NEYMAN'S $C(\alpha)$ TEST FOR UNOBSERVED HETEROGENEITY

JIAYING GU
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

ABSTRACT. A unified framework is proposed for tests of unobserved heterogeneity in parametric statistic models based on Neyman's $C(\alpha)$ approach. Such tests are irregular in the sense that the first order derivative of the log likelihood with respect to the heterogeneity parameter is identically zero, and consequently the conventional Fisher information about the parameter is zero. Nevertheless, local asymptotic optimality of the $C(\alpha)$ tests can be established via LeCam's differentiability in quadratic mean and the limit experiment approach. This leads to local alternatives of order $n^{-1/4}$. The scalar case result is already familiar from existing literature and we extend it to the multi-dimensional case. The new framework reveals that certain regularity conditions commonly employed in earlier developments are unnecessary, i.e. the symmetry or third moment condition imposed on the heterogeneity distribution. Additionally, the limit experiment for the multi-dimensional case suggests modifications on existing tests for slope heterogeneity in cross sectional and panel data models that lead to power improvement. Since the $C(\alpha)$ framework is not restricted to the parametric model and the test statistics do not depend on the particular choice of the heterogeneity distribution, it is useful for a broad range of applications for testing parametric heterogeneity.

1. INTRODUCTION

Neyman's (1959) $C(\alpha)$ test can be viewed as a generalization of Rao's (1948) score test in the presence of nuisance parameters and thus provides a unified framework for parametric statistical inference. We will see that many of the existing tests for neglected parameter heterogeneity can also be formulated as $C(\alpha)$ tests and share common features. However, for these tests the usual score function is identically zero under the null hypothesis, and conventional Fisher information is thus zero. Fortunately, in these cases the second derivative of the log likelihood is non-degenerate and approximations based on it can be used to form a modified version of LeCam's differentiability in quadratic mean (DQM) condition. Local asymptotic normality (LAN) theory, then leads to local asymptotic optimality results for the $C(\alpha)$ test in such settings under local alternatives of order $n^{-1/4}$.

Date: October 4, 2014.

Department of Economics, University of Illinois at Urbana-Champaign, 214 David Kinley Hall, 1407 W. Gregory Dr., Urbana, Illinois 61801, MC-707, USA. Tel: +1-267-994-1519. Fax: +1-217-244-6571. Email Address: gu17@illinois.edu. I would like to thank Roger Koenker for his continued support and encouragement. I would also like to thank Andreas Hagemann, Marc Hallin, Keisuke Hirano, Stanislav Volgushev, two anonymous referees and the participants at the Midwest Econometrics Group meeting 2012 and the Boneyard Conference 2013 at the University of Illinois for valuable comments and useful discussion. I gratefully acknowledge financial support from NSF grant SES-11-53548 and the Paul Boltz summer Fellowship. All errors are my own.

We find that LeCam's limit experiment perspective is very useful in analyzing tests for neglected heterogeneity especially in the multi-dimensional setting. It allows us to first develop optimal test statistics for the Gaussian limit and then extend them to the corresponding asymptotic $C(\alpha)$ test. The one-sided nature of the limit experiment reveals that we require the mixture of χ^2 asymptotics which leads to power improvement compared to the conventional χ^2 type test. This finding is relevant to the Information Matrix test and some of the recent applications to slope heterogeneity testing in panel data models.

We focus initially on the case of a scalar heterogeneity parameter. Although some of the results are already familiar in the literature, the use of the LeCam framework is new and it leads to a set of less restrictive assumptions and sheds light on why reparameterization leads to unnecessary conditions employed in previous literature. Discussing the scalar case in the LeCam framework also facilitates the extension to multivariate settings which is described at the end of Section 2 and is the major contribution of the paper. In Section 3 we consider four different examples. In the first example, the $C(\alpha)$ tests for parameter heterogeneity in Poisson regression model under two slightly different alternative specifications lead to tests introduced in Lee (1986). The second example considers testing for slope heterogeneity in cross sectional linear regression models; the $C(\alpha)$ test in this setting shares many features of the Breusch and Pagan (1979) LM test, but the positivity constraints revealed via the limit experiment suggest a modification that leads to a power gain. We then illustrate an example using the $C(\alpha)$ test to jointly test for heterogenous location and scale parameters in Gaussian panel data models. Lastly, we compare the $C(\alpha)$ test for slope heterogeneity in panel data model to the test considered in Pesaran and Yamagata (2008). For a wide range of N and T, the $C(\alpha)$ test, since it pays explicit attention to the positivity constraints under the alternative, enjoys a power improvement.

The $C(\alpha)$ test for heterogeneity formulated in this paper is very similar to the setup used in some previous development. In a seminal paper, Chesher (1984) points out the score test for unobserved parametric heterogeneity is identical to White's (1982) Information Matrix (IM) test. Cox (1983) obtains similar results using a more general mixture model. These papers can be viewed as important extensions of a somewhat neglected example on testing for parameter heterogeneity in Poisson models in Neyman and Scott (1966). Moran (1973) investigates the asymptotic behavior of these score tests. However, as we will show in Section 4, the parameterization adopted in Moran (1973) and also in Chesher (1984) requires some unnecessary additional assumptions, i.e. the zero third moment or symmetry of the heterogeneity distribution, even though it delivers the same score function formed based on the second derivative as the $C(\alpha)$ test constructed here. The explanation is that the likelihood under their parameterization obtains the same expansion of the likelihood under the $C(\alpha)$ test parameterization only if symmetry holds. In addition, even though the score function is the same, the positivity constraints lead to a different decision rule with mixture of χ^2 asymptotics in contrast to the conventional χ^2 test for the IM test. Furthermore, there are situations where the $C(\alpha)$ test for unobserved heterogeneity is no longer identical to the IM test and we illustrate some conditions for equivalence to hold in Section 4. Lastly, a Monte Carlo simulation is carried out to evaluate the power performance for various examples.

2. The $C(\alpha)$ test for unobserved parameter heterogeneity

Neyman (1959) introduces the $C(\alpha)$ test with the consideration that hypotheses testing problems in applied research often involve several nuisance parameters. In these composite testing problems, most powerful tests do not exist, motivating search for an optimal test procedure that yields the highest power among the class of tests obtaining the same size. Neyman's locally asymptotically optimality result for the $C(\alpha)$ test employs regularity conditions inherited from the conditions used by Cramér (1946) for showing consistency of MLE and some further restrictions on the testing function to allow for replacing the unknown nuisance parameters by its $\sqrt{n}$-consistent estimators. It is the confluence of these Cramér conditions and the maintained significance level α that gives the name to the $C(\alpha)$ test.

2.1. $C(\alpha)$ test in regular cases.

In regular cases, where all the score functions with respect to parameters in the model are non-degenerate and the Fisher information matrix is non-singular, the $C(\alpha)$ test is constructed as follows. Suppose we have $X_1, \ldots, X_n$ as i.i.d. random variables with density $p(x; \xi, \theta)$ where θ are nuisance parameters belonging to $\Theta \subset \mathbb{R}^p$ and ξ are parameters under test that belong to $\Xi \subset \mathbb{R}^q$. For densities satisfing the regularity conditions (Neyman (1959, Definition 3)), we consider testing the hypothesis $H_0 : \xi = \xi_0$ against $H_a : \xi \in \Xi \setminus \{\xi_0\}$ while nuisance parameters $\theta \in \Theta$ are left unspecified. We define the conventional score functions as

$$C_{\xi,n} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \nabla_\xi \log p(X_i; \xi, \theta)|_{\xi=\xi_0}$$

$$C_{\theta,n} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \nabla_\theta \log p(X_i; \xi, \theta)|_{\xi=\xi_0}$$

and denote the corresponding Fisher information matrix as,

$$I = \begin{pmatrix} I_{\xi\xi} & I_{\xi\theta} \\ I_{\theta\xi} & I_{\theta\theta} \end{pmatrix}.$$

Since nuisance parameters θ are left unspecified by H_0, Neyman (1959) shows that for the test statistic to have the same asymptotic behavior when we replace the nuisance parameters θ by any $\sqrt{n}$-consistent estimator $\hat{\theta}_n$, it is necessary and sufficient for the test statistics to be orthogonal to $C_{\theta,n}$. For example, the "residual" score, which constitutes the vector of projecting $C_{\xi,n}$ onto the space spanned by the score vector $C_{\theta,n}$, denoted by

$$g_n(\theta) = C_{\xi,n} - I_{\xi\theta} I_{\theta\theta}^{-1} C_{\theta,n},$$

provides such a test function with variance $I_{\xi.\theta} \equiv I_{\xi\xi} - I_{\xi\theta} I_{\theta\theta}^{-1} I_{\theta\xi}$. Given a $\sqrt{n}$-consistent estimator $\hat{\theta}_n$ for θ, the $C(\alpha)$ test

$$T_n(\hat{\theta}_n) = g_n(\hat{\theta}_n)^\top I_{\xi.\theta}^{-1} g_n(\hat{\theta}_n)$$

is then asymptotically χ^2_q under H_0 and is optimal for local alternatives of the form $\xi_n = \xi_0 + \delta/\sqrt{n}$. When $\hat{\theta}_n$ is the restricted maximum likelihood estimator of θ, $C_{\theta,n}$ is zero and the $C(\alpha)$ test reduces to Rao's score test. The component $I_{\xi\theta} I_{\theta\theta}^{-1} I_{\theta\xi}$ subtracted from the information $I_{\xi\xi}$ for ξ measures the amount of information lost due to not knowing

the nuisance parameters (see e.g. Bickel, Klaassen, Ritov, and Wellner (1993), section 2.4).

2.2. **Testing for unobserved parameter heterogeneity.** The $C(\alpha)$ test for unobserved heterogeneity is usually formulated under a random parameter model. Following Neyman and Scott (1966) we will focus initially on testing homogeneity of a scalar parameter against the alternative that the parameter is random. Consider having i.i.d. random variables $X_1, \ldots, X_n$, with each X_i having density function $p(x; \lambda_i)$. Heterogeneity of the model is introduced by regarding the individual specific λ_i as a random parameter of the form,

$$\lambda_i = \lambda_0 + \tau \xi U_i,$$

where the unobserved U_i's are independent random variables with common distribution function, F, satisfying moment conditions $\mathbb{E}(U) = 0$, $\mathbb{V}(U) = 1$. The parameter τ is a known finite scale parameter, which allows us to rescale the variance for U to be unity. It is not restrictive to assume τ known, as we will see later that τ does not enter the test statistics. It is cancelled out when the test function is studentized by its standard deviation. The hypothesis we would like to test is $H_0 : \xi = 0$, which implies $\lambda_i = \lambda_0$ for all i's. The alternative hypothesis is $H_a : \xi \neq 0$.

Under the above setup, the standard $C(\alpha)$ test described in Section 2.1 breaks down because the score function for ξ for each individual observation x_i, defined as the first order logarithmic derivative of the density function with respect to ξ, is identically zero under the null, hence the Fisher information is also zero,

$$\frac{\partial}{\partial \xi} \log \int p(x_i; \lambda_0 + \tau \xi u) dF(u) \,|_{\xi=0} = \tau \int u dF(u) \frac{p'(x_i; \lambda_0)}{p(x_i; \lambda_0)} = 0.$$

However, in circumstances like this, we can compute the second-order derivative, denoted as $s_i(\lambda_0)$ below,

$$s_i(\lambda_0) := \frac{\partial^2}{\partial \xi^2} \log \int p(x_i; \lambda_0 + \tau \xi u) dF(u) \,|_{\xi=0} = \tau^2 \int u^2 dF(u) \frac{p''(x_i; \lambda_0)}{p(x_i; \lambda_0)} = \tau^2 \frac{p''(x_i; \lambda_0)}{p(x_i; \lambda_0)}.$$

The normed sum of these independent second-order derivatives, $s(\lambda_0) = \frac{1}{\sqrt{n}} \sum_i s_i(\lambda_0)$, can be shown to be asymptotically normally distributed with mean zero and variance $\mathbb{E}(s_1^2(\lambda_0))$ under H_0 by the central limit theorem and by noticing that $\mathbb{E}(p''(x_i; \lambda_0)/p(x_i; \lambda_0)) = 0$ as a consequence of differentiating $\int p(x; \lambda) dx = 1$ as a function of λ twice. This leads to a close analogy with the classical theorem, in which $s(\lambda_0)$ acts as the score function and the variance $\mathbb{E}(s_1^2(\lambda_0))$ plays the role of the Fisher information in the irregular setting considered here.

In regular cases, score tests exploit the fact that if the null hypothesis is false, the gradient of the log likelihood should not be close to zero. Clearly this fails in the irregular case, because no matter how data is generated, the gradient is always zero. It is natural then to make use of the curvature information provided by the second-order derivative for inference. If the null is false, one expects the second-order derivative to be positive. We will see that this second-order score function plays the essential role of constructing the $C(\alpha)$ test for

unobserved heterogeneity. The positivity condition also anticipates that the $C(\alpha)$ test will be one-sided. The goal of the remaining part of this section is to show that the optimality of the $C(\alpha)$ test, as in the regular case, is still preserved under this irregularity and its asymptotic theory, although different from the regular cases in certain perspectives, still takes a simple form.

2.3. **Asymptotic optimality of the $C(\alpha)$ test for parameter heterogeneity.** Under the irregularity discussed above, in order to establish the optimality of the test statistics based on the second-order score function, one could consider modifying the Cramér type regularity conditions in Neyman (1959, Definition 3), requiring the density function to be five times differentiable pointwise and impose a Lipschitz condition on the fifth order derivative with respect to the parameter under test. The main motivation is to obtain a quadratic approximation of the log likelihood ratio using the second-order score function through a higher order Taylor expansion. To be more specific, using the example in Section 2.2 as an illustration, for local alternatives $\lambda_i = \lambda_0 + \tau\xi_n U_i$, with ξ_n be a sequence that converges to zero at certain rate, we have the following Taylor expansion of the log likelihood ratio,

$$\begin{aligned}\Lambda_n = \textstyle\sum_i \log \frac{p(x_i;\lambda_i)}{p(x_I;\lambda_0)} = \frac{\xi_n^2\tau^2}{2!}\mathbb{E}(U^2)\sum_i s_i(\lambda_0) + \frac{\xi_n^3\tau^3}{3!}\mathbb{E}(U^3)\sum_i \frac{\nabla_\lambda^3 p(x_i;\lambda_0)}{p(x_i;\lambda_0)} \\ + \frac{\xi_n^4\tau^4}{4!}\left[\mathbb{E}(U^4)\sum_i \frac{\nabla_\lambda^4 p(x_i;\lambda_0)}{p(x_i;\lambda_0)} - 3\mathbb{E}(U^2)^2\sum_i s_i^2(\lambda_0)\right] + o_P(1).\end{aligned}$$

Let ξ_n be of order $n^{-1/4}$ and provided the third and fourth moments of U are finite in addition to the zero mean and unit variance assumption, we obtain a quadratic approximation of the log-likelihood. More details of such regularity conditions can be found in Rotnitzky, Cox, Bottai, and Robins (2000), in which they consider the maximum likelihood estimation of ξ in the irregular cases in a very general context. Lindsay (1995, Chapter 4) also has a brief discussion of this.

An alternative formulation, rooted in LeCam's local asymptotic normality (LAN) theory, can be based on his differentiability in quadratic mean (DQM) condition. The latter condition is less stringent in regular cases: while Cramér conditions assume the density to be three times differentiable and impose a Lipschitz condition on the third order derivative, the DQM condition only requires first order differentiability and the derivative to be square integrable in $\mathcal{L}_2$ space. Pollard (1997) provides a nice discussion of the DQM condition in these regular cases. This is the new approach we take for analyzing the asymptotic behavior of the $C(\alpha)$ test for heterogeneity. We will show below that by modifying the DQM condition slightly, we can obtain the local asymptotic normality of the log-likelihood ratio and establish the asymptotic optimality of the $C(\alpha)$ test for the irregular cases under assumptions much weaker than those suggested by the classical Neyman's approach. One prominent example for which the classical conditions fail while the DQM conditions are satisfied is the double exponential location model with $p_\theta(x) = f(x-\theta)$ and $f(x) = \frac{1}{2}\exp(-|x|)$. For this model, the density function f is not differentiable at 0 but it satisfies the DQM condition. We would thus have no difficulty constructing a test for homogeneity in the location parameter for this model under the LeCam type conditions.

Suppose we have a random sample $(X_1, \ldots, X_n)$ with density function $p(x; \xi, \theta)$ with respect to some measure μ. The joint distribution of this i.i.d. random sample will be denoted as $P_{n,\xi,\theta}$, which is the product of n copies of the marginal distribution $P(x; \xi, \theta)$.

Assumption 1. The density function p satisfies the following conditions:

(1) ξ_0 is an interior point of Ξ
(2) For all $\theta \in \Theta \subset \mathbb{R}^p$ and $\xi \in \Xi \subset \mathbb{R}$, the density is twice continuously differentiable with respect to ξ and once continuously differentiable with respect to θ for μ-almost all x.
(3) Denoting the first two derivatives of the density with respect to ξ evaluated under the null as $\nabla_\xi p(x; \xi_0, \theta)$ and $\nabla^2_\xi p(x; \xi_0, \theta)$, we have $\mathbb{P}\left(\nabla_\xi p(x; \xi_0, \theta) = 0\right) = 1$ and $\mathbb{P}\left(\nabla^2_\xi p(x; \xi_0, \theta) \neq 0\right) > 0$ for all $\theta \in \Theta \subset \mathbb{R}^p$.
(4) Denoting the derivative of the density with respect to θ evaluated under the null as $\nabla_\theta p(x; \xi_0, \theta)$, for any p-dimensional vector a, $\mathbb{P}\left(\nabla^2_\xi p(x; \xi_0, \theta) \neq a^\top \nabla_\theta p(x; \xi_0, \theta)\right) > 0$.

