Semiparametric Testing with Highly Persistent Predictors*

Bas J.M. Werker[1] and Bo Zhou[2]

[1]Econometrics and Finance Group, Tilburg University

[2]Department of Economics and Finance, Durham University

Abstract

We address the issue of semiparametric efficiency in the bivariate regression problem with a highly persistent predictor, where the joint distribution of the innovations is regarded an infinite-dimensional nuisance parameter. Using a structural representation of the limit experiment and exploiting invariance relationships therein, we construct invariant point-optimal tests for the regression coefficient of interest. This approach naturally leads to a family of feasible tests based on the componentwise ranks of the innovations that can gain considerable power relative to existing tests under non-Gaussian innovation distributions, while behaving equivalently under Gaussianity. When an i.i.d. assumption on the innovations is appropriate for the data at hand, our tests exploit the efficiency gains possible. Moreover, we show by simulation that our test remains well behaved under some forms of conditional heteroskedasticity.

JEL classification: C12, C14

Keywords: predictive regression, limit experiment, LABF, maximal invariant, rank statistics.

*We thank Gaia Becheri for significant input on an earlier version of this paper. We also thank Peter Boswijk, Feike Drost, Ramon van den Akker, two referees, the associate editor, and participants at the European Conferences of the Econometrics Community (EC2) conference, Amsterdam, Dec 2017; Aarhus University, Sep 2018 for helpful comments.

1 Introduction

Over the past two decades, inference for the bivariate regression model with a highly persistent predictor has been well studied under the assumption of bivariate Gaussian innovations. Several procedures have been proposed in the econometric literature, see Cavanagh et al. (1995), Campbell and Yogo (2006), Jansson and Moreira (2006), Elliott et al. (2015), and Moreira and Mourão (2016). These inference procedures are all constructed based on the assumption of Gaussian innovations and, while their validity has been established under weaker assumptions, the asymptotic power of all these procedures cannot go beyond the Gaussian power envelope.

In the present paper we show that, when the application supports an additional assumption of serially independent innovations, sizable power gains are possible beyond the Gaussian power envelope. We establish this result by studying in detail the invariance structures that are present in the limiting experiment associated with the predictive regression model. This leads to a semiparametric power envelop which, under non-Gaussian innovation distributions, lies above the Gaussian power envelope. In that case, even without knowing the innovation distribution, our method dominates existing QMLE-based methods.

Our results precisely quantify the statistical efficiency gains from non-Gaussian innovation distributions when innovations are serially independent in predictive regression models. Under such, arguably restrictive assumption, we construct semiparametrically optimal (in a sense to be made precise later) tests. Whether in concrete applications the assumption of serially independence is warranted, is an empirical question. When it is, it can, as our results show, be exploited leading to sizable power gains (of, as Section 5 shows, up to 30% under Student-t_3 innovation distributions). Symmetrically, to make an informed choice, we study the behavior of our test when the innovations are not i.i.d. but exhibit conditional heteroskedasticity as often found in (financial) applications. Section 5.2 shows that, for the deviations studied, our test still has desirable size and power properties.

We note that our conceptual ideas reach further. We could, for instance, allow for serial dependence along the lines of Zhou et al. (2019) where an AR-type model on the error is imposed. Conditional heterogeneity could formally be addressed along the lines of Ling et al. (2003) where a GARCH-type structure on the error is imposed; or following Boswijk et al. (2005) where the (potentially nonstationary) volatility is estimated nonparametrically. These relaxations would technically be non-trivial and are left for future research. Note that, in view of the robustness-efficiency trade-off (see, e.g., Müller, 2011), an i.i.d. assumption on the innovations

ultimately driving the error term is not avoidable. Our test gives the empirical researchers an additional option: an improved power when innovations are i.i.d. and non-Gaussian.

The study of (optimal) semiparametric inference in the predictive regression model is complicated by the nonstandard asymptotic behavior induced by the local-to-unity asymptotics on the persistence parameter. More precisely, the associated likelihood ratios are of the *Locally Asymptotically Brownian Functional* (LABF) form in (see Jeganathan, 1995) and henceforth outside the conventional *Locally Asymptotically Normality* (LAN) world. As a consequence, the usual semiparametric approach based on projecting the score of the parameter of interest on the tangent space of nuisance scores is not straightforward. In particular, the model does not feature an adaptiveness property, which complicates its analysis. Jansson (2008) deals with the unit root testing problem, which also admits the LABF form, by guessing and then proving a least favorable direction of parametric submodels. An alternative approach has been proposed for the unit root testing problem in Zhou et al. (2019) and generalized to other common types of limiting experiments in Zhou (2020). In the present paper we apply these techniques to the predictive regression model.

The key idea is to exploit invariance structures in a so-called "structural" representation of the limit experiment. This approach sets us apart from most of the statistical and econometric literature where invariance arguments are used in the sequence of experiments. Instead, we obtain procedures which are invariant in the *limit* experiment, thereby making the analysis tractable and applicable to many models. Furthermore, the unique bivariate nature of the predictive regression model leads to a nonstandard multivariate structure in the associated limit experiment (see Theorem 3.1). Therefore, we present the approach in detail in the present paper.

Our contribution is twofold. First, we derive the semiparametric power envelope for (asymptotically) invariant tests in case the predictor's persistence level is assumed to be known, based on the structural LABF limit experiments. More precisely, Girsanov's theorem, combined with the limiting likelihood ratios for LABF experiments, leads to a description of the limit experiment by stochastic differential equations (SDEs). The observations in the limit experiment correspond to the limits of partial-sum processes of the innovations and score functions in the predictive regression model. In this structural representation of the limit experiment, we find that the nuisance parameters induced by the density function of the innovations only appear in the drifts of the driving Brownian motions. This leads to an invariance restriction by taking the Brownian bridges (which are invariant with respect to

these drifts) of these processes, and allows us to remove the nonparametric nuisance parameter (the density f of the innovations). We show that this also generates the *maximal invariant*. In this way, we avoid the problem of explicitly finding the least-favorable submodel. The likelihood of the maximal invariant immediately, by the Neyman-Pearson lemma, leads to the semiparametric power envelope.

Second, we propose a family of semiparametric feasible tests that has desirable properties. These tests are constructed using (asymptotically) sufficient statistics that are based on the increments of innovations, their component-wise ranks, and a pair of chosen marginal reference densities for both innovations including a reference correlation parameter. The ranks appear naturally as rank-based partial-sum score processes which weakly converge to the Brownian bridge that is invariant w.r.t. the density perturbation parameters. To further eliminate the remaining nuisance parameter, namely the predictor's persistence level, we employ the *Approximate Least Favorable Distribution* (ALFD) approach proposed by Elliott et al. (2015). We also follow their suggestion to switch to standard asymptotic approximations when the persistence parameter is far from unity. This helps to control the size of our tests under both non-stationarity and stationarity, see Appendix C. The tests thus obtained are semiparametric in the sense that, for all (fixed) innovation densities allowed, the asymptotic size is correct regardless of the choices of the marginal reference densities or the reference correlation.

Next to their validity, our test are more powerful than existing tests when the true innovation density is non-Gaussian. In particular, we compare our test to Elliott et al. (2015) (henceforth denoted as EMW), which is based on Gaussian likelihood ratios (see also Jansson and Moreira, 2006). Our asymptotic analysis using invariance arguments shows that, under non-Gaussian innovations, the EMW test actually is measurable with respect to an invariant in the limit that is not *maximally* invariant. As a result, under non-Gaussianity, we can construct tests that outperform the Gaussian power envelope and, thus, outperform the EMW test; see Remark 3.2. The power improvement depends on the choices of the marginal reference densities: when they are "closer" to the true marginal densities, we gain more power (and, again, while always having the desired size). Additionally, if one fixes the marginal reference densities to be Gaussian, our test is generally still more powerful than the EMW test under non-Gaussian innovation density; while under Gaussian innovation density, our test performs equivalently to the EMW test. This property is often referred to as the Chernoff-Savage result (see Chernoff and Savage (1958)). In the present LABF setting we have not been able to formally prove this Chernoff-Savage result, but our simulations indicate that this property nevertheless

may hold.

Our rank-based test can be regarded a generalized version of quasi-likelihood ratio tests which take the reference density to be Gaussian. The extra freedom to choose the reference density also comes with the cost of actually choosing it. However, we note that, in line with traditional quasi-likelihood methods, one can always choose the Gaussian reference density. Based on the classical Chernoff-Savage result, we conjecture that our rank-based procedure will then always outperform the quasi-likelihood procedure. This is confirmed by simulations and intuition, but, as discussed below, given the non-standard limiting experiment structure, we have not been able to prove this formally. Alternative, one could study a plug-in estimator where the reference density is nonparametrically estimated. We do not study this formally in the present paper; however, see Section 5.3 for some simulation results. Similarly, one may envision an approach where one pre-tests the residuals for, e.g., high kurtosis and chooses a references density based on that pre-test result.

The paper is organized as follows. In Section 2, we introduce the model and testing problem under consideration. In Section 3, we develop the asymptotic power envelope for test that are (asymptotically) invariant with respect to the innovation density f, assuming the predictor's persistence parameter γ is known. This development is based on the theory of limit experiments (see, e.g., Le Cam (1986) and Van der Vaart (2000)) and a structural version for models of LABF likelihood ratios (see Zhou et al. (2019)). In particular, this section explains where our power gains come from, see Remark 3.2. In Section 4, we employ the ALFD approach proposed by Elliott et al. (2015), among several available choices in the literature, to eliminate the nuisance parameter γ. In Section 5, we report large- and small-sample performances of our tests under both i.i.d. and conditional heteroskedastic errors. Section 6 concludes. All proofs are gathered in the appendix.

2 Model

Let y_t denote a random variable, observable at time t, that we wish to predict at time $t-1$ using an observable explanatory variable x_{t-1}. We consider the predictive regression model

$$y_t = \mu + \beta x_{t-1} + \varepsilon_t^y, \tag{1}$$

$$x_t - \alpha = \gamma(x_{t-1} - \alpha) + \varepsilon_t^x, \tag{2}$$

with $x_0 = 0$.[1] The parameter space is given by $\mu \in \mathbb{R}$, $\alpha \in \mathbb{R}$, $\beta \in \mathbb{R}$, and $\gamma \in (-1, 1]$. We have observations available for $t = 1, \ldots, T$.

Equation (2) features, along the lines of Cavanagh et al. (1995) and Jansson and Moreira (2006), an intercept α. However, as μ is a nuisance parameter in our model, the intercept α can be subsumed in μ without affecting inference on β. Indeed, our test statistics will only depend on the *increments* of x_t, denoted by Δx_t, and their associated ranks and, thus, they are invariant with respect to α. We therefore omit α in the rest of this paper.

To eliminate the nuisance intercept parameter μ in (1), one can directly impose an invariance restriction in the sequence of predictive regression experiments. For instance, the Jansson and Moreira (2006) test is based on the maximal invariant statistic $(y_2 - y_1, y_3 - y_1, \ldots, y_T - y_1)'$. In the present paper, our statistic is only based on y_t's through their ranks and, thus, also enjoys finite-sample invariance w.r.t. μ. To simplify notation, we set $\mu = 0$ throughout the paper and nowhere assume $\mathrm{E}_f(\varepsilon_t^y) = 0$. We will need to impose $\mathrm{E}_f(\varepsilon_t^x) = 0$: allowing for deterministic trends in x_t would lead to an entirely different asymptotic analysis.

Summarizing, as outlined in the introduction, we assume that the innovations $\varepsilon_t = (\varepsilon_t^y, \varepsilon_t^x)'$ are independent and identically distributed (i.i.d.) with (bivariate) density f satisfying the following condition.

Assumption 1. *(a)* $\mathrm{E}_f(\varepsilon_t^x) = 0$ *and* $\mathrm{Var}_f(\varepsilon_t) = \begin{pmatrix} \sigma_y^2 & \rho\sigma_y\sigma_x \\ \rho\sigma_y\sigma_x & \sigma_x^2 \end{pmatrix}$ *is a finite positive-definite matrix.*

(b) The density f is absolutely continuous with a.e. derivative $\dot{f} = \begin{pmatrix} \dot{f}_y \\ \dot{f}_x \end{pmatrix}$.

(c) The (standardized) Fisher information for location,

$$J_f = \begin{pmatrix} J_{f_{yy}} & J_{f_{yx}} \\ J_{f_{yx}} & J_{f_{xx}} \end{pmatrix} = \mathrm{E}_f\left(\ell_f \ell_f'\right),$$

where ℓ_f is the (standardized) location score function

$$\ell_f = \begin{pmatrix} \sigma_y \ell_{f_y} \\ \sigma_x \ell_{f_x} \end{pmatrix} = \begin{pmatrix} -\sigma_y \dot{f}_y / f \\ -\sigma_x \dot{f}_x / f \end{pmatrix},$$

[1] Note that this assumption on the initial value x_0 could possibly be relaxed to the weaker assumption $T^{-1/2} x_0 = o_{\mathrm{P}}(1)$ under $\beta = 0$ and $\gamma = 1$. One can possibly proceed along the lines of Müller and Elliott (2003); see also a remark on this point in Section 4 of Jansson and Moreira (2006). We keep the assumption $x_0 = 0$ for simplicity.

is finite.[2]

(d) $f > 0$. □

Let $\mathfrak{F}$ denote the set of densities satisfying Assumption 1.

The Fisher information J_f and scores ℓ_f for location are standardized in the sense that they are actually those related to ε_t^y/σ_y and ε_t^x/σ_x. As a result, ℓ_f and J_f do not depend on σ_y or σ_x. Note, however, that they both still depend on the correlation between the innovations ε_t^y and ε_t^x, i.e., they still depend on ρ.

We are interested in (optimal) tests for the (composite) null hypothesis

$$\mathrm{H}_0 : \beta = 0,\ \gamma \in (-1, 1],\ f \in \mathfrak{F}, \tag{3}$$

versus the one-sided alternative

$$\mathrm{H}_1 : \beta > 0,\ \gamma \in (-1, 1],\ f \in \mathfrak{F}. \tag{4}$$

As the literature focuses on test derived using an assumed Gaussian innovation density, we will throughout this paper consider Gaussian densities as a special case. This will allow us to make explicit where the power improvements come from in the case of non-Gaussian, serially independent, innovations $(\varepsilon^y, \varepsilon^x)$.

Remark 2.1 (Gaussian f). In case f is zero-mean bivariate Gaussian with correlation matrix $\mathbf{R} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, Assumption 1 is satisfied with $\ell_f(\varepsilon^y, \varepsilon^x) = \mathbf{R}^{-1} \begin{pmatrix} \varepsilon^y/\sigma_y \\ \varepsilon^x/\sigma_x \end{pmatrix}$ and $J_f = \mathbf{R}^{-1}$.

2.1 Local perturbations

Following the by now standard approach in the literature, we study the limit experiment in the sense of Hájek-Le Cam by considering local alternatives for all model parameters, that is, for both the parameter of interest β and the nuisance parameters (γ and f). For β and γ the appropriate rates of convergence are well known, see, e.g., Elliott and Stock (1994), Campbell and Yogo (2006), or Jansson and Moreira (2006). More precisely, we consider a T^{-1}-localization rate for β and γ, i.e.,

$$\beta = \beta^{(T)}(b) = \frac{b}{T}\frac{\sigma_y}{\sigma_x}, \quad \gamma = \gamma^{(T)}(c) = 1 + \frac{c}{T}, \tag{5}$$

[2] Being a Fisher information for location, J_f is automatically nonsingular and positive definite, see Mayer-Wolf et al. (1990, Theorem 2.3).

with $b \in \mathbb{R}$ and $c \in (-\infty, 0]$.[3] Observe that the local perturbation for b features a scaling by σ_y/σ_x. This ensures that the limit experiment will not depend on σ_y and σ_x (although it still depends on ρ).

The nuisance parameter f is infinite dimensional, so it is somewhat more involved to describe its relevant local perturbations. Introduce the separable Hilbert space

$$\mathrm{L}_2^{0,f} = \mathrm{L}_2^{0,f}(\mathbb{R}^2, \mathcal{B}) = \left\{ h \in \mathrm{L}_2^f(\mathbb{R}^2, \mathcal{B}) \mid \mathrm{E}_f h(\varepsilon) = 0,\ \mathrm{E}_f \varepsilon^x h(\varepsilon) = 0 \right\}, \tag{6}$$

where $\mathrm{L}_2^f(\mathbb{R}^2, \mathcal{B})$ denotes, the space of Borel-measurable functions $h : \mathbb{R}^2 \to \mathbb{R}$ satisfying $\mathrm{E}_f h^2(\varepsilon) = \int_{\mathbb{R}^2} h^2(\varepsilon) f(\varepsilon) \mathrm{d}\varepsilon < \infty$. The model assumption $\mathrm{E}_f(\varepsilon_t^x) = 0$ induces the restriction that local perturbations for f are orthogonal to the first component of ε: $\mathrm{E}_f \varepsilon^x h(\varepsilon) = 0$.

