UNOBSERVED HETEROGENEITY IN ECONOMIC MODELS: TESTING AND ESTIMATION

BY

JIAYING GU

DISsertation

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Economics in the Graduate College of the University of Illinois at Urbana-Champaign, 2015

Urbana, Illinois

Doctoral Committee:

Professor Roger Koenker, Chair
Professor Anil Bera
Professor EunYi Chung
Professor Keisuke Hirano, University of Arizona
Professor Xiaofeng Shao
Professor Dave Zhao
Abstract

This dissertation consists three chapters with a central theme on unobserved heterogeneity in economic models. The first two chapters discuss tests for parameter homogeneity against general alternatives. In the first chapter, I propose a unified framework for score tests of parameter homogeneity based on Neyman’s C(α) 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(α) tests can be established via LeCam’s differentiability in quadratic mean and the limit experiment approach. 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.

The second chapter focuses on the likelihood ratio test for the same class of testing problems. The test statistic is based on estimation of general (nonparametric) mixture models using the Kiefer and Wolfowitz (1956) maximum likelihood method. Recent developments in convex optimization are shown to dramatically improve upon earlier EM methods for computation of these estimators, and new results on the large sample behavior of likelihood ratios involving such estimators yield a tractable form of asymptotic inference. The computation efficiency also allows the use of a bootstrap method to determine critical values that are shown to work better than the asymptotic critical values in finite samples and consistency of the bootstrap procedure is formally proved. We compare performance of the likelihood ratio test with that of the C(α) test proposed in the first chapter and identify circumstances in which each is preferred.

The last chapter discusses estimation method for models with unobserved heterogeneity. In particular, the empirical Bayes methods for Gaussian compound decision problems involving longitudinal data are considered. The methods are first illustrated with some simulation examples and then with an application to models of income dynamics. Using PSID data we estimate a simple dynamic model of earnings that incorporates bivariate heterogeneity in intercept and variance of the innovation process. Profile likelihood is employed to estimate an AR(1) parameter controlling the persistence of the innovations. We find that persistence is relatively modest when we permit heterogeneity in variances. Evidence of negative dependence between individual intercepts and variances is revealed by the nonparametric estimation of the mixing distribution, and has important consequences for forecasting future income trajectories.
献给我的父母。
Acknowledgments

Looking back, I find it rather lucky that in 2008 I failed to find a job in the banking industry. This circumstance leads me to the graduate school which has been a very rewarding experience. On this journey, I am greatly indebted to my advisor Roger Koenker. Without his continuous support and encouragement, this thesis would not have been completed. It is my privilege to work with him and observe his dedication and passion in his academic pursuit. His outstanding scholarship has inspired me to stay curious and pursue an academic career.

I would also like to take this opportunity to thank many people that are involved in the process. First of all, I would like to mention my master’s thesis advisor, Professor Tilak Abeysinghe, who taught my first Econometrics course and encouraged me to go to the United States for a Ph.D. degree. I would also like to thank my thesis committee: Professor Anil Bera, EunYi Chung, Keisuke Hirano, Xiaofeng Shao and Dave Zhao for their suggestions and comments. I am grateful for many of my friends: Yu-Hsien Kao, Dongwoo Kim, Blake Riley and Rachel Shafer, studying with them makes the first year less painful than it could be; Qinping Feng has been a very good companion in the Econometrics lab, where this thesis is produced; Andreas Hagemann, Shu Shen, Tom Parker, Antonio Galvao, Carlos Lamarche, Sergey Popov among many others have given many useful advice. Last but not least, Jielu Yao is a friend that goes back a long way and has even more courage on her own Ph.D. journey, I thank her for her precious friendship.

Finally, I wish to thank my parents who have sacrificed a lot for encouraging me to do what I love and pursue my passions. I also wish to thank Stanislav Volgushev for his love and endless support. I dedicate this thesis to my parents.
# Table of Contents

Overview ................................................................. 1

Chapter 1  Neyman’s C(\(\alpha\)) Test for Unobserved Heterogeneity ................. 3
    1.1 Introduction ..................................................... 3
    1.2 C(\(\alpha\)) test for unobserved parameter heterogeneity: scalar case .......... 4
    1.3 C(\(\alpha\)) test for parameter heterogeneity in higher dimensions .......... 13
    1.4 Examples ........................................................ 20
    1.5 Reparameterization and connection to the information matrix test .......... 24
    1.6 Simulation ..................................................... 27
    1.7 Conclusion ..................................................... 28
    1.8 Proofs .......................................................... 31

Chapter 2  Likelihood Ratio Test for Mixture Models .................................. 41
    2.1 Introduction ..................................................... 41
    2.2 Likelihood ratio tests for mixture models ..................................... 42
    2.3 Parametric bootstrap method for critical values ............................... 47
    2.4 Neyman’s C(\(\alpha\)) tests for mixture models .............................. 51
    2.5 The EM test for mixture models ..................................... 52
    2.6 Simulation ..................................................... 55
    2.7 Conclusion ..................................................... 57
    2.8 Proofs .......................................................... 58

Chapter 3  Unobserved Heterogeneity in Income Dynamics ............................ 73
    3.1 Introduction ..................................................... 73
    3.2 Empirical Bayes: a brief overview ........................................ 74
    3.3 Estimating Gaussian mixture models with longitudinal data ................. 77
    3.4 Heterogeneous income dynamics ........................................ 87
    3.5 Conclusion ..................................................... 97
    3.6 Proofs .......................................................... 99

References ............................................................. 102
Overview

The central theme of my dissertation is to develop flexible methods for studying unobserved heterogeneity. Focusing on this, the three chapters are closely related. The first two chapters consider statistical tests for unobserved heterogeneity. Despite the general acceptance of modeling heterogeneity, it is sometimes attractive if a parsimonious model with homogenous parameter would suffice to describe the data, especially when we already have some heterogeneity in the model but are considering the necessity of incorporating more. This makes statistical tests for parameter heterogeneity valuable. On the other hand, naturally, we are then interested in flexible method to estimate the model with heterogeneity if we indeed reject the hypothesis of homogeneity. This leads to the third chapter which proposes estimation strategy with a particular focus on models with longitudinal data.

The first chapter proposes a unified framework for tests of unobserved heterogeneity in parametric statistic models based on Neyman’s $C(\alpha)$ approach. I find that 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. The common approach in previous literature, encountering such irregularity, is to reparameterize to bring the problem back to the classical case. However, I find that such reparameterization leads to an unnecessary zero third moment or symmetry assumption on the heterogeneity distribution. I show that dealing with irregularity directly, local asymptotic optimality of the $C(\alpha)$ test can be established via LeCam’s differentiability in quadratic mean and the limit experiment approach. In contrast to the classical Cramér type of regularity conditions employed by Neyman (1959) for the $C(\alpha)$ test, the LeCam framework provides a set of much cleaner and less restrictive assumptions. I also extend the framework to multiple dimensional case. The limit experiment perspective is very useful in this extension in the sense that we can first develop optimal test for the Gaussian limit and then extend to the asymptotic $C(\alpha)$ test. The one-sided nature revealed by the limit experiment suggests modifications of existing multiple dimensional tests in such settings, e.g. the information matrix test of White (1982), that leads to power improvement.

The second chapter considers the likelihood ratio test, which is a natural alternative to the score test proposed in the first chapter for testing parameter heterogeneity. The score test is attractive for its computational convenience, however, recent developments in convex optimization are shown to dramatically improve upon earlier EM methods for computation of these estimators, and new results on the large sample behavior of likelihood ratios involving such estimators yield a tractable
form of asymptotic inference. The computation efficiency also allows the use of a bootstrap method to determine critical values that are shown to work better than the asymptotic critical values in finite samples. We compare performance of the LRT in this chapter with the $C(\alpha)$ test in the previous chapter and identify circumstances in which each is preferred.

Once we develop inference methods for unobserved heterogeneity, it is natural to consider how to estimate these features. It turns out that in applied econometrics, it is well accepted that individual latent effects are important components of many economic models, especially with longitudinal data. However, incorporating them in completely nonrestrictive way, i.e. the fixed effect approach, leads to inconsistent estimates for other global parameters in the model as noted by Neyman and Scott (1948) for panel data models and Nickel (1981) for dynamic panel data models. On the other hand, imposing arbitrary parametric distributional assumption on heterogeneity, i.e. the classical random effect approach, is also unsound. As shown by Heckman and Singer (1984), the global parameters are sensitive to ad hoc parametric assumptions. It is therefore desirable to seek a balance with a minimum amount of assumed structure on the unobserved heterogeneity part of the model while maintaining good statistical properties for the global parameters. This leads to several recent proposals in imposing group structure on fixed effect models, notably Bester and Hansen (2013) and Bonhomme and Manresa (2014). Instead I take a random effect outlook in the third chapter, building on Robbins (1956) and Kiefer and Wolfowitz (1956). Formulating the individual heterogeneity with a general mixture model allows nonparametric identification of the heterogeneity distribution and the likelihood framework also provides efficient estimation and inference of the global parameters via profiling. Additionally, it is attractive from a prediction point of view to maintain the random effect perspective of unobserved heterogeneity. It is a general impression in the literature that random effect model involving mixtures and integration operation is very hard to compute. Recent advances using modern optimization method for general mixture models by Koenker and Mizera (2014) opens way to a wide range of applications of these methods.

We first illustrate these methods with some simulation examples and then with an application to models of income dynamics. Using PSID data we estimate a simple dynamic model of earnings that incorporates bivariate heterogeneity in intercept and variance of the innovation process. Profile likelihood is employed to estimate an AR(1) parameter controlling the persistence of the innovations. We find that persistence is relatively modest, $\hat{\rho} \approx 0.48$, when we permit heterogeneity in variances. Evidence of negative dependence between individual intercepts and variances is revealed by the nonparametric estimation of the mixing distribution, and has important consequences for forecasting future income trajectories.
Chapter 1

Neyman’s C(α) Test for Unobserved Heterogeneity

1.1 Introduction

Neyman’s (1959) C(α) 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(α) 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(α) test in such settings under local alternatives of order $n^{-1/4}$. The limit experiment perspective is very useful especially in the multi-dimensional setting. It allows us to first develop the best test statistics for the Gaussian limit and then extend to the corresponding asymptotic C(α) test. The one-sided nature of the limit experiment reveals that we require a mixture of χ² asymptotics which leads to power improvement compared to the conventional χ² type test. This finding is relevant to the information matrix test and some of the recent applications to slope heterogeneity test 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 the 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 in Section 1.3. In Section 1.4 we consider four different examples. The first example, the C(α) 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(α) test in this setting shares much similarity with 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(α) test to jointly test in Gaussian panel data models for heterogenous location and scale parameters. Lastly, we compare the C(α) 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(α) test, as it pays explicit
attention to the positivity constraints under the alternative, enjoys a power improvement.

The $C(\alpha)$ test for heterogeneity formulated in this chapter 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 further development to a somewhat neglected example on testing for parameter heterogeneity in a Poisson model in Neyman and Scott (1966). Moran (1973) investigates the asymptotic behavior of these score tests. However, as we will show in Section 1.5, the parameterization adopted in Moran (1973) and also Chesher (1984) requires some unnecessary additional assumptions, even though it delivers the same test statistics as the $C(\alpha)$ test constructed here. We conclude in Section 1.5 that the $C(\alpha)$ test for unobserved heterogeneity is not always identical to the IM test, and illustrate some conditions for equivalence to hold. Lastly, a Monte Carlo simulation is carried out to evaluate the power performance for various examples.

1.2 $C(\alpha)$ test for unobserved parameter heterogeneity: scalar case

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 $\alpha$ that gives the name to the $C(\alpha)$ test.

1.2.1 $C(\alpha)$ 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 $\theta$ are nuisance parameters belonging to $\Theta \subset \mathbb{R}^p$ and $\xi$ are parameters under test that belong to $\Xi \subset \mathbb{R}^q$. For densities satisfying 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 \( \theta \) 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 \( \theta \) 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 \( g_n(\hat{\theta}_n) = C_{\xi, n} - I_{\xi\theta} I^{-1}_{\theta\theta} C_{\theta, n} \), provides such a test function with variance \( I_{\xi\theta} \equiv I_{\xi\xi} - I_{\xi\theta} I^{-1}_{\theta\theta} I_{\theta\xi} \). Given a \( \sqrt{n} \)-consistent estimator \( \hat{\theta}_n \) for \( \theta \), the \( C(\alpha) \) test

\[ T_n(\hat{\theta}_n) = g_n(\hat{\theta}_n) \top I^{-1}_{\xi\theta} g_n(\hat{\theta}_n) \]

is then asymptotically \( \chi^2_q \) 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 \( \theta \), \( C_{\theta, n} \) is zero and the \( C(\alpha) \) test reduces to Rao’s score test. The component \( I_{\xi\theta} I^{-1}_{\theta\theta} I_{\theta\xi} \) subtracted from the information \( I_{\xi\xi} \) for \( \xi \) 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).

### 1.2.2 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 \( \lambda_i \) 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 \( E(U) = 0, \ V(U) = 1 \). The parameter \( \tau \) is a known finite scale parameter, which allows us to rescale the variance for \( U \) to be unity. It is not restrictive to assume \( \tau \) known, as we will see later that \( \tau \) 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 breaks down because the score function for \( \xi \),
for each individual observation \( x_i \), defined as the first order logarithmic derivative of the density function with respect to \( \xi \), is identically zero under the null, hence the Fisher information is also zero,

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

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 \( E(s_i^2(\lambda_0)) \) under \( H_0 \) by the central limit theorem and by noticing that \( E(p''(x_i; \lambda_0)/p(x_i; \lambda_0)) = 0 \) as a consequence of differentiating \( \int p(x; \lambda) dx = 1 \) as a function of \( \lambda \) twice. This leads to a close analogy with the classical theorem, in which \( s(\lambda_0) \) acts as the score function and the variance \( E(s_i^2(\lambda_0)) \) 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.

### 1.2.3 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 1.2.2 as an illustration, for local alternatives \( \lambda_i = \lambda_0 + \tau \xi_n u_i \), with \( \xi_n \) be a sequence that converges to zero at certain rate, we have the
following Taylor expansion of the log likelihood ratio,
\[
\Lambda_n = \sum_i \log \frac{p(x_i; \lambda_0)}{p(x_i; \lambda)} = \frac{\xi_i^2 \tau^2}{2} \sum_i s_i(\lambda_0) + \frac{\xi_i^3 \tau^3}{3} \sum_i \frac{\nabla^3 p(x_i; \lambda_0)}{p(x_i; \lambda_0)} + \frac{\xi_i^4 \tau^4}{4} \left[ \sum_i \frac{\nabla^4 p(x_i; \lambda_0)}{p(x_i; \lambda_0)} - 3 \sum_i \frac{\nabla^2 p(x_i; \lambda_0)}{p(x_i; \lambda_0)} \right] + o_p(1).
\]
Let \( \xi_n \) 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 \( \xi \) 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 \( L_2 \) 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 \( \mu \). 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 1.2.1.** The density function \( p \) satisfies the following conditions:

1. The null value \( \xi_0 \) is an interior point of \( \Xi \).

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 \( \xi \) and once continuously differentiable with respect to \( \theta \) for \( \mu \)-almost all \( x \).

3. Denoting the first two derivatives of the density with respect to \( \xi \), evaluated under the null as \( \nabla_\xi p(x; \xi_0, \theta) \) and \( \nabla^2_\xi p(x; \xi_0, \theta) \), we have \( \mathbb{P} (\nabla_\xi p(x; \xi_0, \theta) = 0) = 1 \) and \( \mathbb{P} (\nabla^2_\xi p(x; \xi_0, \theta) \neq 0) > 0 \) for all \( \theta \in \Theta \subset \mathbb{R}^p \).

4. Denoting the derivative of the density with respect to \( \theta \) evaluated under the null as \( \nabla_\theta p(x; \xi_0, \theta) \), for any \( p \)-dimensional vector \( a \), \( \mathbb{P} (\nabla^2_\xi p(x; \xi_0, \theta) \neq a^T \nabla_\theta p(x; \xi_0, \theta)) > 0 \).
Remark 1.2.2. Here ξ is the parameter under test and θ is the vector of nuisance parameters. The list of regularity conditions in Assumption 1.2.1 tailors the standard conditions for a regular C(α) 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(α) test statistics to be non-degenerate.

Under Assumption 1.2.1, we can now define the modified DQM condition that is crucial for establishing the local asymptotic normality of the model.

Definition 1.2.3. The density p(x; ξ, θ) satisfies the modified differentiability in quadratic mean condition at (ξ_0, θ) if there exists a vector v(x) = (v_ξ(x), v_θ(x))\top ∈ \mathcal{L}_2(μ) such that as (ξ_n, θ_n) → (ξ_0, θ),

\[ \int |\sqrt{p(x; ξ_n, θ_n)} - \sqrt{p(x; ξ_0, θ)} - h_n^\top v(x)|^2 dμ(x) = o(\|h_n\|^2) \]

where h_n = ((ξ_n - ξ_0)^2, (θ_n - θ)^\top)^\top. Here \| \cdot \| denotes the Euclidean norm and \mathcal{L}_2(μ) denotes the \mathcal{L}_2 space of square integrable functions with respect to measure μ.

Furthermore, let β(h_n) be the mass of the part of p(x; ξ_n, θ_n) that is p(x; ξ_0, θ)-singular, then as (ξ_n, θ_n) → (ξ_0, θ),

\[ \frac{β(h_n)}{\|h_n\|^2} \to 0 \]

Usually the vector v(x) contains derivatives of the square root of density \sqrt{p(x; ξ_n, θ_n)} with respect to each parameter evaluated under their null values. Definition 1.2.3 modifies the classical DQM condition such that whenever the first order derivative is degenerately 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, . . . , X_n) with joint distribution P_{n, ξ_n, θ_n} in which ξ_n = ξ_0 + δ_1 n^{-1/4} and θ_n = θ + δ_2 n^{-1/2}.

Theorem 1.2.4. Suppose (X_1, . . . , X_n) are i.i.d. random variables with joint distribution P_{n, ξ_n, θ_n} and the density satisfies Assumption 1.2.1 and the modified DQM condition with

\[ v(x) = (v_ξ(x), v_θ^\top(x)) = \left( \frac{1}{4} \nabla_ξ^2 p(x; ξ_0, θ) |_{p(x; ξ_0, θ) > 0}, \frac{1}{2} \nabla_θ p(x; ξ_0, θ)^\top |_{p(x; ξ_0, θ) > 0} \right)^\top, \]

then for fixed δ_1 and δ_2, the log-likelihood ratio has the following quadratic approximation under
where $t = (\delta_1^2, \delta_2^2)^T$, 

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

and

$$J = 4 \int (vv^T) d\mu(x) = \begin{pmatrix} \text{Cov}(S_{\xi,n}^2, S_{\theta,n}^2) & \text{Cov}(S_{\xi,n}, S_{\theta,n}^T) \\ \text{Cov}(S_{\theta,n}, S_{\theta,n}^T) & \text{Cov}(S_{\theta,n}^2, S_{\theta,n}^2) \end{pmatrix} = \begin{pmatrix} J_{\xi\xi} & J_{\xi\theta} \\ J_{\theta\xi} & J_{\theta\theta} \end{pmatrix}.$$ 

Corollary 1.2.5. With $S_n$ and $J$ defined as in Theorem 1.2.4, we have

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

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

Theorem 1.2.4 shows that under Assumption 1.2.1, 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 $\xi_0$ and the $n^{-1/2}$ neighborhood of the nuisance parameter $\theta$. The joint normality of the vector $S_n$, as established in Corollary 1.2.5, 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 $\xi$ 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 1.2.1, 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 \( P_\vartheta \), if the sequence of the local model \( P_{n, \vartheta_n} \) 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 statistic is matched with the optimal testing procedure in the Gaussian shift limit experiment, hence establishing its optimality.

**Theorem 1.2.6.** Let \( E_n \) be a sequence of experiments based on i.i.d. random variables \((X_1, \ldots, X_n)\) with joint distribution \( P_{n, \xi_n, \theta_n} \) on the sample space \((X_n, A_n)\). We further index the sequence of experiment by \( t = (\delta_1^2, \delta_2^2)^\top \in \mathbb{R}_+ \times \mathbb{R}^p \). The log-likelihood ratio of the sequence of models satisfies,

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

with the score vector \( S_n \) defined as in Theorem 1.2.4 converging in distribution under the null to \( N(0, J) \). Then the sequence of experiments \( E_n \) converges to the limit experiment based on observing one sample from \( Y = t + v \), where \( v \sim N(0, J^{-1}) \). 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 1.2.7.** Under \( H_0 \), \( Z_n \) has distribution \( N(0, 1) \). 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 \( N(\delta_1^2(J_{\xi \xi} - J_{\xi \theta} J_{\theta \theta}^{-1} J_{\theta \xi})^{1/2}, 1) \).

The optimal test statistic \( Z_n \) takes the form of a \( C(\alpha) \) test. It projects the second-order score \( S_{\xi, n} \) for \( \xi \) onto the space spanned by the first-order score vector \( S_{\theta, n} \) for \( \theta \). 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 \( \delta_1 \), which is the first element in the one sample \( Y \).

One common feature of \( C(\alpha) \) 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 \( \delta_1 \). This is not surprising given that the shift parameter corresponding to \( \xi \) 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 \( \delta_1^2 \). We lose the sign information.
on $\delta_1$, and the asymptotically optimal test, if rejects the null, fails to distinguish whether the
development is from the left or from the right (this phenomenon is also emphasized in Rotnitzky,
Cox, Bottai, and Robins (2000)). Let $Y = (Y_1, Y_2)^T$ where the partition is such that $Y_1$ is a scalar
and $Y_2 \in \mathbb{R}^p$ as in Theorem 1.2.6. In the Gaussian limit experiment based on the one sample
from $Y \sim N(t, J^{-1})$, the one-sided test, rejecting $H_0$ if $Y_1 > \Phi^{-1}(1 - \alpha)(J_{\xi \xi} - J_{\xi \theta}J_{\theta \theta}^{-1}J_{\xi \theta})^{-1/2}$, is
the uniformly most powerful test. Since the sequence that converges to the rescaled first element
$(J_{\xi \xi} - J_{\xi \theta}J_{\theta \theta}^{-1}J_{\xi \theta})^{1/2}Y_1$ is exactly $Z_n$, it implies that the asymptotically optimal $C(\alpha)$ test rejects
$H_0$ if $Z_n > \Phi^{-1}(1 - \alpha)$ for any level $\alpha$. Observe that for $\alpha < 0.5$, this is equivalent to rejecting $H_0$
if $(0 \vee Z_n)^2 > c_\alpha$, where $c_\alpha$ is the $(1 - \alpha)$-quantile of $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ and $\chi_0^2$ is a degenerate distribution
with mass 1 at 0. No solution exists for $c_\alpha$ if $\alpha > 0.5$ although this is of little relevance in practice.
We mention the mixture of $\chi^2$ asymptotics just to be more cohesive with the multi-dimensional
extension later. The weight $1/2$ associated with $\chi_0^2$ 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 1.2.2. The $C(\alpha)$ test statistic $Z_n$, constructed from the second-order score
for $\xi$, exploits information of the curvature of the log-likelihood function in the neighborhood of
$\xi_0$. Since at $\xi = \xi_0$, the gradient of the log-likelihood function with respect to $\xi$ 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 1.6, we show that paying attention to this one-sided
feature 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 $\chi^2$ 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 likelihood ratio test for mixture models and the $C(\alpha)$ test is considered in the
next chapter.

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 1.2.1. The similar irregularity also arises in tests for symmetry in normal-skew distribution and is investigated in Hallin and Ley (2014). 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.

1.2.4 Replacing the nuisance parameter by a $\sqrt{n}$-consistent estimator

Notice that the optimal test statistic $Z_n$ we obtained in Theorem 1.2.6 is a function of $\theta$, to make the test statistic feasible under unknown nuisance parameters, we need to replace $\theta$ by some estimator $\hat{\theta}$. In order to ensure that the asymptotics for the test statistic $Z_n$ in Corollary 1.2.7 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

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

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 $\theta$. Neyman’s assumptions are rather strong, for example, he requires the density to be three times differentiable with respect to $\theta$ and also moments of the gradient of $g$ with respect to $\theta$ 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. Our Assumption 1.2.8 below on $g(x, \theta)$ implies the conditions of the Type IV function in Andrews (1994) with $p = 2$.

**Assumption 1.2.8.** There exists some $\delta > 0$ such that for any $\eta, \eta' \in U_\delta(\theta)$ we have for some $\gamma > 0$

$$|g(x, \eta) - g(x, \eta')| \leq \|\eta - \eta'\|^{\gamma} H(x)$$
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 \( n \in \mathbb{N} \), \( \sup_n \mathbb{E}_{P_{n, \xi_n, \theta}} H^2(X) < \infty \). Additionally for some \( c_n = o(1) \),

\[
n^{1/2} \mathbb{E}_{P_{n, \xi_n, \theta}} \left[ H(X) \mathbb{1}_{\{ H(X) > n^{1/2} c_n \}} \right] = o(1).
\]

**Theorem 1.2.9.** Under Assumption 1.2.8, if \( \hat{\theta} \) is a \( \sqrt{n} \)-consistent estimator for \( \theta \), then

\[
|Z_n(\hat{\theta}) - Z_n(\theta)| = o_p(1)
\]

**1.3 \( C(\alpha) \) test for parameter heterogeneity in higher dimensions**

It is of interest to generalize the scalar \( C(\alpha) \) tests of unobserved parameter heterogeneity to higher dimensions. 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 \( H_a : \mu_i \geq 0 \) 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 result on likelihood ratio test (LRT) to cases in which the null value of the parameters under test lies on the boundary of the parameter space. Perlman (1969) and Hillier (1986) among many others consider variants of Gaussian LRT under restricted alternatives. Alternatively, Abelson and Tukey (1963) proposes tests based on the idea of maximin contrast. This is further extended by Schaafsma and Smid (1966) which introduces the optimality concept of “most stringent somewhere most powerful” (MSSMP) test and King and Wu (1997) applies it in a linear regression setting.

Neither LRT nor MSSMP test uniformly dominates each other, but both are shown to be substantially more powerful than the usual \( \chi^2 \) 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 \( \chi^2 \) 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 \( \chi^2 \) 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 \( \chi^2 \) with different degrees of freedom as their asymptotic distribution. One disadvantage of the LRT is that the weights of these \( \chi^2 \)'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.

In the rest of this section, we will first present in details the joint test for heterogeneity in dimension two and then elaborate on more general cases. The limiting distribution of the test statistics under the null involves some nuisance parameters and we propose a simulation-based method to estimate these weights and then find critical values. The method is shown to work well in practice.

1.3.1 Two-dimensional \( C(\alpha) \) test for parameter heterogeneity

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, \( \lambda_i \), of the model. Under the alternative, they take the form, \( \lambda_{ki} = \theta_k + \xi_k U_{ki} \), for \( k = 1, 2 \). Let the covariance matrix for \( U_i = (U_{1i}, U_{2i}) \) be \( \Omega \). Without loss of generality, we let the diagonal element of \( \Omega \) be unity. Under \( H_0 \), \( \xi_k = 0 \), so that \( \lambda_k \)'s are homogenous across individuals taking value \( \theta_k \).

The density function satisfies Assumption 1.2.1 such that the first-order score vector for \( \xi_1 \) and \( \xi_2 \) are vanishing due to the zero mean assumption for \( U_i \) but the second-order score matrix is non-vanishing. It also satisfies the modified DQM condition so that the model is locally asymptotically normal. Typically the score function for \( (\xi_1, \xi_2) \) then consists of all distinct elements in the second-order score matrix. Depending on the assumption on \( \Omega \), some of the elements become zero. For example, if \( \Omega \) is a diagonal matrix, which implies that \( U_1 \) is mutually independent to \( U_2 \), then the off-diagonal terms of the second-order score matrix for \( \xi \) are zero. If \( \Omega \) 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-diagonal terms correspond to shift parameters that can take value over the whole real line. This implies that if \( \Omega \) is not diagonal, then the Gaussian limit experiment has only a subset of the shift parameters that have positivity constraints under the alternative. Theorem 1.3.1 gives the general theory on constructing the \( C(\alpha) \) statistics with the subsequent Corollary 1.3.2 discussing the special case if \( \Omega \) is diagonal.

