Departamento de
Economía
www.eco.uc3m.es

Unidad de
Excelencia
María de Maeztu

June 19th 2022

To Whom it May Concern

Hereby, I certify that Antonio Raiola is a Ph.D student at Universidad Carlos III de Madrid under my supervision.

Miguel A. Delgado
Proffessor of Econometrics.

Chi-Squared Testing for Conditional Moment Restrictions (DRAFT)

Antonio Raiola[*]
University Carlos III de Madrid

June 25, 2023

Abstract

We propose testing conditional moment restrictions (CMRs) using Chi-squared (χ^2) testing procedures. After partitioning the data into cells, χ^2 tests assess whether the discrepancy between the observed means of a generalized residual with the expected vectors of zeroes, under the null, within each cell arose by chance. This is equivalent to test regression specifications by comparing the average response variable with the expected mean under the null within each cell. In contrast to existing omnibus procedures, χ^2 tests offer straightforward implementation and exhibit a standard limit null distribution, ensuring accurate size. To enhance the power of the tests, we propose different partitioning algorithms that leverage information about the null hypothesis and, if specified, the alternative hypotheses. We show that even when the partition depends on the data, under mild restrictions on the complexity of the partitioning algorithm, the tests limit null distribution remains standard. Montecarlo simulations show the good performance of the χ^2 tests compared to omnibus proposals, particularly in high-dimensional environments and against high-frequency alternatives. Our proposed tests offer practical advantages of implementation ease but also exhibit robust statistical properties, making them highly effective in various scenarios.

[*] Universidad Carlos III de Madrid, Calle Madrid 126, 28903 Getafe (Madrid), Spain. Email: araiola@eco.uc3m.es

1 Introduction

Typically, conditional moment restrictions (CMR) arise from the analysis of tautological models entailing parametric specifications of some conditional moment. Models defined by CMRs include models where regression and heteroskedasticity are simultaneously parameterized without restrictions on the distribution, models identified by instrumental variables (see, e.g., Newey [1990]), non-linear simultaneous equations models, transformation models like the Box-Cox transformation or the accelerated failure time model (see Horowitz [1996], for instance), hazard models, to mention but a few. Testing the correctness of the posited conditional moment specification is a fundamental preliminary step for valid inferential claims about the model parameters.

Tests for regression specifications are the most common application of CMR testing in the literature. In these, given a random sample $\{Y_i, X_i\}_{i=1}^n$ from a response variable Y and a d_x-dimensional vector of explanatory variables X, the aim consists of validating a parametric specification of the regression function, $m(x) = \mathbb{E}\left[Y|X\right]$,

$$H_0 : m = m_{\theta_0} \text{ a.s. for some } \theta_0 \in \Theta$$

where $m_\theta(\cdot)$ is a parametric specification of $m(\cdot)$ indexed by elements of a suitable parameter space $\Theta \subset \mathbb{R}^{d_\theta}$.

There is a vast literature of tests for H_0 building upon omnibus tests for probability distribution model specifications, which we broadly categorize into two classes: minimum-distance tests and tests based on smoothers.

The formers compare an empirical integrated measure of the regression function, $\hat{M}_0(x) = n^{-1}\sum_{i=1}^n Y_i\tau(X_i, x)$, with its version imposing the restrictions under the null, $\hat{M}_\theta(x) = n^{-1}\sum_{i=1}^n m_\theta(X_i)\tau(X_i, x)$, where $\tau(\cdot, x)$ is a properly chosen kernel function. The most common choice of $\tau(\cdot, x)$ in the literature have been the indicator function, $\tau(X, x) = \mathbb{I}_{[-\infty,x]}(X)$ (see, e.g., Stute [1997], Koul and Stute [1999], Li et al. [2003], among many others) and the exponential function, $\tau(X, x) = \exp(ix'X)$, where $i = \sqrt{-1}$ denotes the imaginary unit (see Bierens [1982] and Bierens [1990]).

Tests based on smoothers, instead, compare a non-parametric estimate of the regression function $\hat{m}(x)$ with the model under the null; see Eubank and Spiegelman [1990], Hardle and Mammen [1993], Fan and Li [1996], Koul and Ni [2004], Guerre and Lavergne [2005], Li et al. [2016] for some examples. González-Manteiga and Crujeiras [2013] reviews the developments of the two approaches for testing regression specifications, while Delgado et al. [2006] extend them to general CMR testing.

These tests have their strength and limitations. The main advantage is the omnibus property for which the tests have theoretical power against any deviation from the null. However, they suffer from several limitations that restrict their use in practice.

Minimum-distance tests have been shown to possess local power only against alternatives in an unknown finite-dimensional space (see Escanciano [2009]). They suffer from the curse of dimensionality and have limited power toward high-frequencies alternatives (Durbin and Knott [1972]). Their limit null distribution is non-pivotal, and obtaining critical values requires using bootstrap techniques (Stute et al. [1998]).

Smoother-based tests are also ineffective in high-dimensional environments due to the curse of dimensionality, they have trivial power against alternatives approaching the null at the parametric rate, and the size properties depend in practice on the choice of the smoothing parameter. Despite having pivotal limit null distribution, they are often implemented with bootstrap techniques.

In the context of testing probability distributions, there is a third approach that has been overlooked within the framework of regression specification (or CMR) testing. This approach involves the use of Chi-squared (χ^2) tests. Essentially, if the variable X takes values from a finite set, $\mathcal{X}$ say, and the parameter θ_0 is known (i.e., under simple hypothesis), one can test H_0 by comparing the average value of Y for each $x \in \mathcal{X}$ with $m_{\theta_0}(x)$. In the case of general covariates, χ^2 tests expand on this principle by partitioning the data into cells and evaluating whether the difference between the observed average Y and the corresponding expected average within each cell, under the specification in the null, can be attributed to chance.

Unlike the other two approaches, χ^2 tests are expressed in familiar terms for applied econometrician, they are easy to implement and possess standard limit null distribution with excellent size accuracy.

Under mild restrictions on the partitioning algorithm complexity, the asymptotic properties of the test remain unchanged even when the cells' partition boundaries depend randomly on the data. Indeed, the main feature of χ^2 tests is the dependence of the set of detectable alternatives on the chosen classes, which motivates incorporating information about the model in the partitioning procedure. We consider, for instance, partitioning the data along the distribution of fitted values where the relationship between the residuals and the fitted values changes sign. Testing with this partitioning formalizes the statistical practice of looking at the residual-fitted values scatter plot to determine model misspecification (Tsai et al. [1998]). Montecarlo simulations show that χ^2 tests using this partitioning method outperform minimum-distance tests, particularly against high-frequencies deviations from the null. We also consider the partitioning through Neyman-Pearson classes (see Greenwood and Nikulin [1996] and Balakrishnan et al. [2013]) resulting from the points in $\mathcal{X}$ where the model under the null and the alternative cross. If the alternative is unspecified, the partitioning can be done using a non-parametric estimate of $m(\cdot)$ and splitting over the points where $\hat{m}(x) - m(x) = 0$. χ^2 tests built in this way are, in some sense, hybrid tests exploiting the information of the smoothers to aggregate residuals rather than generating them, as with smoothers-based tests. We show that such tests respond to optimality criteria.

The structure of the paper is as follows: In the next section, we introduce the tests for regression specifications. The regression framework allows a natural interpretation of the tests and drastically reduces the notational burden required for the more general case; in Section 3, we discuss the asymptotic properties of the tests under general dependence of the partition cells boundaries from the data; in Section 4, we introduce a set of partitioning algorithms designed to enhance the power of the tests. These are based on a lower-dimensional projection of the data, such as to avoid the curse of dimensionality and improve the ability of the tests to detect departures from the null; in Section 5, we analyze the behavior of the tests against Pitman's local alternatives. The main focus of the discussion is how the partition choice (and the number of cells) influence the sensitivity of the tests in detecting deviations from the null hypothesis; in Section 6, we present a Monte Carlo study that the finite sample performances of the proposed χ^2 tests. This study demonstrates the practical effectiveness of the tests in various scenarios; in the last section, we extend the tests to general CMR. The extension only requires a few additional adjustments.

2 Chi-Squared Tests for CMR

Let $\{Z_i\}_{i=1}^n = \{Y_i, X_i\}_{i=1}^n$ be an i.i.d sample from the $\mathbb{R}^{1+d_x}$-valued random vector $Z = (Y, X)$ with distribution P, where Y is the response variable and X is the d_x-th dimensional vector of explanatory variables with support $\mathcal{X} \subset \mathbb{R}^{d_x}$. If $\mathbb{E}\left[Y^2\right] < \infty$, the associated regression function, $m(x) := \mathbb{E}[Y|X = x]$, is well-defined and characterizes the (a.s.) optimal predictor of Y, in a mean square error sense. In this section, we analyze the problem of testing that $m(\cdot)$ belongs to a class of parametric regression functions,

$$H_0: \ m \in \{m_\theta : \theta \in \Theta\}, \tag{1}$$

for a suitable parameter space $\Theta \subset \mathbb{R}^{d_\theta}$. Thus, under H_0, there exists a $\theta_0 \in \Theta$ such that

$$\int_A Y dP = \int_A m_{\theta_0}(X) dP \ \text{ for all } A \in \sigma(X) \tag{2}$$

where $\sigma(X)$ is the sigma-field generated by X. Recall that (2) is the definition of the regression function for the specification in H_0 (e.g., see definition 34.1 in Billingsley [2013]), which is equivalent to the following orthogonality conditions,

$$H_0: \int_A \varepsilon_{\theta_0}(Z) dP = 0 \ \text{ for all } A \in \sigma(X),$$

with $\varepsilon_\theta(z) = y - m_\theta(x)$ denoting the regression error.

Intuitively, if X takes value in a finite set, and θ_0 is known (i.e. under simple hypothesis), one can test H_0 by simply comparing the average value of Y for each $x \in \mathcal{X}$ with $m_{\theta_0}(x)$. For any type of covariates, once the data is partitioned into L cells, say, the χ^2 test assess whether the difference between the expected and observed averages in each cell arose by chance.

Let $\mathbb{C}$ be a class of measurable sets in $\mathcal{X}$ from which the cells of each partition are drawn, and denote as $\mathbb{D}$ the class of partitions of $\mathcal{X}$ comprised of L sets from $\mathbb{C}$ (L is fixed for all n); that is,

$$\mathbb{D} = \left\{ \boldsymbol{\gamma} = (\gamma_1, ..., \gamma_L) \in \mathbb{C}^L : \bigcup_{l=1}^{L} \gamma_l = \mathcal{X},\ \gamma_l \bigcap \gamma_f = \emptyset,\ \forall l \neq f \right\}, \tag{3}$$

where γ_l and γ_f denote sets of the partition $\boldsymbol{\gamma}$. Troughout, we denote as $\mathbf{I}_{\boldsymbol{\gamma}}(x) = (\mathbb{I}_{\gamma_1}(x), ..., \mathbb{I}_{\gamma_L}(x))'$ the vector of indicator functions over the sets of $\boldsymbol{\gamma}$.

The building block of the χ^2 test statistics is the standardized vector of differences between the sample averages of Y (observed averages) and the corresponding average imposing the specification under H_0 in each cell,

$$\hat{\Phi}_{\boldsymbol{\gamma}}(\theta) = \sqrt{n}\left(\hat{\boldsymbol{\mu}}^0_{\boldsymbol{\gamma}} - \hat{\boldsymbol{\mu}}_{\boldsymbol{\gamma}}(\theta)\right) = \sqrt{n}\begin{bmatrix} \hat{\mu}^0_1 - \hat{\mu}_1(\theta) \\ \cdot \\ \cdot \\ \hat{\mu}^0_L - \hat{\mu}_L(\theta) \end{bmatrix}, \tag{4}$$

where $\hat{\mu}^0_l = \int_{\gamma_l} \hat{M}_0(dx) = n^{-1}\sum_{i=1}^n Y_i \mathbb{I}_{\gamma_l}(X_i)$ and $\hat{\mu}_l(\theta) = \int_{\gamma_l} \hat{M}_\theta(dx) = n^{-1}\sum_{i=1}^n m_\theta(X_i)\mathbb{I}_{\gamma_l}(X_i)$ are the empirical integrated regression function and its version under the null, respectively, evaluated at γ_l.

Of course, tests based on (4) are not omnibus but designed for detecting deviations from H_0 of the type,

$$H_1(\gamma) : \boldsymbol{\mu}^0_{\boldsymbol{\gamma}} - \boldsymbol{\mu}_{\boldsymbol{\gamma}}(\theta) \neq 0 \ \text{ for all } \theta \in \Theta,$$

where $\boldsymbol{\mu}^0_{\boldsymbol{\gamma}} = \mathbb{E}\left[Y \mathbb{I}_{\boldsymbol{\gamma}}(X)\right]$ and $\boldsymbol{\mu}_{\boldsymbol{\gamma}}(\theta) = \mathbb{E}\left[m_\theta(X)\mathbb{I}_{\boldsymbol{\gamma}}(X)\right]$. In other words, they have trivial power when the average difference between $m(\cdot)$ and $m_\theta(\cdot)$ in each cell is zero (despite $m \neq m_\theta$ a.s.). The partitions, therefore, cover a fundamental role in implementing the test and provide a flexible tool to exploit the information given by the model or additional information available to the researcher. One, for instance, might consider Neyman-Pearson cells (Balakrishnan et al. [2013]) resulting from the points in $\mathcal{X}$ where the model under the null and a pre-specified alternative parametrization of $m(\cdot)$ meet (see Example 1 below). Under the pre-specified alternative, Neyman-Pearson classes maximize the difference between $m(\cdot)$ and $m_\theta(\cdot)$ in each cell.

Example 1 (Neyman-Pearson Classes, Balakrishnan et al. [2013]) *Consider testing the linear model $m_{\theta_0}(X_i) = \theta_{00} + X_i\theta_{01}$ against the alternative specification,*

$$H_1 : \tilde{m}_{\theta_0}(X_i) = \theta_{00} + \theta_{01}X_i + \theta_{02}\sin\left(\frac{50X_i}{2\pi}\right)$$

where θ_0 is a known vector. Neyman-Pearson classes split $\mathcal{X}$ over the points where $\theta_{02}\sin\left(\frac{50X_i}{2\pi}\right) = 0$. As a result, under H_1, $m_{\theta_0}(\cdot)$ is strictly bigger or strictly smaller than $\tilde{m}_{\theta_0}(\cdot)$ within each cell, and most cell-specific errors have the same sign, implying that the average error of a single cell is larger than the average error over the union of two contiguous cells (in absolute terms). As a matter of fact, in this example, the average error of (any) two contiguous cells is close to zero.

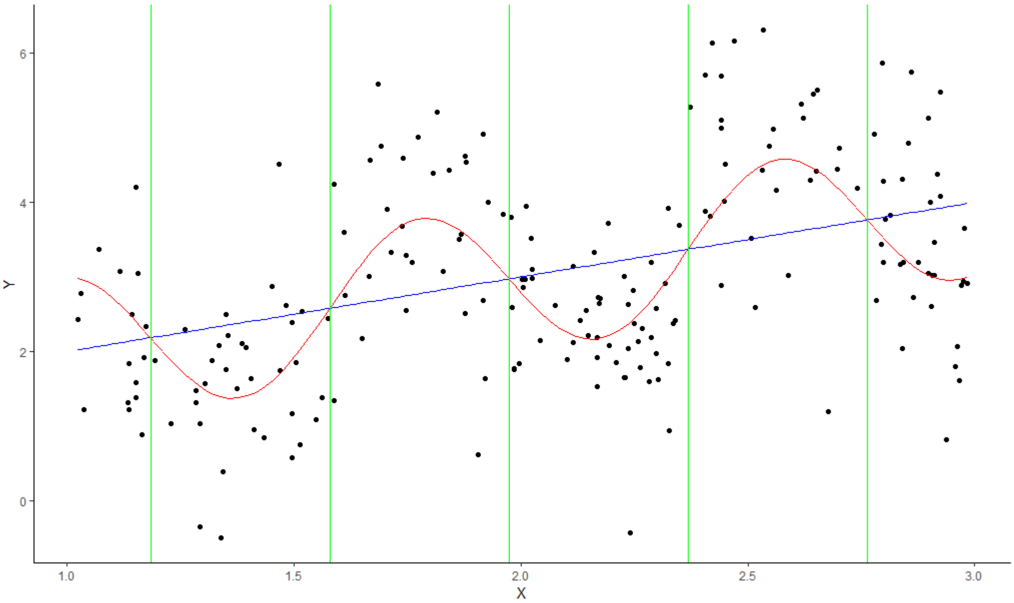

Figure 1: The graph depicts a random draw from the model under H_1 with $\theta_0^* = (1, 1, 1)$. The green lines depict the points where the model under the null (blue line) and under the alternative (red line) meet.

Under simple hypothesis, i.e. when θ_0 is known, by the central limit theorem, under the null,

$$\hat{\Sigma}_\gamma(\theta_0)^{-1/2}\hat{\Phi}_\gamma(\theta_0) \xrightarrow{d} N(0, I_L),$$

where $\hat{\Sigma}_\gamma(\theta) = n^{-1}\sum_{i=1}^n \varepsilon_\theta^2(Z_i)\mathbf{I}_\gamma(X_i)\mathbf{I}_\gamma(X_i)'$ estimates

$$\Sigma_{\gamma,0} := \text{Avar}\left(\hat{\Phi}_\gamma(\theta_0)\right) = \mathbb{E}\left[\varepsilon_{\theta_0}^2(Z)\mathbf{I}_\gamma(X)\mathbf{I}_\gamma(X)'\right] = \text{diag}\{\sigma_{0,1}^2, ..., \sigma_{0,L}^2\}, \qquad (5)$$

under H_0, with $\sigma_{0,l}^2 = \sigma_l^2(\theta_0)$, and $\sigma_l^2(\theta) = n\mathbb{E}\left(\hat{\mu}_l^0 - \hat{\mu}_l(\theta)\right)^2 = \mathbb{E}\left[\varepsilon_\theta^2(Z)\mathbb{I}_\gamma(X_i)\right]$.