The separability of the Hilbert space $\mathrm{L}_2^{0,f}$ ensures the existence of a countable orthonormal basis h_k, $k \in \mathbb{N}$, such that each h_k is bounded and two times continuously differentiable with bounded derivatives; see, e.g., Rudin (1987, Theorem 3.14). Therefore, any function $h \in \mathrm{L}_2^{0,f}$ can be written as $h = \sum_{k=1}^\infty \eta_k h_k$, for some $\eta = (\eta_k)_{k\in\mathbb{N}} \in \ell_2 = \{(z_k)_{k\in\mathbb{N}} \mid \sum_{k=1}^\infty z_k^2 < \infty\}$. Besides the space ℓ_2, we also need the space c_{00} which is defined as the subset of sequences with finite support, i.e.,

$$c_{00} = \left\{ (z_k)_{k\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}} \,\middle|\, \sum_{k=1}^\infty 1\{z_k \neq 0\} < \infty \right\}. \tag{7}$$

Observe that c_{00} is a dense subspace of ℓ_2. It is introduced only in the asymptotic analysis to avoid convergence of infinite-dimensional processes and possibly induced mathematical complications, see Section 2.2. However, the restriction $\eta \in c_{00}$ will not affect our conclusions. Indeed, considering $\eta \in c_{00}$ restricts our analysis to a subset of all semiparametric models which potentially makes the obtained upper bound higher. However, as we are able to show that this higher upper bound is (point-wisely) attainable by feasible tests for arbitrary innovation density in sequence, see Remark 3.1, it constitutes the semiparametric power envelope and the test is semiparametrically optimal.

We model local perturbations to the innovation density f as

$$f_\eta^{(T)}(e) = f(e)\left(1 + \frac{1}{\sqrt{T}} \sum_{k=1}^\infty \eta_k h_k(e)\right) \text{ for all } e \in \mathbb{R}^2, \tag{8}$$

[3] We use here the common approach in the literature to restrict the nuisance parameter c to $(-\infty, 0]$. We conjecture that all results remain valid, with the obvious modifications, in case one would choose the larger parameter space $c \in \mathbb{R}$; see, e.g., Moreira and Mourão (2016).

where $\eta \in c_{00}$. We thus use a standard localization rate $T^{-1/2}$ for the bivariate density f. Indeed, Proposition 3.1 below shows that all the above rates are appropriate in the sense that they lead to contiguous alternatives for the induced probability measures as T tends to infinity.

In order to show that the above localization of the innovation density is valid, we need to establish that $f_\eta^{(T)} \in \mathfrak{F}$. This is the content of the next proposition.

Proposition 2.1. *Let $f \in \mathfrak{F}$ and $\eta \in c_{00}$, then there exists a finite integer $\widetilde{T}$ such that for all $T \geq \widetilde{T}$ we have $f_\eta^{(T)} \in \mathfrak{F}$.*

The proof uses exactly the same arguments as in the proof of Proposition 3.1 in Zhou et al. (2019), but with support $\mathbb{R}^2$ instead of $\mathbb{R}$. It is therefore omitted.

In terms of the local parameters b, c, and η, the hypothesis of interest becomes

$$\mathrm{H}_0 : b = 0,\, c \in \mathbb{R},\, \eta \in c_{00}, \tag{9}$$

versus the one-sided alternative

$$\mathrm{H}_1 : b > 0,\, c \in \mathbb{R},\, \eta \in c_{00}. \tag{10}$$

2.2 Partial-sum processes

In order to derive the limiting experiment for the predictive regression model, we need to introduce some partial-sum processes and study their asymptotic behavior. We denote by $\mathrm{P}_{b,c,\eta;f}^{(T)}$ the law of $(y_1, x_1)', \ldots, (y_T, x_T)'$ under the model (1)–(2), where the parameters β and γ are given by (5) and the innovation density is given by (8). Formally, we define the sequence of experiments of interest as

$$\mathcal{E}^{(T)}(f) := \left(\Omega^{(T)}, \mathcal{F}^{(T)}, \left\{\mathrm{P}_{b,c,\eta;f}^{(T)} : b, c \in \mathbb{R}, \eta \in c_{00}\right\}\right), \quad T \in \mathbb{N}, \tag{11}$$

where $\Omega^{(T)} := \mathbb{R}^{2\times T}$ and $\mathcal{F}^{(T)} := \mathcal{B}(\mathbb{R}^{2\times T})$. We denote the expectation taken under the measure $\mathrm{P}_{0,0,0;f}^{(T)}$ by $\mathrm{E}^{(T)}$.

Let us already mention that we will also introduce a collection of probability measures $\mathbb{P}_{b,c,\eta}$, defined on a probability space $(\Omega, \mathcal{F})$, representing the limit experiment $\mathcal{E}(f)$ in Section 3.1 below; see (26). We will denote he expectation taken under the measure $\mathbb{P}_{0,0,0}$ by $\mathbb{E}$. That is, $\mathrm{P}^{(T)}$ and $\mathrm{E}^{(T)}$ refer to finite-sample distributions in the sequence of experiments, while $\mathbb{P}$ and $\mathbb{E}$ refer to distributions in the limit experiment.

As a final ingredient for our analysis, we introduce some partial-sum processes that we use throughout to link the sequence of experiments $\mathcal{E}^{(T)}(f)$ to the limit experiment $\mathcal{E}(f)$. In particular, define, with $\Delta x_t := x_t - x_{t-1}$, the partial-sum

processes[4]

$$W_{\varepsilon}^{(T)}(s) := \frac{1}{\sqrt{T}} \sum_{t=1}^{\lfloor sT \rfloor} \frac{\Delta x_t}{\sigma_x}, \tag{12}$$

$$W_{\ell_{f_y}}^{(T)}(s) := \frac{1}{\sqrt{T}} \sum_{t=1}^{\lfloor sT \rfloor} \sigma_y \ell_{f_y}(y_t, \Delta x_t), \tag{13}$$

$$W_{\ell_{f_x}}^{(T)}(s) := \frac{1}{\sqrt{T}} \sum_{t=1}^{\lfloor sT \rfloor} \sigma_x \ell_{f_x}(y_t, \Delta x_t), \tag{14}$$

$$W_{h_k}^{(T)}(s) := \frac{1}{\sqrt{T}} \sum_{t=1}^{\lfloor sT \rfloor} h_k(y_t, \Delta x_t), \quad k \in \mathbb{N}. \tag{15}$$

Here we standardize the first three partial-sum processes by the standard deviations σ_y and σ_x in order to make their limits scale invariant. Under $\mathrm{P}_{0,0,0;f}^{(T)}$, by the Functional Central Limit Theorem (see also Lemma A.1), we have

$$\begin{pmatrix} W_{\varepsilon}^{(T)}(s) \\ W_{\ell_{f_y}}^{(T)}(s) \\ W_{\ell_{f_x}}^{(T)}(s) \\ W_{h}^{(T)}(s) \end{pmatrix} \Rightarrow \begin{pmatrix} W_{\varepsilon}(s) \\ W_{\ell_{f_y}}(s) \\ W_{\ell_{f_x}}(s) \\ W_{h}(s) \end{pmatrix}, \quad s \in [0,1], \tag{16}$$

where the Brownian motions W_{ε}, $W_{\ell_{f_y}}$, $W_{\ell_{f_x}}$ and W_h are defined on the common probability space $(\Omega, \mathcal{F}, \mathbb{P}_{0,0,0})$. We have to be precise about the notion of weak convergence adopted in (16) as W_h is infinite dimensional. In line with stochastic process theory, we mean that all finite-dimensional subprocesses of $W_h^{(T)}$ weakly converges in the space $D^{M+3}[0,1]$ with the uniform topology, where M is the dimension of the finite-dimensional subprocess considered. This is precisely because we take the local parameter η to be in c_{00}. For the sake of convenient notation, we write the seemingly infinite-dimensional convergence (16). As argued above, we are ultimately able to attain the semiparametric power envelope induced under the restriction $\eta \in c_{00}$ so that we can claim semiparametric optimality.

Next, define the column vectors $J_{f_y h} = (J_{f_y h_k})_{k \in \mathbb{N}}$ and $J_{f_x h} = (J_{f_x h_k})_{k \in \mathbb{N}}$, where $J_{f_y h_k} := \mathrm{E}_f \left[\sigma_y \ell_{f_y}(\varepsilon_t) h_k(\varepsilon_t)\right]$ and $J_{f_x h_k} := \mathrm{E}_f \left[\sigma_x \ell_{f_x}(\varepsilon_t) h_k(\varepsilon_t)\right]$. As we have the equalities $\mathrm{E}_f \left[\varepsilon_t^x \ell_{f_y}(\varepsilon_t)\right] = -\sigma_y \int_{\mathbb{R}^2} \varepsilon^x \frac{\dot{f}_y(\varepsilon)}{f(\varepsilon)} f(\varepsilon) \mathrm{d}\varepsilon = -\sigma_y \int_{\mathbb{R}^2} \varepsilon^x \dot{f}_y(\varepsilon) \mathrm{d}\varepsilon = 0$ and $\mathrm{E}_f \left[\varepsilon_t^x \ell_{f_x}(\varepsilon_t)\right] = -\sigma_x \int_{\mathbb{R}^2} \varepsilon^x \frac{\dot{f}_x(\varepsilon)}{f(\varepsilon)} f(\varepsilon) \mathrm{d}\varepsilon = -\sigma_x \int_{\mathbb{R}^2} \varepsilon^x \dot{f}_x(\varepsilon) \mathrm{d}\varepsilon = \sigma_x \int_{\mathbb{R}^2} f(\varepsilon) \mathrm{d}\varepsilon = \sigma_x$, the behavior of the Brownian motions W_{ε}, $W_{\ell_{f_y}}$, $W_{\ell_{f_x}}$ and W_h is described by the

[4] One may consider partial sum processes that start at $t = 2$ in order to make them exactly invariant to translations in x_t. This would, clearly, have no effect on our asymptotic results.

covariance matrix

$$\operatorname{Var}\begin{pmatrix} W_\varepsilon(1) \\ W_{\ell_{f_y}}(1) \\ W_{\ell_{f_x}}(1) \\ W_h(1) \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & J_{f_{yy}} & J_{f_{yx}} & J'_{f_y h} \\ 1 & J_{f_{yx}} & J_{f_{xx}} & J'_{f_x h} \\ 0 & J_{f_y h} & J_{f_x h} & I_\infty \end{pmatrix}, \tag{17}$$

where I_∞ denotes the ∞-dimensional identity matrix. The scaling by σ_x and σ_y introduced in (12)–(15) is indeed such that the covariance matrix (17) does not depend on σ_x or σ_y. Again, it still depends on ρ through the various J matrices.

Recall that the functions h_k form an orthonormal basis for all zero-mean finite-variance functions that are orthogonal to ε_t^x. In view of the covariance matrix (17), we may thus write, for $s \in [0,1]$,

$$W_{\ell_{f_y}}(s) = J'_{f_y h} W_h(s), \tag{18}$$

$$W_{\ell_{f_x}}(s) = W_\varepsilon(s) + J'_{f_x h} W_h(s). \tag{19}$$

Consequently, we also have

$$\operatorname{Var}\left[W_{\ell_{f_y}}(1)\right] = J_{f_{yy}} = J'_{f_y h} J_{f_y h}, \tag{20}$$

$$\operatorname{Var}\left[W_{\ell_{f_x}}(1)\right] = J_{f_{xx}} = 1 + J'_{f_x h} J_{f_x h}, \tag{21}$$

$$\operatorname{Cov}\left[W_{\ell_{f_y}}(1), W_{\ell_{f_x}}(1)\right] = J_{f_{yx}} = J'_{f_y h} J_{f_x h}. \tag{22}$$

We again consider the special case of a Gaussian density f.

Remark 2.2 (Gaussian f). In the situation of Gaussian f as discussed in Remark 2.1, we may write the decomposition (21) as $W_{\ell_{f_x}} = W_\varepsilon - \frac{\rho}{\sqrt{1-\rho^2}} W_\perp$ where $W_\perp$ is the standard Brownian motion generated by the increments $\left(\varepsilon^y/\sigma_y - \rho\varepsilon^x/\sigma_x\right)/\sqrt{1-\rho^2}$. Indeed, W_ε and $W_\perp$ are independent (calculate the correlation of the increments that generate both processes). Thus, we also find $J'_{f_x h} W_h(s) = -\rho W_\perp$ and the decomposition (21) becomes $J_{f_{xx}} = 1 + \frac{\rho^2}{1-\rho^2} = \frac{1}{1-\rho^2} = J_{f_{yy}}$. Moreover, we have $W_{\ell_{f_y}} = \frac{1}{\sqrt{1-\rho^2}} W_\perp$ and $J_{f_{yx}} = -\frac{\rho}{1-\rho^2}$.

3 Eliminating the nuisance parameter f by invariance

We first focus on eliminating the nuisance parameter f from the testing problem outlined in Section 2. We will see that this can be handled using invariance arguments in the *limit* experiment, which we derive in Section 3.1. In Section 4, we consider the nuisance parameter γ.

We take the following steps in this section:

1. Provide a structural representation of the limit experiment (Section 3.1).
2. Characterize maximally invariant test statistics in this limit experiment (Section 3.2).
3. Provide a structural representation of the invariant limit experiment (Section 3.3).
4. Provide a feasible version of the asymptotically invariant test statistics to be applied in the sequence of predictive regression experiments (Section 3.4).

These steps also show that, to eliminate the nuisance parameter f, instead of studying invariance restrictions in the *sequence* of finite-sample experiments, we only impose them in the *limit* experiment. Unlike for the location parameter μ (of ε_t^y), this limiting invariance property of the parameter f does not follow directly from exact finite-sample invariance properties. Notably, the existing tests in the literature share this feature, as they also (implicitly) impose the invariance restriction in the limit, though not in the sequence; see Remark 3.2. As far as we know, all existing tests belong to the class of asymptotically invariant (w.r.t. f) tests, while our test is semiparametrically optimal in the model we study. Section 5 shows that this approach leads to considerable power gains in case the innovations are non-Gaussian, while no power is lost under Gaussianity.

3.1 A Structural Representation of the Limit Experiment

We consider the limit experiment corresponding to the predictive regression model (1)–(2) using the local perturbations (5) and (8), i.e., the limit of the experiments $\mathcal{E}^{(T)}(f)$ indexed by T, by studying the asymptotic behavior of the induced likelihood ratios. We expand the likelihood ratio around $(\beta, \gamma, \eta) = (0, 1, 0)$ and derive its limit in the following proposition, which can be interpreted as a generalization of Lemma 4 in Jansson and Moreira (2006) by including non-Gaussian distributions and perturbations thereof.[5]

Proposition 3.1. *Fix $f \in \mathfrak{F}$. Consider the local parameters $b \in \mathbb{R}$, $c \in \mathbb{R}$, and $\eta \in c_{00}$. Then,*

(i) Under $\mathrm{P}^{(T)}_{0,0,0;f}$, the log-likelihood ratio of the predictive regression experiment satisfies, as $T \to \infty$,

$$\log \frac{\mathrm{dP}^{(T)}_{b,c,\eta;f}}{\mathrm{dP}^{(T)}_{0,0,0;f}} = \Delta^{(T)}(b, c, \eta) - \frac{1}{2}\mathcal{Q}^{(T)}(b, c, \eta) + o_{\mathrm{P}}(1), \tag{23}$$

[5] As preparation for the results in Section 4, we allow in this proposition for local perturbations with respect to γ even though, in the present section, γ is assumed to be known.

where

$$\begin{aligned}\Delta^{(T)}(b,c,\eta) = {} & \frac{b}{T}\sum_{t=1}^{T}\frac{x_{t-1}}{\sigma_x}\sigma_y\ell_{f_y}(y_t,\Delta x_t) + \frac{c}{T}\sum_{t=1}^{T}x_{t-1}\ell_{f_x}(y_t,\Delta x_t)\\ & + \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\sum_{k}\eta_k h_k(y_t,\Delta x_t),\\ \mathcal{Q}^{(T)}(b,c,\eta) = {} & \left(b^2 J_{f_{yy}} + c^2 J_{f_{xx}} + 2bcJ_{f_{yx}}\right)\frac{1}{T^2}\sum_{t=1}^{T}\frac{x_{t-1}^2}{\sigma_x^2}\\ & + \left(2bJ'_{f_y h}\eta + 2cJ'_{f_x h}\eta\right)\frac{1}{T^{3/2}}\sum_{t=1}^{T}\frac{x_{t-1}}{\sigma_x} + \eta'\eta.\end{aligned}$$

(ii) Still under $\mathrm{P}^{(T)}_{0,0,0;f}$*, as* $T\to\infty$*, we have*

$$\log\frac{\mathrm{dP}^{(T)}_{b,c,\eta;f}}{\mathrm{dP}^{(T)}_{0,0,0;f}} \Rightarrow \mathcal{L}(b,c,\eta) = \Delta(b,c,\eta) - \frac{1}{2}\mathcal{Q}(b,c,\eta), \tag{24}$$

where

$$\begin{aligned}\Delta(b,c,\eta) &= b\int_0^1 W_\varepsilon(s)\mathrm{d}W_{\ell_{f_y}}(s) + c\int_0^1 W_\varepsilon(s)\mathrm{d}W_{\ell_{f_x}}(s) + \eta' W_h(1)\\ &= \int_0^1 W_\varepsilon(s)\left(bJ_{f_y h} + cJ_{f_x h}\right)'\mathrm{d}W_h(s) + c\int_0^1 W_\varepsilon(s)\mathrm{d}W_\varepsilon(s) + \eta' W_h(1),\\ \mathcal{Q}(b,c,\eta) &= \left(b^2 J_{f_{yy}} + c^2 J_{f_{xx}} + 2bcJ_{f_{yx}}\right)\int_0^1 W_\varepsilon(s)^2\mathrm{d}s\\ &\quad + \eta'\eta + \left(2bJ'_{f_y h}\eta + 2cJ'_{f_x h}\eta\right)\int_0^1 W_\varepsilon(s)\mathrm{d}s\\ &= \int_0^1 \left|\left(bJ_{f_y h} + cJ_{f_x h}\right)W_\varepsilon(s) + \eta\right|^2\mathrm{d}s + c^2\int_0^1 W_\varepsilon(s)^2\mathrm{d}s.\end{aligned}$$

(iii) For every $b,c\in\mathbb{R}$ *and* $\eta\in c_{00}$*, under* $\mathbb{P}_{0,0,0}$*,* $\mathbb{E}[\exp(\mathcal{L}(b,c,\eta))] = 1$*.*

A proof of Proposition 3.1 is provided in Appendix B, but let us give a brief sketch here. Part (i) is immediate from an informal Taylor expansion of the log-likelihood ratios and, formally, follows from Hallin et al. (2015), which provides generally applicable sufficient conditions for the quadratic expansion of likelihood ratios with densities that are differentiable in quadratic mean (DQM). This DQM condition is implied, for location models, by the absolutely continuity of the innovation density function and finiteness of the associated Fisher information, i.e., precisely the content of Assumption 1. A detailed discussion can be found in Le Cam (1986, Section 17.3) or Yang and Le Cam (2000, Section 7.3). Part (ii) follows from the continuous mapping theorem applied to the weak convergence in (16). Both forms of the central sequence Δ and quadratic term $\mathcal{Q}$ follow from (18) and (19). Part (iii) follows from standard stochastic calculations concerning Doléans-Dade exponentials. To see this, note that W_h and W_ε are independent in view of (17) and, thus, have vanishing quadratic covariation.

