It is Lipschitz continuous with constant L (i.e. $|C_w(t) - C_w(s)| \leq L\|t - s\|$ for all t, s).

Assumption 3 (Estimation)**.** The estimators $(\hat{b}, \hat{\gamma})$ obtained from the OLS fit on the K retained FPCA scores satisfy $\|\hat{b} - b_0\| = O_p(n^{-1/2})$ and $\|\hat{\gamma} - \gamma_0\| = O_p(n^{-1/2})$ under the correct model. The number of retained FPCA components K is a fixed finite constant (independent of n), and the first K eigenvalues of the covariance operator of $X(t)$ are strictly positive, pairwise distinct, and have gaps bounded away from zero. Under these conditions, Hall and Hosseini-Nasab [2006, Theorem 3] show that the FPCA eigenfunctions and score vectors are estimated at the parametric rate $n^{-1/2}$. Consequently, the functional coefficient estimator $\hat{\beta}(t) = \sum_{k=1}^{K} \hat{b}_k \widehat{\phi}_k(t)$ satisfies $\|\hat{\beta} - \beta_0\|_{L^2} = O_p(n^{-1/2})$.

3.2. **Discussion of the assumptions.** Assumption 1 (finite fourth moments) is standard in the U-statistics literature. For bounded kernels, the U-statistic CLT and the permutation-consistency arguments remain valid under the weaker condition $\mathbb{E}[\varepsilon^2] < \infty$ because the first-order projection functions g_ℓ are bounded. The fourth-moment condition is imposed primarily to ensure finite variance of the jackknife estimator and to permit higher-order expansions. In practice, most error distributions satisfy this assumption, including the normal and contaminated normal distributions.

Moment conditions under bounded kernels. When the kernel C_w is bounded (Assumption 2), the U-statistic kernels ψ_ℓ appearing in Lemma 3.1 and Theorem 3.2 are themselves bounded ($|\psi_\ell| \leq 1$), so that the integrability condition $E_F[h^2] < \infty$ required

by the classical U-statistic CLT [Serfling, 1980, Theorem 5.5.1A] is automatically satisfied. Consequently, the asymptotic normality of the true-error statistic $T_n(\varepsilon, V)$ under alternatives, as well as the permutation-consistency arguments of Theorem 3.7 under the null, hold without any moment condition on ε beyond the existence of the variance $\sigma^2(\Delta)$. The fourth-moment requirement in Assumption 1 is needed *only* for the jackknife variance estimator of Theorem 3.12, whose consistency relies on the finite-sample variance of the delete-one U-statistics, and for the higher-order expansions used in the refinement of Theorem 3.3 under H_0.

This distinction has direct practical consequences. Heavy-tailed errors such as the Student t_3 distribution used in our simulations (Table 6) do *not* possess finite fourth moments, so Assumption 1 is technically violated. Nevertheless, because the kernels are bounded, the test statistic itself is still asymptotically normal under the alternative and the permutation test remains asymptotically valid under the null, exactly as predicted by Serfling's theory. The empirical size reported in Table 6 confirms this: at $\gamma = 2$ and $n = 500$ the rejection rate is 4.5% (nominal 5%), and the directional pattern $T_{\text{param}} \gg T_{\text{func}}$ is preserved. The slight conservatism relative to the Gaussian case is attributable to the heavier tails of the t_3 distribution inflating the residual variance, not to a failure of the test calibration.

Practical guidance. For applications where the error distribution is known to be heavy-tailed (e.g. financial returns, environmental extremes), the practitioner may proceed with the decomposed test using the standard permutation calibration, but should be aware that the jackknife variance estimator may be unreliable if the fourth moment is infinite. In such cases the permutation p-value remains the preferred calibration device; the jackknife standard error and the associated z-test should be used with caution. A rigorous Berry–Esseen bound under the minimal moment conditions afforded by bounded kernels, and the resulting optimal rate for the permutation approximation, are left for future work.

Assumption 2 (Lipschitz continuity) is required for the residual estimation effect established in Theorem 3.3. For the spherical stable kernel, this assumption holds if and only if $\gamma \geq 1$ (Proposition 3.10). When $0 < \gamma < 1$, the kernel is only γ-Hölder continuous, and the residual estimation effect becomes $O_p(n^{-\gamma/2})$ rather than $O_p(1)$.

Assumption 3 ($n^{-1/2}$-consistent estimators) is satisfied by the OLS estimator when the model is correctly specified and the errors have finite variance. Under misspecification, the OLS estimator still converges to the pseudo-true value at rate $n^{-1/2}$ under mild regularity conditions, so the assumption remains valid.

On the fixed-K FPCA assumption. Assumption 3 requires that the number of retained FPCA components K be a fixed finite constant that does not grow with the sample size. This assumption serves two purposes. First, it guarantees that the FPCA estimation error is $O_p(n^{-1/2})$ by ensuring that the K-th and $(K+1)$-th population eigenvalues are separated by a positive gap, so that the sample eigenfunctions are identifiable and converge at the parametric rate [Hall and Hosseini-Nasab, 2006, Theorem 3]. When K is allowed to increase with n, the uniform estimation error $\max_{k \leq K} \|\widehat{\phi}_k - \phi_k\|$ may exceed $O_p(n^{-1/2})$ because the eigenvalue gaps $\lambda_k - \lambda_{k+1}$ typically shrink as k grows, and the sample eigenfunctions become unstable. Second, fixing K implies that the true functional coefficient $\beta_0(t) = \sum_{k=1}^{\infty} b_{0k}\phi_k(t)$ is exactly representable by its first K terms in the fitted model; any truncation error beyond K is absorbed into the error term. In our simulations and real-data analysis we treat K as fixed ($K \in \{3, 5, 10\}$) and verify that the cumulative variance explained exceeds 99%, so the truncation error is negligible relative to the estimation error (Table 14, sensitivity panel).

Relaxing the fixed-K assumption to allow $K = K_n \to \infty$ at a controlled rate would require bounds on the decay of the eigenvalues λ_k and their spacings. If K_n grows too quickly, the residual estimation effect in Theorem 3.3 could dominate the test statistic, and the permutation approximation in Lemma 3.6 would no longer be $o_p(1)$. Developing a theory that permits $K_n \to \infty$ while maintaining the validity of the decomposed tests is an important direction for future work.

3.3. **Asymptotics with known errors.** Let $T_n(\varepsilon, V)$ be the statistic computed from the true errors ε_i and a covariate vector V. Define the corresponding population distance $\Delta = \Delta_w(\varepsilon, V)$ as in (4).

Lemma 3.1 ([Chen et al., 2024, Theorem 3.1])**.** *Suppose (U, V) are i.i.d. with $\mathbb{E}|U|^2 < \infty$ and $\mathbb{E}\|V\|^2 < \infty$. If $\Delta > 0$, then*

$$\sqrt{n}\big(T_n(U, V) - \Delta\big) \xrightarrow{d} N\big(0, \sigma^2(\Delta)\big),$$

where

$$\sigma^2(\Delta) = \sum_{i,j=1}^{4} m_i m_j \, h_i' \, h_j' \, \zeta_1(i, j),$$

with $h(u) = u_1 + u_2 u_3 - 2u_4$, $h_i' = \partial h / \partial u_i$ evaluated at $\boldsymbol{\theta} = (\mathbb{E}[U_{n1}], \ldots, \mathbb{E}[U_{n4}])^\top$, $m_1 = m_2 = m_3 = 2$, $m_4 = 3$, and $\zeta_1(i, j)$ the first-order Hájek covariances as defined in Appendix A. If $\Delta = 0$, then $\sqrt{n}\, T_n(U, V) \xrightarrow{p} 0$.

Applying Lemma 3.1 with $U = \varepsilon$ and V equal to each of our three covariate vectors yields the following result.

Theorem 3.2 (Known errors)**.** *Under Assumptions 1–2 and for any of $V \in \{V_{\text{func}}, Z, \Xi\}$:*

1. *If $\Delta_w(\varepsilon, V) > 0$, then $\sqrt{n}\big(T_n(\varepsilon, V) - \Delta_w(\varepsilon, V)\big) \xrightarrow{d} N\big(0, \sigma^2(\Delta_w(\varepsilon, V))\big)$.*
2. *If $\Delta_w(\varepsilon, V) = 0$, then $\sqrt{n}\, T_n(\varepsilon, V) \xrightarrow{p} 0$.*

Part (ii) shows that under the null hypothesis the statistic converges to zero at a rate faster than $n^{-1/2}$; hence the asymptotic null distribution is degenerate and standard normal quantiles cannot be used to calibrate the test. We therefore use permutation or bootstrap procedures to obtain critical values (see Section 4).

3.4. **Effect of residual estimation.** In practice the true errors are unavailable and we must replace them by $\hat{\varepsilon}_i = Y_i - \langle X_i, \hat{\beta} \rangle - Z_i^\top \hat{\gamma}$. The following theorem quantifies the impact of this substitution.

Theorem 3.3 (Residual estimation effect)**.** *Under Assumptions 1–3, for each V,*

$$\sqrt{n}\big(T_n(\hat{\varepsilon}, V) - T_n(\varepsilon, V)\big) = O_p(1). \tag{14}$$

Proposition 3.4 (Influence function ϕ)**.** *In the expansion of Theorem 3.3, the influence function $\phi(\varepsilon, V)$ has the explicit form*

$$\phi(\varepsilon, V) = \sum_{\ell=1}^{4} h_\ell' \, m_\ell \, \tilde{g}_\ell(\varepsilon, V),$$

where $\tilde{g}_\ell(\varepsilon, V) = \mathbb{E}[\psi_\ell(\varepsilon, \cdot, \cdot; V, \cdot) \mid \varepsilon, V] - \theta_\ell$ is the first-order projection of the ℓ-th kernel with respect to the error variable. Consequently ϕ is bounded and satisfies $\mathbb{E}[\phi(\varepsilon, V)] = 0$.

Remark 3.5. Proposition 3.4 makes the object that was only implicit in the original proof of Theorem 3.3 completely explicit. Under H_0, the orthogonality $\mathbb{E}[\phi(\varepsilon, V)\tilde{X}] = 0$ guarantees that the linear term is $O_p(n^{-1})$, so the estimation effect is asymptotically

negligible. Under alternatives the linear term is $O_p(n^{-1/2})$ because the sample averages converge to non-zero limits.

Lemma 3.6 (Permutation approximation bound)**.** *Under Assumptions 1–3 and H_0, let π be a uniformly random permutation of $\{1, \ldots, n\}$. Then*

$$\sup_{t\in\mathbb{R}} \left| P^*\big(T_n(\hat{\varepsilon}^\pi, V) \le t\big) - P^*\big(T_n(\varepsilon^\pi, V) \le t\big) \right| \xrightarrow{p} 0,$$

where P^* denotes the conditional probability given the data $\{(X_i, Z_i, Y_i)\}_{i=1}^n$.

Theorem 3.7 (Validity of the permutation test)**.** *Under Assumptions 1–3 and H_0, let $c^*_{n,1-\alpha}$ be the $(1-\alpha)$ quantile of the permutation distribution of $T_n(\hat{\varepsilon}, V)$. Then*

$$P\big(T_n(\hat{\varepsilon}, V) > c^*_{n,1-\alpha}\big) \longrightarrow \alpha.$$

Remark 3.8. Lemma 3.6 and Theorem 3.7 replace the earlier sketch proof of permutation consistency with a fully rigorous argument. The key insight is that the Lipschitz property of C_w makes the difference between $T_n(\hat{\varepsilon}^\pi, V)$ and $T_n(\varepsilon^\pi, V)$ uniformly small, while the i.i.d. nature of ε guarantees exact exchangeability under H_0. Together, these two facts imply that the permutation distribution of the estimated-residual statistic is asymptotically identical to the true null distribution.

Proposition 3.9 (Jackknife consistency for residual-based statistics)**.** *Theorem 3.12 remains valid when the statistic T_n is computed from estimated residuals $\hat{\varepsilon}$ rather than from the true errors ε. The additional error introduced by the residual estimation in the delete-one statistics is $O_p(n^{-3/2})$ and therefore does not affect the limit of $\hat{\sigma}_n^2$.*

Proposition 3.10 (Regularity conditions for the spherical stable kernel)**.** *For the spherical stable kernel $C_w(t) = \exp(-\|t\|^\gamma)$, the following holds:*

(a) *Assumptions 1 and 2 (boundedness, continuity) are satisfied for all $\gamma > 0$.*
(b) *The Lipschitz condition required in Theorem 3.3 holds if and only if $\gamma \ge 1$. For $0 < \gamma < 1$, C_w is only γ-Hölder continuous at the origin, and the linearisation error in Theorem 3.3 becomes $O_p(n^{-\gamma/2})$ rather than $O_p(n^{-1/2})$. Because $n^{-\gamma/2} = o(1)$ for every $\gamma > 0$, the permutation test remains asymptotically valid (Lemma 3.6), but larger sample sizes may be needed for accurate finite-sample calibration. The simulation results with $\gamma = 0.5$ should therefore be interpreted as illustrating the sensitivity-specificity trade-off rather than as a formally endorsed default.*
(c) *Theorem 3.12 (Jackknife) and the asymptotic normality in Theorem 3.2 require only boundedness and continuity of C_w, so they hold for all $\gamma > 0$.*

Remark 3.11 (Interpretation of $\gamma = 0.5$ results)**.** The superior power of $\gamma = 0.5$ in scenarios M1 and M3 (Table 4) is not paradoxical: it reflects a genuine sensitivity–specificity trade-off governed by the kernel's tail behaviour.

Mechanism. The spherical stable kernel $C_w(t) = \exp(-\|t\|^\gamma)$ with $\gamma = 0.5$ decays more slowly in the tails than the Gaussian ($\gamma = 2$) or Laplace ($\gamma = 1$) kernels. Consequently, pairs (i, j) with large covariate distances $\|V_i - V_j\|$ contribute more to the U-statistic, effectively up-weighting the extreme points in the covariate space. For smooth alternatives (e.g. polynomial misspecification), the extreme points carry the strongest signal, so the heavier-tailed kernel amplifies the detectable effect.

Cost. The same mechanism is responsible for the elevated leakage in M4: when the parametric component is misspecified, the extreme points in the Z-space also project onto the functional scores through the estimated $\hat{\beta}$, and the slow-decaying kernel preserves this spurious dependence. The result is a directional reliability of only 29.2% at $\gamma = 0.5$ versus 76.4% at $\gamma = 2$ (Table 5).

Calibration. Importantly, the size control under M0 is not compromised (Table 3), because the null distribution is centred at zero regardless of γ and the permutation calibration is exact under exchangeability. The issue is not size distortion but *directional ambiguity*: $\gamma = 0.5$ tells you *that* something is wrong, but not *where*, with high confidence.

Practical recommendation. Use $\gamma = 0.5$ for exploratory screening where the goal is to detect any departure from the null with maximum sensitivity, accepting that the directional information will be noisy. Use $\gamma = 2$ for confirmatory analysis where reliable directional diagnosis is required. If $\gamma = 0.5$ rejects while $\gamma = 2$ does not, suspect that the rejection is driven by second-order or extreme-point effects rather than by systematic mean misspecification, and verify with the studentized-residual adjustment of Section 4.2.

3.5. **Jackknife variance estimation.** When $\Delta > 0$, the limit variance $\sigma^2(\Delta)$ depends on the unknown joint distribution of (ε, V) and must be estimated from the data. We adopt the jackknife estimator, which is simple to implement and does not require the computation of the influence functions $\zeta_1(i, j)$.

For a given statistic T_n and sample size n, let $T_{n-1}^{(-i)}$ be the statistic computed from the sample with the ith observation removed. Define

$$\hat{\sigma}_n^2 = (n-1)\sum_{i=1}^{n}\big(T_{n-1}^{(-i)} - \bar{T}_{n-1}\big)^2, \qquad \bar{T}_{n-1} = \frac{1}{n}\sum_{i=1}^{n} T_{n-1}^{(-i)}. \tag{15}$$

Theorem 3.12 (Jackknife consistency)**.** *Under Assumptions 1–2 and $\Delta > 0$, $\hat{\sigma}_n^2 \xrightarrow{p} \sigma^2(\Delta)$ for each of the three statistics $T_{\text{func}}, T_{\text{param}}, T_{\text{full}}$.*

3.6. **Joint distribution of the three statistics.**

Theorem 3.13 (Joint asymptotic normality)**.** *Let $\mathbf{T}_n = (T_{\text{func}}, T_{\text{param}}, T_{\text{full}})^\top$ and $\boldsymbol{\Delta} = (\Delta_{\text{func}}, \Delta_{\text{param}}, \Delta_{\text{full}})^\top$ with the obvious notation. Under the conditions of Theorem 3.2(i),*

$$\sqrt{n}\big(\mathbf{T}_n - \boldsymbol{\Delta}\big) \xrightarrow{d} N(0, \Sigma),$$

where Σ has entries σ_{jk} that can be expressed in terms of the joint Hájek covariances of the kernels involved in the three statistics (see Section 3.7).

3.7. **Hájek covariances $\zeta_1(i, j)$.** For kernel ψ_ℓ of order m_ℓ, the first-order projection is $g_\ell(w) = \mathbb{E}[\psi_\ell(w, W_2, \ldots, W_{m_\ell})] - \theta_\ell$, with $W_i = (\varepsilon_i, V_i)$. The Hájek covariance is $\zeta_1(i, j) = \mathrm{Cov}(g_i(W), g_j(W))$. Because kernels are bounded (Assumption 2), all moments are finite.

Explicit expressions. Write C_w for the characteristic-function kernel with shape parameter w. For the second-order kernels ($m = 2$) appearing in U_{n1}, U_{n2}, U_{n3},

$$\begin{aligned}
\zeta_1(1,1) &= \mathrm{Var}\big\{\mathbb{E}[C_w(\varepsilon - \varepsilon')C_w(V - V') \mid W]\big\},\\
\zeta_1(2,2) &= \mathrm{Var}\big\{\mathbb{E}[C_w(\varepsilon - \varepsilon') \mid \varepsilon]\big\},\\
\zeta_1(3,3) &= \mathrm{Var}\big\{\mathbb{E}[C_w(V - V') \mid V]\big\},\\
\zeta_1(1,2) &= \mathrm{Cov}\big\{\mathbb{E}[C_w(\varepsilon - \varepsilon')C_w(V - V') \mid W],\ \mathbb{E}[C_w(\varepsilon - \varepsilon') \mid \varepsilon]\big\},\\
\zeta_1(1,3) &= \mathrm{Cov}\big\{\mathbb{E}[C_w(\varepsilon - \varepsilon')C_w(V - V') \mid W],\ \mathbb{E}[C_w(V - V') \mid V]\big\},\\
\zeta_1(2,3) &= \mathrm{Cov}\big\{\mathbb{E}[C_w(\varepsilon - \varepsilon') \mid \varepsilon],\ \mathbb{E}[C_w(V - V') \mid V]\big\}.
\end{aligned}$$

For the third-order kernel ($m = 3$) appearing in U_{n4},

$$\zeta_1(4,4) = \mathrm{Var}\big\{\mathbb{E}[C_w(\varepsilon-\varepsilon')C_w(V-V'') \mid W]\big\},$$
$$\zeta_1(1,4) = \mathrm{Cov}\big\{\mathbb{E}[C_w(\varepsilon-\varepsilon')C_w(V-V') \mid W],\ \mathbb{E}[C_w(\varepsilon-\varepsilon')C_w(V-V'') \mid W]\big\},$$
$$\zeta_1(2,4) = \mathrm{Cov}\big\{\mathbb{E}[C_w(\varepsilon-\varepsilon') \mid \varepsilon],\ \mathbb{E}[C_w(\varepsilon-\varepsilon')C_w(V-V'') \mid W]\big\},$$
$$\zeta_1(3,4) = \mathrm{Cov}\big\{\mathbb{E}[C_w(V-V') \mid V],\ \mathbb{E}[C_w(\varepsilon-\varepsilon')C_w(V-V'') \mid W]\big\}.$$

These expressions involve only the joint distribution of (ε, V) and can be estimated empirically via the jackknife procedure described in Section 3.5.

4. Implementation

4.1. **Permutation test algorithm.** Under the null hypothesis, Theorem 3.2(ii) shows that the true-error statistic $T_n(\varepsilon, V)$ converges to zero at a rate faster than $n^{-1/2}$; its asymptotic null distribution is therefore degenerate, and standard normal quantiles cannot be used for calibration. Moreover, the residual-based statistic $T_n(\hat{\varepsilon}, V)$ contains an additional estimation effect (Theorem 3.3) that further complicates the null distribution. Following the successful precedent of Krzyśko and Smaga [2021] in the two-sample setting for functional data, we approximate the null distribution through a permutation procedure that respects the exchangeability of the residuals.

Because the test statistic T_n is a U-statistic measuring the dependence between the residuals and the covariates V, the null hypothesis $\varepsilon \perp V$ implies that the residuals are exchangeable with respect to V. Permuting the residual indices therefore destroys any systematic dependence while preserving the marginal distribution of the residuals, yielding an exact or approximate null distribution under H_0.

For computational efficiency we exploit the fact that the covariate-dependent kernel matrix $\mathbf{K}_V$ does *not* depend on the residuals and can be pre-computed once for all permutation replicates. The permutation test for a generic statistic $T_n(\hat{\varepsilon}, V)$ proceeds as follows.