Thus, taking for granted that $\rho_l = \mathbb{E}\left[\mathbb{I}_{\gamma_l}(X)\right] > 0$ for all l, under H_0,

$$\hat{\chi}_{\gamma,0}^2(\theta_0) \xrightarrow{d} \chi_L^2,$$

where

$$\hat{\chi}_{\gamma,0}^2(\theta) = \hat{\Phi}_\gamma(\theta)'\hat{\Sigma}_\gamma(\theta_0)^{-1}\hat{\Phi}_\gamma(\theta) = n\sum_{l=1}^{L}\frac{\left(\hat{\mu}_l^0 - \hat{\mu}_l(\theta)\right)^2}{\sigma_l^2} \tag{6}$$

When θ_0 is unknwon, the criterion in (6) suggests the following minimum distance estimator (hereafter, grouped GMM estimator),

$$\hat{\theta}_{\gamma,\tilde{\theta}} = \arg\min_{\theta\in\Theta} \chi_{\gamma,\tilde{\theta}}^2(\theta) \tag{7}$$

where

$$\hat{\chi}_{\gamma,\tilde{\theta}}^2(\theta) = \hat{\Phi}_\gamma(\theta)'\hat{\Sigma}_\gamma(\tilde{\theta})^{-1}\hat{\Phi}_\gamma(\theta) = n\sum_{l=1}^{L}\frac{\left(\hat{\mu}_l^0 - \hat{\mu}_l(\theta)\right)^2}{\hat{\sigma}_l^2(\tilde{\theta})},$$

$\hat{\sigma}_l^2(\theta) = n^{-1}\sum_{i=1}^n \varepsilon_\theta^2(Z_i)\mathbb{I}_{\gamma_l}(X_i)$, and $\tilde{\theta}$ is some initial $\sqrt{n}$-consistent estimator of θ_0. Henceforth, we drop the dependence on $\tilde{\theta}$.

The estimator $\hat{\theta}_\gamma$, analogous to the limited information estimator (or multinomial maximum-likelihood estimator) of the χ^2 test developed by Pearson [1900] and Fisher [1925] for goodness-of-fit distribution model checking (see Cramér [1946]), is a non-linear GLS on the aggregated data,

$$\hat{\theta}_\gamma = \left[\sum_{l=1}^{L}\frac{\hat{\mu}_l^\circ(\tilde{\theta})\hat{\mu}_l^\circ(\tilde{\theta})'}{\hat{\sigma}_l^2(\tilde{\theta})}\right]^{-1}\sum_{l=1}^{L}\frac{\hat{\mu}_l^\circ(\tilde{\theta})\hat{\mu}_l^0}{\hat{\sigma}_l^2(\tilde{\theta})}$$

with $\hat{\mu}_l^\circ(\theta) = n^{-1}\sum_{i=1}^n \nabla m_\theta(X_i)\mathbb{I}_{\gamma_l}(X_i)$ and $\nabla m_{\bar{\theta}} = d/d\theta m_\theta|_{\theta=\bar{\theta}}$. It belongs to the class of minimum-distance estimators considered by Koul and Ni [2004], with the main difference that the regressogram is used instead of kernels (and the weighting for heteroskedasticity to improve efficiency). Indeed, it is straightforward to see that,

$$\hat{\chi}_\gamma^2(\theta) = n\sum_{l=1}^{L}\frac{\left(\frac{\hat{\mu}_l^0}{\hat{\rho}_l} - \frac{\hat{\mu}_l(\theta)}{\hat{\rho}_l}\right)^2}{\left(\frac{\hat{\sigma}_l(\tilde{\theta})}{\hat{\rho}_l}\right)^2} = n\sum_{l=1}^{L}\frac{\left(\bar{\mu}_l^0 - \bar{\mu}_l(\theta)\right)^2}{\bar{\sigma}_l^2(\tilde{\theta})},$$

where $\hat{\rho}_l = n^{-1}\sum_{i=1}^{n}\mathbb{I}_{\gamma_l}(X_i)$ is the empirical measure of the l-th cell, and $\bar{\mu}_l^0$, $\bar{\mu}_l(\theta)$, and $\bar{\sigma}_l^2(\tilde{\theta})$ are the respective regressogram estimates of $m(x)$, $m_\theta(x)$, and $\mathrm{Var}(\varepsilon_{\theta_0}|X = x)$ for each $x \in \gamma_l$. Under suitable regularity conditions and the null H_0,

$$\sqrt{n}(\hat{\theta}_{\boldsymbol{\gamma}} - \theta_0) \xrightarrow{d} N\left(0, \left[\boldsymbol{\mu}^{\circ}_{\boldsymbol{\gamma},0}(\Sigma_{\boldsymbol{\gamma},0})^{-1}\boldsymbol{\mu}^{\circ\prime}_{\boldsymbol{\gamma},0}\right]^{-1}\right),$$

where $\boldsymbol{\mu}^{\circ}_{\boldsymbol{\gamma},0} = \boldsymbol{\mu}^{\circ}_{\boldsymbol{\gamma}}(\theta_0)$, with $\boldsymbol{\mu}^{\circ}_{\boldsymbol{\gamma}}(\theta) = (\mu^{\circ}_1(\theta), ..., \mu^{\circ}_L(\theta))'$ and $\mu^{\circ}_l(\theta) = \mathbb{E}\left[\hat{\mu}^{\circ}_l(\theta)\right]$ denoting the matrix of partial derivatives of $\boldsymbol{\mu}_{\boldsymbol{\gamma}}(\theta)$.

The minimized criterion, which we refer to as a $\hat{\chi}^2$ test statistics,

$$\hat{\chi}^2_{\boldsymbol{\gamma}} \coloneqq \min_{\theta\in\Theta}\hat{\chi}^2_{\boldsymbol{\gamma}}(\theta) = \hat{\chi}^2_{\boldsymbol{\gamma}}(\hat{\theta}_{\boldsymbol{\gamma}}), \tag{8}$$

is a J-test on the set of the L, out of the many, orthogonality conditions implied by the null,

$$\mathbb{E}\big[m(X)\mathbb{I}_{\gamma_l}(X)\big] = \mathbb{E}\big[m_{\theta_0}(X)\mathbb{I}_{\gamma_l}(X)\big] \text{ for all } l \in \{1, 2, .., L\}.$$

Thus, under the null H_0, and for $L > d_\theta$,

$$\hat{\chi}^2_{\boldsymbol{\gamma}} \xrightarrow{d} \chi^2_{L-d_\theta}.$$

The $\hat{\chi}^2$ test has a simple implementation: after estimating $\hat{\Sigma}_{\boldsymbol{\gamma}}(\tilde{\theta})$, one has to minimize the quadratic criterion and compare it with the appropriate quantile of the limit null distribution. However, the grouped GMM estimator requires global identification of θ_0 from the set of partitioned moments and the necessary order condition $L > d_\theta$. In cases where d_θ is large and n is relatively small, the asymptotic nature of the test hampers the use of too fine partitions, rendering the χ^2 test impractical.

Of course, it is also well motivated, as suggested in classical goodness-of-fit χ^2 tests (e.g. Nikulin [1973]), to verify the null using the Wald testing principle based on $\hat{\Phi}_{\boldsymbol{\gamma}}(\tilde{\theta})$ for some other $\sqrt{n}$-consistent estimator $\tilde{\theta}$,

$$\hat{\mathcal{W}}_{\boldsymbol{\gamma}}(\hat{\theta}) \coloneqq \hat{\Phi}_{\boldsymbol{\gamma}}(\tilde{\theta})\widehat{\mathrm{Avar}}^{-}\left(\hat{\Phi}_{\boldsymbol{\gamma}}(\tilde{\theta})\right)\hat{\Phi}_{\boldsymbol{\gamma}}(\tilde{\theta}), \tag{9}$$

where $\widehat{\text{Avar}}^{-}\left(\hat{\Phi}_{\gamma}(\tilde{\theta})\right)$ is a consistent estimator of some generalized inverse of $\text{Avar}\left(\hat{\Phi}_{\gamma}(\tilde{\theta})\right)$. Therefore, under the null and regularity conditions,

$$\hat{\mathcal{W}}_{\gamma}(\tilde{\theta}) \xrightarrow{d} \chi^2_{rank\left(\text{Avar}\left(\hat{\Phi}_{\gamma}(\tilde{\theta})\right)\right)}.$$

The Wald test allows the use of any $\sqrt{n}$ consistent estimator of θ_0, including minimum-distance estimators such as the one of Domínguez and Lobato [2004]. Notably, these estimators do not require any additional identification assumptions beyond those already provided by the null hypothesis.

Taking for granted the asymptotic linearity of the estimator (see Assumption 3 in the next section), the covariance matrix of $\hat{\Phi}_{\gamma}(\tilde{\theta})$ is characterized as,

$$\text{Avar}\left(\hat{\Phi}_{\gamma}(\tilde{\theta})\right) = \begin{bmatrix} I_L & -\boldsymbol{\mu}^{\circ\prime}_{\gamma,0} \end{bmatrix} \begin{bmatrix} \Sigma_{\gamma,0} & C_{\gamma,0} \\ C'_{\gamma,0} & L_0 \end{bmatrix} \begin{bmatrix} I_L \\ -\boldsymbol{\mu}^{\circ}_{\gamma,0} \end{bmatrix}, \tag{10}$$

where $C_{\gamma,0} = \mathbb{E}\left[\varepsilon_{\theta_0}(Z)\mathbf{I}_{\gamma}(X)l_{\theta_0}(Z)'\right]$, $L_0 = \mathbb{E}\left[l_{\theta_0}(Z)l_{\theta_0}(Z)'\right]$, and $l_{\theta_0}(\cdot)$ is the influence function of $\tilde{\theta}$.

When $\text{Avar}\left(\hat{\Phi}_{\gamma}(\tilde{\theta})\right)$ is full rank (e.g., if $\varepsilon_{\theta_0}(\cdot)\mathbf{I}_{\gamma}(\cdot)$ and $l_{\theta_0}(\cdot)$ have linearly independent components), the Wald test can be performed on any finite splitting of the data. In this case, a valid choice of $\widehat{\text{Avar}}^{-}\left(\hat{\Phi}_{\gamma}(\tilde{\theta})\right)$ is given by the inverse of the plug-in estimator,

$$\hat{W}_{\gamma}(\tilde{\theta}) = \begin{bmatrix} I_L & -\hat{\boldsymbol{\mu}}^{\circ}_{\gamma}(\tilde{\theta}) \end{bmatrix} \begin{bmatrix} \hat{\Sigma}_{\gamma}(\tilde{\theta}) & \hat{C}_{\gamma}(\tilde{\theta}) \\ \hat{C}_{\gamma}(\tilde{\theta})' & \hat{L}(\tilde{\theta}) \end{bmatrix} \begin{bmatrix} I_L \\ \hat{\boldsymbol{\mu}}^{\circ\prime}_{\gamma}(\tilde{\theta}) \end{bmatrix}, \tag{11}$$

where $\hat{\boldsymbol{\mu}}^{\circ}_{\gamma}(\theta) = (\hat{\mu}^{\circ}_{1}(\theta), ..., \hat{\mu}^{\circ}_{L}(\theta))$, $\hat{C}_{\gamma}(\theta) = n^{-1}\sum_{i=1}^{n} l_{\theta}(Z_i)\varepsilon_{\theta}(Z_i)\mathbf{I}_{\gamma}(X_i)'$, and $\hat{L}(\theta) = n^{-1}\sum_{i=1}^{n} l_{\theta}(Z_i)l_{\theta}(Z_i)'$.

In the event that the covariance matrix is rank deficient, if the Moore-Penrose inverse of $\hat{W}_{\gamma}(\tilde{\theta})$, denoted as $\hat{W}^{+}_{\gamma}(\tilde{\theta})$, has rank converging in probability to the one of $\text{Avar}\left(\hat{\Phi}_{\gamma}(\tilde{\theta})\right)$, then $\hat{W}^{+}_{\gamma}(\tilde{\theta}) \xrightarrow{p} \text{Avar}^{+}\left(\hat{\Phi}_{\gamma}(\tilde{\theta})\right)$ (Theorem 2 of Andrews [1987]). However, this need not be the case (see, e.g., Schott [2016] page 222-224) and more complex methods might be required.

Notice that the estimation of the covariance matrix, when $\text{Avar}\left(\hat{\Phi}_{\gamma}(\tilde{\theta})\right)$ is full rank, can be avoided by using a random normalizing weighting matrix, as in Kuan and Lee [2006] (see also Kiefer et al. [2000] for an early reference). Similarly, the over-identification test of Lee et al. [2014] provides a robust version of the $\hat{\chi}^2$ test which does not require estimating $\hat{\Sigma}_\gamma(\tilde{\theta})$ (and, thus, $\tilde{\theta}$). In these cases, the limit null distribution is non-standard but pivotal.

In practice, however, the partition depends randomly on the data. This is the case, for instance, when the partition imposes approximately the same number of observations in each cell or when the cell boundaries depend on the unknown model parameters. Extending the convergence of the statistics above to these cases requires restricting the complexity of the partitioning algorithm, as we discuss in the next section.

3 Data-dependent Cells

When studying the large sample behavior of the statistics, it is crucial to address the inherent influence of the data on the selection of cells (Watson [1959]). Moore and Spruill [1975] were among the first to address this concern in the distribution model check literature providing a rigorous derivation of the null distribution of χ^2 tests with rectangular data-dependent cells. In a more general setting, Pollard [1979] established the result for cells of arbitrary form by utilizing a central limit theorem for empirical measures and Andrews [1988] extended the methodology to conditional distribution testing. In this section, we provide a similar result for the more general CMR testing framework. Specifically, we show that the grouped GMM estimator and the tests keep standard limit distribution when the partition is built with data-dependent cells.

To state the convergence results in a self-contained fashion, we list a minimal set of assumptions consisting of smoothness conditions and restrictions on the partitioning algorithm complexity.

Assumption 1 *(a) $\{Z_i = (Y_i, X_i')'\}_{i=1}^n$ is a sequence of i.i.d. random vectors with $\mathbb{E}|Y_i| < \infty$; (b) $\mathbb{E}\left[\varepsilon_{\theta_0}^2\right] < C$, with $C < \infty$; (c) Θ is a compact subset of $\mathbb{R}^{d_\theta}$ and θ_0 is an interior point of Θ.*

Assumption 2 *$m_\theta(\cdot)$ is twice continuously differentiable in a neighborhood Θ_0 of θ_0, with $\Theta_0 \subset \Theta$. The gradient, $\nabla m_\theta(\cdot) = d/d\theta m_\theta(\cdot)$, is bounded by a square-integrable function $R(\cdot)$ such that $\sup_{\theta\in\Theta_0} |\nabla^{(j)} m_\theta(\cdot)| \leq R(\cdot)$ for all $j \in \{1, .., d_\theta\}$, where $\nabla^{(j)}$ denotes the j-th partial derivative, and $\mathbb{E}\left[R(X)^2\right] < \infty$.*

Assumption 3

(a) The estimator $\tilde{\theta}$ satisfies the following asymptotic expansion under the null,

$$\sqrt{n}(\tilde{\theta} - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} l_{\theta_0}(Z_i) + o_p(1)$$

where $\mathbb{E}\left[l_{\theta_0}(Z)\right] = 0$ a.s., and $L_0 = \mathbb{E}\left[l_{\theta_0}(Z) l_{\theta_0}(Z)'\right]$ is a finite and non-singular matrix.

(b) The vector-valued function $l_\theta(\cdot)$ is twice continuously differentiable in a neighborhood Θ_0 of θ_0 with first partial derivatives bounded by a square-integrable function $R_2(\cdot)$ such that $\sup_{\theta\in\Theta_0} |\nabla^{(j)} l_\theta(\cdot)| \leq R_2(\cdot)$ for all $j \in \{1, .., d_\theta\}$ and $\mathbb{E}\left[R_2(Z)^2\right] < \infty$.

Assumptions 1 and 2 are common in the model check literature with omnibus tests (see Stute and Zhu [2002], for instance). Compared to papers developing χ^2 tests based on probability model (e.g., Tauchen [1985]), we require slightly higher smoothness condition of the regression function but leave completely unrestricted the data distribution. Assumption 3(a) holds for most of the estimators used in practice, such as least square or GMM estimators, as well as for identification-robust minimum-distance estimators. While Assumption 3(b) is a technical requirement for the consistency of the plug-in estimator $\hat{W}_\gamma(\tilde{\theta})$.

We also state the necessary global identification and finite-variance conditions for the consistency and asymptotic normality of the grouped GMM estimator.