To proceed, we denote the second-order score vector (all distinct elements in the second or-
der score matrix stacked into a vector) for \((\xi_1, \xi_2)\) as \((S_{\xi_1^2, n}, S_{\xi_2^2, n}, S_{\xi_1 \xi_2, n})\). The first two correspond to the diagonal terms and the last the off-diagonal term. More specifically, under regularity conditions, they are \(S_{\xi_k^2, n} = \frac{1}{2\sqrt{n}} \sum_i \frac{\nabla \xi_k^2 p(x_i; \xi_0, \theta)}{p(x_i; \xi_0, \theta)} I[p(x_i; \xi_0, \theta) > 0]\) for \(k = 1, 2\) and \(S_{\xi_1 \xi_2, n} = \frac{1}{\sqrt{n}} \sum_i \frac{\nabla \xi_1 \xi_2 p(x_i; \xi_0, \theta)}{p(x_i; \xi_0, \theta)} I[p(x_i; \xi_0, \theta) > 0]\). Let the first-order score for \(\theta\) be \(S_{\xi, n}\) and \(S_{\theta, n} = \frac{1}{\sqrt{n}} \sum_i \frac{\nabla \theta p(x_i; \xi_0, \theta)}{p(x_i; \xi_0, \theta)} I[p(x_i; \xi_0, \theta) > 0]\). 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 \(3 \times 3\) block matrix. The residual score for \(\xi\), 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 \(\Sigma = J_{\xi \xi} - J_{\xi \theta} J_{\theta \theta}^{-1} J_{\theta \xi}\). The partition of \(\Sigma\) is such that \(\Sigma_{(11)}\) collects covariance terms for the first two elements in \(\tilde{S}_{\xi, n}\).

**Theorem 1.3.1.** Let \(v_n\) be the sequence of experiments based on i.i.d. random variable \((X_1, \ldots, X_n)\) with joint distribution \(P_{n, \xi, \theta, \theta_0}\) 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 \((X_n, A_n)\). The log-likelihood ratio of the sequence of experiment satisfies,

\[
\log \left( \frac{dP_{n, \xi, \theta, \theta_0}}{dP_{n, \xi_0, \theta_0}} \right) = t^\top S_n - \frac{1}{2} t^\top J t + o_p(1),
\]

with \(S_n = (S_{\xi_1^2, n}, S_{\xi_2^2, n}, S_{\xi_1 \xi_2, n}, S_{\theta, n}^\top) \sim N(0, J)\). Then the limit experiment of \(v_n\) is based on observing one sample from \(Y = t + v\) with \(t = (\delta_1^2, \delta_2^2, 2\delta_1 \delta_2, \delta_2^2)^\top \in \mathbb{R}_+^4 \times \mathbb{R} \times \mathbb{R}^p\) and \(v \sim N(0, J^{-1})\). We would like to jointly test \(H_0 : \delta_1 = \delta_2 = 0\) against the alternative \(H_1 : \delta_1 \neq 0\) 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 \(\Lambda\) be the Cholesky decomposition 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 \(\rho\) is the correlation coefficient between \(u_{1n}\) and \(u_{2n}\) and \(\nu_1\) and \(\nu_2\) 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 \( w_{3n} := \Sigma_{(22)}^{-1/2} \tilde{S}_{\xi_1, \xi_2, n} \). The \( C(\alpha) \) test statistic is one of the following four cases:

\[
T_n = \begin{cases} 
  w_{1n}^2 + w_{2n}^2 + w_{3n}^2 & \text{if } w_{1n} \geq \frac{\rho}{\sqrt{1-\rho^2}} w_{2n}, w_{2n} \geq 0 \\
  w_{2n}^2 + w_{3n}^2 & \text{if } w_{2n} \leq 0, w_{1n} \geq 0 \\
  (\rho w_{1n} + \sqrt{1-\rho^2} w_{2n})^2 + w_{3n}^2 & \text{if } -\frac{\sqrt{1-\rho^2}}{\rho} w_{2n} \leq w_{1n} \leq \frac{\rho}{\sqrt{1-\rho^2}} w_{2n}, w_{2n} \geq 0 \\
  w_{3n}^2 & \text{if } w_{1n} \leq 0, w_{2n} \leq -\frac{\rho}{\sqrt{1-\rho^2}} w_{1n} 
\end{cases}
\]

Under \( H_0 \), the asymptotic distribution of \( T_n \) follows \( (\frac{1}{2} - \frac{\beta}{2\pi})X_1^2 + \frac{1}{2}X_2^2 + \frac{\beta}{2\pi}X_3^2 \) with \( \beta = \cos^{-1}(\rho) \).

**Corollary 1.3.2.** If \( S_{\xi_1, \xi_2, n} = 0 \), then the log likelihood ratio of the sequence of experiment reduces to

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

with \( S_n = (S_{\xi_1, \xi, n}, S_{\xi_2, \xi, n}, S_{\Theta, \xi, n})^\top \sim \mathcal{N}(0, J) \). Then the limit experiment of \( u_n \) is based on observing one sample from \( Y = t + v \) with \( t = (\delta_1^2, \delta_2^2, \delta_3^2)^\top \in \mathbb{R}_+^3 \times \mathbb{R}^P \) and \( v \sim \mathcal{N}(0, J^{-1}) \). Proceed as in Theorem 1.3.1 with \( u_n = (\tilde{S}_{\xi_1, \xi, n}, \tilde{S}_{\xi_2, \xi, n})^\top \) and find the corresponding Cholesky decomposition \( \Lambda \) for \( \Sigma_{(11)} \) and \( w_n = \Lambda^{-1} u_n \). Under \( H_0 \), the asymptotic distribution of \( T_n \) follows \( (\frac{1}{2} - \frac{\beta}{2\pi})X_0^2 + \frac{1}{2}X_1^2 + \frac{\beta}{2\pi}X_3^2 \) with \( \beta = \cos^{-1}(\rho) \).

**Remark 1.3.3.** Essentially the construction of the test statistics \( T_n \) can be reduced to solving a quadratic programming problem. When all shift parameters are constrained to be positive under the alternative as in Corollary 1.3.2, the test statistics is found by solving

\[
T_n = w_n^\top w_n - \inf_{\mu \in \mathbb{R}_+^P} (w_n - \Lambda^\top \mu)^\top (w_n - \Lambda^\top \mu).
\]

Under the more general setting where only a subset of the shift parameters are constrained to be positive as in Theorem 1.3.1, we observe the quadratic form of \( \tilde{S}_{\xi, n} \) can be partitioned into two independent parts as \( \tilde{S}_{\xi, n} \Sigma^{-1} \Sigma_{(12)} \Sigma^{-1} u_n + \tilde{S}_{\xi_1, \xi_2, n} \Sigma_{(22)}^{-1} \tilde{S}_{\xi_1, \xi_2, n} \) and the test statistics is constructed by solving

\[
T_n = u_n^\top \Sigma_{(12)}^{-1} u_n - \inf_{\mu \in \mathbb{R}_+^P} (u_n - \Sigma_{(12)}^\top \mu)^\top \Sigma_{(12)}^{-1} (u_n - \Sigma_{(12)}^\top \mu) + w_{3n}^2
\]

The quadratic programming construction of the \( C(\alpha) \) test statistics can be easily extended to cases with dimensions higher than two as we discuss below.
1.3.2 General multi-dimensional $C(\alpha)$ test

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 $\theta$. LeCam’s third lemma implies that asymptotically $\tilde{S}_{\xi,n}$ follows $\mathcal{N}(0, \Sigma)$ under $H_0$ and it follows $\mathcal{N}(\Sigma(\delta_1^2, \ldots, \delta_q^2, (2\delta_j\delta_k)_{j\neq k})^\top, \Sigma)$ under local alternative. The construction of the $C(\alpha)$ test is to find

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

(1.3.1)

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 constrained 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 solution can be easily found by using the R package quadprog, Turlach and Weignessel (2013). To proceed, we need to characterize the limiting distribution of $T_n$.

The test statistic follows a mixture of $\chi^2$ distribution asymptotically under the null, albeit with more complex weights. In the simplest case, if both $J$ and $\Omega$ happen to be diagonal matrices, then all off-diagonal terms in the second-order score matrix drop and the weights take a very simple exact form. For $\xi \in \Xi \subset \mathbb{R}^q$ and let the residual score for $\xi$ be $\tilde{S}_{\xi,n}$ with its covariance matrix as $\Sigma_q$. The diagonality of $J$ implies diagonality of $\Sigma_q$. The optimal test statistic for $H_0 : \xi_1 = \cdots = \xi_q = 0$ against $H_\alpha : \xi_i \neq 0$ 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 $H_0$, $T_n \sim \sum_{i=0}^q (\frac{q}{i})2^{-q}\chi_i^2$ and critical values can be easily found. 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 $\chi^2$ 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 optimal detection of random coefficient in autoregressive models.

When $\Sigma_q$ is not diagonal, the weights in the mixture of $\chi^2$ depend on $\Sigma_q$. For cases where all off-diagonal terms in $\Omega$ are non-zero, we have the general form of the limiting distribution for $T_n$ solved from (1.3.1) as

$$T_n^{H_0} \sim \chi_{q(q-1)/2}^2 + \sum_{i=0}^q \omega_i \chi_i^2$$

with $\omega_i$ equals to the probability that the number of non-zero elements among the first $q$ entries of the minimizer $\mu^*$ in (1.3.1) equals to $i$. There exists closed form solutions for these weights when the dimensions are not too high (see for example Shapiro (1985)), but they take very complicated forms involving partial correlation coefficients. Given the simple interpretation of these weights, we can find them via Monte Carlo methods. For example, we can simulate from a multivariate normal distribution with mean zero and covariance matrix $\Sigma_q$. For each sample vector, solve the quadratic problem (1.3.1) and record the number of non-zero entries in the solution $\mu^*$. The sample frequency
of these counts serves as good estimates for the mixing weights. In Table 1.1 we illustrate with \( q = 5 \) assuming both \( \Sigma_q \) and \( \Omega \) are diagonal matrices. This is a case where we know exactly what is the mixing weights and the corresponding critical values \( c_\alpha \) for a given level \( \alpha \). It illustrates that the simulation-based method for finding critical value works well in practice.

<table>
<thead>
<tr>
<th>Mixing Weights</th>
<th>Critical Values</th>
<th>( \alpha = 1% )</th>
<th>( \alpha = 5% )</th>
<th>( \alpha = 10% )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Theoretical</td>
<td>0.031</td>
<td>0.156</td>
<td>0.312</td>
<td>0.312</td>
</tr>
<tr>
<td>Simulation</td>
<td>0.031</td>
<td>0.158</td>
<td>0.313</td>
<td>0.311</td>
</tr>
</tbody>
</table>

Table 1.1: Comparison between theoretical and simulation-based limiting distribution of the test statistics under the null for \( q = 5 \). The first six columns of the table compare the mixing weights \( \omega_i \) for \( i = 0, \ldots, q \) and the last three columns compare critical values for different size \( \alpha \). Simulation is based on 5000 random draws from a multivariate normal distribution and solve the quadratic problem in (1.3.1) for each draw. The estimated weights correspond to the frequency of the respective number of non-zero elements in the solutions.

### 1.3.3 Local power comparison

A popular alternative test for parameter heterogeneity is White’s (1982) information matrix (IM) test and Chesher (1984) provides an insightful score test interpretation to the IM test in this context. The test statistic takes the familiar quadratic form of the score function \( \tilde{S}_{n}^{T} \Sigma^{-1} \tilde{S}_{n} \) and the rejection region is constructed using critical values from a \( \chi^2 \) distribution. The crucial difference between the \( C(\alpha) \) test and the IM test is that the former pays explicit attention to the positivity constrains as implied by the corresponding limit experiment. Both tests obtain the correct asymptotic size, but the \( C(\alpha) \) test enjoys a power improvement. Below we compare the power function of both tests under local alternatives as to see how much power improvement is theoretically possible. In the scalar case, this is to compare the one-sided test versus the two-sided test. For the multi-dimensional case, it is slightly more involving to characterize the asymptotic power curve since it requires finding the mixture of \( \chi^2 \) distribution with non-centrality parameters for the test statistics under local alternatives. Both cases illustrate that the asymptotic power curve of the \( C(\alpha) \) test dominates that of the IM test and the power improvement is substantial.

#### 1.3.3.1 Scalar case

Based on results in Corollary 1.2.7, the asymptotic power function of the \( C(\alpha) \) test is

\[
\pi_n^{(1)}(\delta) = P_{H_1}(Z_n \geq c_\alpha) \rightarrow 1 - \Phi(c_\alpha - \delta)
\]

with \( Z_n \sim_{H_0} N(0, 1) \) and \( Z_n \sim_{H_a} N(\delta, 1) \) where \( \delta > 0 \) indexes the location shift under the local alternatives and the critical value \( c_\alpha = \Phi^{-1}(1 - \alpha) \). If instead, we adopt a two-sided test, the
asymptotic power function is

\[ \pi_n^{(2)}(\delta) = \mathbb{P}_{H_1}(|Z_n| \geq \bar{c}_\alpha) \rightarrow 1 - \Phi(\bar{c}_\alpha - \delta) + \Phi(-\bar{c}_\alpha - \delta) \]

with \( \bar{c}_\alpha = \Phi^{-1}(1 - \alpha/2) \).

### 1.3.3.2 Multi-dimensional case

We will illustrate the power comparison using the simplest independent case with two dimensions. Consider the score vector, \( S_{\xi,n} = (S_{\xi_1,n}, S_{\xi_2,n})^\top \), that follows a multivariate normal distribution \( N_2(0, I_2) \) asymptotically under \( H_0 \) and \( N_2((\delta, \delta)^\top, I_2) \) under local alternative with \( \delta > 0 \) indexing the location shift. The \( C(\alpha) \) test, applying results in Corollary 1.3.2, rejects the null if \( T_n = (0 \vee S_{\xi,n})^\top (0 \vee S_{\xi,n}) \geq c_\alpha \) with \( c_\alpha \) being the \((1 - \alpha)\) quantile of the distribution \( \frac{1}{4}X_0^2 + \frac{1}{2}X_1^2 + \frac{1}{4}X_2^2 \). This leads to the asymptotic power function as

\[ \pi_n^{(1)}(\delta) = \mathbb{P}_{H_1}(T_n \geq c_\alpha) \rightarrow 1 - \left( \omega_0\chi_0^2(c_\alpha) + \omega_1\chi_1^2_{n, ncp = \delta^2}(c_\alpha) + \omega_2\chi_2^2_{n, ncp = 2\delta^2}(c_\alpha) \right). \]

The mixing weights are found as \( \omega_0 = \mathbb{P}_{H_1}(S_{\xi_1,n} \leq 0, S_{\xi_2,n} \leq 0) = \Phi(-\delta)^2 \) and \( \omega_1 = \mathbb{P}_{H_1}(S_{\xi_1,n} \leq 0, S_{\xi_2,n} \geq 0) + \mathbb{P}_{H_1}(S_{\xi_1,n} \geq 0, S_{\xi_2,n} \leq 0) = 2\Phi(-\delta)(1 - \Phi(-\delta)) \) and \( \omega_2 = \mathbb{P}_{H_1}(S_{\xi_1,n} \geq 0, S_{\xi_2,n} \geq 0) = (1 - \Phi(-\delta))^2 \).

The IM test rejects \( S_{\xi,n}^\top S_{\xi,n} \) using critical value \( \bar{c}_\alpha \) as the \((1 - \alpha)\) quantile of the \( \chi_2^2 \) distribution, which yields its asymptotic power function as

\[ \pi_n^{(2)}(\delta) \rightarrow 1 - \chi_2^2_{n, ncp = 2\delta^2}(\bar{c}_\alpha) \]

Figure 1.1 compares the limit of \( \pi_n^{(1)}(\delta) \) and \( \pi_n^{(2)}(\delta) \) for both cases. The asymptotic power curve of the \( C(\alpha) \) test dominates that of the IM test and the power improvement is substantial.

---

Figure 1.1: Theoretical asymptotic power curves of the IM test and the \( C(\alpha) \) test for both scalar and two-dimensional cases. Parameter \( \delta \) indexes location shift in the distribution of the corresponding test statistics in local alternatives.
1.4 Examples

In this section, we describe four examples of using the C(\(\alpha\)) 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(\(\alpha\)) test serves as a unification of many tests already available. As another example, Kiefer (1984) and Lancaster (1985) develop tests for parametric heterogeneity in Cox proportional hazard model both of which can be formulated as C(\(\alpha\)) 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(\(\alpha\)) test offers most interesting departures from existing work.

1.4.1 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.

1.4.1.1 Second Moment Test

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

\[
\lambda_i = \lambda_0 e^{\xi U_i} = \exp(x_i' \beta + \xi U_i)
\]

where \(U_i\) are i.i.d. with distribution \(F\), zero mean and unit variance. We have set \(\tau\) 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, \(\beta_0\), given the assumed form for \(\lambda_i\), can be regarded as a random coefficient. We would like to test \(H_0 : \xi = 0\) against \(H_a : \xi \neq 0\) with \(\beta\) as the unspecified nuisance parameters. Since the first-order score with respect to \(\xi\) vanishes, this problem falls into the framework we considered in Section 1.2.

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 mo-
ment with the second moment for the Poisson model under $H_0$. We reject $H_0$ when $(0 \vee Z_n)^2 > c_\alpha$ with $c_\alpha$ as the critical value from the mixture of $\chi^2$.

**Remark 1.4.1.** 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, $\lambda_i$, 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 $\lambda_i$ as a function of the observed covariates and the unobservable $U_i$. This leads us to the next example.

### 1.4.1.2 Second Factorial Moment Test

If instead, under the same setup as we have in Section 1.4.1.1, we assume,

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

The residual score for $\xi$ 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 $V(g(Y_i, \beta)) = 2$. Replacing $\beta$ by its restricted MLE $\hat{\beta}$, the locally optimal $C(\alpha)$ test is

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

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. Notice again that only overdispersion is possible when deviating from the null and the mixture of $\chi^2$ asymptotics is employed.

### 1.4.2 Joint test for slope heterogeneity in linear regression model

We consider a linear cross sectional model,

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

where $\beta_i$ is a $p \times 1$ vector and $u_i \sim \text{IIDN}(0, \sigma^2)$. In addition, we assume $\beta_{ki} = \beta_{k0} + \xi_k U_{ki}$ for $k = 1, \ldots, p$. 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 \( \xi \), and for the nuisance parameters \( \theta = (\alpha, \beta^T, \sigma^2)^T \):

\[
\begin{align*}
S_{\xi,i} & = (\hat{u}_i^2/\hat{\sigma}^2 - 1)z_i/\hat{\sigma}^2 \\
S_{\sigma^2,i} & = (\hat{u}_i^2/\hat{\sigma}^2 - 1)/2\hat{\sigma}^2 \\
S_{\beta,i} & = \frac{\hat{u}_i^2}{\hat{\sigma}^2}x_i \\
S_{\alpha,i} & = \hat{u}_i/\hat{\sigma}^2
\end{align*}
\]

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

\[
LM = \frac{1}{2} \left( \sum_i \tilde{z}_i f_i \right)^T \left( \sum_i \tilde{z}_i \tilde{z}_i^T \right)^{-1} \left( \sum_i \tilde{z}_i f_i \right)
\]

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 \( C(\alpha) \) test takes the same score function for \( \xi \) and \( \theta \), but pays explicit attention to the positivity constraints on those terms in \( S_{\xi,i} \) that are inherited from the diagonal terms of \( x_i x_i^T \).

We can easily find the residual score for \( \xi \) as

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

and the associated information matrix as \( \Sigma = 2(\sum_i \tilde{z}_i \tilde{z}_i^T)/N\hat{\sigma}^4 \). Partition \( \hat{S}_{\xi,n} \) and \( \Sigma \) such that \( \hat{S}_{(1)} \) and \( \Sigma_{(1)} \) correspond to the elements inherited from the diagonal elements of \( x_i x_i^T \) and proceed as in Theorem 1.3.1.

### 1.4.3 Joint test for location and scale heterogeneity in Gaussian panel data model

In this example, we consider a two dimensional \( C(\alpha) \) 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 \( \epsilon_{it} \sim \text{IIDN}(0, 1) \), \( \mu_i = \mu_0 + \xi_i U_{i1} \) and \( \sigma_i^2 = \sigma_0^2 \exp(\xi_2 U_{i2}) \geq 0 \). 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

\[
f_i = \int \int \left( \frac{1}{2\pi\sigma_0^2 \exp(\xi_2 U_{2i})} \right)^{T/2} \exp \left( -\sum_{t=1}^T \frac{(y_{it} - \mu_0 - \xi_1 U_{i1})^2}{2\sigma_0^2 \exp(\xi_2 U_{2i})} \right) dF_1(U_{i1})dF_2(U_{2i})
\]
The respective score for \((\xi_1, \xi_2)\) and the nuisance parameters \((\mu_0, \sigma_0^2)\) are

\[
\begin{align*}
v_{1i} &= \nabla^2_{\xi_1} \log f_i | \xi_1 = \xi_2 = 0 = \left( \frac{\bar{y}_i - \mu_0}{\sigma_0^2/T} \right)^2 - \frac{1}{\sigma_0^2/T} \\
v_{2i} &= \nabla^2_{\xi_2} \log f_i | \xi_1 = \xi_2 = 0 = (Z_i - T/2)^2 - Z_i \\
v_{3i} &= \nabla \mu_0 \log f_i | \xi_1 = \xi_2 = 0 = \frac{\bar{y}_i - \mu_0}{\sigma_0^2/T} \\
v_{4i} &= \nabla \sigma_0^2 \log f_i | \xi_1 = \xi_2 = 0 = (Z_i - T/2)/\sigma_0^2
\end{align*}
\]

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

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

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

with

\[
\begin{align*}
t_{1n} &= (2NT(T-1)/\hat{\sigma}_1^4)^{-1/2} \left( \sum_i (\bar{y}_i - \hat{\mu}_0)^2 - \frac{NT}{\hat{\sigma}_1^2} \right) \\
t_{2n} &= (NT(T/2 + 1))^{-1/2} \left( \sum_i (Z_i - T/2)^2 - \frac{NT}{2} \right)
\end{align*}
\]

We reject \(H_0\) for \(T_n > c_\alpha\) where \(c_\alpha\) is the \((1 - \alpha)\)-quantile of \(\frac{1}{4}\chi_0^2 + \frac{1}{2}\chi_1^2 + \frac{1}{4}\chi_2^2\).

**Remark 1.4.2.** 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 is included in Section 1.8.) 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 1.3.1. Notice the above test is valid for the large \(N\) fixed \(T\) setting, and the local alternative for \(\xi_n\) is of order \(N^{-1/4}\). If \(T\) also tends to infinity, then the local alternative for \(\xi_n\) is of order \(N^{-1/4}T^{-1/2}\).

### 1.4.4 Test for slope heterogeneity in large panels

Example 1.4.3 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,

\[
y_{it} = \alpha_i + \beta_i^\top x_i + \epsilon_{it},
\]
with $\beta_i$ being a $p \times 1$ vector. The null hypothesis of interest is $H_0: \beta_i = \beta$ for all $i$ against $H_a: \beta_i \neq \beta_j$ for at least one pair of $i \neq j$. The PY test is

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

with $M_\tau$ being the familiar demean matrix and $\tilde{S} = \sum_i ((\hat{\beta}_i - \hat{\beta}_{WFE})^T X_i^T M_\tau X_i (\hat{\beta}_i - \hat{\beta}_{WFE}) / \hat{\sigma}_i^2$ 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 $\hat{\sigma}_i^2$. As both $N$ and $T$ go 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.

For the same model, we can also construct the $C(\alpha)$ score test for heterogeneity in coefficients. Since the $C(\alpha)$ test is a score test which only requires estimating the null model, we can consider the large $N$ fixed $T$ setting. Assuming $\beta_{ki} = \beta_{k0} + \xi_k U_i$ for $k = 1, \ldots, p$. The score function for $\xi$, $S_{\xi,n}$, consists the distinct $p(p+1)/2$ elements of the second-order score matrix, which takes the form $\frac{1}{\sqrt{N}} \sum_i (X_i^T M_\xi \tilde{\xi}^T M_\tau X_i / \hat{\sigma}_i^2 - X_i^T M_\tau X_i / \hat{\sigma}_i^2)$ with nuisance parameters replaced by MLEs. The elements of $S_{\xi,n}$ are asymptotically jointly normal with mean zero and covariance matrix $\Sigma$ under $H_0$ and by LeCam’s third lemma, they jointly follow $N(\Sigma t, \Sigma)$ under the local alternative $(\xi_{j,n} = \xi_j + \delta_j N^{-1/4}, j = 1, \ldots, p)$ with $t = (\delta_1^2, \ldots, \delta_p^2, (2\delta_\tau \delta_k)_{j \neq k})^T$ as discussed in Section 1.3. Not surprisingly, given the connection to the score test shown by Chesher (1984), this shares considerable similarity to the White’s (1982) information matrix test. However, the IM test rejects $H_0$ if $S_{\xi,n}^T \Sigma^{-1} S_{\xi,n}$ exceeds the critical value from $\chi^2_{p(p+1)/2}$ at nominal level $\alpha$, while the $C(\alpha)$ test modifies the IM test by adjusting for positivity constraints on $t$ for the respective elements in the score function. We do not repeat the steps here in applying Theorem 1.3.1. In the simulation section, we compare the $C(\alpha)$, the IM test and the PY test and the results show that the $C(\alpha)$ test enjoys a power gain compared to the other two tests. Given the fact that there exists no uniformly most powerful test for the limit experiment in the multi-dimensional case, it is inconclusive that the power of the $C(\alpha)$ test always dominates that of the PY test. A closer look at the analysis of the PY test under local alternatives indicates that the shift parameter is also necessarily positive (see Section 3.2 of Pesaran and Yamagata (2008)), and hence the PY test is also inherently an over-dispersion test. This suggests using the one-sided instead of the two-sided critical values from standard normal distribution for the PY test will not affect size but improves power.

### 1.5 Reparameterization and connection to the information matrix test

#### 1.5.1 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 1.2.2, without loss of generality, we set \( \xi_0 = 0 \) and the density function as \( p(x; \lambda_0 + \tau \sqrt{\eta} U_1) \). Cox (1983) tests for heterogeneity of \( \lambda_i \) by testing \( H_0 : \eta = 0 \) against \( H_1 : \eta > 0 \). Chesher (1984) takes the same model assuming \( U_1 \) 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 \( \eta \), 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 test thus derived will be identical to the C(α) test using the original parameterization that \( \lambda_i = \lambda_0 + \tau \xi U_1 \). However, the second order derivative for \( \eta \) is unbounded unless we impose an additional moment condition on \( U \), that \( \mathbb{E}(U^3) = 0 \) (See proof in Section 1.8). This condition is implicitly satisfied in Chesher (1984) because of his symmetry assumption on \( U \). Moran (1973) also employed this zero third moment condition but remarked that it was hard to rationalize. One explanation for this extra condition is that the original, more natural specification on the random parameter \( \lambda_i = \lambda_0 + \tau \xi U_1 \) with \( \xi \in \mathbb{R} \) is not equivalent to the reparameterization \( \lambda_i = \lambda_0 + \tau \sqrt{\eta} U_1 \) with \( \eta \in \mathbb{R}_+ \) unless \( U \) has a symmetric distribution. As we have seen, the \( \xi \) parameterization has the advantage that no symmetry or higher moment conditions on \( U \) are necessary.

1.5.2 Connection to the information matrix test

Chesher (1984) was the first to point out that White’s (1982) information matrix test is a score test for unobserved heterogeneity. Since Chesher (1984) can be viewed as a reparameterized C(α) test, it is of interest to investigate the connection between the C(α) test for heterogeneity in general and the IM test. We show that the C(α) test for heterogeneity nests the IM test as a special case.

Take again the example in Section 1.2.2, \( Y_1, \ldots, Y_n \) are i.i.d. random variables each with density function \( p(y; \lambda_i) \). The parameter \( \lambda_i \) is a random parameter and we assume it now takes a more general form \( \lambda_i = \lambda_0 + \xi k(\lambda_0) U_1 \) to incorporate both additive and multiplicative specifications. For example, if \( k(\lambda_0) = 1 \), we have the additive form \( \lambda_i = \lambda_0 + \xi U_1 \), 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 \( \lambda_0 \) is known. Theorem 1.2.4 then implies the following expansion of the log-likelihood function, provided that \( \xi_n = O(n^{-1/4}) \),

\[
\log l = \sum_i \log \int p(y_i; \lambda_i) dF(u) = \sum_i \log p(y_i; \lambda_0) + \frac{1}{2} \xi_n^2 \mathbb{E}(U_i^2) \sum_i k(\lambda_0)^2 \frac{\nabla \lambda \! p(y_i; \lambda_0)}{p(y_i; \lambda_0)} + O_P(1)
\]

The first order derivative of \( l \) with respect to \( \xi_n \) is zero evaluated under \( \xi_n = 0 \), and the second-
order score is
\[
\frac{\partial^2}{\partial \xi_n^2} l_{\xi_n=0} = \sum_i k(\lambda_0)^2 \frac{\nabla^2_\lambda p(y_i; \lambda_0)}{p(y_i; \lambda_0)}.
\]
If \( \lambda_0 \) is unknown, we find the corresponding score for \( \lambda_0 \) 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 \( \lambda_i \), as \( \xi^2 E(U_i^2) k(\lambda_0)^2 \). In a regression model with covariates, \( \lambda_0 \) will then be a function of the covariates with coefficients \( \beta \).