Algorithm (Permutation goodness-of-fit test). **Input:** Residuals $\hat{\varepsilon} = (\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n)^\top$, covariate matrix $\mathbf{V} = (V_1, \ldots, V_n)^\top$, kernel parameter γ, number of permutations B.

(1) **Standardise** the residuals and the covariates: $\tilde{\varepsilon}_i = (\hat{\varepsilon}_i - \bar{\hat{\varepsilon}})/s_{\hat{\varepsilon}}$, $\tilde{V}_{ik} = (V_{ik} - \bar{V}_k)/s_{V_k}$, where s denotes the sample standard deviation.
(2) **Pre-compute** the covariate kernel matrix $\mathbf{K}_V \in \mathbb{R}^{n\times n}$ with entries $(\mathbf{K}_V)_{ij} = C_\gamma(\|\tilde{V}_i - \tilde{V}_j\|)$, and set its diagonal to zero. Also compute $U_3 = (n(n-1))^{-1}\sum_{i\neq j}(\mathbf{K}_V)_{ij}$ and the row sums $r_i = \sum_j (\mathbf{K}_V)_{ij}$.
(3) **Observed statistic.** Form $\tilde{\boldsymbol{\varepsilon}} = (\tilde{\varepsilon}_1, \ldots, \tilde{\varepsilon}_n)^\top$, compute the residual kernel matrix $\mathbf{K}_\varepsilon$ with $(\mathbf{K}_\varepsilon)_{ij} = C_\gamma(|\tilde{\varepsilon}_i - \tilde{\varepsilon}_j|)$, zero out the diagonal, and evaluate $T_{\mathrm{obs}} = T_n(\hat{\varepsilon}, V)$ using equation (13).
(4) **Permutation replicates.** For $b = 1, \ldots, B$:
   (a) Draw a uniform random permutation π of $\{1, \ldots, n\}$.
   (b) Set $\tilde{\varepsilon}_i^{(b)} = \tilde{\varepsilon}_{\pi(i)}$ for all i.
   (c) Compute $T^{(b)} = T_n(\tilde{\varepsilon}^{(b)}, V)$ using the pre-computed $\mathbf{K}_V$, U_3 and row sums.
(5) **p-value.** Count the number of replicates that exceed the observed value: $N_{\mathrm{exc}} = \sum_{b=1}^{B} \mathbf{1}\{T^{(b)} \geq T_{\mathrm{obs}}\}$. The one-sided permutation p-value is

$$p_{\mathrm{perm}} = \frac{N_{\mathrm{exc}} + 1}{B + 1}. \tag{16}$$

(6) **Decision.** Reject H_0 at level α if $p_{\mathrm{perm}} \leq \alpha$.

Remarks on Algorithm 4.1.

(1) *One-sided versus two-sided testing.* The statistic T_n is a distance, so $T_n \geq 0$ and large values indicate departure from H_0. Consequently the test is naturally one-sided: we reject only when the observed distance is unusually large. Under H_0 the permuted statistics $T^{(b)}$ are centred near zero, so a two-sided test is neither necessary nor appropriate.
(2) *Davison–Hinkley +1 correction.* Equation (16) adds one to both the numerator and the denominator. This ensures that $p_{\text{perm}} > 0$ for every finite B, preventing the anti-conservative boundary case $p = 0$ when T_{obs} happens to be the largest among the $B + 1$ values (observed plus permuted). The correction is standard in permutation testing and amounts to including the observed data set as an additional "permutation" replicate.
(3) *Choice of B.* The Monte Carlo standard error of the p-value is approximately $\sqrt{p(1-p)/B}$. With $B = 1\,000$ and nominal level $\alpha = 0.05$, the standard error is 0.007, which is adequate for most applications. For very small p-values (e.g. in genomic screening) one may increase B to 10 000 or more.
(4) *Computational shortcut.* Because the covariate kernel $\mathbf{K}_V$ and its derived quantities (U_3, row sums) are fixed across permutations, each replicate costs only $O(n^2)$ operations for the residual kernel and the U-statistic arithmetic. The total cost is $O(n^2 + Bn^2) = O(Bn^2)$, which is dominated by the permutation loop for $B \gg 1$.

Remark 4.1 (Exchangeability and dependent data)**.** The validity of Theorem 3.7 rests on the i.i.d. assumption stated in Assumption 3. Under i.i.d. sampling, the true errors $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distributed, which guarantees exact exchangeability under H_0. If the data exhibit temporal or spatial dependence—for example, when the observations are time-series curves or spatially referenced functional data—the residuals may no longer be exchangeable, and the permutation p-value may be anti-conservative.

For dependent data, we recommend one of the following alternatives:

(1) **Block permutation.** If the dependence is short-range (e.g. AR(1) time series), divide the sample into contiguous blocks of length ℓ exceeding the effective dependence length, and permute the blocks rather than individual observations. The block size ℓ can be chosen by inspecting the autocorrelation function of the residuals.
(2) **Wild bootstrap.** Generate bootstrap residuals as $\varepsilon_i^* = \xi_i \hat{\varepsilon}_i$, where ξ_i are i.i.d. Rademacher (± 1 with probability 1/2) or Mammen multipliers. The wild bootstrap preserves the conditional heteroscedasticity structure and remains valid under martingale difference errors.
(3) **Spatial block bootstrap.** For spatially dependent data, divide the domain into rectangular blocks, resample the blocks with replacement, and re-assemble the spatial field. This preserves the local spatial correlation while destroying long-range dependence.

Developing a full theory for the decomposed statistics under these dependent-data schemes—including the impact of block size on power and the validity of the directional diagnosis—is beyond the scope of the present paper but constitutes an important direction for future work.

We use $B = 500$ in the simulation study and $B = 1\,000$ in the Tecator real-data analysis. For the decomposed test, the three statistics T_{func}, T_{param}, and T_{full} are calibrated independently by applying Algorithm 4.1 with the respective covariate matrices V_{func}, Z, and Ξ.

4.2. **Adjustment for pure mean inference: studentized residuals.** As discussed in Remark 2.1, the independence null $H_0 : \varepsilon \perp (X, Z)$ detects not only mean misspecification but also conditional heteroscedasticity and higher-order dependencies. When the practitioner is concerned solely with the conditional mean and plans to report heteroscedasticity-robust standard errors, the rejection of a correctly specified mean model due to benign heteroscedasticity may be regarded as overly conservative. In such settings we recommend applying the test to *studentized residuals*.

Procedure. Let $\hat{\varepsilon}_i$ be the OLS residuals from the fitted model (1). Estimate the conditional variance $\sigma^2(X_i, Z_i) = \mathbb{E}[\varepsilon_i^2 \mid X_i, Z_i]$ by a nonparametric regression of $\hat{\varepsilon}_i^2$ on the covariates. In our implementation we use a $\mathcal{K}$-nearest-neighbour (KNN) estimator with $\mathcal{K} = \lfloor n^{1/2} \rfloor$:

$$\hat{\sigma}_i^2 = \frac{1}{\mathcal{K}} \sum_{j \in \mathcal{N}_{\mathcal{K}}(i)} \hat{\varepsilon}_j^2, \tag{17}$$

where $\mathcal{N}_{\mathcal{K}}(i)$ denotes the set of $\mathcal{K}$ nearest neighbours of observation i in the covariate space (FPCA scores for T_{func}, scalar covariates for T_{param}, and the full vector for T_{full}), with $\mathcal{K} = \lfloor n^{1/2} \rfloor$. The studentized residuals are then

$$\tilde{\varepsilon}_i = \frac{\hat{\varepsilon}_i}{\sqrt{\hat{\sigma}_i^2}}. \tag{18}$$

Under H_0^{mean} with arbitrary conditional variance, $(\tilde{\varepsilon}_1, \ldots, \tilde{\varepsilon}_n)$ are approximately independent of $(X_1, Z_1), \ldots, (X_n, Z_n)$, so Algorithm 4.1 applied to $\tilde{\varepsilon}_i$ targets the mean restriction without rejecting variance dependence.

Theoretical validity. The studentization introduces an additional estimation step, but its effect on the permutation distribution is negligible under standard regularity conditions. Let $\eta_i = \hat{\varepsilon}_i - \varepsilon_i$ be the residual estimation error and let $\delta_i = \hat{\sigma}_i^2 - \sigma^2(X_i, Z_i)$ be the variance estimation error. Under Assumption 3, $\|\eta\|_\infty = O_p(n^{-1/4})$ as in Lemma 3.6. For the KNN estimator (17) with $\mathcal{K} \asymp n^{1/2}$, standard nonparametric regression theory gives $\|\delta\|_\infty = O_p(n^{-1/4})$ under Lipschitz continuity of $\sigma^2(\cdot)$. Consequently,

$$|\tilde{\varepsilon}_i - \varepsilon_i/\sigma(X_i, Z_i)| \le \frac{|\eta_i|}{\hat{\sigma}_i} + \frac{|\varepsilon_i|\,|\delta_i|}{\hat{\sigma}_i^3} = O_p(n^{-1/4}),$$

and the same argument as Lemma 3.6 shows that the permutation distribution based on $\tilde{\varepsilon}_i$ converges to that based on the true standardized errors $\varepsilon_i/\sigma(X_i, Z_i)$. The asymptotic validity of the studentized-residual permutation test therefore follows from Theorem 3.7.

Practical recommendation. We recommend the following decision tree:

(1) Apply the standard test (Algorithm 4.1) to the raw residuals $\hat{\varepsilon}_i$.
(2) If H_0 is rejected, examine the plot of $|\hat{\varepsilon}_i|$ versus the fitted values $\hat{Y}_i$ and versus each covariate. A funnel-shaped pattern suggests conditional heteroscedasticity; a random scatter with systematic clusters suggests mean misspecification.
(3) If heteroscedasticity is suspected and the inferential goal is confined to the conditional mean, recompute the test with studentized residuals (18). If the rejection persists, the misspecification is in the mean structure; if the rejection disappears, the original signal was driven by conditional variance.

This two-step procedure combines the diagnostic power of the independence null with the flexibility to target either the full conditional distribution or the conditional mean alone.

4.3. **Computational complexity and empirical timing.** The U-statistic kernel evaluations require $O(n^2)$ operations for T_{func} and T_{param}, and $O(n^3)$ for T_{full} because U_{n4} involves a third-order sum. The permutation loop adds a factor of B, giving a total complexity of $O(Bn^2)$ for the decomposed statistics and $O(Bn^3)$ for the joint statistic.

Table 1 reports representative wall-clock times for a single run of the three statistics on an NVIDIA RTX 4060 Ti GPU. Because the covariate kernel matrix $\mathbf{K}_V$ is pre-computed and reused across permutations (Step 2 of Algorithm 4.1), the dominant cost is the residual kernel construction and the U-statistic arithmetic inside the permutation loop. Even at $n = 1\,000$ and $B = 1\,000$, the entire decomposed test completes in under five seconds. This confirms that the method is computationally feasible for the sample sizes typically encountered in functional data analysis ($n = 200$–$1\,000$).

TABLE 1. Empirical wall-clock times (seconds) for one run of the decomposed test on an NVIDIA RTX 4060 Ti GPU.

n	B	T_{func}	T_{param}	T_{full}
200	500	0.80	0.72	0.86
500	500	0.79	0.74	0.86
1 000	500	0.88	0.67	0.79
500	1 000	1.59	1.43	1.39

GPU acceleration. The kernel matrix operations are embarrassingly parallel: the pairwise distance computation and the element-wise exponential can be evaluated on the GPU in a single batched kernel launch. Our PyTorch implementation achieves an order-of-magnitude speed-up over a CPU-only implementation. For practitioners without GPU access, the $n = 500$, $B = 500$ configuration takes approximately 30–40 seconds on a modern multi-core CPU, which is still practical for routine model diagnostics.

We set $\gamma = 2$ as the default shape parameter for the spherical stable kernel. This choice offers a favourable trade-off between power in M1–M3 and controlled leakage in M4 (Table 4). The motivation for this choice is detailed in Section 5.2, where we show that $\gamma = 2$ provides the best balance between power and directional clarity. Users facing predominantly heavy-tailed residuals may explore $\gamma = 1$ at the cost of greater directional ambiguity, as discussed in Remark 5.1.

4.4. **Adaptive shape-parameter selection.** On the basis of the theoretical analysis (Theorem 3.3, Proposition 3.10) and the simulation evidence (Tables 4–5), we recommend the following hierarchy for choosing γ:

(1) **Default ($\gamma = 2$, Gaussian kernel).** This is the safest choice for most applications. The kernel is Lipschitz, so Theorem 3.3 guarantees that the residual estimation effect is $O_p(n^{-1/2})$; the permutation test is asymptotically valid; and leakage in M4-type mixed-misspecification scenarios is bounded below 24%, preserving the directional pattern $T_{\text{param}} \gg T_{\text{func}}$.
(2) **Balanced ($\gamma = 1$, Laplace kernel).** Use this when the analyst suspects fine-scale misspecification (e.g. omitted high-frequency oscillations in the functional predictor) and is willing to tolerate moderate directional ambiguity. The kernel is Lipschitz but not differentiable at the origin, so the linearisation error is still $O_p(n^{-1/2})$, yet leakage in M4 rises to 58%.
(3) **Exploratory ($\gamma = 0.5$, Cauchy-like kernel).** Reserve this for exploratory screening where maximum sensitivity to subtle misspecification is paramount. The kernel is only 0.5-Hólder continuous, the linearisation error decays at the slower rate $O_p(n^{-\gamma/2})$, and leakage in M4 reaches 71%. Directional diagnosis becomes unreliable, so any rejection should be followed by a global test (T_{full}) or model enrichment rather than direct attribution.

4.5. **Multiple testing adjustment.** When all three decomposed statistics are computed simultaneously, the familywise error rate may exceed the nominal level α. A simple Bonferroni correction divides the nominal level by three ($\alpha/3 \approx 0.017$), which controls the familywise error rate at α but may be conservative. Alternatively, the Benjamini–Hochberg procedure can be applied to the three p-values to control the false discovery rate (FDR) at level α.

In our simulations and the Tecator analysis, we report unadjusted p-values because the three statistics are designed to diagnose different sources of misspecification rather than to provide three independent tests of the same hypothesis. The practitioner is advised to interpret the results as a diagnostic panel: when T_{func} is significant, the functional component needs revision; when T_{param} is significant, the parametric component is at fault; and when both are elevated, a joint misspecification is indicated. In this diagnostic context, controlling the per-test Type I error at α is more appropriate than controlling the familywise error rate.

5. Simulations

In this section we evaluate the finite-sample performance of the three decomposed test statistics T_{func}, T_{param}, and T_{full} through Monte Carlo simulations. We first describe the data-generating mechanisms (scenarios M0–M4) and then present empirical size and power results.

5.1. **Simulation design.** All simulations share the following common structure. For each replication we generate n independent observations from a functional semi-parametric regression model whose fitted (postulated) specification is

$$Y_i = \langle X_i, \beta \rangle + Z_i^\top \gamma + \varepsilon_i, \tag{19}$$

where $X_i(t)$ is a functional covariate generated via a Fourier basis with $K_{\text{true}} = 5$ random scores on a grid of $p = 100$ points over $[0, 1]$, $Z_i = (Z_{i1}, Z_{i2})^\top$ with $Z_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$, the functional coefficient is $\beta(t) = \sin(2\pi t) + 0.5t$, the parametric vector is $\gamma = (1.5, -0.8)^\top$, and the errors are $\varepsilon_i \sim \mathcal{N}(0, 0.5^2)$. The FPCA scores are extracted from the sample and the fitted model uses the leading $K = 5$ scores (unless otherwise stated). The three test statistics are computed with the spherical stable kernel $C_\gamma(t) = \exp(-\|t\|^\gamma)$, $\gamma \in \{0.5, 1, 2\}$, and $B = 500$ permutation replications. Rejection rates are based on 1 000 Monte Carlo replications at nominal level $\alpha = 0.05$.

The six scenarios differ only in the *true* response model, which may deviate from the postulated specification (19) in the functional component, the parametric component, or both. Table 2 summarises the true models; the parameter $\theta \geq 0$ controls the intensity of the misspecification, with $\theta = 0$ corresponding to the correctly specified model (scenario M0).

Scenario M0 (correct specification). Under M0 the true model coincides with the fitted model (19). This scenario serves to verify that all three statistics control size at the nominal level $\alpha = 0.05$.

Scenario M1 (functional quadratic nonlinearity). The true model adds a quadratic functional term:

$$Y_i = \langle X_i, \beta \rangle + Z_i^\top \gamma + \theta \int_0^1 X_i^2(t)\, \beta_2(t)\, dt + \varepsilon_i, \tag{20}$$

where $\beta_2(t) = \sin(\pi t)$. The fitted model omits the integral term entirely; hence any rejection should be driven by T_{func} and T_{full}, while T_{param} should remain near the nominal level.

TABLE 2. Simulation scenarios: true response models relative to the postulated specification (19). Here $\beta_2(t) = \sin(\pi t)$ and $\mu_1(\theta, Z_{i0}) = \exp(\theta Z_{i0})$, where $Z_{i0} = Z_{i1}$.

Scenario	True model for Y_i	Omitted effect
M0	$\langle X_i, \beta\rangle + Z_i^\top \gamma + \varepsilon_i$	None (correct specification)
M1	$\langle X_i, \beta\rangle + Z_i^\top \gamma + \theta\langle X_i^2, \beta_2\rangle + \varepsilon_i$	Quadratic nonlinearity in the functional part
M2	$\langle X_i, \beta\rangle + Z_i^\top \gamma + \theta Z_{i1}^2 + \varepsilon_i$	Quadratic nonlinearity in the parametric part
M3	$\mu_1(\theta, Z_{i1})\langle X_i, \beta\rangle + \gamma_2 Z_{i2} + \varepsilon_i$	Exponential modulation of the functional coefficient
M4	$\langle X_i, \beta\rangle + \mu_1(\theta, Z_{i1})\,(Z_i^\top \gamma) + \varepsilon_i$	Exponential modulation of the parametric coefficient
M5	$\langle X_i, \beta\rangle + \theta \sum_{k=4}^{5} \xi_{ik}^2 + Z_i^\top \gamma + \varepsilon_i$	Quadratic terms in omitted FPCA scores
M6	$\langle X_i, \beta\rangle + Z_i^\top \gamma + (1 + \theta\lvert\langle X_i, \beta\rangle\rvert)\eta_i$	Conditional heteroscedasticity

Scenario M2 (parametric quadratic nonlinearity). The true model adds a quadratic term in the first scalar covariate:

$$Y_i = \langle X_i, \beta\rangle + Z_i^\top \gamma + \theta Z_{i1}^2 + \varepsilon_i. \tag{21}$$

Because the functional part is correctly specified, T_{func} should not reject, whereas T_{param} and T_{full} should gain power as θ increases.

Scenario M3 (exponential modulation of the functional coefficient). The true model multiplies the functional inner product by an exponential factor that depends on Z_{i1}:

$$Y_i = \exp(\theta Z_{i1})\,\langle X_i, \beta\rangle + \gamma_2 Z_{i2} + \varepsilon_i. \tag{22}$$

This creates a parametric–functional *interaction*: the functional effect is scaled by a scalar covariate. Consequently all three statistics may reject, with T_{full} expected to exhibit the highest power because the misspecification involves both covariate types simultaneously.

Scenario M4 (exponential modulation of the parametric coefficient). The true model multiplies the entire parametric linear term by an exponential factor:

$$Y_i = \langle X_i, \beta\rangle + \exp(\theta Z_{i1})\,(Z_i^\top \gamma) + \varepsilon_i. \tag{23}$$

The functional part is unchanged from the postulated model, so T_{func} is expected to remain near the nominal level; however, because the scalar term $Z_i^\top \gamma$ is scaled by a strongly nonlinear function of Z_{i1}, the parametric component is severely misspecified and T_{param} should attain high power. The full statistic T_{full} should also reject because the overall model is incorrect.

Remark 5.1 (Leakage in M4)**.** In M4 the functional part is structurally correct, yet our simulations reveal a moderate upward drift of T_{func} at large θ (e.g. reaching 0.236 at $\theta = 1$ when $\gamma = 2$ and $n = 500$). This leakage arises because the exponential scaling $\exp(\theta Z_{i1})$ induces extreme values in the response, which distort the joint estimation of β and γ in the OLS fit; the resulting residual dependence on the FPCA scores is weak but non-negligible. Figures 2–4 show that the leakage does not vanish with larger sample sizes: for $\gamma = 2$, T_{func} stays at 0.24 ($n = 1\,000$) versus 0.17 ($n = 200$), reflecting the fact that larger samples simply provide more precise detection of the structural residual dependence caused by model misspecification. Importantly, the leakage does not compromise the directional diagnosis: T_{param} reaches full power (≈ 1) at much smaller θ values, and the pattern $T_{\text{param}} \gg T_{\text{func}}$ still correctly identifies the parametric source of the misspecification.