Assumption 2' *(a) $\Sigma_{\gamma,0}$ is positive definite; (b) $\mathbb{E}\left[m(X)\mathbf{I}_\gamma\right] = \mathbb{E}\left[m_\theta(X)\mathbf{I}_\gamma\right]$ if and only if $\theta = \theta_0$; (c) $\left[\boldsymbol{\mu}^{\circ}_{\gamma,0}(\Sigma_{\gamma,0})^{-1}\boldsymbol{\mu}^{\circ\prime}_{\gamma,0}\right]^{-1}$ is non-singular.*

Following Pollard [1979] and Andrews [1988], the data-dependent partitions are modeled as random functions over a class of properly restricted measurable sets (other approaches, like the one of Tauchen [1985], postulate the dependence of cells trough a finite dimensional parameter). Specifically, we equip $\mathbb{C}$ with the topology generated by the $L^2(F_x)$ semi-norm, F_x being the distribution of X under P, and give $\mathbb{D}$ the corresponding product topology. This means that two set C_1, C_2 in $\mathcal{X}$ are close if $F_x(C_1\tilde{\Delta}C_2)$ is small, where $F_x(C) = \int_C dF_x$ and $\tilde{\Delta}$ denotes the symmetric difference operator, $C_1\tilde{\Delta}C_2 = C_1 \cup C_2 \backslash C_1 \cap C_2$. Then, for each sample size n, the corresponding partition is given by a measurable mapping $\hat{\boldsymbol{\gamma}}$ from the underlying probability space to $\mathbb{D}$ converging in probability to some fixed partition of cells $\boldsymbol{\gamma}$ in $\mathbb{D}$; that is, for all $\epsilon > 0$,

$$P\left(F_x(\hat{\gamma}_l\tilde{\Delta}\gamma_l) > \epsilon\right) \to 0 \text{ as } n \to \infty, \text{ for all } l = 1, 2, ..., L.$$

Assumption 4 $\hat{\boldsymbol{\gamma}} \xrightarrow{p} \boldsymbol{\gamma}$ *for some fixed set of cells* $\boldsymbol{\gamma} \in \mathbb{D}$

Crucially, deriving the limit null distribution requires bounding the complexity of the partitions employed for the test construction. We do so by assuming that the cells are drawn from a class with finite Vapnik-Cervonenkis (VC) dimension.

Assumption 5 $\mathbb{C}$ *is a VC class of sets.*

The assumption is convenient because is independent of the data distribution but general enough for the purpose at hand. For instance, algorithms generating a finite number of straight edges and the class of hyper ellipsoids are VC classes. Furthermore, unions, intersections, differences, and complements of VC classes are also VC classes (Andrews [1988] and Pollard [1984] provide a thorough discussion, see also Section 2.6 in Van Der Vaart [1996]). A less stringent condition, assumes $\mathbb{C}$ being a Donsker class for the underlying probability measure (Pollard [1979]). While allowing for a wider range of admissible partitioning, Donsker assumptions are harder to verify in practice.

The following theorems show that the asymptotic distribution of the grouped GMM estimator and of the test statistics are unaffected by data-dependent cells. All the proofs are relegated to the appendix.

Theorem 1 *Let Assumptions 1, 2, 2', 4, 5 hold. Then, under the null hypothesis* H_0*,*

$$\sqrt{n}(\hat{\theta}_{\hat{\boldsymbol{\gamma}}} - \theta_0) \xrightarrow{d} N\left(0, \left[\boldsymbol{\mu}^{\circ}_{\boldsymbol{\gamma},0}(\Sigma_{\boldsymbol{\gamma},0})^{-1}\boldsymbol{\mu}^{\circ\prime}_{\boldsymbol{\gamma},0}\right]^{-1}\right),$$

Theorem 2 *Let Assumptions 1, 2, 4, 5, and the null hypothesis H_0 hold. Then,*

(a) Under Assumption 2' and $L > d_\theta$,

$$\hat{\chi}^2_{\hat{\gamma}} \xrightarrow{d} \chi^2_{L-d_\theta}.$$

(b) Under Assumption 3,

$$\hat{\mathcal{W}}_{\hat{\gamma}}(\tilde{\theta}) \xrightarrow{d} \chi^2_{rank\left(\text{Avar}\left(\hat{\Phi}_\gamma(\tilde{\theta})\right)\right)}.$$

If we further restrict the cells' dependence from the data, the $\hat{\chi}^2$ test, the Wald test, and the grouped GMM estimator with random cells are asymptotically equivalent to their fixed cell counterparts.

Theorem 3 *Under Assumptions 1 to 5 and the null hypothesis H_0, if*

$$\sqrt{n}\mathbb{E}\left[\varepsilon_{\theta_0}(Z)(\mathbf{I}_{\hat{\gamma}}(X) - \mathbf{I}_\gamma(X)\right] = o_p(1), \tag{12}$$

then, $\sqrt{n}(\hat{\theta}_{\hat{\gamma}} - \hat{\theta}_\gamma) = o_p(1)$, $\chi^2_{\hat{\gamma}} = \chi^2_\gamma + o_p(1)$, and $\hat{\mathcal{W}}_{\hat{\gamma}}(\tilde{\theta}) = \hat{\mathcal{W}}_\gamma(\tilde{\theta}) + o_p(1)$.

The restriction in equation (12) holds automatically when the partitioning depends only on the vector of covariates, as is the case for any unsupervised clustering method (see chapter 14 of Hastie et al. [2009]), or if it is independent of the data used for testing, such as when the partitioning algorithm is trained on a subset of the total observations. We illustrate some partitioning procedures in the next section, where we discuss partitioning based on unsupervised clustering and model-based methods, exploiting the information on the null and about pre-specified alternatives.

4 Partitioning Procedures

The first method presented below is an algorithm for splitting multi-variate data through the union of statistically equivalent blocks (SEB) (Gessaman [1970]). Other algorithms, such as the division in cubic cells or k nearest neighbor clustering are equally valid choices. The SEB algorithm is simple to apply and agrees with the statistical practice of dividing the data into equiprobable cells (see Greenwood and Nikulin [1996]) when testing against

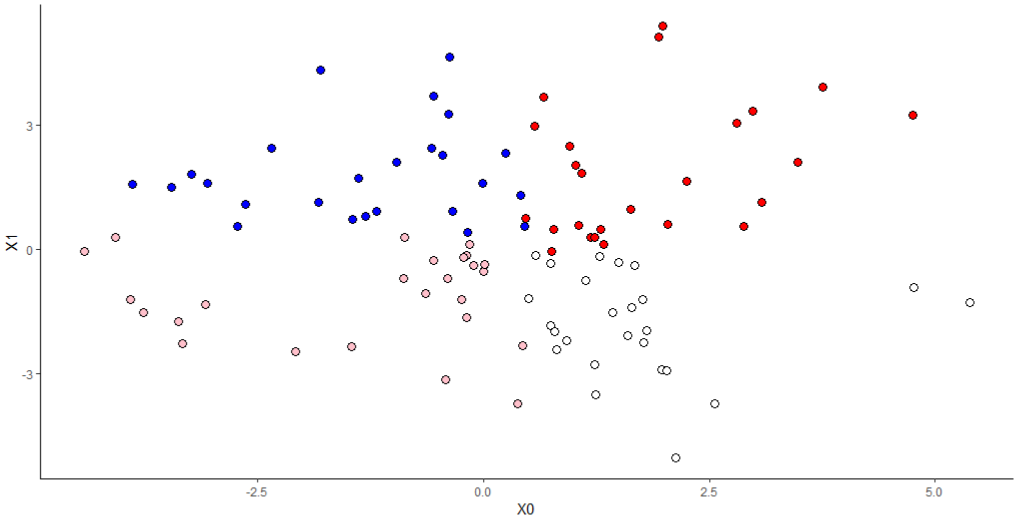

Figure 2: SEB partitioning of random sample from a bivariate standard normal with $S = 2$.

unknown probability distributions. We discuss its application to non-parametric and model-based projections of the data. The aim of these methods is double: on the one hand splitting on projections allows to reduce the curse of dimensionality; on the other hand we expect better power properties by incorporating the CMR and information about possible departures from the null to make the classes.

4.1 Statistically Equivalent Blocks (SEB)

Let $\mathbb{X}$ denote the data matrix, with rows given by $\{X_i\}_{i=1}^n$, of dimension $n \times d_x$; the procedure consists of sequentially sorting the observations based on the value of each column and grouping them at each iteration in S blocks, $S > 1$. After the initial sorting, the marginal support is split into S blocks containing $\lfloor n/S \rfloor$ observations. In the second step, the observations in the s-th block are sorted based on the second column and then again split into S blocks. Proceeding in the obvious fashion, the algorithm generates a partition of the data with a total number of cells, L, equal to S^{d_x}. As the number of cells rises exponentially with d_x, applying directly the procedure to $\mathbb{X}$ is problematic for large or moderate dimension of the covariates set. Even for small choices of S, the final partition may be too fine for the validity of the test's asymptotic approximation.

To avoid excessively fine partitions, the data can be split by running the SEB algorithm on lower-dimensional projection of the data. There are several techniques to reduce the dimension of a matrix, classified into linear and non-linear methods, depending on the type of projection involved. The formers include Principal Components analysis (PCA), Linear Discriminant Analysis, Singular Value Decomposition, etc. The latters Kernel PCA, Multidimensional Scaling, Isomapping, etc.

We briefly describe the PCA method, which is employed in the Montecarlo simulations of the last section. Let $\tilde{\mathbb{X}}$ be the re-scaled version of $\mathbb{X}$ where each value is standardized by the mean and variance of the column and consider the $d_x \times q$ matrix V_q containing the first q eigenvectors of $\tilde{\mathbb{X}}\tilde{\mathbb{X}}'$ (meaning the eigenvectors associated to the q largest eigenvalues). Using the projection of $\tilde{\mathbb{X}}$ on the first q principal components,

$$Z_q = \tilde{\mathbb{X}} V_q,$$

and the fact that V_q is orthornormal, $V_q V_q' = I_d$, we determine the q-dimensional reduction of $\tilde{\mathbb{X}}$ as,

$$\tilde{\mathbb{X}}_q = Z_q V_q'.$$

When $q = d_x$, the approximation is exact, $\tilde{\mathbb{X}} = \tilde{\mathbb{X}}_q$.

4.2 Fitted Values Method

Testing H_0 is equivalent to checking that $m_{\theta_0}(\cdot)$ is the best predictor of Y in a mean-square error sense. If the model is correctly specified, the vector of residuals, $\{\varepsilon_{\tilde{\theta}}(Z_i)\}_{i=1}^n$, resembles the vector of errors, $\{\varepsilon_{\theta_0}(Z_i)\}_{i=1}^n$, and, thus, tends to be mean-independent from the vector of optimal predictors $\{m_{\tilde{\theta}}(X_i)\}_{i=1}^n$. However, when the model is misspecified, irregularities in the relationship between the two variables arise. For instance, if we fit a linear model when the relationship between Y and X is quadratic, $Y = X^2 + \epsilon$ say, the resulting relationship between $\{\varepsilon_{\tilde{\theta}}(Z_i)\}_{i=1}^n$ and $\{m_{\tilde{\theta}}(X_i)\}_{i=1}^n$ will also be quadratic; that is, $\varepsilon_{\tilde{\theta}}(Z_i) = (m_{\tilde{\theta}}(X_i)/\tilde{\theta})^2 - m_{\tilde{\theta}}(X_i) + \epsilon_i$. More in general, if the relationship between Y and X is non-linear, the residuals depends non-linearly on the fitted values.

This suggests a simple strategy to incorporate the information of the model under the null in the partitioning algorithm. After estimating the unknown parameters θ_0, the data is split along the vector of fitted values by looking at the scatter plot and cutting (approximately) over the points where the relationship between $\varepsilon_{\tilde{\theta}}(\cdot)$ and $m_{\tilde{\theta}}(\cdot)$ changes sign. The procedure is based on the statistical practice of looking at the residual-fitted values scatter plot to investigate model misspecification (e.g. Tsai et al. [1998]). By exploiting the dependence between the two estimates, partitioning with fitted values generates large aggregate residuals under the alternative and, thus, boosts the power of the test. To avoid building irregular cells with high variability in the number of observations, one can directly apply the SEB algorithm on the vector of fitted values, choosing the number of cells that maximize the aggregated residual (see Example 2 below). In the Montecarlo simulations of the next section, we show that tests built with this technique are very sensible to deviations from the null.

Example 2 (SEB on fitted values)

In this example, we consider testing the null of linearity $m_\theta(X_i) = 1 + \theta' X_i$ when the model is generated by high-frequency deviations around the slope,

$$Y = \theta_{01}' X_i + 1.2 \sin\left(\frac{50 X_i' \theta_{02}}{2\pi}\right) + \sigma(X_i)\epsilon$$

where, $\epsilon \sim N(0,1)$, $\sigma(X_i) = (\exp(X_1 + X_2 + X_3)/\mathbb{E}\left[\exp(X_1 + X_2 + X_3)\right])^{1/2}$, $X_i = (X_{i,1}, ..., X_{i,5})$, and θ_{01}, θ_{02} are vectors of ones. The scatter plots below show the relationship between the residuals and the fitted values under the null (left panel) and the alternative model (right panel). The blue line is a non-parametric curve[1] fitting $\mathbb{E}\left[\varepsilon_{\tilde{\theta}}(Z) | m_{\tilde{\theta}}(X)\right]$. As it happens, under the alternative and for large enough L, the groups generated by the SEB alghorithm on the vector of fitted values (green lines) spread unevenly around zero (red line) resulting in larger aggregate residuals.

[1] The non-parametric fits in Figure 3 and Figure 4 below are obtained using the bin scatter regression package in R, 'binsreg()', of Cattaneo et al. [2019]

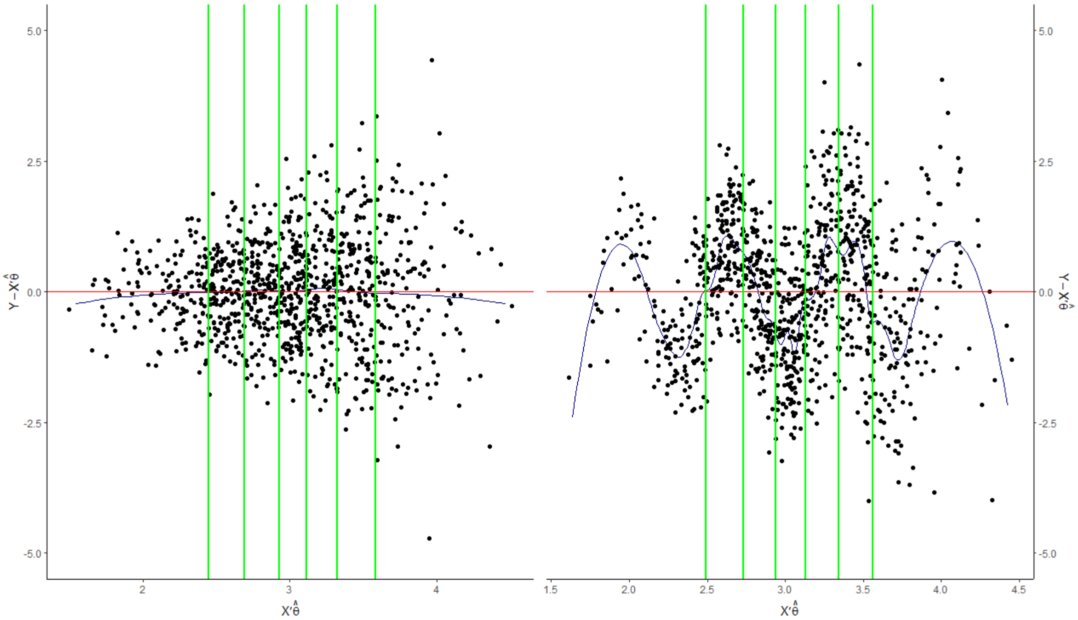

Figure 3: Scatter plot of $\varepsilon_{\tilde{\theta}}(\cdot)$ versus $m_{\tilde{\theta}}(\cdot)$ under the null (left panel) and under the alternative (right panel). The green lines depict the partitioning trough the SEB alghorithm with $L = 7$. The blue line is a non-parametric curve fitting the relationship between the two variables.

4.3 Neyman-Pearson Classes

Frequently, the researcher's primary concern is rejecting a subset of deviations critical for the application at hand. For instance, when estimating Mincer's earning regression, it is common to compare it with alternative specifications of the log-income profile (e.g. Polachek et al. [2008]). Similarly, in the Cox [1972] model, a parsimonious parametric specification of the baseline hazard is often compared to more flexible models (e.g. Seetharaman and Chintagunta [2003]).

When the interest is rejecting toward a given alternative parametric specification,

$$H_1 : m(x) = \tilde{m}_{\theta_0^*}(x) \;\text{ a.s. for some } \theta_0^* \in \Theta^* \subset \mathbb{R}^{d_{\theta^*}},$$

we partition the data using Neyman-Pearson classes, splitting over the points where $m_{\tilde{\theta}}(\cdot) = \tilde{m}_{\tilde{\theta}^*}(\cdot)$. By doing so, the difference between the average prediction under the null and the alternative model in each class is the largest possible, making it easier to detect deviations from the null toward H_1.

If, otherwise, the alternative is left unrestricted, we use non-parametric estimators to determine the optimal split. To motivate the procedure, notice that H_0 is equivalent to,

$$m_{\theta_0}(x) = m(x) \text{ a.s. for some } \theta_0 \in \Theta$$

where $m(x) = \mathbb{E}\left[Y|X = x\right]$. Therefore, if a non-parametric estimator of $m(\cdot)$, $\hat{m}(\cdot)$ say, is available, we implement Neyman-Pearson classes toward the non-parametric alternative by taking the points $x \in \mathcal{X}$ where $\hat{m}(x) - m_{\tilde{\theta}}(x) = 0$.