Remark. Here ξ is the parameter under test and θ is the vector of nuisance parameters. The list of regularity conditions in Assumption 1 tailors the standard conditions for a regular $C(\alpha)$ test to the heterogeneity test we consider here. In particular, condition (3) reflects the irregularity of these tests that the first order logarithmic derivative with respect to ξ vanishes but the second-order derivative is non-vanishing. Condition (2) secures existence of the respective derivatives. Condition (4) rules out the case where there is a perfect linear relationship between the second-order score for ξ and the score for θ. It ensures the new Fisher information thus defined to be non-singular and the $C(\alpha)$ test statistics to be non-degenerate.

Under Assumption 1, we can now define the modified DQM condition that is crucial for establishing the local asymptotic normality of the model.

Definition 1. The density $p(x; \xi, \theta)$ satisfies the modified differentiability in quadratic mean condition at (ξ_0, θ) if there exists a vector $v(x) = (v_\xi(x), v_\theta^\top(x))^\top \in \mathcal{L}_2(\mu)$ such that as $(\xi_n, \theta_n) \to (\xi_0, \theta)$,

$$\int |\sqrt{p(x; \xi_n, \theta_n)} - \sqrt{p(x; \xi_0, \theta)} - h_n^\top v(x)|^2 d\mu(x) = o(\|h_n\|^2)$$

where $h_n = ((\xi_n - \xi_0)^2, (\theta_n - \theta)^\top)^\top$. Here $\|\cdot\|$ denotes the Euclidean norm and $\mathcal{L}_2(\mu)$ denotes the $\mathcal{L}_2$ space of square integrable functions with respect to measure μ.

Furthermore, let $\beta(h_n)$ be the mass of the part of $p(x; \xi_n, \theta_n)$ that is $p(x; \xi_0, \theta)$-singular, then as $(\xi_n, \theta_n) \to (\xi_0, \theta)$,

$$\frac{\beta(h_n)}{\|h_n\|^2} \to 0$$

Usually the vector $v(x)$ contains derivatives of the square root of density $\sqrt{p(x; \xi_n, \theta_n)}$ with respect to each parameter evaluated under their null value. Definition 1 modifies the

classical DQM condition such that whenever the first order derivative is identically zero for certain parameters, it is differentiated again until it is nonvanishing. The corresponding terms in h_n also need to be raised to the same power. For the heterogeneity test, the score function with respect to ξ is of second order and its associated term in h_n is hence quadratic. This further implies that the contiguous alternatives must be $O(n^{-1/4})$. For the following theorems, we will thus focus on the sequence of local models on $(X_1, \dots, X_n)$ with joint distribution P_{n,ξ_n,θ_n} in which $\xi_n = \xi_0 + \delta_1 n^{-1/4}$ and $\theta_n = \theta + \delta_2 n^{-1/2}$.

Theorem 1. Suppose $(X_1, \dots, X_n)$ are i.i.d. random variables with joint distribution P_{n,ξ_n,θ_n} and the density satisfies Assumption 1 and the modified DQM condition with

$$\nu(x) = (\nu_\xi(x), \nu_\theta^\top(x))^\top = \left(\frac{1}{4} \frac{\nabla_\xi^2 p(x;\xi_0,\theta)}{\sqrt{p(x;\xi_0,\theta)}} \mathbb{I}_{[p(x;\xi_0,\theta)>0]}, \frac{1}{2} \frac{\nabla_\theta p(x;\xi_0,\theta)^\top}{\sqrt{p(x;\xi_0,\theta)}} \mathbb{I}_{[p(x;\xi_0,\theta)>0]} \right)^\top,$$

then for fixed δ_1 and δ_2, the log-likelihood ratio has the following quadratic approximation under the null:

$$\Lambda_n = \log \frac{dP_{n,\xi_n,\theta_n}}{dP_{n,\xi_0,\theta}} = t^\top S_n - \frac{1}{2} t^\top J t + o_P(1)$$

where $t = (\delta_1^2, \delta_2^\top)^\top$,

$$S_n = (S_{\xi,n}, S_{\theta,n}^\top)^\top = \left(\frac{2}{\sqrt{n}} \sum_i \frac{\nu_\xi(x_i)}{\sqrt{p(x_i;\xi_0,\theta)}}, \frac{2}{\sqrt{n}} \sum_i \frac{\nu_\theta^\top(x_i)}{\sqrt{p(x_i;\xi_0,\theta)}} \right)^\top$$

and

$$J = 4 \int (\nu\nu^\top) d\mu(x) = \begin{pmatrix} \mathbb{E}(S_{\xi,n}^2) & \mathrm{Cov}(S_{\xi,n}, S_{\theta,n}^\top) \\ \mathrm{Cov}(S_{\xi,n}, S_{\theta,n}) & \mathbb{E}(S_{\theta,n} S_{\theta,n}^\top) \end{pmatrix} \equiv \begin{pmatrix} J_{\xi\xi} & J_{\xi\theta} \\ J_{\theta\xi} & J_{\theta\theta} \end{pmatrix}.$$

Corollary 1. With S_n and J defined as in Theorem 1, we have

$$S_n \overset{P_{n,\xi_0,\theta}}{\rightsquigarrow} \mathcal{N}(0, J),$$

and hence the sequence of models P_{n,ξ_n,θ_n} is locally asymptotically normal (LAN) at (ξ_0, θ) with S_n being interpreted as the score vector and J as the associated Fisher information matrix. Furthermore, P_{n,ξ_n,θ_n} is mutually contiguous to $P_{n,\xi_0,\theta}$.

Theorem 1 shows that under Assumption 1, the modified DQM condition is sufficient for obtaining a quadratic approximation of the log-likelihood ratio for the sequence of local models in the $n^{-1/4}$ neighborhood of the null value ξ_0 and the $n^{-1/2}$ neighborhood of the nuisance parameter θ. The joint normality of the vector S_n, as established in Corollary 1, further indicates the LAN property of this sequence of models. It is important to note that the vector S_n, in which the degenerately zero first-order score function for ξ is replaced by the corresponding second-order derivative of the log-likelihood, acts as the score vector

in this irregular case. Naturally, J has the interpretation of the Fisher information matrix. Under Assumption 1, since we rule out perfect dependence between $S_{\xi,n}$ and $S_{\theta,n}$ in condition (4), J is non-singular.

Having established the LAN property of this sequence of local models, we can now make use of LeCam's (1972) limit experiment theory to show that the $C(\alpha)$ test is locally asymptotically optimal in the scalar case.

Following the definitions given in LeCam (1972) and van der Vaart (1998), an experiment $\mathcal{E}$ indexed by a parameter set H is a collection of probability measures $\{P_h : h \in H\}$ on the sample space $(\mathcal{X}, \mathcal{A})$. A sequence of experiments $\mathcal{E}_n = (\mathcal{X}_n, \mathcal{A}_n, P_{n,h} : h \in H)$ is said to converge to a limit experiment $\mathcal{E} = (\mathcal{X}, \mathcal{A}, P_h : h \in H)$ if the likelihood ratio process for $\mathcal{E}_n$, $\frac{dP_{n,h}}{dP_{n,h_0}}(X_n)$, converges in distribution to the likelihood ratio of the limit experiment, $\frac{dP_h}{dP_{h_0}}(X)$, for h in every finite subset $I \subset H$ and for every null value $h_0 \in H$. A common feature is that many sequences of experiments produce a Gaussian limit experiment. One important example is that for i.i.d. sample from a smooth parametric model with distribution P_ϑ, if the sequence of the local model P_{n,ϑ_n} in which $\vartheta_n = \vartheta_0 + r_n\delta$ with r_n as the appropriate norming rate is locally asymptotically normal, then it has a Gaussian shift experiment as its limit.

The advantage of establishing the limit experiment is several fold. First, the limit experiment is often easier to analyze than the original sequence of models. Second, the limit experiment provides a bound for the optimal estimation (in terms of lower bound on the asymptotic variance) or testing procedure (in terms of upper bound on the asymptotic power) one could achieve in the original model. Third, by the asymptotic representation theory (van der Vaart (1998, Chapter 9)), any sequence of statistics that converges in the original experiment can be matched in the limit experiment and they share identical asymptotic behavior. We will show in particular that the $C(\alpha)$ test statistics is matched with the optimal testing procedure in the Gaussian shift limit experiment, hence establishing its optimality. We first focus on the scalar case, leaving the multi-dimensional case to a separate discussion.

Theorem 2. Let $\mathcal{E}_n$ be a sequence of experiments based on i.i.d. random variables $(X_1, \ldots, X_n)$ with joint distribution P_{n,ξ_n,θ_n} on the sample space $(\mathcal{X}_n, \mathcal{A}_n)$. We further index the sequence of experiment by $t = (\delta_1^2, \delta_2^\top)^\top \in \mathbb{R}_+ \times \mathbb{R}^p$. The log-likelihood ratio of the sequence of models satisfies,

$$\log\left(\frac{dP_{n,\xi_n,\theta_n}}{dP_{n,\xi_0,\theta}}\right) = t^\top S_n - \tfrac{1}{2} t^\top J t + o_P(1),$$

with the score vector S_n defined as in Theorem 1 converging in distribution under the null to $\mathcal{N}(0, J)$. Then the sequence of experiments $\mathcal{E}_n$ converges to the limit experiment based on observing one sample from $Y = t + \nu$, where $\nu \sim \mathcal{N}(0, J^{-1})$. The locally asymptotically optimal statistic for testing $H_0 : \delta_1 = 0$ vs. $H_a : \delta_1 \neq 0$ is

$$Z_n = (J_{\xi\xi} - J_{\xi\theta} J_{\theta\theta}^{-1} J_{\theta\xi})^{-1/2} (S_{\xi,n} - J_{\xi\theta} J_{\theta\theta}^{-1} S_{\theta,n}).$$

Corollary 2. Under H_0, Z_n has distribution $\mathcal{N}(0,1)$. Under H_a, by applying LeCam's third lemma (see e.g. van der Vaart (1998, Example 6.7)), it follows a shifted normal distribution $\mathcal{N}(\delta_1^2(J_{\xi\xi} - J_{\xi\theta}J_{\theta\theta}^{-1}J_{\theta\xi})^{1/2}, 1)$.

The optimal test statistic Z_n takes the form of a C(α) test. It projects the second-order score $S_{\xi,n}$ for ξ onto the space spanned by the first-order score vector $S_{\theta,n}$ for θ. It is the sequence of statistics from the original experiment that can be matched with the optimal test statistic in the limit Gaussian experiment for inference on δ_1, which is the first element in the one sample Y.

One common feature of C(α) heterogeneity tests is that the limit distribution under local alternative is always a right-shifted normal distribution even if we have a two-sided alternative hypothesis for δ_1. This is not surprising given that the shift parameter corresponding to ξ in the Gaussian limit experiment is a quadratic term $\delta_1^2 \in \mathbb{R}_+$. In other words, the best inference procedure one could possibly achieve in the limit experiment is for δ_1^2. We lose the sign information on δ_1, and the asymptotically optimal test, if rejects the null, fails to distinguish whether the deviation is from the left or from the right (this phenomenon is also emphasized in Rotnitzky, Cox, Bottai, and Robins (2000)). Let $Y = (Y_1, Y_2)^\top$ where the partition is such that Y_1 is a scalar and $Y_2 \in \mathbb{R}^p$ as in Theorem 2. In the Gaussian limit experiment based on the one sample from $Y \sim \mathcal{N}(t, J^{-1})$, the one-sided test, rejecting H_0 if $Y_1 \geqslant \Phi^{-1}(1-\alpha)(J_{\xi\xi} - J_{\xi\theta}J_{\theta\theta}^{-1}J_{\theta\xi})^{-1/2}$, is the uniformly most powerful test. Since the sequence that converges to the rescaled first element $(J_{\xi\xi} - J_{\xi\theta}J_{\theta\theta}^{-1}J_{\theta\xi})^{1/2}Y_1$ is exactly Z_n, it implies that the asymptotic C(α) test rejects H_0 if $Z_n \geqslant \Phi^{-1}(1-\alpha)$ for any level α. Observe that for $\alpha < 0.5$, this is equivalent to rejecting H_0 if $(0 \vee Z_n)^2 > c$, where c is the $(1-\alpha)$-quantile of $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ and χ_0^2 is a degenerate distribution with mass 1 at 0. No solution exists for c if $\alpha > 0.5$ although this is of little relevance in practice. We mention the mixture of χ^2 asymptotics just to be more cohesive with the multi-dimensional extension later. The weight 1/2 associated with χ_0^2 is due to the fact that Z_n takes negative values with probability 1/2 under H_0.

There is another intuitive interpretation of the one-sidedness of the test, as we have already anticipated in Section 2.2. The C(α) test Z_n, constructed from the second-order score for ξ, exploits information of the curvature of the log-likelihood function. Since at $\xi = \xi_0$, the gradient of the log-likelihood function with respect to ξ is always zero, it depends on the sign of the second-order derivative to determine whether the null point is a local maximum or a local minimum. Only positive values of Z_n indicates the null point as a local minimum of the log-likelihood function, leading to a rejection of the null hypothesis. As $n \to \infty$, due to normality of Z_n, only half the time we get the "correct" curvature allowing us to reject the null. In the simulation exercise in Section 4, we show that paying attention to this one-sided feature in the decision rule gives more power on testing for parameter heterogeneity.

For the random parameter model, one could of course also consider a likelihood ratio test as an alternative testing strategy for heterogeneity. Among many others, Chen, Chen, and Kalbfleisch (2001) considers a modified likelihood ratio test for homogeneity in finite

mixture models, which is very close to the setup we consider in this paper. They also obtain a mixture of χ^2 asymptotics for their likelihood ratio test statistics. Their modified LRT can be viewed as an asymptotically equivalent testing procedure in mixture models to the $C(\alpha)$ test considered here. The latter, however, inheriting the nice feature of the score test, is much easier to compute. Furthermore, the $C(\alpha)$ test statistics does not depend on the specification of F as long as the moment conditions are satisfied. This can be viewed as a merit of the test because it has power for a large class of alternative models. On the other hand, it can also be viewed as its disadvantage because rejecting the hypothesis does not provide information on what plausible alternatives might be. Comparison between the general LR test for mixture models and the $C(\alpha)$ test is considered in Gu, Koenker, and Volgushev (2013).

The result established thus far is not confined to the heterogeneity test problem. It is applicable whenever the first-order score for the parameter under test vanishes but the second-order score is non-degenerate. There is another possible scenario for the score test to break down, in which none of the first-order score function is vanishing, but there is linear dependence among them, and thus the Fisher information matrix becomes singular. This is the case discussed in considerable detail in Lee and Chesher (1986). Models with selection bias and the stochastic production frontier models fall into this class. They propose an extremum test which is based on the determinant of the matrix of the second-order derivatives of the log likelihood function and show the asymptotic optimality of the test. The extremum test can essentially be reformulated, using a reparameterization slightly different from what the authors suggested in the paper (i.e. choose k to be 1 in Lee and Chesher (1986, p. 132)), to fit into the conditions described in Assumption 1. The similar irregularity also arises in test for symmetry in normal-skew distribution and is investigated in Hallin and Ley (2013). The reparameterization is a Gram-Schmidt orthogonalization in the same spirit of Rotnitzky, Cox, Bottai, and Robins (2000, Section 4.4). The $C(\alpha)$ test can then be constructed and asymptotic optimality of the test follows.

2.4. **Replacing the nuisance parameter by a $\sqrt{n}$-consistent estimator.** Notice that the optimal test statistic Z_n we obtained in Theorem 2 is a function of θ, to make the test statistic feasible under unknown nuisance parameters, we need to replace θ by some estimator $\hat{\theta}$. In order to ensure that the asymptotics for the test statistic Z_n in Corollary 2 is still valid, it suffices to show that $Z_n(\hat{\theta}) - Z_n(\theta) = o_P(1)$ both under the null and local alternatives. There are various ways to obtain this result. The classical approach taken in Neyman (1959) was to make additional differentiability and bound conditions on the test function $g(x_i, \theta)$, which is defined as

$$g(x_i;\theta) = (J_{\xi\xi} - J_{\xi\theta}J_{\theta\theta}^{-1}J_{\theta\xi})^{-1/2}\Big(\frac{2v_\xi(x_i)}{\sqrt{p(x_i;\xi_0,\theta)}} - J_{\xi\theta}J_{\theta\theta}^{-1}\frac{2v_\theta(x_i)}{\sqrt{p(x_i;\xi_0,\theta)}}\Big),$$

such that $Z_n(\theta) = \frac{1}{\sqrt{n}}\sum_i g(x_i,\theta)$. Details of these assumptions can be found in Neyman (1959, Definition 3 (ii) (iii)) and we will not replicate them here. When the conditions are satisfied, Taylor expansion of $Z_n(\hat{\theta})$ around $Z_n(\theta)$ yields the desired results for $\hat{\theta}$ being any $\sqrt{n}$-consistent estimator for θ. Neyman's assumptions are rather strong, for example, he requires the density to be three times differentiable with respect to θ and also moments of

the gradient of g with respect to θ to be continuous. LeCam proposes a discretization trick which works as long as the model satisfies a uniform LAN condition and the $\sqrt{n}$-consistent estimator satisfies an asymptotic discreteness property. The trick is quite standard in one-step estimation problems. Our approach, using more modern probability theory, is to view the difference $Z_n(\hat{\theta}) - Z_n(\theta)$ as an empirical process. More precisely, we make the following assumption on the test function $g(x, \theta)$ to establish the equicontinuity of the empirical process. In fact our Assumption 2 below on $g(x, \theta)$ implies the conditions of the Type IV function in Andrews (1994) with $p = 2$. We include these results here for completeness.