Part (iii) of Proposition 3.1 ensures that we can introduce a collection of probability measures $\mathbb{P}_{b,c,\eta}$ on the measurable space $(\Omega, \mathcal{F})$ (on which the Brownian motions W_ε, $W_{\ell_{f_y}}$, $W_{\ell_{f_x}}$ and W_h are defined) by the Radon-Nikodym derivative

$$\frac{\mathrm{d}\mathbb{P}_{b,c,\eta}}{\mathrm{d}\mathbb{P}_{0,0,0}} = \exp \mathcal{L}(b, c, \eta), \tag{25}$$

where $\mathcal{L}(b, c, \eta)$ is defined in (24). Then, in the sense of Hájek-Le Cam (see, for instance, Van der Vaart (2000), Chapter 9), the sequence of predictive regression experiments, indexed by sample size T, weakly converges to the limit experiment described by the measures $\mathbb{P}_{b,c,\eta}$. We formally define this limit experiment by

$$\mathcal{E}(f) := \left(\Omega, \mathcal{F}, \left\{\mathbb{P}_{b,c,\eta} : b, c \in \mathbb{R}, \eta \in c_{00}\right\}\right), \tag{26}$$

where $\Omega := C[0,1] \times C[0,1] \times C[0,1] \times C^{\mathbb{N}}[0,1]$ and $\mathcal{F} := \mathcal{B}_{\mathcal{C}} \otimes \mathcal{B}_{\mathcal{C}} \otimes \mathcal{B}_{\mathcal{C}} \otimes (\otimes_{k=1}^{\infty} \mathcal{B}_{\mathcal{C}})$.

The following statement is an immediate consequence of Proposition 3.1.

Corollary 3.1. *Let $f \in \mathfrak{F}$, then the sequence of experiments $\mathcal{E}^{(T)}(f)$ converges to the limit experiment $\mathcal{E}(f)$ as $T \to \infty$.*

Although the log-likelihood ratios $\mathcal{L}(b, c, \eta)$ formally describe the limiting experiment, it is more insightful to provide, what we call, a structural representation. This structural representation provides a fixed-horizon continuous-time model for which the likelihoods are exactly equal to $\exp(\mathcal{L}(b, c, \eta))$. From a statistical point of view, the induced experiments are thus equal. The result follows from an immediate application of Girsanov's theorem to the Radon-Nikodym derivates (24). Its proof is therefore omitted.

Theorem 3.1. *Fix $f \in \mathfrak{F}$. Let, under $\mathbb{P}_{0,0,0}$, Z_ε, and Z_h be zero-drift Brownian motions with covariance according to the first and last row and column of (17). The limit experiment $\mathcal{E}(f)$ can be described as: observe $\{(W_\varepsilon(s), W_h(s)) : s \in [0,1]\}$ generated by*

$$\mathrm{d}W_\varepsilon(s) = cW_\varepsilon(s)\mathrm{d}s + \mathrm{d}Z_\varepsilon(s), \tag{27}$$

$$\mathrm{d}W_h(s) = (bJ_{f_y h} + cJ_{f_x h})W_\varepsilon(s)\mathrm{d}s + \eta \mathrm{d}s + \mathrm{d}Z_h(s). \tag{28}$$

A few remarks can be made in relation to Theorem 3.1. First, note that for $b = c = 0$ and $\eta = 0$, we obtain $W_\varepsilon = Z_\varepsilon$ and $W_h = Z_h$. Secondly, the theorem essentially states that while $(W_\varepsilon, W_h')'$ is a zero-drift Brownian motion under $\mathbb{P}_{0,0,0}$, it becomes an Ornstein-Uhlenbeck process under $\mathbb{P}_{b,c,\eta}$, where the log-likelihood ratio $\log(\mathrm{d}\mathbb{P}_{b,c,\eta}/\mathrm{d}\mathbb{P}_{0,0,0})$ equals $\mathcal{L}(b, c, \eta)$. Observe in particular that local perturbations of the innovation density f, as described by η, only affect the drift in (28). We will consider inference procedures that are invariant with respect to η in the

limit experiment. In terms of the (sequence of) predictive regression model(s) this consequently translates into invariance with respect to (local perturbations in) the innovation density f.

In view of (18)–(19), we may also write

$$\mathrm{d}W_{\ell_{f_y}}(s) = (bJ_{f_{yy}} + cJ_{f_{yx}})W_\varepsilon(s)\mathrm{d}s + J'_{f_yh}\eta\mathrm{d}s + \mathrm{d}Z_{\ell_{f_y}}(s), \tag{29}$$

$$\mathrm{d}W_{\ell_{f_x}}(s) = (bJ_{f_{yx}} + cJ_{f_{xx}})W_\varepsilon(s)\mathrm{d}s + J'_{f_xh}\eta\mathrm{d}s + \mathrm{d}Z_{\ell_{f_x}}(s), \tag{30}$$

where $Z_{\ell_{f_x}}$ and $Z_{\ell_{f_y}}$ are zero-drift Brownian motions under $\mathbb{P}_{0,0,0}$. However, these equations do not contain any additional information, precisely given (18) and (19). Nevertheless, they will turn out useful when describing the likelihood ratio of the maximal invariant $\mathcal{M}$ to be introduced below in (33).

3.2 Maximal Invariant

In the limit experiment $\mathcal{E}(f)$, the parameter $b \in \mathbb{R}$ is the parameter of interest, while $c \in \mathbb{R}$ and $\eta \in c_{00}$ are nuisance parameters. Observe that the nuisance parameter η appears only in the drift of the SDEs in Theorem 3.1. This suggests an invariance restriction in line with the approach in Zhou et al. (2019) for unit root testing.

To be specific, we first introduce, for $\eta \in c_{00}$, the transformations $\mathfrak{g}_\eta : C^{\mathbb{N}}[0,1] \to C^{\mathbb{N}}[0,1]$ by

$$[\mathfrak{g}_\eta(W)](s) = W(s) - \eta s, \tag{31}$$

for $W \in C^{\mathbb{N}}[0,1]$ and all $s \in [0,1]$. The transformation $\mathfrak{g}_\eta$ adds a drift $s \mapsto -\eta s$ to W. Thus, Theorem 3.1 implies that the law of $(W_\varepsilon, (\mathfrak{g}_\eta(W_h))')'$ under $\mathbb{P}_{b,c,0}$ is the same as the law of $(W_\varepsilon, W'_h)'$ under $\mathbb{P}_{b,c,\eta}$.[6] Denote by $\mathfrak{G}_\eta$ the group of transformations $\mathfrak{g}_\eta$ for $\eta \in c_{00}$. We can now characterize the maximal invariant with respect to $\mathfrak{G}_\eta$ in the limit experiment $\mathcal{E}(f)$.

For any process W, we define the associated *bridge process* by

$$B^W(s) := W(s) - sW(1), \tag{32}$$

for all $s \in [0,1]$. Then, one readily verifies

$$\begin{aligned} B^{\mathfrak{g}_\eta(W)}(s) &= [\mathfrak{g}_\eta(W)](s) - s[\mathfrak{g}_\eta(W)](1) \\ &= W(s) - \eta s - s(W(1) - \eta) \\ &= W(s) - sW(1) \\ &= B^W(s). \end{aligned}$$

[6] By (19) and (18), the same holds for $W_{\ell_{f_x}}$ and $W_{\ell_{f_y}}$.

As a result, the bridges B^{W_h} are invariant under the transformations $\mathfrak{g}_\eta$.

Define the mapping M by $M(W_\varepsilon, W_h) := (W_\varepsilon, B^{W_h})$. It then follows that statistics that are measurable with respect to the σ-field

$$\mathcal{M} = \sigma\left(M(W_\varepsilon, W_h)\right) = \sigma\left(W_\varepsilon, B^{W_h}\right), \tag{33}$$

are invariant with respect to $\mathfrak{g}_\eta$ for all $\eta \in c_{00}$. Moreover, in the following theorem, we show $\mathcal{M}$ to be *maximally* invariant. Its proof is, again, provided in Appendix B.

Theorem 3.2. *In the limit experiment $\mathcal{E}(f)$, for $\eta \in c_{00}$, the σ-field $\mathcal{M}$ in (33) is maximally invariant with respect to $\mathfrak{G}_\eta$.*

3.3 A Structural Representation of the Invariant Limit Experiment

Theorem 3.2 implies that any inference invariant with respect to $\mathfrak{G}_\eta$ must be measurable with respect to $\mathcal{M}$; see, e.g., Lehmann and Romano (2006, Theorem 6.2.1). Therefore, by the Neyman-Pearson lemma, inference based on the likelihood ratio with respect to $\mathcal{M}$ yields the power envelope for invariant tests in the limit experiment $\mathcal{E}(f)$. The following result provides this likelihood ratio.

Theorem 3.3. *Fix $f \in \mathfrak{F}$. Then the likelihood ratios in the limit experiment $\mathcal{E}(f)$ restricted to the maximal invariant $\mathcal{M}$ are given by*

$$\exp \mathcal{L}_{\mathcal{M}}(b,c) := \frac{\mathrm{d}\mathbb{P}^{\mathcal{M}}_{b,c}}{\mathrm{d}\mathbb{P}^{\mathcal{M}}_{0,0}} = \mathbb{E}\left[\frac{\mathrm{d}\mathbb{P}_{b,c,\eta}}{\mathrm{d}\mathbb{P}_{0,0,0}}|\mathcal{M}\right] = \exp\left(\Delta_{\mathcal{M}}(b,c) - \frac{1}{2}\mathcal{Q}_{\mathcal{M}}(b,c)\right), \tag{34}$$

where

$$\begin{aligned}
\Delta_{\mathcal{M}}(b,c) &= \int_0^1 W_\varepsilon(s)\left(bJ_{f_y h} + cJ_{f_x h}\right)' \mathrm{d}B^{W_h}(s) + c\int_0^1 W_\varepsilon(s)\mathrm{d}W_\varepsilon(s) \qquad (35)\\
&= b\int_0^1 W_\varepsilon(s)\mathrm{d}B_{\ell_{f_y}}(s) + c\left(\int_0^1 W_\varepsilon(s)\mathrm{d}B_{\ell_{f_x}}(s) + W_\varepsilon(1)\overline{W_\varepsilon}\right),\\
\mathcal{Q}_{\mathcal{M}}(b,c) &= \left(bJ_{f_y h} + cJ_{f_x h}\right)^2 \int_0^1 \left(W_\varepsilon(s) - \overline{W_\varepsilon}\right)^2 \mathrm{d}s + c^2\int_0^1 W_\varepsilon(s)^2\mathrm{d}s \qquad (36)\\
&= \left(b^2 J_{f_{yy}} + c^2(J_{f_{xx}} - 1) + 2bcJ_{f_{yx}}\right)\left(\overline{W_\varepsilon^2} - (\overline{W_\varepsilon})^2\right) + c^2\left(\overline{W_\varepsilon}\right)^2,
\end{aligned}$$

with $\overline{W_\varepsilon^2} = \int_0^1 W_\varepsilon(s)^2\mathrm{d}s$ and $\overline{W_\varepsilon} = \int_0^1 W_\varepsilon(s)\mathrm{d}s$.

The proof is provided in Appendix B. The first ways to write $\Delta_{\mathcal{M}}(b,c)$ and $\mathcal{Q}_{\mathcal{M}}(b,c)$ make explicit that the likelihood factorizes in a conditional likelihood given W_ε and the marginal likelihood of W_ε. Both second ways to write $\Delta_{\mathcal{M}}(b,c)$ and $\mathcal{Q}_{\mathcal{M}}(b,c)$ follow from (18)–(19) and (20)–(22). Those are the versions that we use below to construct our feasible test statistics. Theorem 3.3 also immediately

yields the semiparametric power envelope, still for fixed c, that we do not present in detail for brevity.

The restriction to invariant tests removes the nuisance parameter η from the testing problem. Indeed, the likelihood ratio (34) no longer depends on η. Therefore, we can formally define the limit experiment restricted to the maximal invariance $\mathcal{M}$ as

$$\mathcal{E}_{\mathcal{M}}(f) := \left(\Omega, \mathcal{M}, \left\{\mathbb{P}^{\mathcal{M}}_{b,c} : b, c \in \mathbb{R}\right\}\right). \tag{37}$$

Again, the likelihood ratios $\mathrm{d}\mathbb{P}^{\mathcal{M}}_{b,c}/\mathrm{d}\mathbb{P}^{\mathcal{M}}_{0,0}$ can also be interpreted as Girsanov transformations. We state this as a corollary as the result follows immediately from calculating the bridges corresponding to $W_{\ell_{f_y}}$ and $W_{\ell_{f_x}}$ in Theorem 3.3.

Corollary 3.2. *Fix $f \in \mathfrak{F}$. Let, under $\mathbb{P}^{\mathcal{M}}_{0,0}$, Z_ε and Z_h be zero-drift Brownian motions with covariance according to the first and last row and column of (17). The limit experiment $\mathcal{E}_{\mathcal{M}}(f)$ can be described as follows: we observe, with $B^{W_h}(s) = W_h(s) - sW_h(1)$, $\left\{\left(W_\varepsilon(s), B^{W_h}(s)\right) : s \in [0,1]\right\}$ with (W_ε, W_h) generated by*

$$\mathrm{d}W_\varepsilon(s) = cW_\varepsilon(s)\mathrm{d}s + \mathrm{d}Z_\varepsilon(s), \tag{38}$$

$$\mathrm{d}W_h(s) = (bJ_{f_y h} + cJ_{f_x h})W_\varepsilon(s)\mathrm{d}s + \mathrm{d}Z_h(s). \tag{39}$$

The difference between Corollary 3.2 and Theorem 3.1 is twofold. First, besides the process W_ε, the observation in the invariant limit experiment in Corollary 3.2 is only the Brownian bridge B^{W_h} and not the complete Brownian motion W_h. Second, as a consequence of this, the nuisance parameter η disappeared from (39).

Corollary 3.2 does not provide, as far as we know, a further invariance structure that can be used to eliminate the nuisance parameter c. As a result, we rely, in Section 4, on the so-called Approximate Least Favorable Distribution method to deal with this last nuisance parameter.

We conclude this section by again considering the special case of a Gaussian innovation density f. This also shows where exactly our power gains, under serially independent innovations, come from relative to the Gaussian procedures in, for instance, Jansson and Moreira (2006).

Remark 3.1 (Attainability of the Semiparametric Power Envelope). One may expect the semiparametric power envelope to be formally attainable by a likelihood-ratio test constructed using a nonparametric estimate of the score function ℓ_f. Intuitively, the argument is as follows. Rewrite $\int_0^1 W_\varepsilon(s)\mathrm{d}B_{\ell_{f_y}}(s) = \int_0^1 \left(W_\varepsilon(s) - \overline{W_\varepsilon}\right)\mathrm{d}W_{\ell_{f_y}}(s)$. Hence, even though there is a bias a (at rate $\sqrt{T}$) in the estimated score function, this bias will be canceled out automatically since $\int_0^1 \left(W_\varepsilon(s) - \overline{W_\varepsilon}\right)\mathrm{d}\left(as + W_{\ell_{f_y}}(s)\right) = \int_0^1 \left(W_\varepsilon(s) - \overline{W_\varepsilon}\right)\mathrm{d}W_{\ell_{f_y}}(s)$. The same argument applies to the term $\int_0^1 W_\varepsilon(s)\mathrm{d}B_{\ell_{f_x}}(s)$.