Example 3 (Parametric and Non-parametric Neyman-Pearson Classes) *Consider again testing the linear model, $m_{\theta_0}(x) = \theta_{00} + X_i\theta_{01}$ for some $\theta_0 \in \Theta$, against either one of the alternatives below,*

$$\begin{aligned} &H_1^a : \tilde{m}_{\theta_0^*}(x) = \theta_{00}^* + \theta_{01}^* X_i + \theta_{02}^* \sin\left(\frac{50X_i}{2\pi}\right) \;\; \textit{for some}\; \theta_0^* \in \Theta^*, \\ &H_1^b : m(x) \neq m_\theta \;\; \textit{for all}\; \theta \in \Theta, \end{aligned}$$

where θ_0 and θ_0^* are unknwon vectors. Against H_1^a (left panel of Figure 4), after obtaining the respective OLS estimates of the model under the null and under the alternative, Neyman-Pearson classes split over the points $x \in \mathcal{X}$ where $(\tilde{\theta}_{00} - \tilde{\theta}_{00}^*) + (\tilde{\theta}_{01} - \tilde{\theta}_{01}^*)x - \tilde{\theta}_{02}^* \sin(50x/2\pi) = 0$. Against H_1^b (right panel of Figure 4), we partition the data over the $x \in \mathcal{X}$ where $\tilde{\theta}_{00} + \tilde{\theta}_{01}x = \hat{m}(x)$, $\hat{m}(\cdot)$ being a non-parametric estimator of the regression curve.

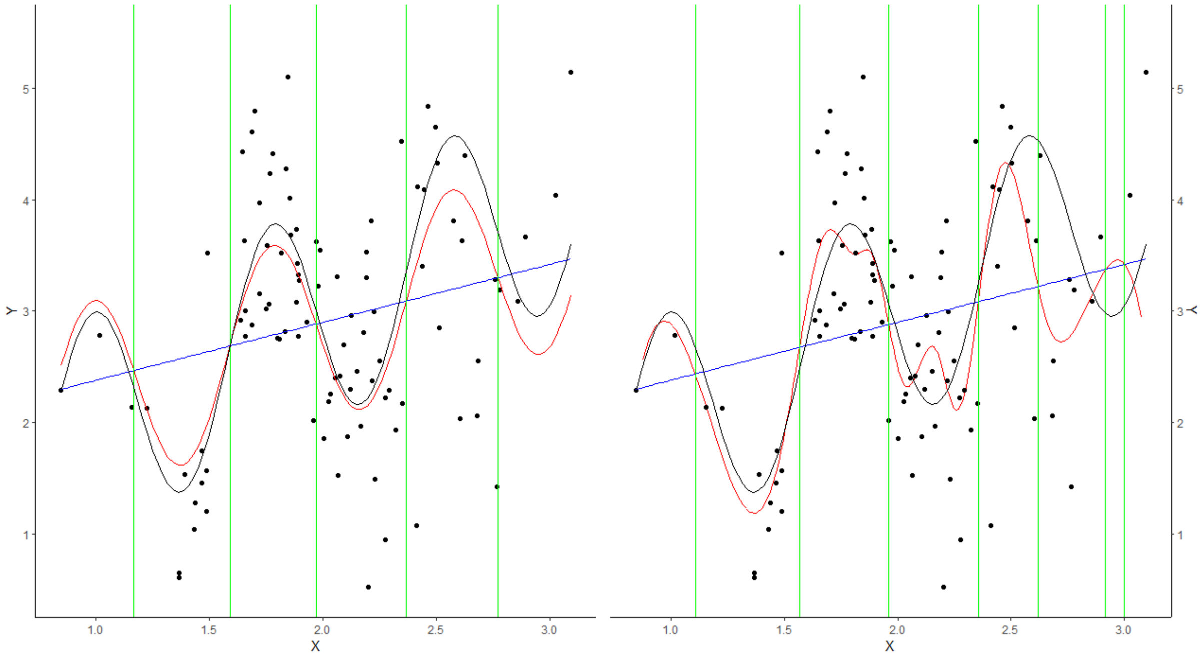

Figure 4: Neyman-Pearson classes with parametric (left panel) and non-parametric (right panel) estimates of the regression function. The green line depicts the points where the estimates under the null (blue line) and the estimates under the alternative (red line) meet. The data is generated under H_1^a (black line) with $\theta^* = (1, 1, 1)$.

Notably, the difference $\hat{m}(\cdot) - m_{\tilde{\theta}}(\cdot)$ is the building block of every omnibus test using smoothers. Chi-squared tests built with non-parametric Neyman-Pearson classes are, in some sense, hybrid tests, where the information given by the non-parametric comparison is used for aggregating the residuals rather than generating them. Indeed, the equivalence,

$$\begin{aligned}\mathbb{E}\left[\varepsilon_{\theta_0}(Z)|X\right] = 0 \iff & \mathbb{E}\left[\big(Y - m_{\theta_0}(X)\big)\mathbb{I}\{m(X) - m_{\theta_0}(X) > 0\}\right] = \\ & \mathbb{E}\left[\big(Y - m_{\theta_0}(X)\big)\mathbb{I}\{m(X) - m_{\theta_0}(X) \leq 0\}\right] = 0,\end{aligned}$$

shows that a parametric test against

$$H_1(\gamma_\theta) : \mathbb{E}\big[m(X)\mathbb{I}\{X \in \gamma_{\theta,l}\}\big] \neq \mathbb{E}\big[m_\theta(X)\mathbb{I}\{X \in \gamma_{\theta,l}\}\big] \;\text{ for all } \theta \in \Theta \text{ and some } l \in \{1, 2\},$$

where $\gamma_{\theta,1} = \{x \in \mathcal{X} : m(X) - m_\theta(X) > 0\}$ and $\gamma_{\theta,2} = \{x \in \mathcal{X} : m(X) - m_\theta(X) \leq 0\}$, is an omnibus test for H_0. Therefore, as far as the partitioning restrictions of Section 4 hold, by Theorem 3, a χ^2 test based on $\hat{\Phi}_{\hat{\gamma}}(\theta)$, where $\hat{\gamma} \xrightarrow{p} \gamma_{\theta_0}$ and $\hat{\Phi}_{\hat{\gamma}}(\theta_0) = \hat{\Phi}_{\gamma_{\theta_0}}(\theta_0) + o_p(1)$ under the null, is omnibus. Notice that condition (12) of Theorem 3, in this case, require the estimation of $\hat{\gamma}$ on an independent subsample of the data.

5 Power Analysis (WIP)

The power analysis of tests in the GMM class, such as χ^2 tests, commonly assumes a sequence of alternatives, known as Pittman drifts, where the unconditional moments converge to zero at the parametric rate. More general approaches, like the ones of Tauchen [1985] and Newey [1985], examine the power properties of the tests under distributional perturbations using different techniques such as Frechet and Gateaux differentiation.

In this paper, we focus on studying the local power of the test statistics under functional Pittman drift. The objective of our analysis is to investigate how the partition choice and the number of cells influence the non-centrality parameters and the overall power of the tests. Specifically, we consider the sequence of local alternatives,

$$H_{1,n} : m(x) = m_{\theta_0}(x) + \frac{1}{\sqrt{n}} h(x) \text{ a.s.,} \tag{13}$$

where $h(X)$ is a random variable representing departures from the null hypothesis with $\mathbb{E}\left[h(X)^2\right] < \infty$, $0 < P(h(X) = 0) < 1$, and $h(\cdot)$ is a differentiable function for all $x \in \mathcal{X}$.

Under simple hypothesis, after denoting as $\delta(\boldsymbol{\gamma}) = (\delta_1, ..., \delta_L)$, with $\delta_l = \mathbb{E}\left[h(X)\mathbb{I}_{\gamma_l}(X)\right]$, the vector of distances between the null and the alternative specification in each cell of $\boldsymbol{\gamma}$, the non-centrality parameter of both the Wald and the $\hat{\chi}^2$ test (which are numerically equivalent under simple hypothesis) is given by,

$$d_{\boldsymbol{\gamma}}(\Sigma_{\boldsymbol{\gamma},0}) = \delta(\boldsymbol{\gamma})'(\Sigma_{\boldsymbol{\gamma},0})^{-1}\delta(\boldsymbol{\gamma}) = \sum_{l=1}^{L} \frac{\delta_l^2}{\sigma_{0,l}^2}. \tag{14}$$

The quantity $d_{\boldsymbol{\gamma}}(\Sigma_{\boldsymbol{\gamma},0})$ can be intuitively understood in geometric terms. It represents the weighted area between the null specification and $m(\cdot)$, where the weights are determined by the average level of heteroskedasticity within each cell. In simpler words, the contribution of a cell to $d_{\boldsymbol{\gamma}}(\Sigma_{\boldsymbol{\gamma},0})$ is smaller when the within-cell noise is larger. This is because in such cases, the test is unable to distinguish the misspecification from the inherent noise.

Equation (14) highlights an important trade-off when choosing L. On one hand, the non-centrality parameter $d_{\boldsymbol{\gamma}}(\Sigma_{\boldsymbol{\gamma},0})$ increases as the number of cells grows. Specifically, if $q_l \neq 1$ for any $l = 1, .., L$, the inequality,

$$\frac{\delta_l^2}{q_l} + \frac{\delta_f^2}{q_f} \geq \frac{(\delta_l + \delta_f)^2}{q_l + q_f},$$

demonstrates that $d_{\boldsymbol{\gamma}}(\Sigma_{\boldsymbol{\gamma},0})$ rises for nested partitions. In other words, as the partition becomes finer with more cells, the test becomes more capable of detecting smaller deviations from the null hypothesis, resulting in increased power.

On the other hand, the power of the test under the alternative hypothesis $H_{1,n}$ decreases as the number of cells L increases. This decline in power is due to the higher variability of the limit null distribution associated with a larger number of cells. As the number of cells grows, the limit null distribution becomes more spread out, leading to a higher chance of observing test statistics falling in less extreme regions of the distribution and, thus, not rejecting.

If instead of the non-centrality parameter, we consider the euclidean norm of the drift,

$$d_{\boldsymbol{\gamma}}(I_L) = \delta(\boldsymbol{\gamma})'\delta(\boldsymbol{\gamma}) = \sum_{l=1}^{L} \delta_l^2 = \sum_{l=1}^{L} \mathbb{E}\left[h(X)\mathbb{I}_{\gamma_l}(X)\right]^2,$$

then, for any pair δ_l, δ_f such that $\operatorname{sgn}(\delta_l) = \operatorname{sgn}(\delta_f)$, the inequality $(\delta_l + \delta_f)^2 \geq \delta_l^2 + \delta_f^2$, suggests that an optimal partitioning for $d_{\boldsymbol{\gamma}}(I_L)$ is given by two cells containing the points where $h(x) = \sqrt{n}(m(x) - m_{\theta_0}(x))$ takes only positive and negative values, respectively; that is, the optimal partitioning of $d(\gamma, I_L)$ is given by two Neyman-Pearson classes. Below we formalize this intuition for the case $d_x = 1$.

Proposition 1 *Let $d_x = 1$, then $d(\delta, I_L) = ||\boldsymbol{\mu}^{\circ}_{\gamma,0}||$ is maximized by two Neyman-Pearson classes, $\gamma^* = \{\gamma_i^*\}_{i=1}^2$,*

$$\gamma_1^* = \{x \in \mathcal{X} : h(x) \geq 0\} \quad \gamma_2^* = \{x \in \mathcal{X} : h(x) < 0\}$$

This suggests that, in some sense, Neyman-Pearson classes correspond to optimization criteria for the non-centrality parameter. Of course, a rigorous optimization result should account for the dependence of the denominators on the cell boundaries. Doing so, has the inevitable consequence of determining splitting conditions that depend on the heteroskedasticity function.

Under composite hypothesis, ...

6 Montecarlo Study

The data is generated as,

$$Y_i = \theta' X_i + 0.8 \sin\left(\frac{c\sum_{j=1}^{d_x} X_{j,i}}{2\pi}\right) + \sigma(X_i)\epsilon_i$$

where $X_i = (X_{1,i}, ..., X_{d_x,i})$ the error distributes normally and independent from X, $\epsilon|X \sim N(0,1)$,

$$\sigma^2(X) = \frac{g(X)}{\mathbb{E}[g(X)]},$$

and $g(X) = e^{a(X_1+X_2+X_3)}$, with $a \in \{0, 0.5, 1\}$. The covariates $X_{j,i}$ are uniformly distributed with mean zero and variance $\sigma_x \in \{1/12, 1\}$. When σ_x is large, θ_0 is estimated more precisely but the errors have bigger variance. The parameters a and c control the level of heteroskedasticity in the errors, with the heteroskedasticity function normalized such that $\mathbb{E}[\sigma^2(X)] = 1$, and $c \in \{0, 10, 20, 50\}$, and the deviations from the null hypothesis, respectively. In the null model, we have $c = 0$, indicating no departures from linearity. As the value of c increases, the deviations occur at higher frequencies, becoming harder to distinguish from the sampling error. To avoid excessive computations, under the alternative, i.e. $c \neq 0$, we fix $a = 1$. The other parameters are set as follows: $\theta = \mathbf{1}$, where $\mathbf{1}$ is a $d_x \times 1$ vector of ones, $n \in \{100, 200, 500, 1000\}$, and $d_x \in \{5, 10\}$.

The $\hat{\chi}^2$ test is built as described in Section 2 using the grouped GMM estimator,

$$\hat{\theta}_\gamma = \left[\sum_{l=1}^{L} \frac{\bar{X}_l \bar{X}_l'}{\hat{\sigma}_l^2(\tilde{\theta})}\right]^{-1} \sum_{l=1}^{L} \frac{\bar{X}_l \hat{\mu}_l^0}{\hat{\sigma}_l^2(\tilde{\theta})}.$$

where $\bar{X}_l = n^{-1}\sum_{i=1}^{n} X_i \mathbb{I}_{\gamma_l}(X_i)$, and $\tilde{\theta}$ is the OLS estimator.

Given our data-generating process, we can easily show that the influence function of $\tilde{\theta}$, $l_{\theta_0}(z) = \mathbb{E}\left[XX'\right]^{-1} x\varepsilon_{\theta_0}(z)$, and $\varepsilon_{\theta_0}(\cdot)\mathbf{I}_\gamma(z)$ have linearly independent components (however, the independence does not hold when X includes a constant term). Since the asymptotic variance Avar $\left(\hat{\Phi}\boldsymbol{\gamma}(\tilde{\theta})\right)$ is full rank, the Wald test is implemented using the plug-in estimator $\hat{W}\boldsymbol{\gamma}(\tilde{\theta})$,

$$\begin{aligned}\hat{W}_{\boldsymbol{\gamma}}(\tilde{\theta}) = \mathbb{E}_n\left[\varepsilon_{\tilde{\theta}}(Z_i)^2\mathbf{I}_{\boldsymbol{\gamma}}(X_i)\mathbf{I}_{\boldsymbol{\gamma}}(X_i)'\right] - \mathbb{E}_n\left[\mathbf{I}_{\boldsymbol{\gamma}}(X_i)X_i'\right]\mathbb{E}_n\left[X_iX_i'\right]^{-1}\mathbb{E}_n\left[\varepsilon_{\tilde{\theta}}(Z_i)X_i\mathbf{I}_{\boldsymbol{\gamma}}(X_i)'\right] - \\ - \mathbb{E}_n\left[\varepsilon_{\tilde{\theta}}(Z_i)^2\mathbf{I}_{\boldsymbol{\gamma}}(X_i)X_i'\right]\mathbb{E}_n\left[X_iX_i'\right]^{-1}\mathbb{E}_n\left[X_i\mathbf{I}_{\boldsymbol{\gamma}}(X_i)'\right] + \\ + \mathbb{E}_n\left[\mathbf{I}_{\boldsymbol{\gamma}}(X_i)X_i'\right]\mathbb{E}_n\left[X_iX_i'\right]^{-1}\mathbb{E}_n\left[\varepsilon_{\tilde{\theta}}(Z_i)2X_iX_i'\right]\mathbb{E}_n\left[X_iX_i'\right]^{-1}\mathbb{E}_n\left[X_i\mathbf{I}_{\boldsymbol{\gamma}}(X_i)'\right].\end{aligned}$$

The two tests are compared with two minimum-distance tests based on marked residuals processes indexed by real vectors (see Stute [1997] and Stute and Zhu [2002]). Let $R_{1,n}(x_1)$ and $R_{2,n}(x_2)$ be the marked empirical processes of the residuals respectively indexed by $x_1 \in \mathbb{R}^{d_x}$ and by $x_2 \in \mathbb{R}$,

$$\begin{aligned}R_{1,n}(x_1) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(Y_i - X_i'\tilde{\theta}\right)\mathbb{I}\{X_i \leq x_1\}, \\ R_{2,n}(x_2) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(Y_i - X_i'\tilde{\theta}\right)\mathbb{I}\{X_i'\tilde{\theta} \leq x_2\}.\end{aligned}$$

The test statistics consist of functionals $\psi(\cdot)$ of $R_{1,n}(\cdot)$ and $R_{2,n}(\cdot)$. In these simulations, we only consider the Kolmogorov-Smirnov functional,

$$\begin{aligned}\mathbf{KS1} &= \sup_{x\in\mathbb{R}^{d_x}} |R_{1,n}(x)|, \\ \mathbf{KS2} &= \sup_{x\in\mathbb{R}} |R_{2,n}(x)|.\end{aligned}$$

Since these tests have non-pivotal limiting distribution, we approximate it with a Wild bootstrap procedure illustrated below (see Stute et al. [1998]):

1. Estimate the model under the null and obtain $X_i'\tilde{\theta}$.

2. Extract $X_i^* = X_i$ and $Y_i^* = X_i'\tilde{\theta} + \varepsilon_{\tilde{\theta}}(Z_i)V_i^*$, where $\{V_i^*\}_{i=1}^n$ is an i.i.d sample from a distribution which assigns masses $(\sqrt{5}+1)/2\sqrt{5}$ and $(\sqrt{5}-1)/2\sqrt{5}$ to the points $(1-\sqrt{5})/2$ and $(\sqrt{5}+1)/2$.