Assumption 2. There exists some $\delta > 0$ such that for any $\eta, \eta' \in U_\delta(\theta)$ we have for some $\gamma > 0$

$$|g(x, \eta) - g(x, \eta')| \leqslant \|\eta - \eta'\|^\gamma H(x)$$

for $P_{n,\xi_n,\theta}$-almost all x (for every $n \in \mathbb{N}$) where H is square integrable with respect to $P_{n,\xi_n,\theta}$ for all $n \in \mathbb{N}$, $\sup_n \mathbb{E}_{P_{n,\xi_n,\theta}} H^2(X) < \infty$ and additionally for some $c_n = o(1)$, $n^{1/2}\mathbb{E}_{P_{n,\xi_n,\theta}}[H(X)\mathbb{I}_{\{H(X) > n^{1/2}c_n\}}] = o(1)$.

Theorem 3. Under Assumption 2, if $\hat{\theta}$ is a $\sqrt{n}$-consistent estimator for θ, then

$$|Z_n(\hat{\theta}) - Z_n(\theta)| = o_P(1)$$

2.5. $C(\alpha)$ **test for parameter heterogeneity in higher dimensions.** It is of interest to generalize the $C(\alpha)$ tests of unobserved parameter heterogeneity to higher dimensions in the irregular case. For example, in a linear regression model, we may want to jointly test for slope heterogeneity for more than one covariates. When panel data is available, we may want to test for heterogeneity in the slope coefficients in the presence of individual variances, see for example Pesaran and Yamagata (2008). The main challenge comes from the one-sidedness of the test. Fortunately, the limit experiment turns out to be multivariate Gaussian with location shifts in each coordinate (or in a subset of coordinates) towards the right tail. This naturally requires us to look for optimal tests for deviations of the location parameters of the multivariate Gaussian from zero restrictions to the positive orthant.

To be more specific, suppose the limit multivariate Gaussian experiment has mean vector $(\mu_1, \ldots, \mu_q)$, we would like to test $H_0 : \mu_i = 0$ for $i = 1, \ldots, q$ against the alternative $H_a : \mu_i \geqslant 0$ for $i = 1, \ldots, q$ with at least one inequality holds strictly. Unlike in the univariate case where the one-sided test is optimal in the sense of being uniformly most powerful and hence the asymptotic analogue $C(\alpha)$ test obtains the same optimality locally asymptotically, there exists no uniformly optimal test for the multivariate case. There are two dominant options in the literature. The likelihood ratio test has been studied by many authors. Chernoff (1954) extends the classical Wilks's result on likelihood ratio test (LRT) to cases in which the null value of the parameters under test lie on the boundary of the parameter space. Bartholomew (1961), Nüesch (1966) and Perlman (1969) among many others consider variants of Gaussian LRT under restricted alternatives. Hillier (1986) provides details for the LRT with dimension equal to three. Self and Liang (1987) give some further examples for LRT with nuisance parameters. Alternatively, Abelson and Tukey (1963) propose tests based on the idea of maximin contrast and this is further extended

by Schaafsma and Smid (1966) who introduce the optimality concept of "most stringent somewhere most powerful" (MSSMP) test.

Neither LRT nor MSSMP test uniformly dominates each other, but both are shown to be substantially more powerful than the usual χ^2 or F test for the multivariate Gaussian case. We construct the $C(\alpha)$ test by extending the LRT via the limit experiment into its local asymptotic version. It allows a direct power comparison to the usual χ^2 test, i.e. the Information Matrix test, which ignores the positivity constraints. It is also closer to the historical development of generalization of the regular $C(\alpha)$ test to the multidimensional case by Bühler and Puri (1966), which can be viewed as the asymptotic analogue of the usual χ^2 test in the Gaussian limit experiment for testing $\mu_i = 0$ against non-constrained alternative $\mu_i \neq 0$. Additionally, as we will show, the $C(\alpha)$ test can also be easily adapted if only a subset of the shift parameters are subject to positivity constraints.

The LRT statistics for these one-sided test problems in multi-dimensions all obtain a mixture of χ^2 with different degrees of freedom as their asymptotic distribution. One disadvantage of the LRT is that the weights of these χ^2's get complicated very quickly as dimension increases in most cases. In contrast, the MSSMP test has a standard normal asymptotics. However, it is hard to adapt the MSSMP test to situations where only a subset of the shift parameters are subject to constraints.

We will present in details the joint test for heterogeneity in dimension two as an illustration and comment on the more general case. Suppose again we have i.i.d. random sample $(X_1, \ldots, X_n)$ with density $p(x; \xi, \theta)$. The parameters under test are now $\xi = (\xi_1, \xi_2) \in \Xi \subset \mathbb{R}^2$. They take null value $\xi_0 = (\xi_{10}, \xi_{20})$ and $\theta \in \Theta \subset \mathbb{R}^p$ are the nuisance parameters. For heterogeneity tests in particular, we consider testing for heterogeneity of a vector of parameters, λ_i, of the model. Under the alternative, they take the form, $\lambda_{ki} = \theta_k + \tau\xi_k U_{ki}$, for $k = 1, 2$. Let the covariance matrix for $U_i = (U_{1i}, U_{2i})$ be Ω. Without loss of generality, we let the diagonal element of Ω be unity. Under $H_0, \xi_k = 0$, so that λ_k's are homogenous across individuals taking value θ_k.

The density function satisfies Assumption 1 such that the first-order score vector for ξ_1 and ξ_2 are vanishing due to the zero mean assumption for U_i but each elements in the second-order score matrix are non-vanishing. It also satisfies the modified DQM condition so that the model is locally asymptotically normal. Typically the score function for (ξ_1, ξ_2) then consists of all distinct elements in the second-order score matrix. Depending on the assumption on Ω, some of the elements become zero. For example, if Ω is a diagonal matrix, which implies that U_{1i} is mutually independent to U_{2i}, then the off-diagonal terms of the second-order score matrix for ξ are zero. If Ω has non-zero off-diagonal elements, then the corresponding cross terms in the score matrix are also non-vanishing and need to be included.

It is crucial to distinguish the above-mentioned two scenarios, since the diagonal terms in the score matrix correspond to the shift terms in the Gaussian limit experiment that are

subject to positivity constraints, while the off-diagnonal terms correspond to shift parameters that can take value over the whole real line. This implies that if Ω is not diagonal, then the Gaussian limit experiment has only a subset of the shift parameters that have positivity constraints under the alternative. Theorem 4 gives the general theory on constructing the C(α) statistics with the subsequent Corollary 3 discussing the special case if Ω is diagonal.

To proceed, we denote the second-order score vector (all distinct elements in the second order score matrix stacked into a vector) for (ξ_1, ξ_2) as $(S_{\xi_1^2,n}, S_{\xi_2^2,n}, S_{\xi_1\xi_2,n})$. The first two corresponds to the diagonal terms and the last the off-diagonal term. Let the first-order score for θ be $S_{\theta,n}$. More specifically, under regularity conditions, they are $S_{\xi_k^2,n} = \frac{1}{2\sqrt{n}}\sum_i \frac{\nabla^2_{\xi_k\xi_k} p(x_i;\xi_0,\theta)}{p(x_i;\xi_0,\theta)}\mathbb{I}_{[p(x_i;\xi_0,\theta)>0]}$ for $k = 1, 2$; and $S_{\xi_1\xi_2,n} = \frac{1}{2\sqrt{n}}\sum_i \frac{\nabla^2_{\xi_1\xi_2} p(x_i;\xi_0,\theta)}{p(x_i;\xi_0,\theta)}\mathbb{I}_{[p(x_i;\xi_0,\theta)>0]}$ and $S_{\theta,n} = \frac{1}{\sqrt{n}}\sum_i \frac{\nabla_\theta p(x_i;\xi_0,\theta)^\top}{p(x_i;\xi_0,\theta)}\mathbb{I}_{[p(x_i;\xi_0,\theta)>0]}$. Let the associated information matrix be denoted as, $J = \begin{pmatrix} J_{\xi\xi} & J_{\xi\theta} \\ J_{\theta\xi} & J_{\theta\theta} \end{pmatrix}$, with $J_{\xi\xi}$ being a 3×3 block matrix. The residual score for ξ, similar to the scalar case, is found to be

$$\tilde{S}_{\xi,n} = \begin{pmatrix} \tilde{S}_{\xi_1^2,n} \\ \tilde{S}_{\xi_2^2,n} \\ \tilde{S}_{\xi_1\xi_2,n} \end{pmatrix} := \begin{pmatrix} S_{\xi_1^2,n} \\ S_{\xi_2^2,n} \\ S_{\xi_1\xi_2,n} \end{pmatrix} - J_{\xi\theta}J_{\theta\theta}^{-1}S_{\theta,n}$$

and the covariance matrix for $\tilde{S}_{\xi,n}$ is $\Sigma = J_{\xi\xi} - J_{\xi\theta}J_{\theta\theta}^{-1}J_{\theta\xi} := \begin{pmatrix} \Sigma_{(11)} & \Sigma_{(12)} \\ \Sigma_{(21)} & \Sigma_{(22)} \end{pmatrix}$. The partition of Σ is such that $\Sigma_{(11)}$ collects covariance terms for the first two elements in $\tilde{S}_{\xi,n}$.

Theorem 4. Let υ_n be the sequence of experiments based on i.i.d. random variable $(X_1, \ldots, X_n)$ with joint distribution P_{n,ξ_n,θ_n} with $\xi_n = (\xi_{10}, \xi_{20}) + (\delta_1, \delta_2)n^{-1/4}$ and $\theta_n = \theta + \delta_3 n^{-1/2}$ on the sample space $(\mathcal{X}_n, \mathcal{A}_n)$. The log-likelihood ratio of the sequence of experiment satisfies,

$$\log\left(\frac{dP_{n,\xi_n,\theta_n}}{dP_{n,\xi_0,\theta}}\right) = t^\top S_n - \frac{1}{2}t^\top J t + o_p(1),$$

with $S_n = (S_{\xi_1^2,n}, S_{\xi_2^2,n}, S_{\xi_1\xi_2,n}, S_{\theta,n}^\top)^\top \sim \mathcal{N}(0, J)$. Then the limit experiment of υ_n is based on observing one sample from $Y = t + \nu$ with $t = (\delta_1^2, \delta_2^2, 2\delta_1\delta_2, \delta_3^\top)^\top \in \mathbb{R}_+^2 \times \mathbb{R} \times \mathbb{R}^p$ and $\nu \sim \mathcal{N}(0, J^{-1})$. We would like to jointly test $H_0 : \delta_1 = \delta_2 = 0$ against the alternative $H_a : \delta_1 \neq 0$ or $\delta_2 \neq 0$. Let $u_n := (u_{1n}, u_{2n})^\top = (\tilde{S}_{\xi_1^2,n}, \tilde{S}_{\xi_2^2,n})^\top - \Sigma_{(12)}\Sigma_{(22)}^{-1}\tilde{S}_{\xi_1\xi_2,n}$ and let Λ be the Cholesky decompositon of $\Sigma_{11.2} := \Sigma_{(11)} - \Sigma_{(12)}\Sigma_{(22)}^{-1}\Sigma_{(21)}$, that is

$$\Lambda = \begin{pmatrix} \sqrt{\nu_1} & 0 \\ \rho\sqrt{\nu_2} & \sqrt{\nu_2}\sqrt{1-\rho^2} \end{pmatrix}$$

where ρ is the correlation coefficient between u_{1n} and u_{2n} and ν_1 and ν_2 are their respective variances. Define $w_n = (w_{1n}, w_{2n})^\top$ as

$$w_n \equiv \Lambda^{-1}u_n = \begin{pmatrix} u_{1n}/\sqrt{\nu_1} \\ (1-\rho^2)^{-1/2}(u_{2n}/\sqrt{\nu_2} - \rho u_{1n}/\sqrt{\nu_1}) \end{pmatrix}$$

and let $w_{3n} := \Sigma_{(22)}^{-1/2} S_{\xi_1\xi_2,n}$. The $C(\alpha)$ test statistic is one of the following four cases:

$$T_n = \begin{cases} w_{1n}^2 + w_{2n}^2 + w_{3n}^2 & \text{if } w_{1n} \geqslant \frac{\rho}{\sqrt{1-\rho^2}} w_{2n}, w_{2n} \geqslant 0 \\ w_{1n}^2 + w_{3n}^2 & \text{if } w_{2n} \leqslant 0, w_{1n} \geqslant 0 \\ (\rho w_{1n} + \sqrt{1-\rho^2} w_{2n})^2 + w_{3n}^2 & \text{if } -\frac{\sqrt{1-\rho^2}}{\rho} w_{2n} \leqslant w_{1n} \leqslant \frac{\rho}{\sqrt{1-\rho^2}} w_{2n} \\ & w_{2n} \geqslant 0 \\ w_{3n}^2 & \text{if } w_{1n} \leqslant 0, w_{2n} \leqslant -\frac{\rho}{\sqrt{1-\rho^2}} w_{1n} \end{cases}$$

Under H_0, the asymptotic distribution of T_n follows $(\frac{1}{2} - \frac{\beta}{2\pi})\chi_1^2 + \frac{1}{2}\chi_2^2 + \frac{\beta}{2\pi}\chi_3^2$ with $\beta = \cos^{-1}(\rho)$.

Corollary 3. If $S_{\xi_1\xi_2,n} = 0$, then the log likelihood ratio of the sequence of experiment reduces to

$$\log\left(\frac{dP_{n,\xi_n,\theta_n}}{dP_{n,\xi_0,\theta}}\right) = t^\top S_n - \frac{1}{2} t^\top J t + o_p(1),$$

with $S_n = (S_{\xi_1^2,n}, S_{\xi_2^2,n}, S_{\theta,n}^\top)^\top \sim \mathcal{N}(0, J)$. Then the limit experiment of v_n is based on observing one sample from $Y = t + \nu$ with $t = (\delta_1^2, \delta_2^2, \delta_3^\top)^\top \in \mathbb{R}_+^2 \times \mathbb{R}^p$ and $\nu \sim \mathcal{N}(0, J^{-1})$. Proceed as in Theorem 4 with $u_n = (\tilde{S}_{\xi_1^2,n}, \tilde{S}_{\xi_2^2,n})^\top$ and find the corresponding Cholesky decomposition Λ for $\Sigma_{(11)}$ and $w_n = \Lambda^{-1} u_n$. Under H_0, the asymptotic distribution of T_n follows $(\frac{1}{2} - \frac{\beta}{2\pi})\chi_0^2 + \frac{1}{2}\chi_2^1 + \frac{\beta}{2\pi}\chi_2^2$ with $\beta = \cos^{-1}(\rho)$.

Remark. When dimension gets higher, the construction of the $C(\alpha)$ test follows the similar idea. We first find residual score $\tilde{S}_{\xi,n}$ for $(\xi_1, \ldots, \xi_q)$ by projecting away the effect of the score of θ. LeCam's third lemma implies that asymptotically $\tilde{S}_{\xi,n}$ follows $\mathcal{N}(0, \Sigma)$ under H_0 and $\mathcal{N}(\Sigma(\delta_1^2, \ldots, \delta_q^2, (2\delta_j\delta_k)_{j\neq k})^\top, \Sigma)$ under local alternative. The construction of the $C(\alpha)$ test is to find

$$T_n = \tilde{S}_{\xi,n}\Sigma^{-1}\tilde{S}_{\xi,n} - \inf_{\mu\in\mathcal{C}} (\Sigma^{-1}\tilde{S}_{\xi,n} - \mu)^\top \Sigma (\Sigma^{-1}\tilde{S}_{\xi,n} - \mu) \tag{1}$$

where the cone $\mathcal{C} = \mathbb{R}_+^q \times \mathbb{R}^{q(q-1)/2}$, the space of the vector $(\delta_1^2, \ldots, \delta_q^2, (2\delta_j\delta_k)_{j\neq k})^\top$. We observe that T_n is the LR statistics treating $\Sigma^{-1}\tilde{S}_{\xi,n}$ as the single observation in the limit experiment (See a similar idea in Silvapulle and Silvapulle (1995)). The w_n worked out in Theorem 4 and Corollary 3 are explicit solution for (1) when $q = 2$. For $q > 2$, the solution for μ, and therefore the test statistics T_n, can be easily found by using the R package `quadprog`, Turlach and Weignessel (2013).