Compare the discussion in Jansson (2008, Section 6) for the unit root testing problem and Zhou (2020, Section 2) for general LAN, LAMN, and LABF experiments.

Remark 3.2 (Gaussian f). In the situation of Gaussian f, Remark 2.1 and Remark 2.2 imply that $B_{\ell_{f_y}}$ and $B_{\ell_{f_x}}$ are linear combinations of B_ε and $B_\perp$ (the Brownian bridges generated by W_ε and $W_\perp$, respectively). As a result, the optimal invariant procedures are measurable with respect to W_ε and $B_\perp$. Using the same conditional expectation calculation, the associated log-likelihood ratio of the Gaussian σ-field, $\mathcal{M}_{\text{Gaussian}} = \sigma(W_\varepsilon, B_\perp)$, leads to the Gaussian log-likelihood ratio in Jansson and Moreira (2006, Lemma 3). As $B_\perp$ is spanned by B^{W_h}, the σ-field $\mathcal{M}_{\text{Gaussian}}$ is also invariant w.r.t η (or f), but it is not *maximally* invariant. As a consequence, under non-Gaussianity, this leads to an efficiency loss in statistical inference.

Note that all existing tests in the literature are (essentially) based on the Gaussian likelihood of the generally non-maximally invariant $\mathcal{M}_{\text{Gaussian}}$, e.g., Jansson and Moreira (2006) and Elliott et al. (2015). Therefore, these tests belong to the class of asymptotically invariant tests. This invariance imposed in the limiting experiment is associated to invariance w.r.t. the innovation density f in the sequence as η represents local perturbations precisely of f. Indeed, we have the convergence $W_\varepsilon^{(T)}(s) \Rightarrow W_\varepsilon(s)$ and the one associated to $W_\perp$ for all $f \in \mathfrak{F}$, hence, η will not enter the associated equation (27) in the limiting experiment. See Müller (2011) for a more comprehensive analysis of this convergence.

3.4 Rank-based asymptotically invariant statistics

The elimination of the nuisance parameter η is performed in the limit experiment $\mathcal{E}(f)$ and leads to $\mathcal{E}_{\mathcal{M}}(f)$. We now show how this elimination can be mimicked in the actual predictive regression model of interest, i.e., in $\mathcal{E}^{(T)}(f)$. It is reasonable to expect that exploiting the asymptotic invariance structures also works "well" for the sequence of experiments. The claim will be substantiated by the simulation results in Section 5.

In line with the vast literature on rank-based inference, the appearance of the Brownian Bridges $B_{\ell_{f_x}}$ and $B_{\ell_{f_y}}$ in Corollary 3.2, naturally suggests to use statistics that are based on ranks of the innovations ε_t^y and ε_t^x in the predictive regression model. Indeed, we will follow that route. However, in the present situation we deal with bivariate innovations $(\varepsilon_t^y, \varepsilon_t^x)$ which complicates the analysis considerably relative to models with univariate innovations that are mostly studied in the literature.

As the true innovation density f is unknown, we actually base our test statistic

on an assumed (so-called *reference*) density g that also satisfies Assumption 1. Let g_y and g_x denote the marginal densities for the first, respectively, second component of g. The bivariate nature of the innovations $(\varepsilon_t^y, \varepsilon_t^x)$ implies that we cannot deal with a completely general reference bivariate density g. Thus, we choose marginal reference densities g_y and g_x, and a reference correlation parameter ρ_g. For the marginal reference densities, we impose the standard condition in the rank-based inference literature, see, e.g., Theorem 13.5 in Van der Vaart (2000).

Assumption 2. *The marginal* reference densities g_i, $i = \{y, x\}$, *are strictly positive, absolutely continuous with derivative* $\dot{g}_i$ *and* $J_{g_i} := \int (\dot{g}_i/g_i)^2 g_i < \infty$. *Moreover, we have*

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \left(-\frac{\dot{g}_i}{g_i} \left(G_i^{-1} \left(\frac{t}{T+1} \right) \right) \right)^2 = J_{g_i}, \tag{40}$$

where G_i^{-1} *is the inverse cumulative distribution function associated to* g_i.

Moreover, given an additionally chosen reference correlation $\rho_g \in (-1, 1)$, we define the associated bivariate reference score function

$$\ell_g(\varepsilon^y, \varepsilon^x) := \left(\ell_{g_y}(\varepsilon^y, \varepsilon^x), \ell_{g_x}(\varepsilon^y, \varepsilon^x) \right)' \tag{41}$$

where

$$\begin{aligned}
\ell_{g_y}(\varepsilon^y, \varepsilon^x) &= -\left(\frac{\dot{g}_y}{g_y}(\varepsilon^y) - \rho_g \frac{\dot{g}_x}{g_x}(\varepsilon^x) \right) \Big/ (1 - \rho_g^2), \\
\ell_{g_x}(\varepsilon^y, \varepsilon^x) &= -\left(\frac{\dot{g}_x}{g_x}(\varepsilon^x) - \rho_g \frac{\dot{g}_y}{g_y}(\varepsilon^y) \right) \Big/ (1 - \rho_g^2).
\end{aligned}$$

The linearity of the reference score functions ℓ_{g_y} and ℓ_{g_x} is key to the analysis that follows. It implies that, when using component-wise ranks of the innovations $(\varepsilon^y, \varepsilon^x)$, the resulting rank-based processes converge to a bivariate Brownian bridge. Despite its seemingly restrictive nature, the linearity allows to fully exploit the invariance structures embedded in the predictive regression model of interest, leading to sizable power gains (see Section 5).

Now, let $R_{y,t}$ denote the rank of y_t (among $y_1, \ldots, y_T$), while $R_{x,t}$ denotes the rank of $\Delta x_t = x_t - x_{t-1}$ (among $\Delta x_1, \ldots, \Delta x_T$). Note that the pairs $(R_{y,t}, R_{x,t})$ equal the (component-wise) ranks of $(\varepsilon_t^y, \varepsilon_t^x)$ under $\beta = 0$ and $\gamma = 0$. We define the bivariate partial sum process of the rank-based scores by

$$\begin{aligned}
B_{\ell_g}^{(T)}(s) &= \left(B_{\ell_{gy}}^{(T)}(s), B_{\ell_{gx}}^{(T)}(s) \right)' \\
&:= \frac{1}{\sqrt{T}} \sum_{t=1}^{\lfloor sT \rfloor} \ell_g \left(G_y^{-1} \left(\frac{R_{y,t}}{T+1} \right), G_x^{-1} \left(\frac{R_{x,t}}{T+1} \right) \right),
\end{aligned} \tag{42}$$

for $s \in [0, 1]$. Following the semiparametric literature (see, e.g., Bickel (1982), Jansson (2008) and Zhou et al. (2019)), we study the limit behavior of $B_{\ell_g}^{(T)}$ for

the predictive regression model in Section 2 with fixed, but arbitrary, $f \in \mathfrak{F}$ (i.e., $\eta = 0$). This, in turn, implies the limit behavior of the rank-based statistic proposed in Section 4.

First, we establish the limiting behavior of $B_{\ell_g}^{(T)}$ under $\mathrm{P}_{0,0,0;f}^{(T)}$ for any $f \in \mathfrak{F}$. Its proof is organized in Appendix B.

Proposition 3.2. *Suppose $\varepsilon_t = (\varepsilon_t^y, \varepsilon_t^x)'$ are i.i.d. innovations with density $f \in \mathfrak{F}$. Let g_y and g_x be reference densities that satisfy Assumption 2 and fix the reference correlation ρ_g. Then, under $\mathrm{P}_{0,0,0;f}^{(T)}$, we have*

$$\left(W_\varepsilon^{(T)}, W_{\ell_f}^{(T)}, B_{\ell_g}^{(T)}\right)' \Rightarrow \left(W_\varepsilon, W_{\ell_f}, B_{\ell_g}\right)', \tag{43}$$

where $W_{\ell_f} := (W_{\ell_{f_y}}, W_{\ell_{f_x}})'$ and B_{ℓ_g} is a bivariate Brownian bridge, i.e., $B_{\ell_g}(s) := W_{\ell_g}(s) - sW_{\ell_g}(1)$, with W_{ℓ_g} a zero-drift Brownian motion. The covariance of W_{ℓ_g} with W_ε and W_{ℓ_f} is given by

$$\mathrm{Var}\begin{pmatrix} W_\varepsilon(1) \\ W_{\ell_f}(1) \\ W_{\ell_g}(1) \end{pmatrix} = \begin{pmatrix} 1 & \boldsymbol{e}_1' & \boldsymbol{\sigma}_{\varepsilon g}' \\ \boldsymbol{e}_1 & J_f & J_{fg} \\ \boldsymbol{\sigma}_{\varepsilon g} & J_{gf} & J_g \end{pmatrix}, \tag{44}$$

where

$$\begin{aligned}
\boldsymbol{e}_1 &= (0,1)', \\
\boldsymbol{\sigma}_{\varepsilon g} &= (\sigma_{\varepsilon g_y}, \sigma_{\varepsilon g_x})' = \mathrm{E}_f\left[\varepsilon_t^x \ell_g\left(G_y^{-1}(F_y(\varepsilon_t^y)), G_x^{-1}(F_x(\varepsilon_t^x))\right)\right], \\
J_{fg} &= J_{gf}' = \mathrm{E}_f\left[\ell_f(\varepsilon_t^y, \varepsilon_t^x)\ell_g\left(G_y^{-1}(F_y(\varepsilon_t^y)), G_x^{-1}(F_x(\varepsilon_t^x))\right)'\right], \\
J_g &= \mathrm{E}_f\left[\ell_g\left(G_y^{-1}(F_y(\varepsilon_t^y)), G_x^{-1}(F_x(\varepsilon_t^x))\right)\ell_g\left(G_y^{-1}(F_y(\varepsilon_t^y)), G_x^{-1}(F_x(\varepsilon_t^x))\right)'\right].
\end{aligned}$$

The above result is classical for univariate rank statistics. In the present paper, we use component-wise bivariate ranks. One complication is that the matrix J_g depends on f through its copula. This implies that, like J_g, it will have to be estimated in applications; see also the discussion of Zhou (2020, Theorem 3.1).

Next, we provide the limit behavior of $B_{\ell_g}^{(T)}$ under the law $\mathrm{P}_{b,c,0;f}^{(T)}$, with $b \in \mathbb{R}$ and $c \in (-\infty, 0]$. A proof is again provided in Appendix B.

Corollary 3.3. *Impose the same conditions as in Proposition 3.2. For any $f \in \mathfrak{F}$, $b \in \mathbb{R}$, and $c \in (-\infty, 0]$, we have*

$$\left(W_\varepsilon^{(T)}, W_{\ell_f}^{(T)}, B_{\ell_g}^{(T)}\right)' \Rightarrow \left(W_\varepsilon, W_{\ell_f}, B_{\ell_g}\right)', \tag{45}$$

under $\mathrm{P}_{b,c,0;f}^{(T)}$, where the law of $(W_\varepsilon, W_{\ell_f}, B_{\ell_g})'$ is implied by

$$\mathrm{d}W_\varepsilon(s) = cW_\varepsilon(s)\mathrm{d}s + \mathrm{d}Z_\varepsilon(s), \tag{46}$$

$$\mathrm{d}W_{\ell_f}(s) = J_f(b,c)'W_\varepsilon(s)\mathrm{d}s + \mathrm{d}Z_{\ell_f}(s), \tag{47}$$

$$\mathrm{d}W_{\ell_g}(s) = J_{gf}(b,c)'W_\varepsilon(s)\mathrm{d}s + \mathrm{d}Z_{\ell_g}(s), \tag{48}$$

with $(Z_\varepsilon, Z_{\ell_f}, Z_{\ell_g})'$ *a zero-drift Brownian motion with variance identical to (44).*

We use the rank-based processes $B_{\ell_g}^{(T)}$ to replace B_{ℓ_f} in the likelihood ratio in Theorem 3.3; see Section 4.2 for details. In line with Remark 3.1, one could contemplate to use reference densities $\hat{f}$ based on a non-parametric estimate of the true innovation density, but we leave a formal analysis for future work. As we will see in Section 5, even for incorrectly chosen reference densities (that is, for $g \neq f$), our procedure features power gains over existing Gaussian based procedures. These gains come from the assumption that the error term ε_t is driven by some i.i.d. innovations, which may possibly be maintained in empirical work. It is important to note that choosing a reference density $g \neq f$ does not affect the validity of our test. The test will be of the appropriate level irrespective of the reference densities g_y and g_x chosen (provided they satisfy Assumption 2). But, likelihood ratio tests based on Theorem 3.3 still feature the nuisance parameter c. We deal with this in the next section.

Remark 3.3 (Invariance Restriction and Rank Statistics). Consider a reference density $g \neq f$ as a perturbed version of f with perturbation h_g defined by $g = f(1+h_g)$. The expectation of scores based on the reference density g under the true density f, $\xi := \mathbb{E}_f[-\dot{g}/g(\varepsilon)]$, is generally not zero. As a consequence, g-based quasi-likelihood/score inference becomes invalid. When considering local alternatives, i.e., $g = f_\eta^{(T)}$ (see equation (8)) within our asymptotic framework, we will have $\xi = J_{fb}'\eta$ with $J_{fb} = (J_{f_y h}, J_{f_x h})$.

To get valid inference invariant w.r.t. ξ, besides the $\hat{f}$-based approach mentioned in Remark 3.1, one can also consider the use of rank-based statistics. Specifically, if one replaces the innovations ε_t^y's by their rank-based versions as above, i.e., by $G_y^{-1}(R_{y,t}/(T+1))$'s, the average (over all observations) of the associated scores will be non-random (as each rank simply occurs once) and close to zero (because of the equally-spaced nature of the ranks). Consequently, an innovation-based partial sum process will become tied down at $s = 1$ just as, in the limit experiment, the Brownian bridge can be seen as a tied down Brownian motion. Note that any non-zero drift ξ in the partial sum process (as discussed above) will be removed as well by this transformation. As a result, the inference procedure becomes invariant with respect to perturbations in the density f. Said differently, the zero-mean (at $s = 1$) property of rank-based scores enables us to choose any reference density g while preserving the validity of the inference procedure. The choice set for g could also include a nonparametrically estimated density $\hat{f}$ (see the simulation evidence in Section 5.3) although a formal analysis is beyond the scope of the present paper.

4 Eliminating the nuisance parameter γ by ALFD

In the previous section, we have developed the semiparametric power envelope for tests on b that are invariant with respect to η, under the assumption that c is known. We now address the question of testing the regression coefficient β in case γ is treated as a nuisance parameter as well.

As argued in the the discussion following Corollary 3.2, we conjecture that the nuisance parameter c cannot be dealt with using invariance arguments. Various alternative methods to deal with nuisance parameters in testing problems have been used in the literature. In relation to the predictive regression model at hand, we mention the Bonferroni method (Cavanagh et al. (1995) and Campbell and Yogo (2006)); tests based on a conditional unbiasedness condition (Jansson and Moreira (2006)); and tests based on a numerically calculated Approximate Least Favorable Distribution (ALFD) as more recently proposed in Elliott et al. (2015). All these techniques apply to the Gaussian likelihood ratio statistic in Remark 3.2.

These approaches have different advantages and disadvantages. Campbell and Yogo (2006) proposes a modified Bonferroni method to eliminate the nuisance parameter c, leading to a simple yet more powerful test than the Cavanagh et al. (1995) test. However, as pointed out by Phillips (2014), inference based on Bonferroni bounds can be severely undersized when the predictor is "far away" from being a unit root process ($\gamma << 1$). In such a case, confidence intervals obtained by inverting the test may end up having essentially zero coverage probability. Jansson and Moreira (2006) develops an approach conditional on specific auxiliary statistics—the terms (only) associated with c in the Gaussian likelihood ratio—and derives an optimal test in the class of conditionally unbiased tests. Nevertheless, such a conditional unbiasedness constraint narrows the considered class and rules out some more powerful tests. Consequently, as shown by the simulation results of Jansson and Moreira (2006), the associated test has relatively low power compared to the Campbell and Yogo (2006) test under most alternatives.

Recently, Elliott et al. (2015) proposes a numerical algorithm to determine an ALFD of the nuisance parameter c to optimize weighted average power over some compact interval (of c). Note that with respect to our parameter of interest b, we consider point-optimal test and do not use weighted powers over a discretized space to avoid the induced computational complexities. On one hand, the ALFD yields an upper bound of the weighted average power for all valid tests. On the other hand, integrating out the likelihood statistic w.r.t. the ALFD leads to a "nearly optimal" test whose power is close to the upper bound. Moreover, by switching to standard

asymptotic approximations in case γ appears to be far from unity, the associated test can achieve better size and power performances for all $c \in (-\infty, 0]$ (e.g., across the parameter space $\gamma \in (-1, 1]$). Therefore, we employ this ALFD approach in the present paper, together with the switching mechanism (see Appendix C), to our rank-based likelihood statistics in (56).[7] This leads to tests that are of correct size for all relevant c and have good power performance. We confirm these properties by simulations in Section 5.