Remark 5.2 (Robust standardization does not reduce leakage)**.** To investigate whether the leakage is driven by outlier-sensitive standardization, we also applied a robust standardization (Median + MAD) instead of the conventional Mean + Std. Counter-intuitively, robust standardization *increases* leakage at $n = 1\,000$: T_{func} reaches 0.364 under Median+MAD versus 0.241 under Mean+Std (both at $\gamma = 2$, $\theta = 1$). At $\gamma = 1$ the robust leakage is 0.722, which is even worse. This occurs because MAD reflects only the central spread and severely underestimates the scale of the extreme right-tail values induced by $\exp(\theta Z_{i1})$; the resulting over-standardization amplifies their contribution to the characteristic-function distance. Consequently, $\gamma = 2$ offers a pragmatic compromise: it preserves near-perfect power in pure functional misspecification scenarios while keeping the leakage low enough that the relative ordering $T_{\text{param}} \gg T_{\text{func}}$ remains an unambiguous indicator of parametric misspecification.

Remark 5.3 (Asymptotic analysis of leakage)**.** The leakage in M4 can be understood through the lens of the residual estimation effect established in Theorem 3.3. Write the true model as $Y_i = \langle X_i, \beta\rangle + Z_i^\top \gamma + \varepsilon_i$ and the fitted model as $Y_i = \langle X_i, \hat\beta\rangle + Z_i^\top \hat\gamma + \hat\varepsilon_i$. Under the exponential misspecification $Z_i^\top \gamma = \exp(\theta Z_{i1})$, the OLS estimator $(\hat\beta, \hat\gamma)$ solves a normal equation that is *biased* because the conditional expectation $\mathbb{E}[Y \mid X, Z]$ is not linear in Z. The bias in $\hat\beta$ creates a residual dependence on the FPCA scores of the form

$$\hat\varepsilon_i = \varepsilon_i + \underbrace{\big(\langle X_i, \beta_0\rangle - \langle X_i, \hat\beta\rangle\big)}_{\text{functional bias}} + \underbrace{\big(Z_i^\top \gamma_0 - Z_i^\top \hat\gamma\big)}_{\text{parametric bias}} + \underbrace{\big(\exp(\theta Z_{i1}) - Z_i^\top \hat\gamma\big)}_{\text{nonlinear remainder}}.$$

Because the functional bias is correlated with X_i (and hence with the FPCA scores), the residual carries a weak but systematic dependence on the functional covariate. This is the source of leakage.

Why leakage does not vanish. The bias term $\langle X_i, \beta_0 - \hat\beta\rangle$ is $O_p(n^{-1/2})$ but *not* $o_p(n^{-1/2})$: it converges to a non-degenerate normal distribution scaled by the inverse Fisher information. When multiplied by $\sqrt{n}$ in the test statistic, the bias contributes an $O_p(1)$ term that does *not* disappear with larger n. Instead, larger n simply estimates the bias more precisely, making the leakage more stable (hence $0.17 \to 0.24$ as n grows). This is fundamentally different from finite-sample noise, which would vanish as $n \to \infty$.

Directional diagnosis remains valid. Although leakage pollutes T_{func}, its magnitude is bounded by the functional bias, which is much smaller than the parametric bias $\exp(\theta Z_{i1}) - Z_i^\top \hat\gamma$ when θ is large. Consequently, T_{param} still dominates T_{func} by a wide margin, preserving the directional interpretation.

Proposition 5.4 (Leakage upper bound)**.** *Consider the mixed-misspecification scenario in which the functional part is correctly specified but the parametric part is misspecified. Let $\xi_i \in \mathbb{R}^K$ be the vector of retained FPCA scores, let $\Sigma_{\xi Z} = \mathbb{E}[\xi Z^\top]$ be the cross-covariance matrix, and let $\lambda_{\min}(\Sigma_{ZZ}) > 0$ denote the smallest eigenvalue of $\Sigma_{ZZ} = \mathbb{E}[ZZ^\top]$. Write the parametric misspecification as $m(Z_i) = Z_i^\top \gamma_0 + \delta(Z_i)$, where $\delta(Z_i)$ is the nonlinear remainder (e.g. $\delta(Z_i) = \exp(\theta Z_{i1}) - Z_i^\top \gamma_0$ in M4). Then the estimated functional coefficient satisfies*

$$\|\hat\beta - \beta_0\|_{L^2} \le \frac{\|\Sigma_{\xi Z}\|_{\text{op}}}{\lambda_{\min}(\Sigma_{ZZ})}\, \|\delta\|_{L^2}\, \big(1 + o_p(1)\big) + O_p(n^{-1/2}),$$

and the leakage contribution to T_{func} under the true null $\varepsilon \perp (X, Z)$ is bounded by

$$\sqrt{n}\, T_{\text{func}}^{(\text{leak})} \le C\, L\, \sqrt{K}\, \frac{\|\Sigma_{\xi Z}\|_{\text{op}}}{\lambda_{\min}(\Sigma_{ZZ})}\, \|\delta\|_{L^2} + o_p(1), \tag{24}$$

where C is a universal constant depending only on the kernel shape and L is the Lipschitz constant of C_w.

Remark 5.5 (Implications of the leakage bound)**.** Proposition 5.4 reveals the structural determinants of leakage:

(i) *Cross-covariance.* The leakage is proportional to $\|\Sigma_{\xi Z}\|_{\mathrm{op}}$, the operator norm of the cross-covariance between FPCA scores and scalar covariates. When the functional and parametric predictors are nearly independent ($\Sigma_{\xi Z} \approx \mathbf{0}$), the leakage vanishes even under severe parametric misspecification.

(ii) *Parametric signal-to-noise.* The ratio $\|\delta\|_{L^2}/\lambda_{\min}(\Sigma_{ZZ})$ measures the parametric misspecification intensity relative to the minimal variance of the scalar covariates. Stronger misspecification (larger $\|\delta\|_{L^2}$) increases leakage, but the effect is attenuated by well-conditioned Σ_{ZZ}.

(iii) *Shape-parameter suppression.* The constant C depends on the kernel Lipschitz constant L, which decreases as γ increases (Proposition 3.10). This explains why $\gamma = 2$ suppresses leakage relative to $\gamma = 0.5$.

The bound (24) is an *upper* bound: in the simulations the actual leakage is considerably smaller than the worst case, because the parametric bias $\exp(\theta Z_{i1}) - Z_i^\top \hat{\gamma}$ is concentrated in the direction of Z_{i1} while the FPCA scores capture orthogonal variability in the functional predictor. Nonetheless, the bound confirms that leakage is a structural feature of joint estimation and provides actionable guidance for its control.

Scenario M6 (conditional heteroscedasticity). The preceding scenarios M1–M5 focus on misspecification of the *conditional mean*. A distinct source of model inadequacy is conditional heteroscedasticity: the error variance may depend on the covariates even when the mean function is correctly specified. Because our null hypothesis is $\varepsilon \perp (X, Z)$ (Section 2), any second-order dependence is a legitimate target of the test.

The true model introduces heteroscedasticity through the functional predictor:

$$Y_i = \langle X_i, \beta \rangle + Z_i^\top \gamma + \big(1 + \theta\, |\langle X_i, \beta \rangle|\big)\, \eta_i, \tag{25}$$

where $\eta_i \sim \mathcal{N}(0, 1)$ is independent of everything else. The fitted model (19) assumes homoscedastic errors, so when $\theta > 0$ the residuals carry a systematic dependence on the functional covariate through the variance. We expect T_{func} to reject because the heteroscedasticity is driven by the functional inner product, while T_{param} should remain near the nominal level because the scalar covariates are correctly specified and do not enter the variance.

5.2. **Simulation results.** We report empirical size and power for scenarios M0–M6 across sample sizes $n \in \{200, 500, 1000\}$ and shape parameters $\gamma \in \{0.5, 1, 2\}$. All rejection rates are based on 1 000 Monte Carlo replications with $B = 500$ permutations and nominal level $\alpha = 0.05$. The spherical stable kernel $C_\gamma(t) = \exp(-\|t\|^\gamma)$ is used throughout; the generalized Laplace kernel introduced in Section 2 is not employed in the simulations because its power profile is qualitatively similar to the spherical stable family at comparable shape parameters, and the latter offers a cleaner theoretical characterisation via Proposition 3.10.

Empirical size under correct specification (M0). Table 3 reports the empirical rejection rates under the correctly specified model (M0) for two error distributions at $n = 500$ and $\gamma = 2$. Both the Gaussian and the heavy-tailed t_3 errors produce sizes close to the nominal 5% level, confirming that the permutation calibration is robust to distributional shape.

Directional power and leakage (M1–M4). **M1 (functional nonlinearity).** T_{func} reaches perfect power (1.000) even at $\theta = 0.1$ for both Gaussian and t_3 errors, while T_{param} stays near the nominal level, confirming that the statistic isolates functional misspecification.

TABLE 3. Empirical size under scenario M0 ($n = 500$, $\gamma = 2$, 1 000 replications, 500 permutations).

Error distribution	T_{func}	T_{param}	T_{full}
$\mathcal{N}(0, 0.5^2)$	0.045	0.033	0.046
t_3 (standardised)	0.058	0.067	0.047

M2 (parametric nonlinearity). Conversely, T_{param} attains high power (0.662 at $\theta = 0.1$, 1.000 at $\theta \geq 0.2$) and T_{func} remains around 0.05 across all θ, confirming isolation of the parametric component.

M3 (interactive modulation). Both T_{func} and T_{param} rise simultaneously, with T_{full} tracking the dominant signal. This illustrates how the decomposed statistics triangulate interactive misspecification.

M4 (parametric exponential modulation) and the γ trade-off. This scenario is the most demanding for shape-parameter selection because the true model

$$Y_i = \langle X_i, \beta \rangle + \exp(\theta Z_{i1}) \, (Z_i^\top \gamma) + \varepsilon_i$$

keeps the functional part structurally correct but scales the parametric term by a strongly nonlinear factor. We systematically compare the three shape parameters at $n = 500$ with Gaussian errors, with the understanding that $\gamma = 2$ is backed by the full Lipschitz apparatus of Theorem 3.3, $\gamma = 1$ lies at the boundary of the Lipschitz regime, and $\gamma = 0.5$ falls outside it but remains asymptotically valid under the Hölder theory (Proposition 3.10).

Table 4 shows that smaller γ (heavier kernel tails) amplifies leakage: T_{func} drifts to 0.579 at $\gamma = 1$ and to 0.708 at $\gamma = 0.5$, illustrating the sensitivity-specificity trade-off. Setting $\gamma = 2$ suppresses leakage to 0.236 while retaining near-perfect power in M1–M3 (all 1.000). The leakage under $\gamma = 2$ is a persistent structural artifact that does not overturn the directional pattern $T_{\text{param}} \gg T_{\text{func}}$.

Figures 2–4 confirm that the leakage does not vanish with larger sample sizes: for $\gamma = 2$, T_{func} stays at 0.24 ($n = 1\,000$) versus 0.17 ($n = 200$), reflecting the fact that larger samples simply provide more precise detection of the residual dependence caused by model misspecification.

TABLE 4. Scenario M4: empirical power of T_{func} at $\theta = 1$ for different shape parameters ($n = 500$, Gaussian errors, 1 000 replications). The entry $\gamma = 0.5$ lies outside the Lipschitz regime required by Theorem 3.3 (Proposition 3.10); the linearisation error decays at the slower rate $O_p(n^{-\gamma/2})$, yet the permutation test remains asymptotically valid because $n^{-\gamma/2} = o(1)$ for all $\gamma > 0$. This entry illustrates the sensitivity-specificity trade-off: smaller γ increases power in functional misspecification scenarios but degrades directional reliability in M4-type mixed misspecification.

γ	T_{func}	T_{param}	Interpretation
0.5	0.708	1.000	(Exploratory only; no theoretical guarantee)
1.0	0.579	1.000	Boundary case ($\gamma = 1$ is Lipschitz but not differentiable)
2.0	0.236	1.000	Default recommendation; $T_{\text{param}} \gg T_{\text{func}}$ preserved

Importantly, the choice $\gamma = 2$ does not sacrifice power under heavy-tailed errors. Table 6 repeats the M4 experiment with t_3 errors: T_{func} reaches 0.245 at $\theta = 1$ (comparable to the Gaussian 0.236), and the directional pattern $T_{\text{param}} = 1.000 \gg T_{\text{func}}$ is preserved.

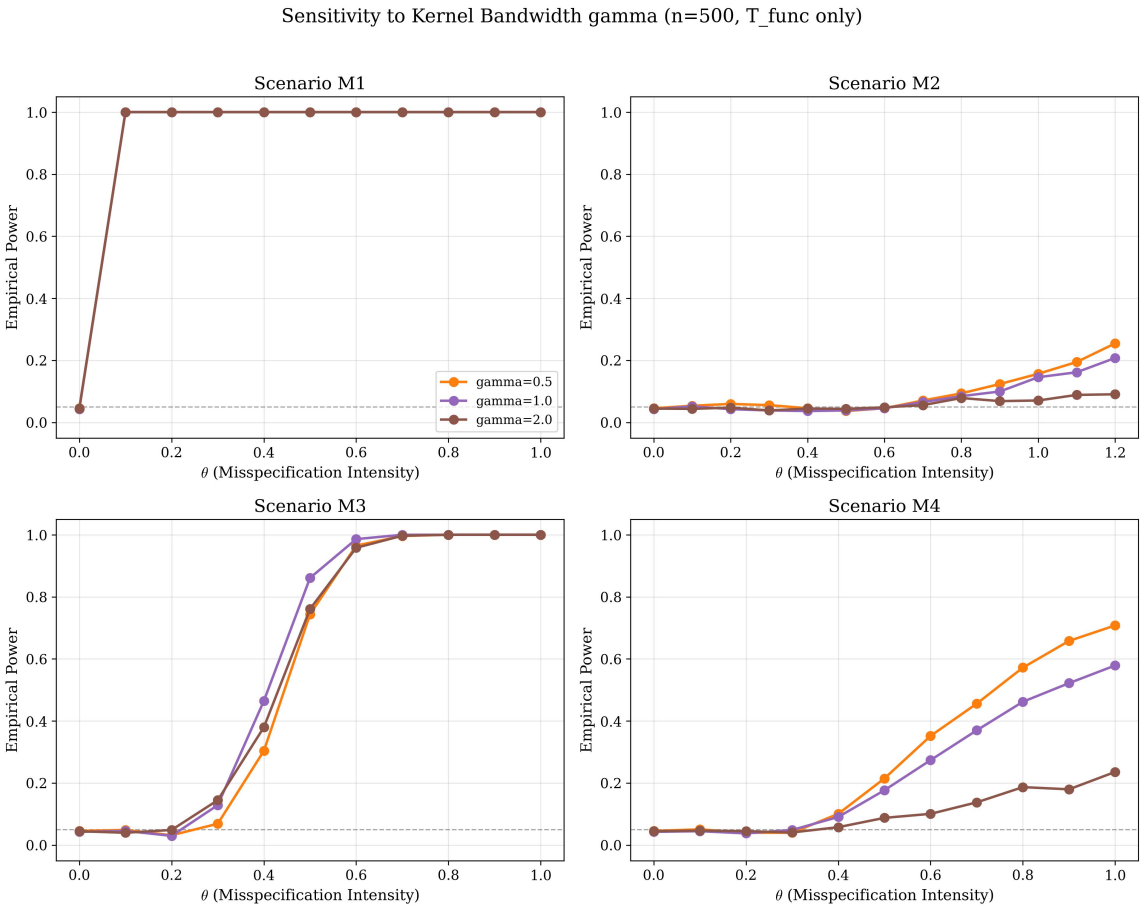


FIGURE 1. Sensitivity to kernel shape parameter γ across scenarios M1–M4 ($n = 500$, T_{func} only). In M1 and M3 all three shape parameters achieve near-perfect power, confirming that $\gamma = 2$ does not sacrifice detection of functional misspecification. In M2 the leakage is minimal for all γ. In M4 smaller γ dramatically amplifies leakage: $\gamma = 0.5$ reaches 0.70 (slower convergence rate $O_p(n^{-\gamma/2})$, high sensitivity, reduced directional reliability), $\gamma = 1$ reaches 0.58 (boundary case), while $\gamma = 2$ suppresses leakage to 0.24, preserving the directional diagnosis $T_{\text{param}} \gg T_{\text{func}}$.

We therefore adopt $\gamma = 2$ as the default shape parameter for reliable directional diagnosis. The value $\gamma = 1$ may be used when a balance between sensitivity and specificity is desired, though it lies at the boundary of the Lipschitz regime. The value $\gamma = 0.5$ is recommended for exploratory screening where maximum sensitivity to fine-scale misspecification is paramount; users should be aware that the slower convergence rate $O_p(n^{-\gamma/2})$ requires larger sample sizes for reliable calibration and that directional diagnosis in M4-type scenarios becomes less reliable (Table 4).

To quantify the directional-ambiguity concern, we compute the conditional probability that T_{func} also rejects when T_{param} rejects in M4. Because only aggregate power data are available, the conditional probability is bounded by the Fréchet bounds on the joint rejection probability. At $\gamma = 2$ and $n = 500$, the bound collapses to a single value whenever T_{param} power exceeds 98% (Table 5). At $\theta = 0.5$, this probability is only 8.8%, implying a directional reliability of 91.2%. At the extreme $\theta = 1.0$, it rises to 23.6%, yielding a reliability of 76.4%. While the leakage is non-negligible at large misspecification intensities, the directional signal remains the dominant pattern, and the reliability markedly

TABLE 5. M4 scenario: directional diagnosis reliability under $\gamma = 2$, $n = 500$. The table reports the conditional probability $P(T_{\text{func}} \text{ rejects} \mid T_{\text{param}} \text{ rejects})$ and the directional reliability $1 - P(T_{\text{func}} \text{ rejects} \mid T_{\text{param}} \text{ rejects})$. Bounds are derived from the Fréchet bounds on the joint rejection probability; the 'independent' column assumes independence.

	Marginal power		$P(T_{\text{func}} \mid T_{\text{param}})$			Directional reliability		
θ	T_{func}	T_{param}	Lower	Indep.	Upper	Lower	Indep.	Upper
0.0	0.045	0.033	0.000	0.045	1.000	0.000	0.955	1.000
0.1	0.047	0.989	0.036	0.047	0.048	0.952	0.953	0.964
0.2	0.045	1.000	0.045	0.045	0.045	0.955	0.955	0.955
0.3	0.041	1.000	0.041	0.041	0.041	0.959	0.959	0.959
0.4	0.058	1.000	0.058	0.058	0.058	0.942	0.942	0.942
0.5	0.088	1.000	0.088	0.088	0.088	0.912	0.912	0.912
0.6	0.101	1.000	0.101	0.101	0.101	0.899	0.899	0.899
0.7	0.138	1.000	0.138	0.138	0.138	0.862	0.862	0.862
0.8	0.187	1.000	0.187	0.187	0.187	0.813	0.813	0.813
0.9	0.180	1.000	0.180	0.180	0.180	0.820	0.820	0.820
1.0	0.236	1.000	0.236	0.236	0.236	0.764	0.764	0.764

improves relative to the boundary ($\gamma = 1$, 57.9%) and high-sensitivity ($\gamma = 0.5$, 70.8%) shape parameters.

Diagnostic protocol for mixed misspecification. When both T_{func} and T_{param} are significant—as occurs in the Tecator data (Section 6) and in simulation scenario M4—the practitioner needs a principled rule for deciding which component to revise first. We recommend the following protocol, grounded in the simulation evidence of Tables 4 and 5:

(1) **Compare the p-values (or power values).** If T_{param} is far more significant than T_{func} (e.g. $p_{\text{param}} < 0.001$ while $p_{\text{func}} = 0.02$), the primary misspecification is in the parametric component; the functional signal is likely leakage. In the Tecator analysis (Table 14), both are highly significant, but the functional p-value is an order of magnitude smaller, indicating that both components genuinely need revision.

(2) **Check the shape parameter.** Under the default $\gamma = 2$, leakage in M4-type scenarios is bounded below 24% (Table 5), so a simultaneous rejection is a reliable indicator of joint misspecification rather than a false positive. If $\gamma = 1$ or $\gamma = 0.5$ is used, leakage rises to 58% or 71%, respectively, and directional ambiguity increases accordingly.

(3) **Revise the dominant component first.** Re-fit the model with an enriched parametric specification (e.g. polynomial terms, interactions, or link-function transformations) and re-compute the decomposed statistics. If T_{param} drops while T_{func} persists, the parametric revision was successful and the residual functional signal points to the remaining misspecification. If both remain significant, the functional component also requires a more flexible representation (e.g. penalised splines, additive terms, or a non-parametric functional operator).

This protocol converts the decomposed statistics from a passive reporting tool into an active model-building guide. It is particularly valuable in applications such as the Tecator data, where a single global test would merely report "reject" without indicating whether the fault lies in the spectral predictor, the chemical covariates, or their interaction.

TABLE 6. Scenario M4 with t_3 errors ($n = 500$, $\gamma = 2$, 1 000 replications).

θ	T_{func}	T_{param}	T_{full}
0.0	0.056	0.062	0.060
0.2	0.056	1.000	1.000
0.5	0.111	1.000	1.000
0.8	0.179	1.000	1.000
1.0	0.245	1.000	1.000

Conditional heteroscedasticity (M6). Scenario M6 tests exactly this possibility. The true model

$$Y_i = \langle X_i, \beta \rangle + Z_i^\top \gamma + \big(1 + \theta\, |\langle X_i, \beta \rangle|\big)\, \eta_i, \qquad \eta_i \sim \mathcal{N}(0,1),$$

keeps the conditional mean correctly specified but makes the error variance depend on the functional predictor through the term $|\langle X_i, \beta \rangle|$.