The test statistic, when dimension grows, continues to follow a mixture of χ^2 distribution asymptotically under the null, albeit with more complex weights. In the simplest case, if both J and Ω happen to be diagonal matrices, then all off-diagonal terms in the second-order score matrix drop and the weights take a very simple form. For $\xi \in \Xi \subset \mathbb{R}^q$ and let the residual score for ξ be $\tilde{S}_{\xi,n}$ with its covariance matrix as Σ_q. The diagonality of

J implies diagonality of Σ_q. The optimal test statistic for $H_0 : \xi_1 = \cdots = \xi_q = 0$ against $H_a : \xi_i \neq 0$ for at least one i is

$$T_n = (0 \vee \tilde{S}_{\xi,n})^\top \Sigma_q^{-1} (0 \vee \tilde{S}_{\xi,n})$$

Under H_0, $T_n \sim \sum_{i=0}^{q} \binom{q}{i} 2^{-q} \chi_i^2$. As q becomes large, paying attention to the one-sided nature of the test achieves much better power performance than simply using the inner product of $\tilde{S}_{\xi,n}$ and the χ^2 asymptotics, because the latter wastes $1 - (1/2)^q$ portion of the type-I error. This point is also stressed in Akharif and Hallin (2003) on the optimal detection of random coefficient in autoregressive models.

3. Examples

In this section, we describe four examples of using the C(α) test for unobserved parameter heterogeneity in various models. The first Poisson regression example leads to similar test statistics already familiar in the literature. This is to illustrate that the C(α) test serves as a unification of many tests already available. As another example not fully elaborated here, Kiefer (1984) and Lancaster (1985) develop tests for parametric heterogeneity in Cox proportional hazard model both of which can be formulated as C(α) tests. Some of these familiar tests are derived under very specific assumptions on the heterogeneity distribution F. As we have already noted, this is not necessary as long as some very mild moment conditions are satisfied. All the other three examples are multi-dimensional cases, as this is the area where we think the limit experiment and the C(α) test offers most interesting departures from existing work.

3.1. **Tests for overdispersion in Poisson Regression.**

Overdispersion tests for Poisson models constitute the most common example on test of parameter heterogeneity. Such a test was proposed in Fisher (1950) and also serves as the motivating example in Neyman and Scott (1966). We will consider two distinct versions of the test for unobserved heterogeneity in the conditional mean function of the Poisson regression model.

3.1.1. *Second Moment Test.*

Suppose we have $(Y_1, \ldots, Y_n)$ as i.i.d. random variables follow Poisson distribution with mean parameter λ_i. We further assume that

$$\lambda_i = \lambda_{0i} e^{\xi U_i} = \exp(x_i' \beta + \xi U_i)$$

where U_i are i.i.d. with distribution F, zero mean and unit variance. We have set τ to be 1 without loss of generality. The x_i's are covariates of the Poisson regression model including an intercept term. These covariates could be viewed as observed heterogeneity in the mean function, while U_i, since it is not explained by the covariates, is unobserved heterogeneity. Thus, the intercept coefficient, β_0, given the assumed form for λ_i, can be regarded as a random coefficient. We would like to test $H_0 : \xi = 0$ against $H_a : \xi \neq 0$ with β as the unspecified nuisance parameters. Since the first-order score with respect to ξ vanishes, this problem falls into the framework we considered in Section 2.

With some straightforward calculation and the nuisance parameters replaced by their MLEs, we find the $C(\alpha)$ test statistic as

$$Z_n = \frac{\sum_i [(y_i - \exp(x_i'\hat{\beta}))^2 - \exp(x_i'\hat{\beta})]}{\sqrt{2\sum_i \exp(2x_i'\hat{\beta})}}$$

We call this a second moment test because Z_n is essentially comparing the sample second moment with the second moment for the Poisson model under H_0. We reject H_0 when $(0 \vee Z_n)^2 > c_\alpha$ with c_α as the critical value from the mixture of χ^2.

Remark. The $C(\alpha)$ test constructed above is identical to the first test statistic proposed in Lee (1986) for overdispersion in Poisson regression models. In his derivation, Lee assumed that the Poisson mean parameter, λ_i, follows a Gamma distribution with certain mean-variance ratio. The Poisson-Gamma compound distribution then leads to a negative binomial model. As Lee noted (p.700), the same test statistic can also be derived under some other distribution in addition to the Gamma distribution (See also Dean and Lawless (1989)). From the $C(\alpha)$ perspective, the test statistic does not depend on the distribution of U, as long as the moment conditions are satisfied. However, the form of the test statistic does depend on the particular specification on λ_i as a function of the observed covariates and the unobservable U_i. This leads us to the next example.

3.1.2. *Second Factorial Moment Test.* If instead, under the same setup as we have in 3.1.1, we assume,

$$\lambda_i = \lambda_{0i}\left(1 + \xi U_i/\sqrt{\lambda_{0i}}\right)$$

The residual score for ξ is now found to be, with $\lambda_{0i} = \exp(x_i'\beta)$,

$$g(y_i, \beta) = \left[y_i(y_i - 1) - 2\lambda_{0i}(y_i - \lambda_{0i}) - \lambda_{0i}^2\right]/\lambda_{0i}$$

and $\mathbb{V}(g(Y_i, \beta)) = 2$. Replacing β by its restricted MLE $\hat{\beta}$, the locally optimal $C(\alpha)$ test is

$$Z_n = \frac{1}{\sqrt{2n}}\sum_i \left[y_i(y_i - 1) - \hat{\lambda}_{0i}^2\right]/\hat{\lambda}_{0i}$$

The test statistic Z_n is comparing the second sample factorial moment with that induced by the Poisson model under the null. Note that this test reduces to the second moment test if there are no covariates. Noticing again that only overdispersion is possible when deviating from the null, one-sided alternatives and the mixture of χ^2 asymptotics is employed.

3.2. **Joint test for slope heterogeneity in linear regression model.** We consider a linear cross sectional model,

$$y_i = x_i^\top \beta_i + u_i,$$

where β_i is a $p \times 1$ vector and $u_i \sim \text{IID}\mathcal{N}(0, \sigma^2)$. In addition, we assume $\beta_{ki} = \beta_{k0} + \xi_k U_{ki}$ for $k = 2, \ldots, p$. Without loss of generality, we impose $U_{ki} = U_i$ for all k and U_i has mean zero and unit variance. As discussed earlier, this implies we need to include all distinct

elements in the second order score matrix. Replacing nuisance parameters by their MLEs, it is easy to find the respective score for ξ and for the nuisance parameters $\theta = (\beta^\top, \sigma^2)^\top$:

$$\begin{aligned} S_{\xi,i} &= (\hat{u}_i^2/\hat{\sigma}^2 - 1)z_i/\hat{\sigma}^2 \\ S_{\sigma^2,i} &= (\hat{u}_i^2/\hat{\sigma}^2 - 1)/2\hat{\sigma}^2 \\ S_{\beta,i} &= \tfrac{\hat{u}_i}{\hat{\sigma}^2}x_i \end{aligned}$$

where $\hat{u}_i = y_i - x_i^\top \hat{\beta}$ and z_i is the vector of length $p(p-1)/2$ that consists distinct elements of $x_i x_i^\top$. The same testing problem is considered in the seminal paper by Breusch and Pagan (1979) who propose the LM test taking the form

$$LM = \frac{1}{2}(\sum_i z_i f_i)^\top (\sum_i z_i z_i^\top)^{-1}(\sum_i z_i f_i)$$

with $f_i = \hat{u}_i^2/\hat{\sigma}^2 - 1$. Under H_0, the LM statistic follows $\chi^2_{p(p-1)/2}$ asymptotically.

The C(α) test takes the same score function for ξ and θ, but pays explicit attention to the positivity constraints on those terms in $S_{\xi,i}$ that are inherited from the diagonal terms of $x_i x_i^\top$. We can easily find the residual score for ξ as

$$\tilde{S}_{\xi,n} = \frac{1}{\sqrt{n}}\sum_i (z_i - \bar{z})(\hat{u}_i^2/\hat{\sigma}^2 - 1)/\hat{\sigma}^2$$

and the associated Information matrix as $\Sigma = 2(\sum_i (z_i - \bar{z})(z_i - \bar{z})^\top)/N\hat{\sigma}^4$. Partition $\tilde{S}_{\xi,n}$ and Σ such that $\tilde{S}_{(1)}$ and $\Sigma_{(11)}$ correspond to the elements inherited from the diagonal elements of $x_i x_i^\top$ and proceed as in Theorem 4. In the simulation section we give a comparison between the C(α) test and the LM test which provides some encouraging evidence of power improvement.

3.3. **Joint test for location and scale heterogeneity in Gaussian panel data model.** In this example, we consider a two dimensional C(α) test for parameter heterogeneity in a Gaussian panel data model. The model is assumed to be

$$y_{it} = \mu_i + \sigma_i \epsilon_{it}$$

with $\epsilon_{it} \sim \text{IID}\mathcal{N}(0,1)$, $\mu_i = \mu_0 + \xi_1 U_{1i}$ and $\sigma_i^2 = \sigma_0^2 \exp(\xi_2 U_{2i}) \geqslant 0$. For convenience, we assume the random variables U_{ki} are i.i.d. with distribution F_k for $k = 1, 2$. Both U_1 and U_2 have zero mean and unit variance and are assumed to be independent for simplicity.

The unconditional density of observing $(y_{i1}, \ldots, y_{iT})$ is

$$f_i = \int\int \Big(\frac{1}{2\pi\sigma_0^2 \exp(\xi_2 u_{2i})}\Big)^{T/2} \exp\left(-\sum_{t=1}^{T} \frac{(y_{it} - \mu_0 - \xi_1 u_{1i})^2}{2\sigma_0^2 \exp(\xi_2 u_{2i})}\right) dF_1(u_{1i}) dF_2(u_{2i})$$

The respective score for (ξ_1, ξ_2) and the nuisance parameters (μ_0, σ_0^2) are

$$\begin{aligned}
\nu_{1i} &= \nabla^2_{\xi_1} \log f_i|_{\xi_1=\xi_2=0} &&= (\tfrac{\bar{y}_{i\cdot}-\mu_0}{\sigma_0^2/T})^2 - \tfrac{1}{\sigma_0^2/T} \\
\nu_{2i} &= \nabla^2_{\xi_2} \log f_i|_{\xi_1=\xi_2=0} &&= (Z_i - \tfrac{T}{2})^2 - Z_i \\
\nu_{3i} &= \nabla_{\mu_0} \log f_i|_{\xi_1=\xi_2=0} &&= \tfrac{\bar{y}_{i\cdot}-\mu_0}{\sigma_0^2/T} \\
\nu_{4i} &= \nabla_{\sigma_0^2} \log f_i|_{\xi_1=\xi_2=0} &&= (Z_i - \tfrac{T}{2})/\sigma_0^2
\end{aligned}$$

where $\bar{y}_{i\cdot}$ is the sample mean defined as $\sum_{t=1}^T y_{it}/T$ and $2Z_i = \sum_{t=1}^T (y_{it}-\mu_0)^2/\sigma_0^2 \sim \chi^2_T$.

Replacing the nuisance parameters by their MLEs, the optimal $C(\alpha)$ test for $H_0 : \xi_1 = \xi_2 = 0$ against $H_a : \xi_i \neq 0$ for at least one i is:

$$T_n = (0 \vee t_{1n})^2 + (0 \vee t_{2n})^2$$

with

$$\begin{aligned}
t_{1n} &= (2NT(T-1)/\hat{\sigma}_0^4)^{-1/2} \left(\textstyle\sum_i (\frac{\bar{y}_{i\cdot}-\hat{\mu}_0}{\hat{\sigma}_0^2/T})^2 - \frac{NT}{\hat{\sigma}_0^2} \right) \\
t_{2n} &= (NT(T/2+1))^{-1/2} \left(\textstyle\sum_i (Z_i - T/2)^2 - \frac{NT}{2} \right)
\end{aligned}$$

We reject H_0 for $T_n > c_\alpha$ where c_α is the $(1-\alpha)$-quantile of $\frac{1}{4}\chi^2_0 + \frac{1}{2}\chi^2_1 + \frac{1}{4}\chi^2_2$.

Remark. The first component t_{1n} of the test statistics may be recognized again as the test for individual effect in Gaussian panel data model proposed by Breusch and Pagan (1980). The second component t_{2n} is equivalent to a single parameter $C(\alpha)$ test for a Gamma model with heterogenous scale parameter. (Analytical derivation details appear in the Appendix B.) The factorization provided by the Gaussian model leads to simple asymptotics of the test statistics. Introducing dependence between the random effects u_1 and u_2 will add an extra score function which is the cross term in the second order score matrix, $\nabla^2_{\xi_1\xi_2} \log f_i$. In this case, we proceed as in Theorem 4. Notice the above test is valid for the large N fixed T setting, and the local alternative for ξ_n is of order $N^{-1/4}$. If T also tends to infinity, then the local alternative for ξ_n is of order $N^{-1/4}T^{-1/2}$.

3.4. **Test for slope heterogeneity in large panels.** Example 3.3 above tests for randomness in individual location and variances. Perhaps a more realistic application is to allow for individual effects and the group-wise heteroscedasticity in the error but test for randomness in the slope coefficients. This problem has been considered in Swamy (1970) and is recently revived in Pesaran and Yamagata (2008) (hereafter PY). The PY test is a standardized version of Swamy (1970) under large N large T setting. The model is assumed to be,

$$y_{it} = \alpha_i + x_{it}^\top \beta_i + \epsilon_{it},$$

with β_i being a $p \times 1$ vector. The null hypothesis of interest is $H_0 : \beta_i = \beta$ for all i against $H_1 : \beta_i \neq \beta_j$ for at least one pair of $i \neq j$. The PY test is

$$\tilde{\Delta}^{PY} = \sqrt{\frac{N(T+1)}{T-k-1}} \left(\frac{N^{-1}\tilde{S} - k}{\sqrt{2k}} \right)$$

with M_τ being the familiar demean matrix and $\tilde{S} = \sum_i (\hat{\beta}_i - \hat{\beta}_{WFE})^\top X_i^\top M_\tau X_i (\hat{\beta}_i - \hat{\beta}_{WFE})/\tilde{\sigma}_i^2$ where $\hat{\beta}_i$ is the within estimator for each individual regression and $\hat{\beta}_{WFE}$ is

the proper pooled estimator that accounts for individual specific variance $\tilde{\sigma}_i^2$. Su and Chen (2013) gives an LM test interpretation for $\tilde{S}$ in the PY test that,

$$\tilde{S} = \sum_i \hat{\epsilon}_i^\top M_\tau X_i (X_i^\top M_\tau X_i)^{-1} X_i^\top M_\tau \hat{\epsilon}_i / \tilde{\sigma}_i^2$$

with $\hat{\epsilon}_{it} = M_\tau(y_{it} - x_{it}^\top \hat{\beta}_{WFE})$. This is the LM test statistic for considering the regression $\hat{\epsilon}_{it} = \alpha_i + (x_{it} - \bar{x}_i)^\top \phi_i + \eta_{it}$ and test for $\phi_i = 0$ for all i. As both N and T goes to infinity, with proper re-centering and standardization, the resulting PY test has a standard normal asymptotics under H_0 and the authors recommend a two-sided test for inference.

In the large N large T setting, we can also construct the C(α) score test for heterogeneity in coefficients. Assuming again $\beta_{ki} = \beta_{k0} + \xi_k U_{ki}$ for $k = 1, \ldots, p$. The score function for ξ, $S_{\xi,n}$, is the distinct $p(p+1)/2$ elements of the second-order score matrix, which takes the form $\frac{1}{\sqrt{N}} \sum_i (X_i^\top M_\tau \hat{\epsilon}_i \hat{\epsilon}_i^\top M_\tau X_i / \hat{\sigma}_i^4 - X_i^\top M_\tau X_i / \hat{\sigma}_i^2)$ with nuisance parameters replaced by MLEs. The elements of $S_{\xi,n}$ are asymptotically jointly normal with mean zero and covariance matrix Σ under H_0 and by LeCam's third lemma, they jointly follow $\mathcal{N}(\Sigma t, \Sigma)$ under the local alternative $(\xi_{j,n} = \xi_j + \delta_j N^{-1/4} T^{-1/2}, j = 1, \ldots, p)$ with $t = (\delta_1^2, \ldots, \delta_p^2, (2\delta_j \delta_k)_{j \neq k})^\top$ as discussed in Section 2.5. Not surprisingly, given the connection to the score test shown by Chesher (1984), this shares considerable similarity to the White (1982) Information Matrix test. However, the IM test rejects H_0 if $S_{\xi,n} \Sigma^{-1} S_{\xi,n}$ exceeds the critical value from $\chi^2_{p(p+1)/2}$ at nominal level α, while the C(α) test modifies the IM test by adjusting for positivity constraints in t for the respective elements in the score function. We do not repeat the steps here in applying Theorem 4. In the simulation section, we compare the C(α), the IM test and the PY test and the results show that the C(α) test enjoys a power gain compared to the other two tests. It is also worth mentioning that the advantage of the C(α) test is that we only need to estimate under the null model. In addition, the test can be derived in the same way for large N and fixed T setting, except the local alternative for ξ_n is then of order $N^{-1/4}$.