4.1 The Approximately Least Favorable Distribution (ALFD) Approach

In Section 3 we used invariance arguments to reduce the predictive regression testing problem towards log-likelihood ratio of the form (34) where b is the parameter of interest to be tested and c is a nuisance parameter. We briefly outline, in the present section, how the Approximate Least Favorable Distribution approach in Elliott et al. (2015) works in our setting.

Rewrite the log-likelihood ratio of the maximal invariant $\mathcal{M}$ in Theorem 3.3 as

$$\mathcal{L}_{\mathcal{M}}(b, c) = bS_1 + cS_2 - \frac{1}{2}\left((b, c)J_f(b, c)' - c^2\right) S_3 - \frac{1}{2}c^2 S_4,$$

where

$$S_1 = \int_0^1 W_\varepsilon(s)\mathrm{d}B_{\ell_{f_y}}(s), \quad S_2 = \int_0^1 W_\varepsilon(s)\mathrm{d}B_{\ell_{f_x}}(s) + W_\varepsilon(1)\overline{W_\varepsilon}, \tag{49}$$
$$S_3 = \overline{W_\varepsilon^2} - \left(\overline{W_\varepsilon}\right)^2 \quad \text{and} \quad S_4 = \overline{W_\varepsilon^2}.$$

One can thus consider the four-dimensional sufficient statistic $S := (S_1, S_2, S_3, S_4)$. For notational simplicity, in the present section, we denote by $F_{b,c}(S)$ the distribution of S under $\mathbb{P}_{b,c}$. The hypothesis of interest is

$$\mathrm{H}_0 : b = 0, \quad c \in (-\infty, 0] \quad \text{versus} \quad \mathrm{H}_1 : b > 0, \quad c \in (-\infty, 0]. \tag{50}$$

Note that, thus, both the null and the alternative hypothesis are composite. We first discuss elimination of the nuisance parameter c under the alternative and, subsequently, its elimination under the null.

[7] We expect that other approaches based on likelihood ratios, e.g., the approaches of Campbell and Yogo (2006) and Jansson and Moreira (2006), will apply here as well. This is because (i) the semiparametric likelihood ratio $\mathcal{L}_{\mathcal{M}}(b, c)$ in Theorem 3.3 is as the general version of the Gaussian likelihood ratio in Jansson and Moreira (2006, Lemma 3), thus when the true density is Gaussian, the former reduces to the latter; (ii) its rank-based proxy in (56) has the same structure (exponential family); and (iii) the asymptotic behaviors of the associated rank-based processes are known and consistently estimable.

To eliminate the nuisance parameter c under the alternative, a standard approach is to consider a so-called weighted average power (see, e.g., Andrews and Ploberger (1994))

$$\text{WAP}(\varphi) = \int_c \left(\int_S \varphi(S) \mathrm{d}F_{b,c}(S) \right) \mathrm{d}\Lambda_1(c), \tag{51}$$

where φ is some test function for the problem above and Λ_1 is a probability weighting measure for $c \in (-\infty, 0]$. The weighting measure Λ_1 can be chosen by the researcher and reflects the weights that she assigns to various values of c under the alternative. Due to Fubini's Theorem, we have

$$\text{WAP}(\varphi) = \int_S \varphi(S) \mathrm{d} \int_c F_{b,c}(S) \mathrm{d}\Lambda_1(c), \tag{52}$$

which leads to the simple alternative hypothesis $\mathrm{H}_{1;\Lambda_1}$, under which the distribution of S is given by the mixture $F_{b;\Lambda_1}(S) = \int F_{b,c}(S) \mathrm{d}\Lambda_1(c)$. In this way, the testing problem is reduced to testing H_0 against $\mathrm{H}_{1;\Lambda_1}$.

Subsequently, in order to eliminate the nuisance parameter c under the null we proceed as follows. Again we impose a probability weighting measure Λ_0 for c and introduce the simple null hypothesis, denoted $\mathrm{H}_{0;\Lambda_0}$, under which the distribution of S is given by $F_{b;\Lambda_0}(S) = \int F_{b,c}(S) \mathrm{d}\Lambda_0(c)$. Now we define the test $\varphi_{\bar{b};\Lambda}$ by

$$\varphi_{\bar{b},\Lambda_0}(S) = \begin{cases} 1 & \text{if} \quad \mathrm{d}F_{\bar{b},\Lambda_1}(S) > \kappa \mathrm{d}F_{0,\Lambda_0}(S), \\ 0 & \text{if} \quad \mathrm{d}F_{\bar{b},\Lambda_1}(S) \leq \kappa \mathrm{d}F_{0,\Lambda_0}(S), \end{cases} \tag{53}$$

where the critical value κ is chosen to obtain the desired size. By the Neyman-Pearson Lemma, $\varphi_{\bar{b},\Lambda_0}$ is point optimal at $b = \bar{b}$, for the problem of testing the null $\mathrm{H}_{0;\Lambda_0}$ against the alternative $\mathrm{H}_{1;\Lambda_1}$.

The problem of choosing Λ_0 is, unfortunately, more complicated than that of choosing Λ_1. The reason is that we want to control the rejection probability of the test, not only under $\mathrm{H}_{0;\Lambda_0}$, but for all values of $c \in (-\infty, 0]$. In general there is no reason to expect that a level-α test under $\mathrm{H}_{0;\Lambda_0}$ is of correct size for the entire null hypothesis H_0. However, for some specific choices of Λ_0 this statement is true, and such a distribution is called a *least-favorable distribution*; see, e.g., Lehmann and Romano (2006), Theorem 3.8.1. Formally, a distribution Λ_0^* is called *least favorable* if the most powerful level-α test (53) for testing $\mathrm{H}_{0;\Lambda_0^*}$ against $\mathrm{H}_{1;\Lambda_1}$ is of the desired size for the (entire) null hypothesis H_0. Moreover, once more by Theorem 3.8.1 in Lehmann and Romano (2006), the test $\varphi_{\bar{b},\Lambda_0^*}$ is also point optimal (at $b = \bar{b}$) for this problem. A least-favorable distribution Λ_0^* exists in most of the usual statistical problems. conditions that ensure this and associated references can be found in Section 3.8 of Lehmann and Romano (2006).

As, in most cases, the least-favorable distribution Λ_0^* is not easily obtained, Elliott et al. (2015) propose a numerical method to find, what they call, an "Approximate Least Favorable Distribution" (ALFD). The ALFD is defined as follows.

Definition 1. *An ϵ-ALFD is a probability distribution $\Lambda_0^{*\epsilon}$ over $(-\infty, 0]$ satisfying*

(i) the Neyman-Pearson test (53) with $\Lambda = \Lambda_0^{*\epsilon}$ and critical value $\kappa = \kappa^*$, i.e., $\varphi_{\bar{b},\Lambda_0^{*\epsilon}}$, is of size α under $\mathrm{H}_{0;\Lambda_0^{*\epsilon}}$ and has power $\bar{\pi}$ against $\mathrm{H}_{1;\Lambda_1}$;

(ii) there exists $\kappa^{*\epsilon}$ such that the test (53) with $\Lambda = \Lambda_0^{*\epsilon}$ and $\kappa = \kappa^{*\epsilon}$, $\varphi^{\epsilon}_{\bar{b},\Lambda_0^{*\epsilon}}$, is of level α under H_0, and has power of at least $\bar{\pi} - \epsilon$ against $\mathrm{H}_{1;\Lambda_1}$.

The test $\varphi^{\epsilon}_{\bar{b},\Lambda_0^{*\epsilon}}$ (in particular, the ALFD $\Lambda_0^{*\epsilon}$ and the critical value $\kappa^{*\epsilon}$) is exactly what we are looking for, once we have set the weights Λ_1 of interest for the alternative hypothesis. Besides the size control under H_0, the definition above also ensures that the test $\varphi^{\epsilon}_{\bar{b},\Lambda_0^{*\epsilon}}$ enjoys a near-optimality property with a relatively small power loss (less than ϵ).

Note that even for a given (small) value of ϵ, the ALFD $\Lambda_0^{*\epsilon}$ is not necessarily "close" to the least favorable distribution Λ_0^*. Actually, (possibly infinitely) many pairs of $(\Lambda_0^{*\epsilon}, \kappa^{*\epsilon})$ may satisfy Definition 1. The details about how to implement the numerical algorithm to determine a pair of $(\Lambda_0^{*\epsilon}, \kappa^{*\epsilon})$ (henceforth the test $\varphi_{\bar{b},\Lambda_0^{*\epsilon}}$) for a small ϵ can be found in Section 3 and Appendix A of Elliott et al. (2015). As the nuisance parameter space $c \in (-\infty, 0]$ is unbounded, we also need to "switch" back to standard test statistics (i.e., in the stationary case) for large values of $|c|$. We provide in Appendix C the details about our test for the standard part of the limit experiment $\mathcal{E}(f)$.

4.2 Putting it all together

Putting everything together, our test for the predictive regression model is based on applying the ALFD approach to the rank-based counterpart (using Proposition 3.2) of the asymptotically point-optimal invariant derived in Theorem 3.3.

We thus replace, in the sufficient statistic $S = (S_1, S_2, S_3, S_4)$ in (49), W_ε, $B_{\ell_{fy}}$, and $B_{\ell_{fx}}$ by $W_\varepsilon^{(T)}$, $B_{\ell_{gy}}^{(T)}$, and $B_{\ell_{gx}}^{(T)}$, leading to the feasible rank-based statistic

$$S_g^{(T)} := \left(S_{g,1}^{(T)}, S_{g,2}^{(T)}, S_{g,3}^{(T)}, S_{g,4}^{(T)}\right), \tag{54}$$

where

$$
\begin{aligned}
S_{g,1}^{(T)} &= \int_0^1 W_\varepsilon^{(T)}(s)\mathrm{d}B_{\ell_{g_y}}^{(T)}(s), \\
S_{g,2}^{(T)} &= \int_0^1 W_\varepsilon^{(T)}(s)\mathrm{d}B_{\ell_{g_x}}^{(T)}(s) + W_\varepsilon^{(T)}(1)\int_0^1 W_\varepsilon^{(T)}(s)\mathrm{d}s, \\
S_{g,3}^{(T)} &= \int_0^1 W_\varepsilon^{(T)}(s)^2\mathrm{d}s - \left(\int_0^1 W_\varepsilon^{(T)}(s)\mathrm{d}s\right)^2, \\
S_{g,4}^{(T)} &= \int_0^1 W_\varepsilon^{(T)}(s)^2\mathrm{d}s.
\end{aligned}
$$

To make the log-likelihood ratio $\mathcal{L}_\mathcal{M}$ in (34) fully feasible, we also have to deal with J_f. From Kagan and Landsman (1999) we know that J_f is diagonalized by the Cholesky root of the correlation matrix $\mathbf{R}_g$. Therefore, we replace J_f by

$$
J_p = \begin{pmatrix} J_{p_{yy}} & J_{p_{yx}} \\ J_{p_{yx}} & J_{p_{xx}} \end{pmatrix} := \mathbf{R}_g^{-\frac{1}{2}\prime}\,\mathrm{diag}\{J_{g_y}, J_{g_x}\}\mathbf{R}_g^{-\frac{1}{2}}, \tag{55}
$$

where J_{g_y} and J_{g_x} are the Fisher information of the chosen marginal reference densities defined in Assumption 2, and $\mathbf{R}_g$ is the correlation matrix based on the chosen reference correlation ρ_g, i.e., $\mathbf{R}_g := \left(\begin{smallmatrix} 1 & \rho_g \\ \rho_g & 1 \end{smallmatrix}\right)$. We recommend to use a consistent estimate of ρ as ρ_g regarding the power of the test, although any choice of ρ_g would lead to correct sizes. This leads to our feasible rank-based log-likelihood statistic

$$
\mathcal{L}_g^{(T)}(\bar{b}, c) := \bar{b}S_{g,1}^{(T)} + cS_{g,2}^{(T)} - \frac{1}{2}\left((\bar{b}, c)J_p(\bar{b}, c)' - c^2\right)S_{g,3}^{(T)} - \frac{1}{2}c^2 S_{g,4}^{(T)}, \tag{56}
$$

where $\bar{b}$ serves as a (chosen) fixed alternative point for the quasi-likelihood statistic (see, e.g., Elliott et al. (1996)). The nuisance parameter c will be removed by the ALFD approach, which is the reason that we do not need to similarly introduce a fixed alternative $\bar{c}$ for it. We provide the limit behavior of $\mathcal{L}_g^{(T)}(\bar{b}, c)$ in the following proposition.

Proposition 4.1. *Suppose $\varepsilon_t = (\varepsilon_t^y, \varepsilon_t^x)'$ are i.i.d. innovations with density $f \in \mathfrak{F}$. Let g_y and g_x be reference densities that satisfy Assumption 2 and fix the reference correlation ρ_g. Then, for $\bar{b} \in \mathbb{R}$, $b \in \mathbb{R}$ and $c \in (-\infty, 0]$, under $\mathrm{P}_{b,c,0;f}^{(T)}$, we have*

$$
\mathcal{L}_g^{(T)}(\bar{b}, c) \Rightarrow \mathcal{L}_g(\bar{b}, c),
$$

where

$$
\mathcal{L}_g(\bar{b}, c) := \bar{b}S_{g,1} + cS_{g,2} - \frac{1}{2}\left((\bar{b}, c)J_p(\bar{b}, c)' - c^2\right)S_{g,3} - \frac{1}{2}c^2 S_{g,4} \tag{57}
$$

with

$$
\begin{aligned}
&S_{g,1} = \int_0^1 W_\varepsilon(s)\mathrm{d}B_{\ell_{g_y}}(s), \quad S_{g,2} = \int_0^1 W_\varepsilon(s)\mathrm{d}B_{\ell_{g_x}}(s) + W_\varepsilon(1)\overline{W_\varepsilon}, \\
&S_{g,3} = \overline{W_\varepsilon^2} - \left(\overline{W_\varepsilon}\right)^2 \quad \text{and} \quad S_{g,4} = \overline{W_\varepsilon^2},
\end{aligned}
$$

where the behavior of W_ε and W_{ℓ_g} is given in (46) and (48), respectively.

We omit the proof of Proposition 4.1 since it directly follows from the weak convergence in Corollary 3.3, the continuous mapping theorem, and the rank-based stochastic integral convergence result in the proof of Zhou et al. (2019, Lemma 4.1). This proposition guarantees the validity of our test under the null $b = 0$, $c \in (-\infty, 0]$. Together with Corollary 3.3 it also establishes the power under local alternatives with $b > 0$.

Although not explicit in the above, observe that the statistic $\mathcal{L}_g^{(T)}$ in (56) still depends on σ_x through $W_\varepsilon^{(T)}$ defined in (12). We will simply replace σ_x by its sample counterpart below. As long as this estimator is consistent, the continuous mapping theorem shows that this replacement has no asymptotic consequences. The statistic does not depend on σ_y, but it does depends on the reference correlation ρ_g.

Now, applying the ALFD algorithm to $\mathcal{L}_g^{(T)}(\bar{b}, c)$, we obtain a distribution $\Lambda_{0,g}^{*\epsilon}$ and critical value $\kappa_{g,n}$ such that the test

$$\varphi_{g,n}(S_g^{(T)}, \rho_g) = \begin{cases} 1 & \text{if} \quad \int \mathcal{L}_g^{(T)}(\bar{b}, c) \mathrm{d}\Lambda_1(c) > \kappa_{g,n} \int \mathcal{L}_g^{(T)}(0, c) \mathrm{d}\Lambda_{0,g}^{*\epsilon}(c) \\ 0 & \text{if} \quad \int \mathcal{L}_g^{(T)}(\bar{b}, c) \mathrm{d}\Lambda_1(c) < \kappa_{g,n} \int \mathcal{L}_g^{(T)}(0, c) \mathrm{d}\Lambda_{0,g}^{*\epsilon}(c) \end{cases} \tag{58}$$

is of size α.

In order to get the appropriate critical values of the test, note that we need consistent estimates, under the null, of J_g and J_{fg}.[8] We need these in order to ensures the feasibility of the numerically determined pair $(\Lambda_0^{*\epsilon}, \kappa^{*\epsilon})$. In applications J_g and J_{fg} can easily be estimated, however, in the Monte Carlo study below we estimate J_g and J_{fg} based on the known. This is necessary as we cannot afford to determine a pair $(\Lambda_0^{*\epsilon}, \kappa^{*\epsilon})$ for each repetition in the simulation. That would be too intensive computationally.