3. Estimate

$$R^*_{1,n}(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (Y^*_i - \theta^* X_i)\, \mathbb{I}\{X_i \leq x\},$$
$$R^*_{2,n}(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (Y^*_i - \theta^* X_i)\, \mathbb{I}\{X_i'\hat{\theta} \leq x\},$$

   θ^* being the OLS estimator on the bootstrap sample, and obtain the bootstrap functionals $\psi(R^*_{1,n})$ and $\psi(R^*_{2,n})$.

4. Repeat (2) and (3) B times. The critical point for the nominal level α are given by the $B(1-\alpha)$-th order statistics of $\{\psi(R^*_{1,n,b})\}_{b=1}^{B}$ and $\{\psi(R^*_{2,n,b})\}_{b=1}^{B}$.

In all the simulations we set $B = 500$ bootstrap repetitions.

We consider partitioning with SEB on the first q principal Components (PSEB) and SEB on the estimated regression function under the null (FIT). Notice that both type of partitioning take values in a VC class (see, e.g., Problem 14 on p.152 of Van Der Vaart [1996]). The PSEB and FIT approaches respectively generate $L_1 = S_1^q$ cells and $L_2 = S_2$ cells, S_1 and S_2 being user-chosen parameters. To make the tests comparable we force PSEB and FIT to generate the same number of cells in all the simulations. Specifically, we set $S_1 = 2$, $q = \lceil \log_2(d_x + 2) \rceil$ ($\lceil x \rceil$ being the ceiling function), and $S_2 = S_1^q$, resulting in $L = 8$ cells for $d_x = 5$ and $L = 16$ cells for $d_x = 10$.

It is worth remarking that, in practice, the grouping on fitted values, as discussed in Section 4, takes the L cuts in the residual-fitted values scatter plot that maximize the aggregate residuals, while in these simulations we just ”blindly” split using SEB with a fixed value of L. Thus, the perfomance of the tests using the FIT method are expected to be suboptimal to an ”eyeballing” partitioning.

Incidentally, the marked residual process indexed by partitions generated with the FIT method,

$$\hat{\Phi}_{\hat{\gamma}}(\tilde{\theta}) = n^{-1/2} \sum_{i=1}^{n} \mathbb{I}\{X_i'\tilde{\theta} \in C_l\} \left[Y_i - X_i'\tilde{\theta} \right],$$

where $\bigcup_{l=1}^{L} C_l = \mathbb{R}$, is a finite cells version of the process used in the Stute and Zhu [2002] test.

Table 1 presents the rejection rates under the null hypothesis, revealing that both the χ^2 tests and the omnibus tests exhibit size bias for small sample sizes, high levels of heteroskedasticity, and highly volatile regressors. However, as the sample size increases, the bias tends to diminish for all the tests (with the exception of $KS1$ with $a = 1$ and $d_x = 10$). It is important to note that the validity of the χ^2 tests asymptotic approximation relies on the number of observations within each cell, and therefore observing bias when testing over $L = 16$ cells is not surprising. Furthermore, the bias tends to decrease as the ratio n/L becomes larger, especially for higher degrees of heteroskedasticity. Interestingly, the $\hat{\chi}^2$ test reach the nominal size faster than the Wald test.

In terms of power (Table 2), the χ^2 tests with FIT grouping demonstrate superior performance overall. The second best performing test, $KS2$, outperforms the Wald test with FIT grouping only when both σ_x and n are small. It is worth noting that when σ_x is large, the omnibus proposals perform relatively poorly compared to the χ^2 tests with FIT grouping. As is known in the literature, omnibus tests have limited power against high-frequency alternatives ($c = 50$), while both the Wald and the $\hat{\chi}^2$ with FIT exhibit a higher probability of detecting the alternative.

The Wald test consistently outperforms the $\hat{\chi}^2$ test across all the tested scenarios. However, it is important to note that the comparison between the two tests is influenced by their distinct limit null distributions. Intuitively, the Wald test leverages more information by implicitly incorporating the moments used for estimation, resulting in higher power. Conversely, the Wald test's limit null distribution has more degrees of freedom, leading to lower power. While in our specific setting the trade-off clearly favors the Wald test, it remains unclear in general which test would perform better. Similar considerations apply to classical χ^2 tests for probability distribution model checking (see, e.g., Moore and Spruill [1975]).

Table 1: Size

					PSEB		**FIT**			
σ_x	n	d_x	a	L	$\hat{\chi}^2$	$\hat{\mathcal{W}}$	$\hat{\chi}^2$	$\hat{\mathcal{W}}$	**KS1**	**KS2**
0.083	100	5	0.0	8	0.044	0.036	0.054	0.052	0.052	0.064
0.083	100	5	0.5	8	0.053	0.037	0.053	0.043	0.056	0.076
0.083	100	5	1.0	8	0.059	0.041	0.059	0.041	0.054	0.055
0.083	100	10	0.0	16	0.050	0.041	0.061	0.025	0.016	0.087
0.083	100	10	0.5	16	0.047	0.034	0.060	0.034	0.015	0.077
0.083	100	10	1.0	16	0.053	0.030	0.056	0.022	0.012	0.076
0.083	200	5	0.0	8	0.039	0.045	0.047	0.043	0.055	0.046
0.083	200	5	0.5	8	0.052	0.045	0.057	0.048	0.058	0.053
0.083	200	5	1.0	8	0.054	0.052	0.044	0.051	0.051	0.062
0.083	200	10	0.0	16	0.034	0.043	0.069	0.052	0.029	0.067
0.083	200	10	0.5	16	0.071	0.050	0.048	0.036	0.020	0.059
0.083	200	10	1.0	16	0.053	0.042	0.062	0.048	0.013	0.073
0.083	500	5	0.0	8	0.044	0.046	0.063	0.049	0.042	0.048
0.083	500	5	0.5	8	0.044	0.047	0.056	0.053	0.057	0.051
0.083	500	5	1.0	8	0.048	0.052	0.052	0.051	0.059	0.053
0.083	500	10	0.0	16	0.048	0.053	0.055	0.057	0.049	0.059
0.083	500	10	0.5	16	0.046	0.056	0.055	0.052	0.045	0.052
0.083	500	10	1.0	16	0.060	0.058	0.052	0.040	0.036	0.054
0.083	1000	5	0.0	8	0.039	0.050	0.051	0.050	0.061	0.062
0.083	1000	5	0.5	8	0.039	0.047	0.048	0.062	0.057	0.061
0.083	1000	5	1.0	8	0.053	0.045	0.056	0.049	0.056	0.056
0.083	1000	10	0.0	16	0.056	0.061	0.056	0.052	0.050	0.058
0.083	1000	10	0.5	16	0.064	0.054	0.048	0.051	0.046	0.041
0.083	1000	10	1.0	16	0.046	0.050	0.043	0.046	0.053	0.042
1	100	5	0.0	8	0.051	0.048	0.040	0.048	0.051	0.057
1	100	5	0.5	8	0.046	0.054	0.050	0.042	0.055	0.062
1	100	5	1.0	8	0.041	0.027	0.030	0.023	0.033	0.060
1	100	10	0.0	16	0.042	0.038	0.060	0.034	0.013	0.059
1	100	10	0.5	16	0.047	0.032	0.054	0.027	0.009	0.081
1	100	10	1.0	16	0.031	0.010	0.036	0.009	0.001	0.069
1	200	5	0.0	8	0.063	0.055	0.042	0.043	0.060	0.059
1	200	5	0.5	8	0.050	0.040	0.053	0.045	0.048	0.049
1	200	5	1.0	8	0.046	0.035	0.044	0.037	0.046	0.057
1	200	10	0.0	16	0.057	0.045	0.054	0.036	0.032	0.068
1	200	10	0.5	16	0.052	0.040	0.053	0.037	0.007	0.063
1	200	10	1.0	16	0.032	0.020	0.035	0.017	0.003	0.062
1	500	5	0.0	8	0.047	0.050	0.046	0.055	0.053	0.056
1	500	5	0.5	8	0.047	0.050	0.056	0.046	0.060	0.052
1	500	5	1.0	8	0.045	0.038	0.032	0.043	0.044	0.044
1	500	10	0.0	16	0.050	0.040	0.057	0.043	0.040	0.062
1	500	10	0.5	16	0.055	0.048	0.046	0.055	0.023	0.061
1	500	10	1.0	16	0.052	0.048	0.032	0.020	0.004	0.057
1	1000	5	0.0	8	0.055	0.065	0.056	0.056	0.056	0.063
1	1000	5	0.5	8	0.051	0.050	0.057	0.075	0.057	0.056
1	1000	5	1.0	8	0.057	0.058	0.055	0.044	0.060	0.061
1	1000	10	0.0	16	0.050	0.050	0.049	0.054	0.043	0.052
1	1000	10	0.5	16	0.050	0.044	0.041	0.039	0.037	0.050
1	1000	10	1.0	16	0.049	0.043	0.048	0.036	0.012	0.047

Estimated size at the nominal level $\alpha = 0.05$ of the $\hat{\chi}^2$ test and Wald test under PSEB and FIT grouping. σ_x, n, k, L, respectively denote the regressors variance, the sample size, the number of covariates, and the number of cells; a regulates the degree of heteroskedasticity.

Table 2: Power

					PSEB		**FIT**			
σ_x	n	d_x	c	L	$\hat{\chi}^2$	$\hat{\mathcal{W}}$	$\hat{\chi}^2$	$\hat{\mathcal{W}}$	**KS1**	**KS2**
0.083	100	5	10	8	0.127	0.304	0.352	0.595	0.600	0.648
0.083	100	5	20	8	0.153	0.142	0.515	0.683	0.121	0.436
0.083	100	5	50	8	0.054	0.042	0.109	0.116	0.058	0.081
0.083	100	10	10	16	0.081	0.056	0.368	0.356	0.042	0.513
0.083	100	10	20	16	0.055	0.039	0.263	0.260	0.015	0.310
0.083	100	10	50	16	0.054	0.042	0.042	0.038	0.009	0.074
0.083	200	5	10	8	0.206	0.702	0.651	0.980	0.907	0.952
0.083	200	5	20	8	0.283	0.284	0.848	0.993	0.166	0.846
0.083	200	5	50	8	0.069	0.047	0.303	0.416	0.063	0.120
0.083	200	10	10	16	0.112	0.118	0.775	0.949	0.119	0.869
0.083	200	10	20	16	0.045	0.048	0.644	0.890	0.023	0.615
0.083	200	10	50	16	0.058	0.048	0.105	0.120	0.027	0.070
0.083	500	5	10	8	0.357	0.977	0.894	1.000	1.000	1.000
0.083	500	5	20	8	0.468	0.489	0.966	1.000	0.350	1.000
0.083	500	5	50	8	0.044	0.061	0.773	0.951	0.055	0.375
0.083	500	10	10	16	0.207	0.311	0.976	1.000	0.368	1.000
0.083	500	10	20	16	0.050	0.069	0.961	1.000	0.040	0.997
0.083	500	10	50	16	0.049	0.049	0.526	0.749	0.031	0.110
0.083	1000	5	10	8	0.522	1.000	0.956	1.000	1.000	1.000
0.083	1000	5	20	8	0.596	0.602	0.991	1.000	0.648	1.000
0.083	1000	5	50	8	0.058	0.053	0.956	1.000	0.051	0.849
0.083	1000	10	10	16	0.353	0.593	0.996	1.000	0.678	1.000
0.083	1000	10	20	16	0.070	0.067	0.996	1.000	0.061	1.000
0.083	1000	10	50	16	0.052	0.047	0.948	1.000	0.049	0.368
1	100	5	10	8	0.041	0.036	0.666	0.974	0.056	0.146
1	100	5	20	8	0.049	0.031	0.360	0.563	0.047	0.073
1	100	5	50	8	0.045	0.039	0.051	0.041	0.050	0.068
1	100	10	10	16	0.034	0.026	0.474	0.770	0.003	0.117
1	100	10	20	16	0.027	0.026	0.165	0.198	0.001	0.068
1	100	10	50	16	0.035	0.021	0.061	0.032	0.004	0.064
1	200	5	10	8	0.061	0.067	0.853	1.000	0.054	0.240
1	200	5	20	8	0.051	0.043	0.642	0.946	0.052	0.083
1	200	5	50	8	0.037	0.041	0.071	0.072	0.041	0.063
1	200	10	10	16	0.041	0.031	0.824	1.000	0.002	0.161
1	200	10	20	16	0.049	0.033	0.499	0.827	0.006	0.085
1	200	10	50	16	0.049	0.035	0.041	0.050	0.010	0.062
1	500	5	10	8	0.075	0.085	0.956	1.000	0.052	0.751
1	500	5	20	8	0.048	0.040	0.908	0.999	0.042	0.135
1	500	5	50	8	0.046	0.053	0.110	0.178	0.054	0.073
1	500	10	10	16	0.045	0.039	0.976	1.000	0.018	0.414
1	500	10	20	16	0.042	0.031	0.915	0.999	0.016	0.130
1	500	10	50	16	0.044	0.046	0.134	0.176	0.016	0.059
1	1000	5	10	8	0.104	0.113	0.980	1.000	0.060	0.998
1	1000	5	20	8	0.049	0.054	0.972	1.000	0.060	0.267
1	1000	5	50	8	0.046	0.048	0.168	0.360	0.065	0.079
1	1000	10	10	16	0.045	0.043	0.997	1.000	0.030	0.868
1	1000	10	20	16	0.052	0.045	0.992	1.000	0.034	0.198
1	1000	10	50	16	0.047	0.047	0.320	0.615	0.031	0.069

Estimated power at the nominal level $\alpha = 0.05$ of the $\hat{\chi}^2$ test and Wald test under PSEB and FIT grouping. σ_x, n, d_x, L, respectively denote the regressors variance, the sample size, the number of covariates, and the number of cells; c governs departures from the null of linearity.

7 General CMRs (WIP)

To extend the analysis to general moment restrictions, we introduce a vector of response variables, Y, taking values in $\mathcal{Y} \subset \mathbb{R}^{d_y}$, with $d_y \geq 1$, and a generalized residual vector (Wooldridge [1990]), $\boldsymbol{\varepsilon}_\theta(\cdot) : \mathbb{R}^{d_y} \times \mathbb{R}^{d_x} \to \mathbb{R}^{d_\varepsilon}$ with $\boldsymbol{\varepsilon}_\theta(\cdot) = (\varepsilon_{1,\theta}(\cdot), ..., \varepsilon_{d_\varepsilon,\theta}(\cdot))'$, defining parametric relationships between Y and X. The null hypothesis is defined as before,

$$\int_A \boldsymbol{\varepsilon}_{\theta_0}(Z)dP = 0 \text{ for all } A \in \sigma(X). \tag{15}$$

The generality of this framework allows testing for a wide range of econometric models such as regression and heteroskedasticity models, transformation models like the Box-Cox transformation or the accelerated failure time model (see Horowitz [1996], for instance), simultaneous equation model identified by instrumental variables (Newey [1990]), etc.

When the dimension of the generalized residual is bigger than one, it might be optimal to consider a partition for each component of $\boldsymbol{\varepsilon}_\theta(\cdot)$. In particular, for each $j \in \{1, ..., d_\varepsilon\}$, let $\mathbb{D}_j$ be a class of partitions of $\mathcal{X}$ comprised of L_j sets from $\mathbb{C}$ (L_j is fixed for all n); that is,

$$\mathbb{D}_j = \left\{ \boldsymbol{\gamma}_j = (\gamma_{j,1}, ..., \gamma_{j,L_j})' \in \mathbb{C}^{L_j} : \cup_{l=1}^{L_j} \gamma_{j,l} = \mathcal{X},\ \gamma_{j,l} \cap \gamma_{j,f} = \emptyset,\ \forall l \neq f \right\}, \tag{16}$$

We let $\mathcal{E}_\theta(\cdot)$ be the $\bar{L} \times \bar{L}$ block diagonal matrix of generalized residuals with main diagonal elements given by $\{\varepsilon_{j,\theta}(\cdot) I_{L_j}\}_{j=1}^{d_\varepsilon}$, where $\bar{L} = \sum_j L_j$. If $L_1 = L_2 = \cdots = L_{d_\varepsilon} = L$, then $\mathcal{E}_\theta(\cdot) = \text{diag}[\boldsymbol{\varepsilon}_\theta(\cdot)] \otimes I_L$, where $\text{diag}[\boldsymbol{\varepsilon}_\theta(\cdot)] = \text{diag}\{\varepsilon_{1,\theta}(\cdot), ..., \varepsilon_{d_\varepsilon,\theta}(\cdot)\}$ is the $d_\varepsilon \times d_\varepsilon$ diagonal matrix with the components of $\boldsymbol{\varepsilon}_\theta(\cdot)$ on the main diagonal and $\otimes$ denotes the Kronecker product.