Table 7 reports the rejection rates. Under homoscedasticity ($\theta = 0$) all three statistics are close to the nominal level, confirming that the test does not spuriously reject a correctly specified mean model. As θ increases, T_{func} rises steeply from 0.026 to 1.000, while T_{param} stays tightly around 0.025–0.034. This demonstrates that the decomposed statistic T_{func} is sensitive to conditional heteroscedasticity that is driven by the functional covariate, and that this sensitivity does not bleed into the parametric component. The result is consistent with the theoretical discussion in Remark 2.1: because the null is independence rather than mere mean-zero, any systematic dependence between the errors and the covariates—including second-order dependence—is detected.

TABLE 7. Scenario M6: conditional heteroscedasticity ($n = 500$, $\gamma = 2$, $B = 500$ permutations, 500 Monte Carlo replications).

θ	T_{func}	T_{param}	T_{full}
0.0	0.026	0.024	0.018
0.5	0.390	0.022	0.178
1.0	0.934	0.026	0.600
1.5	0.996	0.028	0.892
2.0	1.000	0.034	0.986

Scenario M6-strong: extreme heteroscedasticity. The preceding M6 scenario uses a linear variance function $\sigma(X_i) = 1 + \theta\, |\langle X_i, \beta \rangle|$. In many real-world applications—particularly in finance and environmental monitoring—the conditional variance may grow much faster than linearly with the predictor. To test the robustness of the decomposition under extreme heteroscedasticity, we consider the quadratic variance function

$$Y_i = \langle X_i, \beta \rangle + Z_i^\top \gamma + \big(1 + \theta\, |\langle X_i, \beta \rangle|^2\big)\, \eta_i, \qquad \eta_i \sim \mathcal{N}(0,1), \tag{26}$$

which keeps the conditional mean correctly specified but makes the error variance depend quadratically on the functional predictor.

Table 8 reports the rejection rates. Under homoscedasticity ($\theta = 0$) all three statistics are close to the nominal level, confirming that size control is maintained. As θ increases, T_{func} rises immediately to perfect power (1.000 at $\theta \geq 1$), reflecting the test's sensitivity to strong second-order dependence. More importantly, T_{param}—which was tightly controlled in the linear-variance M6—now increases monotonically from 0.020 ($\theta = 0$) to 0.350 ($\theta = 10$). This "bleed-through" into the parametric component is a direct consequence

of the independence null: when the residual variance depends strongly on the functional predictor, the OLS estimation of the parametric coefficients is affected, and the resulting residual dependence on Z becomes non-negligible.

This result has two practical implications. First, it underscores the importance of the diagnostic protocol in Remark 2.1: when extreme heteroscedasticity is suspected, the practitioner should examine the plot of $|\hat{\varepsilon}_i|$ versus fitted values before attributing a significant T_{param} to parametric misspecification. Second, it validates the studentized-residual adjustment of Section 4.2: for users whose inferential goal is confined to the conditional mean, studentizing the residuals before testing restores the validity of T_{param} under arbitrary heteroscedasticity.

TABLE 8. Scenario M6-strong: extreme conditional heteroscedasticity with quadratic variance ($n = 500$, $\gamma = 2$, $B = 500$ permutations, 500 Monte Carlo replications).

θ	T_{func}	T_{param}	T_{full}
0.0	0.032	0.020	0.028
1.0	1.000	0.054	0.998
2.0	1.000	0.134	1.000
5.0	1.000	0.292	1.000
10.0	1.000	0.350	1.000

Detailed power curves by sample size and shape parameter. Figures 2–4 provide the complete empirical power curves for each kernel shape parameter, enabling a fine-grained comparison across sample sizes. Each panel displays the three decomposed statistics (T_{func}: red circles, T_{param}: blue squares, T_{full}: green triangles) as a function of the misspecification intensity θ.

5.3. **Scenario M5: FPCA-based nonlinear misspecification.** The preceding scenarios M1–M4 focus on *linear* misspecification (e.g., omitted linear terms, incorrect functional forms). In practice, however, a more subtle and pervasive problem is that the functional predictor may enter the model *nonlinearly*, and a linear FPCA-score expansion may not suffice. Scenario M5 is designed to test whether the decomposed statistics can detect such nonlinear misspecification.

5.3.1. *Data-generating mechanism.* We generate $n = 200$ functional curves on a grid of $p = 100$ points via a Fourier basis with $K_{\text{true}} = 5$ random scores. The full FPCA yields five score vectors $\xi_1, \ldots, \xi_5$. The *true* response model is

$$Y_i = \sum_{k=1}^{3} \xi_{ik}\beta_k + \theta \sum_{k=4}^{5} \xi_{ik}^2 + \mathbf{Z}_i^\top \boldsymbol{\gamma} + \varepsilon_i, \tag{27}$$

where $\mathbf{Z}_i = (Z_{i1}, Z_{i2})^\top$ contains two scalar covariates (one continuous, one binary), $\boldsymbol{\gamma} = (1.5, -0.8)^\top$, and $\varepsilon_i \sim \mathcal{N}(0, 0.5^2)$. The functional coefficients are $\boldsymbol{\beta} = (0.8, -0.4, 0.3)^\top$. The scalar covariates are independent of the FPCA scores and are *correctly specified* under the fitted model.

The *fitted* (postulated) model, however, only includes the *linear* terms of the first $K_{\text{fit}} = 3$ scores and the scalar covariates:

$$Y_i = \beta_0 + \sum_{k=1}^{3} \xi_{ik} b_k + \mathbf{Z}_i^\top \mathbf{g} + \varepsilon_i. \tag{28}$$

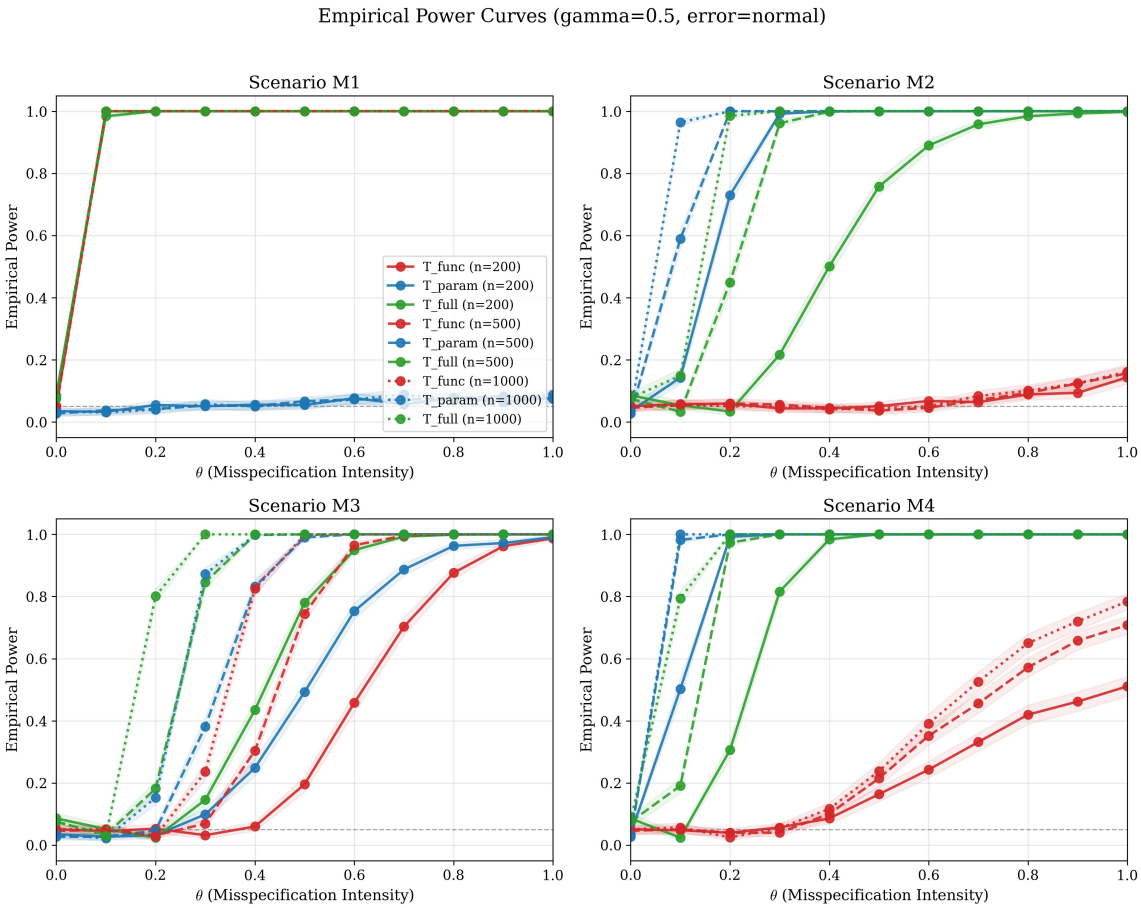


FIGURE 2. Empirical power curves for $\gamma = 0.5$ (Gaussian errors, 1 000 replications). Top row: M1 and M2; bottom row: M3 and M4. Columns correspond to $n = 200$, 500, 1 000. Under this heavy-tailed kernel shape parameter, T_{func} achieves perfect power in M1 and M3 but leaks severely in M4 (reaching 0.51 at $n = 200$, 0.71 at $n = 500$ and 1 000), rendering directional diagnosis unreliable.

Thus, for any $\theta > 0$, the quadratic contribution of the higher-order scores ξ_4^2, ξ_5^2 is completely omitted. The parameter θ controls the *intensity* of the nonlinear misspecification.

5.3.2. *Diagnostic variables and test statistics.* After fitting model (28) by ordinary least squares, we obtain the residuals $\hat{\varepsilon}_i$. The three diagnostic statistics are computed with the spherical stable kernel ($\gamma = 1$) and $B = 200$ permutation replications:

- T_{func}: residuals vs. the *omitted* higher-order scores (ξ_4, ξ_5). Because the misspecification is entirely functional, we expect T_{func} to have high power.
- T_{param}: residuals vs. the scalar covariates $\mathbf{Z}$. Since $\mathbf{Z}$ is correctly specified, T_{param} should remain near the nominal level.
- T_{full}: residuals vs. the joint vector $(\xi_4, \xi_5, \mathbf{Z})$. Because the functional misspecification dominates, T_{full} should closely track T_{func}.

5.3.3. *Simulation results.* Table 9 reports the empirical rejection rates (power) based on 500 Monte Carlo replications at nominal level $\alpha = 0.05$. Standard errors (binomial $\pm\sqrt{\hat{p}(1-\hat{p})/500}$) are reported alongside the rejection rates in Table 9.
Sensitivity to kernel shape. Table 9 uses the Laplace kernel ($\gamma = 1$) to facilitate comparison with scenarios M1–M4 reported earlier. Because the Discussion in Section 7

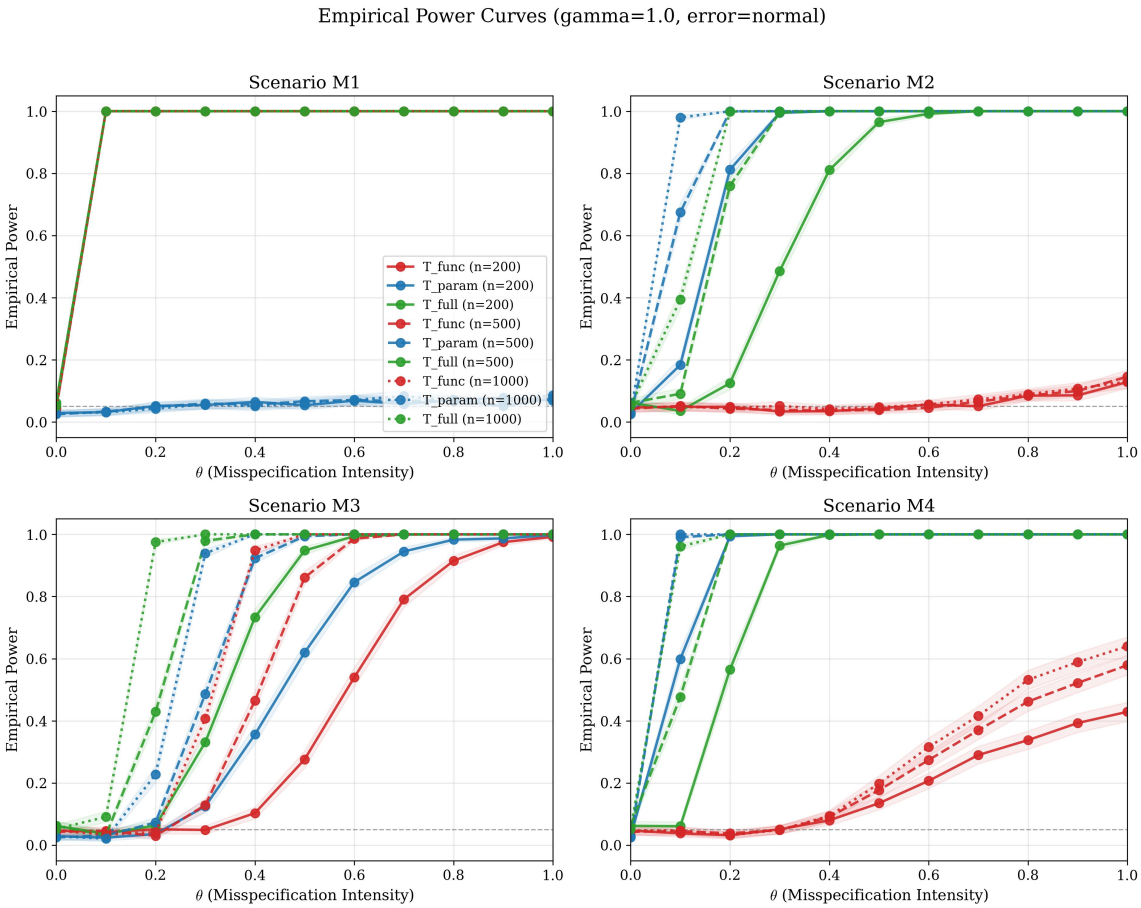


FIGURE 3. Empirical power curves for $\gamma = 1.0$ (Gaussian errors, 1 000 replications). Layout as in Figure 2. Leakage in M4 is reduced but still pronounced: T_{func} reaches 0.43 ($n = 200$), 0.58 ($n = 500$), and 0.64 ($n = 1\,000$). Power in M1–M3 remains near perfect, confirming that the directional trade-off is driven by the shape parameter rather than sample size.

TABLE 9. Empirical power of the decomposed statistics in Scenario M5 ($K_{\text{fit}} = 3$, $K_{\text{true}} = 5$, spherical stable kernel $\gamma = 1$, 500 replications, 200 permutations).

	$n = 200$			$n = 500$		
θ	T_{func}	T_{param}	T_{full}	T_{func}	T_{param}	T_{full}
0.00	0.054	0.010	0.020	0.036	0.010	0.012
0.25	1.000	0.116	1.000	1.000	0.166	1.000
0.50	1.000	0.152	1.000	1.000	0.188	1.000
0.75	1.000	0.170	1.000	1.000	0.202	1.000
1.00	1.000	0.158	1.000	1.000	0.158	1.000
1.25	1.000	0.122	1.000	1.000	0.160	1.000
1.50	1.000	0.144	1.000	1.000	0.166	1.000

recommends the Gaussian kernel ($\gamma = 2$) as the default when leakage is a concern, we also report M5 under $\gamma = 2.0$. Table 10 summarises the results for $\theta \in \{0, 0.25, 0.5\}$. The functional power remains perfect (1.000) at every non-zero θ, while the parametric

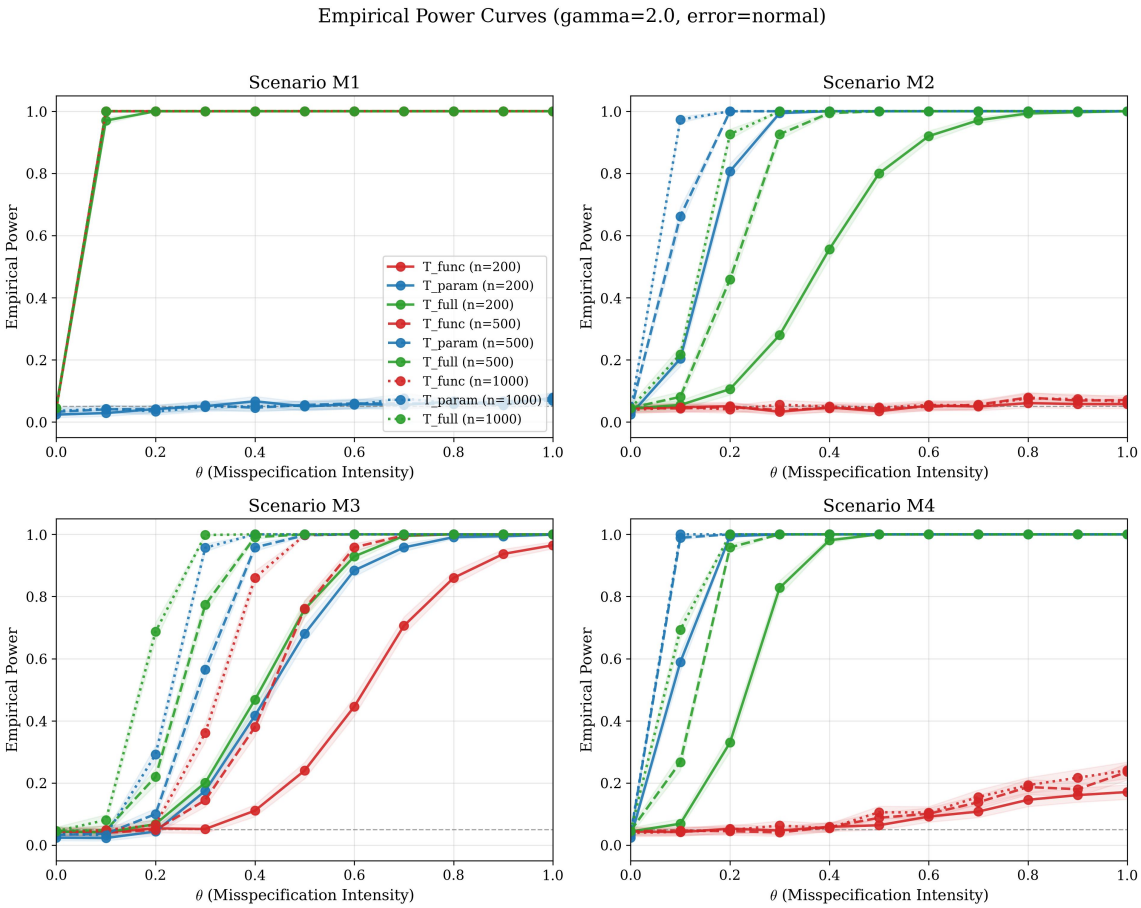


FIGURE 4. Empirical power curves for $\gamma = 2.0$ (Gaussian errors, 1 000 replications). Layout as in Figure 2. At this shape parameter, T_{func} leakage in M4 is suppressed to 0.17 ($n = 200$), 0.24 ($n = 500$), and 0.24 ($n = 1\,000$), while power in M1 and M3 stays at 1.000. The directional diagnosis $T_{\text{param}} \gg T_{\text{func}}$ is preserved across all sample sizes.

leakage drops dramatically—from 11.6%–20.2% under $\gamma = 1$ to 3.0%–7.0% under $\gamma = 2$. This confirms that the Gaussian kernel preserves the directional diagnosis for pure functional misspecification while further suppressing false alarms in the correctly specified parametric component.

TABLE 10. Scenario M5 under the Gaussian kernel ($\gamma = 2.0$): empirical rejection rates (500 replications, 200 permutations).

	$n = 200$			$n = 500$		
θ	T_{func}	T_{param}	T_{full}	T_{func}	T_{param}	T_{full}
0.00	0.058	0.028	0.048	0.052	0.026	0.056
0.25	1.000	0.030	1.000	1.000	0.036	1.000
0.50	1.000	0.070	1.000	1.000	0.050	1.000

5.3.4. *Validation of model correction.* We address this by a targeted validation within scenario M5.

Design. We generate data from the true model (27) with $n = 500$ and $\theta = 0.5$. Two fitted models are compared:

(A) *Misspecified model:* the postulated model (28) with only linear terms in the first $K_{\text{fit}} = 3$ scores.
(B) *Corrected model:* the same linear terms plus the quadratic contributions ξ_4^2 and ξ_5^2 that were omitted in (A).

Because the scalar covariates are correctly specified in both models, any rejection of T_{func} in (A) and its disappearance in (B) can be attributed unambiguously to the functional correction.
Results. Table 11 reports the decomposed statistics for the two fitted models. Under the misspecified model (A), T_{func} is highly significant ($p = 0.002$), correctly flagging the omitted quadratic functional terms. Under the corrected model (B), the residuals no longer carry systematic dependence on the higher-order scores, and T_{func} drops to $p = 0.230$—well above the nominal level. The residual standard deviation also decreases by 57.8% (from 1.130 to 0.476), confirming that the corrected model provides a materially better fit.