5 A Monte Carlo Study

In this section, we explore by Monte Carlo the size and power properties of our test (58), combined with the switching approach detailed in Appendix C, (labeled WZ) relative to the Gaussian quasi-likelihood counterpart in Elliott et al. (2015) (labeled EMW). From the theoretical results, both tests should enjoy good size properties but the WZ test should exhibit larger power in case the true innovation

[8] A simple consistent estimator for J_g would be the sample covariance of the rank-based scores ℓ_g defined in (41) and a direct rank-based estimator for J_{fg} can be found in Cassart et al. (2010). It is worth noting that the consistency automatically also holds under local alternatives due to Le Cam's third lemma.

distribution is not Gaussian. Under Gaussian innovation distribution, both tests should have similar power.

Section 5.1 provides simulations under the predictive regression model studied formally in this paper. Section 5.2 provides results of our test under conditional heteroskedasticity. Finally, Section 5.3 provides results when the reference density used in the test is estimated.

5.1 Simulations under maintained i.i.d. assumption

We simulate the model (1)–(2) with $\mu = 2$, $\sigma_y = 3$, $\sigma_x = 3$, and $\rho = -0.5$. All results reported in this section are based on 10,000 replications.

For the ALFD approach in Elliott et al. (2015), we choose a discrete weighting distribution Λ_1 in (51) where each of the chosen 57 points

$$c \in \{0, -0.25^2, -0.5^2, \ldots, -14^2\}$$

of the support have equal weight. The same 57 points are also as the support of $\Lambda_0^{*\epsilon}$. For the test statistic in (58), we choose a fixed alternative $\bar{b} = B(1.645)$ where the power is about 50%. For the reference correlation ρ_g, we use the simple sample correlation of $\hat{\epsilon}_t^y$ and $\hat{\epsilon}_t^x$ under the null, where $\hat{\epsilon}_t^y = y_t - \sum_{t=1}^T y_t$ and $\hat{\epsilon}_t^x$ is the residual of the regression of x_t on x_{t-1}.

We present the power curves in two ways. The first presentation follows Elliott et al. (2015). Specifically, for the data generating process, we let the local nuisance parameter c (which governs the persistence of the predictor) take 21 values $c \in \{0, -10, -20, \ldots, -200\}$. And to have roughly similar power for each value of c, we transform the parameter b by

$$b = B(\delta) := \delta \sqrt{\frac{-2c + 6}{1 - \rho^2}}, \quad \text{for} \quad c < 0. \tag{59}$$

Alternatives for β are now characterized by different values of δ. The null hypothesis H_0 corresponds to $\delta = 0$, and we let the parameter of interest b take three alternatives: $\delta \in \{1, 2, 3\}$. Secondly, we present power curves where we fix the nuisance parameter $c = -25$ and plot the rejection rates for $\delta \in [0, 6]$. The significance level α is chosen to be 5% in all cases.

In Figure 1, we reports the large-sample ($T = 2,000$) size and power properties of our rank-based WZ test and the EMW test, for different combinations of the true density f and the marginal reference densities g_y and g_x. The upper-left subplot reports the case where f is a multivariate t_3 density, while g_y and g_x are both univariate t_3 densities. Both the EMW test and the WZ test are of correct size for all chosen values of c. Under the alternative hypothesis (i.e., for $\delta \in \{1, 2, 3\}$),

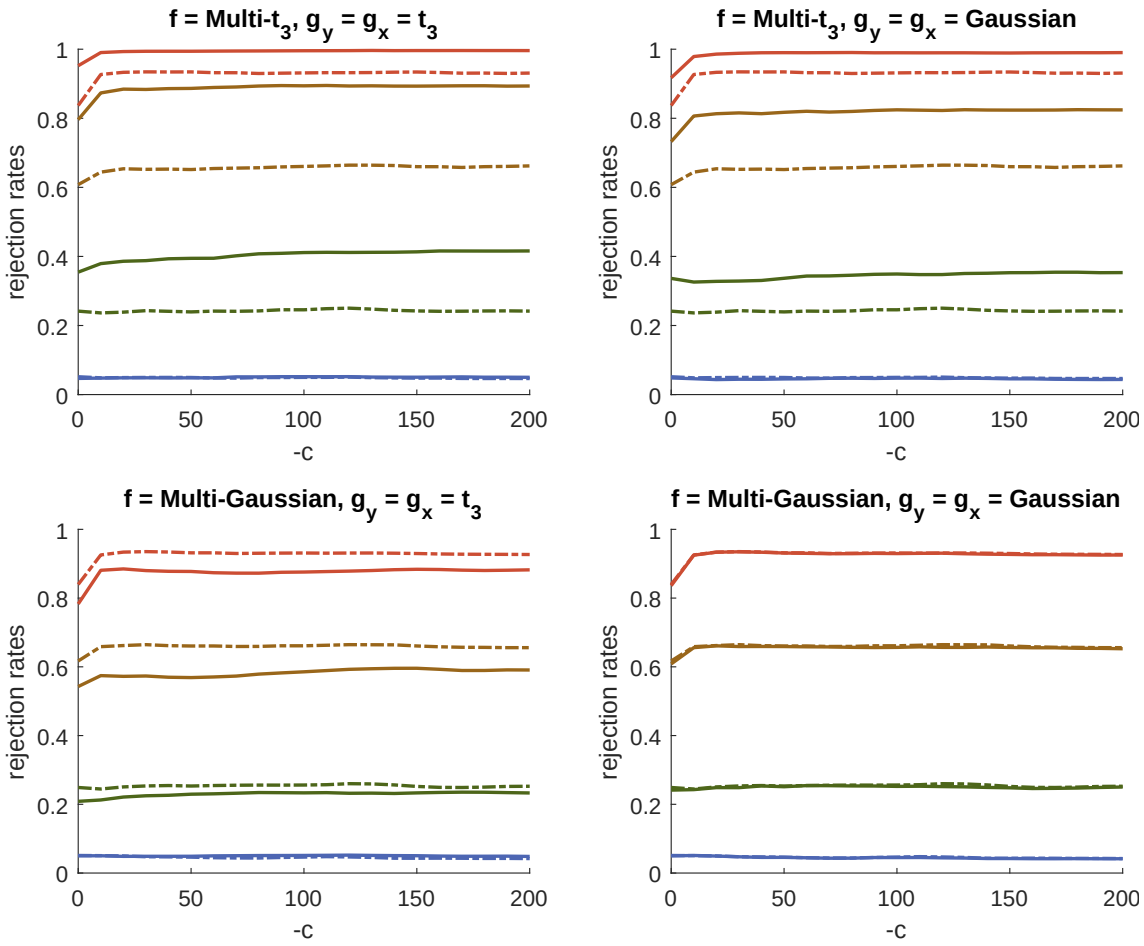

Figure 1: Rejection rates of the WZ test (solid lines) and the EMW test (dashed lines) for different values of $\delta = 0$, 1, 2, and 3, corresponding to lines in blue, green, brown, and red, respectively. For all the four cases, the correlation is -0.5. The sample size is 2,000.

the WZ test is more powerful than the EMW test. Taking the alternative $\delta = 2$ as example, for most values of c, the power of the EMW test is about 65% while the WZ test attains about 90% power. In the upper-right subplot, we keep f unchanged and let g_y and g_x both be Gaussian. Both tests provide correct size and, again, the WZ test is more powerful than the EMW test. However, compared to the upper-left subplot, we observe that the WZ test suffers a small power loss when choosing reference densities that are further away from the true ones. When f is Gaussian, the WZ test with Gaussian marginal reference densities shares almost the same size and power performances as the EMW test, as shown by the bottom-left subplot. The bottom-right subplot presents the case when f is Gaussian, while the marginal reference densities g_y and g_x are univariate t_3. In this case, the WZ test is less

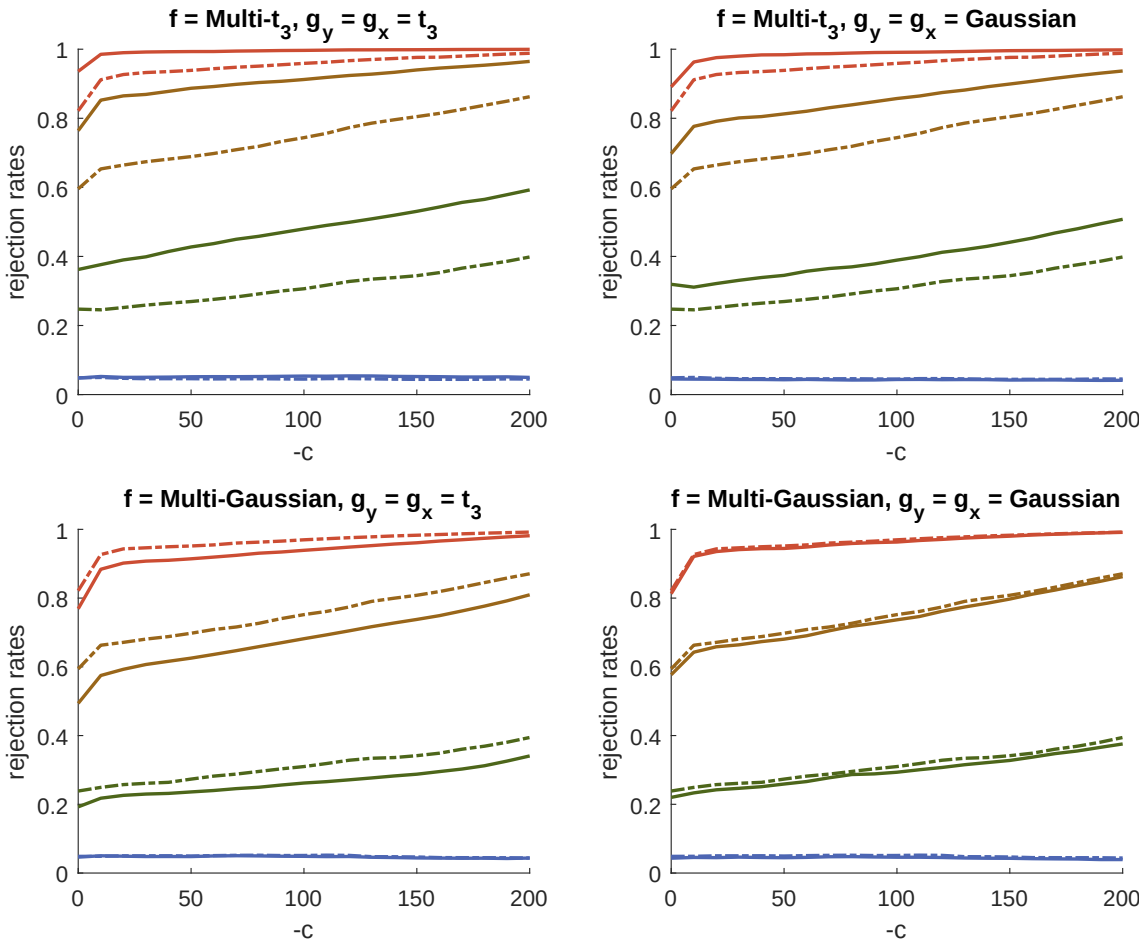

Figure 2: Rejection rates of the WZ test (solid lines) and the EMW test (dashed lines) for different values of $\delta = 0$, 1, 2, and 3, corresponding to lines in blue, green, brown, and red, respectively. For all the four cases, the correlation is -0.5. The sample size is 200.

powerful than the EMW test. In practice, we may want to avoid this power loss by pre-testing the residuals under the null hypothesis. We study this in Section 5.3.

Actually, one can always use Gaussian reference densities as a conservative choice, which is based on a (numerical) Chernoff and Savage (1958) result — keeping the marginal reference densities g_y and g_x Gaussian, the WZ test is always more powerful than the EMW test when f is non-Gaussian, and it works as well as the EMW test when f is Gaussian. A formal proof of this result in LABF-type experiments is still an open question, but we show that this property holds in some more simulations. In Figure 5, we fix g_y and g_x to be Gaussian, and choose four different multivariate innovation distributions: (i) Gaussian copula with Laplace marginal distributions (top-left, labeled Multi-Laplace); (ii) Multivariate Pearson

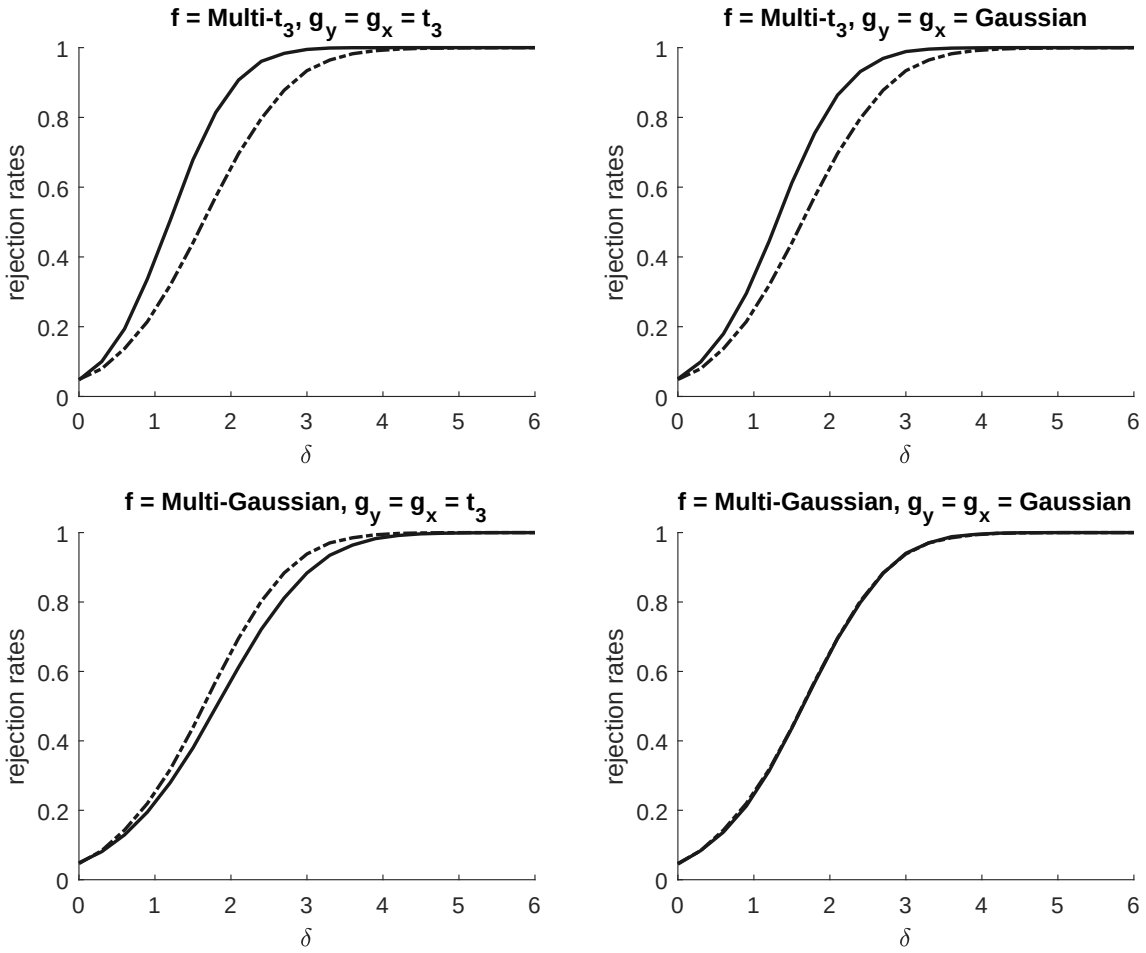

Figure 3: Rejection rates of the WZ test (solid lines) and the EMW test (dashed lines) for fixed value of $c = -25$ and different values of $\delta \in [0, 6]$. For all the four cases, the correlation is -0.5. The sample size is 2,000.

distribution with skewness 3 and kurtosis 36 (top-right, labeled Multi-Pearson); (iii) Gaussian copula with t_3 distribution for the first dimension and Gaussian distribution for the second dimension (bottom-left, labeled Multi-combo1); and (iv) t_3 copula with Gaussian for the first dimension and t_3 for the second dimension (bottom-right, labeled Multi-combo2). These simulations support the Chernoff-Savage result and also show that the further away the true distribution is from Gaussian, the more power can be gained by the WZ test. Moreover, case (iv) in the bottom-right subplot shows that actually the power we gain by the WZ test is from the innovation of the first dimension, ε_t^y. When the distribution of ε_t^y is Gaussian, we do as well as the EMW test. We conjecture that inference for β in the predictive regression model (1)-(2) is adaptive with respect to the marginal density of ε_t^x, when γ is eliminated by the ALFD approach in Elliott et al. (2015).