White’s (1982) information matrix test under regression setting, on the other hand, is con-
structed based on the following moment conditions:
\[
E \left[ \text{vech} \left( \nabla^2_\beta \log p(y; \lambda_0(x_i, \beta)) + \nabla_\beta \log p(y; \lambda_0(x_i, \beta)) \nabla^T_\beta \log p(y; \lambda_0(x_i, \beta)) \right) \right] = 0
\]
where \( \text{vech} \) 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
\[
IM = \sum_i \left[ \frac{\nabla^2_\beta p(y; \lambda_0(x_i, \beta))}{p(y; \lambda_0(x_i, \beta))} \nabla_\beta \lambda_0(x_i, \beta) \nabla^T_\beta \lambda_0(x_i, \beta) + \frac{\nabla_\lambda p(y; \lambda_0(x_i, \beta))}{p(y; \lambda_0(x_i, \beta))} \nabla^2_\beta \lambda_0(x_i, \beta) \right]
\]

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:
\[
C \nabla_\beta \lambda_0(x_i, \beta) \nabla^T_\beta \lambda_0(x_i, \beta) = k(\lambda_0) k(\lambda_0)^T
\]
\[
\sum_i \frac{\nabla_\lambda p(y; \lambda_0(x_i, \beta))}{p(y; \lambda_0(x_i, \beta))} \nabla^2_\beta \lambda_0(x_i, \beta) = 0
\]
where \( C \) is a non-zero constant. We give several examples below as illustrations.

**Example 1.5.1.** Normal regression with \( Y_i \sim N(\mu_i, 1) \), where \( \mu_i = \mu_0 + \xi k(\mu_0) U_i \) and \( \mu_0 = x_i' \beta \). Note that \( \nabla_\beta \mu_0 \nabla^T_\beta \mu_0 = x_i x_i^T \) and \( \nabla^2_\beta \mu_0 = 0 \). Considering only the IM test based on the
intercept term, it is equivalent to the \( C(\alpha) \) test for heterogeneity in \( \beta_0 \) if \( k(\mu_0) = C \neq 0 \). If considering all elements in the IM test, the equivalence holds if \( x_i x_i^T = k(\mu_0) k(\mu_0)^T \). In this case,
the \( C(\alpha) \) test is multivariate, testing for homogeneity for all coefficients \( \beta \) in \( \mu_0 \).

**Example 1.5.2.** Poisson regression with \( Y_i \sim \text{Poi}(\lambda_i) \), where \( \lambda_i = \lambda_0 + \xi k(\lambda_0) U_i \) and \( \lambda_0 = \exp(x_i' \beta) \).
Considering only the IM test for the intercept term, we have \( \nabla_\beta \lambda_0 = \nabla^2_\beta \lambda_0 = \lambda_0 \). If \( \beta \)'s are replaced by their MLEs, the second identity for equivalence holds because the normal equation for
solving MLE of \( \beta_0 \) gives

\[
\sum_i \frac{\nabla \lambda p(y; \lambda_{0i})}{p(y; \lambda_{0i})} \nabla^2 \beta_0 \lambda_{0i} = \sum_i \frac{\nabla \lambda p(y; \lambda_{0i})}{p(y; \lambda_{0i})} \nabla \beta_0 \lambda_{0i} = 0
\]

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 \( \xi \), which leads to the second moment test for the Poisson regression model as discussed in Section 1.4.1.1. There are of course many other possible specifications for the conditional mean function of \( \lambda_{0i} \) which would lead to other equivalence conditions. We do not delve further into details here, but refer 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 specifications 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 \( \chi^2 \) asymptotics for the IM test. The simulation shows that overlooking the intrinsic one-sidedness of alternatives sacrifices power.

### 1.6 Simulation

We first revisit the Poisson regression model to illustrate the points made in Section 1.5.2. As discussed in Example 1.5.2 and also in Section 1.4.1, 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 \( \chi^2 \), and the information matrix test, using critical values from \( \chi^2 \) 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 \( 1.5 h \) with probability \( 2/3 \) and \( -3 h \) 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 1.2, 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 \( \chi^2 \) 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).

We then 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 1.2 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 have 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 \( \beta_i \) distributions. For 21 distinct equally spaced values of \( h \in [0, 1/3] \), the first example assumes \( \beta_i \sim N(0, h^2) \) and the second assumes \( \beta_i \) taking two possible values \( \{-h, 2h\} \) with probability \( (2/3, 1/3) \). Results are presented in Figure 1.3. 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.

### 1.7 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 the classical approach in Neyman (1959). The multi-dimensional extension suggests a modification on the usual \( \chi^2 \) 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
Figure 1.2: 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.

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 1.2 would be to incorporate density
Table 1.2: 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 $\epsilon_{it} \sim \text{IID} N(0, \sigma_i^2)$ and $\sigma_i^2 \sim U(1, 2)$. Both regressors are $N(0, 0.5^2)$. Under the null, $\beta_i = 1$ for all $i$ and under the alternative, $\beta_i \sim N(1, 0.15^2)$. The PY test is based on a two sized $N(0, 1)$ test, the IM test is based on $\chi^2_3$ test and the $C(\alpha)$ test on a mixture of $\chi^2$ test. All tests are conducted at 5\% nominal level with 5000 replications.

<table>
<thead>
<tr>
<th>T</th>
<th>N</th>
<th>Size</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>PY</td>
<td>IM</td>
</tr>
<tr>
<td>20</td>
<td>30</td>
<td>0.051</td>
<td>0.040</td>
</tr>
<tr>
<td>30</td>
<td>30</td>
<td>0.049</td>
<td>0.050</td>
</tr>
<tr>
<td>50</td>
<td>30</td>
<td>0.047</td>
<td>0.050</td>
</tr>
<tr>
<td>100</td>
<td>30</td>
<td>0.045</td>
<td>0.056</td>
</tr>
<tr>
<td>20</td>
<td>50</td>
<td>0.041</td>
<td>0.037</td>
</tr>
<tr>
<td>30</td>
<td>50</td>
<td>0.053</td>
<td>0.054</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>0.046</td>
<td>0.057</td>
</tr>
<tr>
<td>100</td>
<td>50</td>
<td>0.047</td>
<td>0.060</td>
</tr>
<tr>
<td>20</td>
<td>100</td>
<td>0.049</td>
<td>0.047</td>
</tr>
<tr>
<td>30</td>
<td>100</td>
<td>0.048</td>
<td>0.051</td>
</tr>
<tr>
<td>50</td>
<td>100</td>
<td>0.049</td>
<td>0.058</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>0.049</td>
<td>0.062</td>
</tr>
<tr>
<td>20</td>
<td>200</td>
<td>0.052</td>
<td>0.049</td>
</tr>
<tr>
<td>30</td>
<td>200</td>
<td>0.050</td>
<td>0.046</td>
</tr>
<tr>
<td>50</td>
<td>200</td>
<td>0.049</td>
<td>0.055</td>
</tr>
<tr>
<td>100</td>
<td>200</td>
<td>0.046</td>
<td>0.058</td>
</tr>
</tbody>
</table>

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 1.2.3. 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 $\xi_n$ converging to the null value $\xi_0$. 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.
Figure 1.3: 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 $\alpha_i \sim \mathcal{U}(0,1)$ and $\epsilon_{it} \sim \text{IID } \mathcal{N}(0, \sigma_i^2)$ and $\sigma_i^2 \sim \mathcal{U}(1,2)$. Both regressors are normal variable with mean zero and standard deviation 0.2 and 0.5 respectively. The solid line with circles is the power curve for the PY test, the crossed curve for the information matrix test with $\chi^2$ asymptotics and the curve with triangle signs for the $C(\alpha)$ test with mixture of $\chi^2$ asymptotics.

### 1.8 Proofs

Before proceeding to the proof for Theorem 1.2.4, 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 $v_\xi$ and $v_\theta$ be shorthand for $v_\xi(x_i)$ and $v_\theta(x_i)$. Let $\| \cdot \|$ be $L_2(\mu)$-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 $r_n(x_i, \xi_n, \theta_n) = f_n - f_0 - h_n^\top v(x_i)$ and denote $R_i = r_n(x_i, \xi_n, \theta_n)/f_0$.

**Lemma 1.8.1.** Under Assumption 1.2.1 and the modified DQM condition, we have the following:

1. $\sum_i R_i^2 = o_P(1)$
2. $\mathbb{E}(v(X)/f_0) = 0$
3. $2 \sum_i R_i = -\frac{1}{2} t^\top J t + o_P(1)$
4. $n^{-1/2} \sum_i R_i v_\xi/f_0 = o_P(1), \ n^{-1/2} \sum_i R_i v_\theta/f_0 = o_P(1)$
5. $\max_{1 \leq i \leq n} |R_i| = o_P(1)$
6. $\max_{1 \leq i \leq n} \left| \frac{2}{\sqrt{n}} \frac{v_\xi}{f_0} \right| = o_P(1), \ \max_{1 \leq i \leq n} \left| \frac{2}{\sqrt{n}} \frac{v_\theta}{f_0} \right| = o_P(1)$
Proof of (1). Under the modified DQM condition, the Markov inequality yields,
\[
\mathbb{P}(\sum_i R_i^2 > \epsilon) \leq e^{-2n\mathbb{E}(R_i^2)} = e^{-2n \int r_n^2(x; \xi_n, \theta_n) d\mu(x)} \to 0.
\]

Proof of (2) and (3). Since both \(f_n\) and \(f_0\) are objects with \(L_2(\mu)\)-norm 1
\[
0 = \|f_n\|_{\mu,2}^2 - \|f_0\|_{\mu,2}^2
= (\xi_n - \xi_0)^4 \|v_\xi\|_{\mu,2}^2 + (\theta_n - \theta)^T \|v_\theta\|_{\mu,2}^2 (\theta_n - \theta) + \|r_n\|_{\mu,2}^2 + 2 \langle (\theta_n - \theta)^T v_\theta, r_n \rangle
+ 2(\xi_n - \xi_0)^2 (\theta_n - \theta)^T \langle v_\theta, v_\xi \rangle + 2(\xi_n - \xi_0)^2 \langle v_\xi, r_n \rangle + 2(\xi_n - \xi_0)^2 \langle f_0, v_\xi \rangle
+ 2(\theta_n - \theta)^T \langle f_0, v_\theta \rangle + 2 \langle f_0, r_n \rangle
\]

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 \(v_\xi\) and \(v_\theta\) are square integrable with respect to measure \(\mu\) by assumption, \(\langle v_\xi, r_n \rangle = o(1/\sqrt{n})\) and \(\langle v_\theta, r_n \rangle = o(1/\sqrt{n})\). 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
\[
\langle f_0, v_\xi \rangle = \langle f_0, v_\theta \rangle = 0
\]
This proves (2) since \(0 = \langle f_0, v_\xi \rangle = \mathbb{E}(v_\xi(X)/f_0)\). Similar argument shows \(\mathbb{E}(v_\theta(X)/f_0) = 0\). Hence,
\[
2 \langle f_0, r_n \rangle = - (\xi_n - \xi_0)^4 \|v_\xi\|_{\mu,2}^2 - (\theta_n - \theta)^T \|v_\theta\|_{\mu,2}^2 (\theta_n - \theta)
- 2(\xi_n - \xi_0)^2 (\theta_n - \theta)^T \langle v_\theta, v_\xi \rangle + o(1/n)
= -\frac{1}{4n} t^T J t + o(1/n)
\]
with \(t^T = (\delta_1^2, \delta_2^2)\).

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
\[
2 \sum_i R_i = 2n \mathbb{E}(R_1) + o_P(1)
= 2n \langle f_0, r_n \rangle + o_P(1)
= 2n (\frac{1}{4n} t^T J t + o(1/n)) + o_P(1)
= -\frac{1}{4} t^T J t + o_P(1)
\]

Proof of (4). By Hölder’s inequality,
\[
\sum_i R_i \cdot \frac{2}{\sqrt{n}} f_0 v_\xi \leq \sqrt{\sum_i R_i^2 \sum_i \left(\frac{2}{\sqrt{n}} f_0 v_\xi\right)^2} = o_P(1)O_P(1) = o_P(1)
\]
Similar argument admits the second result.
Proof of (5).\[
\mathbb{P}( \max_{1 \leq i \leq n} |R_i| > \epsilon) \leq n \mathbb{P}( |R_1|^2 > \epsilon^2) \leq \epsilon^{-2} n \mathbb{E}(R_1^2) \rightarrow 0
\]

Proof of (6).\[
\mathbb{P}( \max_{1 \leq i \leq n} 2v_{\xi}/f_0 > \epsilon \sqrt{n}) \leq n \mathbb{P}( |2v_{\xi}/f_0| > \epsilon \sqrt{n}) \\
\leq \epsilon^{-2} \mathbb{E} \left( (2v_{\xi}(X_1)/f_0)^2 \right) \mathbb{I}_{|2v_{\xi}/f_0| > \epsilon \sqrt{n}} \rightarrow 0
\]

Similar argument admits the second statement. \(\square\)

Proof of Theorem 1.2.4. 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 1.2.1, we have the following Taylor expansion:
\[
f_n = f_0 + (\xi_n - \xi_0)^2 v_{\xi} + (\theta_n - \theta)^\top v_\theta + r_n(x_i; \xi_n, \theta_n).
\]

Denoting \(w_i = 2(f_n/f_0 - 1)\), we have
\[
w_i = 2(\xi_n - \xi_0)^2 v_{\xi}/f_0 + 2(\theta_n - \theta)^\top v_\theta + 2R_i.
\]

To show that under the modified DQM condition, the log-likelihood ratio admits a quadratic approximation, we use results in Lemma 1.8.1.

The log-likelihood ratio can be represented as
\[
\Lambda_n = \sum_i \log \frac{p(x_i; \xi_n, \theta_n)}{p(x_i; \xi_0, \theta)} = \sum_i 2 \log \frac{f_n}{f_0} = \sum_i 2 \log(1 + w_i/2) \\
= \sum_i w_i - \frac{1}{4} \sum_i w_i^2 + \frac{1}{2} \sum_i w_i^2 \beta(w_i)
\]

with \(\beta(x) \rightarrow 0\) as \(x \rightarrow 0\).

Using (3) in Lemma 1.8.1 and with \(S_n = (S_{\xi,n}, S_{\theta,n})^\top\) and \(J\) defined in Theorem 1.2.4, we have
\[
\sum_i w_i = 2 \frac{\delta_2}{\sqrt{n}} \sum_i \frac{v_{\xi}}{f_0} + 2 \frac{\delta_2}{\sqrt{n}} \sum_i \frac{v_\theta}{f_0} + 2 \sum_i R_i = t^\top S_n - \frac{1}{4} t^\top J t + o_p(1)
\]

Using (1) and (4) in Lemma 1.8.1, we have
\[
\sum_i w_i^2 = \sum_i \left( \frac{2\delta_2}{\sqrt{n}} \frac{v_{\xi}}{f_0} + \frac{2\delta_2}{\sqrt{n}} \frac{v_\theta}{f_0} + 2R_i \right)^2 \\
= t^\top J t + o_p(1) + 4 \sum_i R_i^2 + 4 \sum_i R_i \left( \frac{2\delta_2}{\sqrt{n}} \frac{v_{\xi}}{f_0} + \frac{2\delta_2}{\sqrt{n}} \frac{v_\theta}{f_0} \right) \\
= t^\top J t + o_p(1)
\]

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 1.8.1, we have
\begin{align*}
\mathbb{P}\left( \max_{1 \leq i \leq n} |w_i| > \epsilon \right) & \leq \delta_1^2 \mathbb{P}\left( \max_{1 \leq i \leq n} \frac{\sqrt{n} v_i}{\tau_0} > \epsilon \right) + \delta_2^2 \mathbb{P}\left( \max_{1 \leq i \leq n} \frac{\sqrt{n} v_0}{\tau_0} > \epsilon \right) \\
& \quad + 2 \mathbb{P}\left( \max_{1 \leq i \leq n} |R_i| > \epsilon \right) \to 0
\end{align*}

Since when \( w_i \to 0, \beta(w_i) \to 0 \), we have \( \max_{1 \leq i \leq n} |\beta(w_i)| = o_P(1) \). By Hölder’s inequality,

\[ \sum_i w_i^2 \beta(w_i) \leq \max_{1 \leq i \leq n} |\beta(w_i)| \sum_i w_i^2 = o_P(1)O_P(1) = o_P(1). \]

Therefore, the log-likelihood ratio is approximated by

\[ \Lambda_n = \sum_i w_i - \frac{1}{4} \sum_i w_i^2 + \frac{1}{2} \sum_i w_i^2 \beta(w_i) = t^T S_n - \frac{1}{4} t^T J t - \frac{1}{2} t^T J t + o_P(1) = t^T S_n - \frac{1}{2} t^T J t + o_P(1) \]

\( \square \)

**Proof of Corollary 1.2.5.** Since \( S_n \) is a normed iid sum, by the central limit theorem,

\[ S_n \xrightarrow{P_{n,\xi_0,}\theta} N(0, J) \]

The zero asymptotic mean of \( S_n \) is provided by (2) in Lemma 1.8.1, then the asymptotic variance for \( S_n \) is \( J \) as defined in Theorem 1.2.4.

The quadratic approximation for \( \Lambda_n \) established in Theorem 1.2.4 together with the joint normality of \( S_n \) leads to the LAN property of the sequence of model \( P_{n,\xi_n,\theta_n} \). Furthermore, we have

\[ \Lambda_n \xrightarrow{P_{n,\xi_0,}\theta} N\left(-\frac{1}{2} t^T J t, t^T J t\right). \]

By LeCam’s first lemma (see e.g. van der Vaart (1998, Lemma 6.4)), \( P_{n,\xi_n,\theta_n} \) and \( P_{n,\xi_0,\theta} \) are mutually contiguous.

\( \square \)

**Proof of Theorem 1.2.6.** The sequence of experiments \( \mathcal{E}_n \) converges to a shifted Gaussian \( N(t, J^{-1}) \) as a result of Theorem 9.4 in van der Vaart (1998). The log-likelihood ratio process of observing one sample from \( N(t, J^{-1}) \) is

\[ \log \frac{dN(t, J^{-1})}{dN(0, J^{-1})}(Y) = t^T J Y - \frac{1}{2} t^T J t \]

It suffices to show that \( J^{-1} S_n \) converges to the distribution of \( Y \) under the null. Corollary 1.2.5 establishes \( S_n \xrightarrow{P_{n,\xi_0,}\theta} N(0, J) \), we thus have \( J^{-1} S_n \xrightarrow{P_{n,\xi_0,}\theta} N(0, J^{-1}) \).

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,

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

Notice the rescaling in $Z_n$ is needed to obtain a unit asymptotic variance for the test statistic. \hfill \Box

**Proof of Corollary 1.2.7.** Since $\xi$ is a scalar and $S_n \overset{p}{\to} N(0, I)$ under $H_0$, it is immediate that the asymptotic null distribution for $Z_n$ is $N(0, 1)$.

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 $\theta$ is left unspecified as in the null, hence we set $\delta_2 = 0$ in the log-likelihood ratio expansion. Under $H_0$,

$$(Z_n, A_n) \overset{p}{\to} \delta_{\theta} \mathcal{N}\left(\begin{pmatrix} 0 \\ -\frac{1}{2} \delta_1^2 J_{\xi \xi} \end{pmatrix}, \begin{pmatrix} 1 & \sigma_{12} \\ \sigma_{12} & \delta_1^2 J_{\xi \xi} \end{pmatrix}\right)$$

with $\sigma_{12} = \text{Cov}(Z_n, A_n) = \delta_1^2 (J_{\xi \xi} - J_{\xi \theta} J_{\theta \xi}^{-1} J_{\theta \theta})^{1/2}$. With $\delta_2 = 0$, Corollary 1.2.5 implies that $P_n, \xi_n, \theta$ are mutually contiguous to $P_n, \xi_0, \theta$, then LeCam’s third lemma implies,

$$Z_n \overset{p}{\to} \mathcal{N}(\sigma_{12}, 1).$$

\hfill \Box

**Proof of Theorem 1.2.9.** Define the class of functions:

$$F_n := \left\{ x \mapsto (g(x, \theta) - g(x, \eta)) \mid \|\theta - \eta\| \leq \delta_n \right\}.$$

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

$$\left| Z_n(\hat{\theta}) - Z_n(\theta) \right| \leq \sup_{f \in F_n} |G_n(f)|$$

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

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

$$\|\eta - \eta N_n\| \leq \varepsilon^{1/\gamma}$$

35
The condition on $g$ in Assumption 1.2.8 implies

$$|g(x, \eta) - g(x, \eta N_n)| \leq \|\eta - \eta N_n\|^{\gamma} H(x) \leq \epsilon H(x)$$

It follows that the bracketing number, $N_1(\|\cdot\|_{2, \mathcal{F}_n, \mathcal{L}_2(P_n, \xi_n, \theta)})$ is bounded from above by $C_2 e^{-p/\gamma}$.

Furthermore, the assumption also implies that for $f \in \mathcal{F}_n$, $\|f\|_{p_n, 2} \leq \delta_n^{1/2} \|H\|_{p_n, 2}$ 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

$$\mathbb{E}_{P_{n, \xi_n, \theta}} \left( \sup_{f \in \mathcal{F}_n} |G_n(f)| \right) \leq \int_{\mathcal{F}_n} (\delta_n^{1/2} \|H\|_{p_n, 2} + \sqrt{n} \mathbb{E}_{P_{n, \xi_n, \theta}} [H(X) I[H(X) > \sqrt{n} \alpha(\delta_n^\gamma)]])$$

where the bracketing integral is defined as

$$J_{\mathcal{F}_n}(\delta_n^{1/2}, \mathcal{F}_n, \mathcal{L}_2(P_n, \xi_n, \theta)) = \int_0^{\delta_n^{1/2}} \sqrt{1 + \log N_1(\|\cdot\|_{p_n, 2, \mathcal{F}_n, \mathcal{L}_2(P_n, \xi_n, \theta)} )} d\epsilon$$

and

$$\alpha(\delta_n^\gamma) = \delta_n^{1/2} \|H\|_{p_n, 2} / \sqrt{1 + \log N_1(\|\cdot\|_{p_n, 2, \mathcal{F}_n, \mathcal{L}_2(P_n, \xi_n, \theta)} )}.$$

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

$$J_{\mathcal{F}_n}(\delta_n^{1/2}, \mathcal{F}_n, \mathcal{L}_2(P_n, \xi_n, \theta)) \leq \int_0^{\delta_n^{1/2}} 1 + \log(C_2 e^{-p/\gamma}) d\epsilon \to 0$$

Since $H(x)$ is square integrable for all $n$ by Assumption 1.2.8, the first term goes to zero.

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

$$\alpha(\delta_n^\gamma) \geq \frac{\delta_n^{1/2} \|H\|_{p_n, 2}}{\sqrt{1 + \log(C_2 \delta_n^{-p})}} := k_n \to 0$$

As long as $k_n$ converges to zero slower than $c_n$, Assumption 1.2.8 ensures that the second term also tends to zero.

The last step is to check that $\sup_{f \in \mathcal{F}_n} \frac{1}{\sqrt{n}} \sum_{i=1}^n |G_{P_{n, \xi_n, \theta}}(f(X_i))| = o(1)$ so that $\sup_{f \in \mathcal{F}_n} |G_n(f)|$ is the correct upper bound. This is trivially true under the null, where $\xi_n = \xi_0$ for all $n \in \mathbb{N}$, since $\mathbb{E}_{P_{n, \xi_0, \theta}}(g(X_i, \theta)) = \mathbb{E}_{P_{n, \xi_0, \theta}}(g(X_i, \tilde{\theta})) = 0$. 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

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

36
Denote \( p_n = p(x; \xi_n, \theta) \) and \( p_0 = p(x; \xi_0, \theta) \), we have the following expansion

\[
\sqrt{n} \int (g(x, \eta) - g(x, \theta)) p_n(\eta) \, d\eta = \sqrt{n} \int \left( (g(x, \eta) - g(x, \theta))(\sqrt{p_0} + (\xi_n - \xi_0)^2 v_\xi(x) + r_n) \right) \sqrt{p_n} \, d\eta
\]

The last two terms are \( o(1) \) uniformly over \( \eta \) for \( \|\eta - \theta\| \leq \delta_n \) due to the DQM condition in Definition 1.2.3 and assumption on \( g \) in Assumption 1.2.8. Since Cauchy-Schwarz inequality implies that with respect to \( L_2(\mu) \)-norm,

\[
|\int (g(x, \eta) - g(x, \theta)) \sqrt{p_n} v_\xi(x) \, d\eta| \leq \| (g(x, \eta) - g(x, \theta)) \sqrt{p_0} \|_{\mu, 2} \| v_\xi \|_{\mu, 2} \leq \|\eta - \theta\| \gamma H \| p_n, 2 \| \| v_\xi \|_{\mu, 2} = o(1).
\]

Similarly,

\[
|\sqrt{n} \int (g(x, \eta) - g(x, \theta)) \sqrt{p_n} r_n \, d\eta| \leq \| (g(x, \eta) - g(x, \theta)) \sqrt{p_0} \|_{\mu, 2} \| r_n \|_{\mu, 2} = o(1).
\]

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 1.3.1.** As in the proof of Theorem 1.2.6, the limit of the sequence \( v_n \) is a shifted Gaussian experiment \( Y \sim N(t, J^{-1}) \) but now with \( t^T = (\delta_1^2, \delta_2^2, 2\delta_1\delta_2, \delta_3^2) \). An equivalent limit experiment observes \( X \sim N(Jt, J) \) with \( X = JY \), because the likelihood ratio process of \( dN(t, J^{-1}) \) is identical to that of \( dN(t^T J^{-1} J) \).

To be more explicit, denoting the first three elements of \( X \) to be \( X_\xi \), and the rest to be \( X_\theta \), we have under the alternative,

\[
\begin{pmatrix} X_\xi \\ X_\theta \end{pmatrix} \overset{d}{=} N \left( \begin{pmatrix} J_{\xi \xi} & J_{\xi \theta} \\ J_{\theta \xi} & J_{\theta \theta} \end{pmatrix} \begin{pmatrix} t_\xi \\ t_\theta \end{pmatrix}, J \right)
\]

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

To focus on testing for zero restrictions on \( t_\xi \), we find the conditional distribution of \( X_\xi \) on \( X_\theta \) to be

\[
\tilde{X}_\xi = X_\xi - J_{\xi \theta} J_{\theta \theta}^{-1} X_\theta \overset{d}{=} N((J_{\xi \xi} - J_{\xi \theta} J_{\theta \theta}^{-1} J_{\theta \xi}) t_\xi, J_{\xi \xi} - J_{\xi \theta} J_{\theta \theta}^{-1} J_{\theta \xi}).
\]

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 $N(0, \Sigma)$ with $\Sigma = J_{\xi\xi} - J_{\xi\theta} (J_{\theta\theta}^{-1}) J_{\theta\xi}$, and under local alternative, its asymptotic distribution is $N(S^T, \Sigma)$. Notice we can decompose $\tilde{S}_{\xi,n} \Sigma^{-1} \tilde{S}_{\xi,n}$ into two independent pieces as $u_n^T \Sigma_{11.2} u_n + w_{3n}^T w_{3n}$. Let the Cholesky decomposition of $\Sigma_{11.2}$ be such that $\Lambda^T \delta = \Sigma_{11.2}^T \eta$ and $w_n \sim \mathcal{N}(\Delta^T \eta, \Sigma)$. Since $(\delta_1^2, \delta_2^2) \in \mathbb{R}_+^2$ and $(\eta_1, \eta_2) = \Lambda^T (\delta_1^2, \delta_2^2)$, the feasible parameter set is therefore the convex cone defined as,

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

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 $\odot$, the test statistics is the sum of squares of the elements of $w_n$ and $w_{3n}$ which are mutually independent:

$$T_n = w_{1n}^2 + w_{2n}^2 + w_{3n}^2 \sim \chi^2_3$$

**Case 2:** When the test statistic falls into area $\odot$, we need to project $w_n$ onto the convex cone $\odot$, 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
\[ T_n = (\rho^2 w_{1n} + \rho \sqrt{1 - \rho^2} w_{2n})^2 + (\rho \sqrt{1 - \rho^2} w_{1n} + (1 - \rho^2) w_{2n})^2 + w_{3n}^2 \]

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,
\[ T_n = 0 + w_{3n}^2 \sim \chi^2_1 \]

The asymptotic distribution of the \( C(\alpha) \) test statistics is a mixture of \( \chi^2 \)'s, for which the weights are characterized by the probability of falling into different regions. The angle \( \beta \) 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
\[ \left( \frac{1}{2} - \frac{\beta}{2\pi} \right) \]

Analytical Derivation for Section 1.4.3. The information matrix for \( (\xi, \theta) = (\xi_1, \xi_2, \mu_0, \sigma_0^2) \) is
\[
I = \begin{pmatrix}
I_{\xi\xi} & I_{\xi\theta} \\
I_{\theta\xi} & I_{\theta\theta}
\end{pmatrix} = \frac{N T}{\sigma_0^4} \begin{pmatrix}
2T & \sigma_0^2 & 0 & 1 \\
\sigma_0^2 & (T + 3)\sigma_0^4/2 & 0 & \sigma_0^2/2 \\
0 & 0 & \sigma_0^2 & 0 \\
1 & \sigma_0^2/2 & 0 & 1/2
\end{pmatrix}
\]

We further find
\[
I_{\xi,\theta} = I_{\xi,\xi} - I_{\xi,\theta} I^{-1}_{\theta,\theta} I_{\theta,\xi} = \begin{pmatrix}
2NT(T - 1)/\sigma_0^4 & 0 \\
0 & NT(T/2 + 1)
\end{pmatrix}
\]

and
\[
I_{\xi,\theta} I^{-1}_{\theta,\theta} = \begin{pmatrix}
0 & 2 \\
0 & \sigma_0^2
\end{pmatrix}
\]

As we have remarked in Section 1.3, the diagonality of \( I_{\xi,\theta} \) provides much convenience for finding the optimal test statistics. Denote
\[
T_n := \begin{pmatrix} t_{1n} \\ t_{2n} \end{pmatrix} = I_{\xi,\theta}^{-1/2} \begin{pmatrix}
\sum_i v_{1i} - 2 \sum_i v_{4i} \\
\sum_i v_{2i} - \sigma_0^2 \sum_i v_{4i}
\end{pmatrix} = \begin{pmatrix}
(2NT(T - 1)/\sigma_0^4)^{-1/2} \left( \sum_i (\bar{y}_i - \mu_0)^2 - NT/\sigma_0^2 \right) \\
(NT(T/2 + 1))^{-1/2} \left( \sum_i (Z_i - T/2)^2 - NT/2 \right)
\end{pmatrix}
\]

Replacing \( (\mu_0, \sigma_0^2) \) by their MLEs yields the joint \( C(\alpha) \) test.
Claim in Section 1.5. Here we provide the detail derivation for the claim in Section 1.5 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 1.8.2.** 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 \( \eta \) evaluated under \( \eta = 0 \) is unbounded unless \( E(U^3) = 0 \) and \( E(U^4) < \infty \).