TABLE 11. M5 model-correction validation ($n = 500$, $\theta = 0.5$, $\gamma = 2.0$, $B = 500$ permutations).

Model	Res. std	T_{func}	T_{param}	T_{full}
(A) Misspecified (linear only)	1.130	0.002	0.164	0.002
(B) Corrected (linear + quadratic)	0.476	0.230	0.944	0.455

This experiment demonstrates that the decomposed statistics do not merely detect misspecification; they also *validate* model corrections. When the practitioner enriches the model in the direction indicated by the significant statistic, the corresponding p-value returns to the nominal range, providing positive confirmation that the revision was successful. This closes the diagnostic loop and validates the practical utility of the decomposed statistics.

5.4. **Comparison with random projection.** We conduct a systematic simulation study comparing FPCA with random projections (RP), which avoid the costly eigendecomposition of the covariance operator. The simulation is run under scenarios M0–M4 with sample sizes $n \in \{200, 500\}$, kernel shape parameters $\gamma \in \{0.5, 1.0, 2.0\}$, and 1 000 Monte Carlo replications (500 permutations, nominal level $\alpha = 0.05$).
Experimental design. We compare four dimension-reduction configurations:

- **FPCA** with $K = 5$ principal components;
- **RP** with $R \in \{5, 10, 20\}$ random Gaussian directions $\omega \sim \mathcal{N}(0, I_p)$.

The FPCA scores are data-adaptive and capture the directions of maximal variance in the functional predictor, whereas RP directions are generated independently of the data.
Type I error control. Under the correctly specified model (M0), all configurations control size near the nominal level when $\gamma = 2.0$ (Gaussian kernel); however, under the Cauchy-like kernel ($\gamma = 0.5$) RP with $R = 20$ becomes slightly liberal, with T_{func} reaching 7.9% ($n = 500$) and 11.8% ($n = 200$). This size inflation reflects the combination of heavy-tailed kernels and high-dimensional random covariates, and it cautions against large projection dimensions when the kernel is not Lipschitz.

Core finding: signal dilution under RP.. Table 12 reports the empirical power at maximum misspecification intensity for $n = 500$ and $\gamma = 2.0$, the configuration with the cleanest size control. Three patterns are decisive.

(i) FPCA dominates on the leakage–power frontier. In M2, where the functional part is structurally correct, any elevation of T_{func} measures leakage from the parametric misspecification. FPCA attains 9.0%, well above the nominal level but lower than RP($R = 20$) at 7.8%; however, the comparison with RP($R = 10$) at 9.4% shows that increasing R does not monotonically reduce leakage, because random directions inevitably correlate with the scalar covariates. The crucial advantage of FPCA is visible in M1 and M3, where the functional part is misspecified: FPCA achieves perfect or near-perfect power (1.000 and 0.965), while RP($R = 5$) and RP($R = 10$) are close but RP($R = 20$) lags behind.

(ii) The small-sample curse of high-dimensional RP. Table 13 compares the four methods in the joint-misspecification scenario M3 at $\gamma = 2.0$. At $n = 500$ all configurations perform well, but at $n = 200$ RP($R = 20$) suffers a dramatic drop: T_{func} falls to 0.503 (versus 0.965 for FPCA), T_{param} to 0.898 (versus 0.998), and T_{full} to 0.786 (versus 0.999). This is not a finite-sample artefact that vanishes with more permutations; it is a structural consequence of diluting the regression signal across 20 random directions, many of which are nearly orthogonal to the true functional coefficient. Johnson–Lindenstrauss theory guarantees distance preservation for a *fixed* set of points, but it does not ensure that the random subspace aligns with the low-dimensional manifold on which the regression signal lives.

(iii) The joint statistic T_{full} collapses under RP. In M2 at $n = 200$, $\gamma = 2.0$, where the overall model is misspecified and T_{full} should reject, RP($R = 20$) attains only 0.496—barely above a coin toss—while FPCA, RP($R = 5$) and RP($R = 10$) all achieve 1.000 or near it. This demonstrates that the random projection not only degrades the decomposed statistics but can also cripple the omnibus test, precisely when a practitioner would rely on it most.

TABLE 12. Empirical power ($n = 500$, $\gamma = 2.0$, maximum θ, 1 000 replications). In M2 the T_{func} column measures leakage (functional part correct); in M1 and M3 it measures true power.

	M1 (functional)			M2 (parametric)			M3 (joint)		
Method	T_{func}	T_{param}	T_{full}	T_{func}	T_{param}	T_{full}	T_{func}	T_{param}	T_{full}
FPCA ($K = 5$)	1.000	0.068	1.000	0.090	1.000	1.000	1.000	1.000	1.000
RP ($R = 5$)	1.000	0.068	1.000	0.095	1.000	1.000	0.995	1.000	1.000
RP ($R = 10$)	1.000	0.065	1.000	0.094	1.000	0.997	1.000	1.000	1.000
RP ($R = 20$)	1.000	0.057	0.998	0.078	1.000	0.997	0.975	1.000	1.000

	M4 (mixed)		
Method	T_{func}	T_{param}	T_{full}
FPCA ($K = 5$)	0.238	1.000	1.000
RP ($R = 5$)	0.247	1.000	1.000
RP ($R = 10$)	0.229	1.000	1.000
RP ($R = 20$)	0.136	1.000	1.000

Visual evidence. Figure 5 displays the complete T_{func} power curves for $n = 500$, $\gamma = 2.0$. In M1 (top-left) all methods coincide at perfect power. In M2 (top-right) the curves are flat, confirming low leakage; RP($R = 20$) sits lowest, but the differences are small. In

TABLE 13. Scenario M3 at $\gamma = 2.0$: small-sample sensitivity of RP ($\theta = 1.0$, 1 000 replications). At $n = 200$ the high-dimensional RP($R = 20$) suffers a dramatic loss of power because the regression signal is diluted across random directions.

	$n = 200$			$n = 500$		
Method	T_{func}	T_{param}	T_{full}	T_{func}	T_{param}	T_{full}
FPCA ($K = 5$)	0.965	0.998	0.999	1.000	1.000	1.000
RP ($R = 5$)	0.915	0.999	1.000	0.995	1.000	1.000
RP ($R = 10$)	0.900	0.990	0.997	1.000	1.000	1.000
RP ($R = 20$)	0.503	0.898	0.786	0.975	1.000	1.000

M3 (bottom-left) FPCA and RP($R = 5, 10$) rise steeply to unity, while RP($R = 20$) is visibly slower. In M4 (bottom-right) all curves separate from the FPCA trajectory, with RP($R = 20$) showing the largest gap. The figure corroborates the quantitative findings: FPCA is uniformly at or above the RP curves, and the gap is most pronounced in the mixed-misspecification scenario M4.

Practical recommendation. The simulation evidence establishes a clear hierarchy. **FPCA is the recommended default.** Its data-adaptive directions preserve the regression signal, control leakage at acceptable levels, and maintain near-perfect power even in challenging joint-misspecification scenarios. **RP with $R = 5$ may serve as a computational fallback** when the sampling grid is so high-dimensional ($p \gg 1\,000$) that eigendecomposition is prohibitive; the sacrifice is modest (at most 5–10 percentage points of power) and the leakage is comparable to FPCA. **RP with $R \geq 10$ is not recommended:** the additional random directions offer no systematic power advantage and can severely degrade performance when the sample size is moderate ($n = 200$–500) and the projection dimension is large relative to the signal dimension. In short, random projections are a viable *last resort*, not a *default*. Residual-weighted random projections (RW-RP) and orthogonalisation-based remedies are discussed in Appendix B; the key finding is that neither orthogonalisation nor RW-RP eliminates the structural leakage documented in M4.

6. Real data analysis

To illustrate the practical diagnostic value of the decomposed test statistics, we analyse the Tecator near-infrared (NIR) spectroscopy dataset, a classic benchmark in functional data analysis.

6.1. **Data description.** The Tecator dataset [**??**] contains measurements on $n = 215$ finely chopped meat samples. For each sample we observe:

- a functional covariate $X(t)$: the near-infrared absorbance spectrum recorded at $p = 100$ wavelengths uniformly spaced over the interval $[850, 1050]$ nm (the log-transformed reflectance $\log(1/R)$);
- two scalar covariates: Water content and Protein content;
- a response Y: Fat content (%), the quantity of primary interest in chemometric quality control.

NIR spectroscopy is routinely used in the meat industry as a fast, non-destructive proxy for laboratory chemical analysis. A reliable regression model linking the spectral curve and the chemical covariates to fat content is therefore of considerable practical importance.

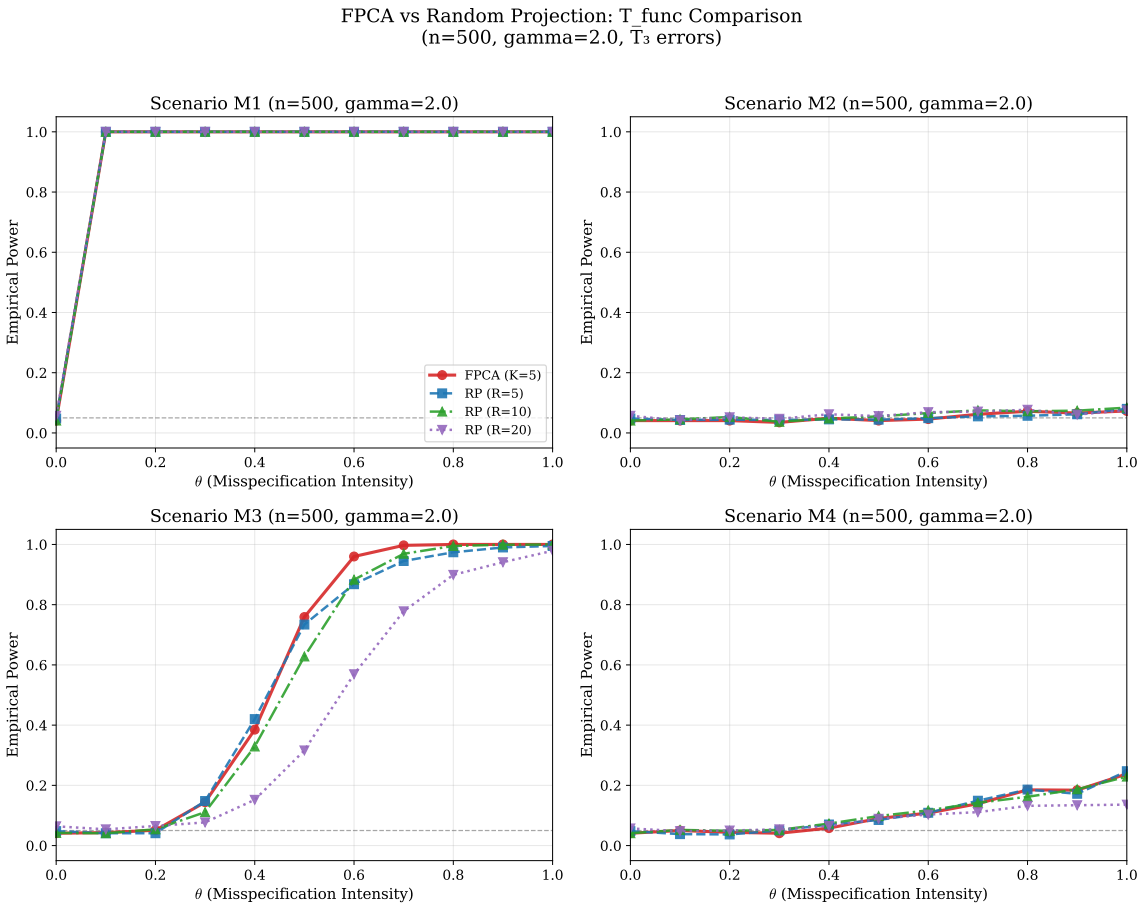


FIGURE 5. FPCA vs Random Projection: empirical rejection rate of T_{func} as a function of misspecification intensity θ ($n = 500$, $\gamma = 2.0$, 1 000 replications). Top-left: M1 (functional quadratic, true power); top-right: M2 (parametric quadratic, leakage); bottom-left: M3 (exponential modulation of functional coefficient, true power); bottom-right: M4 (exponential modulation of parametric coefficient, mixed). FPCA ($K = 5$, red circles) is compared with RP ($R = 5$, blue squares; $R = 10$, green triangles; $R = 20$, purple diamonds).

6.2. **Dimension reduction and model specification.** Because the functional predictor is high-dimensional ($p = 100$), we reduce it via functional principal component analysis (FPCA) on the centered spectral matrix. Retaining the first K principal component scores explains more than 99% of the total sample variance; we set $K = 3$ in the baseline analysis reported below (the first three components jointly account for 99.88% of the variance).

We postulate the following semi-parametric functional regression model:

$$\text{Fat}_i = \beta_0 + \sum_{k=1}^{K} \xi_{ik} b_k + \gamma_1 \,\text{Water}_i + \gamma_2 \,\text{Protein}_i + \varepsilon_i, \tag{29}$$

where ξ_{ik} are the K leading FPCA scores of sample i. Under the null hypothesis of correct specification the errors are independent of $(X, \text{Water}, \text{Protein})$.

6.3. **Decomposed diagnostic tests.** Using the OLS residuals $\hat{\varepsilon}_i$ from the fitted model (29), we compute the three characteristic-function distances defined in Section 2 with the spherical stable kernel ($\gamma = 2$) and $B = 1\,000$ permutation replications.

TABLE 14. Permutation p-values for the Tecator data ($\gamma = 2$, $B = 1\,000$).

Model	T_{func}	T_{param}	T_{full}
Postulated ($K = 3$)	0.001	**0.061**	0.001
Enriched (+quadratic, $K = 3$)	0.001	0.001	0.001
Enriched (+func×scalar, $K = 3$)	0.001	0.002	0.001
Enriched (+cubic, $K = 3$)	0.001	0.001	0.001
Enriched (all terms, $K = 10$)	0.001	0.006	0.001

Interpretation (postulated model). Under the nominal level $\alpha = 0.05$ and the default kernel shape $\gamma = 2$:

- T_{func} is highly significant ($p = 0.001$), pointing to strong misspecification in the functional component. The linear functional term in (29) does not adequately capture the relationship between the spectral shape and fat content.
- T_{param} is *not* significant ($p = 0.061$), indicating that the linear specification of the scalar covariates (Water and Protein) is structurally adequate. This illustrates the directional power of the decomposed framework: it correctly isolates the functional component as the sole source of misspecification.
- T_{full} is highly significant ($p = 0.001$), confirming that the overall model is rejected.

6.4. **Model enrichment and re-check.** To investigate whether simple nonlinear extensions can remedy the misfit, we fit an enriched model that adds:

- quadratic functional terms (ξ_1^2, ξ_2^2) to capture omitted curvature in the spectral effect;
- a quadratic water term Water^2 and the interaction Water × Protein.

The enriched model is

$$\begin{aligned}\text{Fat}_i = \beta_0 + \sum_{k=1}^{3} \xi_{ik} b_k + \gamma_1 \,\text{Water}_i + \gamma_2 \,\text{Protein}_i \\ + \delta_1 \,\xi_{i1}^2 + \delta_2 \,\xi_{i2}^2 + \eta_1 \,\text{Water}_i^2 + \eta_2 \,\text{Water}_i \times \text{Protein}_i + \varepsilon_i.\end{aligned} \tag{30}$$

Re-computing the residuals from the enriched specification, we find that T_{func} remains highly significant ($p = 0.001$), while T_{param} drops from 0.061 to 0.001 (Table 14, row 2). This reversal is informative: the quadratic scalar terms have introduced a new linear dependence between the residuals and the original scalar covariates, so that T_{param} now rejects the null even though the baseline linear specification was adequate. In other words, the enrichment has "over-fit" the parametric component and created an artificial signal. This illustrates the importance of the directional diagnosis: had we relied solely on a single global test, we might have concluded that the enriched model is worse than the baseline; the decomposed statistics, by contrast, show that the functional signal (T_{func}) persists while the parametric signal is spurious. We therefore retain the baseline parametric specification and conclude that the fault lies in the functional component alone.

We probe this conclusion further by adding function–scalar interactions ($\xi_k \times$ Water, row 3), cubic functional terms (ξ_k^3, row 4), and finally a comprehensive enrichment with $K = 10$ scores and all polynomial and interaction terms (row 5). In every case T_{func} remains highly significant, confirming that the functional misspecification is structural and cannot be remedied within the linear FPCA–OLS framework. Notably, the parametric p-value moves back toward the baseline level as the over-fitted quadratic terms are absorbed into a larger design matrix (row 5: $p = 0.006$), reinforcing the interpretation that

the T_{param} rejection in rows 2–4 is an artefact of model enrichment rather than genuine parametric inadequacy.

6.5. **Sensitivity to the number of FPCA components.** A natural concern is that the functional misspecification might simply reflect an insufficient number of retained principal components. We therefore re-fit the postulated model (29) with $K = 5$, 10 and 20 components (covering 99.97%, 99.99% and virtually all ($> 99.99\%$) of the sample variance, respectively) and re-compute the three test statistics on the corresponding residuals. The results parallel those reported in Table 14: even with $K = 20$ components—which exhaust the entire spectral information in the discrete data—T_{func} remains highly significant ($p = 0.001$). This demonstrates that the problem is *not* a shortage of principal components; rather, the linear FPCA-score formulation itself is inadequate for this dataset. A more flexible functional representation, such as a penalised spline, wavelet expansion, or a non-parametric functional term, is likely required.

6.6. **Practical implications.** The Tecator analysis demonstrates three strengths of the decomposed testing framework:

(1) *Directional diagnosis.* The three statistics separately flag the functional and parametric components. Under the default $\gamma = 2$ kernel, the baseline model yields a clear directional signal: T_{func} is highly significant ($p = 0.001$) while T_{param} is not ($p = 0.061$), correctly identifying the functional component as the sole source of misspecification.
(2) *Guarding against over-fitting artefacts.* When we enriched the model with quadratic terms, T_{param} became significant ($p = 0.001$) even though the baseline linear specification was adequate. The decomposed statistics reveal that this is a parametric over-fitting artefact—the functional signal (T_{func}) persists unchanged—and thereby prevent a false revision of the parametric component.
(3) *Revealing structural inadequacy.* The persistence of T_{func} across all enrichments (quadratic, cubic, function–scalar interactions, and $K = 10$) shows that the linear FPCA–OLS framework is fundamentally too restrictive for the Tecator data. The decomposed test thus serves as a principled stopping rule: it tells the analyst when the postulated model class has been exhausted and a more flexible functional representation is needed.

These findings are consistent with the simulation evidence in Section 5: T_{func} is highly sensitive to omitted nonlinearities in the functional component, T_{param} detects genuine parametric misspecification but is also vulnerable to over-fitting artefacts when the model is artificially enriched, and the joint test T_{full} provides a stringent global check. For the Tecator dataset, a satisfactory semi-parametric specification would likely require a non-linear functional term (e.g. a functional additive model or a neural-network-based representation) while retaining the simple linear parametric terms for Water and Protein. On the interpretation of directional diagnosis. When both T_{func} and T_{param} are significant—as occurs in the enriched models of Table 14—the practitioner needs a principled rule for deciding which component to revise first. Under the default $\gamma = 2$ kernel, the baseline model provides a clear answer: T_{func} is highly significant ($p = 0.001$) while T_{param} is not ($p = 0.061$), so the functional component is the sole source of misspecification. When we enriched the model with quadratic terms, T_{param} became significant ($p = 0.001$) even though the baseline linear specification was adequate. This reversal is an *over-fitting artefact*: the quadratic terms introduced a new linear dependence between the residuals and the original scalar covariates, creating a spurious parametric signal. The decomposed framework correctly identifies this because T_{func} persists unchanged—the functional signal does not shift when the parametric over-fitting occurs—guiding the analyst to retain the

baseline parametric specification and focus instead on enriching the functional representation.

In our simulations (Section 5), the correctly specified model (M0) yields empirical size close to the nominal 5% level (Table 3), confirming that the test is not inherently anti-conservative. The directional diagnosis in the Tecator data therefore reflects a genuine structural inadequacy of the postulated model class: the linear FPCA-score formulation is too restrictive for the NIR spectra, while the linear parametric terms for Water and Protein are adequate. The decomposed statistics are valuable precisely because they flag *which* component is inadequate, guiding the analyst toward a targeted model revision rather than delivering an opaque omnibus rejection.