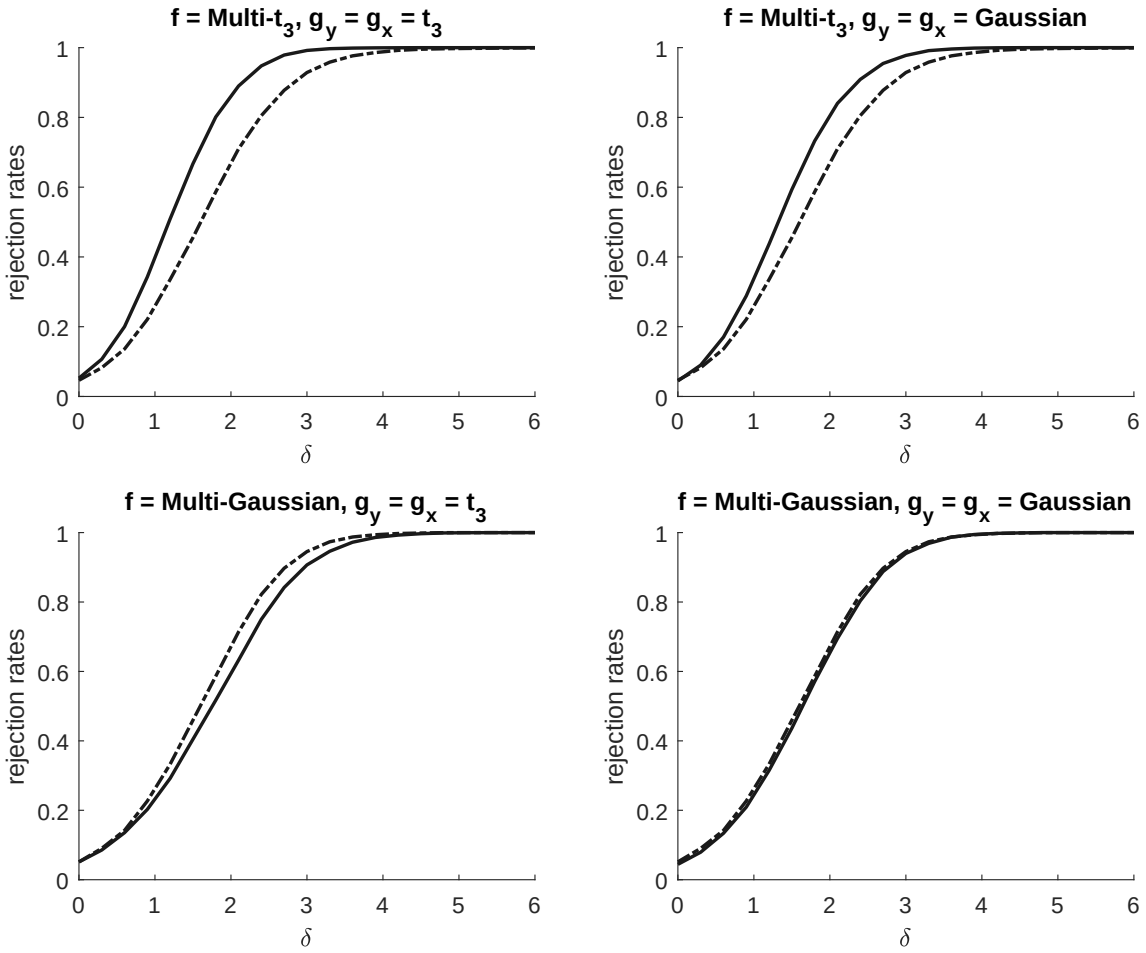

Figure 4: Rejection rates of the WZ test (solid lines) and the EMW test (dashed lines) for fixed value of $c = -25$ and different values of $\delta \in [0, 6]$. For all the four cases, the correlation is -0.5. The sample size is 2,000.

In Figure 3 and Figure 4, we present the powers of the WZ test and the EMW test (for fixed $c = -15$ and for $\delta \in [0, 6]$) under the same settings as in Figure 1 and Figure 2, respectively. The results show the power gain of the WZ test over the EMW test for all alternative values of c.

We also provide some small-sample ($T = 200$) results for both tests in Figure 2 and Figure 6 (the small-sample counterparts of Figure 1 and Figure 5, respectively). The conclusions are similar: both tests are of good size (all around 4.5%) using the same combinations of $\Lambda_0^{*\epsilon}$ and κ_g. The WZ test still gains considerable power in the case of non-Gaussian densities, though the gain is slightly smaller than in the large-sample case. This once more shows the additional information present, when supported by the application at hand, of an i.i.d.-ness assumption on the innovations. Appendix D provides additional simulation results in Figure D.1 and

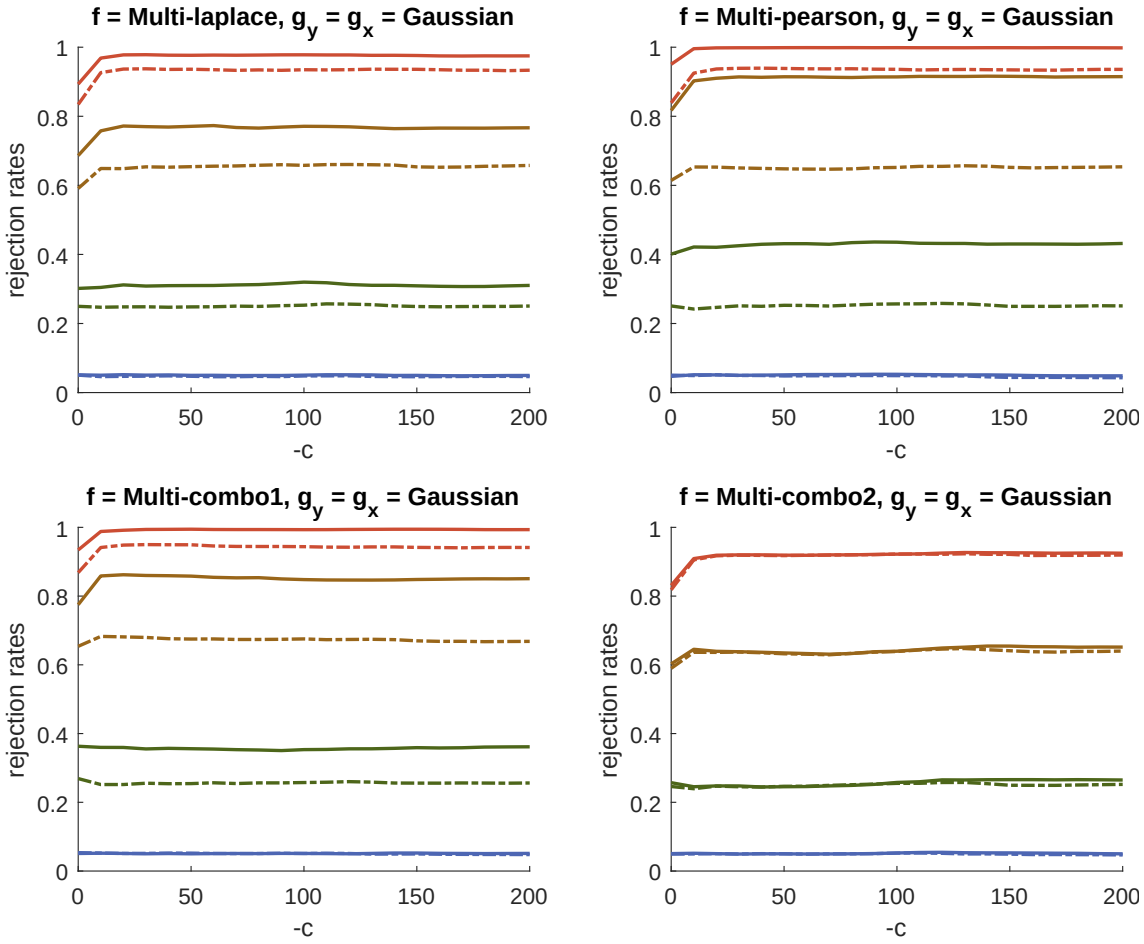

Figure 5: Rejection rates of the WZ test (solid lines) and the EMW test (dashed lines) for different values of $\delta = 0$, 1, 2, and 3, corresponding to lines in blue, green, brown, and red, respectively. For all the four cases, $\rho = -0.5$ and $T = 2,000$.

Figure D.2 for Figure 5 and Figure 6 using $c = -15$ and $\delta \in [0, 6]$, respectively.

Finally, we repeat the simulations of Figure 1 and Figure 2, but for $\rho = -0.9$, in Figure D.3 and Figure D.4 respectively in Appendix D. These simulations confirm our previous conclusions about the WZ test: correct sizes, power gain under non-Gaussian f, the Chernoff-Savage result, and decent small-sample performances.

5.2 Simulations under conditional heteroskedasticity

In many (financial) applications the maintained assumption of i.i.d. innovations will not be satisfied. We therefore study, by simulation, the behavior of the tests when the innovations exhibit conditional heteroskedasticity. The tests are identical to those in the previous sections, thus not adapted to deal with possible heteroskedasticity.

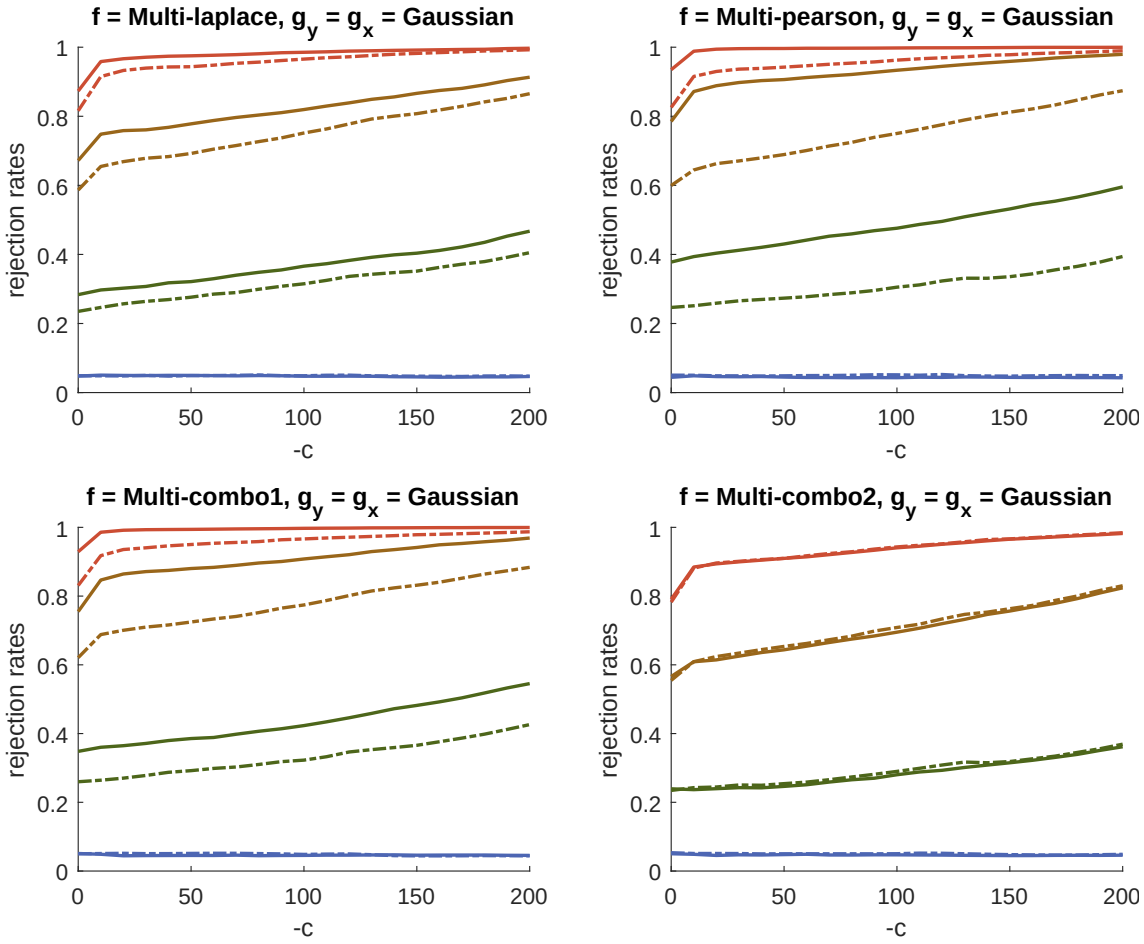

Figure 6: Rejection rates of the WZ test (solid lines) and the EMW test (dashed lines) for different values of $\delta = 0$, 1, 2, and 3, corresponding to lines in blue, green, brown, and red, respectively. For all the four cases, $\rho = -0.5$ and $T = 200$.

Keeping everything else unchanged, we replace the i.i.d. innovations $(\varepsilon_t^y, \varepsilon_t^x)'$, by a univariate GARCH(1,1) model (i) for ε_t^y only; or (ii) for both ε_t^y and ε_t^x in the data generating process. Formally, we choose

$$\varepsilon_t^y = \sqrt{1-\rho^2}\varepsilon_{1,t} + \rho\varepsilon_{2,t},$$
$$\varepsilon_t^x = \varepsilon_{2,t},$$

where, for case (ii), $\varepsilon_{1,t}$ and $\varepsilon_{2,t}$ are independently generated by the GARCH(1,1) model

$$\varepsilon_{j,t} = \nu_{j,t}\sqrt{h_{j,t}},$$
$$h_{j,t} = 1 + 0.07\varepsilon_{j,t-1} + 0.92h_{j,t-1},$$

for $j = 1, 2$, where $\nu_{j,t}$'s are i.i.d. innovations. For case (i), we let $\varepsilon_{2,t}$ be i.i.d. and

independent of $\varepsilon_{1,t}$. The joint density of $\nu_{1,t}$ and $\nu_{2,t}$ is denoted by f. The GARCH parameters are chosen based on common empirical findings.

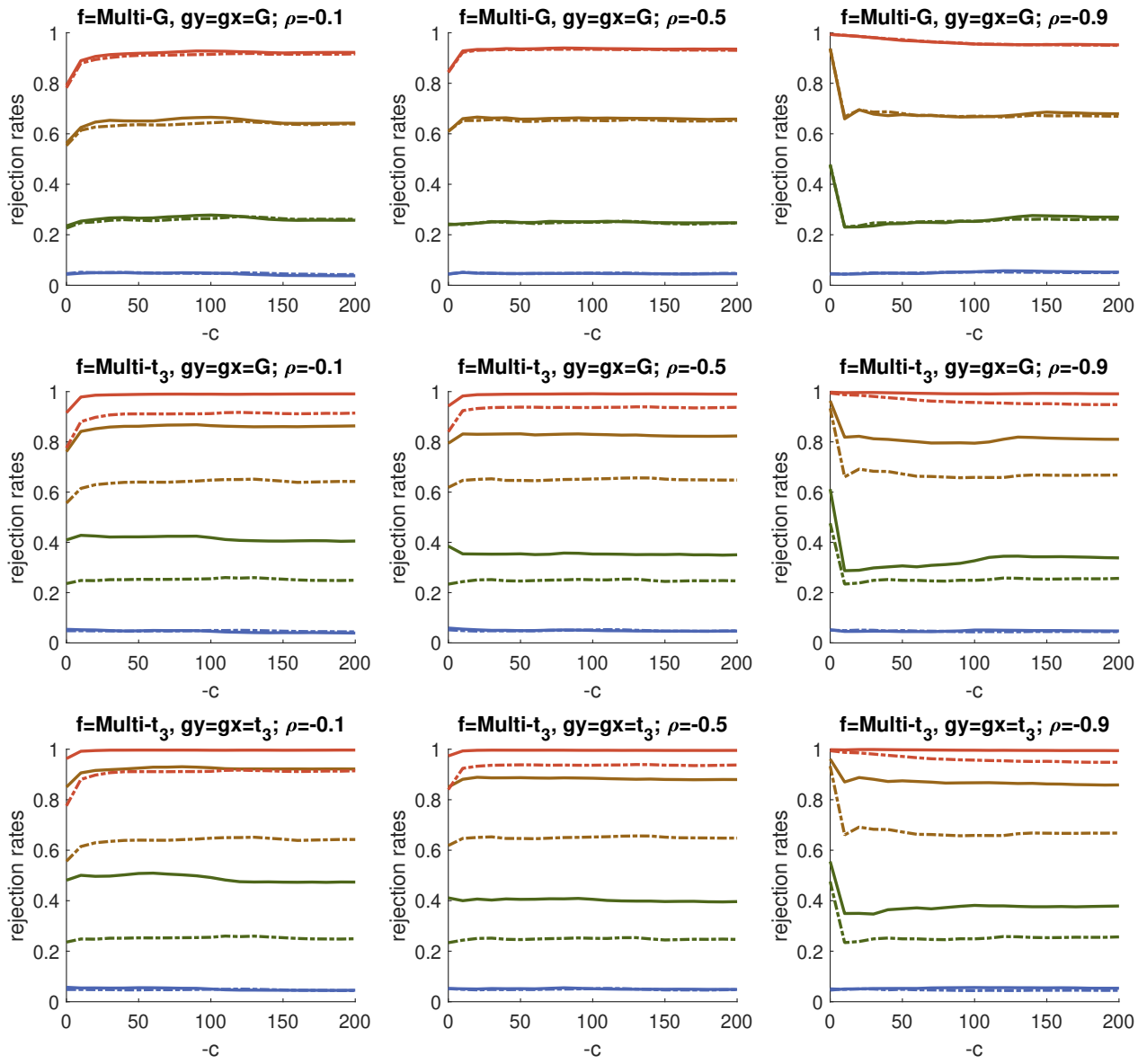

Figure 7: Rejection rates of the WZ test (solid lines) and the EMW test (dashed lines) for different values of $\delta = 0$, 1, 2, and 3, corresponding to lines in blue, green, brown, and red, respectively, under *heteroskedasticity*. For all the four cases, $T = 2,000$.