**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 \( \eta \) is

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

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

The second order derivative is

\[
\nabla^2_\eta l|_{\eta=0} = \frac{\tau^2 \sqrt{\pi} \int \nabla^3_\lambda p(y; \lambda_0) u^2 dF(u) \tau \int \nabla_\lambda^2 p(y; \lambda_0) u dF(u)}{4 \eta \int p(y; \lambda_0) dF(u)} \bigg|_{\eta=0} - \left( \nabla_\eta l|_{\eta=0} \right)^2
\]

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

\[
\nabla^2_\eta l|_{\eta=0} = \frac{\tau^4}{12} \left[ \frac{\int \nabla^4_\lambda p(y; \lambda_0) p(y; \lambda_0) dF(u)}{p(y; \lambda_0)} - 3 \int \left( \frac{\nabla_\lambda^2 p(y; \lambda_0)}{p(y; \lambda_0)} \right)^2 \frac{\nabla^2_\lambda^4 p(y; \lambda_0)}{p(y; \lambda_0)} \right] < \infty
\]

\( \square \)
Chapter 2

Likelihood Ratio Test for Mixture Models

2.1 Introduction

Given a simple parametric density model, \( p(x|\mu) \), for iid observations, \( X_1, \ldots, X_n \), there is a natural temptation to complicate the model by allowing the parameter, \( \mu \), to vary with \( i \). In the absence of other information (e.g. observable covariate) that would distinguish the observations from one another it may be justifiable to view the \( \mu \)'s as drawn at random. Inference for such mixture models is complicated by a variety of problems, notably their lack of identifiability. Two dominant approaches exist: Neyman’s \( C(\alpha) \) test (discussed in Chapter 1) and the likelihood ratio test (LRT) that will be discussed in this chapter. The \( C(\alpha) \) test is particularly attractive for testing homogeneity against general forms of heterogeneity for the parameter \( \mu \). Such tests have a somewhat irregular but still relatively simple asymptotic theory, and are generally easy to compute. The LRT, in contrast, has a much more complicated limiting behavior, and is generally more difficult to compute.

We will argue that recent developments in convex optimization have dramatically reduced the computational burden of the LRT approach for general, nonparametric alternatives. Following Laird (1978), prior efforts to compute the Kiefer-Wolfowitz MLE for general nonparametric mixture models have employed some variant of the EM algorithm. However, Koenker and Mizera (2014) have recently shown that interior point methods for general convex optimization provide a much more efficient, more accurate computational approach. A second impediment to the use of LRT methods for general mixture problems has been the lack of a tractable limiting distribution theory. Extending recent work of Gassiat (2002), Liu and Shao (2003) and Azaïs, Gassiat, and Mercadier (2009), we develop below an easily simulated method of computing limiting critical values for the LRT statistic for testing homogeneity in general nonparametric mixture models. However as we find in simulations, these limiting critical values do not serve as a good approximation in moderate samples. We then propose a parametric bootstrap method to determine critical values. We formally prove the consistency of the bootstrap method and illustrate its size and power performance in simulations.

Together, these new developments provide a unified framework for likelihood ratio testing for mixtures. We also briefly review the theory of \( C(\alpha) \) testing for homogeneity developed in Chapter 1 and the EM test recently proposed by Chen and Li (2009) and Li, Chen, and Marriott (2009).

This chapter is based on joint work with Roger Koenker and Stanislav Volgushev.
The EM test has a connection to the C(α) test and shares the same limiting null distribution. Simulation comparisons of the LRT with C(α) and the EM test demonstrate that the LRT can be a highly effective complementary approach.

2.2 Likelihood ratio tests for mixture models

Before discussing the likelihood ratio test for general mixture models, our first concern is to obtain a reliable maximum likelihood estimator for these models that leads to a viable maximum likelihood evaluation for computing the LR test statistics. Lindsay (1995) offers a comprehensive overview of the vast literature on mixture models. He traces the idea of maximum likelihood estimation of a nonparametric mixing distribution \( F \), given random samples from the mixture density,

\[
g(x) = \int p(x|\mu) dF(\mu),
\]

(2.2.1)
to Robbins (1950). Kiefer and Wolfowitz (1956) established the consistency property of the maximum likelihood estimator of \( F \) (hereafter, Kiefer-Wolfowitz MLE) and yet only with Laird (1978) did a viable computational strategy emerge for it. The EM method proposed by Laird has been employed extensively in subsequent work, e.g. Heckman and Singer (1984) and Jiang and Zhang (2009), even though it has been widely criticized for its slow convergence. Recently Koenker and Mizera (2014) have noted that the Kiefer-Wolfowitz estimator can be formulated as a convex optimization problem and solved very efficiently by interior point methods. Recent work by Gassiat (2002), Liu and Shao (2003) and Azaïs, Gassiat, and Mercadier (2009) has clarified the limiting behavior of the LRT for general class of alternatives, and taken together these developments offer a fresh opportunity to explore the viability of the LRT for inference on mixtures.

It seems ironic that many of the difficulties inherent in maximum likelihood estimation of finite parameter mixture models vanish when we consider nonparametric mixtures. The notorious multimodality of parametric likelihood surfaces is replaced by a much simpler, strictly convex optimization problem possessing a unique, unimodal solution. It is of obvious concern that consideration of such a wide class of alternatives may depress the power of associated tests; we will see that while there is some loss of power when compared to more restricted parametric LRTs, the loss is typically modest, a small price to pay for power against a broader class of alternatives. We will see that comparing with C(α) tests that are also designed to detect general alternatives, the LRT can be competitive.

2.2.1 Maximum likelihood estimation of general mixtures

Suppose that we have iid observations, \( X_1, \cdots, X_n \) from the mixture density (2.2.1), the Kiefer-Wolfowitz MLE requires us to solve,

\[
\min_{F \in S} \left\{ - \sum_{i=1}^{n} \log g(x_i) \right\},
\]
where $\mathcal{G}$ is the (convex) set of all mixing distributions. The problem is one of minimizing the sum of convex functions subject to linear equality and inequality constraints. The dual to this (primal) convex program proves to be somewhat more tractable from a computational viewpoint, and takes the form,

$$
\max_{\nu \in \mathbb{R}^n} \left\{ \sum_{i=1}^{n} \log \nu_i \mid \sum_{i=1}^{n} \nu_i p(x_i|\mu) \leq n, \text{ for all } \mu \right\}
$$

See Lindsay (1983) and Koenker and Mizera (2014) for further details. This variational form of the problem may still seem rather abstract since it appears – even in the dual – that we need to check an infinite number of values of $\mu$, for each choice of the vector, $\nu$. However, it suffices in applications to consider a fine grid of values $\{\mu_1, \cdots, \mu_m\}$ and write the primal problem as

$$
\min_{f \in \mathbb{R}^m, g \in \mathbb{R}^n} \left\{ -\sum_{i=1}^{n} \log (g_i) \mid Af = g, f \in S \right\}
$$

where $A$ is an $n$ by $m$ matrix with elements $p(x_i|\mu_j)$ and $S = \{s \in \mathbb{R}^m|1^\top s = 1, \ s \geq 0\}$ is the unit simplex. Thus, $\hat{f}_j$ denotes the estimated mixing density evaluated at the grid point, $\mu_j$ and $\hat{g}_i$ denotes the estimated mixture density evaluated at $x_i$. The dual problem in this discrete formulation becomes,

$$
\max_{\nu \in \mathbb{R}^n} \left\{ \sum_{i=1}^{n} \log \nu_i \mid A^\top \nu \leq n1_m, \ \nu \geq 0 \right\}.
$$

Primal and dual solutions are immediately recoverable from the solution to either problem. Interior point methods such as those provided by PDCO of Saunders (2003) and Mosek of Andersen (2010), are capable of solving dual formulations of typical problems with $n = 200$ and $m = 300$ in less than one second. The R empirical Bayes package REBayes, Koenker (2013), is available for download from CRAN. It is based on the RMosek package of Friberg (2012), and was used for all of the computations reported below. We have compared this approach with other proposals including those of Lesperance and Kalbfleisch (1992) and Groeneboom, Jongbloed, and Wellner (2008), but thus far have found nothing competitive in terms of speed and accuracy.

Solutions to the nonparametric MLE problem of Kiefer and Wolfowitz produce estimates of the mixing distribution, $F$, that are discrete and possessing only a few mass points. A theoretical upper bound on the number of these atoms of $n$ was established already by Lindsay (1983), but in practice the number is actually observed to be far fewer. It may seem surprising, perhaps even disturbing, that even when the true mixing distribution has a smooth density, our estimates of that density is discrete with only a handful of atoms. This may appear less worrying if we consider a more explicit example. Suppose that we have a location mixture of Gaussians,

$$
g(x) = \int \phi(x - \mu) dF(\mu),
$$

so we are firmly in the deconvolution business, a harsh environment notorious for its poor convergence rates. One interpretation of this is that good approximations of the mixture density $g$ can
be achieved by relatively simple discrete mixtures with only a few atoms. For many applications estimation of \( g \) is known to be sufficient: this is quite explicit for example for empirical Bayes compound decision problems where Bayes rules depend entirely on the estimated \( \hat{g} \). Of course given our discrete formulation of the Kiefer-Wolfowitz problem, we can only identify the location of atoms up to the scale of the grid spacing, but we believe that the \( m \approx 300 \) grid points we have been using in the simulations reported below are probably adequate for most applications. For testing this assertion is supported by the fact that finer grids, when employed, exert a negligible impact on the LR statistic. Recently, Dicker and Zhao (2014) has shown that with \( m = \sqrt{n} \), the Hellinger distance between \( \hat{g} \) and \( g \) is bounded by \( O_p(\log n / \sqrt{n}) \).

Given a reliable maximum likelihood estimator for the general nonparametric mixture model it is of obvious interest to know whether an effective likelihood ratio testing strategy can be developed. This question has received considerable prior attention, again Lindsay (1995) provides an authoritative overview of this literature. However, more recently work by Gassiat (2002), Liu and Shao (2003) and Azaïs, Gassiat, and Mercadier (2009) have revealed new features of the asymptotic behavior of the likelihood ratio for mixture settings that enable one to derive asymptotic critical values for the LRT.

### 2.2.2 Asymptotic theory of likelihood ratios for general mixtures

Consider a parametric family of distributions that have density \( p(\cdot | \mu) \) with respect to some sigma-finite measure \( \lambda \) and parameters from the parameter set \( \Theta \). Our aim is to test whether the i.i.d. sample \( X_1, \ldots, X_n \) is generated from the distribution with density \( p(\cdot | \mu_0) \) for some \( \mu_0 \in \Theta \) against the general alternative that \( X_1, \ldots, X_n \) is generated from a mixture of the form \( p(\cdot) := \int_{\Theta} p(\cdot | \mu) d\eta(\mu) \) for some non-degenerate distribution \( \eta \) on \( \Theta \) (non-degenerate in the sense that \( \eta \) is not a one-point distribution). Consider the following set of distributions on \( \Theta \)

\[
\mathcal{G} := \{ \eta | \eta \text{ distribution on } \Theta \}
\]

Denote by \( \delta(\mu) \) the Dirac measure in the point \( \mu \). Define the log-likelihood function corresponding to the measure \( \eta \) as

\[
\ell_n(\eta) := \sum_{i=1}^{n} \log p_\eta(X_i). 
\]

The likelihood ratio test statistic is given by

\[
L_n := \sup_{\eta} \ell_n(\eta) - \sup_{\mu \in \Theta} \ell_n(\delta(\mu))
\]

where the first supremum is taken over all probability distributions \( \eta \) on \( \Theta \). To derive the asymptotic distribution of the likelihood ratio under the null, assume that the data are generated from a
measure with density $p(\cdot | \mu_0)$ for some $\mu_0 \in \Theta$. Consider the decomposition

$$L_n = \sup_{\eta} \ell_n(\eta) - \ell_n(\delta(\mu_0)) + \ell_n(\delta(\mu_0)) - \sup_{\mu \in \Theta} \ell_n(\delta(\mu)).$$

The second term in this decomposition can be handled by classical parametric theory. Under suitable regularity conditions we obtain

$$\sup_{\mu \in \Theta} \ell_n(\delta(\mu)) - \ell_n(\delta(\mu_0)) = \frac{1}{2} \left( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I(\mu_0)^{-1/2} \ell'(X_i|\mu_0) \left\| \right\| + o_P(1) \tag{2.2.2} \right)
$$

with $I(\mu)$ being the Fisher information matrix and $\ell'(X_i|\mu) := \nabla_\mu \log p_{\delta(\mu)}(X_i)$. Handling the first part in the decomposition of $L_n$ is more challenging. Expansions for this term can be derived under various sets of conditions (see for example Gassiat (2002), Liu and Shao (2003) and Azaïs, Gassiat, and Mercadier (2009)). For the sake of a simple presentation we will follow Gassiat (2002).

For $\eta \in \mathcal{G}, \mu \in \Theta, \eta \neq \delta(\mu)$ let

$$s_{\eta,\mu}(x) := \left( \frac{p_{\eta}(x)}{p_{\delta(\mu)}(x)} - 1 \right) / \left\| \frac{p_{\eta}}{p_{\delta(\mu)}} - 1 \right\|_{2, \delta(\mu)}$$

where we defined $\|f\|_{2,\eta} := (\int \int f^2(x)p(x|z)dzp(z)dv(x))^{1/2}$. Also, define the class of functions

$$\mathcal{F} := \{x \mapsto s_{\eta,\mu}(x) \; | \; \eta \in \mathcal{G}, \eta \neq \delta(\mu_0)\}.$$

For $\eta \in \mathcal{G}, \eta \neq \delta(\mu_0)$ define

$$G_n(\eta) := n^{-1/2} \sum_{i=1}^{n} s_{\eta,\mu_0}(X_i)$$

and note that by construction $E[s_{\eta,\mu_0}(X_i)] = 0$. Now a slight modification of the proof of Theorem 3.1 in Gassiat (2002) leads to the following result for the asymptotic behavior of the likelihood ratio test.

**Theorem 2.2.1.** Assume $X_1, ..., X_n$ are generated from $p(\cdot | \mu_0)$ and that $G_n \sim \mathcal{G}$ in $\ell^\infty(\mathcal{G} \setminus \delta(\mu_0))$ for a centered Gaussian process $\mathcal{G}$ with covariance structure that takes the form $E[G(\eta)G(\nu)] = E[s_{\eta,\mu_0}(X_i)s_{\eta,\mu_0}(X_i)]$. Then

$$2 \left( \sup_{\eta} \ell_n(\eta) - \ell_n(\delta(\mu_0)) \right) = \sup_{\eta \in \mathcal{G}} \left( \max \left\{ G_n(\eta), 0 \right\} \right)^2 + o_p(1).$$

If additionally (2.2.2) holds and $\ell'(X_i|\mu_0)$ is square integrable,

$$2L_n \sim \sup_{\eta \in \mathcal{G}} \left( \max \left\{ G(\eta), 0 \right\} \right)^2 - Y_1^2.$$

Here $Y_1 \sim \mathcal{N}(0, 1)$ and $(\mathcal{G}, Y_1)$ is jointly normal with covariance taking the form $E[\mathcal{G}(\eta)Y_1] = 0$.
\( \mathbb{E}[s_{\eta \mu}(X) I(\mu)^{-1/2} \ell'(X|\mu)] \). Here, by jointly normal we mean that for any collection \( \eta_1, \ldots, \eta_k \in \mathcal{G} \) the random vector \( (Y_1, G(\eta_1), \ldots, G(\eta_k)) \) follows a centered multivariate normal distribution with the covariance described above.

### 2.2.3 Asymptotic critical values

In order to apply the above limiting result in practice, we need to know how to obtain critical values from the asymptotic distribution. For the purpose of illustration, we consider the following normal mixture example.

**Example 2.2.2.** Consider mixtures of \( N(\mu, 1) \) distributions and assume that \( M = [L, U] \) with \( 0 \in M \) is the support for the location parameter. According to the above result, computations in Azaïs, Gassiat, and Mercadier (2009) show that the asymptotic distribution of the log-likelihood ratio test statistic \( L_n \) under the null of \( X_i \sim N(0, 1) \) i.i.d. is given by

\[
D = \left( \sup_{\eta \in \mathcal{G}} (V_{\eta})_+ \right)^2 - Y_1^2
\]

where \( (V_{\eta})_{\eta \in \mathcal{G}} \) is the Gaussian process given by

\[
V_{\eta} := \left( \sum_{k=1}^{\infty} \frac{Y_k \kappa_k(\eta)}{(k!)^{1/2}} \right) \left( \sum_{k=1}^{\infty} \frac{\kappa_k^2(\eta)}{k!} \right)^{1/2}
\]

with \( Y_1, Y_2, \ldots \) denoting i.i.d. \( N(0, 1) \) distributed random variables, \( \kappa_k(\eta) := \int_M \mu^k \eta(\mu) d\mu \) and \( x_+ \) denoting the positive part of \( x \).

There exists a simpler expression for the distribution of \( D \). More precisely, we will demonstrate that

\[
D \overset{D}{=} \sup_{\eta \in \mathcal{G}} \left( \left( \sum_{k=2}^{\infty} \frac{Y_k \kappa_k(\eta)}{(k!)^{1/2}} \right)^2 / \sum_{k=2}^{\infty} \frac{\kappa_k^2(\eta)}{k!} \right).
\]

(2.2.3)

The detailed derivation is provided in Section 2.8. Approximating the distribution function \( \eta \) on \( M \) by a discrete distribution function with masses \( p_1, \ldots, p_N \) on a fine grid \( m_1, \ldots, m_N \) leads to the approximation

\[
D \approx \sup_{p_1, \ldots, p_N} \left( \left( \sum_{j=1}^{N} p_j \sum_{k=2}^{\infty} \frac{Y_k m_{j}^k}{(k!)^{1/2}} \right)_+ \right)^2 / \sum_{i,j=1}^{N} p_i p_j \sum_{k=2}^{\infty} \frac{(m_i m_j)^k}{k!}.
\]

In particular, maximizing the right-hand side with respect to \( p_1, \ldots, p_N \) under the constraints \( p_i \geq 0, \sum p_i = 1 \) for fixed grid \( m_1, \ldots, m_N \) can be formulated as a quadratic optimization problem of the form

\[
\min_p p^T A p \quad \text{under} \quad p_i \geq 0, \ p^T b = 1
\]
where \( p = (p_1, \ldots, p_N) \), \( A_{ij} = \sum_{k=2}^{\infty} \frac{(m_1 m_i)^k}{k!}, \quad b_i = \sum_{k=2}^{\infty} \frac{Y_k m^k_i}{i!}, \) if \( \max b_i > 0 \). If \( \max b_i \leq 0 \), we can set \( D = 0 \). This suggests a practical way of simulating critical values after replacing the infinite sum by a finite approximation and avoiding the grid point 0. Table 2.1 below contains simulated critical values in some particular settings. All results are based on 10,000 simulation runs with the sums for \( A \) and \( b \) cut off at \( k = 25 \) and grids with 200 points equally spaced points excluding the point 0.

<table>
<thead>
<tr>
<th>M</th>
<th>90%</th>
<th>95%</th>
<th>99%</th>
</tr>
</thead>
<tbody>
<tr>
<td>[-1,1]</td>
<td>2.75</td>
<td>3.95</td>
<td>6.93</td>
</tr>
<tr>
<td>[-2,2]</td>
<td>3.90</td>
<td>5.37</td>
<td>8.71</td>
</tr>
<tr>
<td>[-3,3]</td>
<td>5.34</td>
<td>6.87</td>
<td>10.46</td>
</tr>
<tr>
<td>[-4,4]</td>
<td>6.38</td>
<td>8.32</td>
<td>11.91</td>
</tr>
</tbody>
</table>

Table 2.1: Simulated asymptotic critical values for the asymptotic null distribution for various sets \( M \).

To explore the finite sample performance of the above method we begin with an experiment to compare the critical values of the LRT of homogeneity in the Gaussian location model with the simulated asymptotic critical values in Table 2.1. We consider sample sizes \( n \) taking values in \( \{100, 500, 1000, 5000, 10000\} \) and four choices of the domain of the MLE of the mixture is estimated: \( \{[-j,j] : j = 1, \ldots, 4\} \). We maintain a grid spacing of 0.01 for the mixing distribution on these domains for each of these cases for the Kiefer-Wolfowitz MLE. Results are reported in Table 2.2. For the three largest sample sizes we bin the observations into 300 and 500 equally spaced bins respectively. It is noted that the empirical critical values are consistently smaller than those simulated from the asymptotic theory. There appears to be a tendency for the empirical critical values to increase with \( n \), but this tendency is rather weak. This finding is perhaps not entirely surprising in view of the slow rates of convergence established elsewhere in the literature, see e.g. Bickel and Chernoff (1993) and Hall and Stewart (2005). This implies that these simulated critical values are not likely to work well in size control, which motivates us to consider an alternative bootstrap based method in determining critical values in the next section.

### 2.3 Parametric bootstrap method for critical values

The parametric bootstrap method for testing parameter homogeneity we are about to introduce is not a new idea. In finite mixture models, it has been proposed by McLachlan (1987) and Chen and Chen (2001). However, to the best of our knowledge, this is the first time that the bootstrap method is formally shown to be consistent for LRT with general mixture models.

The parametric bootstrap approach to determine critical values for the distribution of \( L_n \) is defined as follows.

1. Compute the maximum likelihood estimator \( \hat{\mu} := \arg\max_{\mu \in \Theta} L_n(\delta(\mu)) \).
Table 2.2: Critical values for likelihood ratio test of Gaussian parameter homogeneity: The first five rows of the table report empirical critical values based on 1000 replications of the LRT using the Kiefer-Wolfowitz estimate of the nonparametric Gaussian location mixture distribution. Results for sample sizes 5,000 and 10,000 were computed by binning the observations into 300, 500 equally spaced bins respectively. Restriction of the domain of the mixing distribution is indicated by the column labels. The last row reproduces the simulated asymptotic critical values reported in Table 2.1.

<table>
<thead>
<tr>
<th>n</th>
<th>cval(.90)</th>
<th>cval(.95)</th>
<th>cval(.99)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[-1,1]</td>
<td>[-2,2]</td>
<td>[-3,3]</td>
</tr>
<tr>
<td>100</td>
<td>2.09</td>
<td>2.69</td>
<td>2.80</td>
</tr>
<tr>
<td>500</td>
<td>2.22</td>
<td>2.69</td>
<td>2.80</td>
</tr>
<tr>
<td>1,000</td>
<td>2.67</td>
<td>3.46</td>
<td>3.72</td>
</tr>
<tr>
<td>5,000</td>
<td>2.68</td>
<td>3.56</td>
<td>3.91</td>
</tr>
<tr>
<td>10,000</td>
<td>2.41</td>
<td>3.11</td>
<td>3.29</td>
</tr>
<tr>
<td>∞</td>
<td>2.75</td>
<td>3.90</td>
<td>5.34</td>
</tr>
</tbody>
</table>

2. Repeatedly generate data $Y_1, ..., Y_n \sim p(\cdot|\hat{\mu})$ i.i.d.

3. Compute the $\alpha$-quantile $q_{n,\alpha}$ of the test statistic $L_n$ applied to the data $Y_1, ..., Y_n$.

The null of parameter homogeneity is rejected if $L_n > q_{n,\alpha}$. To prove that this bootstrap procedure leads to a valid test, we need to show that the quantile $q_{n,\alpha}$ converges to the quantile $q_\alpha$ of the asymptotic distribution of $L_n$ conditionally on the data almost surely under the null. Since $\hat{\mu} \to \mu_0$ almost surely under rather weak assumptions, the statement of Theorem 2.3.4 given below will imply consistency of the parametric bootstrap.

Fix an arbitrary sequence of points $\mu_n$ in $\Theta$ with $\mu_n \to \mu_0 \in \Theta$, as $n \to \infty$. Assume that $\Theta \subset \mathbb{R}^k$ and define for $\epsilon > 0$ $\Theta^\epsilon$ as the $\epsilon$-enlargement of $\Theta$ with respect to Euclidean distance. Denote by $\mathcal{G}^\epsilon$ the set of non-degenerate distributions on $\Theta^\epsilon$. To each measure $\eta \in \mathcal{G}$ define the measure $\eta_n$ through $\eta_n(A) = \eta(A - \mu + \mu_n)$ for all Borel sets $A \subset \Theta$. From now on, assume that $X_1, ..., X_n$ are i.i.d. $\sim p(\cdot|\mu_n)$ and define the following sequence of processes indexed by $\mathcal{G}^\epsilon$

$$G_n(\eta) := n^{-1/2} \sum_{i=1}^{n} s_{\eta_n,\mu_n}(X_i).$$

Also, write $\ell_n(\eta) := \sum_{i=1}^{n} \log p_\eta(X_i)$ (this is different from the setting in Section 2.2.2 since here the distribution of the bootstrap sample $X_1, ..., X_n$ changes with $n$). To analyze the asymptotic behavior of $L_n$ under triangular arrays, consider the decomposition

$$L_n = \sup_{\eta} \ell_n(\eta) - \ell_n(\delta(\mu_n)) + \ell_n(\delta(\mu_n)) - \sup_{\mu \in \Theta} \ell_n(\delta(\mu)).$$

Classical results suggest that under suitable regularity conditions the second part in the above

48
decomposition should take the form

$$\sup_{\mu \in \Theta} \ell_n(\delta(\mu)) - \ell_n(\delta(\mu_n)) = \frac{1}{2} \left( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I(\mu_n)^{-1/2} q'(X_i|\mu_n) \right)^2 + o_P(1) \quad (2.3.1)$$

provided that \( \mu_n \to \mu_0 \). Various conditions ensuring the above representation exist, and we are not going into details here. The main challenge is to derive an expansion for the first part of \( L_n \) for data that are generated from a triangular array. This is established in Theorem 2.3.4 under the following set of assumptions:

**Assumption 2.3.1.** Assume that

$$\left( G_n, \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I(\mu_n)^{-1/2} q'(X_i|\mu_n) \right) \sim (G, Y_1)$$

in \( \ell^\infty(\mathcal{G}^c) \) where \((G, Y_1)\) are jointly centered normal with covariance structure of the form

$$E[G(\eta_1)G(\eta_2)] = \int_{\mathbb{R}} s_{\eta_1,\mu_0}(x)s_{\eta_2,\mu_0}(x)p_{\delta(\mu_0)}(x)d\nu(x),$$

$$E[G(\eta)Y_1] = \int_{\mathbb{R}} s_{\eta,\mu_0}(x)I(\mu_0)^{-1/2} q'(x|\mu_0)p_{\delta(\mu_0)}(x)d\nu(x).$$

Additionally, assume that for \( \epsilon \downarrow 0 \) we have

$$\sup_{\eta \in \mathcal{G}^c} \inf_{\tilde{\eta} \in \mathcal{G}} |G(\eta) - G(\tilde{\eta})| = o_P(1).$$

**Assumption 2.3.2.** Letting \( s_{\eta,\mu,-} := \min(0, s_{\eta,\mu}) \) we have that

$$\sup_{\eta \in \mathcal{G}^c} \left| \frac{1}{n} \sum_{i=1}^{n} (s_{\eta_n,\mu_n}(X_i) - 1) \right| + \left| \frac{1}{n} \sum_{i=1}^{n} (s_{\eta_n,\mu_n,-}(X_i) - \|s_{\eta,\mu,-}\|_2 \delta(\mu)) \right| = o_P(1).$$

**Assumption 2.3.3.** For every \( n \in \mathbb{N} \), assume that the class of functions

$$\mathcal{F}_n := \left\{ x \mapsto s_{\eta_n,\mu_n}(x) \mid \eta \in \mathcal{G}^c \right\}$$

admits a square integrable envelope function \( F_n \) such that \( \max_{i=1,\ldots,n} F_n(X_i) = o_P(n^{1/2}) \) and \( \limsup_{n \to \infty} E F_n^2(X_i) < \infty \).