The χ^2 test statistics can be expressed as quadratic forms of

$$\hat{\Phi}_{\boldsymbol{\gamma}}(\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathcal{E}_\theta(Z_i) \mathbf{I}_{\boldsymbol{\gamma}}(X_i), \tag{17}$$

where $\mathbf{I}_{\boldsymbol{\gamma}}(\cdot) = (\mathbf{I}'_{\boldsymbol{\gamma}_1}, ..., \mathbf{I}'_{\boldsymbol{\gamma}_{d_\varepsilon}})'$ is the vector of indicator functions over all the partitions.

The covariance matrices of $\hat{\Phi}_{\gamma}(\theta_0)$ and $\hat{\Phi}_{\gamma}(\tilde{\theta})$ under the null are given by,

$$\Sigma_{\gamma,0} = \mathbb{E}\left[\mathcal{E}_{\theta_0}(Z)\mathbf{I}_{\gamma}(X)\mathbf{I}_{\gamma}(X)'\mathcal{E}_{\theta_0}(Z)\right], \tag{18}$$

and,

$$\operatorname{Avar}\left(\hat{\Phi}_{\gamma}(\tilde{\theta})\right) = \begin{bmatrix} I_{\bar{L}} & -\boldsymbol{\mu}^{\circ}_{\gamma,0} \end{bmatrix} \begin{bmatrix} \Sigma_{\gamma,0} & C_{\gamma,0} \\ C'_{\gamma,0} & L_0 \end{bmatrix} \begin{bmatrix} I_{\bar{L}} \\ -\boldsymbol{\mu}^{\circ\prime}_{\gamma,0} \end{bmatrix}, \tag{19}$$

where $\boldsymbol{\mu}^{\circ}_{\gamma,0} = \mathbb{E}\left[\nabla\mathcal{E}_{\theta_0}(Z)\mathbf{I}_{\gamma}(X)\right]$ is the Jacobian matrix, $C'_{\gamma,0} = \mathbb{E}\left[\mathcal{E}_{\theta_0}(Z)\mathbf{I}_{\gamma}(X)l_{\theta_0}(Z)'\right]$, and L_0 is defined as before.

As $\Sigma_{\gamma,0}$ is not diagonal anymore (unless $\boldsymbol{\varepsilon}_{\theta}(\cdot)$ has orthogonal components and $L = 1$, or $d_{\varepsilon} = 1$), the advantages of the $\hat{\chi}^2$ test, in terms of implementability, are loss in the general framework. Nonetheless, the $\hat{\chi}^2$ test is preferable to the Wald test when the rank of $\operatorname{Avar}\left(\hat{\Phi}_{\gamma}(\tilde{\theta})\right)$ is unkown.

[EFFICIENCY CONSIDERATIONS]

References

Donald WK Andrews. Asymptotic results for generalized wald tests. *Econometric Theory*, 3(3):348–358, 1987.

Donald WK Andrews. Chi-square diagnostic tests for econometric models: theory. *Econometrica: Journal of the Econometric Society*, pages 1419–1453, 1988.

Narayanaswamy Balakrishnan, Vassilly Voinov, and Mikhail Stepanovich Nikulin. *Chi-squared goodness of fit tests with applications.* Academic Press, 2013.

Herman J Bierens. Consistent model specification tests. *Journal of Econometrics*, 20(1): 105–134, 1982.

Herman J Bierens. A consistent conditional moment test of functional form. *Econometrica: Journal of the Econometric Society*, pages 1443–1458, 1990.

Patrick Billingsley. *Convergence of probability measures.* John Wiley & Sons, 2013.

Matias D Cattaneo, Richard K Crump, Max H Farrell, and Yingjie Feng. On binscatter. *arXiv preprint arXiv:1902.09608*, 2019.

David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

Harald Cramér. *Mathematical methods of statistics*, volume 26. Princeton university press, 1946.

James Davidson. *Stochastic limit theory: An introduction for econometricians.* OUP Oxford, 1994.

Miguel A Delgado, Manuel A Domínguez, and Pascal Lavergne. Consistent tests of conditional moment restrictions. *Annales d'Économie et de Statistique*, pages 33–67, 2006.

Manuel A Domínguez and Ignacio N Lobato. Consistent estimation of models defined by conditional moment restrictions. *Econometrica*, 72(5):1601–1615, 2004.

James Durbin and Martin Knott. Components of cramér–von mises statistics. i. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):290–307, 1972.

J Carlos Escanciano. On the lack of power of omnibus specification tests. *Econometric Theory*, 25(1):162–194, 2009.

Randall L Eubank and Clifford H Spiegelman. Testing the goodness of fit of a linear model via nonparametric regression techniques. *Journal of the American Statistical Association*, 85(410):387–392, 1990.

Yanqin Fan and Qi Li. Consistent model specification tests: omitted variables and semiparametric functional forms. *Econometrica: Journal of the econometric society*, pages 865–890, 1996.

Ronald Aylmer Fisher. Theory of statistical estimation. In *Mathematical proceedings of the Cambridge philosophical society*, pages 700–725. Cambridge University Press, 1925.

MP Gessaman. A consistent nonparametric multivariate density estimator based on statistically equivalent blocks. *The Annals of Mathematical Statistics*, 41(4):1344–1346, 1970.

Evarist Giné and Joel Zinn. Some limit theorems for empirical processes. *The Annals of Probability*, pages 929–989, 1984.

Wenceslao González-Manteiga and Rosa M Crujeiras. An updated review of goodness-of-fit tests for regression models. *Test*, 22:361–411, 2013.

Priscilla E Greenwood and Michael S Nikulin. *A guide to chi-squared testing*, volume 280. John Wiley & Sons, 1996.

Emmanuel Guerre and Pascal Lavergne. Data-driven rate-optimal specification testing in regression models. *The Annals of Statistics*, 33(2):840–870, 2005.

Wolfgang Hardle and Enno Mammen. Comparing nonparametric versus parametric regression fits. *The Annals of Statistics*, pages 1926–1947, 1993.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Joel L Horowitz. Semiparametric estimation of a regression model with an unknown transformation of the dependent variable. *Econometrica: Journal of the Econometric Society*, pages 103–137, 1996.

Nicholas M Kiefer, Timothy J Vogelsang, and Helle Bunzel. Simple robust testing of regression hypotheses. *Econometrica*, 68(3):695–714, 2000.

Hira L Koul and Pingping Ni. Minimum distance regression model checking. *Journal of Statistical Planning and Inference*, 119(1):109–141, 2004.

Hira L Koul and Winfried Stute. Nonparametric model checks for time series. *The Annals of Statistics*, 27(1):204–236, 1999.

Chung-Ming Kuan and Wei-Ming Lee. Robust m tests without consistent estimation of the asymptotic covariance matrix. *Journal of the American Statistical Association*, 101 (475):1264–1275, 2006.

Wei-Ming Lee, Chung-Ming Kuan, and Yu-Chin Hsu. Testing over-identifying restrictions without consistent estimation of the asymptotic covariance matrix. *Journal of Econometrics*, 181(2):181–193, 2014.

Hongjun Li, Qi Li, and Ruixuan Liu. Consistent model specification tests based on k-nearest-neighbor estimation method. *Journal of Econometrics*, 194(1):187–202, 2016.

Qi Li, Cheng Hsiao, and Joel Zinn. Consistent specification tests for semiparametric/nonparametric models based on series estimation methods. *Journal of Econometrics*, 112(2):295–325, 2003.

David S Moore and Marcus C Spruill. Unified large-sample theory of general chi-squared statistics for tests of fit. *The Annals of Statistics*, pages 599–616, 1975.

Whitney K Newey. Generalized method of moments specification testing. *Journal of econometrics*, 29(3):229–256, 1985.

Whitney K Newey. Efficient instrumental variables estimation of nonlinear models. *Econometrica: Journal of the Econometric Society*, pages 809–837, 1990.

Whitney K Newey and Daniel McFadden. Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245, 1994.

Mikhail Stepanovich Nikulin. Chi-square test for continuous distributions with location and scale parameters. *Teoriya Veroyatnostei i ee Primeneniya*, 18(3):583–591, 1973.

Karl Pearson. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900.

Solomon W Polachek et al. Earnings over the life cycle: The mincer earnings function and its applications. *Foundations and Trends® in Microeconomics*, 4(3):165–272, 2008.

David Pollard. General chi-square goodness-of-fit tests with data-dependent cells. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 50(3):317–331, 1979.

David Pollard. *Convergence of stochastic processes*. Springer Science & Business Media, 1984.

C Radhakrishna Rao and Sujit Kumar Mitra. Generalized inverse of a matrix and its applications. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Theory of Statistics*, volume 6, pages 601–621. University of California Press, 1972.

James R Schott. *Matrix analysis for statistics*. John Wiley & Sons, 2016.

PB Seetharaman and Pradeep K Chintagunta. The proportional hazard model for purchase timing: A comparison of alternative specifications. *Journal of Business & Economic Statistics*, 21(3):368–382, 2003.

W Stute, W González Manteiga, and M Presedo Quindimil. Bootstrap approximations in model checks for regression. *Journal of the American Statistical Association*, 93(441): 141–149, 1998.

Winfried Stute. Nonparametric model checks for regression. *The Annals of Statistics*, pages 613–641, 1997.

Winfried Stute and Li-Xing Zhu. Model checks for generalized linear models. *Scandinavian Journal of Statistics*, 29(3):535–545, 2002.

George Tauchen. Diagnostic testing and evaluation of maximum likelihood models. *Journal of Econometrics*, 30(1-2):415–443, 1985.

Chih-Ling Tsai, Zongwu Cai, and Xizhi Wu. The examination of residual plots. *Statistica Sinica*, pages 445–465, 1998.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Jon Wellner Van Der Vaart. *Weak convergence*. Springer, 1996.

Geoffrey S Watson. Some recent results in chi-square goodness-of-fit tests. *Biometrics*, pages 440–468, 1959.

Jeffrey M Wooldridge. A unified approach to robust, regression-based specification tests. *Econometric Theory*, 6(1):17–43, 1990.

A Appendix A

A.1 Lemmas

We first state auxiliary lemmas for the propositions and theorems in the main text. We let $\rightsquigarrow$ denote weak convergence on $l^{\infty}(\mathbb{D})$ (see definition 13.3 in Van Der Vaart [1996], hereafter VW), where $l^{\infty}(\mathbb{D})$ is the space of all real-valued functions that are uniformly bounded on $\mathbb{D}$, and $\xrightarrow{d}$ denote convergence of real-valued random variables. Troughout, to highlight the dependency on the partition, we denote as $\hat{\Phi}_{\theta}(\boldsymbol{\gamma}) := \hat{\Phi}_{\boldsymbol{\gamma}}(\theta)$ and $\hat{\Phi}_{0}(\boldsymbol{\gamma}) := \hat{\Phi}_{\boldsymbol{\gamma}}(\theta_0)$.

Lemma 1 *Under the null H_0,*

(a) if Assumption 1, 2, and 2' hold, then $\Sigma_{\boldsymbol{\gamma}}(\tilde{\theta}) = \Sigma_{\boldsymbol{\gamma},0} + o_p(1)$.

(b) if Assumption 1, 2, and 3 hold, then $\hat{W}_{\boldsymbol{\gamma}}(\tilde{\theta}) = \text{Avar}\left(\hat{\Phi}_{\boldsymbol{\gamma}}(\tilde{\theta})\right) + o_p(1)$

Lemma 2 *Under the null hypothesis* H_0*, Assumptions 1, and 5,*

$$\hat{\Phi}_0(\cdot) \rightsquigarrow \Phi_0(\cdot) \ \ \textit{as a process on } l^{\infty}(\mathbb{D}),$$

where $\Phi_0(\cdot)$ *is an* $\mathbb{R}^L$*-valued Gaussian process with zero mean vector and covariance structure given by,*

$$\mathbb{E}\left[\Phi_0(\boldsymbol{\gamma})\Phi_0(\tilde{\boldsymbol{\gamma}})\right] = \mathbb{E}\left[\varepsilon_{\theta_0}(Z)^2 \mathbf{I}_{\boldsymbol{\gamma}}(X)\mathbf{I}_{\tilde{\boldsymbol{\gamma}}}(X)'\right] \ \ \forall \boldsymbol{\gamma}, \tilde{\boldsymbol{\gamma}} \in \mathbb{D}.$$

Lemma 3 *Under the null hypothesis* H_0*, and Assumptions 1-5, it holds that:*

(a) $\sup_{\boldsymbol{\gamma}\in\mathbb{D}} \left|\hat{\Phi}_{\tilde{\theta}}(\boldsymbol{\gamma}) - (\hat{\Phi}_0(\boldsymbol{\gamma}) - \hat{\boldsymbol{\mu}}^{\circ\prime}_{\boldsymbol{\gamma}}(\theta_0)\sqrt{n}(\hat{\theta} - \theta_0))\right| = o_p(1).$

(b) $\hat{\boldsymbol{\mu}}^{\circ}_{\tilde{\boldsymbol{\gamma}}}(\theta_0) = \boldsymbol{\mu}^{\circ}_{\boldsymbol{\gamma},0} + o_p(1).$

Lemma 4 *Under the null* H_0*, and Assumptions 4,5,*

(a) if Assumption 1, 2, and 2' hold, then $\Sigma_{\tilde{\boldsymbol{\gamma}}}(\tilde{\theta}) = \Sigma_{\boldsymbol{\gamma},0} + o_p(1).$

(b) if Assumption 1, 2, and 3 hold, then $\hat{W}_{\tilde{\boldsymbol{\gamma}}}(\tilde{\theta}) = \text{Avar}\left(\hat{\Phi}_{\boldsymbol{\gamma}}(\tilde{\theta})\right) + o_p(1)$

A.2 Proofs

For any class of functions $\mathcal{F}$, we denote as $\{P_n f : f \in \mathcal{F}\}$ the empirical measure indexed by $\mathcal{F}$, such that $P_n f = n^{-1}\sum f(Z_i)$; alike, we use Pf for the population measure, $Pf = \int f(Z)dP(Z)$. We say that a class of functions is: i) Glivenko-Cantelli for P (hereafter, P-GC) whenever $\sup_{f\in\mathcal{F}} |P_n - P|f = o_p(1)$; ii) P-Donsker if $\{\sqrt{n}(P_n - P)f : f \in \mathcal{F}\}$ converge in distribution to a tight random element in the space $l^{\infty}(\mathcal{F})$. Throughout, we refer to both classes of sets with finite VC dimension and classes of functions with finite VC subgraph dimension as VC classes. These classes, having uniformly bounded covering numbers (Theorem 2.6.7 in VW), are Glivenko-Cantelli and Donsker (see Theorem 2.4.3 and 2.5.2 in VW) for any probability measure on the sample space, provided that they have integrable and square-integrable envelope function, respectively.

Proof of Lemma 1. For the first part of the Lemma, by the weak law of large numbers (WLLN) and a mean value theorem argument (MVT), suffices to show that

$$\frac{1}{n}\sum_{i=1}^{n}\left(\varepsilon_{\tilde{\theta}}^{2}(Z_i)-\varepsilon_{\theta_0}^{2}(Z_i)\right)\mathbb{I}_{\gamma_l}(X_i)=I+II+III=o_p(1)$$

for each $l\in 1,2,...,L$, where,

$$\begin{aligned}I&=(\tilde{\theta}-\theta_0)'\frac{1}{n}\sum_{i=1}^{n}\nabla m_{\bar{\theta}}(X_i)\nabla m_{\bar{\theta}}(X_i)'\mathbb{I}_{\gamma_l}(X_i)(\hat{\theta}-\theta_0),\\ II&=\frac{2}{n}\sum_{i=1}^{n}m_{\theta_0}(X_i)\nabla m_{\bar{\theta}}(X_i)'\mathbb{I}_{\gamma_l}(X_i)(\tilde{\theta}-\theta_0),\\ III&=\frac{2}{n}\sum_{i=1}^{n}Y_i(m_{\tilde{\theta}}(X_i)-m_{\theta_0}(X_i))\mathbb{I}_{\gamma_l}(X_i),\end{aligned}$$

and $|\bar{\theta}-\theta_0|\leq|\tilde{\theta}-\theta_0|$. The triangle inequality, Assumption 2, and the consistency of $\tilde{\theta}$ show that, $|I|\leq d_{\theta}^{2}\left\|\tilde{\theta}-\theta_0\right\|^{2}n^{-1}\sum_{i=1}^{n}R(X_i)^2=o_p(1)$, where $\|\cdot\|$ denotes the euclidean norm. By a similar reasoning,

$$\begin{aligned}|II|&\leq d_{\theta}\left\|\tilde{\theta}-\theta_0\right\|\frac{2}{n}\sum_{i=1}^{n}m_{\theta_0}(X_i)R(X_i)\\ &\leq d_{\theta}\left\|\tilde{\theta}-\theta_0\right\|\left(\mathbb{E}\left[Y^2\right]\right)^{1/2}\left(\mathbb{E}\left[R(X)^2\right]\right)^{1/2}+o_p(1)=o_p(1)\end{aligned}$$

where the last inequality follows from the WLLN, the law of iterated expectation, and Cauchy-Schwarz inequality. Finally, after expanding again around θ_0 it is easy to see that $|III|\leq d_{\theta}\left\|\tilde{\theta}-\theta_0\right\|2n^{-1}\sum_{i=1}^{n}Y_iR(X_i)\mathbb{I}_{\gamma_l}(X_i)=o_p(1)$.