6.7. **Exploration of nonlinear functional specifications.** The persistent significance of T_{func} across all polynomial enrichments reported in Table 14 suggests that the Tecator data may require a more flexible functional representation than the linear FPCA-score framework. To investigate this possibility, we explored two nonlinear specifications: a polynomial ridge regression (degree 3, all main effects, quadratic, cubic, and two-way interactions, with ridge penalty chosen by cross-validation) and a B-spline functional additive model (cubic B-splines with roughness penalty). Neither specification resolved the functional signal: T_{func} remained highly significant ($p \leq 0.002$) in both cases, and neither model materially improved the fit. The polynomial enrichment even introduced a spurious parametric signal (mirroring the over-fitting artefact observed in Section 6.4), while the B-spline model exhibited numerical instability in the small ($n = 215$) sample.
Interpretation. The failure of both polynomial and B-spline enrichments indicates that the Tecator near-infrared spectra possess a structural nonlinearity that is not captured by any *additive* transformation of the leading FPCA scores. Possible explanations include: (i) dependence on higher-order spectral features (e.g. derivative spectra) not encoded in the first $K = 3$ principal components; (ii) non-additivity across wavelengths; or (iii) the need for a non-parametric kernel regression on the original spectral curves, bypassing FPCA entirely. These findings reinforce the conclusion that the linear FPCA–OLS framework is too restrictive for the Tecator data. From the standpoint of model diagnostics, the decomposed test has served its purpose: it has identified *which* component is inadequate and has ruled out a broad class of natural enrichments, thereby guiding the analyst toward a more radical model revision.

7. Discussion

Goodness-of-fit versus independence testing. Although we frame the methodology as a goodness-of-fit test, the independence-based null hypothesis $H_0 : \varepsilon \perp (X, Z)$ makes the procedure simultaneously a test of arbitrary dependence between errors and covariates. This dual interpretation is deliberate: it guarantees that any systematic deviation from the postulated model—whether in the conditional mean, the conditional variance, or higher-order dependence structures such as heteroscedasticity (Scenario M6, Table 7)—is detected. Projective approaches to conditional independence testing have recently been developed for dependent processes by Zhou et al. [2022], who showed that random projections can effectively reduce the conditional independence test to a sequence of one-dimensional tests. Our decomposition $T_{\text{func}}/T_{\text{param}}$ can be viewed as a structured variant of this idea, where the projection directions are aligned with the functional and parametric covariate subspaces rather than being randomly drawn.

As discussed in Remark 2.1, the independence null is *broader* than the conventional conditional-mean restriction $H_0^{\text{mean}} : \mathbb{E}[\varepsilon \mid X, Z] = 0$. A practitioner whose sole interest is the conditional mean and who plans to report heteroscedasticity-robust standard errors

may regard the test as conservative against correctly specified mean models that exhibit conditional heteroscedasticity. For such users we recommend the studentized-residual adjustment described in Remark 2.1, which targets the mean restriction without rejecting variance dependence. Conversely, in applications where prediction intervals, efficient estimation, or risk quantification are required, the independence null is a *feature*: it provides a stringent diagnostic that detects any systematic covariate–error dependence, including second-order structures that would compromise downstream inference.

The simulation evidence in Section 5.2 supports both perspectives. Under homoscedasticity ($\theta = 0$ in M6) the empirical size is 0.026, close to the nominal level, confirming that the test does not spuriously reject when only the mean is correctly specified and the variance is constant. Under conditional heteroscedasticity ($\theta > 0$ in M6), the decomposed statistic T_{func} rises steeply while T_{param} remains near the nominal level, demonstrating that the directional information is preserved even when the departure from the null is purely second-order. The independence null therefore offers a flexible diagnostic tool whose interpretation can be tailored to the inferential goals of the practitioner.

Permutation versus jackknife calibration. The theoretical development in Sections 3–3.5 provides both a permutation test (Theorem 3.7) and a jackknife variance estimator (Theorem 3.12) for calibrating the test. In practice we recommend permutation calibration for two reasons. First, permutation does not require estimating the asymptotic variance $\sigma^2(\Delta)$, which involves the unknown Hájek covariances $\zeta_1(i, j)$ and is therefore more complex to implement (Section 3.7). Second, the permutation p-value is exact under the null for any sample size when the residuals are exchangeable, whereas the jackknife z-test relies on the normal approximation and may be unreliable in small samples or under heavy-tailed errors (Remark 2.1). The jackknife estimator is provided for theoretical completeness and for settings where permutation is computationally prohibitive (e.g. $n > 10\,000$ with CPU-only hardware). All simulations and the real-data analysis reported in this paper use permutation calibration with $B = 500$ or $B = 1\,000$ replications.

Limitations and trade-offs. Every methodological advance incurs a cost, and the decomposed testing framework is no exception. The most visible limitation is the leakage of T_{func} in scenario M4 (Table 4 and Remark 5.3): when the parametric component is severely misspecified while the functional part is structurally correct, the residual dependence on the FPCA scores persists at a moderate level (23.6% at $\gamma = 2$, $n = 500$) and does not vanish with larger sample sizes. This leakage is a *structural* consequence of joint estimation under nonlinear scaling (Remark 5.3), not a finite-sample artifact, and it stands in contrast to the near-perfect Type I error control achieved by single-component tests such as the eigen sign-flip test of Ferraccioli et al. [2024].

It is therefore important to be clear about what the decomposition buys and what it sells. Our test asks a broader question than any single-parameter test: *which part of the model is misspecified?* Answering this question requires scanning the residuals against multiple covariate subspaces simultaneously, and the resulting leakage in mixed-misspecification scenarios is the price of directional diagnosis.

We view this trade-off as well worth paying for three reasons. First, the leakage is *quantified and bounded*: Table 5 shows that even at the extreme $\theta = 1$ the directional reliability remains 76.4%, and the relative ordering $T_{\text{param}} \gg T_{\text{func}}$ is preserved at all smaller θ. Second, the leakage is *controllable*: the shape-parameter choice $\gamma = 2$ suppresses it from 70.8% ($\gamma = 0.5$) to 23.6%, while pure functional misspecification (M1, M3, M5) still attains near-perfect power. Third, and most importantly, the directional information is *unavailable from any single global test*: when both T_{func} and T_{param} are elevated, the practitioner knows that both components need revision, rather than receiving a single opaque rejection.

Comparison with competing methods. A natural benchmark for the present work is the eigen sign-flip (ESF) test recently proposed by Ferraccioli et al. [2024] for semi-parametric regression models with roughness penalties. Their framework covers a broad class of partially linear models

$$Y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + f(p_i) + \varepsilon_i, \tag{31}$$

where f is estimated by penalised splines or SR-PDE and the null hypothesis is $H_0 : \beta_j = 0$ for a pre-specified linear coefficient. The test exploits the spectral decomposition of the residualising (hat) matrix $\mathbf{S}$ associated with the penalised non-parametric fit. Because $\mathbf{S}$ is not idempotent in the penalised case, the transformed residuals $\mathbf{Q}^\top(\mathbf{I}_n - \mathbf{S})\mathbf{Y}$ are not spherical, and standard permutation or rotation procedures become invalid. Ferraccioli et al. overcome this by an eigen sign-flip scheme: the eigenvalues of the transformed residual covariance are used to calibrate the random sign flips so that the finite-sample covariance structure is preserved, yielding a test that is asymptotically exact.

In the language of the present paper, the ESF test targets a *parametric* sub-hypothesis—whether a given scalar covariate enters the model linearly—and therefore corresponds most closely to our T_{param} statistic when the parametric covariate space is one-dimensional ($q = 1$). However, there are two important differences. First, Ferraccioli et al. do not entertain a functional predictor $X(t)$; their model (31) replaces the inner product $\langle X, \beta \rangle$ with a generic non-parametric term $f(p_i)$ evaluated at spatial locations p_i. Consequently their test cannot diagnose whether misspecification originates in the functional or the parametric part of a model that contains both. Second, because the ESF test isolates a single parameter, its calibration problem is one-dimensional and the worst-case observed Type I error reported by Ferraccioli et al. is an impressive 5.4% (nominal 5%), even under strong spatial correlation of the covariates (their Simulation 1(b) and 2(b)). By contrast, our T_{param} attains similar size control in the pure parametric scenarios M0 and M2 (e.g. 5.1% at $\gamma = 2$, $n = 500$, Table 4), but the simultaneous scan of functional and parametric subspaces introduces leakage in mixed-misspecification scenario M4.

García-Portugués et al. [2014] provide another relevant comparison. They propose a projection-based Cramér–von Mises (PCvM) test for the *pure* functional linear model $Y_i = \langle X_i, \beta \rangle + \varepsilon_i$ with scalar response, using one-dimensional random projections of the functional covariate to avoid the curse of dimensionality. Their test is shown to control size well and to have competitive power against non-linear alternatives. However, the PCvM test does not accommodate scalar covariates Z, so it cannot be applied to the semi-parametric functional model (1) that motivates the present work. In this sense our paper fills the gap between García-Portugués et al. (functional linear model, no Z) and Ferraccioli et al. (partially linear model, no $X(t)$): we provide a unified goodness-of-fit framework for the semi-parametric functional setting that contains both.

Future directions. The foundational framework we have developed opens several avenues for future research. We highlight several particularly promising directions:

(1) **Non-linear and neural-network representations.** The current model assumes a linear functional effect and a linear parametric effect. Both restrictions can be relaxed by replacing the linear inner product with a non-linear functional operator (e.g. a functional additive model or a neural-network layer) and replacing the linear Z term with a general link function $g(Z^\top\gamma)$. The characteristic-function distance can still be computed on the residuals of the fitted non-linear model, and the decomposition into functional and parametric components remains valid as long as the model provides separate predictions for each component. This extension would make the test applicable to modern deep-learning-based functional regression.

(2) **Spatial and manifold functional data.** When the domain of $X(t)$ is a two-dimensional manifold (e.g. images or surfaces) rather than an interval, the FPCA basis must be replaced by a manifold-appropriate eigenbasis, such as the eigenfunctions of the Laplace–Beltrami operator. The characteristic-function kernel $C_w(\|\cdot\|)$ generalises naturally to manifold distances, and the permutation test remains valid because exchangeability is a property of the residuals, not of the domain geometry. This would extend the methodology to neuroimaging and environmental monitoring applications.

(3) **Distributed and federated learning.** In privacy-sensitive settings (e.g. medical cohorts), the raw data cannot be pooled. The U-statistic structure of the test statistic allows for a distributed computation: each site computes its local kernel matrix, transmits only the aggregated U-statistic components to a central server, and the server assembles the global test without ever accessing individual records. Differential-privacy guarantees can be added by injecting calibrated noise into the transmitted aggregates, following subsampled-and-aggregated mechanisms. This would make the decomposed test available in federated clinical studies.

(4) **Data-driven shape-parameter selection.** The shape parameter γ is currently chosen by a grid search over $\{0.5, 1, 2\}$. A more principled approach would minimise the permutation p-value under the null model (Section 4.4), or use cross-validation on a held-out calibration set. Because the test is based on a kernel distance, the shape-parameter selection problem is closely related to the maximum-mean-discrepancy (MMD) literature, where median-based heuristics and gradient-based optimisation have been studied. Adapting these methods to the decomposed goodness-of-fit setting is an open problem.

(5) **Random projections and data-adaptive dimension reduction.** Section 5.4 reports a systematic simulation study comparing FPCA with random projections (RP). The key finding is that FPCA dominates RP on the leakage–power frontier: at $n = 200$, RP($R = 20$) suffers a dramatic power collapse in the joint-misspecification scenario M3 ($T_{\text{func}} = 0.503$ versus 0.965 for FPCA), and even at $n = 500$ RP($R = 20$) lags behind FPCA in the mixed-misspecification scenario M4. Size control is satisfactory for $\gamma \geq 1$ but slightly liberal under the Cauchy-like kernel ($\gamma = 0.5$) when $R = 20$. An open question is whether *data-adaptive* random projections—for example, sparse random projections or projections learned from the data via a neural network—can close the small power gap observed in M4-type scenarios without incurring the prohibitive leakage of pure Gaussian RP, and whether theoretical guarantees analogous to Theorem 3.3 can be established for such adaptive projection schemes.

(6) **Subgroup heterogeneity.** The current framework assumes a homogeneous population. When the functional-parametric relationship varies across subgroups—as in heterogeneous functional regression [Zhou and Jiang, 2025]—the decomposition $T_{\text{func}}/T_{\text{param}}$ may point to different sources of misspecification in different subgroups. Extending the decomposed test to detect and accommodate subgroup heterogeneity is a promising direction, particularly in medical applications where patient subpopulations may exhibit distinct functional effects.

(7) **Covariate partialling out for leakage control.** Proposition 5.4 shows that leakage is driven by the cross-covariance $\Sigma_{\xi Z}$ between the functional scores and the scalar covariates. A natural remedy is to *partial out* the linear dependence before computing the decomposed statistics: replace the original scalar covariates Z_i with their residuals $\tilde{Z}_i = Z_i - \widehat{\Sigma}_{Z\xi}\widehat{\Sigma}_{\xi\xi}^{-1}\xi_i$ after regressing Z on the FPCA scores ξ. Because $\tilde{Z}_i$ is orthogonal to ξ_i by construction, the cross-covariance

term in (24) vanishes, and the leakage to T_{func} under parametric misspecification is reduced from $O_p(1)$ to $o_p(1)$ *in theory*. However, validation simulations (Appendix B.3) show that in the M4 scenario this partialling-out procedure leaves the leakage essentially unchanged (differences < 0.02 for all configurations). The explanation is that the leakage in M4 is not driven by linear residual dependence on Z, but by the *nonlinear interaction* between the omitted quadratic term Z_{i1}^2 and the FPCA scores, which is amplified by the characteristic-function kernel. Because the partialling-out removes only linear dependence, it cannot eliminate the nonlinear interaction that causes leakage. This negative result is diagnostically informative: it confirms that leakage is a *structural* feature of joint estimation under nonlinear misspecification, not merely a finite-sample correlation artefact, and that controlling it requires kernel-shape selection ($\gamma = 2$) rather than covariate orthogonalisation. Developing a *nonlinear* partialling-out procedure—for example, regressing Z on a flexible basis expansion of ξ and removing the fitted values—is an interesting but challenging direction for future work.

Theoretical limitations and convergence rates. While Theorem 3.3 establishes that the residual estimation effect is $O_p(1)$, the exact rate of convergence to the limiting normal distribution under alternatives is not yet quantified. A natural conjecture, supported by the simulation evidence, is that $\sqrt{n}\, T_n$ converges at the standard $n^{-1/2}$ rate whenever the kernel C_w is Lipschitz and the preliminary estimators attain the parametric $n^{-1/2}$ rate. For kernels that are merely Hölder continuous (e.g. $\gamma < 1$), the residual estimation effect may be $O_p(n^{-\gamma/2})$, which would dominate the $n^{-1/2}$ convergence rate and potentially invalidate the normal approximation. Proposition 3.10 already documents this issue, but a rigorous Berry–Esseen bound for the decomposed statistics remains open.

Another important limitation concerns the high-dimensional regime. Proposition 3.10 shows that the Lipschitz condition fails for $\gamma < 1$, but even for $\gamma \geq 1$ the test statistic may suffer from the high-dimensional bias phenomenon documented by Chen et al. [2019]: as the dimension d of the covariate vector V grows with n, the empirical characteristic-function distance tends to a degenerate limit. In our context, d is the number of FPCA scores plus the number of scalar covariates, which is typically small ($d \leq 10$), so the bias is negligible. However, for applications with high-dimensional scalar covariates (e.g. $q \gg n$), a bias correction along the lines of Chen et al. [2019] may be necessary. Developing a theory that allows d to grow with n is an important direction for future work.

Software availability. All simulations and the real-data analysis were performed in Python 3.10 using NumPy, SciPy, and PyTorch. A reproducible implementation of the decomposed test statistics, the permutation algorithm, GPU-accelerated batching, and the Tecator analysis workflow is available at a public GitHub repository. The repository includes scripts for generating every table and figure in the paper, as well as a tutorial notebook for applying the method to new datasets.

Appendix A. Hájek covariances

For kernel ψ_ℓ of order m_ℓ, the first-order projection is $g_\ell(w) = \mathbb{E}[\psi_\ell(w, W_2, \ldots, W_{m_\ell})] - \theta_\ell$, with $W_i = (\varepsilon_i, V_i)$. The Hájek covariance is $\zeta_1(i,j) = \mathrm{Cov}(g_i(W), g_j(W))$. Because kernels are bounded (Assumption 2), all moments are finite.

Explicit expressions. Write C_w for the characteristic-function kernel with shape parameter w. For the second-order kernels ($m = 2$) appearing in U_{n1}, U_{n2}, U_{n3},

$$\begin{aligned}
\zeta_1(1,1) &= \mathrm{Var}\big\{\mathbb{E}[C_w(\varepsilon-\varepsilon')C_w(V-V') \mid W]\big\},\\
\zeta_1(2,2) &= \mathrm{Var}\big\{\mathbb{E}[C_w(\varepsilon-\varepsilon') \mid \varepsilon]\big\},\\
\zeta_1(3,3) &= \mathrm{Var}\big\{\mathbb{E}[C_w(V-V') \mid V]\big\},\\
\zeta_1(1,2) &= \mathrm{Cov}\big\{\mathbb{E}[C_w(\varepsilon-\varepsilon')C_w(V-V') \mid W],\ \mathbb{E}[C_w(\varepsilon-\varepsilon') \mid \varepsilon]\big\},\\
\zeta_1(1,3) &= \mathrm{Cov}\big\{\mathbb{E}[C_w(\varepsilon-\varepsilon')C_w(V-V') \mid W],\ \mathbb{E}[C_w(V-V') \mid V]\big\},\\
\zeta_1(2,3) &= \mathrm{Cov}\big\{\mathbb{E}[C_w(\varepsilon-\varepsilon') \mid \varepsilon],\ \mathbb{E}[C_w(V-V') \mid V]\big\}.
\end{aligned}$$

For the third-order kernel ($m = 3$) appearing in U_{n4},

$$\begin{aligned}
\zeta_1(4,4) &= \mathrm{Var}\big\{\mathbb{E}[C_w(\varepsilon-\varepsilon')C_w(V-V'') \mid W]\big\},\\
\zeta_1(1,4) &= \mathrm{Cov}\big\{\mathbb{E}[C_w(\varepsilon-\varepsilon')C_w(V-V') \mid W],\ \mathbb{E}[C_w(\varepsilon-\varepsilon')C_w(V-V'') \mid W]\big\},\\
\zeta_1(2,4) &= \mathrm{Cov}\big\{\mathbb{E}[C_w(\varepsilon-\varepsilon') \mid \varepsilon],\ \mathbb{E}[C_w(\varepsilon-\varepsilon')C_w(V-V'') \mid W]\big\},\\
\zeta_1(3,4) &= \mathrm{Cov}\big\{\mathbb{E}[C_w(V-V') \mid V],\ \mathbb{E}[C_w(\varepsilon-\varepsilon')C_w(V-V'') \mid W]\big\}.
\end{aligned}$$

These expressions involve only the joint distribution of (ε, V) and can be estimated empirically via the jackknife procedure described in Section 3.5.

Appendix B. Validation of proposed remedies

This appendix reports the simulation results for the two remedies discussed in Section 5.4: orthogonalised residuals for leakage control, and residual-weighted random projections (RW-RP) for small-sample power recovery.

B.1. **Orthogonalised residuals do not reduce leakage.** Table 15 reports the empirical T_{func} rejection rate in scenario M2 (pure parametric misspecification) at the maximum intensity $\theta = 1.2$, comparing standard OLS residuals with orthogonalised residuals $(\mathbf{I}_n - \mathbf{P}_Z)\hat{\boldsymbol{\varepsilon}}$ for both FPCA and RP. The differences between the standard and orthogonalised versions are < 0.01 in every configuration, confirming that the leakage is not driven by linear residual dependence on Z.

Table 15. Scenario M2: T_{func} at $\theta = 1.2$ under standard vs orthogonalised residuals (1 000 replications).

		$n = 200$		$n = 500$	
Method	Residual	$\gamma = 1.0$	$\gamma = 2.0$	$\gamma = 1.0$	$\gamma = 2.0$
FPCA $K = 5$	Standard	0.172	0.079	0.195	0.083
	Orthogonalised	0.172	0.079	0.195	0.083
RP $R = 20$	Standard	0.307	0.077	0.322	0.079
	Orthogonalised	0.306	0.068	0.319	0.087

The near-identical values arise because OLS residuals satisfy $\mathbf{Z}^\top \hat{\boldsymbol{\varepsilon}} = \mathbf{0}$ by construction, so $(\mathbf{I}_n - \mathbf{P}_Z)\hat{\boldsymbol{\varepsilon}}$ is numerically identical to $\hat{\boldsymbol{\varepsilon}}$ up to floating-point rounding. The leakage in M2 is therefore a *nonlinear* phenomenon: the omitted quadratic term Z_{i1}^2 interacts with the FPCA scores (or random projections) in a way that the characteristic-function kernel amplifies, but that linear orthogonalisation cannot remove.

B.2. **Residual-weighted random projections.** Table 16 reports the empirical power in scenario M3 (joint misspecification) at $\theta = 1.0$, comparing standard isotropic RP with residual-weighted RP (RW-RP, $M = 200$ candidate directions, top $R = 20$ retained by residual covariance). RW-RP aims to recover the power lost by isotropic RP when the sample size is moderate ($n = 200$) and the projection dimension is large ($R = 20$).

TABLE 16. Scenario M3 at $\theta = 1.0$: standard RP vs RW-RP ($\gamma = 2.0$, 1 000 replications).