We can show that Assumption 2.3.1-2.3.3 hold for some standard models. Details for verification of these assumptions for location mixture of normals and the mixture of Poisson are included in Section 2.8. We now state our main result.
Theorem 2.3.4. Under Assumptions 2.3.1-2.3.3 we have

\[ 2 \sup_{\eta} \left( \ell_n(\eta) - \ell_n(\mu_n) \right) = \sup_{\eta \in \mathcal{G}} \left( \max \left\{ \mathcal{G}_n(\eta), 0 \right\} \right)^2 + o_p(1). \]

If additionally (2.3.1) holds we have

\[ 2 \left( \sup_{\eta} \ell_n(\eta) - \sup_{\mu \in \Theta} \ell_n(\delta(\mu)) \right) \sim R := \sup_{\eta \in \mathcal{G}} \left( \max(\mathcal{G}(\eta), 0) \right)^2 - Y_1^2. \]

Theorem 2.3.4 suggests that critical values based on the parametric bootstrap should lead to an asymptotic level \( \alpha \) test of homogeneity. In order to prove the validity of the bootstrap procedure, we need an additional result on the properties of the distribution of \( R \). More precisely, it requires that the distribution of \( R \), say \( F_R \), is continuous around \( F_R^{-1}(\alpha) \) (note that if \( F_R \) is discontinuous at \( F_R^{-1}(\alpha) \), convergence of \( q_n, \alpha \) to \( F_R^{-1}(\alpha) \) does not necessarily follow from Theorem 2.3.4). The following theorem completes this last step.

Theorem 2.3.5. Let the assumptions of Theorem 2.2.1 hold, and additionally assume that

\[ \mathbb{E}[\mathcal{G}(\delta(\mu_0 + 1/n))\mathcal{G}(\delta(\mu_0 - 1/n))] \to -1 \text{ and there exists } \eta \in \mathcal{G} \text{ such that } \mathbb{E}[\mathcal{G}(\eta)Y_1] \neq \pm 1. \]

Then the distribution of \( R \) is continuous on \((0, \infty)\) and \( \mathbb{P}(R > 0) \geq 1/4 \). In particular, the bootstrap works for tests with level \( \alpha \leq 1/4 \).

Remark 2.3.6. For mixture models with densities of the form \( p(\cdot | \mu) = p(\cdot - \mu) \) there is an alternative way of simulating quantiles of the LR test. The key observation is that, assuming that we allow for an arbitrary support of the mixing distribution, the distribution of the likelihood ratio test under the null does not depend on the location of the true parameter. Thus the following procedure provides a way to conduct an exact test for parameter homogeneity when the support is unrestricted.

1. Repeatedly generate data \( Y_1, \ldots, Y_n \sim p(\cdot | 0) \) i.i.d. for \( B \) times. For each bootstrap sample, compute the LR test statistics \( L_{n,b} \) for \( b = 1, \ldots, B \).

2. Compute the \( \alpha \)-quantile \( q_{n,\alpha} \) of all \( L_{n,b}, b = 1, \ldots, B \).

The null of parameter homogeneity is rejected if \( L_n > q_{n,\alpha} \).

It is important to keep in mind that this invariance property will hold only if we consider an unrestricted support. In the case of Gaussian location mixtures, it is well known that the likelihood ratio test statistic with mixing distributions of unbounded support diverges to infinity (see Hartigan (1985)). This indicates that such a test might have a lower power than a test with a restricted support for the mixture. However, the divergence rate is extremely slow, and so in practice the power loss might not be relevant. Table 2.3 tabulates the bootstrap critical values for the null distribution of the LR test statistics for testing homogeneity of the Gaussian location parameter based on this observation. \( B \) bootstrap samples of size \( n \) is generated from standard normal distribution and the critical values are found based on the empirical distribution of the
corresponding likelihood ratio test statistics. To evaluate size performance of using these bootstrap critical values, we apply the LRT on a random sample \(X_1, \ldots, X_n \sim N(1, 1)\) for homogeneity versus general mixture on the location parameter. The last row of Table 2.4 reports the size performance of the LRT with these tabulated bootstrap critical values. In the same table, we also report the size performance of the C(\(\alpha\)) test and the EM test that will be discussed in the next section.

<table>
<thead>
<tr>
<th></th>
<th>90%</th>
<th>95%</th>
<th>99%</th>
</tr>
</thead>
<tbody>
<tr>
<td>n=100</td>
<td>3.14</td>
<td>4.60</td>
<td>8.12</td>
</tr>
<tr>
<td>n=200</td>
<td>3.15</td>
<td>4.48</td>
<td>7.21</td>
</tr>
<tr>
<td>n=500</td>
<td>3.44</td>
<td>4.69</td>
<td>7.84</td>
</tr>
</tbody>
</table>

Table 2.3: Bootstrap critical values for likelihood ratio test of homogeneity of Gaussian location parameter: For various sample sizes, the bootstrap critical values are found following the procedure described in Remark 2.3.6 with \(B = 2,000\).

<table>
<thead>
<tr>
<th></th>
<th>n = 100</th>
<th>n = 200</th>
<th>n = 500</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>90%</td>
<td>95%</td>
<td>99%</td>
</tr>
<tr>
<td>EM</td>
<td>0.088</td>
<td>0.044</td>
<td>0.010</td>
</tr>
<tr>
<td>C((\alpha))</td>
<td>0.103</td>
<td>0.050</td>
<td>0.018</td>
</tr>
<tr>
<td>LRT</td>
<td>0.072</td>
<td>0.038</td>
<td>0.008</td>
</tr>
</tbody>
</table>

Table 2.4: Size performance for various tests for homogeneity of the Gaussian location parameter: Independent samples of different sizes are generated from \(N(1, 1)\). We consider test for homogeneity versus general alternative. The EM test is as proposed in Chen and Li (2009) using the R code provided by the second author on http://sas.uwaterloo.ca/~p4li/software/index.html for the EM test for Gaussian mixture with known variance. The C(\(\alpha\)) test uses critical values from \(\frac{1}{2}X^2_0 + \frac{1}{2}X^2_1\) null distribution. LRT uses bootstrap critical values tabulated in Table 2.3. Results are based on 6,000 repetition.

### 2.4 Neyman’s C(\(\alpha\)) tests for mixture models

As we have discussed in details in Chapter 1, Neyman’s C(\(\alpha\)) tests can be applied to test for homogeneity in mixture models. The test statistics typically still take a simple form although their theory requires some substantial amendment from the regular cases due to the singularity of the score function. We refer the theoretical details to Chapter 1 and use the following two examples to illustrate the construction of the C(\(\alpha\)) test for parameter homogeneity. Both tests lead to an over-dispersion test. In the Gaussian case, the test compares the sample variance with the variance under the null hypothesis. In the Poisson case, we reject the null of homogeneity if there exists over-dispersion in the sample variance in comparison to the sample mean.
Example 2.4.1. Consider testing for homogeneity in the Gaussian location mixture model with independent observations $X_i \sim N(\mu_i, 1), i = 1, \cdots, n$. Assume that $\mu_i = \mu_0 + \tau \xi U_i$, for known $\tau$, and iid $U_i \sim F$ with $EU = 0$ and $VU = 1$. We would like to test $H_0 : \xi = 0$ with the location parameter $\mu_0$ treated as a nuisance parameter. The second-order score for $\xi$ is found to be, $\nabla^2_\xi \log p(x|\mu_0, \xi = 0) = \tau^2((x - \mu_0)^2 - 1)$ and the first-order score for $\mu_0$ is, $\nabla_{\mu_0} \log p(x|\mu_0, \xi = 0) = (x - \mu_0)$. Note that under the null, $\text{cov}(\nabla^2_\xi \log p(X|0, \mu_0), \nabla_{\mu_0} \log p(X|0, \mu_0)) = 0$. Thus, we have the locally asymptotically optimal $C(\alpha)$ test as

$$Z_n = \frac{1}{\sqrt{2n}} \sum_{i=1}^{n} ((X_i - \mu_0)^2 - 1)$$

The obvious estimate for the nuisance parameter is the sample mean, and we reject the null hypothesis when $(0 \lor Z_n)^2 > c_\alpha$. The test statistic $Z_n$ depends on the sample variance of $X$. Under the general alternative model, we have $\text{Var}(X) = E_{\mu}[\text{Var}(X|\mu)] + \text{Var}_\mu[E(X|\mu)] = 1 + \text{Var}(\mu)$. Under the alternative, the magnitude of $Z_n$ solely depends on $\sqrt{n} \text{Var}(\mu)$.

Example 2.4.2. Consider now testing for homogeneity of the mean parameter in the Poisson model with independent observations $X_i \sim f(|\lambda_i), i = 1, \cdots, n$ with $f(x|\lambda) = \frac{\lambda^x \exp(-\lambda)}{x!}$. Assume that $\lambda_i = \lambda_0 \exp(\tau \xi U_i)$, for known $\tau$, and iid $U_i \sim F$ with $EU = 0$ and $VU = 1$. We would like to test $H_0 : \xi = 0$ with the mean parameter $\lambda_0$ treated as a nuisance parameter. The second-order score for $\xi$, is found to be, $\nabla^2_\xi \log f(x|\lambda_0, \xi = 0) = \tau^2((x - \lambda_0)^2 - \lambda_0)$ and the first-order score for $\lambda_0$ is, $\nabla_{\lambda_0} \log f(x|\lambda_0, \xi = 0) = (x - \lambda_0)/\lambda_0$. Note that under the null, $\text{cov}(\nabla^2_\xi \log f(X|\lambda_0, 0), \nabla_{\lambda_0} \log f(X|\lambda_0, 0)) = \lambda_0$. Thus, we have the locally asymptotically optimal $C(\alpha)$ test as

$$Z_n = \frac{1}{\sqrt{2n}} \sum_{i=1}^{n} \frac{(X_i - \lambda_0)^2 - \lambda_0 - (X_i - \lambda_0)}{\lambda_0}$$

The obvious estimate for the nuisance parameter $\lambda_0$ is the sample mean $\bar{X}$, which further reduces $Z_n = \frac{1}{\sqrt{2n}} \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2 - \bar{X}}{\bar{X}}$ and we reject the null hypothesis when $(0 \lor Z_n)^2 > c_\alpha$. The test statistic $Z_n$ depends on the ratio of the sample variance and sample mean of $X$. Under the alternative model, we have $\text{Var}(X) = E(\lambda) + \text{Var}(\lambda)$ and $E(X) = E(\lambda)$. The magnitude of the test statistics $Z_n$ under the alternative is determined by the ratio $\sqrt{n} \text{Var}(\lambda)/E(\lambda)$.

2.5 The EM test for mixture models

Most of the literature on testing for parameter homogeneity focuses on likelihood ratio test for finite mixture models. A very closely related inquiry involves testing the order of a finite mixture model since testing homogeneity versus general alternative mixture models can be reduced to testing whether the underlying mixture model has more than one component. There has been a continuous development of this line of work by Chen and Li (2009), Li, Chen, and Marriott (2009) and Li and Chen (2010), who propose an EM approach for testing the order of a finite mixture model. They
show that the limiting distribution of the EM test under the null takes a simple form and thus circumvents the challenges associated with likelihood ratio testing for finite mixture models, that it has a complicated null distribution while requiring stronger assumptions. Since their approach is closely related to the nonparametric likelihood ratio test and the construction of the EM test builds upon the EM algorithm for solving mixture models which contrasts with the alternative convex optimization approach proposed in Koenker and Mizera (2014) as advocated here, we provide a detailed discussion of the EM test. We also points out an interesting connection of the EM test with the $C(\alpha)$ test.

The construction of the EM test, as the name suggests, relies on the EM algorithm. Since the EM algorithm is known to exhibit slow convergence in high dimensional settings, the EM test restricts the alternative to a special class of mixing distribution with fixed order and only conduct a finite number of iteration of the algorithm. As pointed out in Li and Chen (2010), if the true order of the mixing distribution under the alternative exceeds the imposed order, the method may be inefficient (i.e. the likelihood can be increased further if one searches over a larger space of distributions). On the contrary, the LR test using the Kiefer-Wolfowitz MLE does not require knowledge of the number of support points for the alternative finite mixture model, and the convex optimization formulation greatly facilitates the computation of the LRT statistics.

To avoid the likelihood diverging to infinity (see Hartigan (1985)), the EM test introduces a penalty function on the mixing weights that prevents such occurrences. By doing so the test also removes the compactness condition usually required for the likelihood ratio test statistics as in Gassiat (2002) and also required in our setting, at the cost of introducing new tuning parameters.

Our presentation of the EM test below follows Li and Chen (2010) closely. Consider a finite mixture model with a parametric base distribution with density $f(x|\mu)$. The marginal density can therefore be written as

$$g(x) = \int f(x|\mu)dF(\mu)$$

with $F(\mu) = \sum_{h=1}^{m} \alpha_h 1(\mu_h \leq \mu)$ as the discrete mixing distribution with $m$ support points at locations $\{\mu_1, \ldots, \mu_m\}$. We are interested in testing

$$H_0 : m = m_0 \text{ versus } H_A : m > m_0$$

When $m_0 = 1$, we have the special case of homogeneity testing. We will focus on this case from now on. Although the alternative hypothesis contains a large class of mixing distributions, the EM test starts by restricting to the following class of mixing distribution with exactly two support points.

$$\Omega_2(\beta) = \{\beta 1(\mu_1 \leq \mu) + (1-\beta) 1(\mu_2 \leq \mu) : \mu_1, \mu_2 \in I\}$$

where $\beta \in (0, 0.5]$. The set $I = (-\infty, +\infty)$ is the support for $\{\mu_1, \mu_2\}$. The mixing distribution in the set $\Omega_2$ corresponds to an alternative model with exactly two support points that split the

---

In the R code for the EM test supplied by the first author of Li and Chen (2010) on http://sas.uwaterloo.ca/~p4li/software/index.html, it is taken to be the data support.
original one support point into two separate ones with weights \((\beta, 1 - \beta)\).

For each mixing distribution \(\Psi \in \Omega_2(\beta)\), the penalized log-likelihood function is defined to be

\[
\text{pl}_n(\Psi) = \sum_{i=1}^{n} \log f(X_i; \Psi) + p(\beta)
\]

where \(p(\beta)\) is a continuous penalty function such that it is maximized at value 0.5 and goes to negative infinity as \(\beta\) goes to 0 or 1. The penalty function avoids the possibility that maximizing \(\text{pl}_n(\Psi)\) puts arbitrarily small probability mass at a specific location.

The EM test construction starts with a set \(B\) that contains \(J\) different \(\beta\) values. Feeding each \(\beta\) into the penalized log-likelihood \(\text{pl}_n(\Psi)\) we can find the optimal solution for \((\hat{\mu}_1, \hat{\mu}_2)\) which forms a solution \(\hat{\Psi} \in \Omega_2(\beta)\). The \(\hat{\Psi}\) thus obtained for each \(\beta\) is used as the initial values for the EM algorithm iteration updating all the parameters for a better fit of \(\Psi\) that maximizes the modified likelihood function. The EM algorithm iteration continues for \(K = 3\) steps and at each iteration the modified likelihood ratio \(M_n^{(k)}(\beta)\) is defined as below, among all \(K\) iterations and for each \(\beta \in B\),

\[
M_n^{(k)}(\beta) = 2\{\text{pl}_n(\Psi^{(k)}(\beta)) - \sum_i \log f(X_i; \hat{\Psi}_0)\}, k = 1, \ldots, K,
\]

with \(\hat{\Psi}_0\) is the maximum likelihood estimator for the null model. The EM test statistic is then found to be the maximum across different \(\beta\)

\[
\text{EM}_n = \max\{M_n^{(K)}(\beta) : \beta \in B\}
\]

The EM test is shown to have a simple mixture of \(\chi^2\) limiting distribution \(\sum_{h=0}^{m_0} q_h \chi^2_h\) under the null. For \(m_0 = 1\), it simplifies to \(\frac{1}{2} \chi_0^2 + \frac{1}{2} \chi_1^2\).

It is worth noting that the EM test can be shown to be asymptotically equivalent to a \(C(\alpha)\) test that tests for randomness of the parameter \(\theta\). More specifically, consider the example with \(m_0 = 1\). The homogenous null model is thus \(f(X_i; \Psi_0) = f(X_i; \mu_0)\). Theorem 2 of Li and Chen (2010) shows that for any fixed finite \(K\), as \(n \to \infty\),

\[
\text{EM}_n = \sup_{\nu \geq 0} \left\{ 2\nu \sum_{i=1}^{n} \tilde{b}_{2i} - n\nu^2 \tilde{B}_{22} \right\} + o_p(1)
\]  

with \(\tilde{b}_{2i} = b_{2i} - B_{2i}B_{1i}^{-1}b_{1i}\) and \(\tilde{B}_{22} = B_{22} - B_{2i}B_{1i}^{-1}B_{12}\). Consider the parametrization which introduces randomness of \(\mu_0\), \(\mu_i = \mu_0 + \xi U_i\) as we used to construct the \(C(\alpha)\) test discussed earlier. We can easily show that \(b_{2i}\) is nothing but the second order score function for \(\xi\), and \(b_{1i}\) is the score function for the nuisance parameter \(\theta_0\), and \(B_{ij}\) is the corresponding elements in the Fisher information matrix. Solving the quadratic problem in (2.5.1), we get

\[
\text{EM}_n = Z_n + o_p(1)
\]
with \( Z_n = \tilde{B}_2^{-1} \left( n^{-1/2} \sum_{i=1}^{n} \tilde{b}_{2i} \vee 0 \right)^2 \), which is exactly the \( C(\alpha) \) test.

When \( m_0 > 1 \), the same idea applies except that we need a \( C(\alpha) \) test of multiple dimensions. The null model is then written as \( f(X_i; \Psi_0) = \sum_{h=1}^{m_0} \alpha_h f(X_i; \mu_h) \). Introduce randomness in each of the support points \( \mu_h \) via \( \tilde{\mu}_h = \mu_h + \tilde{\xi}_h U_h \) and test for \( H_0 : \tilde{\xi}_h = 0, \forall h \) versus \( H_1 : \sum_{h=1}^{m_0} \tilde{\xi}_h \neq 0 \).

The nuisance parameters consist \( (\alpha_h, \mu_h) \) for \( h = 1, \ldots, m_0 \). The \( C(\alpha) \) test constructed using Theorem 1.3.1 in Chapter 1 can be shown to be asymptotically equivalent to the EM test.

The most attractive feature of the EM test and shared by the \( C(\alpha) \) test is that they have a simple limiting distribution, a mixture of \( \chi^2 \), in contrast to the supremum of Gaussian process limit for the nonparametric LRT. But given the proposed bootstrap method, this disadvantage of the nonparametric LRT is alleviated. In the next section we compare the size and power performance of the general LRT with the EM test and the \( C(\alpha) \) test for different mixture models in simulations.

### 2.6 Simulation

To compare power of the \( C(\alpha) \), the EM test and LRT to detect heterogeneity in the Gaussian location model we conducted four distinct experiments. Two were based on variants of the Chen (1995) example with the discrete mixing distribution \( F(\mu) = (1-\lambda)\delta_h/(1-\lambda) + \lambda \delta_{-h/\lambda} \). In the first experiment we set \( \lambda = 1/3 \), as in the original Chen example, in the second experiment we set \( \lambda = 1/20 \). We consider four tests: (i.) the \( C(\alpha) \) as described in Example 2.4.1, (ii.) a parametric version of the LRT in which only the value of \( h \) is assumed to be unknown and the relative probabilities associated with the two mass points are known; this enables us to relatively easily find the MLE, \( \hat{h} \) by separately optimizing the likelihood on the positive and negative half-line and taking the best of the two solutions, (iii.) the Kiefer-Wolfowitz LRT computed with equally spaced binning on the support of the sample, and finally as benchmark (iv.) the classical Kolmogorov-Smirnov test of normality (v.) the EM test for one component versus two components.

The sample size in all the power comparisons was taken to be 200, with 10,000 replications. We consider 21 distinct values of \( h \) for each of the experiments equally spaced on the respective plotting regions.

In the left panel of Figure 2.1 we illustrate the results for the first experiment with \( \lambda = 1/3 \): with the location invariant property of the Gaussian mixture model, we use the bootstrap critical values in Table 2.3 for the nonparametric LRT. The EM test, \( C(\alpha) \) and the parametric LRT are essentially indistinguishable in this experiment, and both have slightly better performance than the nonparametric LRT. All four of these tests perform substantially better than the Kolmogorov-Smirnov test. In the right panel of Figure 2.1 we have results of another version of the Chen example, except that now \( \lambda = 1/20 \), so the mixing distribution is much more skewed. Still \( C(\alpha) \) does well for small values of \( h \), but for \( h \geq 0.07 \) the two LRT procedures, which are now essentially indistinguishable, dominate. The performance of the EM test lies in between the \( C(\alpha) \) test and the nonparametric LRT test. Again, the KS test performance is poor compared to the other tests explicitly designed for the mixture setting.

In Figure 2.2 we illustrate the results of two additional experiments, both of which are based
on mixing distributions with densities with respect to Lebesgue measure. On the left we consider \( F(\mu, h) = I(-h < \mu < h)/(2h) \). Again, we can reduce the parametric LRT to optimizing separately over the positive and negative half-lines to compute the MLE, \( \hat{h} \). This would seem to give the parametric LRT a substantial advantage over the Kiefer-Wolfowitz nonparametric MLE, however as is clear from the figure, there is little difference in their performance. Again, the \( C(\alpha) \) test and the EM test are somewhat better than either of the LRTs, but the difference is modest. In the right panel of Figure 2.2 we have a similar setup, except that now the mixing distribution is Gaussian with scale parameter \( h \), and again the ordering is very similar to the uniform mixing case. In both of the latter experiments, the parametric LRT is somewhat undersized at the null; so we made an empirical size adjustment.

We also consider the power performance of the the above mentioned tests for Poisson mixture models. Similarly to the Gaussian case, the Poisson mean parameter has the discrete mixing distribution \( F(\mu) = 2(1-\lambda)\exp(\delta h/(1-\lambda)) + 2\lambda\exp(\delta -h/\lambda) \). We consider \( \lambda = 1/3 \) and \( \lambda = 1/20 \) case. The \( C(\alpha) \) test is constructed as described in Example 2.4.2. Since Poisson mixture model does not take a location shift form, we resort to the parametric bootstrap method described in Section 2.3 to determine the critical value. Figure 2.3 shows the power for the \( C(\alpha) \) test, the EM test and the KW-LRT for different values of \( h \). Again, we observe similar pattern of the power curves as in the Gaussian case. For more extreme mixing distribution, the KW-LRT dominates the other two tests by quite a substantial margin.

Figure 2.1: Power comparison of several tests of parameter homogeneity: The left panel illustrates empirical power curves for four tests of parameter homogeneity for the Chen (1995) mixture with \( \lambda = 1/3 \), in the right panel we illustrate the power curves for the same four tests for the Chen mixture with \( \lambda = 1/20 \). Note that in the more extreme (right) setting, the LRTs outperform the \( C(\alpha) \) test.
2.7 Conclusion

We have seen that the Neyman $C(\alpha)$ test provides a simple, powerful, albeit somewhat irregular, strategy for constructing tests of parameter homogeneity. In contrast, likelihood ratio testing for mixture models has been somewhat inhibited by their apparent computational difficulty, as well as the complexity of its asymptotic theory. Recent developments in convex optimization have dramatically reduced the computational effort of earlier EM methods, and new theoretical developments have led to practical simulation methods for large sample critical values for the Kiefer-Wolfowitz nonparametric version of the LRT. Local asymptotic optimality of the $C(\alpha)$ test assures that it is highly competitive in most circumstances, but we have illustrated a class of examples where the LRT has a slight edge. The two approaches are complementary; clearly there is little point in testing for heterogeneity if there is no mechanism for estimating models under the alternative. Our LRT approach obviously provides a direct pathway to estimation of the mixture model under the alternative. Since parametric mixture models are notoriously tricky to estimate, it is a remarkable fact that the nonparametric formulation of the MLE problem a la Kiefer-Wolfowitz can be solved quite efficiently – even for large sample sizes by binning – and effectively used as an alternative testing procedure. We hope that these new developments will encourage others to explore these methods.
Figure 2.3: Power comparison of several tests of parameter homogeneity for Poisson mixture models: The figure illustrates empirical power curves for three tests of parameter homogeneity for a discrete mixtures of Poisson. The discrete mixing distribution is specified as \( F(\mu) = 2(1-\lambda) \exp(\delta_{h/(1-\lambda)}) + 2\lambda \exp(\delta_{-h/\lambda}) \) with \( \lambda = 1/3 \) in the left panel and \( \lambda = 1/20 \) in the right panel for \( h \) taking 21 different values. The critical values for LRT are based on the bootstrap method. We adopt the warp bootstrap method in Giacomini, Politis, and White (2013) to speed up the Monte Carlo simulation. The empirical power curve is based on 5,000 repetition.

2.8 Proofs

Proof of (2.2.3). Given a measure \( \eta \in \mathcal{G}, \eta \neq \delta(\mu_0) \) define \( V(\eta) := \sum_{k=2}^{\infty} \frac{\kappa_k^2(\eta)}{k!} \). Also, define for \( n \in \mathbb{N} \) and \( \alpha \in [-N, N] \) the probability measure \( \tilde{\eta}_n := p_n \delta_{c_n} + (1 - p_n) \eta \) with \( p_n := 1 - V(\eta)/n \) and \( c_n := \frac{1-p_n}{p_n} (\alpha - \kappa_1(\eta)) \) [the dependence of \( p_n, c_n \) on \( \eta \) is suppressed in the notation]. Note that for \( n \) sufficiently large we have \( \tilde{\eta}_n \in \mathcal{G} \) for all \( \alpha \in [-N, N] \). Moreover, by construction \( \kappa_1(\tilde{\eta}_n) = \alpha(1 - p_n) \) and

\[
\kappa_k(\tilde{\eta}_n) = \kappa_k(\eta)(1 - p_n) + (1 - p_n)\left(\frac{1-p_n}{p_n}\right)^{k-1}(\alpha - \kappa_1(\eta))^k
\]

for \( n \in \mathbb{N} \). This implies for \( n \) sufficiently large we have a.s.

\[
|\alpha Y_1 + \sum_{k=2}^{\infty} \frac{Y_k \kappa_k(\tilde{\eta}_n)}{(k!)^{1/2}}| \sum_{k=1}^{\infty} \frac{Y_k \kappa_k(\tilde{\eta}_n)}{(k!)^{1/2}} | \leq \frac{1-p_n}{p_n} \sum_{k=2}^{\infty} \frac{|Y_k| \tilde{C}^k}{\sqrt{k!}} \left(\frac{1-p_n}{p_n}\right)^{k-2} \leq \frac{2\tilde{C}^2 V(\eta)}{n} \sum_{k=2}^{\infty} \frac{|Y_k|}{\sqrt{k!}}
\]

and

\[
|\alpha^2 + \sum_{k=2}^{\infty} \frac{\kappa_k^2(\tilde{\eta}_n)}{k!} - \frac{1}{(1-p_n)^2} \sum_{k=1}^{\infty} \frac{\kappa_k^2(\tilde{\eta}_n)}{k!}| \leq CV(\eta) \frac{C}{n}
\]

for finite constants \( C, \tilde{C} \) depending only on \( N \) but not on \( \alpha \) and \( \eta \) [note that \( \eta \in \mathcal{G} \) has support contained in \([L, U]\)]. Thus for every \( N < \infty, \epsilon > 0 \) we have with probability at least \( 1 - \epsilon \) [this
follows by choosing $n$ sufficiently large]

\[
\sup_{\eta \in \mathcal{G}} \frac{\sum_{k=1}^{\infty} \frac{Y_k \kappa_k(\eta)}{(k!)^{1/2}}}{\left( \sum_{k=1}^{\infty} \frac{k^2(\eta)}{k!} \right)^{1/2}} \geq \sup_{\alpha \in [-N,N]} \sup_{\eta \in \mathcal{G}} \frac{\alpha Y_1 + \sum_{k=2}^{\infty} \frac{Y_k \kappa_k(\eta)}{(k!)^{1/2}}}{\left( \alpha^2 + \sum_{k=2}^{\infty} \frac{k^2(\eta)}{k!} \right)^{1/2}} - \epsilon.
\]

Next, observe that $N$ can be chosen so large that with probability at least $1 - f(\epsilon)$

\[
\sup_{\alpha \in \mathbb{R} \setminus [-N,N]} \sup_{\eta \in \mathcal{G}} \frac{\alpha Y_1 + \sum_{k=2}^{\infty} \frac{Y_k \kappa_k(\eta)}{(k!)^{1/2}}}{\left( \alpha^2 + \sum_{k=2}^{\infty} \frac{k^2(\eta)}{k!} \right)^{1/2}} \leq |Y_1| + \epsilon
\]

where $f(\alpha) \to 0$ for $\alpha \to 0$. Finally, note that

\[
\sup_{\eta \in \mathcal{G}} \frac{\sum_{k=1}^{\infty} \frac{Y_k \kappa_k(\eta)}{(k!)^{1/2}}}{\left( \sum_{k=1}^{\infty} \frac{k^2(\eta)}{k!} \right)^{1/2}} \geq |Y_1| \quad \text{a.s.}
\]

[consider the sequence of measures $\eta_n = \delta_{\text{sign}(Y_1)/n} \in \mathcal{G}$.]