4. Reparameterization and connection to the Information matrix Test

4.1. **Reparameterization.** A common strategy in prior literature to circumvent the irregularity, that the first-order score function is degenerately zero, is to reparameterize the model. In fact, this is the advice given in the original Neyman (1959) C(α) paper (Section 9, p. 225) and also in Cox and Hinkley (1974, p. 117-118). For the heterogeneity tests considered in this paper in particular, Cox (1983) and Chesher (1984) adopt such a reparameterization by letting $\eta = \xi_0 + (\xi - \xi_0)^2$. Reconsidering the example in Section 2.2, without loss of generality, we set $\xi_0 = 0$ and have the density function as $p(x; \lambda_0 + \tau\sqrt{\eta} U_i)$. Cox (1983) tests for heterogeneity of λ_i by testing $H_0 : \eta = 0$ against $H_1 : \eta > 0$. Under H_0, this is to test whether $\mathrm{Var}(\lambda) = 0$. Chesher (1984) takes the same model assuming U_i follows a symmetric location-scale distribution. A more recent treatment, focusing on random individual effects in panel data models by Bennala, Hallin, and Paindaveine (2012) also uses the same reparametrization but adopts a less stringent LeCam framework.

At first sight, reparameterization avoids the irregularity of having a degenerate score function. The first order derivative with respect to η, albeit an undefined $\frac{0}{0}$ function, can be evaluated by the l'Hôpital's rule. As long as $\mathbb{E}(U^2)$ is non-zero, the score function is nonvanishing. The score function thus derived also involves the second derivative and is identical to the score function in the $C(\alpha)$ test using the original parameterization that $\lambda_i = \lambda_0 + \tau\xi U_i$. However, the second order derivative for η is unbounded unless we impose an additional moment condition on U, that $\mathbb{E}(U^3) = 0$ (See the derivation in the Appendix C). This condition is implicitly satisfied in Chesher (1984) because of the symmetry distribution assumption on U. Moran (1973) also employed this zero third moment condition and remarked that it was hard to rationalize. One explanation for this extra condition is that the original, more natural specification on the random parameter $\lambda_i = \lambda_0 + \tau\xi U_i$ with $\xi \in \mathbb{R}$ is not equivalent to the reparameterization $\lambda_i = \lambda_0 + \tau\sqrt{\eta}U_i$ with $\eta \in \mathbb{R}_+$ unless U has a symmetric distribution. When symmetry does not hold for the distribution of U, the likelihood does not obtain a proper expansion around η. As we have seen, the ξ parameterization has the advantage that no symmetry or higher moment conditions on U are necessary.

4.2. **Connection to the Information Matrix test.** Chesher (1984) was the first to point out that White's (1982) Information Matrix (IM) test is a score test for unobserved heterogeneity. Since Chesher (1984) can be viewed as a reparameterized $C(\alpha)$ test, it is of interest to investigate the connection between the $C(\alpha)$ test for heterogeneity in general and the IM test. We show that the $C(\alpha)$ test for heterogeneity nests the IM test as a special case.

Take again the example in Section 2.2, $Y_1, \dots Y_n$ are i.i.d. random variables each with density function $p(y; \lambda_i)$. The parameter λ_i is a random parameter and we assume it now takes a more general form $\lambda_i = \lambda_0 + \xi k(\lambda_0)U_i$ to incorporate both additive and multiplicative specifications. For example, if $k(\lambda_0) = 1$, we have the additive form $\lambda_i = \lambda_0 + \xi U_i$, while if $k(\lambda_0) = \lambda_0$, then the multiplicative form. The function $k(\lambda_0)$ thus allows flexible specification for the random parameter.

For simplicity and to fix ideas, we first assume λ_0 is known. Theorem 1 then implies the following expansion of the log-likelihood function, provided that $\xi_n = O(n^{-1/4})$,

$$l = \sum_i \log \int p(y_i; \lambda_i) dF(u) = \sum_i \log p(y_i; \lambda_0) + \frac{1}{2}\xi_n^2 \mathbb{E}(U_i^2) \sum_i k(\lambda_0)^2 \frac{\nabla_\lambda^2 p(y_i; \lambda_0)}{p(y_i; \lambda_0)} + O_p(1)$$

The first order derivative of l with respect to ξ_n is zero evaluated under $\xi_n = 0$, and the second-order score is

$$\frac{\partial^2}{\partial \xi_n^2} l|_{\xi_n = 0} = \sum_i k(\lambda_0)^2 \frac{\nabla_\lambda^2 p(y_i; \lambda_0)}{p(y_i; \lambda_0)}.$$

If λ_0 is unknown, we find the corresponding score for λ_0 and take the projection step to get the $C(\alpha)$ test. This is very close to the approximation in Cox (1983) except we allow for a more flexible variance function for the random parameter λ_i, as $\xi^2 \mathbb{E}(U_i^2) k(\lambda_0)^2$. In a regression model with covariates, λ_0 will then be a function of the covariates with coefficients β.

White's (1982) Information Matrix test under regression setting, on the other hand, is constructed based on the following moment conditions:

$$\mathbb{E}\left[\mathcal{vech}\left(\nabla_\beta^2 \log p(y;\lambda_0(x_i,\beta)) + \nabla_\beta \log p(y;\lambda_0(x_i,\beta))\nabla_\beta^\top \log p(y;\lambda_0(x_i,\beta))\right)\right] = 0$$

where $\mathcal{vech}$ is the operator which stacks the elements in the lower triangular part of a symmetric matrix. Using the chain rule, we see that the IM test statistic uses the following sample analogue of the moment condition

$$IM = \sum_i \Big[\frac{\nabla_\lambda^2 p(y;\lambda_0(x_i,\beta))}{p(y;\lambda_0(x_i,\beta))}\nabla_\beta\lambda_0(x_i,\beta)\nabla_\beta^\top\lambda_0(x_i,\beta) + \frac{\nabla_\lambda p(y;\lambda_0(x_i,\beta))}{p(y;\lambda_0(x_i,\beta))}\nabla_\beta^2\lambda_0(x_i,\beta)\Big]$$

There are various forms for the IM test in the literature (see Davidson and MacKinnon (1998)), we focus on the efficient score version, in which all the nuisance parameters are replaced by their restricted MLEs. For the $C(\alpha)$ test to be equivalent to the efficient score version of the IM test, it is sufficient to have the following two identities:

$$C\nabla_\beta\lambda_0(x_i,\beta)\nabla_\beta^\top\lambda_0(x_i,\beta) = k(\lambda_0)k(\lambda_0)^\top$$
$$\textstyle\sum_i \frac{\nabla_\lambda p(y;\lambda_0(x_i,\beta))}{p(y;\lambda_0(x_i,\beta))}\nabla_\beta^2\lambda_0(x_i,\beta) = 0$$

where C is a non-zero constant. We give several examples below as illustrations.

Example 4.1. *Normal regression with* $Y_i \sim \mathcal{N}(\mu_i, 1)$*, where* $\mu_i = \mu_{0i} + \xi k(\mu_{0i})U_i$ *and* $\mu_{0i} = x_i'\beta$*.*

Note that $\nabla_\beta\mu_{0i}\nabla_\beta^\top\mu_{0i} = x_i x_i^\top$ and $\nabla_\beta^2\mu_{0i} = 0$. Considering only the IM test based on the intercept term, it is equivalent to the $C(\alpha)$ test for heterogeneity in β_0 if $k(\mu_{0i}) = C \neq 0$. If considering all elements in the IM test, the equivalence holds if $x_i x_i^\top = k(\mu_{0i})k(\mu_{0i})^\top$. In this case, the $C(\alpha)$ test is multivariate, testing for homogeneity for all coefficients β in μ_{0i}.

Example 4.2. *Poisson regression with* $Y_i \sim \mathsf{Poi}(\lambda_i)$*, where* $\lambda_i = \lambda_{0i} + \xi k(\lambda_{0i})U_i$ *and* $\lambda_{0i} = \exp(x_i'\beta)$*.*

Considering only the IM test for the intercept term, we have $\nabla_{\beta_0}\lambda_{0i} = \nabla_{\beta_0}^2\lambda_{0i} = \lambda_{0i}$. If β's are replaced by their MLEs, the second identity for equivalence holds because the normal equation for the MLE of β_0 gives

$$\sum_i \frac{\nabla_\lambda p(y;\lambda_{0i})}{p(y;\lambda_{0i})}\nabla_{\beta_0}^2\lambda_{0i} = \sum_i \frac{\nabla_\lambda p(y;\lambda_{0i})}{p(y;\lambda_{0i})}\nabla_{\beta_0}\lambda_{0i} = 0$$

Therefore, the IM test is equivalent to the $C(\alpha)$ test if $k(\lambda_{0i}) = \lambda_{0i}$ which is satisfied for the multiplicative alternative $\lambda_i = \lambda_{0i}(1 + \xi U_i)$. This specification is a first order linear approximation of the alternative form $\lambda_i = \lambda_{0i}\exp(\xi U_i)$ for small ξ, which leads to the second moment test for the Poisson regression model as discussed in Section 3.1.1. There are of course many other possible specifications for the conditional mean function of λ_{0i} which would lead to other equivalence conditions. We do not delve into further details here,

but refer the readers to Cameron and Trivedi (1998, Chapter 5) and Dean (1992) for more elaborated discussions on count data models.

In summary, when the model contains covariates, the functional form of the $C(\alpha)$ test is equivalent to the IM test only under a particular alternative specification, provided that the nuisance parameters are also replaced by their corresponding restricted MLEs. When the model does not contain covariates, the IM test will always be equivalent to the $C(\alpha)$ test because the function $k(\lambda_0)$ is no longer individual specific and can be factored out as a constant from the score function. It will then be cancelled when we rescale the score by its standard deviation to form the $C(\alpha)$ test statistic. To see more clearly how different specification affect the power performance of various testing procedures discussed here, especially in cases where the IM test no longer serves as an optimal test, we conduct a Monte Carlo simulation in the next section. It is also important to deviate from the common practice in using the χ^2 asymptotics for the IM test or the LM test for heterogeneity. The simulation shows that overlooking the intrinsic one-sidedness of alternatives sacrifices power.

5. Simulation

We first revisit the Poisson regression model to illustrate the points made in Section 4.2. As discussed in Example 4.2 and also in Section 3.1, when $k(\lambda_{0i})$ takes different functional forms, one finds different optimal test statistics. For two different data generation processes, we compare three testing procedures: the second moment test and the second factorial moment test, both are one-sided tests and use critical value from a mixture of χ^2 and the information matrix test, using critical values from the χ^2 distribution. The first experiment generates data from a Poisson regression model with the conditional moment function as $\lambda_i = \lambda_{0i} + \tau\xi\lambda_{0i}u_i$ and the second with $\lambda_i = \lambda_{0i} + \tau\xi\sqrt{\lambda_{0i}}u_i$. In both cases $\lambda_{0i} = exp(\beta_0 + \beta_1 x_i)$ and $\tau\xi u_i$ has a mixture distribution taking value $1.5h$ with probability $2/3$ and $-3h$ with probability $1/3$. We consider 21 distinct values of h equally spaced and the design of X is fixed for all experiments as a sample drawn from a standard normal distribution. Using other X designs does not change the conclusions. The sample size for all power comparison is fixed at 500 with 10000 replications.

In the left panel of Figure 1, both the second moment test and the information matrix test performs uniformly better than the second factorial moment test. This is to be expected since the second moment test is the optimal test derived using the $C(\alpha)$ framework. The IM test using just the element for the intercept term has an identical test function as the second moment test, but using the one-sided test with mixture of χ^2 critical value gives better power, especially for the 10% level case. On the other hand, the second factorial moment test is superior under the second experiment, although the power for the second moment test and the IM test also converges to unity albeit much more slowly. It is documented in the literature that the IM test has poor size in small samples (Chesher and Spady (1991)) and the $C(\alpha)$ test may be subject to similar criticism. We use size-corrected critical values as suggested by Horowitz (1994).

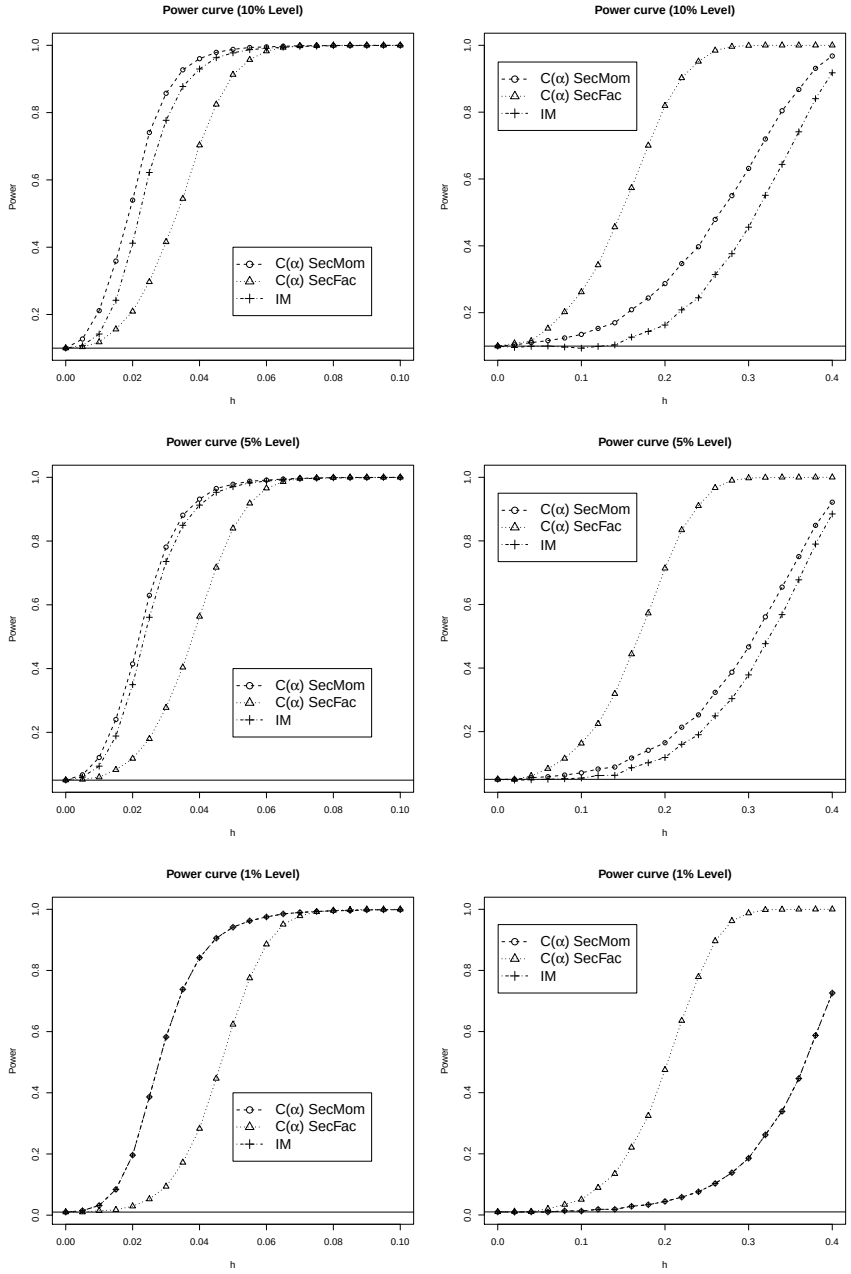


FIGURE 1. Power Comparison of Unobserved Heterogeneity Test for Poisson Regression Model: The left panel corresponds to the first experiment and the right panel to the second. The dotted curve corresponds to the power curve of the second moment test, the curve with triangle signs for the second factorial moment and the crossed curve for the IM test of the intercept term.

We then conduct a power comparison between the $C(\alpha)$ test and the Breusch and Pagan (1979) LM test for Example 3.2. Two different alternative β_i distributions are considered. The first one assumes β_i takes value 0 for $i = 1, \ldots, N/2$ and $cN^{-1/4}$ for $i = N/2+1, \ldots, N$. We let c take 51 distinct values equally spaced from 0 to $\sqrt{50}$. The second case assumes $\beta_i \sim \mathcal{N}(0, \sigma^2)$ with σ taking 21 distinct values from 0 to 1. For simplicity, we consider the case with dimension two, where both x covariates are standard normal variables. The sample size is fixed at 500 with 10000 replications. Figure 2 presents the power curve for the 5 % nominal level. The first experiment has a slightly bigger power gain compared to the second, but in both cases, the $C(\alpha)$ test dominate the power curve of the LM test based on the usual χ^2 asymptotics.

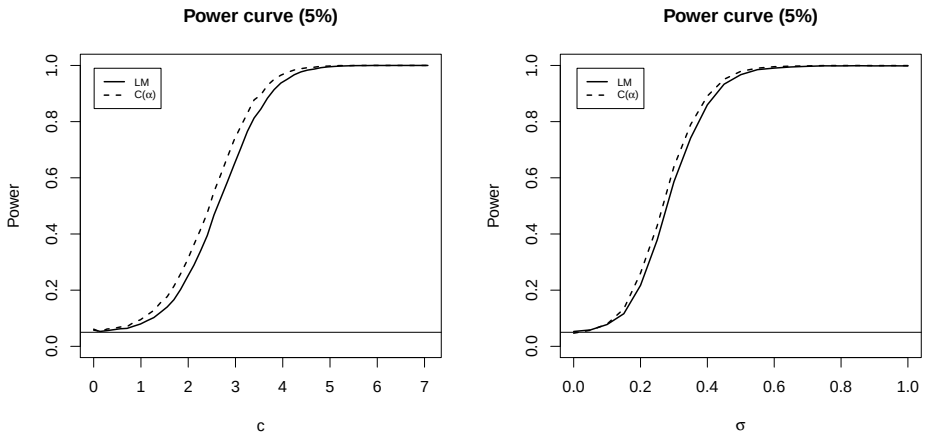


FIGURE 2. Power Comparison of Slope Heterogeneity Test for Linear Regression Model: The left figure corresponds to the first experiment and the right to the second. The dotted curve corresponds to the power curve of the $C(\alpha)$ test based on the mixture of χ^2 asymptotics and the solid curve for the LM test based on the χ^2_3 asymptotics.