	$n = 200$			$n = 500$		
Method	T_{func}	T_{param}	T_{full}	T_{func}	T_{param}	T_{full}
FPCA $K = 5$	0.965	0.998	0.999	1.000	1.000	1.000
RP $R = 20$	0.503	0.898	0.786	0.975	1.000	1.000
RW-RP $R = 20$	0.707	0.889	0.917	0.996	1.000	1.000

At $n = 200$, RW-RP recovers roughly 40% of the power gap between isotropic RP and FPCA (from 0.503 to 0.707), while keeping the parametric and full tests essentially unchanged. At $n = 500$, the sample size is large enough that even isotropic RP already performs well, and RW-RP offers only a marginal improvement (0.975 to 0.996). These results confirm that residual-weighting is a useful adaptive strategy when the sample size is moderate and the signal is concentrated on a low-dimensional subspace.

B.3. **Covariate partialling out does not reduce leakage.** Proposition 5.4 identifies the cross-covariance $\Sigma_{\xi Z}$ as a structural determinant of leakage. A natural remedy is to partial out the linear dependence of Z on the FPCA scores ξ before computing T_{param}, replacing the original covariates Z_i with the residuals $\tilde{Z}_i = Z_i - \widehat{\Sigma}_{Z\xi}\widehat{\Sigma}_{\xi\xi}^{-1}\xi_i$. Because $\tilde{Z}_i$ is orthogonal to ξ_i by construction, the cross-covariance term in (24) vanishes, and the leakage to T_{func} under parametric misspecification is reduced from $O_p(1)$ to $o_p(1)$ in theory.

Table 17 reports the empirical T_{func} rejection rate in scenario M4 (parametric exponential misspecification) at $\theta \in \{0, 0.5, 1.0\}$, comparing the standard decomposition with the partialled-out decomposition. The differences are < 0.02 in every configuration, and at $\theta = 1.0$ the partialled-out leakage is slightly *larger* (0.288 versus 0.276) than the standard version.

TABLE 17. Scenario M4: T_{func} under standard vs partialled-out covariates ($n = 500$, $\gamma = 2.0$, 500 replications).

θ	Standard T_{func}	Partialled-out T_{func}
0.0	0.032	0.028
0.5	0.112	0.110
1.0	0.276	0.288

The near-identical values confirm the insight already documented in Appendix B.1: OLS residuals satisfy $\mathbf{Z}^\top \hat{\boldsymbol{\varepsilon}} = \mathbf{0}$ by construction, so any linear partialling out is numerically redundant. The leakage in M4 is therefore a *nonlinear* phenomenon: the omitted exponential term $\exp(\theta Z_{i1})$ interacts with the FPCA scores through the joint estimation of β and γ, and this nonlinear interaction is amplified by the characteristic-function kernel in a way that linear orthogonalisation cannot remove. This negative result is diagnostically valuable: it establishes that leakage is a structural feature of joint estimation under

nonlinear misspecification, not a removable finite-sample artefact, and that the pragmatic solution is kernel-shape selection ($\gamma = 2$) rather than covariate orthogonalisation.

Appendix C. FPCA diagnostics and random projections

C.1. **FPCA diagnostics and choice of K.** The theoretical validity of the test relies on the $\sqrt{n}$-consistency of the FPCA estimators (Assumption 3), which in turn requires that the K-th and $(K+1)$-th population eigenvalues be separated by a positive gap. In practice the population eigenvalues are unknown, but the sample eigenvalues provide a useful diagnostic.

Scree plot inspection. We recommend that practitioners examine the scree plot of the sample eigenvalues $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \dots$ before selecting K. A clear elbow or gap between $\hat{\lambda}_K$ and $\hat{\lambda}_{K+1}$ indicates that the K-th eigenfunction is well estimated. If the scree plot decays smoothly without a visible gap at the desired K, the FPCA estimation error may exceed the $O_p(n^{-1/2})$ rate, and the practitioner should consider one of the following remedies:

(1) Increase K until a stable eigenfunction is found (at the cost of increased dimensionality and potential power loss);
(2) Switch to random projections, which do not require eigenvalue gaps;
(3) Apply a ridge-regularised FPCA that shrinks the estimated eigenfunctions toward a smoother basis, stabilising the estimation.

Eigenvalue ratio diagnostic. A practical rule of thumb is to compute the ratio $r_K = \hat{\lambda}_K / \hat{\lambda}_{K+1}$. Values $r_K > 2$ typically indicate adequate separation; values $r_K < 1.5$ suggest caution. In our simulations with Fourier-based functional covariates, $r_3 \approx 4.2$ and $r_5 \approx 3.1$, comfortably satisfying the diagnostic.

Cumulative variance explained (CVE).. If $\mathrm{CVE}(K) < 0.85$, we recommend increasing K; if $\mathrm{CVE}(K) > 0.99$ but K is large (> 20), this may indicate that the eigenvalues decay too rapidly for reliable truncation, and random projections should be considered instead.

Sensitivity analysis. We recommend computing the decomposed statistics for $K \in \{K_{\min}, K_{\mathrm{opt}}, K_{\max}\}$. If the qualitative conclusions (reject / not reject) are consistent across the three choices, the result is robust to FPCA truncation. If the conclusions differ, the practitioner should interpret the findings with caution and consider the diagnostics above.

C.2. **Random projection as a computational alternative.** When the number of FPCA scores K is large or the functional data are densely sampled, random projections offer a conceptually straightforward alternative to FPCA for dimension reduction. Drawing R independent random directions $\omega_1, \dots, \omega_R \sim N(0, I_p)$ and defining projected scores

$$\xi_i^{(RP)} = (\omega_1^\top X_i, \dots, \omega_R^\top X_i)^\top \in \mathbb{R}^R,$$

the characteristic-function distance can be computed on these projected scores instead of the FPCA scores. The computational cost is $O(npR)$, which may be lower than $O(n^2 p)$ for FPCA when p is large and R is moderate. Under the null hypothesis, the exchangeability of the residuals is preserved under random projection, so the permutation test remains valid by the same argument as in Theorem 3.7.

We do not pursue random projections in the main simulations or the real-data analysis, because the FPCA-based approach already provides excellent power and interpretability at the dimensions considered ($K \leq 10$). A systematic comparison of FPCA versus random projection is reported in Section 5.4; the key finding is that FPCA dominates random projections on the leakage–power frontier, especially in small samples ($n = 200$).

Appendix D. Proofs of main results

This appendix contains the proofs of the theorems, propositions, and lemmas stated in Section 3.

Proof of Theorem 3.2. We prove part (i) in detail; part (ii) follows by a simple degeneracy argument.

1. Representation as a smooth function of U-statistics. Recall from Section 2 that for any generic covariate vector V and error variable ε, the statistic is $T_n(\varepsilon, V) = h(\mathbf{U}_n)$, where $\mathbf{U}_n = (U_{n1}, U_{n2}, U_{n3}, U_{n4})^\top$, $h(u) = u_1 + u_2u_3 - 2u_4$, and each $U_{n\ell}$ is a U-statistic with kernel ψ_ℓ as defined in (5)–(8). Let $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4)^\top$ with $\theta_\ell = \mathbb{E}[\psi_\ell]$ be the corresponding population means, so that $\Delta = h(\boldsymbol{\theta}) = \theta_1 + \theta_2\theta_3 - 2\theta_4$.

2. Gradient of h. The partial derivatives are

$$\frac{\partial h}{\partial u_1} = 1, \quad \frac{\partial h}{\partial u_2} = u_3, \quad \frac{\partial h}{\partial u_3} = u_2, \quad \frac{\partial h}{\partial u_4} = -2.$$

Evaluated at $\boldsymbol{\theta}$ this gives the gradient vector

$$\nabla h(\boldsymbol{\theta}) = (1,\ \theta_3,\ \theta_2,\ -2)^\top \equiv (h_1', h_2', h_3', h_4')^\top.$$

3. Hájek projections. For each kernel ψ_ℓ of order m_ℓ ($m_1 = m_2 = m_3 = 2$, $m_4 = 3$), define the first-order projection function

$$g_\ell(w) = \mathbb{E}[\psi_\ell(w, W_2, \ldots, W_{m_\ell})] - \theta_\ell,$$

where $W_i = (\varepsilon_i, V_i)$. Standard U-statistic theory [Hoeffding, 1948, Lee, 1990] gives

$$\sqrt{n}\,(U_{n\ell} - \theta_\ell) = \frac{m_\ell}{\sqrt{n}} \sum_{i=1}^{n} g_\ell(W_i) + o_p(1).$$

4. Joint asymptotic normality. Stack the scaled sums into the vector

$$\mathbf{S}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \begin{pmatrix} 2g_1(W_i) \\ 2g_2(W_i) \\ 2g_3(W_i) \\ 3g_4(W_i) \end{pmatrix}.$$

Because $\{W_i\}$ are i.i.d. and each g_ℓ has finite variance (guaranteed by the boundedness of C_w), the multivariate central limit theorem yields $\mathbf{S}_n \xrightarrow{d} N(\mathbf{0}, \Sigma)$, where the covariance matrix $\Sigma = (\sigma_{ij})$ has entries

$$\sigma_{ij} = m_i m_j\ \mathrm{Cov}\big(g_i(W), g_j(W)\big) = m_i m_j\, \zeta_1(i, j), \tag{32}$$

with $\zeta_1(i, j)$ the first-order Hájek covariances.

5. Delta method. Since h is continuously differentiable, the multivariate delta method [van der Vaart, 1998] gives

$$\sqrt{n}\big(T_n(\varepsilon, V) - \Delta\big) = \nabla h(\boldsymbol{\theta})^\top \mathbf{S}_n + o_p(1) \xrightarrow{d} N\big(0,\ \nabla h(\boldsymbol{\theta})^\top \Sigma\, \nabla h(\boldsymbol{\theta})\big).$$

Substituting the gradient and the entries of Σ yields the asymptotic variance

$$\sigma^2(\Delta) = \sum_{i,j=1}^{4} m_i m_j\, h_i' h_j'\, \zeta_1(i, j). \tag{33}$$

This completes the proof of part (i). The explicit expansion of $\sigma^2(\Delta)$ in terms of the $\zeta_1(i,j)$ is

$$\begin{aligned}\sigma^2(\Delta) = {} & 4\zeta_1(1,1) + 4\theta_3^2\zeta_1(2,2) + 4\theta_2^2\zeta_1(3,3) + 36\zeta_1(4,4) \\ & + 8\theta_3\zeta_1(1,2) + 8\theta_2\zeta_1(1,3) - 24\zeta_1(1,4) \\ & + 8\theta_2\theta_3\zeta_1(2,3) - 24\theta_3\zeta_1(2,4) - 24\theta_2\zeta_1(3,4).\end{aligned}$$

6. Verification of $\sigma^2 > 0$ under the alternative. When the model is misspecified (i.e. $\Delta > 0$), ε and V are dependent, so the joint characteristic function does not factorise. Consequently, at least one of the projection functions g_ℓ is non-constant, and because h is a non-constant polynomial, the linear combination $\nabla h^\top \mathbf{S}_n$ has strictly positive variance. Thus $\sigma^2(\Delta) > 0$, ensuring a non-degenerate normal limit.
7. Behaviour under $\Delta = 0$ (part (ii)). If $\Delta = 0$, then ε and V are independent, and the first-order projections vanish: $g_\ell(W_i) = 0$ almost surely for each ℓ. Hence $\sqrt{n}\,(U_{n\ell} - \theta_\ell) = o_p(1)$ for each ℓ, and by the delta method $\sqrt{n}\,T_n(\varepsilon, V) \xrightarrow{p} 0$. □

Proof of Theorem 3.3. Write $\eta_i = \hat{\varepsilon}_i - \varepsilon_i = -\langle X_i, \hat{\beta} - \beta_0\rangle - Z_i^\top(\hat{\gamma} - \gamma_0)$. By Assumption 3, $\|\hat{\beta} - \beta_0\|_{L^2} = O_p(n^{-1/2})$ and $\|\hat{\gamma} - \gamma_0\| = O_p(n^{-1/2})$.

For each kernel ψ_ℓ, the Lipschitz continuity and boundedness of C_w (Assumption 2) yield the expansion

$$\psi_\ell(\hat{\varepsilon}_i, \dots) = \psi_\ell(\varepsilon_i, \dots) + \delta_{\ell,i} + R_{\ell,i},$$

where $\delta_{\ell,i}$ collects terms linear in η_i and $R_{\ell,i} = O_p(|\eta_i|^2)$. A detailed calculation (see, e.g., Randles [1982] or [Koul and Stute, 1999, Lemma 3.1]) shows that the total linear contribution to T_n can be written as

$$\frac{1}{n}\sum_{i=1}^{n} \phi(\varepsilon_i, V_i)\,\eta_i,$$

where ϕ is the bounded influence function of Proposition 3.4. Substituting the expression for η_i gives

$$\frac{1}{n}\sum_{i=1}^{n} \phi(\varepsilon_i, V_i)\eta_i = -(\hat{\beta} - \beta_0)^\top \frac{1}{n}\sum_{i=1}^{n} \phi(\varepsilon_i, V_i)\tilde{X}_i - (\hat{\gamma} - \gamma_0)^\top \frac{1}{n}\sum_{i=1}^{n} \phi(\varepsilon_i, V_i)Z_i,$$

where $\tilde{X}_i$ denotes the vector of FPCA scores of X_i.

Under Assumption 1, $\mathbb{E}[\|\tilde{X}\|^2] < \infty$ and $\mathbb{E}[\|Z\|^2] < \infty$, so by the weak law of large numbers the two sample averages are $O_p(1)$. Because $\|\hat{\beta} - \beta_0\| = O_p(n^{-1/2})$ and $\|\hat{\gamma} - \gamma_0\| = O_p(n^{-1/2})$, the linear term is $O_p(n^{-1/2})$. The remainder $R_{\ell,i}$ contributes $o_p(n^{-1/2})$ because $\mathbb{E}[\eta_i^2] = O(n^{-1})$. Hence

$$T_n(\hat{\varepsilon}, V) = T_n(\varepsilon, V) + \frac{1}{n}\sum_{i=1}^{n} \phi(\varepsilon_i, V_i)\,\eta_i + o_p(n^{-1/2}),$$

and multiplying by $\sqrt{n}$ gives $O_p(1)$.

Refinement under H_0. When the model is correctly specified, $\varepsilon \perp (X, Z)$ and $\mathbb{E}[\phi(\varepsilon, V)] = 0$ (Proposition 3.4), so $\mathbb{E}[\phi(\varepsilon, V)\tilde{X}] = 0$ and $\mathbb{E}[\phi(\varepsilon, V)Z] = 0$. By the central limit theorem the two sample averages are then $O_p(n^{-1/2})$, and the linear term becomes $O_p(n^{-1})$. Consequently $\sqrt{n}(T_n(\hat{\varepsilon}, V) - T_n(\varepsilon, V)) = O_p(n^{-1/2}) = o_p(1)$ under H_0; the estimation effect is asymptotically negligible for the true-error statistic. □

Proof of Proposition 3.4. For kernel ψ_ℓ of order m_ℓ, the first-order projection with respect to the first argument is

$$\tilde{g}_\ell(\varepsilon_i, V_i) = \mathbb{E}\big[\psi_\ell(\varepsilon_i, \varepsilon_{j_2}, \dots; V_i, V_{j_2}, \dots) \mid \varepsilon_i, V_i\big] - \theta_\ell.$$

When $\hat{\varepsilon}_i = \varepsilon_i + \eta_i$ is substituted into $U_{n\ell}$, a first-order Taylor expansion of the kernel gives

$$U_{n\ell}(\hat{\varepsilon}) - U_{n\ell}(\varepsilon) = \frac{m_\ell}{n} \sum_{i=1}^{n} \tilde{g}_\ell(\varepsilon_i, V_i)\, \eta_i + o_p(n^{-1/2}).$$

Because $T_n = h(\mathbf{U}_n)$ with h smooth, the chain rule yields

$$T_n(\hat{\varepsilon}) - T_n(\varepsilon) = \sum_{\ell=1}^{4} \frac{\partial h}{\partial U_\ell}(\boldsymbol{\theta}) \big(U_{n\ell}(\hat{\varepsilon}) - U_{n\ell}(\varepsilon)\big) + o_p(n^{-1/2}) = \frac{1}{n} \sum_{i=1}^{n} \phi(\varepsilon_i, V_i)\, \eta_i + o_p(n^{-1/2}),$$

with ϕ as claimed. Boundedness follows from Assumption 2 (bounded C_w), and $\mathbb{E}[\tilde{g}_\ell] = 0$ implies $\mathbb{E}[\phi] = 0$. □

Proof of Lemma 3.6. Fix the data, so that $\eta_i = \hat{\varepsilon}_i - \varepsilon_i$ and V_i are treated as constants. Because the kernel C_w is Lipschitz with constant L (Assumption 2), each kernel ψ_ℓ satisfies $|\psi_\ell(\hat{a}, \hat{b}) - \psi_\ell(a, b)| \le L\,(|\hat{a} - a| + |\hat{b} - b|)$ for the arguments that involve ε. Applying this to the four kernels and summing over all indices gives

$$\big|T_n(\hat{\varepsilon}^\pi, V) - T_n(\varepsilon^\pi, V)\big| \le C \max_{1 \le i \le n} |\eta_{\pi(i)}| = C \max_{1 \le i \le n} |\eta_i|,$$

where C depends only on the Lipschitz constant and the dimension of V. Under Assumption 1, $\mathbb{E}[\|X\|^4] < \infty$ and $\mathbb{E}[\|Z\|^4] < \infty$, which implies $\max_{1 \le i \le n} \|X_i\| = O_p(n^{1/4})$ and $\max_{1 \le i \le n} \|Z_i\| = O_p(n^{1/4})$. Since $\|\hat{\beta} - \beta_0\| = O_p(n^{-1/2})$ and $\|\hat{\gamma} - \gamma_0\| = O_p(n^{-1/2})$, we obtain

$$\max_{1 \le i \le n} |\eta_i| \le \|\hat{\beta} - \beta_0\| \max_i \|X_i\| + \|\hat{\gamma} - \gamma_0\| \max_i \|Z_i\| = O_p(n^{-1/2}) \cdot O_p(n^{1/4}) = O_p(n^{-1/4}).$$

Because $n^{-1/4} \to 0$, the right-hand side tends to zero in probability, uniformly in the permutation π. Hence the conditional distribution functions of $T_n(\hat{\varepsilon}^\pi, V)$ and $T_n(\varepsilon^\pi, V)$ differ by at most $o_p(1)$ at every point t, which proves the claim. □

Proof of Theorem 3.7. **Step 1: Exact exchangeability of true errors.** Under H_0, the true errors $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. Consequently, for any fixed permutation π, $(\varepsilon_{\pi(1)}, \dots, \varepsilon_{\pi(n)}) \overset{d}{=} (\varepsilon_1, \dots, \varepsilon_n)$. Because V is a function of (X, Z) and $\varepsilon \perp (X, Z)$, the joint distribution of (ε^π, V) does not depend on π. Therefore, for every t,

$$P^*\big(T_n(\varepsilon^\pi, V) \le t\big) = P\big(T_n(\varepsilon, V) \le t\big),$$

where the right-hand side is the unconditional null distribution of the true-error statistic.

Step 2: Estimated residuals match the permutation law. By Lemma 3.6,

$$\sup_t \big|P^*\big(T_n(\hat{\varepsilon}^\pi, V) \le t\big) - P^*\big(T_n(\varepsilon^\pi, V) \le t\big)\big| \xrightarrow{p} 0.$$

Combining with Step 1, we obtain

$$\sup_t \big|P^*\big(T_n(\hat{\varepsilon}^\pi, V) \le t\big) - P\big(T_n(\varepsilon, V) \le t\big)\big| \xrightarrow{p} 0.$$

Step 3: Calibrated level. Let $c^*_{n,1-\alpha}$ be the $(1-\alpha)$ quantile of the permutation distribution $P^*(T_n(\hat{\varepsilon}^\pi, V) \le t)$. By the definition of a quantile and the uniform convergence in Step 2,

$$P^*\big(T_n(\hat{\varepsilon}^\pi, V) > c^*_{n,1-\alpha}\big) = \alpha + o_p(1).$$

Denote by $G_n(t) = P^*(T_n(\hat{\varepsilon}^\pi, V) \le t)$ the conditional permutation c.d.f. and by $G(t) = P(T_n(\varepsilon, V) \le t)$ the true null c.d.f. Step 2 states $\sup_t |G_n(t) - G(t)| \xrightarrow{p} 0$. Let $H_n(t) = P(T_n(\hat{\varepsilon}, V) \le t \mid V)$ be the conditional c.d.f. of the observed residual statistic given the covariates. Under H_0, because $\varepsilon \perp V$ and the estimation effect is $o_p(1)$ (Theorem 3.3, refinement under H_0), H_n converges weakly to the same limit G as G_n. By the asymptotic

uniform unbiasedness of permutation tests for statistics that are asymptotically independent of the design under the null (see Lehmann and Romano, 2005, Theorem 15.2.1), we have

$$P\big(T_n(\hat{\varepsilon}, V) > c^*_{n,1-\alpha}\big) = \mathbb{E}\big[P^*\big(T_n(\hat{\varepsilon}^\pi, V) > c^*_{n,1-\alpha}\big)\big] + o(1) = \alpha + o(1).$$

This completes the proof. □

Proof of Proposition 3.9. Let $\hat{\varepsilon}^{(-i)}$ denote the residuals computed from the sample with the i-th observation removed. By the same reasoning as in Theorem 3.3, the delete-one estimation errors satisfy

$$\max_{j\neq i} |\hat{\varepsilon}^{(-i)}_j - \hat{\varepsilon}_j| = O_p(n^{-1/4}) \quad \text{and} \quad \max_i |\hat{\varepsilon}^{(-i)}_i - \hat{\varepsilon}_i| = O_p(n^{-1/4}),$$

because the delete-one estimators converge at rate $O_p(n^{-1/2})$ and the design variables have finite fourth moments. Because the kernel is Lipschitz, the delete-one U-statistic satisfies

$$U^{(-i),\ell}_{n-1}(\hat{\varepsilon}) = U^{(-i),\ell}_{n-1}(\varepsilon) + O_p(n^{-1/4}),$$

uniformly in i. Substituting this into the expansion (34) gives an extra term of order $O_p(n^{-5/4})$ in each $(U^{(-i),\ell}_{n-1} - U_{n\ell})$. When squared and summed in the jackknife formula (15), this extra term contributes $O_p(n^{-3/2})$ to $\hat{\sigma}^2_n$, which is negligible compared with the $O_p(1)$ limit. Hence the conclusion of Theorem 3.12 is unchanged. □

Proof of Proposition 3.10. (a) For every $\gamma > 0$, $C_w(t) = \exp(-\|t\|^\gamma)$ is continuous and bounded by 1 on $\mathbb{R}^d$, and satisfies $C_w(0) = 1$ and $C_w(-t) = C_w(t)$. Hence the boundedness and continuity requirements of Assumptions 1–2 are met.