Summarizing the findings above, we have shown that for any $\epsilon > 0$ we have with probability arbitrarily close to one:

\[
\sup_{\eta \in \mathcal{G}} \frac{\sum_{k=1}^{\infty} \frac{Y_k \kappa_k(\eta)}{(k!)^{1/2}}}{\left( \sum_{k=1}^{\infty} \frac{k^2(\eta)}{k!} \right)^{1/2}} \geq \sup_{\alpha \in \mathbb{R} \setminus [-N,N]} \sup_{\eta \in \mathcal{G}} \frac{\alpha Y_1 + \sum_{k=2}^{\infty} \frac{Y_k \kappa_k(\eta)}{(k!)^{1/2}}}{\left( \alpha^2 + \sum_{k=2}^{\infty} \frac{k^2(\eta)}{k!} \right)^{1/2}} - \epsilon.
\]

By letting $\epsilon \to 0$ the above can be turned into an almost sure inequality with no $\epsilon$ on the right-hand side. Finally, setting $\alpha = \kappa_1(\eta)$ we see that the converse inequality also holds almost surely. Thus we have shown that

\[
\sup_{\eta \in \mathcal{G}} \frac{\sum_{k=1}^{\infty} \frac{Y_k \kappa_k(\eta)}{(k!)^{1/2}}}{\left( \sum_{k=1}^{\infty} \frac{k^2(\eta)}{k!} \right)^{1/2}} = \sup_{\alpha \in \mathbb{R} \setminus [-N,N]} \sup_{\eta \in \mathcal{G}} \frac{\alpha Y_1 + \sum_{k=2}^{\infty} \frac{Y_k \kappa_k(\eta)}{(k!)^{1/2}}}{\left( \alpha^2 + \sum_{k=2}^{\infty} \frac{k^2(\eta)}{k!} \right)^{1/2}} \quad \text{a.s.}
\]

Define $\beta_k := \frac{\kappa_k(\eta)}{(k!)^{1/2}}$. Fix a realization of $Y_1, Y_2, \ldots$. First, observe that it suffices to consider the supremum over $\eta \in \mathcal{G}$ with $\sum_{k=2}^{\infty} Y_k \beta_k \geq 0$. Fixing $\eta \in \mathcal{G}$ shows that in the case $\sum_{k=2}^{\infty} Y_k \beta_k > 0$ the supremum with respect to $\alpha$ on the right-hand side above is attained for $\alpha^* = Y_1 \frac{\sum_{k=2}^{\infty} \beta_k}{\sum_{k=2}^{\infty} Y_k \beta_k}$, and plugging this into the equation above we obtain

\[
\sup_{\alpha \in \mathbb{R}} \frac{\alpha Y_1 + \sum_{k=2}^{\infty} \frac{Y_k \kappa_k(\eta)}{(k!)^{1/2}}}{\left( \alpha^2 + \sum_{k=2}^{\infty} \frac{k^2(\eta)}{k!} \right)^{1/2}} = \left( Y_1^2 + \frac{\left( \sum_{k=2}^{\infty} Y_k \beta_k \right)^2}{\sum_{k=2}^{\infty} \beta_k^2} \right)^{1/2}
\]
for every $\eta \in \mathcal{G}$ with $\sum_{k=2}^{\infty} Y_k \beta_k > 0$. In the case $\sum_{k=2}^{\infty} Y_k \beta_k = 0$ we obtain $\alpha^* = \text{sign}(Y_1)$. Summarizing the above arguments yields

$$
\left( \sup_{\eta \in \mathcal{G}} \left( \frac{\sum_{k=1}^{\infty} Y_{kk}(\eta)}{\left( \sum_{k=1}^{\infty} \frac{\eta^2 k^2}{\kappa^2} \right)^{1/2}} \right) \right)^2 = Y_1^2 + \sup_{\eta \in \mathcal{G}} \left( \frac{\sum_{k=2}^{\infty} Y_k \beta_k}{\sum_{k=2}^{\infty} \beta_k^2} \right)^2
$$

and this directly implies (2.2.3)

Proof of Theorem 2.3.4. The proof uses arguments from the proof of Theorem 3.1 in Gassiat (2002). Let $\gamma_n := |\mu_n - \mu|$. Observe to each $\eta \in \mathcal{G}$ there exists $\bar{\eta} \in \mathcal{G}^\mathcal{G}$ such that $\bar{n}_n = \eta$. Thus under Assumption 2.3.1 we have

$$
n^{-1} \sup_{\eta \in \mathcal{G}} \left( \sum_{i=1}^{n} s_{\eta, \mu_n}(X_i) \right)^2 \leq n^{-1} \sup_{\eta \in \mathcal{G}^\mathcal{G}} \left( \sum_{i=1}^{n} s_{\eta_n, \mu_n}(X_i) \right)^2 \leq \sup_{\eta \in \mathcal{G}^\mathcal{G}} \mathcal{G}_n(\eta)^2 = O_P(1) \quad (2.8.1)
$$

where the first inequality holds for $n$ sufficiently large. Moreover

$$
\lim_{n \to \infty} \inf_{\eta \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} s_{\eta, \mu_n, -}(X_i) \geq \lim_{n \to \infty} \inf_{\eta \in \mathcal{G}^\mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} s_{\eta_n, \mu_n, -}(X_i) \geq \inf_{\eta \in \mathcal{G}^\mathcal{G}} \|s_{\eta, \mu_n, -}\|_{2, \delta(\mu_n)}^2 > 0
$$

where the equality follows by Assumption 2.3.2 and the inequality follows by the same arguments as (5) in Gassiat (2002). Apply Inequality 1.1 from Gassiat (2002) to obtain

$$
\sup_{\eta \in \mathcal{G}, \ell_n(\eta) - \ell_n(\delta(\mu_n)) > 0} \|\frac{p_\eta}{p_{\delta(\mu_n)}} - 1\|_{2, \delta(\mu_n)} = O_P(n^{-1/2}). \quad (2.8.2)
$$

Applying Assumption 2.3.3 there exists functions $F_n$ such that $\sup_{\eta \in \mathcal{G}} |s_{\eta, \mu_n}(x)| \leq F_n(x)$ and $\sup_{i=1, \ldots, n} F(X_i) = o_P(n^{-1/2})$. Thus there exists $\alpha_n \to \infty$ such that $\sup_{i=1, \ldots, n} F(X_i) = o_P(\alpha_n^{-1} n^{1/2})$. For such a sequence $\alpha_n$ define the sets

$$
M_{n1} := \{\eta \in \mathcal{G} : \ell_n(\eta) - \ell_n(\delta(\mu_n)) > 0\}, \quad M_{n2} := \{\eta \in \mathcal{G} : \|\frac{p_\eta}{p_{\delta(\mu_n)}} - 1\|_{2, \delta(\mu_n)} \leq n^{-1/2} \alpha_n^{1/2}\}
$$

From (2.8.2) we obtain that $M_{n1} \subset M_{n2}$ with probability tending to one. On the other hand a Taylor expansion of $x \mapsto \log(1 + x)$ shows that

$$
\sup_{\eta \in M_{n2}} \ell_n(\eta) - \ell_n(\delta(\mu_n)) = \sup_{\eta \in M_{n2}} \left( \|\frac{p_\eta}{p_{\delta(\mu_n)}} - 1\|_{2, \delta(\mu_n)} \sum_{i=1}^{n} s_{\eta, \mu_n}(X_i) - \frac{1}{2} \|\frac{p_\eta}{p_{\delta(\mu_n)}} - 1\|_{2, \delta(\mu_n)}^2 \sum_{i=1}^{n} s_{\eta, \mu_n}(X_i) \right)
$$

+ $\|\frac{p_\eta}{p_{\delta(\mu_n)}} - 1\|_{2, \delta(\mu_n)}^2 \sum_{i=1}^{n} s_{\eta, \mu_n}^2(X_i) \mathcal{R}(\|\frac{p_\eta}{p_{\delta(\mu_n)}} - 1\|_{2, \delta(\mu_n)} \sum_{i=1}^{n} s_{\eta, \mu_n}(X_i))$
where the remainder function $R$ satisfies $R(u) \to 0$ for $u \to 0$. Now by the definition of $\alpha_n$ we have

$$
\sup_{\eta \in M_{n2}} \frac{1}{n} \sum_{i=1}^{n} s_{\eta,\mu_n}^2(X_i) R \left( \left\| \frac{p_{\eta}}{p_{\delta(\mu_n)}} - 1 \right\|_{L^2(\delta(\mu_n))} s_{\eta,\mu_n}(X_i) \right) \\
\leq \sup_{\eta \in M_{n2}} \frac{1}{n} \sum_{i=1}^{n} s_{\eta,\mu_n}^2(X_i) R \left( \frac{n^{-1/2} \alpha_n^{1/2}}{\alpha_n^{1/2}} \right)
$$

$$= o_p(1) \sup_{\eta \in M_{n2}} \frac{1}{n} \sum_{i=1}^{n} s_{\eta,\mu_n}^2(X_i).
$$

Additionally, Assumption 2.3.2 implies that

$$
\sup_{\eta \in \mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^{n} (s_{\eta,\mu_n}^2(X_i) - 1) \right| \leq \sup_{\eta \in \mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^{n} (s_{\eta,\mu_n}^2(X_i) - 1) \right| = o_p(1).
$$

Thus we see that

$$
\sup_{\eta \in M_{n2}} \ell_n(\eta) - \ell_n(\delta(\mu_n)) = \sup_{\eta \in M_{n2}} \left( \left\| \frac{p_{\eta}}{p_{\delta(\mu_n)}} - 1 \right\|_{L^2(\delta(\mu_n))} \sum_{i=1}^{n} s_{\eta,\mu_n}(X_i) - \frac{n}{2} \left\| \frac{p_{\eta}}{p_{\delta(\mu_n)}} - 1 \right\|_{L^2(\delta(\mu_n))} (1 + r_n) \right)
$$

where $r_n$ does not depend on $\eta$ and $r_n = o_p(1)$. Since $M_{n1} \subset M_{n2}$ with probability tending to one, and since

$$
\sup_{\eta \in \mathcal{G}} \ell_n(\eta) - \ell_n(\delta(\mu_n)) = \sup_{\eta \in M_{n1}} \ell_n(\eta) - \ell_n(\delta(\mu_n)),
$$

it follows that

$$
\sup_{\eta \in \mathcal{G}} \ell_n(\eta) - \ell_n(\delta(\mu_n)) = \sup_{\eta \in M_{n2}} \left( \left\| \frac{p_{\eta}}{p_{\delta(\mu_n)}} - 1 \right\|_{L^2(\delta(\mu_n))} \sum_{i=1}^{n} s_{\eta,\mu_n}(X_i) - \frac{n}{2} \left\| \frac{p_{\eta}}{p_{\delta(\mu_n)}} - 1 \right\|_{L^2(\delta(\mu_n))} (1 + r_n) \right) + o_p(1).
$$

(2.8.3)

Next, observe that for any $\eta \in \mathcal{G}$ we also have $\eta^t := \eta + (1 - t)\delta(\mu_n) \in \mathcal{G}$ for any $t \in [0, 1]$. Additionally, we have

$$
\left\| \frac{p_{\eta^t}}{p_{\delta(\mu_n)}} - 1 \right\|_{L^2(\delta(\mu_n))} = \frac{n}{2} \left\| \frac{p_{\eta}}{p_{\delta(\mu_n)}} - 1 \right\|_{L^2(\delta(\mu_n))}
$$

and by construction $s_{\eta^t,\mu_n}^2 = s_{\eta,\mu_n}^2$. Thus

$$
\sup_{\eta \in M_{n2}} \left( \left\| \frac{p_{\eta}}{p_{\delta(\mu_n)}} - 1 \right\|_{L^2(\delta(\mu_n))} \sum_{i=1}^{n} s_{\eta,\mu_n}(X_i) - \frac{n}{2} \left\| \frac{p_{\eta}}{p_{\delta(\mu_n)}} - 1 \right\|_{L^2(\delta(\mu_n))} (1 + r_n) \right)
$$

$$= \sup_{\eta \in \mathcal{G}} \left( t \left\| \frac{p_{\eta}}{p_{\delta(\mu_n)}} - 1 \right\|_{L^2(\delta(\mu_n))} \sum_{i=1}^{n} s_{\eta,\mu_n}(X_i) - \frac{nt^2}{2} \left\| \frac{p_{\eta}}{p_{\delta(\mu_n)}} - 1 \right\|_{L^2(\delta(\mu_n))} (1 + r_n) \right)
$$

where $c_n(\eta) := n^{-1/2} \alpha_n^{1/2} \left\| \frac{p_{\eta^t}}{p_{\delta(\mu_n)}} - 1 \right\|_{L^2(\delta(\mu_n))}^{-1}$. As soon as $r_n > -1$, which happens with probability tending to one, the supremum of the inner term over $t \geq 0$ is attained at $t = 0$ if $\sum_{i=1}^{n} s_{\eta,\mu_n}(X_i) \leq 0$.
and at
\[ t_n(\eta) := \frac{n^{-1} \sum_{i=1}^{n} s_{\eta, \mu_n}(X_i)}{(1 + r_n) \left\| \frac{p_n}{p_{\delta(\mu_n)}} - 1 \right\|_{2, \delta(\mu_n)}} \]

if \( \sum_{i=1}^{n} s_{\eta, \mu_n}(X_i) > 0 \). Because of (2.8.1) it follows that \( t_n(\eta) \leq c_n(\eta) \) with probability tending to one, so that taken together we have

\[
\sup_{\eta \in M_{n2}} \left( \left\| \frac{p_n}{p_{\delta(\mu_n)}} - 1 \right\|_{2, \delta(\mu_n)} \sum_{i=1}^{n} s_{\eta, \mu_n}(X_i) - \frac{n}{2} \right) \left\| \frac{p_n}{p_{\delta(\mu_n)}} - 1 \right\|_{2, \delta(\mu_n)} (1 + r_n) = \frac{1}{2} \sup_{\eta \in G} \left( \max \left\{ n^{-1/2} \sum_{i=1}^{n} s_{\eta, \mu_n}(X_i), 0 \right\} \right)^2 + o_p(1) \]

Combining this with (2.8.3) yields

\[
\sup_{\eta \in G} \ell_n(\eta) - \ell_n(\delta(\mu_n)) = \frac{1}{2} \sup_{\eta \in G} \left( \max \left\{ n^{-1/2} \sum_{i=1}^{n} s_{\eta, \mu_n}(X_i), 0 \right\} \right)^2 + o_p(1). \tag{2.8.4}
\]

Recall that for each \( \eta \in G \) there exists \( \tilde{\eta} \in G^n \) such that \( \eta = \tilde{\eta} \). Thus

\[
\left| \sup_{\eta \in G} \left( \max \left\{ n^{-1/2} \sum_{i=1}^{n} s_{\nu_n, \delta(\mu)}(X_i), 0 \right\} \right)^2 - \sup_{\eta \in G} \left( \max \left\{ n^{-1/2} \sum_{i=1}^{n} s_{\nu_n, \mu_n}(X_i), 0 \right\} \right)^2 \right| \leq \sup_{\nu \in G^n} \inf_{\eta \in G} \left| \left( n^{-1/2} \sum_{i=1}^{n} s_{\nu_n, \mu_n}(X_i) \right)^2 - \left( n^{-1/2} \sum_{i=1}^{n} s_{\nu_n, \mu_n}(X_i) \right)^2 \right| \leq \sup_{\nu \in G^n} \inf_{\eta \in G} \left| G_n^2(\nu) - G_n^2(\eta) \right| \leq 2 \left( \sup_{\nu \in G^n} \left| G_n(\nu) \right| \right) \left( \sup_{\eta \in G} \inf_{\eta \in G} \left| G_n(\nu) - G_n(\eta) \right| \right) = o_p(1) \tag{2.8.5}
\]

The \( o_p(1) \) in last line above follows from Assumption 2.3.1. More precisely, note that by the Continuous Mapping Theorem applied to the map \( f \mapsto \sup_{\eta \in G^n} \inf_{\tilde{\eta} \in G} |f(\eta) - f(\tilde{\eta})| \) we have for any fixed \( \epsilon > 0 \)

\[ \sup_{\eta \in G} \inf_{\tilde{\eta} \in G} |G_n(\eta) - G_n(\tilde{\eta})| \sim \sup_{\eta \in G} \inf_{\tilde{\eta} \in G} |G(\eta) - G(\tilde{\eta})|. \]

Thus for arbitrary \( \epsilon > 0, t > 0 \) we have

\[ \lim_{n \to \infty} \Pr \left( \sup_{\eta \in G^n} \inf_{\tilde{\eta} \in G} |G_n(\eta) - G_n(\tilde{\eta})| \leq t \right) \leq \Pr \left( \sup_{\eta \in G} \inf_{\tilde{\eta} \in G} |G(\eta) - G(\tilde{\eta})| \leq t \right). \]
and the right-hand side can be made arbitrarily small by letting \( \epsilon \downarrow 0 \). This shows that

\[
\sup_{v \in \mathcal{G}, n} \inf_{\eta \in \mathcal{G}} |G_n(v) - G_n(\eta)| = o_P(1).
\]

Now equations (2.8.4), (2.8.5) yield

\[
2 \sup_{\eta \in \mathcal{G}} (\ell_n(\eta) - \delta_n(\mu_n)) = \sup_{\eta \in \mathcal{G}} \left( \max_{\eta \in \mathcal{G}} \{ G_n(\eta), 0 \} \right)^2 + o_P(1),
\]

and the assertion of the Theorem follows.

\[\square\]

**Proof of Theorem 2.3.5.** First we observe that \( \mathcal{G} \) is the of \( G_n \) under weak convergence in \( \ell^\infty(\mathcal{G} \setminus \delta(\mu_0)) \) and thus tight. Under the assumptions that \( E[G(\delta(\mu_0 + 1/n) - \delta(\mu_0 - 1/n))] \to -1 \) (this follows under classical differentiability conditions), we have \( \sup_{\eta} G(\eta) > 0 \) a.s. so that \( \max(0, \sup_\eta G(\eta)) = \sup_\eta G(\eta) \) a.s. The proof of the assertion consists of three steps. First, we show that the distribution of \( R \) is continuous on \((0, \infty)\) (Claim 2). Second, we provide a lower bound for \( P(R > 0) \). Define

\[
F_y(t) := P \left( \sup_\eta G_n \leq t \big| Y_1 = y \right).
\]

**Claim 1:** For any \( y \in \mathbb{R} \), \( F_y(\cdot) \) is continuous on \((|y|, \infty)\).

Observe that by the joint normality of \((G_\eta)_{\eta \in \mathcal{G}}, Y_1 \) the conditional distribution of \((G_\eta)_{\eta \in \mathcal{G}} \) given \( Y_1 = y \) is that of a tight Gaussian random element with mean \( y\rho_\eta \) where \( \rho_\eta := E[G_\eta Y_1] \) and a covariance function \( \kappa \) that does not depend on \( y \). Let \( \hat{G} \) denote a centered Gaussian process with covariance function \( \kappa \). Then the conditional distribution of \( G \) given \( Y_1 = y \) and the distribution of \((\hat{G}(\eta) + y\rho_\eta)_{\eta \in \mathcal{G}} \) coincide.

Since \( \hat{G} \) is a centered, tight Gaussian process, it follows by the arguments given on page 60-61 of Ledoux and Talagrand (1991) that \( \sup_\eta |\hat{G}_\eta| \) has a continuous distribution on \( \mathbb{R} \) with left support point at 0, that is \( P(\sup_\eta |\hat{G}_\eta| < \epsilon) > 0 \) for all \( \epsilon > 0 \). Since \( P(\sup_\eta \hat{G}_\eta < \epsilon) \geq P(\sup_\eta |\hat{G}_\eta| < \epsilon) \) it follows that also \( P(\sup_\eta \hat{G}_\eta < \epsilon) > 0 \) for all \( \epsilon > 0 \).

According to Tsirel’son (1976), the distribution of \( \sup_\eta (y\rho_\eta + \hat{G}_\eta) \) can only have a jump at the left endpoint of it’s support and has a density to the right of that point. Moreover recall that \( |\rho_\eta| \leq 1 \), for \( y > 0, \epsilon > 0 \)

\[
P(\sup_\eta (y\rho_\eta + \hat{G}_\eta) - y \leq \epsilon) = P(\sup_\eta (y(\rho_\eta - 1) + \hat{G}_\eta) \leq \epsilon) \geq P(\hat{G}(\eta) \leq \epsilon) > 0.
\]

Thus for \( y > 0 \) the distribution of \( \sup_\eta (y\rho_\eta + \hat{G}_\eta) \) has a density on \((y, \infty)\) as by the computation above its left support point is bounded from below by 0. Similarly for \( y < 0 \) we obtain

\[
P(\sup_\eta (y\rho_\eta + \hat{G}_\eta) - |y| \leq \epsilon) = P(\sup_\eta (y\rho_\eta - |y| + \hat{G}_\eta) \leq \epsilon) \geq P(\hat{G}(\eta) \leq \epsilon) > 0.
\]
Thus for all \( y \in \mathbb{R} \) the distribution of \( \sup_{\eta}(y \rho_{\eta} + \tilde{G}_{\eta}) \) has density on \( (|y|, \infty) \) and Claim 1 follows.

**Claim 2:** The distribution of \( (\sup_{\eta} G_{\eta})^2 - Y_1^2 \) is continuous on \( (0, \infty) \).

Let \( 0 < a < b \). Then by continuity of \( F_y \) on \( (|y|, \infty) \)

\[
P\left( (\sup_{\eta} G_{\eta})^2 - Y_1^2 \in [a, b] \right) = \int_\mathbb{R} P\left( (\sup_{\eta} G_{\eta})^2 - y^2 \in [a, b] \mid Y_1 = y \right) \phi(y) \, dy
\]

\[
= \int_\mathbb{R} \left( F_y((y^2 + b)^{1/2}) - F_y((y^2 + a)^{1/2}) \right) \phi(y) \, dy.
\]

Now for \( a \uparrow b > 0 \) we have for every \( y \in \mathbb{R} \) that \( F_y((y^2 + b)^{1/2}) - F_y((y^2 + a)^{1/2}) \to 0 \) since \( (y^2 + b)^{1/2} > |y| \) is a continuity point of \( F_y \). Thus the integral converges to zero by dominated convergence. Since \( b > 0 \) was arbitrary the assertion follows.

**Claim 3:** \( P((\sup_{\eta} G_{\eta})^2 - Y_1^2 > 0) \geq 1/4. \)

By assumption there exists \( \eta_0 \in \mathcal{G} \) such that \( |E[G_{\eta_0} Y_1]| \neq 1 \). Moreover,

\[
P((\sup_{\eta} G_{\eta})^2 - Y_1^2 > 0) \geq P(|G_{\eta_0}| > |Y_1|) = 1/4.
\]

Here, the last inequality follows since \( (G_{\eta_0}, Y_1) \) is a two-dimensional Gaussian vector with non-degenerate distribution.

The assertion follows by combining Claim 2 and Claim 3.

\[ \square \]

**Verification of Assumption 2.3.1 - 2.3.3**

**Example 2.8.1. Location mixture of normals**

Assume that \( \Theta = [a, b] \) for some \( a < 0 < b \) and that the densities \( p \) take the form \( p(x|\mu) = (2\pi)^{-1/2} \exp((-x - \mu)^2/2) \).

In this setting \( \Theta^\varepsilon = [a - \varepsilon, b + \varepsilon] \). Assume without loss of generality that \( \mu_0 = 0 \). Following the arguments in Azaïs, Gassiat, and Mercadier (2009), we find that the score functions \( s_{\eta, \mu} \) can be represented as

\[
s_{\eta, \mu}(x) = \sum_{k=1}^{\infty} \frac{E[(Z - \mu)^k]}{k!} H_k(x - \mu) = \left( \sum_{k=1}^{\infty} \frac{E[(Z - \mu)^k]}{k!} \right)^{1/2}
\]

where \( Z \sim \eta \) and \( (H_k)_{k \in \mathbb{N}} \) denote the Hermite polynomials. In particular, we see that for \( X \sim N(\mu_0, 1) \) we have \( s_{\eta_n, \mu_0}(X) \overset{d}{=} s_{\eta_n, 0}(Y) \) with \( Y \sim N(0, 1) \) (this follows from the definition of \( s_{\eta_n, \mu_0} \)).
Thus the distribution of \((G_n(\eta))_{\eta \in \Theta}\) is equal to that of the process \((\tilde{G}_n(\eta))_{\eta \in \Theta}\) where

\[
\tilde{G}_n(\eta) := n^{-1/2} \sum_{i=1}^{n} s_{\eta,0}(Y_i)
\]

with \(Y_1, ..., Y_n\) i.i.d. \(\sim N(0,1)\). Thus the arguments from the proof of Theorem 3 in Azaïs, Gassiat, and Mercadier (2009) yield \(G_n \sim G\) in \(\ell^\infty(\mathcal{G})\) where the limiting process \(G\) is Gaussian and has a covariance structure of the form

\[
E[G(\eta_1)G(\eta_2)] = \frac{E[\exp(Z_1\tilde{Z}_2)] - 1}{(E[\exp(Z_1\tilde{Z}_1)] - 1)^{1/2}(E[\exp(Z_2\tilde{Z}_2)] - 1)^{1/2}}
\]

where \(Z_1,\tilde{Z}_1 \sim \eta_1, Z_2,\tilde{Z}_2 \sim \eta_2\) and \(Z_1, Z_2, \tilde{Z}_1, \tilde{Z}_2\) are independent. To prove the second part of Assumption 2.3.1, consider the following construction. To each random variable \(Z\) on \([a - \epsilon, b + \epsilon]\) define a transformed random variable \(Y\) through

\[
Y := ZI(Z \in [a, b]) + \frac{M}{M + \epsilon} ZI(Z \notin [a, b]).
\]

where \(M := \min(|a|, b)\). By construction, the support of \(Y\) is contained in \([a, b]\). Denoting the distribution of \(Y\) by \(\nu_{\eta, \epsilon}\), straightforward but tedious calculations show that

\[
\sup_{\eta \in \mathcal{G}} E[(G(\eta) - G(\nu(\eta, \epsilon)))^2] = o(1)
\]

as \(\epsilon \downarrow 0\). By the uniform continuity of the process \(G\) with respect to the metric \(d(\eta, \nu) := (E[(G(\eta) - G(\nu))^2])^{1/2}\) induced by its covariance [see Example 1.5.10 in van der Vaart and Wellner (1996)], this shows that the second part Assumption 2.3.1 also holds.