Lastly, we compare the $C(\alpha)$ test with the Pesaran and Yamagata (2008) test and the Information Matrix test for a Gaussian panel data model. As shown by Pesaran and Yamagata (2008) Table 1, their standardized Swamy test has very nice size and power performance compared to some other existing tests, i.e. the Hausman test and the original Swamy (1970) test for a wide range of N and T. We consider a panel data model with two exogenous regressors and normal errors with individual variances. Table 1 reports the size and power for the $C(\alpha)$ test, the PY test and the Information matrix test. Both the PY test and the $C(\alpha)$ test has correct size and the IM test is slightly on the conservative side. For all N and T combinations, the $C(\alpha)$ test has a significant power gain.

For a further power comparison, we consider with the same model as above for two different β_i distributions. For 21 distinct equally spaced values of $h \in [0, 1/3]$, the first example

assumes $\beta_i \sim \mathcal{N}(0, h^2)$ and the second assumes β_i taking two possible values $\{1-h, 1+2h\}$ with probability $(2/3, 1/3)$. Results are presented in Figure 3. The sample size is fixed for each experiment at $T = 50$ and $N = 100$ with 5000 replications. The $C(\alpha)$ test again exhibits encouraging power improvement compared to the other two tests uniformly for all h values.

		Size			Power		
T	N	PY	$C(\alpha)$	IM	PY	$C(\alpha)$	IM
20	30	0.053	0.040	0.032	0.064	0.072	0.055
30	30	0.049	0.046	0.034	0.082	0.109	0.095
50	30	0.046	0.054	0.041	0.137	0.215	0.179
100	30	0.045	0.048	0.042	0.434	0.604	0.538
20	50	0.045	0.042	0.037	0.070	0.070	0.063
30	50	0.046	0.038	0.033	0.112	0.154	0.125
50	50	0.045	0.052	0.041	0.263	0.379	0.325
100	50	0.047	0.052	0.044	0.625	0.738	0.721
20	100	0.046	0.040	0.034	0.090	0.099	0.084
30	100	0.051	0.039	0.040	0.172	0.224	0.195
50	100	0.046	0.045	0.047	0.406	0.570	0.484
100	100	0.044	0.042	0.046	0.886	0.946	0.945
20	200	0.049	0.035	0.033	0.151	0.171	0.140
30	200	0.045	0.044	0.036	0.317	0.395	0.336
50	200	0.045	0.047	0.041	0.672	0.800	0.773
100	200	0.048	0.048	0.043	0.993	0.998	0.999

TABLE 1. Size and Power comparison between the PY test, the Information Matrix test and the $C(\alpha)$ test for different N and T. Data are generated as $y_{it} = \alpha_i + x_{it}^\top \beta_i + \epsilon_{it}$ with $\alpha_i \sim U(0,1)$ and $\epsilon_{it} \sim \text{IID}\mathcal{N}(0, \sigma_i^2)$ and $\sigma_i^2 \sim U(1,2)$. Both regressors are $\mathcal{N}(0, 0.5^2)$. Under the null, $\beta_{1i} = \beta_{2i} = 1$ for all i and under the alternative, $\beta_{1i} = \beta_{2i} \sim \mathcal{N}(1, 0.15^2)$. The PY test is based on a two sized $\mathcal{N}(0,1)$ test, the IM test is based on χ_3^2 test and the $C(\alpha)$ test on a mixture of χ^2 test. All tests are conducted at 5% nominal level with 5000 replications.

6. Conclusion

We have shown that Neyman's $C(\alpha)$ test provides a unified approach to testing for neglected heterogeneity in parametric models. The irregularity encountered in these testing problems, that the score function is identically zero, can be circumvented by defining a second-order score function. Optimality of this new score function is established by formulating the problem in LeCam's LAN framework and examining the associated limit experiment. This framework provides neater regularity conditions in the irregular problem as compared to classical approach in Neyman (1959). The multi-dimensional extension

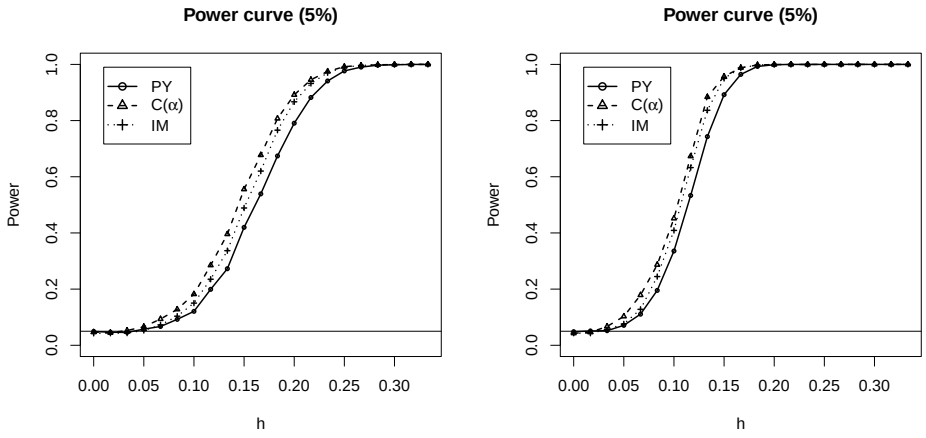


FIGURE 3. Power Comparison of Slope Heterogeneity Test for Gaussian Panel Data Model: The left figure corresponds to the first experiment and the right to the second for different values of h. Data are generated as $y_{it} = \alpha_i + x_{it}^\top \beta_i + \epsilon_{it}$ with $\alpha_i \sim U(0,1)$ and $\epsilon_{it} \sim \text{IID}\mathcal{N}(0, \sigma_i^2)$ and $\sigma_i^2 \sim U(1,2)$. Both regressors are normal variable with mean zero and standard deviation 0.5. The solid line with circles is the power curve for the PY test, the crossed curve for the Information Matrix test with χ^2 asymptotics and the curve with triangle signs for the $C(\alpha)$ test with mixture of χ^2 asymptotics.

suggests a modification on the usual χ^2 test that leads to power improvement in many applications.

The $C(\alpha)$ test inherits the chief merit of the score test, computation is made easy under the null model. In contrast, the likelihood ratio test, in face of the generally unknown heterogeneity distribution F, is computationally challenging. We have also seen that the $C(\alpha)$ test has local power against a wide class of alternatives, that allows us to avoid strict parametric assumptions on F, relying instead on weaker moment conditions. A further advantage of the LeCam framework is that it enables us to dispense with symmetry and higher order moment conditions that have been employed in earlier work.

A straightforward generalization of the theorems in Section 2 would be to incorporate density functions that allow the first $(m-1)$ logarithmic derivatives to vanish. Rotnitzky, Cox, Bottai, and Robins (2000) also discuss estimation problems in this general case under classical MLE type of conditions. In such cases, we can define the m^{th} order derivative of the log density as the score function and require the Pitman-type local alternative to be of order $n^{-1/2m}$. LeCam's DQM condition needs to be modified by raising the corresponding elements in the expansion to m^{th} power, as we did for $m = 2$ in Definition 1. It is curious to observe that only when m is an even integer is the test required to be one-sided and reparameterization is not advisable. When m is odd, we can use reparameterization to

transform the irregular problem back to a regular case, without imposing additional restrictions (i.e. symmetry of the distribution F).

A drawback of the $C(\alpha)$ test, as reflected in Neyman (1979), is that asymptotic optimality of the test is only established under local alternatives. The approximation of the power function, which is characterized by the asymptotic behavior of the test statistics under such alternatives, relies on n tending to infinity and the parameter ξ_n converging to the null value ξ_0. The behavior of the power function for finite samples or fixed alternatives is largely unknown. Some finite sample correction like those pursued in Honda (1988) and Chesher and Spady (1991) is left for future work.

References

Abelson, R., and J. Tukey (1963): "Efficient Utilization of Non-numerical Information in Quatitative Analysis: General Theory and the Case of Simple Order," *Annals of Mathematical Statistics*, 34, 1347–1369.

Akharif, A., and M. Hallin (2003): "Efficient Detection of Random Coefficients in Autoregressive Models," *Annals of Statistics*, 31(2), 675–704.

Andrews, D. (1994): "Empirical Process Methods in Econometrics," in *Handbook of Econometrics, Volume 4*, ed. by R. Engle, and D. L. McFadden. Elsevier.

Bartholomew, D. (1961): "A Test of Homogeneity of Means under Restricted Alternatives," *Journal of the Royal Statistical Society, Series B*, 23, 239–281.

Bennala, N., M. Hallin, and D. Paindaveine (2012): "Pseudo-Gaussian and Rank-based Optimal Tests for Random Individual Effects in Large n Small T Panels," *Journal of Econometrics*, 170, 50–67.

Bickel, P., C. Klaassen, Y. Ritov, and J. Wellner (1993): *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press: Baltimore and London.

Breusch, T., and A. Pagan (1979): "A Simple Test for Heteroscedasticity and Random Coefficient Variation," *Econometrica*, 47(5), 1287–1294.

——— (1980): "The Lagrange Multiplier Test and its Applications to Model Specification in Econometrics," *Review of Economic Studies*, 47, 239–253.

Bühler, W., and P. Puri (1966): "On Optimal Asymptotic Tests of Composite Hypotheses with Several Constraints," *Z. Wahrscheinlichkeittstheorie verw. Geb.*, 5, 71–88.

Cameron, A., and P. Trivedi (1998): *Regression Analysis of Count Data*. Cambridge University Press.

Chen, H., J. Chen, and J. Kalbfleisch (2001): "A Modified Likelihood Ratio Test for Homogeneity in Finite Mixture Models," *Journal of the Royal Statistical Society,Series B*, 63(1), 19–29.

Chernoff, H. (1954): "On the Distribution of the Likelihood Ratio," *The Annals of Mathematical Statistics*, 25(3), 573–578.

Chesher, A. (1984): "Testing for Neglected Heterogeneity," *Econometrica*, 52(4), 865–872.

Chesher, A., and R. Spady (1991): "Asymptotic Expansions of the Information Matrix Test Statistic," *Econometrica*, 59(3), 787–815.

Cox, D. (1983): "Some Remarks on Overdispersion," *Biometrika*, 70(1), 269–274.

Cox, D., and D. Hinkley (1974): *Theoretical Statistics*. Chapman and Hall: London.

Cramér, H. (1946): *Mathematical Methods of Statistics*. Princeton University Press: Princeton, New Jersey.

Davidson, R., and J. MacKinnon (1998): "Graphical Methods for Investigating the Size and Power of Hypothesis Tests," *The Manchester School*, 66(1), 1–26.

Dean, C. (1992): "Testing for Overdispersion in Poisson and Binomial Regression Models," *Journal of the American Statistical Association*, 87, 451–457.

Dean, C., and J. Lawless (1989): "Tests for Detecting Overdispersion in Poisson Regression Models," *Journal of the American Statistical Association*, 84, 467–472.

Fisher, R. (1950): "The Significance of Deviations from Expectation in a Poisson Series," *Biometrika*, 6, 17–24.

Gu, J., R. Koenker, and S. Volgushev (2013): "Testing for Homogeneity in Mixture Models," arXiv: 1302.1805[stat.ME].

Hallin, M., and C. Ley (2013): "Skew-Symmetric Distributions and Fisher Information: The Double Sin of the Skew-Normal," *Bernoulli*, forthcoming.

Hillier, G. (1986): "Joint Tests for Zero Restrictions on Nonnegative Regression Coefficients," *Biometrika*, 73(3), 657–669.

Honda, Y. (1988): "A Size Correction To the Lagrange Multiplier Test for Heteroscedasticity," *Journal of Econometrics*, 38, 375–386.

Horowitz, J. (1994): "Bootstrap-based Critical Values for the Information Matrix Test," *Journal of Econometrics*, 61, 395–411.

Kiefer, N. (1984): "A Simple Test for Heterogeneity in Exponential Models of Duration," *Journal of Labor Economics*, 2(4), 539–549.

Lancaster, T. (1985): "Generalized Residuals and Heterogenous Duration Models with Applications to the Weibull Model," *Journal of Econometrics*, 28, 155–169.

LeCam, L. (1972): "Limits of Experiments," in *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Vol. I.*, pp. 245–261. University of California Press: Berkeley and Los Angeles.

Lee, L. (1986): "Specification Test for Poisson Regression Models," *International Economic Review*, 27(3), 689–706.

Lee, L., and A. Chesher (1986): "Specification testing when score test statistics are identically zero," *Journal of Econometrics*, 31, 121–149.

Lindsay, B. G. (1995): *Mixture Models: Theory, Geometry and Applications*. IMS, Hayward, California.

Moran, P. (1973): "Asymptotic Properties of Homogeneity Tests," *Biometrika*, 60(1), 79–85.

Neyman, J. (1959): "Optimal Asymptotic Tests of Composite Statistical Hypotheses," in *Probability and Statistics, the Harald Cramer Volume*, ed. by U. Grenander. Wiley: New York.

——— (1979): "C(α) Tests and Their Use," *Sankhyā: The Indian Journal of Statistics*, 41, 1–21.

Neyman, J., and E. Scott (1966): "On the Use of C(α) Optimal Tests of Composite Hypotheses," *Bull. Inst. Int. Statist.*, 41(1), 477–497.

Nüesch, P. (1966): "On the Problem of Testing Location in Multivariate Problems for Restricted Alternatives," *Annals of Mathematical Statistics*, 37, 113–119.

Perlman, M. (1969): "One-sided Testing Problems in Multivariate Analysis," *Annals of Mathematical Statistics*, 40, 549–567.

Pesaran, M., and Y. Yamagata (2008): "Testing Slope Homogeneity in Large Panels," *Journal of Econometrics*, 142, 50–93.

Pollard, D. (1997): "Another Look at Differentiability in Quadratic Mean," in *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, ed. by D. Pollard, E. Torgersen, and G. Yang. Springer-Verlag: New York.

Rao, C. (1948): "Large-Sample Test of Statistical Hypotheses Concerning Several Parameters with Applications to Problems of Estimation," *Proceedings of the Cambridge Philosophical Society*, 44, 50–57.

Rotnitzky, A., D. Cox, M. Bottai, and J. Robins (2000): "Likelihood-based Inference with Singular Information Matrix," *Bernoulli*, 6(2), 243–284.

Schaafsma, W., and L. Smid (1966): "Most Stringent Somewhere Most Powerful Tests Against Alternatives Restricted by a Number of Linear Alternatives," *Annals of Mathematical Statistics*, 37, 1161–1172.

Self, S., and K.-Y. Liang (1987): "Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests under Nonstandard Conditions," *Journal of the American Statistical Association*, 82(398), 605–610.

Silvapulle, M., and P. Silvapulle (1995): "A Score Test Against One-Sided Alternatives," *Journal of the American Statistical Association*, 90, 342–349.

Su, L., and Q. Chen (2013): "Testing Homogeneity in Panel Data Models with Interactive Fixed Effects," *Econometric Theory*, 29, 1079–1135.

Swamy, P. (1970): "Efficient Inference in a Random Coefficient Regression Model," *Econometrica*, 38, 311–323.

Turlach, B., and A. Weingessel (2013): "Functions to solve Quadratic Programming Problems," R package version 1.5-5, 2013-04-17.

van der Vaart, A. (1998): *Asymptotic Statistics*. Cambridge University Press.

van der Vaart, A. W., and J. A. Wellner (1996): *Weak Convergence and Empirical Processes - Springer Series in Statistics*. Springer: New York.

White, H. (1982): "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1–25.

Appendix A. Proof of theorems

Before proceeding to the proof for **Theorem 1**, we first prove the following lemma as an adaption to Pollard (1997, Lemma 1). Denote $f_n = \sqrt{p(x_i;\xi_n,\theta_n)}$ and $f_0 = \sqrt{p(x_i;\xi_0,\theta)}$. Let ν_ξ and ν_θ be shorthand for $\nu_\xi(x_i)$ and $\nu_\theta(x_i)$. Let $\|\cdot\|$ be $\mathcal{L}_2(\mu)$-norm and $\langle\cdot,\cdot\rangle$ be the inner product. If it contains a vector, then it is defined as the vector of inner product for each elements. Further, let $r_n(x_i,\xi_n,\theta_n) = f_n - f_0 - h_n^\top \nu(x_i)$ and denote $R_i = r_n(x_i,\xi_n,\theta_n)/f_0$.