For the second part of the lemma, we need to show that $\hat{C}_{\gamma}(\tilde{\theta})=C_{\gamma,0}+o_p(1)$, $\hat{L}(\tilde{\theta})=L_0+o_p(1)$, and $\hat{\boldsymbol{\mu}}_{\gamma}^{\circ}(\tilde{\theta})=\boldsymbol{\mu}_{\gamma,0}^{\circ}+o_p(1)$. By the usual MVT argument and the law of large numbers,

$$L_n=L_0+I+II+II'+o_p(1)$$

with,

$$\|I\| = \left\|\frac{1}{n}\sum_{i=1}^{n}\nabla l_{\bar{\theta}}(Z_i)\left(\hat{\theta}-\theta_0\right)\left(\hat{\theta}-\theta_0\right)'\nabla l_{\bar{\theta}}(Z_i)'\right\| \leq d_\theta^4\left\|\hat{\theta}-\theta_0\right\|^2\frac{1}{n}\sum_{i=1}^{n}R_2^2(Z_i) = o_p(1),$$

$$\begin{aligned}\|II\| &= \left\|\frac{1}{n}\sum_{i=1}^{n}\nabla l_{\bar{\theta}}(Z_i)\left(\hat{\theta}-\theta_0\right)l'_{\theta_0}(Z_i)\right\| \leq d_\theta^2\left\|\hat{\theta}-\theta_0\right\|\frac{1}{n}\sum_{i=1}^{n}\|l_{\theta_0}(Z_i)\|\,R_2(Z_i)\\ &\leq d_\theta^2\left\|\hat{\theta}-\theta_0\right\|\left(\frac{1}{n}\sum_{i=1}^{n}\|l_{\theta_0}(Z_i)\|^2\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n}R_2^2(Z_i)\right)^{1/2} = o_p(1).\end{aligned}$$

Alike, we write $\hat{C}_{\boldsymbol{\gamma}}(\tilde{\theta})$ as,

$$\hat{C}_{\boldsymbol{\gamma}}(\tilde{\theta}) = I - II - III + C_{\gamma,0} + o_p(1),$$

where,

$$\begin{aligned}I &= \frac{1}{n}\sum_{i=1}^{n}l_{\theta_0}(Z_i)\nabla m_{\bar{\theta}}(Z_i)'(\hat{\theta}-\theta_0)\mathbf{I}_{\boldsymbol{\gamma}}(X_i)'\\ II &= \frac{1}{n}\sum_{i=1}^{n}\nabla l_{\bar{\theta}}(Z_i)'(\hat{\theta}-\theta_0)\varepsilon_{\theta_0}(Z_i)\mathbf{I}_{\boldsymbol{\gamma}}(X_i)'\\ III &= \frac{1}{n}\sum_{i=1}^{n}\nabla l_{\bar{\theta}}(Z_i)(\hat{\theta}-\theta_0)(\hat{\theta}-\theta_0)'\nabla m_{\bar{\theta}}(Z_i)\mathbf{I}_{\boldsymbol{\gamma}}(X_i)'\end{aligned}$$

By Assumptions 2, and 3,

$$\begin{aligned}\|I\| &\leq \left\|\hat{\theta}-\theta_0\right\|\frac{1}{n}\sum_{i=1}^{n}\|l_{\theta_0}(Z_i)\|\,\|\nabla m_{\bar{\theta}}(Z_i)\|\,\|\mathbf{I}_{\boldsymbol{\gamma}}(X_i)\|\\ &\leq \sqrt{L}\left\|\hat{\theta}-\theta_0\right\|\left(\frac{1}{n}\sum_{i=1}^{n}\|l_{\theta_0}(Z_i)\|^2\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n}R^2(Z_i)\right)^{1/2} = o_p(1),\end{aligned}$$

An analogous reasoning shows that $\|II\| = o_p(1)$, and,

$$\begin{aligned}\|III\| &\leq \left\|\hat{\theta} - \theta_0\right\|^2 \frac{1}{n}\sum_{i=1}^{n} \|\nabla l_{\bar{\theta}}(Z_i)\| \, \|\nabla m_{\bar{\theta}}(Z_i)\| \, \|\mathbf{I}_{\boldsymbol{\gamma}}(X_i)\| \\ &\leq \sqrt{L} d_\theta^3 \left\|\hat{\theta} - \theta_0\right\|^2 \left(\frac{1}{n}\sum_{i=1}^{n} R^2(Z_i)\right)^{1/2} \left(\frac{1}{n}\sum_{i=1}^{n} R_2^2(Z_i)\right)^{1/2} = o_p(1).\end{aligned}$$

Finally, $\hat{\boldsymbol{\mu}}^\circ_{\boldsymbol{\gamma}}(\tilde{\theta}) = \boldsymbol{\mu}^\circ_{\boldsymbol{\gamma},0} + o_p(1)$ follows from the proof of Lemma 3 below. ■

Proof of Lemma 2. By Lemma 2.6.17 in VW and Assumption 5, both $\mathbb{D}$ and $\{\mathbf{I}_{\boldsymbol{\gamma}}(X) : \gamma \in \mathbb{D}\}$ are VC classes. Therefore, $\mathcal{F} = \{\varepsilon_{\theta_0}(z)\mathbf{I}_{\boldsymbol{\gamma}}(X) : \gamma \in \mathbb{D}\}$ is a VC class (Lemma 2.6.18 in VW), with square integrable envelope function $F = |\varepsilon_{\theta_0}|$, and, hence, is P-Donsker. The convergence of the finite-dimensional distributions (fidis) of $\hat{\Phi}_0(\cdot)$ to those of $\Phi_0(\cdot)$, by the multivariate central limit theorem, characterize the limit process. ■

Proof of Lemma 3. By an MVT argument,

$$\hat{\Phi}_{\hat{\theta}}(\boldsymbol{\gamma}) = \hat{\Phi}_0(\boldsymbol{\gamma}) - I'\sqrt{n}(\hat{\theta} - \theta_0) - \hat{\boldsymbol{\mu}}^{\circ\prime}_{\boldsymbol{\gamma}}(\theta_0)\sqrt{n}(\hat{\theta} - \theta_0)$$

where,

$$I = \hat{\boldsymbol{\mu}}^\circ_{\hat{\boldsymbol{\gamma}}}(\bar{\theta}) - \hat{\boldsymbol{\mu}}^\circ_{\hat{\boldsymbol{\gamma}}}(\theta_0) = \frac{1}{n}\sum_{i=1}^{n}\big(\nabla m_{\bar{\theta}}(X_i) - \nabla m_{\theta_0}(X_i)\big)\mathbf{I}_{\boldsymbol{\gamma}}(X_i)'$$

and $|\bar{\theta} - \theta_0| \leq |\tilde{\theta} - \theta_0|$. The class $\{\nabla m_\theta(x) : \theta \in \Theta\}$ is a collection of continuous mapping, $\theta \to \nabla m_\theta$, over the compact metric space Θ with integrable envelope function $R(\cdot)$ and, therefore, is P-GC (e.g., Example 19.8 in Van der Vaart [2000]). Thus,

$$\begin{aligned}\sup_{\gamma \in \mathbb{D}} \|I\| &\leq \sup_{\gamma \in \mathbb{D}} \frac{1}{n}\sum_{i=1}^{n} \|\nabla m_{\bar{\theta}}(X_i) - \nabla m_{\theta_0}(X_i)\| \, \|\mathbf{I}_{\boldsymbol{\gamma}}(X_i)\| \\ &\leq \sqrt{L}\frac{1}{n}\sum_{i=1}^{n} \|\nabla m_{\bar{\theta}}(X_i) - \nabla m_{\theta_0}(X_i)\| = o_p(1).\end{aligned}$$

where the last equality follows from an application of the uniform law of large numbers (e.g., Davidson [1994], Theorem 21.6). For the second part of the Lemma is sufficient to prove that,

$$|II| = \left| \frac{1}{n} \sum_{i=1}^{n} \nabla^{(j)} m_{\theta_0}(X_i) \big(\mathbf{I}_{\hat{\boldsymbol{\gamma}}}(X_i) - \mathbf{I}_{\boldsymbol{\gamma}}(X_i) \big) \right| = o_p(1),$$

for each $j \in \{1, .., d_\theta\}$. To see this is true, notice that by Assumption 5, $\mathbb{D}\tilde{\Delta}\mathbb{D} = \{\boldsymbol{\gamma}_1 \tilde{\Delta} \boldsymbol{\gamma} : \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2 \in \mathbb{D}\}$ is a class of subsets of unions of VC classes, and hence is VC. Therefore, $\{|\nabla^{(j)} m_{\theta_0}(x)| \mathbf{I}_{\tilde{\gamma}}(x) : \tilde{\gamma} \in \mathbb{D}\tilde{\Delta}\mathbb{D}\}$ is also VC with integrable envelope $R(\cdot)$ and, hence, P-GC. Thus, for each $j \in \{1, .., d_\theta\}$,

$$\begin{aligned} |II| &\leq \frac{1}{n} \sum_{i=1}^{n} |\nabla^{(j)} m_{\theta_0}(X_i)| \mathbf{I}_{\hat{\boldsymbol{\gamma}}\tilde{\Delta}\boldsymbol{\gamma}}(X_i) \\ &\leq \sup_{\tilde{\gamma} \in \mathbb{D}\tilde{\Delta}\mathbb{D}} (P_n - P) |\nabla^{(j)} m_{\theta_0}| \mathbf{I}_{\hat{\boldsymbol{\gamma}}\tilde{\Delta}\boldsymbol{\gamma}} + \mathbb{E}\left[|\nabla^{(j)} m_{\theta_0}(X)| \mathbf{I}_{\hat{\boldsymbol{\gamma}}\tilde{\Delta}\boldsymbol{\gamma}}(X) \right] \\ &= o_p(1) + \mu_R(\hat{\boldsymbol{\gamma}}\tilde{\Delta}\boldsymbol{\gamma}) = o_p(1) \end{aligned}$$

where $\mu_R(\hat{\boldsymbol{\gamma}}\tilde{\Delta}\boldsymbol{\gamma}) = \big(\mu_R(\hat{\gamma}_1 \tilde{\Delta} \gamma_1), ..., \mu_R(\hat{\gamma}_L \tilde{\Delta} \gamma_L) \big)'$, and $\mu_R(A) = \int_A \mathbb{E}\left[R(Z) | X \right] dP(X)$ is a (signed) measure absolutely continuous with respect to P. The last equality follows from Assumption 4. ■

Proof of Lemma 4. For each element on the main diagonal of $\hat{\Sigma}_{\hat{\boldsymbol{\gamma}}}(\tilde{\theta}) - \hat{\Sigma}_{\boldsymbol{\gamma}}(\tilde{\theta})$ write,

$$\frac{1}{n} \sum_{i=1}^{n} \varepsilon_{\hat{\theta}}(Z_i)^2 (\mathbb{I}_{\hat{\gamma}_l}(X_i) - \mathbb{I}_{\gamma_l}(X_i)) = I + II + III$$

where

$$\begin{aligned} I &= \frac{1}{n} \sum_{i=1}^{n} \varepsilon_{\theta_0}(Z_i)^2 (\mathbb{I}_{\hat{\gamma}_l}(X_i) - \mathbb{I}_{\gamma_l}(X_i)) \\ II &= (\hat{\theta} - \theta_0)' \frac{1}{n} \sum_{i=1}^{n} \nabla m_{\bar{\theta}}(X_i) \nabla m_{\bar{\theta}}(X_i)' (\mathbb{I}_{\hat{\gamma}_l}(X_i) - \mathbb{I}_{\gamma_l}(X_i)) (\hat{\theta} - \theta_0) \\ III &= -(\hat{\theta} - \theta_0)' \frac{2}{n} \sum_{i=1}^{n} \varepsilon_{\theta_0}(Z_i) \nabla m_{\bar{\theta}}(X_i) (\mathbb{I}_{\hat{\gamma}_l}(X_i) - \mathbb{I}_{\gamma_l}(X_i)) \end{aligned}$$

The class $\{\varepsilon_{\theta_0}(z)^2 \mathbb{I}_{\tilde{\gamma}_l} : \tilde{\gamma} \in \mathbb{C}\tilde{\Delta}\mathbb{C}\}$ is VC with integrable envelope function $\varepsilon^2_{\theta_0}$ and, hence, is P-GC. Therefore, $|I| \leq \mu_\sigma(\hat{\gamma}_l \tilde{\Delta} \gamma_l) + o_p(1) = o_p(1)$, by Assumption 4. Also, $|II| \leq$

$\sqrt{L}d_\theta^2 \left\|\hat{\theta}-\theta_0\right\|^2 n^{-1}\sum_{i=1}^n R(X_i)^2 = o_p(1)$, by Assumptions 1-2 and the consistency of $\hat{\theta}$, and $|III| \leq \sqrt{L}d_\theta \left\|\hat{\theta}-\theta_0\right\| n^{-1}\sum_{i=1}^n \varepsilon_{\theta_0} R(X_i) = o_p(1)$ by Cauchy-Schwarz inequality. Thus, $\hat{\Sigma}_{\hat{\gamma}}(\tilde{\theta}) = \hat{\Sigma}_{\gamma}(\tilde{\theta}) + o_p(1)$, and the first part of the lemma follows from Lemma 1(a).

For the second part of the lemma, notice that by Lemma 3(b) (and the proof of the first part of Lemma 3), $\hat{\boldsymbol{\mu}}^\circ_{\hat{\gamma}}(\hat{\theta}) = \boldsymbol{\mu}^\circ_{\gamma,0} + o_p(1)$, and for each element of $C_n(\hat{\gamma}) - C_n(\gamma_0)$ it holds that,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \varepsilon_{\hat{\theta}}(Z_i) l_{\hat{\theta},j}(\mathbb{I}_{\hat{\gamma}_l}(X_i) - \mathbb{I}_{\gamma_l}(X_i)) &\leq \left(\frac{1}{n}\sum_{i=1}^n \varepsilon_{\hat{\theta}}(Z_i)^2 \mathbb{I}_{\hat{\gamma}_l \tilde{\Delta}\gamma_l}(X_i)\right)^{1/2} \left(\frac{1}{n}\sum_{i=1}^n l_{\hat{\theta},j}\right)^{1/2} \\ &= o_p(1)O_p(1),\end{aligned}$$

where $l_{\theta,j}$ denotes the j-th component of l_θ and the last equality follows from the first part of this proof and Lemma 1. ■

Proof of Theorem 1. The estimator consistency follows from $\sup_{\theta\in\Theta} |Q_n(\theta) - Q_0(\theta)| = o_p(1)$, where $Q_0(\theta) = \mathbb{E}\left[\varepsilon_\theta(Z)\mathbf{I}_\gamma(X)\right]' (\Sigma_{\gamma,0})^{-1}\mathbb{E}\left[\varepsilon_\theta(Z)\mathbf{I}_\gamma(X)\right]$ and $Q_n(\theta) = n^{-1}\hat{\chi}^2_{\hat{\gamma},\tilde{\theta}}(\theta)$ (see Theorem 2.1 in Newey and McFadden [1994], for instance). Notice that,

$$\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_\theta(Z_i)\mathbf{I}_{\hat{\gamma}}(X_i) - \mathbb{E}\left[\varepsilon_\theta(Z)\mathbf{I}_\gamma(X)\right]\right| \leq \sup_{\theta\in\Theta}|I| + \sup_{\theta\in\Theta}|II| + \sup_{\theta\in\Theta}|III|$$

where

$$\begin{aligned} I &= (P_n - P)\varepsilon_\theta \mathbf{I}_\gamma, \\ II &= \frac{1}{n}\sum_{i=1}^n \varepsilon_{\theta_0}(Z_i)(\mathbf{I}_{\hat{\gamma}}(X_i) - \mathbf{I}_\gamma(X_i)), \\ III &= \frac{1}{n}\sum_{i=1}^n \left[\nabla m_{\bar{\theta}}(Z_i)(\mathbf{I}_{\hat{\gamma}}(X_i) - \mathbf{I}_\gamma(X_i))'\right]'(\theta-\theta_0). \end{aligned}$$

The mapping $\theta \to \varepsilon_\theta$ is continuous over the compact Θ, with

$$\mathbb{E}\left[\sup_{\theta\in\Theta}\varepsilon_\theta(Z)\right] \leq \mathbb{E}\left[\sup_{\bar{\theta},\theta\in\Theta}\varepsilon_{\theta_0}(Z) - \nabla m_{\bar{\theta}}(X)'(\theta-\theta_0)\right] \leq d_\theta \mathbb{E}\left[R(Z)\right] D < \infty,$$

by Assumption 2 and the compactness of Θ, where D denotes the diameter of Θ. Therefore, both $\{\varepsilon_\theta(z) : \theta \in \Theta\}$ and $\{\varepsilon_\theta(z)\mathbf{I}_{\boldsymbol{\gamma}}(X) : \theta \in \Theta\}$ are P-GC classes (see Corollary 8.6 in Giné and Zinn [1984], for instance) and, hence, $\sup_{\theta\in\Theta} |I| = o_p(1)$. Then, by Cauchy-Schwarz inequality and Assumptions 1 and 4,

$$\sup_{\theta\in\Theta} |II| \leq \left(\frac{1}{n}\sum_{i=1}^{n} \varepsilon^2_{\theta_0}(Z_i)\right)^{1/2} \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{I}_{\hat{\boldsymbol{\gamma}}\tilde{\Delta}\boldsymbol{\gamma}}(X_i)\right)^{1/2} = o_p(1).$$