To verify Assumption 2.3.2, observe that \(\tilde{G}_n\) can be identified with the empirical process based on the observations \(Y_1, ..., Y_n\) and indexed by the class of functions \(\mathcal{F} := \{s_{\eta,0}\eta \in \Theta\}\). Weak convergence of \(\tilde{G}_n\) implies that the class \(\mathcal{F}\) is Donsker, and thus \(\mathcal{F}_2\) is Glivenko-Cantelli [see Lemma 2.10.4 in van der Vaart and Wellner (1996)]. Moreover, since \(\mathcal{F}\) is Donsker so is \(\mathcal{F}_- := \{s_{\eta,0,-}\eta \in \Theta\}\), and thus \(\mathcal{F}_2\) is also Glivenko-Cantelli. This shows that Assumption 2.3.2 holds.

Finally, Assumption 2.3.3 follows by the arguments in the proof of Theorem 3 in Azaïs, Gassiat, and Mercadier (2009).

**Example 2.8.2. Mixture of Poisson**

Assume that \(\Theta = [a, b]\) for some \(0 < a < b\) and that the densities \(p\) take the form \(p(k|\mu) = \mu^k e^{-\mu}/k!\) with respect to the counting measure on \(\mathbb{N}\). As stated in Section 3.3 of Azaïs, Gassiat, and Mercadier (2009), the likelihood ratios have the following representation

\[
\frac{p_\eta(x)}{p_\delta(\mu_n)(x)} - 1 = \sum_{k=1}^{\infty} \frac{kE[(Z - \mu_n)^k]}{(k! \mu_n^k)^{1/2}} \frac{C_k(x|\mu_n)}{k} =: \sum_{k=1}^{\infty} \frac{a_k(\mu_n, \eta) C_k(x|\mu_n)}{k}
\]

(2.8.6)
where \( Z \sim \eta \). Here, the functions \( x \mapsto C_k(x|\mu_n) \) are polynomials of order \( k \) which are given by

\[
C_k(x|\mu_n) := \left[ \frac{\mu_n^{k/2}}{(k!)^{1/2}} \right] \frac{d^k}{dz^k} \left( \frac{z}{\mu_n} \right)^x \exp(-z + \mu_n) \bigg|_{z=\mu_n}.
\]

The functions \( (x \mapsto C_k(x|\mu_n))_{k \in \mathbb{N}} \) are centered and orthonormal with respect to \( P_{\delta(\mu_n)} \), i.e. for \( k, \ell \in \mathbb{N} \)

\[
\mathbb{E}_{\mu_n}[C_k(X|\mu_n)] = 0, \quad \mathbb{E}_{\mu_n}[C_k(X|\mu_n)C_\ell(X|\mu_n)] = I[k = \ell]. \tag{2.8.7}
\]

In particular, we have that \( 1 = \mathbb{E}_{\mu_n}C_k^2(X|\mu_n) = \sum_{u \geq 0} C_k^2(u|\mu_n)e^{-\mu_n u}/u! \geq C_k^2(x|\mu_n)e^{-\mu_n x}/x! \) for any \( x \in \mathbb{N}_0 \) so that the series in (2.8.6) converges pointwise. The score functions \( s_{\eta_n,\mu_n} \) can be represented as

\[
s_{\eta_n,\mu_n}(x) = \sum_{k=1}^{\infty} \frac{a_k(\eta_n, \mu_n) C_k(x|\mu_n)}{\kappa W(\eta_n, \mu_n)}, \quad w(\eta_n, \mu_n) := \left( \sum_{\ell=1}^{\infty} \ell^{-2} a_\ell^2(\eta_n, \mu_n) \right)^{1/2}. \tag{2.8.8}
\]

For \( L \geq 2 \), define the approximating function

\[
s_{\eta_n,\mu_n}^{(L)}(x) = \sum_{k=1}^{L} \frac{a_k(\eta_n, \mu_n) C_k(x|\mu_n)}{\kappa W^{(L)}(\eta_n, \mu_n)}, \quad w^{(L)}(\eta_n, \mu_n) := \left( \sum_{\ell=1}^{L} \ell^{-2} a_\ell^2(\eta_n, \mu_n) \right)^{1/2}.
\]

Obviously, the function \( x \mapsto s_{\eta_n,\mu_n}^{(L)}(x) \) is a polynomial of degree \( L \). Later, we will prove the following identities holding for \( L \geq 2 \), some finite \( n_0 \) and a constant \( C \) independent of \( n, \eta, \mu \)

\[
\sup_{\eta \in \mathcal{G}^c} \sup_{n \geq n_0} \left| \frac{w^{(L)}(\eta_n, \mu_n)}{w(\eta_n, \mu_n)} - 1 \right| \leq CL^{-1}, \quad \sup_{\eta \in \mathcal{G}^c} \left| \frac{w^{(L)}(\eta, \mu)}{w(\eta, \mu)} - 1 \right| \leq CL^{-1}, \tag{2.8.9}
\]

\[
\sum_{k \geq 2} \frac{[\mathbb{E}_\eta[(Z - \mu_n)^k]/k!\mu_n^k]^2}{\kappa W^{(L)}(\eta_n, \mu_n)^2} \leq C([\mathbb{E}_\eta[(Z - \mu_n)^2]])^2. \tag{2.8.10}
\]

Additionally, for any fixed \( k \) one obtains by straightforward calculations

\[
\sup_{\eta \in \mathcal{G}^c} |a_k(\eta_n, \mu_n) - a_k(\eta, \mu_0)| \to 0, \quad n \to \infty, \tag{2.8.11}
\]

and for any fixed \( L \geq 2 \) [this will be proved later]

\[
\sup_{\eta \in \mathcal{G}} \left| \frac{w^{(L)}(\eta_n, \mu_n)}{w^{(L)}(\eta, \mu_0)} - 1 \right| \to 0, \quad n \to \infty. \tag{2.8.12}
\]

We now prove the first part of Assumption 2.3.1. For \( L \geq 2 \) define

\[
G^{(L)}(\eta) := \sum_{k=1}^{L} \frac{a_k(\eta, \mu_0) Z_k}{\kappa W^{(L)}(\eta, \mu_0)}, \quad G(\eta) := \sum_{k=1}^{\infty} \frac{a_k(\eta, \mu_0) Z_k}{\kappa W(\eta, \mu_0)}
\]

where \( Z_1, Z_2, \ldots \) i.i.d. \( \sim N(0, 1) \). By an application of Lemma B.1 from Bücher, Dette, and Volgushev
(2011) it suffices to prove the following claims:

(i) For every $L \geq 2$ we have $(G_n s_{\eta, \mu_n}^{(L)})_{\eta \in \mathcal{G}} \rightsquigarrow (G^{(L)})_{\eta \in \mathcal{G}}$ as $n \to \infty$.

(ii) $G^{(L)} \rightsquigarrow G$ as $L \to \infty$.

(iii) For every $\epsilon > 0$ we have $\lim_{L \to \infty} \limsup_{n \to \infty} P^\epsilon \left( \sup_{\eta \in \mathcal{G}} |G_n s_{\eta, \mu_n}^{(L)} - G_n s_{\eta, \mu_n}| = 0 \right).$

For a proof of the third assertion note that

$G_n s_{\eta, \mu_n}^{(L)} - G_n s_{\eta, \mu_n} = \left(1 - \frac{w(\eta, \mu_n)}{w(L)(\eta, \mu_n)}\right) \sum_{k=1}^{\infty} \frac{a_k(\eta, \mu_n)}{k w(\eta, \mu_n)} G_n C_k(\cdot | \mu_n) + \frac{w(\eta, \mu_n)}{w(L)(\eta, \mu_n)} \sum_{k=L+1}^{\infty} \frac{a_k(\eta, \mu_n)}{k w(\eta, \mu_n)} G_n C_k(\cdot | \mu_n).$

The first term in the above decomposition can be bounded as follows

$$\sup_{\eta \in \mathcal{G}} |A_n^{(L)}(\eta, \mu_n)| \leq \sup_{\eta \in \mathcal{G}} \left| \left(1 - \frac{w(\eta, \mu_n)}{w(L)(\eta, \mu_n)}\right) \sum_{k=1}^{\infty} \frac{a_k(\eta, \mu_n)}{k w(\eta, \mu_n)} G_n C_k(\cdot | \mu_n) \right| \leq CL^{-1} \left( \sum_{k=1}^{\infty} \frac{(G_n C_k(\cdot | \mu_n))^2}{k^2} \right)^{1/2} \sup_{\eta \in \mathcal{G}} \left( \sum_{k=1}^{\infty} \frac{a_k^2(\eta, \mu_n)}{w^2(\eta, \mu_n)} \right)^{1/2} \leq CL^{-1} \left( \sum_{k=1}^{\infty} \frac{(G_n C_k(\cdot | \mu_n))^2}{k^2} \right)^{1/2}.$$

Since $E[(G_n C_k(\cdot | \mu_n))^2] = 1$ for all $k \in \mathbb{N}$ by the orthonormality of the $(C_k(\cdot | \mu_n))_{k \in \mathbb{N}}$, we obtain

$$\lim_{L \to \infty} \limsup_{n \to \infty} E \left| \sup_{\eta \in \mathcal{G}} A_n^{(L)}(\eta, \mu_n) \right|^2 = 0.$$

By similar arguments as above we also obtain the bound

$$\sup_{\eta \in \mathcal{G}} |B_n^{(L)}(\eta, \mu_n)| \leq C_1 \left( \sum_{k=L+1}^{\infty} \frac{(G_n C_k(\cdot | \mu_n))^2}{k^2} \right)^{1/2} \sup_{\eta \in \mathcal{G}} \left( \frac{w(\eta, \mu_n)}{w(L)(\eta, \mu_n)} \right) \leq C_2 \left( \sum_{k=L+1}^{\infty} \frac{(G_n C_k(\cdot | \mu_n))^2}{k^2} \right)^{1/2},$$

where the last inequality holds for $n$ sufficiently large. Thus

$$\lim_{L \to \infty} \limsup_{n \to \infty} E \left| \sup_{\eta \in \mathcal{G}} B_n^{(L)}(\eta, \mu_n) \right|^2 \leq \lim_{L \to \infty} C_2 \sum_{k=L+1}^{\infty} \frac{1}{k^2} = 0.$$

Thus assertion (iii) follows. Assertion (ii) can be proved by similar arguments with $Z_k$ replacing
Thus (i)-(iii) are established we see that the first assertion in Assumption 2.3.1 holds and the arguments are omitted for brevity. For the proof of assertion (i), note that for any fixed \( L \) it is easy to verify that

\[
(G_n C_1(\cdot|\mu_n), ..., G_n C_L(\cdot|\mu_n)) \rightsquigarrow (Z_1, ..., Z_L).
\]

To see this, recall that the \( C_k(\cdot|\mu_n) \) are polynomials and that the coefficients of \( C_k(\cdot|\mu_0) \) converge to those of \( C_k(\cdot|\mu_0) \). Weak convergence of \( (G_n s^{(L)}_{n,\mu_n})_{\eta \in \mathcal{G}^e} \) follows by the extended continuous mapping theorem [see Theorem 1.11.1 in van der Vaart and Wellner (1996)] applied to the maps

\[
g_n : (x_1, ..., x_L) \mapsto \left( \sum_{k=1}^{L} \frac{a_k(\eta, \mu_n) x_k}{kw(L)(\eta, \mu_n)} \right)_{\eta \in \mathcal{G}^e}, \quad g : (x_1, ..., x_L) \mapsto \left( \sum_{k=1}^{L} \frac{a_k(\eta, \mu_0) x_k}{kw(L)(\eta, \mu_0)} \right)_{\eta \in \mathcal{G}^e}.
\]

Thus (i)-(iii) are established we see that the first assertion in Assumption 2.3.1 holds and the limiting Gaussian process \( G \) has the following covariance structure,

\[
\mathbb{E}[G(\eta_1)G(\eta_2)] = \frac{\mathbb{E}\left[\exp\left(\frac{L_1 - \mu_1}{L_2 - \mu_2}\right)\right] - 1}{(\mathbb{E}\left[\exp\left(\frac{L_1 - \mu_1}{L_2 - \mu_2}\right)\right] - 1)^{1/2}} \frac{\mathbb{E}\left[\exp\left(\frac{(L_1 - \mu_1)(L_2 - \mu_2)}{\mu_2}\right)\right] - 1}{(\mathbb{E}\left[\exp\left(\frac{(L_1 - \mu_1)(L_2 - \mu_2)}{\mu_2}\right)\right] - 1)^{1/2}}
\]

where \( Z_1, \tilde{Z}_1 \sim \eta_1, Z_2, \tilde{Z}_2 \sim \eta_2 \) and \( Z_1, Z_2, \tilde{Z}_1, \tilde{Z}_2 \) are independent. The second assertion in Assumption 2.3.1 can now be proved by arguments similar to those in Example 2.8.1.

Next, let us verify Assumption 2.3.2. Consider the following decomposition

\[
\mathbb{E} \sup_{\eta \in \mathcal{G}^e} \left| \frac{1}{n} \sum_{i=1}^{n} s^{(L)}_{\eta_n,\mu_n}(X_i) - (s_{\eta_n,\mu_n})^2(X_i) \right| \leq \mathbb{E} \sup_{\eta \in \mathcal{G}^e} \left| \frac{1}{n} \sum_{i=1}^{n} [s_{\eta_n,\mu_n}(X_i) - s^{(L)}_{\eta_n,\mu_n}(X_i)] [s_{\eta_n,\mu_n}(X_i) + s^{(L)}_{\eta_n,\mu_n}(X_i)] \right| \leq \mathbb{E} \left[ \left( \sup_{\eta \in \mathcal{G}^e} \frac{1}{n} \sum_{i=1}^{n} [s_{\eta_n,\mu_n}(X_i) - s^{(L)}_{\eta_n,\mu_n}(X_i)]^2 \right)^{1/2} \right] \leq \mathbb{E} \left[ \left( \sup_{\eta \in \mathcal{G}^e} \frac{1}{n} \sum_{i=1}^{n} [s_{\eta_n,\mu_n}(X_i) - s^{(L)}_{\eta_n,\mu_n}(X_i)]^2 \right)^{1/2} \right] \leq \mathbb{E} \left[ \sup_{\eta \in \mathcal{G}^e} \frac{1}{n} \sum_{i=1}^{n} [s_{\eta_n,\mu_n}(X_i) + s^{(L)}_{\eta_n,\mu_n}(X_i)]^2 \right]. \tag{2.8.13}
\]
Moreover, for \(n\) sufficiently large and some constants \(C_2, \tilde{C}\)

\[
\sup_{\eta \in \mathcal{G}^c} |s_{\eta_n, \mu_n}(X_i) - s_{\eta_n, \mu_n}^{(L)}(X_i)| \leq \sup_{\eta \in \mathcal{G}^c} \left| 1 - \frac{w(\eta, \mu_n)}{w^{(L)}(\eta, \mu_n)} \right| \sqrt{\sum_{k=1}^{\infty} \frac{a_k(\eta, \mu_n)}{k^2 w(\eta, \mu_n)} C_k(X_i | \mu_n)} \\
+ \sup_{\eta \in \mathcal{G}^c} \left| \frac{w(\eta, \mu_n)}{w^{(L)}(\eta, \mu_n)} \right| \sqrt{\sum_{k=L+1}^{\infty} \frac{a_k^2(\eta, \mu_n)}{k^2 w^2(\eta, \mu_n)}} \frac{C_k^2(X_i | \mu_n)}{k^{1/2}} \\
\leq \sup_{\eta \in \mathcal{G}^c} \left| 1 - \frac{w(\eta, \mu_n)}{w^{(L)}(\eta, \mu_n)} \right| \sqrt{\sum_{k=1}^{\infty} \frac{a_k^2(\eta, \mu_n)}{k^2 w^2(\eta, \mu_n)}} \frac{C_k^2(X_i | \mu_n)}{k^{1/2}} \\
+ \sup_{\eta \in \mathcal{G}^c} \left| \frac{w(\eta, \mu_n)}{w^{(L)}(\eta, \mu_n)} \right| \sqrt{\sum_{k=L+1}^{\infty} \frac{a_k^2(\eta, \mu_n)}{k^2 w^2(\eta, \mu_n)}} \frac{C_k^2(X_i | \mu_n)}{k^{1/2}} \\
\leq \tilde{C} L^{-1/2} \left( \sum_{k=1}^{\infty} \frac{C_k^2(X_i | \mu_n)}{k^{1/2}} \right) + C_2 \left( \sum_{k=L+1}^{\infty} \frac{C_k^2(X_i | \mu_n)}{k^{1/2}} \right).
\]

The last identity shows that for some constant \(C_3\) and \(n\) sufficiently large

\[
\mathbb{E} \sup_{\eta \in \mathcal{G}^c} |s_{\eta_n, \mu_n}(X_i) - s_{\eta_n, \mu_n}^{(L)}(X_i)|^2 \leq C_3 \left( L^{-2} + \sum_{k=L+1}^{\infty} \frac{1}{k^2} \right). \tag{2.8.14}
\]

Combining Assumption 2.3.3 with (2.8.13) and (2.8.14) shows that

\[
\limsup_{n \to \infty} \mathbb{E} \sup_{\eta \in \mathcal{G}^c} \left| \frac{1}{n} \sum_{t=1}^{n} s_{\eta_n, \mu_n}^2(X_t) - (s_{\eta_n, \mu_n}^{(L)})^2(X_t) \right| \leq C_4 \left( L^{-2} + \sum_{k=L+1}^{\infty} \frac{1}{k^2} \right). \tag{2.8.15}
\]

Next, observe that by construction we have \(\mathbb{E}[(s_{\eta_n, \mu_n}^{(L)})^2(X_i)] = 1\) for all \(n \in \mathbb{N}, L \geq 2, \eta \in \mathcal{G}^c\).

Moreover simple arguments show that for every fixed \(k, l \in \mathbb{N}\)

\[
\frac{1}{n} \sum_{t=1}^{n} C_k(X_t | \mu_n) C_l(X_t | \mu_n) \overset{P}{\to} I(k = l).
\]

By the extended continuous mapping theorem [see Theorem 1.11.1 in van der Vaart and Wellner (1996)] applied to the maps

\[
g_n : (x_{kl})_{k, l=1, \ldots, L} \mapsto \left( \sum_{k, l=1}^{L} \frac{a_k(\eta_n, \mu_n) a_l(\eta_n, \mu_n) x_{kl}}{k! l! w^{(L)}(\eta_n, \mu_n)^2} \right)_{\eta \in \mathcal{G}^c}
\]

\[
g : (x_{kl})_{k, l=1, \ldots, L} \mapsto \left( \sum_{k, l=1}^{L} \frac{a_k(\eta, \mu_0) a_l(\eta, \mu_0) x_{kl}}{k! l! w^{(L)}(\eta, \mu_0)^2} \right)_{\eta \in \mathcal{G}^c}
\]

it follows that for every \(L \geq 2\)

\[
\sup_{\eta \in \mathcal{G}^c} \left| \frac{1}{n} \sum_{t=1}^{n} ((s_{\eta_n, \mu_n}^{(L)})^2(X_t) - 1) \right| = o_P(1).
\]
Combining this with (2.8.15) proves the first assertion in Assumption 2.3.2. To establish the second part of Assumption 2.3.2, note that for \( x, y \in \mathbb{R} \) we have \(|x_\cdot - y_\cdot| \leq |x - y|\). Thus

\[
\sup_{\eta \in S^c} \left| \frac{1}{n} \sum_{i=1}^{n} s_{\eta, \mu_n,-}(X_i) - (s_{\eta, \mu_n,-}^{(L)})^2(X_i) \right| \leq \sup_{\eta \in S^c} \left| \frac{1}{n} \sum_{i=1}^{n} (s_{\eta, \mu_n,-}(X_i) - s_{\eta, \mu_n,-}^{(L)}(X_i))^2 \right|^{1/2} \left| \frac{1}{n} \sum_{i=1}^{n} (s_{\eta, \mu_n,-}(X_i) + s_{\eta, \mu_n,-}^{(L)}(X_i))^2 \right|^{1/2} \leq \sup_{\eta \in S^c} \left| \frac{1}{n} \sum_{i=1}^{n} (s_{\eta, \mu_n}(X_i) - s_{\eta, \mu_n}^{(L)}(X_i))^2 \right|^{1/2} \left| \frac{2}{n} \sum_{i=1}^{n} 8s_{\eta, \mu_n}(X_i)^2 + 2s_{\eta, \mu_n}(X_i) - s_{\eta, \mu_n}^{(L)}(X_i)^2 \right|^{1/2} .
\]

This combined with (2.8.14) and Assumption 2.3.3 yields

\[
\limsup_{n \to \infty} \mathbb{E} \sup_{\eta \in S^c} \left| \frac{1}{n} \sum_{i=1}^{n} s_{\eta, \mu_n,-}^2(X_i) - (s_{\eta, \mu_n,-}^{(L)})^2(X_i) \right| \leq C_4 \left( L^{-2} + \sum_{k=L+1}^{\infty} \frac{1}{k^2} \right). \tag{2.8.16}
\]

Thus it suffices to show that for each fixed \( L \)

\[
\sup_{\eta \in S^c} \left| \frac{1}{n} \sum_{i=1}^{n} (s_{\eta, \mu_n,-}^{(L)})^2(X_i) - ||s_{\eta, \mu_n,-}^{(L)}||^2_{2, \delta(\mu_n)} \right| = o_p(1) \tag{2.8.17}
\]

and that

\[
\lim_{L \to \infty} \limsup_{n \to \infty} \sup_{\eta \in S^c} ||s_{\eta, \mu_n,-}^{(L)}||^2_{2, \delta(\mu_n)} - ||s_{\eta, \mu_0,-}^{(L)}||^2_{2, \delta(\mu_0)} = 0. \tag{2.8.18}
\]

To prove (2.8.17), define \( y^{(L)}(x) := (1, \ldots, x^L) \) observe that there exists a constant \( C \) [note that \( s_{\eta, \mu_n}^{(L)}(x) \) is a polynomial in \( x \) of degree \( L \)] such that

\[
\sup_{\eta \in S^c} \left| \frac{1}{n} \sum_{i=1}^{n} (s_{\eta, \mu_n,-}^{(L)})^2(X_i) - ||s_{\eta, \mu_n,-}^{(L)}||^2_{2, \delta(\mu_n)} \right| \leq \sup_{b \in \mathbb{R}^{L+1}, \|b\| \leq C} \left| \frac{1}{n} \sum_{i=1}^{n} \left( b^T Y^{(L)}(X_i) \right)^2 I[b^T Y^{(L)}(X_i) \leq 0] - \mathbb{E}[(b^T Y^{(L)}(X_i))^2 I[b^T Y^{(L)}(X_i) \leq 0]] \right|
\]

Weak convergence to zero of the right-hand side can be proved after observing that the class of functions \( y \mapsto (b^T y)^2 I[b^T y \leq 0] : \|b\| \leq C \) is VC and has an envelope \( G \) function which satisfies \( \sup_{n \geq n_0} \mathbb{E}G^2(Y^{(L)}(X_i)) < \infty \) for some \( n_0 < \infty \). Thus convergence of the right-hand side above to zero follows from Theorem 2.8.1 in van der Vaart and Wellner (1996).

Next, let us prove (2.8.18). We begin by proving

\[
\limsup_{n \to \infty} \sup_{\eta \in S^c} \left| ||s_{\eta, \mu_n,-}^{(L)}||^2_{2, \delta(\mu_n)} - ||s_{\eta, \mu_0,-}^{(L)}||^2_{2, \delta(\mu_0)} \right| = 0 \tag{2.8.19}
\]

for every fixed \( L \geq 2 \). Convergence to zero of \( \sup_{\eta \in S^c} \left| ||s_{\eta, \mu_n,-}^{(L)}||^2_{2, \delta(\mu_n)} - ||s_{\eta, \mu_0,-}^{(L)}||^2_{2, \delta(\mu_0)} \right| \) follows

70
from the fact that, for \( V_n \sim \text{Pois}(\mu_n) \), we have for some sequence \( \alpha_n = o(1) \)

\[
\sup_{\eta \in G^s} \left\| s^{(L)}_{\eta, \mu_n} \right\|_{2, \delta(\mu_n)}^2 - \left\| s^{(L)}_{\eta, \mu_0} \right\|_{2, \delta(\mu_0)}^2 \leq \sup_{\|a-b\| \leq \alpha_n, \|a\| \leq \epsilon, \|b\| \leq \epsilon} E \left[ (b^T Y^{(L)}(V_n))^2 I(b^T Y^{(L)}(V_n) \leq 0) - (a^T Y^{(L)}(V_n))^2 I(a^T Y^{(L)}(V_n) \leq 0) \right] \leq 2 C \alpha_n E[\|Y^{(L)}(V_n)\|^4] = o(1)
\]

(2.8.20)

where the last inequality follows from \(|x^2 - y^2| \leq (|x| + |y|)(|x| - |y|)\). Similarly, letting \( V \sim \text{Pois}(\mu_0) \), the second term can be bounded by

\[
\sup_{\eta \in G^s} \left\| s^{(L)}_{\eta, \mu_0} \right\|_{2, \delta(\mu_n)}^2 - \left\| s^{(L)}_{\eta, \mu_0} \right\|_{2, \delta(\mu_0)}^2 \leq \sup_{b \in \mathbb{R}^{L+1}, \|b\| \leq \epsilon} E \left[ (b^T Y^{(L)}(V_n))^2 I(b^T Y^{(L)}(V_n) \leq 0) - (b^T Y^{(L)}(V_0))^2 I(b^T Y^{(L)}(V_0) \leq 0) \right].
\]

Covering \( B := \{b \in \mathbb{R}^{L+1} : \|b\| \leq \epsilon\} \) with a finite number of balls of radius \( \epsilon \) one can reduce the above problem to showing that

\[
E[(b^T Y^{(L)}(V_n))^2 I(b^T Y^{(L)}(V_n) \leq 0)] \rightarrow E[(b^T Y^{(L)}(V_0))^2 I(b^T Y^{(L)}(V_0) \leq 0)]
\]

for any fixed \( b \in B \). Observe that \( V_n \) converges weakly to \( V \). The continuous mapping theorem implies that \((b^T Y^{(L)}(V_n))^2 I(b^T Y^{(L)}(V_n) \leq 0) \) converges weakly to \((b^T Y^{(L)}(V_0))^2 I(b^T Y^{(L)}(V_0) \leq 0) \), and by uniform integrability of the sequence \((b^T Y^{(L)}(V_n))^2 I(b^T Y^{(L)}(V_n) \leq 0) \) this implies convergence of the first moment. Together with (2.8.20) this establishes (2.8.19). Finally, the convergence

\[
\lim_{L \to \infty} \sup_{\eta \in G^s} \left\| s^{(L)}_{\eta, \mu_0} \right\|_{2, \delta(\mu_n)}^2 - \left\| s_{\eta, \mu_0} \right\|_{2, \delta(\mu_0)}^2 = 0
\]

can be proved by similar arguments as (2.8.16) with \( n^{-1} \sum \delta \) replaced by the expectation, the details are omitted for the sake of brevity. This completes the proof of Assumption 2.3.2.

Finally, Assumption 2.3.3 can be verified by a straightforward extension of the arguments in the proof of Theorem 4 of Azaïs, Gassiat, and Mercadier (2009).

Thus it remains to establish (2.8.9)-(2.8.12). We begin by noting that for \( Z \sim \eta \) with \( \eta \) having support contained in \([m, M]\) it follows that \(|Z - \mu_n|^k \leq M^{k-2}(Z - \mu_n)^2\) for \( k \geq 3 \). Thus, as soon as \( \mu_n \in [m, M] \), which is the case for \( n \) sufficiently large, we have

\[
\sum_{k \geq 2} \frac{(E_\eta[(Z - \mu_n)^k])^2}{k! \mu_n^k} \leq (E_\eta[(Z - \mu_n)^2])^2 \sum_{k \geq 2} \frac{M^{2k-4}}{k! m^k} \leq C(E_\eta[(Z - \mu_n)^2])^2.
\]

This shows (2.8.10). Next, observe that

\[
\left( \frac{w^{(L)}(\eta, \mu_n)}{w^{(L)}(\eta, \mu_n)} \right)^2 = \sum_{\ell=1}^{L} \ell^{-2} a_{\ell}^2(\eta_n, \mu_n) = 1 + \sum_{\ell=L+1}^{\infty} \ell^{-2} a_{\ell}^2(\eta_n, \mu_n) \leq \sum_{\ell=1}^{\infty} \ell^{-2} a_{\ell}^2(\eta_n, \mu_n).
\]
Now for $Z \sim \eta$ with $\eta$ having support contained in $[m, M]$ we have as soon as $\mu_n \in [m, M]$

$$\sum_{\ell=L+1}^{\infty} \ell^{-2} a_\ell^2(\eta_n, \mu_n) \leq \sum_{k=L+1}^{\infty} \frac{\left(\mathbb{E}_\eta[(Z-\mu_n)^k]\right)^2}{k! \mu_n^k} \leq 2M^2 \sum_{k \geq L+1} \frac{M^{2k-4}}{k! m^k} \leq C \mathbb{L}^{-1}.$$ 

Since $w(\eta, \mu_n), w^{(L)}(\eta, \mu_n)$ are non-negative, the first part of (2.8.9) follows, and the second part of (2.8.9) can be established by exactly the same arguments. Finally,

$$\left( \frac{w^{(L)}(\eta_n, \mu_n)}{w^{(L)}(\eta, \mu_0)} \right)^2 = \sum_{k=1}^L \frac{(\mathbb{E}_{\eta_n}[(Z-\mu_n)^k])^2}{k! \mu_n^k} \frac{(\mathbb{E}_{\eta}[(Z-\mu_0)^k])^2}{k! \mu_0^k}$$

and by construction $\mathbb{E}_{\eta_n}[(Z - \mu_n)^k] = \mathbb{E}_{\eta}[(Z - \mu_0)^k]$ for all $k \in \mathbb{N}$. Now (2.8.12) follows since $\max_{k=1,\ldots,n} |(\mu_n/\mu_0)^k - 1| \to 0$ as $n \to \infty$. This completes all proofs for the Poisson case.
Chapter 3

Unobserved Heterogeneity in Income Dynamics

3.1 Introduction

Unobserved heterogeneity has become a pervasive concern throughout applied econometrics. Longitudinal data presents special opportunities and challenges for models of unobserved heterogeneity; in virtually all econometric applications involving panel data there will be some form of latent, i.e. unobserved, individual specific effects. Classical econometric methods adopt either a differencing strategy designed to purge these effects, or some form of shrinkage method to mitigate their undesirable “incidental parameter” effect. In this chapter we will describe some new nonparametric empirical Bayes methods for estimation and prediction in panel data models with unobserved heterogeneity.