Lemma 1. Under Assumption 1 and the modified DQM condition, we have the following:

(1) $\sum_i R_i^2 = o_p(1)$
(2) $\mathbb{E}(\nu(X)/f_0) = 0$
(3) $2\sum_i R_i = -\frac{1}{4}t^\top J t + o_P(1)$
(4) $n^{-1/2}\sum_i R_i \nu_\xi/f_0 = o_P(1)$, $n^{-1/2}\sum_i R_i \nu_\theta/f_0 = o_P(1)$
(5) $\max_{1\leqslant i\leqslant n} |R_i| = o_P(1)$
(6) $\max_{1\leqslant i\leqslant n} |\frac{2}{\sqrt{n}}\frac{\nu_\xi}{f_0}| = o_P(1)$, $\max_{1\leqslant i\leqslant n} |\frac{2}{\sqrt{n}}\frac{\nu_\theta}{f_0}| = o_P(1)$

Proof of (1). Under the modified DQM condition, the Markov inequality yields,

$$\begin{aligned}\mathbb{P}(\textstyle\sum_i R_i^2 > \epsilon) &\leqslant \epsilon^{-2} n \mathbb{E}(R_1^2) \\ &= \epsilon^{-2} n \textstyle\int r_n^2(x;\xi_n,\theta_n)\mathrm{d}\mu(x) \to 0.\end{aligned}$$

Proof of (2) and (3). Since both f_n and f_0 are objects with $\mathcal{L}_2(\mu)$-norm 1

$$\begin{aligned}0 &= \|f_n\|_{\mu,2}^2 - \|f_0\|_{\mu,2}^2 \\ &= (\xi_n-\xi_0)^4\|\nu_\xi\|_{\mu,2}^2 + (\theta_n-\theta)^\top\|\nu_\theta\|_{\mu,2}^2(\theta_n-\theta) + \|r_n\|_{\mu,2}^2 + 2\left\langle(\theta_n-\theta)^\top\nu_\theta, r_n\right\rangle \\ &\quad +2(\xi_n-\xi_0)^2(\theta_n-\theta)^\top\langle\nu_\theta,\nu_\xi\rangle + 2(\xi_n-\xi_0)^2\langle\nu_\xi, r_n\rangle + 2(\xi_n-\xi_0)^2\langle f_0,\nu_\xi\rangle \\ &\quad +2(\theta_n-\theta)^\top\langle f_0,\nu_\theta\rangle + 2\langle f_0, r_n\rangle\end{aligned}$$

Let $\{\theta_n,\xi_n\}$ be sequences such that $\theta_n-\theta = O(n^{-1/2})$ and $(\xi_n-\xi_0)^2 = O(n^{-1/2})$. Note that by Cauchy-Schwarz inequality and the fact that both ν_ξ and ν_θ are square integrable with respect to measure μ by assumption, $\langle\nu_\xi, r_n\rangle = o(1/\sqrt{n})$ and $\langle\nu_\theta, r_n\rangle = o(1/\sqrt{n})$. Therefore, the third, fourth and the sixth terms are of order $o(1/n)$. The first, second and fifth terms are of order $O(1/n)$. The ninth term is of order $o(n^{-1/2})$ by Cauchy-Schwarz inequality. The seventh and eighth term are both of order $O(1/\sqrt{n})$, but in order for the identity to hold, they must be of smaller order to balance with other terms. For this to happen, we must have

$$\langle f_0,\nu_\xi\rangle = \langle f_0,\nu_\theta\rangle = 0$$

This proves (2) since $0 = \langle f_0, \nu_\xi \rangle = \mathbb{E}(\nu_\xi(X)/f_0)$. Similar argument shows $\mathbb{E}(\nu_\theta(X)/f_0) = 0$. Hence,

$$\begin{aligned} 2\langle f_0, r_n \rangle &= -(\xi_n - \xi_0)^4 \|\nu_\xi\|_{\mu,2}^2 - (\theta_n - \theta)^\top \|\nu_\theta\|_{\mu,2}^2 (\theta_n - \theta) \\ &\quad -2(\xi_n - \xi_0)^2 (\theta_n - \theta)^\top \langle \nu_\xi, \nu_\theta \rangle + o(1/n) \\ &= -\tfrac{1}{4n} t^\top J t + o(1/n) \end{aligned}$$

with $t^\top = (\delta_1^2, \delta_2^\top)$.

Since $\mathbb{V}(2\sum_i R_i)$ is bounded above by $4\sum_i \mathbb{E}(R_i^2)$, which goes to 0 from (1), we have

$$\begin{aligned} 2\textstyle\sum_i R_i &= 2n\mathbb{E}(R_1) + o_P(1) \\ &= 2n\langle f_0, r_n \rangle + o_P(1) \\ &= 2n\left(-\tfrac{1}{8n} t^\top J t + o(1/n)\right) + o_P(1) \\ &= -\tfrac{1}{4} t^\top J t + o_P(1) \end{aligned}$$

Proof of (4). By Hölder's inequality,

$$\sum_i R_i \frac{2}{\sqrt{n}} \frac{\nu_\xi}{f_0} \leqslant \sqrt{\sum_i R_i^2 \sum_i (\frac{2}{\sqrt{n}} \frac{\nu_\xi}{f_0})^2} = o_P(1) O_P(1) = o_P(1)$$

Similar argument admits the second result.

Proof of (5).

$$\mathbb{P}(\max_{1 \leqslant i \leqslant n} |R_i| > \epsilon) \leqslant n\mathbb{P}(|R_i|^2 > \epsilon^2) \leqslant \epsilon^{-2} n \mathbb{E}(R_i^2) \to 0$$

Proof of (6).

$$\begin{aligned} \mathbb{P}(\max_{1 \leqslant i \leqslant n} |2\nu_\xi / f_0| > \epsilon\sqrt{n}) &\leqslant n\mathbb{P}(|2\nu_\xi/f_0| > \epsilon\sqrt{n}) \\ &\leqslant \epsilon^{-2}\mathbb{E}\left((2\nu_\xi(X_1)/f_0)^2\right) \mathbb{I}_{[|2\nu_\xi/f_0| > \epsilon\sqrt{n}]} \to 0 \end{aligned}$$

Similar argument admits the second statement. ∎

Proof of Theorem 1 We consider $\xi_n = \xi_0 + \delta_1 n^{-1/4}$ and $\theta_n = \theta + \delta_2 n^{-1/2}$ throughout the proof. Under **Assumption 1**, we have the following Taylor expansion:

$$f_n = f_0 + (\xi_n - \xi_0)^2 \nu_\xi + (\theta_n - \theta)^\top \nu_\theta + r_n(x_i; \xi_n, \theta_n).$$

Denoting $w_i = 2(f_n/f_0 - 1)$, we have

$$w_i = 2(\xi_n - \xi_0)^2 \frac{\nu_\xi}{f_0} + 2(\theta_n - \theta)^\top \frac{\nu_\theta}{f_0} + 2R_i.$$

To show that under the modified DQM condition, the log-likelihood ratio admits a quadratic approximation, we use results in **Lemma 1**.

The log-likelihood ratio can be represented as

$$\begin{aligned} \Lambda_n &= \sum_i \log \frac{p(x_i; \xi_n, \theta_n)}{p(x_i; \xi_0, \theta)} = \sum_i 2\log \frac{f_n}{f_0} = \sum_i 2\log(1 + w_i/2) \\ &= \sum_i w_i - \frac{1}{4}\sum_i w_i^2 + \frac{1}{2}\sum_i w_i^2 \beta(w_i) \end{aligned}$$

with $\beta(x) \to 0$ as $x \to 0$.

Using (3) in **Lemma 1** and with $S_n = (S_{\xi,n}, S_{\theta,n}^\top)^\top$ and J defined in **Theorem 1**, we have

$$\sum_i w_i = 2\frac{\delta_1^2}{\sqrt{n}} \sum_i \frac{\nu_\xi}{f_0} + 2\frac{\delta_2^\top}{\sqrt{n}} \sum_i \frac{\nu_\theta}{f_0} + 2\sum_i R_i = t^\top S_n - \frac{1}{4} t^\top J t + o_P(1)$$

Using (1) and (4) in **Lemma 1**, we have

$$\begin{aligned} \textstyle\sum_i w_i^2 &= \textstyle\sum_i \left(\frac{2\delta_1^2}{\sqrt{n}} \frac{\nu_\xi}{f_0} + \frac{2\delta_2^\top}{\sqrt{n}} \frac{\nu_\theta}{f_0} + 2R_i \right)^2 \\ &= t^\top J t + o_P(1) + 4 \textstyle\sum_i R_i^2 + 4 \sum_i R_i \left(\frac{2\delta_1^2}{\sqrt{n}} \frac{\nu_\xi}{f_0} + \frac{2\delta_2^\top}{\sqrt{n}} \frac{\nu_\theta}{f_0} \right) \\ &= t^\top J t + o_P(1) \end{aligned}$$

Lastly, we need to show that $\sum_i w_i^2 \beta(w_i) = o_p(1)$. First note that using (5) and (6) in **Lemma 1**, we have

$$\begin{aligned} \mathbb{P}\left(\max_{1 \leqslant i \leqslant n} |w_i| > \epsilon \right) \leqslant\ & \delta_1^2 \mathbb{P}\left(\max_{1 \leqslant i \leqslant n} \left| \tfrac{2}{\sqrt{n}} \tfrac{\nu_\xi}{f_0} \right| > \epsilon \right) + \delta_2^\top \mathbb{P}\left(\max_{1 \leqslant i \leqslant n} \left| \tfrac{2}{\sqrt{n}} \tfrac{\nu_\theta}{f_0} \right| > \epsilon \right) \\ & + 2\mathbb{P}\left(\max_{1 \leqslant i \leqslant n} |R_i| > \epsilon \right) \to 0 \end{aligned}$$

Since when $w_i \to 0$, $\beta(w_i) \to 0$, we have $\max_{1 \leqslant i \leqslant n} |\beta(w_i)| = o_p(1)$. By Hölder's inequality,

$$\sum_i w_i^2 \beta(w_i) \leqslant \max_{1 \leqslant i \leqslant n} |\beta(w_i)| \sum_i w_i^2 = o_P(1) O_P(1) = o_P(1).$$

Therefore, the log-likelihood ratio is approximated by

$$\begin{aligned} \Lambda_n &= \textstyle\sum_i w_i - \frac{1}{4} \sum_i w_i^2 + \frac{1}{2} \sum_i w_i^2 \beta(w_i) \\ &= t^\top S_n - \tfrac{1}{4} t^\top J t - \tfrac{1}{4} t^\top J t + o_p(1) \\ &= t^\top S_n - \tfrac{1}{2} t^\top J t + o_p(1) \end{aligned}$$

■

Proof of Corollary 1 Since S_n is a normed iid sum, by the central limit theorem,

$$S_n \overset{P_{n,\xi_0,\theta}}{\rightsquigarrow} \mathcal{N}(0, J)$$

The zero asymptotic mean of S_n is provided by (2) in **Lemma 1**, then the asymptotic variance for S_n is J as defined in Theorem 1.

The quadratic approximation for Λ_n established in **Theorem 1** together with the joint normality of S_n leads to the LAN property of the sequence of model P_{n,ξ_n,θ_n}. Furthermore, we have

$$\Lambda_n \overset{P_{n,\xi_0,\theta}}{\rightsquigarrow} \mathcal{N}(-\frac{1}{2} t^\top J t, t^\top J t).$$

By LeCam's first lemma (see e.g. van der Vaart (1998, Lemma 6.4)), P_{n,ξ_n,θ_n} and $P_{n,\xi_0,\theta}$ are mutually contiguous. ■

Proof of Theorem 2 The sequence of experiments $\mathcal{E}_n$ converges to a shifted Gaussian $\mathcal{N}(t, J^{-1})$ as a result of Theorem 9.4 in van der Vaart (1998). The log-likelihood ratio process of observing one sample from $\mathcal{N}(t, J^{-1})$ is

$$\log \frac{d\mathcal{N}(t, J^{-1})}{d\mathcal{N}(0, J^{-1})}(Y) = t^\top JY - \frac{1}{2} t^\top Jt$$

It suffices to show that $J^{-1}S_n$ converges to the distribution of Y under the null. **Corollary 1** establishes $S_n \overset{P_{n,\xi_0,\theta}}{\rightsquigarrow} \mathcal{N}(0, J)$, we thus have $J^{-1}S_n \overset{P_{n,\xi_0,\theta}}{\rightsquigarrow} \mathcal{N}\left(0, J^{-1}\right)$.

The optimal test statistic for $H_0 : \delta_1 = 0$ against $H_a : \delta_1 \neq 0$ in the limit experiment is the first element in Y. The sequence of test statistics from the original experiment $\mathcal{E}_n$ that matches with the first element in Y is the $C(\alpha)$ statistic,

$$Z_n = (J_{\xi\xi} - J_{\xi\theta} J_{\theta\theta}^{-1} J_{\theta\xi})^{-1/2} (S_{\xi,n} - J_{\xi\theta} J_{\theta\theta}^{-1} S_{\theta,n}).$$

Notice the rescaling in Z_n is needed to obtain a unit asymptotic variance for the test statistic. ■

Proof of Corollary 2 Since ξ is a scalar and $S_n \overset{P_{n,\xi_0,\theta}}{\rightsquigarrow} \mathcal{N}(0, J)$ under H_0, it is immediate that the asymptotic null distribution for Z_n is $\mathcal{N}(0, 1)$.

We can now use LeCam's third lemma (see e.g. van der Vaart (1998, Example 6.7)) to derive the asymptotic distribution for Z_n under local alternatives. We are interested in the local alternative that $\xi_n = \xi_0 + \delta_1 n^{-1/4}$ and nuisance parameter θ is left unspecified as in the null, hence we set $\delta_2 = 0$ in the log-likelihood ratio expansion. Under H_0,

$$(Z_n, \Lambda_n) \overset{P_{n,\xi_0,\theta}}{\rightsquigarrow} \mathcal{N}\left(\begin{pmatrix} 0 \\ -\frac{1}{2}\delta_1^4 J_{\xi\xi} \end{pmatrix} \quad \begin{pmatrix} 1 & \sigma_{12} \\ \sigma_{12} & \delta_1^4 J_{\xi\xi} \end{pmatrix}\right)$$

with $\sigma_{12} = \mathrm{Cov}(Z_n, \Lambda_n) = \delta_1^2 (J_{\xi\xi} - J_{\xi\theta} J_{\theta\theta}^{-1} J_{\theta\xi})^{1/2}$. With $\delta_2 = 0$, **Corollary 1** implies that $P_{n,\xi_n,\theta}$ are mutually contiguous to $P_{n,\xi_0,\theta}$, then LeCam's third lemma implies,

$$Z_n \overset{P_{n,\xi_n,\theta}}{\rightsquigarrow} \mathcal{N}(\sigma_{12}, 1).$$

■

Proof of Theorem 3 Define the class of functions:

$$\mathcal{F}_n := \left\{ x \mapsto (g(x, \theta) - g(x, \eta)) \middle| \|\theta - \eta\| \leqslant \delta_n \right\}.$$

If $\hat{\theta}$ is a $\sqrt{n}$-consistent estimator of θ, and $\delta_n = O(n^{-k})$ with $k < 1/2$, we obtain that with probability tending to one

$$\left| Z_n(\hat{\theta}) - Z_n(\theta) \right| \leqslant \sup_{f \in \mathcal{F}_n} |\mathbb{G}_n(f)|$$

where $\mathbb{G}_n(f) := n^{-1/2}\sum_i(f(X_i) - \mathbb{E}f(X_i))$ denotes the empirical process indexed by $\mathcal{F}_n$. Proving $Z_n(\hat{\theta}) - Z_n(\theta) = o_P(1)$ thus amounts to establishing asymptotic equicontinuity of the process $\mathbb{G}_n$ with respect to the Euclidean norm.

Let the parameter space near true θ, $U_{\delta_n}(\theta)$, be covered by balls with radius $\epsilon^{1/\gamma}$, the number of balls can be upper bounded by $C_1\epsilon^{-p/\gamma}$ with C_1 as a constant that does not depend on n and p being the dimension of the nuisance parameter space. Then for $\forall\eta \in U_{\delta_n}(\theta)$, $\exists N_\eta$, such that

$$\|\eta - \eta_{N_\eta}\| \leqslant \epsilon^{1/\gamma}$$

The condition on g in **Assumption 2** implies

$$|g(x,\eta) - g(x,\eta_{N_\eta})| \leqslant \|\eta - \eta_{N_\eta}\|^\gamma H(x) \leqslant \epsilon H(x)$$

It follows that the bracketing number, $N_{[\,]}(\epsilon\|H\|_2, \mathcal{F}_n, \mathcal{L}_2(P_{n,\xi_n,\theta}))$ is bounded from above by $C_2\epsilon^{-p/\gamma}$.