Finally,

$$\begin{aligned}\sup_{\theta\in\Theta} \|III\| &\leq \sup_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \left\|\mathbf{I}_{\hat{\boldsymbol{\gamma}}\tilde{\Delta}\boldsymbol{\gamma}}(X_i)\right\| \|\nabla m_{\bar{\theta}}(X_i)\| \|\theta-\theta_0\| \\ &\leq d_\theta D \frac{1}{n}\sum_{i=1}^{n} \left\|\mathbf{I}_{\hat{\boldsymbol{\gamma}}\tilde{\Delta}\boldsymbol{\gamma}}(X_i)\right\| R(X_i) = o_p(1),\end{aligned}$$

where the second inequality follows from Assumption 2 and the compactness of Θ. This result, together with the consistency of $\hat{\Sigma}_{\hat{\boldsymbol{\gamma}}}(\tilde{\theta})$ implies that $\sup_{\theta\in\Theta} |Q_n(\theta) - Q_0(\theta)| = o_p(1)$ and, therefore, $\hat{\theta}_{\hat{\boldsymbol{\gamma}}} = \theta_0 + o_p(1)$. For the asymptotic normality: by Assumptions 2 and 1(c), the first-order conditions of the minimization problem are satisfied with probability approaching one, $\hat{\boldsymbol{\mu}}^{\circ}_{\hat{\boldsymbol{\gamma}}}(\hat{\theta}_{\boldsymbol{\gamma}})\hat{\Sigma}_{\hat{\boldsymbol{\gamma}}}(\tilde{\theta})^{-1}n^{-1/2}\hat{\Phi}_{\hat{\boldsymbol{\gamma}}}(\hat{\theta}_{\boldsymbol{\gamma}}) = 0$. Expanding $\hat{\Phi}_{\hat{\boldsymbol{\gamma}}}(\hat{\theta}_{\boldsymbol{\gamma}})$ around θ_0 and solving gives the Bahadur representation,

$$\sqrt{n}(\hat{\theta}^{G} - \theta_0) = -\left[\hat{\boldsymbol{\mu}}^{\circ}_{\hat{\boldsymbol{\gamma}}}(\hat{\theta}_{\boldsymbol{\gamma}})\hat{\Sigma}_{\hat{\boldsymbol{\gamma}}}(\tilde{\theta})^{-1}\hat{\boldsymbol{\mu}}^{\circ\prime}_{\hat{\boldsymbol{\gamma}}}(\bar{\theta})\right]^{-1} \hat{\boldsymbol{\mu}}^{\circ}_{\hat{\boldsymbol{\gamma}}}(\hat{\theta}_{\boldsymbol{\gamma}})\hat{\Sigma}_{\hat{\boldsymbol{\gamma}}}(\tilde{\theta})^{-1}\hat{\Phi}_{\hat{\boldsymbol{\gamma}}}(\theta_0) + o_p(1)$$

where $|\bar{\theta} - \theta_0| \leq |\hat{\theta}_{\hat{\boldsymbol{\gamma}}} - \theta_0|$.

From the proof of Lemma 3, it follows that $\hat{\boldsymbol{\mu}}^{\circ}_{\hat{\boldsymbol{\gamma}}}(\theta) - \hat{\boldsymbol{\mu}}^{\circ}_{\hat{\boldsymbol{\gamma}}}(\theta_0) = o_p(1)$ for any $\theta \underset{p}{\rightarrow} \theta_0$. Thus by consistency of $\hat{\theta}_{\hat{\boldsymbol{\gamma}}}$ and Lemma 1(b), it follows that both $\hat{\boldsymbol{\mu}}^{\circ}_{\hat{\boldsymbol{\gamma}}}(\hat{\theta}_{\boldsymbol{\gamma}})$ and $\hat{\boldsymbol{\mu}}^{\circ}_{\hat{\boldsymbol{\gamma}}}(\bar{\theta})$ converge in probability to $\boldsymbol{\mu}^{\circ}_{\boldsymbol{\gamma},0}$. Finally, by Assumptions 4-5, and the uniform continuity of the sample paths of $\Phi_0(\cdot)$,

$$\Phi_0(\hat{\boldsymbol{\gamma}}) \xrightarrow{d} N\left(0, \Sigma_{\boldsymbol{\gamma},0}\right).$$

■

Proof of Theorem 2. By Lemma 3, 4, and Assumption 3 (or Theorem 1 and Assumption 2' for the $\hat{\chi}^2$) both test statistics are asymptotically equivalent to the following quadratic form,

$$q(\hat{\theta}, W, \hat{\gamma}) = \left(\Phi_0(\hat{\boldsymbol{\gamma}}) - \sqrt{n}\boldsymbol{\mu}^{\circ\prime}_{\boldsymbol{\gamma},0}\bar{l}_n\right)' W^{-1}\left(\Phi_0(\hat{\boldsymbol{\gamma}}) - \sqrt{n}\boldsymbol{\mu}^{\circ}_{\boldsymbol{\gamma},0}\bar{l}_n\right), \qquad \text{(A1)}$$

where $\bar{l}_n = \mathbb{E}_n\left[l_{\theta_0}(Z_i)\right]$, W^{-1} is the probability limit of W_n^{-1}, and the couple $(W_n^{-1}, \hat{\theta})$ is equal to $(\hat{\Sigma}_{\hat{\boldsymbol{\gamma}}}(\tilde{\theta}), \hat{\theta}_{\hat{\boldsymbol{\gamma}}})$ in the $\hat{\chi}^2$ test and $\left(\widehat{\text{Avar}^-}\left(\hat{\Phi}_{\boldsymbol{\gamma}}(\tilde{\theta})\right), \tilde{\theta}\right)$ in the Wald test. The functional

$$\phi(z, w, \gamma) = \left(z(\gamma) - \boldsymbol{\mu}^{\circ\prime}_{\boldsymbol{\gamma},0}w\right)' W^{-1}\left(z(\gamma) - \boldsymbol{\mu}^{\circ\prime}_{\boldsymbol{\gamma},0}w\right),$$

mapping $(\hat{\Phi}_0(\cdot), \sqrt{n}\bar{l}, \hat{\boldsymbol{\gamma}})$ into $q(\hat{\theta}, W, \hat{\boldsymbol{\gamma}})$ is continuous with respect to the product topology on $l^\infty(\mathbb{D}) \times \mathbb{R}^L \times \mathbb{D}$ (see Lemma 4 in Andrews [1988]). Thus, Theorem 2 follows by establishing the limit null distribution of $(\Phi_0(\hat{\boldsymbol{\gamma}}) - \boldsymbol{\mu}^{\circ\prime}_{\boldsymbol{\gamma},0}\sqrt{n}\bar{l})$ and an application of the continuous mapping theorem (e.g., Theorem 1.3.6 in VW). Lemma 2, Assumptions 1, 4, 5, and the central limit theorem imply that $(\Phi_0(\cdot), \sqrt{n}\bar{l}, \hat{\gamma})$ is a uniformly tight process on $\mathbb{D}$ with fidis converging weakly to those of $(\Phi_0(\cdot), l_0, \gamma_0)$, where $l_0 \overset{d}{=} N(0, L_0)$ and $\mathbb{E}\left[l_0\phi_0(\boldsymbol{\gamma})\right] = \mathbb{E}\left[l_{\theta_0}(Z)\varepsilon_{\theta_0}(Z)\mathbf{I}_{\boldsymbol{\gamma}}(X)'\right]$. Thus,

$$(\hat{\Phi}_0(\cdot), \sqrt{n}\bar{l}, \hat{\boldsymbol{\gamma}}) \rightsquigarrow (\Phi_0(\cdot), l_0, \gamma_0) \;\text{ on } l^\infty(\mathbb{D}),$$

and by the continuous mapping theorem,

$$q(\hat{\theta}, W, \hat{\gamma}) \xrightarrow{d} Y'W^{-1}Y,$$

where $Y \overset{d}{=} N(0, \Sigma_Y)$ and $\Sigma_Y = \text{Avar}\left(\hat{\Phi}_{\boldsymbol{\gamma}}(\hat{\theta})\right)$. In the Wald test, where W^{-1} is a generalized inverse of Σ_Y, $Y'W^{-1}Y \overset{d}{=} \chi^2_{rank\left(\text{Avar}\left(\hat{\Phi}_{\boldsymbol{\gamma}}(\tilde{\theta})\right)\right)}$ by Theorem 7.3(i) in Rao and Mitra [1972]. The limit null distribution of the $\hat{\chi}^2$ test follows from the fact that $\Sigma_Y = \text{Avar}\left(\hat{\Phi}_{\boldsymbol{\gamma}}(\hat{\theta}_{\hat{\boldsymbol{\gamma}}})\right) = \Sigma_{\boldsymbol{\gamma},0} - \boldsymbol{\mu}^{\circ\prime}_{\boldsymbol{\gamma}}(\boldsymbol{\mu}^{\circ}_{\boldsymbol{\gamma}}\Sigma^{-1}_{\boldsymbol{\gamma},0}\boldsymbol{\mu}^{\circ\prime}_{\boldsymbol{\gamma}})\boldsymbol{\mu}^{\circ}_{\boldsymbol{\gamma}}$ and, thus, $\Sigma^{-1/2}_{\boldsymbol{\gamma},0}(\Sigma_Y)\Sigma^{-1/2}_{\boldsymbol{\gamma},0}$ is idempotent with rank equal to $L - d_\theta$. ■

Proof of Theorem 3. Is sufficient to show that,

$$\hat{\Phi}_0(\hat{\boldsymbol{\gamma}}) - \hat{\Phi}_0(\boldsymbol{\gamma}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_{\theta_0}(Z_i)(\mathbf{I}_{\hat{\gamma}}(X_i) - \mathbf{I}_{\gamma}(X_i)) = o_p(1).$$

To see this, notice that $f_{\hat{\gamma}} = \varepsilon_{\theta_0}\mathbb{I}_{\hat{\gamma}}$ is a random element taking values in the P-Donsker class $\mathcal{F} = \{\varepsilon_{\theta_0}(z)\mathbf{I}_{\gamma}(x) : \gamma \in \mathbb{D}\}$ (see Lemma 2) such that,

$$\int \left(\varepsilon_{\theta_0}(\mathbf{I}_{\hat{\gamma}}(X) - \mathbf{I}_{\gamma}(X))\right)^2 dP(Z) = \mu_\sigma(\hat{\boldsymbol{\gamma}}\tilde{\Delta}\boldsymbol{\gamma}) = o_p(1),$$

where $\mu_\sigma(\hat{\boldsymbol{\gamma}}\tilde{\Delta}\boldsymbol{\gamma})$ is a vector defined analogously to $\mu_R(\hat{\boldsymbol{\gamma}}\tilde{\Delta}\boldsymbol{\gamma})$ and $\mu_\sigma(A) = \int_A \mathbb{E}\left[\varepsilon_{\theta_0}(Z)^2|X\right] dP(X)$ is an absolutely continuous measure with respect to P. Thus, by Lemma 19.24 in Van der Vaart [2000], $\sqrt{n}(P_n - P)(f_{\hat{\gamma}} - f_{\gamma}) = o_p(1)$, and

$$\begin{aligned}\hat{\Phi}_0(\hat{\boldsymbol{\gamma}}) - \hat{\Phi}_0(\boldsymbol{\gamma}) &= \sqrt{n}(P_n - P)(f_{\hat{\gamma}} - f_{\gamma}) + \sqrt{n}P(f_{\hat{\gamma}} - f_{\gamma}) \\ &= \sqrt{n}\mathbb{E}\left[\varepsilon_{\theta_0}(Z)(\mathbf{I}_{\hat{\gamma}}(X) - \mathbf{I}_{\gamma}(X)\right] + o_p(1) = o_p(1),\end{aligned}$$

by assumption (12). ■

Proof of Proposition 1. Let $Q = \text{diag}(q_1, ..., q_L)$, with $q_l > 0$ for all l, and consider

$$D\left(\{a_l\}_{l=0}^{L}, L\right) = \sum_{l=1}^{L} \frac{\left(\int\limits_{a_{l-1}}^{a_l} h(t)f_x(t)dt\right)^2}{q_l} = \sum_{l=1}^{L} \frac{\mathbb{E}\left[h(x)\mathbb{I}\{X \in \gamma_l\}\right]^2}{q_l},$$

where $\gamma_l = \{x \in \mathcal{X} : a_{l-1} \leq x \leq a_l\}$ and $\{a_l\}_{l=0}^{K}$, where $a_l \in \mathcal{X}$ and $a_1 \leq a_2 \leq \cdots \leq a_K$ splits the support of $\mathcal{X}$ into L cells $C_l = \{x \in \mathcal{X} : a_{l-1} \leq x \leq a_l\}$. Fix the partition dimension $L = K$, where K is the number of points in $\mathcal{X}$ where $h(x) = 0$, and let determine the optimal sequence of splitting points $\{a_l^*\}_{l=0}^{K}$. By definition, the first and last element of the sequence are the extrema of the support, $a_0^* = \inf\{x \in \mathcal{X}\}$, $a_K^* = \sup\{x \in \mathcal{X}\}$. The other optimal elements solve,

$$\{a_l^*\}_{l=1}^{K-1} = \operatorname*{arg\,max}_{\{a_l\}_{l=1}^{K-1}\in\mathcal{X}^{K-2}} \frac{1}{2}\sum_{l=1}^{K-1} \frac{\left(\int\limits_{a_{l-1}}^{a_l} h(t)f_x(t)dt\right)^2}{q_l}.$$

The gradient of the objective function, ∇, is a $(K-2)$-dimensional vector with typical elements,

$$[\nabla]_l = \Big(\mathbb{E}\left[h(X)\gamma_{l-1}(X)\right] - \mathbb{E}\left[h(X)\gamma_l\right](X)\Big)h(a_i)f_x(a_l)/q_l$$

with $l \in \{1,...,K-1\}$, and the Hessian,∇_2 say, is given by the $(K-2)\times(K-2)$ matrix with typical elements,

$$[\nabla_2]_{l,f} = \begin{cases} \frac{1}{q_l}\Big(2\left(h(a_l)f(a_l)\right)^2 + \left(\mathbb{E}\left[h(X)\gamma_{l-1}(X)\right] - \mathbb{E}\left[h(X)\gamma_l(X)\right]\right)\left(f(a_l)h(a_l)\right)'\Big) & \text{for } l=f \\ -f(a_f)h(a_f)f(a_l)h(a_l)/q_l & \text{for } |l-f|\leq 1 \\ 0 & \text{for } |l-f|>1 \end{cases}$$

where $(f(a_l)h(a_l))' = d/da_l\left(f(a_l)h(a_l)\right)$ denotes the derivative with respect to a_l, and $l,f\in\{1,...,K-1\}$. Consider the optimal sequence of points $\{a_l^*\}_{l=1}^{K-1}$ defined recursively as,

$$a_1^* : h(a_1^*) = 0, \quad a_l^* = S_l\tilde{a}_l + (1-S_l)a_{l-1}^*,$$

where $S_l = \mathbb{I}\{h(x)=0, x>a_{l-1}^*\}$, and $\tilde{a}_l$ is defined such that $h(\tilde{a}_l)=0$ and the derivative of h at $\tilde{a}_i$ has opposite sign of the derivative of h at a_{l-1}^*, $\operatorname{sgn}(h'(a_l)) = -\operatorname{sgn}(h'(a_{l-1}^*))$, meaning that $\{a_l^*\}$ is the sequence of consecutive points where the function assumes value zero. Then, it is easy to check that the quadratic form of the Hessian at $\{a_l^*\}_{l=1}^{K}$ is negative definite,

$$\begin{aligned} b'\nabla_2 b = &\sum_{l=2}^{K-1}\Big(\mathbb{E}\left[h(X)\gamma_{l-1}(X)\right] - \mathbb{E}\left[h(X)\gamma_l(X)\right]\Big)h'(a_l)f_x(a_l)(b_l^2/q_l) \\ &= \mathbb{E}\left[h(X)\gamma_1(X)\right]h'(a_2)f_x(a_2)(b_1^2/q_1) + \mathbb{E}\left[h(X)\gamma_2(X)\right]\left(h'(a_3)f_x(a_3)(b_2^2/q_2) - h'(a_2)f_x(a_2)(b_1^2/q_1)\right) \\ &+ \mathbb{E}\left[h(X)\gamma_{K-1}(X)\right]\left(h'(a_{K-1})f_x(a_{K-1})(b_{K-2}^2/q_{K-2}) - h'(a_{K-2})f_x(a_{K-2})(b_{K-3}^2/q_{K-3})\right) \\ &- \mathbb{E}\left[h(X)C_K(X)\right]h'(a_{K-1})f_x(a_{K-1})(b_{K-2}^2/q_{K-2}) \leq 0 \ \text{ for all } b\in\mathbb{R}^{K-1}, \end{aligned}$$

with strict inequality when $h'(x) \neq 0$ for some a_l^*. The last inequality follows from the fact that $h'(a_l*) \geq 0$ $(h'(a_l^*) \leq 0)$ implies $\mathbb{E}\left[h(X)\gamma_{l-1}\right] \leq 0$ $(\mathbb{E}\left[h(X)\gamma_{l-1}\right] \geq 0)$. Thus, $\{a_l^*\}_{l=0}^{K}$ is a global maximum for fixed L. Now, notice that, since $Q = I_M$,

$$(\delta_l + \delta_f)^2 \geq \delta_l^2 + \delta_f^2$$

for any pair δ_l, δ_f such that $\operatorname{sgn}(\delta_l) = \operatorname{sgn}(\delta_f)$; therefore, any finite split of $\mathcal{X}$ is dominated by a partition with two classes. A contradiction argument shows that the global maximum is attained by a two-cell partition merging the positive and negative cells characterized by $\{a_j^*\}_{j=0}^{K}$. ■