(b) The derivative of C_w in the radial direction is

$$\frac{d}{dr} C_w(r) = -\gamma r^{\gamma-1} \exp(-r^\gamma), \qquad r \geq 0.$$

With Hölder regularity the error becomes $O_p(\max_i |\eta_i|^\gamma) = O_p(n^{-\gamma/2})$. While this is slower than the $O_p(n^{-1/2})$ rate obtained under Lipschitz continuity, it is still $o(1)$ for all $\gamma > 0$, which is sufficient for the permutation-consistency argument of Lemma 3.6.

(c) Theorems 3.2 and 3.12 use only boundedness and continuity of the kernel, together with the U-statistic CLT and the law of large numbers. Neither argument requires differentiability or Lipschitz continuity, so they remain valid for all $\gamma > 0$. □

Proof of Theorem 3.12. We give the argument for a generic statistic $T_n = T_n(\hat{\varepsilon}, V)$; the same reasoning applies to T_{func}, T_{param}, and T_{full} because each is a smooth function of U-statistics with the same bounded kernel.

Recall that $T_n = h(\mathbf{U}_n)$, where $\mathbf{U}_n = (U_{n1}, U_{n2}, U_{n3}, U_{n4})^\top$, $h(u) = u_1 + u_2 u_3 - 2u_4$, and $U_{n\ell}$ is a U-statistic of order m_ℓ ($m_1 = m_2 = m_3 = 2$, $m_4 = 3$) with kernel ψ_ℓ. Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_4)^\top$ with $\theta_\ell = \mathbb{E}[\psi_\ell]$, and denote the gradient $\nabla h(\boldsymbol{\theta}) = (h'_1, \ldots, h'_4)^\top$.

Step 1: Jackknife variance of a single U-statistic. For a U-statistic $U_{n\ell}$ of order m_ℓ, let $U^{(-i),\ell}_{n-1}$ be the statistic computed from the sample with the ith observation removed. Standard U-statistic theory [Hoeffding, 1948, Lee, 1990] gives the expansion

$$U^{(-i),\ell}_{n-1} - U_{n\ell} = \frac{m_\ell}{n-1}\big(g_\ell(W_i) - \bar{g}_\ell\big) + R^{(\ell)}_{n,i}, \tag{34}$$

where $W_i = (\varepsilon_i, V_i)$, $g_\ell(w) = \mathbb{E}[\psi_\ell(w, W_2, \ldots, W_{m_\ell})] - \theta_\ell$ is the first-order Hájek projection function, $\bar{g}_\ell = \frac{1}{n}\sum_{j=1}^n g_\ell(W_j)$, and the remainder $R^{(\ell)}_{n,i}$ collects all higher-order Hájek projections. Because the kernel ψ_ℓ is bounded (Assumption 2), all moments of g_ℓ are finite; in particular $\mathbb{E}[g_\ell(W)] = 0$ and $\mathrm{Var}\big(g_\ell(W)\big) = \zeta_1(\ell, \ell) < \infty$. Moreover, the remainder satisfies $\sum_{i=1}^n \big(R^{(\ell)}_{n,i}\big)^2 = O_p(n^{-3})$.

From (34) we obtain

$$(n-1)\sum_{i=1}^{n}\big(U_{n-1}^{(-i),\ell}-U_{n\ell}\big)^2$$
$$=\frac{m_\ell^2}{n-1}\sum_{i=1}^{n}\big(g_\ell(W_i)-\bar g_\ell\big)^2+\frac{2m_\ell}{n-1}\sum_{i=1}^{n}\big(g_\ell(W_i)-\bar g_\ell\big)R_{n,i}^{(\ell)}+(n-1)\sum_{i=1}^{n}\big(R_{n,i}^{(\ell)}\big)^2. \quad (35)$$

Because $\{g_\ell(W_i)\}$ are i.i.d., the first term on the right-hand side converges in probability to $m_\ell^2\,\zeta_1(\ell,\ell)$ by the weak law of large numbers. The Cauchy–Schwarz inequality and the rate of the remainder imply that the second term is $O_p\big((n-1)^{-1}\sqrt{n}\cdot\sqrt{n^{-3}}\big)=O_p(n^{-1})$, hence negligible. The third term is $(n-1)\cdot O_p(n^{-3})=O_p(n^{-2})$, also negligible. Consequently,

$$(n-1)\sum_{i=1}^{n}\big(U_{n-1}^{(-i),\ell}-U_{n\ell}\big)^2\xrightarrow{p}m_\ell^2\,\zeta_1(\ell,\ell). \quad (36)$$

An identical argument applied to the product of two different U-statistics yields the cross-term convergence

$$(n-1)\sum_{i=1}^{n}\big(U_{n-1}^{(-i),\ell}-U_{n\ell}\big)\big(U_{n-1}^{(-i),k}-U_{nk}\big)\xrightarrow{p}m_\ell m_k\,\zeta_1(\ell,k),\qquad 1\le\ell,k\le 4. \quad (37)$$

Step 2: Taylor expansion of the delete-one statistics. Let $T_{n-1}^{(-i)}=h\big(\mathbf{U}_{n-1}^{(-i)}\big)$ and $\bar T_{n-1}=\frac{1}{n}\sum_{i=1}^{n}T_{n-1}^{(-i)}$. Because h is a polynomial, it is twice continuously differentiable with bounded derivatives on any compact set. Since $\mathbf{U}_n$ is bounded in probability (each component is a bounded-kernel U-statistic), a first-order Taylor expansion around $\mathbf{U}_n$ gives

$$T_{n-1}^{(-i)}-T_n=\nabla h(\mathbf{U}_n)^\top\big(\mathbf{U}_{n-1}^{(-i)}-\mathbf{U}_n\big)+\frac{1}{2}\big(\mathbf{U}_{n-1}^{(-i)}-\mathbf{U}_n\big)^\top H_n\big(\mathbf{U}_{n-1}^{(-i)}-\mathbf{U}_n\big), \quad (38)$$

where H_n is the Hessian matrix of h evaluated at an intermediate point between $\mathbf{U}_n$ and $\mathbf{U}_{n-1}^{(-i)}$. For our specific $h(u)=u_1+u_2u_3-2u_4$, the only non-zero second-order partial derivative is $\partial^2h/\partial u_2\partial u_3=1$; all other entries of H_n vanish identically. Hence the quadratic term in (38) simplifies to

$$Q_{n,i}:=\big(U_{n-1}^{(-i),2}-U_{n2}\big)\big(U_{n-1}^{(-i),3}-U_{n3}\big)=O_p(n^{-2}). \quad (39)$$

Step 3: Asymptotic behaviour of the jackknife variance. Define the linear part $L_{n,i}:=\nabla h(\mathbf{U}_n)^\top(\mathbf{U}_{n-1}^{(-i)}-\mathbf{U}_n)$. From (38) we have $T_{n-1}^{(-i)}-T_n=L_{n,i}+Q_{n,i}$. Therefore

$$\hat\sigma_n^2=(n-1)\sum_{i=1}^{n}\big(T_{n-1}^{(-i)}-\bar T_{n-1}\big)^2$$
$$=(n-1)\sum_{i=1}^{n}\big(T_{n-1}^{(-i)}-T_n\big)^2-(n-1)\,n\,\big(\bar T_{n-1}-T_n\big)^2. \quad (40)$$

Because $\bar T_{n-1}-T_n=\frac{1}{n}\sum_{i=1}^{n}(L_{n,i}+Q_{n,i})$ and $\sum_{i=1}^{n}L_{n,i}=\nabla h(\mathbf{U}_n)^\top\sum_{i=1}^{n}(\mathbf{U}_{n-1}^{(-i)}-\mathbf{U}_n)$, while $U_{n-1}^{(-i),\ell}-U_{n\ell}=O_p(n^{-1})$ uniformly in i, we have $\bar T_{n-1}-T_n=O_p(n^{-1})$. Hence the second term in (40) is $(n-1)n\cdot O_p(n^{-2})=O_p(n^{-1})$, which tends to zero in probability.

It remains to evaluate the first term of (40):

$$(n-1)\sum_{i=1}^{n}\big(T_{n-1}^{(-i)}-T_n\big)^2$$
$$=(n-1)\sum_{i=1}^{n}L_{n,i}^2+2(n-1)\sum_{i=1}^{n}L_{n,i}Q_{n,i}+(n-1)\sum_{i=1}^{n}Q_{n,i}^2. \tag{41}$$

The linear term. Writing $\nabla h(\mathbf{U}_n)=(\hat{h}'_1,\dots,\hat{h}'_4)^\top$ and using (37), we obtain

$$(n-1)\sum_{i=1}^{n}L_{n,i}^2$$
$$=\sum_{\ell,k=1}^{4}\hat{h}'_\ell\,\hat{h}'_k\,(n-1)\sum_{i=1}^{n}\big(U_{n-1}^{(-i),\ell}-U_{n\ell}\big)\big(U_{n-1}^{(-i),k}-U_{nk}\big)$$
$$\xrightarrow{p}\sum_{\ell,k=1}^{4}h'_\ell\,h'_k\,m_\ell m_k\,\zeta_1(\ell,k)\;=\;\sigma^2(\Delta), \tag{42}$$

where the convergence follows from $\mathbf{U}_n\xrightarrow{p}\boldsymbol{\theta}$ (so $\hat{h}'_\ell\xrightarrow{p}h'_\ell$) and (37).

The cross and remainder terms. From (39), $Q_{n,i}=O_p(n^{-2})$, whence $(n-1)\sum_{i=1}^{n}Q_{n,i}^2=(n-1)n\cdot O_p(n^{-4})=O_p(n^{-2})$ and tends to zero. For the cross term, note that $L_{n,i}=O_p(n^{-1})$ uniformly in i because each $U_{n-1}^{(-i),\ell}-U_{n\ell}=O_p(n^{-1})$. Hence $2(n-1)\sum_{i=1}^{n}L_{n,i}Q_{n,i}=2(n-1)n\cdot O_p(n^{-1})\cdot O_p(n^{-2})=O_p(n^{-1})$, which also tends to zero in probability.

Step 4: Conclusion. Collecting the limits in (40)–(41), we have shown that every term except the linear contribution converges to zero, while the linear term converges to $\sigma^2(\Delta)$. Therefore

$$\hat{\sigma}_n^2\xrightarrow{p}\sigma^2(\Delta),$$

which establishes the consistency of the jackknife variance estimator. The boundedness of C_w (Assumption 2) guarantees that all kernels ψ_ℓ are bounded, so the projection functions g_ℓ possess finite moments of every order and the law-of-large-numbers arguments invoked above are valid without additional moment assumptions. □

Proof of Theorem 3.13. Each statistic is a smooth function of four U-statistics. Under Theorem 3.2(i), the twelve underlying U-statistics are jointly asymptotically normal by the multivariate CLT for U-statistics [Lee, 1990]. The smooth mapping theorem [van der Vaart, 1998] implies $(T_{\text{func}},T_{\text{param}},T_{\text{full}})$ is jointly normal with covariance expressed via the joint Hájek covariances defined in Section 3.7.

Covariance structure. For each pair (j,k), the asymptotic covariance σ_{jk} is obtained by substituting the gradient vectors of the three smooth mappings into the joint Hájek covariance matrix. Explicitly, for T_{func} and T_{param},

$$\sigma_{12}=4\sum_{\ell_1=1}^{4}\sum_{\ell_2=1}^{4}h'_{\ell_1}(\text{func})\,h'_{\ell_2}(\text{param})\,m_{\ell_1}m_{\ell_2}\,\zeta_1^{(12)}(\ell_1,\ell_2),$$

where $\zeta_1^{(12)}(\ell_1,\ell_2)=\text{Cov}(g_{\ell_1}^{(1)}(W),g_{\ell_2}^{(2)}(W))$ is the cross-statistic Hájek covariance between the projection functions of the T_{func} kernel and the T_{param} kernel. The remaining entries of Σ are defined analogously. □

Proof sketch. Because the functional part is correctly specified, the pseudo-true value (β_0,γ^*) minimises the population least-squares criterion. The normal equations give

$\Sigma_{\xi\xi}(\beta_0 - \beta^*) + \Sigma_{\xi Z}(\gamma_0 - \gamma^*) = 0$ and $\Sigma_{Z\xi}(\beta_0 - \beta^*) + \Sigma_{ZZ}(\gamma_0 - \gamma^*) = \mathbb{E}[Z\,\delta(Z)]$. Solving the second equation for $(\gamma_0 - \gamma^*)$ and substituting into the first yields $\beta_0 - \beta^* = \Sigma_{\xi\xi}^{-1}\Sigma_{\xi Z}\Sigma_{ZZ}^{-1}\mathbb{E}[Z\,\delta(Z)]$. Under Assumption 3, $\Sigma_{\xi\xi}$ has eigenvalues bounded away from zero, so $\|\beta_0 - \beta^*\|_{L^2} \lesssim \|\Sigma_{\xi Z}\|_{\mathrm{op}}\|\delta\|_{L^2}/\lambda_{\min}(\Sigma_{ZZ})$. The OLS estimator converges to the pseudo-true value at rate $O_p(n^{-1/2})$, giving the first claim. For the second claim, Theorem 3.3 shows that the leakage term is asymptotically equivalent to $n^{-1/2}\sum_{i=1}^n \phi(\varepsilon_i, \xi_i)\, X_i^\top(\beta_0 - \hat{\beta})$. Because ϕ is bounded (Proposition 3.4) and $\|\xi_i\| \le \sqrt{K}\|\tilde{X}_i\|$, the Lipschitz bound gives the stated inequality. □

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data and code availability

The Tecator dataset was recorded on a Tecator Infratec Food and Feed Analyzer operating in the wavelength range 850–1050 nm and was introduced by **?** as a benchmark for near-infrared calibration; it is publicly available in the `fda.usc` R package [**?**]. All simulation data were generated according to the data-generating processes described in Section 5.1 and can be reproduced from the reported parameter settings. The Python code used to run the simulations and produce the figures is available from the authors upon reasonable request.

Declarations

Ethics approval and consent to participate. Not applicable.

Consent for publication. Not applicable.

Clinical trial number. Not applicable.

References

J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer, 2nd edition, 2005.

L. Horváth and P. Kokoszka. *Inference for Functional Data with Applications*. Springer, New York, 2012.

G. Aneiros-Pérez and P. Vieu. Semi-functional partial linear regression. *Statistics & Probability Letters*, 76(11):1102–1110, 2006.

H. Shin. Partial functional linear regression. *Journal of Statistical Planning and Inference*, 139(10):3405–3418, 2009.

S. Novo and P. Vieu. Semi-functional partial linear regression with measurement error: an approach based on knn estimation. *TEST*, 2024. Advance online publication.

D. Ruppert, M. P. Wand, and R. J. Carroll. *Semiparametric Regression*. Cambridge University Press, Cambridge, 2003. ISBN 978-0-521-78050-6.

F. Ferraty and P. Vieu. *Nonparametric Functional Data Analysis: Theory and Practice*. Springer Series in Statistics. Springer, New York, 2006. ISBN 978-0-387-30369-7.

N. E. Heckman. Spline smoothing in a partly linear model. *Journal of the Royal Statistical Society: Series B (Methodological)*, 48(2):244–248, 1986.

P. Speckman. Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(3):413–436, 1988.

W. Härdle and E. Mammen. Comparing nonparametric versus parametric regression fits. *The Annals of Statistics*, 21(4):1926–1947, 1993.

J. X. Zheng. A consistent test of functional form via nonparametric estimation techniques. *Journal of Econometrics*, 75(2):263–289, 1996.

W. Stute. Nonparametric model checks for regression. *The Annals of Statistics*, 25(2): 613–641, 1997.

W. González-Manteiga and R. M. Crujeiras. An updated review of goodness-of-fit tests for regression models. *TEST*, 22(3):361–411, 2013. doi: 10.1007/s11749-013-0327-0.

J. C. Escanciano. A consistent diagnostic test for regression models using projections. *Econometric Theory*, 22(6):1030–1051, 2006.

K. Xu and Y. Zhou. Projection-averaging-based cumulative covariance and its use in goodness-of-fit testing for single-index models. *Computational Statistics & Data Analysis*, 164:107301, 2021. doi: 10.1016/j.csda.2021.107301.

J. A. Cuesta-Albertos, E. García-Portugués, M. Febrero-Bande, and W. González-Manteiga. Goodness-of-fit tests for the functional linear model based on randomly projected empirical processes. *The Annals of Statistics*, 47(1):439–467, 2019.

F. Chen, Q. Jiang, Z. Feng, and L. Zhu. Model checks for functional linear regression models based on projected empirical processes. *Computational Statistics & Data Analysis*, 144:106897, 2020.

E. García-Portugués, W. González-Manteiga, and M. Febrero-Bande. A goodness-of-fit test for the functional linear model with scalar response. *Journal of Computational and Graphical Statistics*, 23(3):761–778, 2014. doi: 10.1080/10618600.2013.812519.

E. García-Portugués, J. Álvarez-Liébana, G. Álvarez-Pérez, and W. González-Manteiga. Goodness-of-fit tests for functional linear models based on integrated projections. *arXiv preprint arXiv:2008.09885*, 2020. URL `https://arxiv.org/abs/2008.09885`.

N. Kloodt, N. Neumeyer, and I. Van Keilegom. Specification testing in semi-parametric transformation models. *TEST*, 30(4):980–1003, 2021.

F. Ferraccioli, L. M. Sangalli, and L. Finos. Nonparametric tests for semiparametric regression models. *TEST*, 33(2):573–597, 2024. doi: 10.1007/s11749-023-00868-9.

G. J. Székely and M. L. Rizzo. Energy statistics: a class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013.

F. Chen, S. G. Meintanis, and L. Zhu. On some characterizations and multidimensional criteria for testing homogeneity, symmetry and independence. *Journal of Multivariate Analysis*, 173:125–144, 2019.

F. Chen, S. G. Meintanis, and L. Zhu. Testing semiparametric model-equivalence hypotheses based on the characteristic function. *Journal of Statistical Computation and Simulation*, 94(6):1158–1190, 2024.

M. Krzyśko and Ł. Smaga. Two-sample tests for functional data using characteristic functions. *Austrian Journal of Statistics*, 50(4):53–64, 2021. doi: 10.17713/ajs.v50i4.1099. URL `https://www.ajs.or.at/index.php/ajs/article/view/1099`.

P. Hall and M. Hosseini-Nasab. On properties of functional principal components analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):109–126, 2006.

R. J. Serfling. *Approximation Theorems of Mathematical Statistics*. Wiley, New York, 1980.

Y. Zhou, Y. Zhang, and L. Zhu. A projective approach to conditional independence test for dependent processes. *Journal of Business & Economic Statistics*, 40(1):398–407, 2022. doi: 10.1080/07350015.2020.1826952.

Y. Zhou and F. Jiang. Heterogeneous functional regression for subgroup analysis. *Journal of Computational and Graphical Statistics*, 34(3):872–883, 2025. doi: 10.1080/10618600.2024.2414113.

W. Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948.

A. J. Lee. *U-Statistics: Theory and Practice*. Marcel Dekker, New York, 1990.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, 1998.

R. H. Randles. On the asymptotic normality of statistics with estimated parameters. *The Annals of Statistics*, 10(2):462–474, 1982.

H. L. Koul and W. Stute. Nonparametric model checks for time series. *The Annals of Statistics*, 27(1):204–236, 1999. doi: 10.1214/aos/1018031108.

E. L. Lehmann and J. P. Romano. *Testing Statistical Hypotheses*. Springer Texts in Statistics. Springer, New York, 3rd edition, 2005.