Furthermore, the assumption also implies that for $f \in \mathcal{F}_n$, $\|f\|_{P_n,2} \leqslant \delta_n^\gamma\|H\|_{P_n,2}$ with $\mathcal{L}_2(P_{n,\xi_n,\theta})$-norm. We can now apply Theorem 2.14.2 in van der Vaart and Wellner (1996) and get

$$\mathbb{E}_{P_{n,\xi_n,\theta}}(\sup_{f\in\mathcal{F}_n}|\mathbb{G}_n(f)|) \leqslant J_{[\,]}(\delta_n^\gamma, \mathcal{F}_n, \mathcal{L}_2(P_{n,\xi_n,\theta}))\|H\|_{P_n,2} + \sqrt{n}\mathbb{E}_{P_{n,\xi_n,\theta}}[H(X)I\{H(X) > \sqrt{n}a(\delta_n^\gamma)\}]$$

where the bracketing integral is defined as

$$J_{[\,]}(\delta_n^\gamma, \mathcal{F}_n, \mathcal{L}_2(P_{n,\xi_n,\theta})) = \int_0^{\delta_n^\gamma}\sqrt{1 + \log N_{[\,]}(\epsilon\|H\|_{P_n,2}, \mathcal{F}_n, \mathcal{L}_2(P_{n,\xi_n,\theta}))}d\epsilon$$

and

$$a(\delta_n^\gamma) = \delta_n^\gamma\|H\|_2/\sqrt{1 + \log N_{[\,]}(\delta_n^\gamma\|H\|_{P_n,2}, \mathcal{F}_n, \mathcal{L}_2(P_{n,\xi_n,\theta}))}.$$

Provided that $\delta_n \to 0$, we have for n large enough,

$$J_{[\,]}(\delta_n^\gamma, \mathcal{F}_n, \mathcal{L}_2(P_{n,\xi_n,\theta})) \leqslant \int_0^{\delta_n^\gamma} 1 + \log(C_2\epsilon^{-p/\gamma})d\epsilon \to 0$$

Since $H(x)$ is square integrable for all n by **Assumption 2**, the first term goes to zero.

The upper bound for the bracketing number also yields a lower bound for $a(\delta_n^\gamma)$ that is for δ_n sufficiently small,

$$a(\delta_n^\gamma) \geqslant \frac{\delta_n^\gamma\|H\|_{P_n,2}}{\sqrt{1 + \log(C_2\delta_n^{-p})}} := k_n \to 0$$

As long as k_n converges to zero slower than c_n, **Assumption 2** ensures that the second term also tends to zero.

The last step is to check that $\sup_{f\in\mathcal{F}_n}\frac{1}{\sqrt{n}}\sum_i \mathbb{E}_{P_{n,\xi_n,\theta}}(f(X_i)) = o(1)$ so that $\sup_{f\in\mathcal{F}_n}|\mathbb{G}_n(f)|$ is the correct upper bound. This is trivially true under the null, where $\xi_n = \xi_0$ for all $n \in \mathbb{N}$, since

$\mathbb{E}_{P_{n,\xi_0,\theta}}(g(X_i,\theta)) = \mathbb{E}_{P_{n,\xi_0,\theta}}(g(X_i,\hat{\theta})) = 0$. Under local alternatives with $\xi_n = \xi_0 + \delta_1 n^{-1/4}$ and given the i.i.d. assumption on the sample, it suffices to show that

$$\sup_{\|\eta-\theta\|\leqslant\delta_n} \sqrt{n}\int (g(x,\eta)-g(x,\theta))p(x;\xi_n,\theta)dx = o(1)$$

Denote $p_n = p(x;\xi_n,\theta)$ and $p_0 = p(x;\xi_0,\theta)$, we have the following expansion

$$\begin{aligned}
&\sqrt{n}\textstyle\int (g(x,\eta)-g(x,\theta))p_n dx \\
&\qquad = \sqrt{n}\textstyle\int \Big((g(x,\eta)-g(x,\theta))(\sqrt{p_0}+(\xi_n-\xi_0)^2\nu_\xi(x)+r_n\Big)\sqrt{p_n}dx \\
&\qquad = \sqrt{n}\textstyle\int (g(x,\eta)-g(x,\theta))\sqrt{p_0}\sqrt{p_n}dx \\
&\qquad\quad + \sqrt{n}(\xi_n-\xi_0)^2\textstyle\int (g(x,\eta)-g(x,\theta))\sqrt{p_n}\nu_\xi(x)dx \\
&\qquad\quad + \sqrt{n}\textstyle\int (g(x,\eta)-g(x,\theta))\sqrt{p_n}r_n dx
\end{aligned}$$

The last two terms are $o(1)$ uniformly over η for $\|\eta-\theta\| \leqslant \delta_n$ due to the DQM condition in **Definition 1** and assumption on g in **Assumption 2**. Since Cauchy-Schwarz inequality implies that with respect to $\mathcal{L}_2(\mu)$-norm,

$$\begin{aligned}
|\textstyle\int (g(x,\eta)-g(x,\theta))\sqrt{p_n}\nu_\xi(x)dx| &\leqslant \|(g(x,\eta)-g(x,\theta))\sqrt{p_0}\|_{\mu,2}\|\nu_\xi\|_{\mu,2} \\
&\leqslant \|\eta-\theta\|^\gamma \|H\|_{P_n,2}\|\nu_\xi\|_{\mu,2} = o(1).
\end{aligned}$$

Similarly,

$$|\sqrt{n}\textstyle\int (g(x,\eta)-g(x,\theta))\sqrt{p_n}r_n dx| \leqslant \|(g(x,\eta)-g(x,\theta))\sqrt{p_0}\|_{\mu,2}\sqrt{n}\|r_n\|_{\mu,2} = o(1).$$

The first term is also $o(1)$ by expanding $\sqrt{p_n}$ again and applying Cauchy-Schwarz inequality in a similar fashion. ∎

Proof of Theorem 4 As in the proof of **Theorem 2**, the limit of the sequence v_n is a shifted Gaussian experiment $Y \sim \mathcal{N}(t, J^{-1})$ but now with $t^\top = (\delta_1^2, \delta_2^2, 2\delta_1\delta_2, \delta_3^\top)$. An equivalent limit experiment observes $X \sim \mathcal{N}(Jt, J)$ with $X = JY$, because the likelihood ratio process of $\frac{d\mathcal{N}(t,J^{-1})}{d\mathcal{N}(0,J^{-1})}(Y)$ is identical to that of $\frac{d\mathcal{N}(t^\top J,J)}{d\mathcal{N}(0,J)}(X)$.

To be more explicit, denoting the first three elements of X to be X_ξ, and the rest to be X_θ, we have under the alternative,

$$\begin{pmatrix} X_\xi \\ X_\theta \end{pmatrix} \overset{\mathcal{D}}{=} \mathcal{N}\left(\begin{pmatrix} J_{\xi\xi} & J_{\xi\theta} \\ J_{\theta\xi} & J_{\theta\theta} \end{pmatrix}\begin{pmatrix} t_\xi \\ t_\theta \end{pmatrix}, J\right)$$

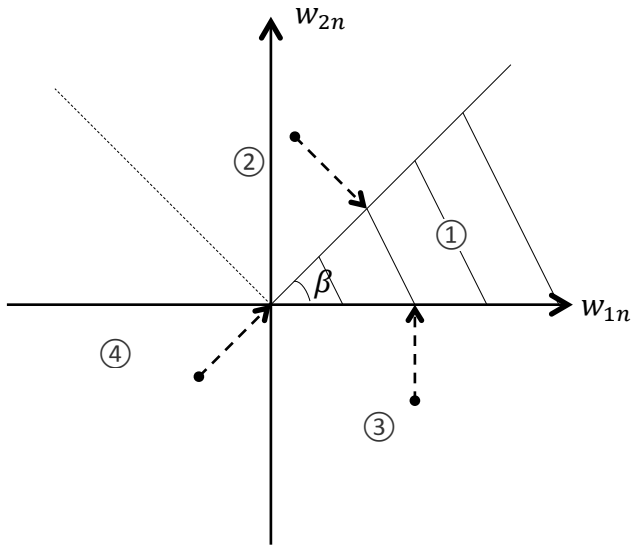
with $t_\xi = (\delta_1^2, \delta_2^2, 2\delta_1\delta_2)^\top$ and $t_\theta = \delta_3^\top$.

To focus on testing for zero restrictions on t_ξ, we find the conditional distribution of X_ξ on X_θ to be

$$\tilde{X}_\xi = X_\xi - J_{\xi\theta}J_{\theta\theta}^{-1}X_\theta \overset{\mathcal{D}}{=} \mathcal{N}((J_{\xi\xi} - J_{\xi\theta}J_{\theta\theta}^{-1}J_{\theta\xi})t_\xi, J_{\xi\xi} - J_{\xi\theta}J_{\theta\theta}^{-1}J_{\theta\xi}).$$

The matched statistic from the original experiment is then

$$\tilde{S}_{\xi,n} = S_{\xi,n} - J_{\xi\theta}J_{\theta\theta}^{-1}S_{\theta,n}$$



Under H_0, $\tilde{S}_{\xi,n}$ follows $\mathcal{N}(0,\Sigma)$ with $\Sigma = J_{\xi\xi} - J_{\xi\theta}J_{\theta\theta}^{-1}J_{\theta\xi}$, and under local alternative, its asymptotic distribution is $\mathcal{N}(\Sigma t_\xi, \Sigma)$.

Notice we can decompose $\tilde{S}_{\xi,n}\Sigma^{-1}\tilde{S}_{\xi,n}$ into two independent pieces as $u_n^\top \Sigma_{11.2} u_n + w_{3n}^\top w_{3n}$. Let the Cholesky decomposition of $\Sigma_{11.2}$ be such that $\Lambda\Lambda^\top = \Sigma_{11.2}$, then $w_n := \Lambda^{-1}u_n \overset{\mathbb{P}_{n,\xi_0,\theta}}{\rightsquigarrow} \mathcal{N}(0,I)$ and $w_n \overset{\mathbb{P}_{n,\xi_n,\theta_n}}{\rightsquigarrow} \mathcal{N}(\Lambda^\top \begin{pmatrix}\delta_1^2\\ \delta_2^2\end{pmatrix}, I)$. Since $(\delta_1^2,\delta_2^2) \in \mathbb{R}_+^2$ and

$$\begin{pmatrix}\eta_1\\ \eta_2\end{pmatrix} := \Lambda^\top \begin{pmatrix}\delta_1^2\\ \delta_2^2\end{pmatrix} = \begin{pmatrix}\delta_1^2\sqrt{\nu_1} + \rho\delta_2^2\sqrt{\nu_2}\\ \sqrt{\nu_2}\sqrt{1-\rho^2\delta_2^2}\end{pmatrix}$$

The feasible parameter set is therefore the convex cone defined as,

$$\left\{(\eta_1,\eta_2) \mid \eta_2 \geqslant 0, \eta_1 - \frac{\rho}{\sqrt{1-\rho^2}}\eta_2 \geqslant 0\right\}.$$

For test statistic taking a value that falls outside of the feasible set, it needs to be projected onto the set. This yields the following four cases as illustrated in the figure.

Case 1: When the value of the test statistic w_n falls into shaded area ①, the test statistics is the sum of squares of the elements of w_n and w_{3n} which are mutually independent:

$$T_n = w_{1n}^2 + w_{2n}^2 + w_{3n}^2 \sim \chi_3^2$$

Case 2: When the test statistic falls into area ②, we need to project w_n onto the convex cone ①, which gives a point with coordinates $(\rho^2 w_{1n} + \rho\sqrt{1-\rho^2}w_{2n}, \rho\sqrt{1-\rho^2}w_{1n} + (1-\rho^2)w_{2n})$. The $C(\alpha)$ test statistic is hence:

$$\begin{aligned} T_n &= (\rho^2 w_{1n} + \rho\sqrt{1-\rho^2}w_{2n})^2 + (\rho\sqrt{1-\rho^2}w_{1n} + (1-\rho^2)w_{2n})^2 + w_{3n}^2 \\ &= (\rho w_{1n} + \sqrt{1-\rho^2}w_{2n})^2 + w_{3n}^2 \sim \chi_2^2 \end{aligned}$$

Case 3: When the test statistic w_n falls in area ③, projecting onto the region ① yields $(w_{1n}, 0)$ and thus,

$$T_n = w_{1n}^2 + w_{3n}^2 \sim \chi_2^2$$

Case 4: Lastly, when w_n falls into region ④, projecting onto region ① yields $(0, 0)$ and hence,

$$T_n = 0 + w_{3n}^2 \sim \chi_1^2$$

The asymptotic distribution of the $C(\alpha)$ test statistics is a mixture of χ^2's, for which the weights are characterized by the probability of falling into different regions. The angle β spanned by the shaded area ① as marked in the figure is $\beta = \cos^{-1}(\rho)$, hence the probability of falling into region ① is $\frac{\beta}{2\pi}$. The probability of falling into ② and ③ is $\frac{1}{2}$, leaves the probability of falling into ④ as $(\frac{1}{2} - \frac{\beta}{2\pi})$. ∎

Appendix B. Computational details in examples

B.1. **Joint test for Gaussian panel data model.** The information matrix for $(\xi, \theta) = (\xi_1, \xi_2, \mu_0, \sigma_0^2)$ is

$$I = \begin{pmatrix} I_{\xi\xi} & I_{\xi\theta} \\ I_{\theta\xi} & I_{\theta\theta} \end{pmatrix} = \frac{NT}{\sigma_0^4} \begin{pmatrix} 2T & \sigma_0^2 & 0 & 1 \\ \sigma_0^2 & (T+3)\sigma_0^4/2 & 0 & \sigma_0^2/2 \\ 0 & 0 & \sigma_0^2 & 0 \\ 1 & \sigma_0^2/2 & 0 & 1/2 \end{pmatrix}$$

We further find

$$I_{\xi.\theta} = I_{\xi\xi} - I_{\xi\theta} I_{\theta\theta}^{-1} I_{\theta\xi} = \begin{pmatrix} 2NT(T-1)/\sigma_0^4 & 0 \\ 0 & NT(T/2+1) \end{pmatrix}$$

and

$$I_{\xi\theta} I_{\theta\theta}^{-1} = \begin{pmatrix} 0 & 2 \\ 0 & \sigma_0^2 \end{pmatrix}.$$

As we have remarked in Section 2.5, the diagonality of $I_{\xi.\theta}$ provides much convenience for finding the optimal test statistics. Denote

$$T_n := \begin{pmatrix} t_{1n} \\ t_{2n} \end{pmatrix} = I_{\xi.\theta}^{-1/2} \begin{pmatrix} \sum_i \nu_{i1} - 2\sum_i \nu_{4i} \\ \sum_i \nu_{2i} - \sigma_0^2 \sum_i \nu_{4i} \end{pmatrix} = \begin{pmatrix} (2NT(T-1)/\sigma_0^4)^{-1/2} \left(\sum_i (\frac{\bar{y}_{i.} - \mu_0}{\sigma_0^2/T})^2 - NT/\sigma_0^2 \right) \\ (NT(T/2+1))^{-1/2} \left(\sum_i (Z_i - T/2)^2 - NT/2 \right) \end{pmatrix}$$

Replacing (μ_0, σ_0^2) by their MLEs yields the joint $C(\alpha)$ test.

Appendix C. Claim in Section 4

Here we provide the detail derivation for the claim in Section 4 that the reparameterization adopted in Chesher (1984) and Cox (1983) for heterogeneity test requires extra moment conditions on U for second derivative of log density with respect to the test parameter to be bounded.

Proposition 1. For iid random variable $Y_1, \ldots, Y_n$ each with density function $\int p(y; \lambda_0 + \tau\sqrt{\eta}u_i)dF(u_i)$, where U_i is a random variable with zero mean and unit variance. The second-order derivative of the log density with respect to η evaluated under $\eta = 0$ is unbounded unless $\mathbb{E}(U^3) = 0$ and $\mathbb{E}(U^4) < \infty$.

Proof Denote the log density as $l = \log \int p(y; \lambda_0 + \tau\sqrt{\eta}u_i)dF(u_i)$. The first order derivative with respect to η is

$$\nabla_\eta l\big|_{\eta=0} = \frac{\tau \int \nabla_\lambda p(y; \lambda_0) u dF(u)}{2\sqrt{\eta} \int p(y; \lambda_0) dF(u)} = \frac{\tau^2}{2}\mathbb{E}(U^2)\frac{\nabla^2_\lambda p(y; \lambda_0)}{p(y; \lambda_0)}$$

The last step is obtained by applying the l'Hôpital's rule.

The second order derivative is

$$\begin{aligned}\nabla^2_\eta l\big|_{\eta=0} &= \frac{\tau^2\sqrt{\eta}\int \nabla^2_\lambda p(y;\lambda_0)u^2 dF(u) - \tau\int \nabla_\lambda p(y;\lambda_0) u dF(u)}{4\eta\sqrt{\eta}\int p(y;\lambda_0)dF(u)}\Big|_{\eta=0} - \Big(\nabla_\eta l\big|_{\eta=0}\Big)^2 \\ &= \frac{\tau^3 \int \nabla^3_\lambda p(y;\lambda_0) u^3 dF(u)}{12\sqrt{\eta}\int p(y;\lambda_0)dF(u)}\Big|_{\eta=0} - \Big(\nabla_\eta l\big|_{\eta=0}\Big)^2\end{aligned}$$

Provided that $\nabla^3_\lambda p(y; \lambda_0)$ is not degenerately zero, $\nabla^2_\eta l$ is unbounded unless $\mathbb{E}(U^3) = 0$ and $\mathbb{E}(U^4) < \infty$ so that we can apply l'Hôpital's rule again and get

$$\nabla^2_\eta l\big|_{\eta=0} = \frac{\tau^4}{12}\left[\mathbb{E}(U^4)\frac{\nabla^4_\lambda p(y; \lambda_0)}{p(y; \lambda_0)} - 3\mathbb{E}(U^2)^2\Big(\frac{\nabla^2_\lambda p(y; \lambda_0)}{p(y; \lambda_0)}\Big)^2\right] < \infty$$

■