In Figure 7, we present case (i) where only the innovations of the response variable, ε^y, exhibit conditional heteroskedasticity, while the predictor innovations

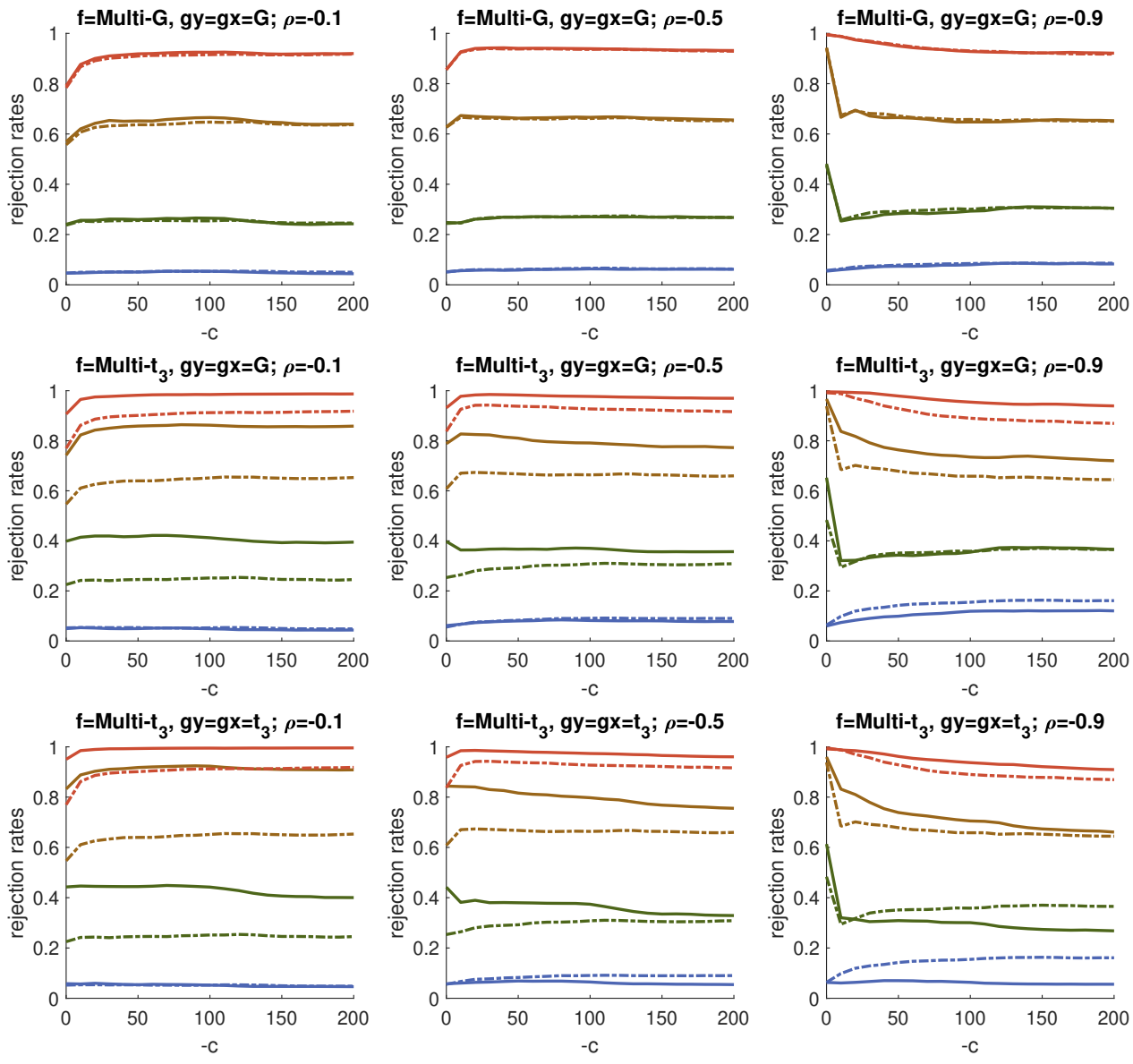

Figure 8: Rejection rates of the WZ test (solid lines) and the EMW test (dashed lines) for different values of $\delta = 0$, 1, 2, and 3, corresponding to lines in blue, green, brown, and red, respectively, under *heteroskedasticity*. For all the four cases, $T = 2,000$.

ε^x are still i.i.d. We show results for three density combinations as mentioned in the title of each subplot and three different values for the correlation of innovations ($\rho = -0.1$, -0.5, and -0.9). In all nine cases, we find that both the EMW and WZ tests still have decent sizes, i.e., heteroskedasticity appearing only in ε^y will not

affect their size performances much. In terms of power, the WZ test outperforms the EMW test under the t_3 distribution, and both tests have similar powers under Gaussianity. In addition, when ε^y is exhibits more heteroskedasticity (i.e., when ρ is close to 0), the WZ test gains more power as heteroskedasticity pushes the unconditional innovation distribution further away from Gaussianity.

Figure 8 presents the results for case (ii) where both innovations, ε^y and ε^x, are heteroskedastic and correlated as modeled above. When ρ is close to zero, the size distortion becomes smaller, while for larger (absolute) values of ρ, both tests become more oversized, especially under heavy-tailed innovation distribution. The WZ test suffers less size distortion than the EMW test under t_3 distributions and, using t_3 reference marginal densities, the size distortion bcomes even smaller (see the bottom panel).

The small-sample counterparts of Figure 7 and Figure 8 with $T = 200$ are provided in Appendix D. We draw conclusions similar to the i.i.d. case in Section 5.1. Additionally, we find that, when both ε^y and ε^x are heteroskedastic and their correlation is close to -1, both the EMW and WZ tests are less over-sized in the small-sample case.

These conclusions above also apply to other GARCH settings with different value chosen for parameters. These simulation results are available upon request.

5.3 Simulations under estimated reference density

In this section, we provide simulation results for the WZ test based on nonparametrically estimated reference densities, i.e., $g_y = \hat{f}_y$ and $g_x = \hat{f}_x$, under the i.i.d. setting as in Section 5.1.

In Figure 9, we compare the WZ test with $g_y = \hat{f}_y$ and $g_x = \hat{f}_x$ (dotted lines) with the EMW test (dashed lines) and with as the WZ test using correctly specified reference marginal densities (solid lines). When both the true and the reference densities are Gaussian (right plot), we see that all three tests perform similarly with decent size and power properties. When the true innovation distribution is Student-t_3, all three tests control the sizes well, while in terms of power, both WZ tests outperform the Gaussian-based EMW test. The WZ test with estimated reference densities suffers a small efficiency loss due to the nonparametric estimation.

Figure 10 provides the small-sample results under the same setting but with sample size $T = 200$. In general, the smaller sample leads to lower size and power for the WZ test with estimated reference densities relative to the large-sample case. But again, when f is heavy-tailed, it can be more powerful than the EMW test.

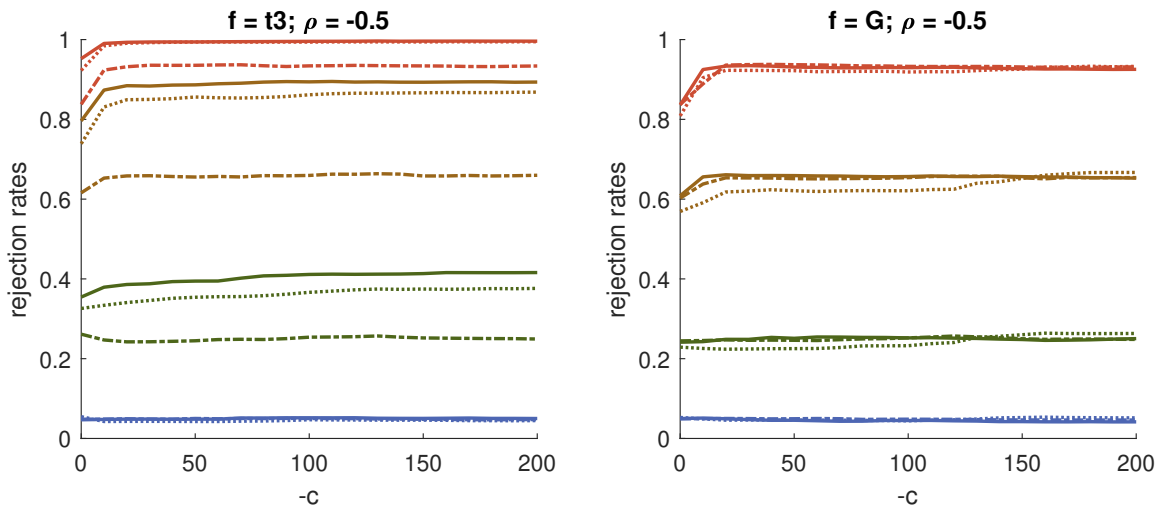

Figure 9: Rejection rates of the WZ test (solid lines), the EMW test with Student-t_3 marginal reference densities (dashed lines), and the EMW test with nonparametrically estimated density $\hat{f}$ (dotted lines) for different values of $\delta = 0$, 1, 2, and 3, corresponding to lines in blue, green, brown, and red, respectively. For all the four cases, $T = 2,000$.

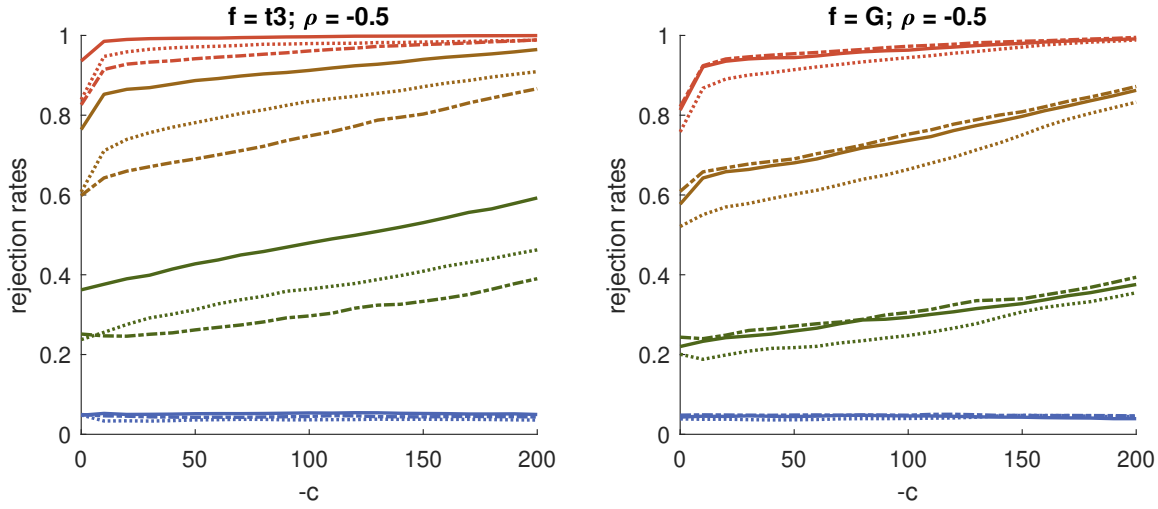

Figure 10: Rejection rates of the WZ test (solid lines), the EMW test with Student-t_3 marginal reference densities (dashed lines), and the EMW test with nonparametrically estimated density $\hat{f}$ (dotted lines) for different values of $\delta = 0$, 1, 2, and 3, corresponding to lines in blue, green, brown, and red, respectively. For all the four cases, $T = 200$.

6 Conclusion

In this paper, we show that there is significant statistical information, when supported by the application at hand, in a maintained assumption of serially indepen-

dent innovations in a predictive regression model. We exploit this information by deriving the (maximal) invariance structures in the associated limit experiment.

Specifically, we first derive the maximal invariant in the (structural) limit experiment where the predictor's persistence parameter is assumed to be known. This leads to the semiparametric power envelope for test that are invariant with respect to the innovation density. The associated likelihood ratio thus gives the semiparametric counterparts of the Gaussian sufficient statistics of Jansson and Moreira (2006). Under non-Gaussianity, larger powers are possible than under Gaussianity; a well-known result in many classical statistical models. To eliminate the predictor's persistence nuisance parameter, we employ the ALFD approach recently proposed in Elliott et al. (2015).

Our analysis naturally leads to statistics based on the bivariate component-wise ranks of the innovations in the model. Our statistics involve a choice of reference densities that is, subject to some mild regularity conditions, largely arbitrary. Irrespective of the choice of reference densities, our test, for any (fixed) innovation density allowed, has correct asymptotic size. Under non-Gaussianity, even with incorrectly specified reference densities, our test have better power properties than existing tests in the literature that are derived under the assumption of Gaussian innovation densities. These alternative tests do not need serially independent innovations and, as a result, we precisely quantify the power improvements possible when such an assumption is supported by the data. Monte Carlo simulations corroborate our asymptotic results and illustrate that the rank-based tests also work well in smaller samples.

References

Andrews, D. W. and Ploberger, W. (1994), "Optimal tests when a nuisance parameter is present only under the alternative," *Econometrica: Journal of the Econometric Society*, 1383–1414.

Bickel, P. J. (1982), "On adaptive estimation," *The Annals of Statistics*, 647–671.

Boswijk, H. P. et al. (2005), "Adaptive testing for a unit root with nonstationary volatility," *UvA-Econometrics Discussion Paper*, 7.

Campbell, J. Y. and Yogo, M. (2006), "Efficient tests of stock return predictability," *Journal of financial economics*, 81, 27–60.

Cassart, D., Hallin, M., Paindaveine, D., et al. (2010), "On the estimation of cross-

information quantities in rank-based inference," in *Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis: A Festschrift in Honor of Professor Jana Jurečková*, Institute of Mathematical Statistics, pp. 35–45.

Cavanagh, C. L., Elliott, G., and Stock, J. H. (1995), "Inference in models with nearly integrated regressors," *Econometric theory*, 11, 1131–1147.

Chernoff, H. and Savage, I. R. (1958), "Asymptotic normality and efficiency of certain nonparametric test statistics," *The Annals of Mathematical Statistics*, 972–994.

Elliott, G., Müller, U. K., and Watson, M. W. (2015), "Nearly optimal tests when a nuisance parameter is present under the null hypothesis," *Econometrica*, 83, 771–811.

Elliott, G., Rothenberg, T. J., and James, H. (1996), "Stock, 1996,"Efficient tests for an autoregressive unit root,"," *Econometrica*, 64, 813–836.

Elliott, G. and Stock, J. H. (1994), "Inference in time series regression when the order of integration of a regressor is unknown," *Econometric theory*, 10, 672–700.

Hallin, M., Van Den Akker, R., and Werker, B. J. (2015), "On quadratic expansions of log-likelihoods and a general asymptotic linearity result," in *Mathematical Statistics and Limit Theorems*, Springer, pp. 147–165.

Jansson, M. (2008), "Semiparametric power envelopes for tests of the unit root hypothesis," *Econometrica*, 76, 1103–1142.

Jansson, M. and Moreira, M. J. (2006), "Optimal inference in regression models with nearly integrated regressors," *Econometrica*, 74, 681–714.

Jeganathan, P. (1995), "Some aspects of asymptotic theory with applications to time series models," *Econometric Theory*, 11, 818–887.

Kagan, A. and Landsman, Z. (1999), "Relation between the covariance and Fisher information matrices," *Statistics & Probability Letters*, 42, 7–13.

Le Cam, L. M. (1986), "Asymptotic methods in statistical theory," .

Lehmann, E. L. and Romano, J. P. (2006), *Testing statistical hypotheses*, Springer Science & Business Media.

Ling, S., McAleer, M., et al. (2003), "On adaptive estimation in nonstationary ARMA models with GARCH errors," *The Annals of Statistics*, 31, 642–674.

Mayer-Wolf, E. et al. (1990), "The Cramér-Rao functional and limiting laws," *The Annals of Probability*, 18, 840–850.

Moreira, M. J. and Mourão, R. (2016), "A critical value function approach, with an application to persistent time-series," *arXiv preprint arXiv:1606.03496*.

Müller, U. K. (2011), "Efficient tests under a weak convergence assumption," *Econometrica*, 79, 395–435.

Müller, U. K. and Elliott, G. (2003), "Tests for unit roots and the initial condition," *Econometrica*, 71, 1269–1286.

Phillips, P. C. (2014), "On confidence intervals for autoregressive roots and predictive regression," *Econometrica*, 82, 1177–1195.

Rudin, W. (1987), "Real and complex analysis," .

Van der Vaart, A. W. (2000), *Asymptotic statistics*, vol. 3, Cambridge university press.

Yang, G. L. and Le Cam, L. (2000), "Asymptotics in statistics: some basic concepts," *Berlin, German: Springer*.

Zhou, B. (2020), "A General Semiparametric Approach for LAN, LAMN, and LABF Experiments," *Working Paper*.

Zhou, B., van den Akker, R., and Werker, B. J. (2019), "Semiparametrically optimal hybrid rank tests for unit roots," *The Annals of Statistics*, 47, 2601–2638.