As stressed in recent work of Efron (2010, 2011), empirical Bayes methods pioneered by Robbins (1951, 1956) provide a statistical framework for many contemporary “big data” applications. Although they predate the development of hierarchical Bayes methods exemplified in the work of Lindley and Smith (1972) and Chamberlain and Leamer (1976), they share many common features. The transition from parametric to nonparametric empirical Bayes methods brings exciting new opportunities that greatly expand the flexibility of existing approaches to panel data modeling and its treatment of unobserved heterogeneity.

We will begin with a brief overview of empirical Bayes methods beginning with Robbins (1951). In Section 3.3 we extend the predominant Gaussian location mixture framework to accommodate nonparametric location and scale mixtures with covariates in the classical Gaussian panel data setting, including some simulation evidence to illustrate the performance of the new methods. Section 3.4 describes an extended application to models of heterogeneous income dynamics that illustrates both estimation and prediction aspects of the new methods including, notably, the introduction of a bivariate joint distribution of unobserved heterogeneity and covariate effects via profile likelihood methods.

In sharp contrast to the classical Gaussian hierarchical Bayes framework for panel data, or its frequentist analogues, the nonparametric mixture formulation of our proposed methods offers a much more flexible view of unobserved heterogeneity while preserving most of the virtues of the likelihood formalism.

This chapter is based on joint work with Roger Koenker.
3.2 Empirical Bayes: a brief overview

Given a simple parametric model, there is a natural temptation to complicate it by admitting that those immutable natural constants that constitute the model’s original parameters might instead be random. One of the earliest examples of this type is the classical Gaussian random effects, compound decision problem introduced by Robbins (1951). We observe independent \( Y_1, \ldots, Y_n \) each Gaussian with known, common variance, \( \theta \) but individual specific means, \( Y_i \sim N(\alpha_i, \theta) \). Our objective is to estimate all the \( \alpha_i \)'s subject to squared error loss,

\[
L_2(\hat{\alpha}, \alpha) = \|\hat{\alpha} - \alpha\|_2^2 = \sum_{i=1}^{n} (\hat{\alpha}_i - \alpha_i)^2.
\]

The naive (unbiased) solution would simply set \( \hat{\alpha}_i = Y_i \), but the usual presumption in such circumstances would be that the observations have some common genesis, and consequently we may be able to “borrow strength” from the full sample to improve upon these myopic predictions based on a single observation.

Suppose we believed that the \( \alpha_i \)'s were drawn iid-ly from the distribution, \( F \), so the \( Y_i \)'s would have convolution density \( g(y) = \int \phi((y - \alpha)/\sqrt{\theta})/\sqrt{\theta}dF(\alpha) \): What would the Reverend Bayes advise? Elementary exponential family theory yields the following proposition. Concise proofs of the propositions appear in Section 3.6.

**Proposition 3.2.1.** For \( Y_i \sim N(\alpha_i, \theta) \) and \( \{\alpha_i\} \text{ iid } F \), the Bayes rule under \( L_2 \) loss is:

\[
\delta(y) = y + \theta g'(y)/g(y) \quad (3.2.1)
\]

and \( \delta(y) \) is non-decreasing in \( y \).

Efron (2011) refers to this expression for \( \delta(y) \) as Tweedie’s formula, citing Robbins’s (1956) attribution of it to M.C.K. Tweedie. Tukey (1974) provides an earlier attribution to Arthur Eddington appearing in Dyson (1926). A major objective of this chapter is to explore the consequences of extending this result to longitudinal settings in which we can estimate heterogeneity of scale as well as location.

Of course one may well ask: where did this \( F \) come from? And this question leads us inevitably toward estimation of the density, \( g \), and hence to the empirical Bayes paradigm. When \( F \) comes from a finite dimensional parametric family there are several familiar special cases.

3.2.1 Some parametric examples

1. Suppose \( \sigma^2 = 1 \) and we believed that the \( \mu_i \)'s were iid \( N(0, \sigma_0^2) \), so the \( Y_i \)'s are iid \( N(0, 1 + \sigma_0^2) \), the Bayes rule would be,

\[
\delta(y) = \left(1 - \frac{1}{1 + \sigma_0^2}\right)y.
\]
Thus, we shrink our naive estimates all toward zero. When \( \sigma_0^2 \) is unknown, \( S = \sum Y_i^2 \sim (1 + \sigma_0^2) \chi_n^2 \), and recalling that an inverse \( \chi_n^2 \) random variable has expectation, \( (n - 2)^{-1} \), we obtain the Stein rule in its simplest form:

\[
\hat{\delta}(y) = \left( 1 - \frac{n - 2}{S} \right) y.
\]

2. When, slightly more generally, \( \mu_i \sim N(\mu_0, \sigma_0^2) \) we shrink instead toward the prior mean,

\[
\delta(y) = \mu_0 + \left( 1 - \frac{1}{1 + \sigma_0^2} \right) (y - \mu_0),
\]

and estimating the prior parameters costs us one degree of freedom, so we obtain the celebrated James-Stein estimator,

\[
\hat{\delta}(y) = \bar{Y}_n + \left( 1 - \frac{n - 3}{S} \right) (y - \bar{Y}_n),
\]

for \( \bar{Y}_n = n^{-1} \sum Y_i \) and \( S = \sum (Y_i - \bar{Y}_n)^2 \).

3. If each observation has its own known variance: \( Y_i \sim N(\mu_i, \sigma_i^2) \) and \( \mu_i \sim N(\mu_0, \sigma_0^2) \), as might be plausible in the case that each \( Y_i \) is from a different measuring device each with known precision, or as in the ubiquitous baseball batting average examples, as in Brown (2008) and Jiang and Zhang (2010), in which binomial variances depend upon a known number of “at bats” in the initial period. In such cases we have the Bayes rule,

\[
\delta(y_i) = \mu_0 + \left( 1 - \frac{\sigma_i^2}{\sigma_0^2 + \sigma_i^2} \right) (y_i - \mu_0)
\]

4. Further generalizing, we may wish to replace \( \mu_0 \) by a function of observable covariates, say \( z_i^\top \beta_0 \). Then, as in Jiang and Zhang (2010), we obtain a positive-part James-Stein estimator,

\[
\hat{\delta}(y_i) = \left( 1 - \frac{p - 2}{\sum (z_i^\top \hat{\beta}/\sigma_i)^2} \right) z_i^\top \hat{\beta} + \left( 1 - \frac{n - p - 2}{\sum (y_i - z_i^\top \hat{\beta})^2/\sigma_i^2} \right) (y_i - z_i^\top \hat{\beta})
\]

where \( p \) denotes the dimension of \( \beta \) and \((u)_+ = u \mathbf{1}(u > 0)\).

5. Another important class of examples arises from the assumption of sparsity, that is, an assertion that most of the \( \mu_i \) are probably zero. Johnstone and Silverman (2004) consider a model in which,

\[
dF(\mu) = (1 - w)\delta_0(\mu) + w\varphi_\nu(\mu)
\]

where with probability \((1 - w)\), \( \mu = 0 \), while with probability \( w \) it is drawn from a density, \( \varphi \), with scale parameter, \( \nu \). They compare performance of several hard and soft thresholding rules in addition to empirical Bayes procedures that estimate the parameters \( w \) and \( \nu \). This
is closely related to an extensive recent literature on more formal Bayesian methods for the Gaussian sequence model, e.g. Castillo and van der Vaart (2012).

The simulation designs of Johnstone and Silverman (2004) have served as a benchmark for several more recent studies of empirical Bayes methods including Brown and Greenshtein (2009), Jiang and Zhang (2009), and Koenker and Mizera (2014), all of which explore non-parametric estimation of the Gaussian mixture model.

3.2.2 Non-parametric estimation of the Gaussian mixture model

When we lack confidence in a particular parametric specification of the mixing distribution, $F$, we are faced with a more serious quandary. It is apparent that we need a non-parametric estimate of $F$, and in our Gaussian location mixture setting this is tantamount to solving a deconvolution problem: find $F$ such that the density,

$$g(y) = \int \phi(y - \alpha) dF(\alpha)$$

matches that of the observed $Y_i$’s. Deconvolution is notoriously difficult as shown by Carroll and Hall (1988) and Fan (1991), but before we despair a second look at Tweedie’s formula (3.2.1) reveals that we may not really need an estimate of $F$. We need only estimate the mixture density $g$, a task that can be accomplished at standard univariate non-parametric convergence rates for smooth densities, and smoothness is ensured by the Gaussian convolution whatever $F$ might be.

Kernel density estimation of $g$ as proposed by Brown and Greenshtein (2009) seems to be the natural approach, but in addition to the familiar, but still unsettling, requirement of choosing a bandwidth, kernel estimators of $g$ have the drawback that they do not enforce the monotonicity of the Bayes rule. The latter failing can be addressed by a further monotonization step, or by a penalization approach as suggested in Koenker and Mizera (2014). However, a more direct approach is possible via the Kiefer-Wolfowitz non-parametric maximum likelihood estimator (KWMLE) for the mixture model. This approach was first proposed by Jiang and Zhang (2009) for the Gaussian compound decision problem, suggesting the EM algorithm as a computational strategy.

Although Kiefer and Wolfowitz (1956) established consistency of their MLE for the mixing distribution $F$, it was not until the appearance of Laird (1978) that a viable computational strategy for the estimator was available. The EM algorithm has remained the standard approach for its computation ever since. Heckman and Singer (1984) constitutes an influential econometric application. However, EM has notoriously slow convergence in such applications, and this fact has seriously inhibited the use of the KWMLE in applications. It introduces what is, in effect, a new smoothing parameter into the computational strategy controlled by the stopping criterion of the algorithm. Koenker and Mizera (2014) have recently proposed an alternative computational method for the KWMLE that circumvents these problems. For a broad class of mixture problems, the Kiefer-Wolfowitz estimator can be formulated as a convex optimization problem and solved efficiently by
modern interior point methods. Quicker, more accurate computation of the KWMLE opens the way to a much wider range of applications of the method for models of heterogeneity.

In the next section we will describe how these methods can be adapted to longitudinal data, first for location and scale mixtures separately, then for location-scale mixtures and finally for location scale mixtures with covariate effects. In contrast to compound decision problems with cross sectional data, richer longitudinal data offers new opportunities permitting more complex structures of unobserved heterogeneity.

3.3 Estimating Gaussian mixture models with longitudinal data

Extending the Gaussian compound decision problem with one location parameter per observation to unbalanced longitudinal observations in which we have $m_i$ observations on each individual is quite straightforward. We will describe this relatively simple setting first, and then gradually introduce heterogeneous variance effects, first with independent prior assumptions and then with a general form of bivariate heterogeneity. Estimation of covariate effects via profile likelihood is then introduced. The section concludes with some simulation evidence intended to illustrate our estimation and prediction methods, leading to an extended application of the methods to models of earning dynamics.

Suppose for convenience that we have unit variance for the noise so $u_{it} \sim N(0, 1)$, and we have,

$$y_{it} = \alpha_i + u_{it}, \quad t = 1, \cdots, m_i, \quad i = 1, \cdots, n.$$ 

Sufficiency can be used to reduce the problem to the sample: $\bar{y}_i = m_i^{-1} \sum_{t=1}^{m_i} y_{it} \sim N(\alpha_i, m_i^{-1})$. When the $\alpha_i$’s are iid from $F$, we can write the log likelihood of the observed $y_{it}$’s as,

$$\ell(F|y) = K(y) + \sum_{i=1}^{n} \log(\sqrt{m_i} \int \phi(\sqrt{m_i}(\bar{y}_i - \alpha)) dF(\alpha))$$

Optimizing over an infinite dimensional $F$ necessitates some form of discrete approximation. As in earlier EM implementations, such as that of Jiang and Zhang (2009), we take $F$ to have a piecewise constant (Lebesgue) density on a relatively fine grid containing the empirical support of the observed $\bar{y}_i$’s. Maximizing the likelihood $\ell(F|y)$ generally yields a small number of discrete mass points whose location is determined obviously only up to the scale of the grid. With a few hundred grid intervals we can obtain a quite accurate estimate. Further refinement is always possible as discussed in Koenker and Mizera (2014), but already with a uniform grid with 300 points we have very precise positioning of the mass points of the mixing distribution, more precise than the statistical accuracy of the mass locations would justify. Letting $f_j : j = 1, \cdots, p$ denote the function values of $dF$ on this grid, we can express the constrained maximum likelihood problem as,

$$\max_{f} \left[ \sum_{i=1}^{n} \log(g_i) \mid g = Af, \sum_{j=1}^{p} f_j \Delta_j = 1, \quad f \geq 0 \right], \quad (3.3.1)$$
where \( A = (A_{ij} = \sqrt{m_i} \int \phi(\sqrt{m_i}(y_i - \alpha_j))) \) and \( \Delta_j \) is the \( j \)th grid spacing. As posed, the problem is evidently convex, and therefore has a unique solution. It is well-known, going back to Kiefer and Wolfowitz (1956) and Laird (1978), that variational solutions to the original problem are discrete with fewer than \( n \) atoms. It is somewhat difficult to appreciate this result by viewing EM solutions, since the number of EM iterations required to obtain an accurate solution would test the patience of even the most diligent researchers. But interior point methods make this discreteness easily apparent. Since the number of non-negligible \( \hat{f}_j > 0 \) obtained is typically much smaller than \( n \), often only a handful of points, even in large samples, this also guides our judgement regarding the adequacy of the original grid. As documented in Koenker and Mizera (2014) solving a small problem of this type with \( n = 200 \) and \( p = 300 \) grid points requires about 1 second for the Mosek optimizer and about 10 minutes to achieve a somewhat less precise solution via EM. Ten minutes may not seem prohibitive, but embedding larger problems of this type in profile likelihood settings where many such solutions are required is another story. Dicker and Zhao (2014) have recently shown that grids with \( p = \sqrt{n} \) yield convergence in Hellinger distance of the mixture density at rate \( O_p(\log n/\sqrt{n}) \), the parametric rate modulo the log term. Unfortunately, little is known at this stage about the convergence properties of the mixing distribution beyond the consistency result of Kiefer and Wolfowitz.

The dual formulation of primal problem (3.3.1) has proven to be somewhat more efficient from a computational viewpoint. The dual can be expressed as

\[
\max_{\nu} \left\{ \sum_{i=1}^{n} \log(\nu_i) | A^\top \nu \leq n1_p, \nu \geq 0 \right\}.
\]

This formulation reveals that we are only required to solve for the \( n \)-dimension vector \( \nu \), albeit subject to an infinite dimensional constraint that we have discretized to an \( p \) dimensional grid, see Koenker and Mizera (2014) for further details.

### 3.3.1 Estimating Gaussian scale mixtures

Gaussian scale mixtures can be estimated in much the same way that we have described for location mixtures. Suppose we now observe,

\[ y_{it} = \sqrt{\theta_i} u_{it}, \quad t = 1, \cdots, m_i, \quad i = 1, \cdots, n \]

with \( u_{it} \sim N(0, 1) \). Sufficiency again reduces the sample to \( n \) observations on \( S_i = m_i^{-1} \sum_{t=1}^{m_i} y_{it}^2 \), and thus \( S_i \) has the gamma distribution with shape parameter, \( r_i = m_i/2 \), and scale parameter \( \theta_i/r_i \), i.e.

\[
\gamma(S_i|r_i, \theta_i/r_i) = \frac{1}{\Gamma(r_i)(\theta_i/r_i)^{r_i}} S_i^{r_i-1} \exp\{-S_i r_i/\theta_i\},
\]
and the marginal density of $S_i$ when the $\theta_i$ are iid from $F$ is

$$g(S_i) = \int \gamma(S_i|r_i, \theta/r_i) dF(\theta).$$

To estimate $F$ we can proceed exactly as before except that now the matrix $A$ has typical element $\gamma(S_i|\theta_j)$ for $\theta_j$ on a fine grid covering the support of the sample $S_i$'s.

### 3.3.2 Estimating Gaussian location-scale mixtures

When both location and scale are heterogeneous we must combine the strategies already described. We should stress that modeling heterogeneity of scale parameters would not be possible with cross sectional data since individuals are then only measured once. The model is now,

$$y_{it} = \alpha_i + \sqrt{\theta_i} u_{it}, \; t = 1, \ldots, m_i, \; i = 1, \ldots, n$$

with $u_{it} \sim N(0, 1)$. We will provisionally assume that $\alpha_i \sim F_\alpha$ and $\theta_i \sim F_\theta$ are independent. Again, we have sufficient statistics:

$$\bar{y}_i|\alpha_i, \theta_i \sim N(\alpha_i, \theta_i/m_i)$$

and

$$S_i|r_i, \theta_i \sim \gamma(S_i|r_i, \theta_i/r_i),$$

where $r_i = (m_i - 1)/2$, and the log likelihood becomes,

$$\ell(F_\alpha, F_\theta|y) = K(y) + \sum_{i=1}^{n} \log \int \int \gamma(S_i|r_i, \theta/r_i) \sqrt{m_i} \phi((\sqrt{m_i}(\bar{y}_i - \alpha_i))/\sqrt{\theta})/\sqrt{\theta} dF_\alpha(\alpha) dF_\theta(\theta)$$

Since the scale component of the log likelihood is additively separable from the location component, we can solve for $\hat{F}_\theta$ in a preliminary step, as in the previous subsection, and then solve for the $\hat{F}_\alpha$ distribution. In fact, under the independent prior assumption, we can re-express the Gaussian component of the likelihood as Student-t and thereby eliminate the dependence on $\theta$ in the Kiefer-Wolfowitz problem for estimating $F_\alpha$. This is highly convenient for estimation purposes, however it should be stressed that prediction restores the interdependence on both $F_\alpha$ and $F_\sigma$ as we discuss in more detail below.

When the independent prior assumption is implausible, and this may be typical of many econometric applications like our income dynamics application, we can construct two dimensional grids. This makes the constraint matrix, $A$, a bit more unwieldy, but raises no new issues of principle. If, as in our empirical application to income dynamics, we permit a general bivariate prior for $(\alpha, \theta)$, the Bayes rule for estimating $\alpha$ under $L_2$ loss takes a considerably more complex form summarized in the following result.

**Proposition 3.3.1.** Suppose that $y_{it}|\alpha_i, \theta_i \sim N(\alpha_i, \theta_i)$ and $(\alpha_i, \theta_i)$ are iid from $H(\alpha, \theta)$, then the
Bayes rule for $\alpha$ conditional on the sufficient statistics $\bar{y}_i$ and $S_i$ is

$$E(\alpha|\bar{y}_i, S_i) = \int E(\alpha|\bar{y}_i, \theta)f(\theta|\bar{y}_i, S_i)d\theta$$

where $E(\alpha|\bar{y}_i, \theta)$ is the Bayes rule of Proposition 3.2.1, for fixed $\theta$, and $f(\theta|\bar{y}_i, S_i)$ denotes the posterior density of $\theta$ for individual $i$ under the prior $H$. The Bayes rule is monotone in $\bar{y}_i$ in the limit as $S_i \to 0$ and $S_i \to \infty$, however for intermediate values of $S_i$ such monotonicity is no longer assured.

Monotonicity rests upon the contribution of $\frac{d}{d\bar{y}} f(\theta|\bar{y}_i, S_i)$, since for fixed $\theta$ the contribution from inner expectation is monotone by Proposition 3.2.1. In the $S_i$ limits the posterior $f(\theta|\bar{y}_i, S_i)$ puts all its mass on the most extreme points of the prior and consequently also produces a monotone Bayes rule for $\alpha$ as a function of $\bar{y}_i$. However, for moderate values of $S_i$ the situation is more complicated, and as we shall see in the empirical section, non-monotonicities can occur.

### 3.3.3 Covariate effects

Having seen how to estimate the Gaussian location-scale mixture model we will now briefly describe how to introduce covariate effects into the model, which now takes the form,

$$y_{it} = x_{it}\beta + \alpha_i + \sqrt{\theta_i}u_{it}.$$

Given a $\beta$ it is easy to see that,

$$\bar{y}_i|\alpha_i, \beta, \theta_i \sim N(\alpha_i + \bar{x}_i\beta, \theta_i/m_i)$$

so the sufficient statistic for $\alpha_i$ is $\bar{y}_i - \bar{x}_i\beta$. Similarly, the sufficient statistic for $\theta_i$ can be defined as,

$$S_i = \frac{1}{m_i - 1} \sum_{t=1}^{m_i} (y_{it} - x_{it}\beta - (\bar{y}_i - \bar{x}_i\beta))^2$$

and $S_i|\beta, \theta_i \sim \gamma(r_i, \theta_i/r_i)$, where as before, $r_i = (m_i - 1)/2$. Apparently, using the familiar panel data terminology, the sufficient statistic for $\alpha_i$ contains the between information, while the within information, deviations from the individual means, is contained in the $S_i$. A note of caution should be added however since the orthogonality of the within and between information enjoyed by the classical Gaussian panel data model no longer holds in this general mixture setting. This can be
seen more clearly by examining the likelihood function,

\[ L(\beta, h) = \prod_{i=1}^{n} g((\alpha, \beta, \theta)|y_{i1}, \ldots, y_{im_i}) \]

\[ = \prod_{i=1}^{n} \int \prod_{t=1}^{m_i} 0^{-1/2}\phi((y_{it} - x_{it}\beta - \alpha)/\sqrt{\theta})h(\alpha, \theta)d\alpha d\theta \]

\[ = K \prod_{i=1}^{n} S_i^{1-r_i} \int (\theta/m_i)^{-1/2}\phi((\bar{y}_i - \bar{x}_i\beta - \alpha)/\sqrt{\theta/m_i})e^{-R_i R_i^T S_i^{-1} g(r_i)}h(\alpha, \theta)d\alpha d\theta \]

where \( R_i = r_i s_i / \theta \) and \( K = \prod_{i=1}^{n} \left( \frac{r_i}{\sqrt{m_i r_i}} (1/\sqrt{2\pi})^{m_i - 1} \right) \).

Even with the independent prior assumption, \( h(\alpha, \theta) = h_\alpha(\alpha)h_\theta(\theta) \), the likelihood does not factor because the Gaussian piece depends on both \( \alpha \) and \( \theta \). However, the fact that \( S_i \), hence the Gamma piece of the likelihood, does not depend on \( \alpha \) provides a convenient estimation strategy by using the Gamma mixture to estimate \( h_\theta \), and a Studentized version of the Gaussian mixture, \( (\bar{y}_i - \bar{x}_i\beta - \alpha_i)/\sqrt{S_i} \sim t_{m_i-1} \), for estimating \( h_\alpha \). Including covariates adapts this estimation strategy: Given a \( \beta \) we can estimate the two mixing distributions and then evaluate the full profile likelihood. We will illustrate this approach in the empirical section, albeit with a more general mixture model that drops the independent prior assumption, and allows for covariates including lagged response. Our approach is related to recent work by Bonhomme and Manresa (2014) on grouped patterns of heterogeneity in panel data, in the sense that both approaches reduce the dimensionality of the heterogeneity distribution substantially, although the estimation methods employed are quite different. Convexity of our likelihood formulation ensures a unique solution and avoids the introduction of further tuning parameters, while the clustering algorithms employed by Bonhomme and Manresa require more delicate attention.

### 3.3.4 Empirical Bayes prediction: some simulation evidence

To develop some intuition about empirical Bayes methods we will consider some simple illustrative simulation examples in this section before turning to our main empirical application.

#### 3.3.4.1 Gaussian location mixtures

Suppose that we have a random sample from the model: \( y_i = \alpha_i + u_i \) with iid \( u_i \sim N(0, 1) \), and iid \( \alpha_i \sim \frac{2}{3}\delta_{-h} + \frac{1}{3}\delta_{2h} \) as in Chen (1995). Here, \( \delta_a \) denotes the distribution with point mass one at the point \( a \). If we were successful in estimating the distribution of \( \alpha_i \), we would expect that Tweedie’s formula (3.2.1) should deliver predictions that correctly shrink the original observations toward their respective \( \alpha_i \)’s. Of course the nature of the shrinkage depends crucially on the loss function as well as the prior. Thus, \( L_1 \) loss yields decisions that are closely related to classification, while \( L_2 \) loss delivers a Bayes rule whose shrinkage is somewhat more mild.

In Figure 3.1 we illustrate the foregoing situation with \( n = 400 \) and \( h = 0.5 \), which represents
a fairly challenging problem since the marginal density is still unimodal. In the left panel we have
the estimated mixing distribution in red solid curve, with the target distribution represented by
the blue dashed curve. The larger of the two actual mass points at $x = -0.5$ is quite accurately
estimated, however the smaller mass point at $x = 1$ is split into two pieces by the Kiefer-Wolfowitz
estimate. The true mixture distribution in blue dashed curve in the middle panel appears to be
reasonably accurately estimated by the red solid curve. The corresponding Bayes rules as derived
in Proposition 3.2.1 in the right panel show that the empirical Bayes rule (in red solid line) shrinks
a little too aggressively in the left tail, and not quite aggressively enough in the right tail, compared
to the omniscient Bayes rule in the blue dashed line. But we should hasten to add that it represents
an enormous improvement over the unbiased (naive) decision rule, $\hat{\alpha}_i = y_i$, depicted in the black
dotted line.

![Figure 3.1](image)

Figure 3.1: Empirical Bayes estimation for the Chen (1995) example: A sample of 400 observations
from the model with $y_i = \alpha_i + u_i$ with iid $u_i \sim N(0, 1)$, and iid $\alpha_i \sim \frac{2}{3}\delta_{-h} + \frac{1}{3}\delta_{2h}$, is illustrated by
the histogram in the middle panel and the “rug plots” in the adjacent panels. The Kiefer-Wolfowitz
estimate of the mixing distribution is illustrated in the left panel with the red solid curve, with the
target distribution in the blue dashed curve. The corresponding estimate of the mixture density and
Bayes rule appear in the other panels contrasted to their dashed target functions. The unbiased,
a naive decision rule is depicted in the right panel in the black dotted line.

Replacing the two mass point distribution by $\alpha_i \sim U[-h, 2h]$ yields an even more challenging
problem. The Kiefer-Wolfowitz estimator tries valiantly to mimic the uniform mixture by a discrete
mixture as illustrated in Figure 3.2. The two point mixing distribution appearing in the left panel
does not seem to be a very satisfactory surrogate for the uniform, but as can be seen in the middle
panel, it does a remarkably good job of imitating the correct mixture density. The Bayes rule
comparison in the right panel again illustrates that the shrinkage in the tails is not ideal, but much
preferable to the naive, unbiased rule.

82
Figure 3.2: Empirical Bayes estimation for the Chen (1995) example: A sample of 400 observations from the model with \( y_i = \alpha_i + u_i \) with iid \( u_i \sim N(0,1) \), iid \( \alpha_i \sim U[-h,2h] \), and \( h = 0.5 \), is illustrated by the histogram in the middle panel and the “rug plots” in the adjacent panels. The Kiefer-Wolfowitz estimate of the mixing distribution is illustrated in the left panel with the red solid curve, with the target distribution in the blue dashed curve. The corresponding estimate of the mixture density and Bayes rule appear in the other panels contrasted to their blue dashed target functions. The unbiased, naive decision rule is depicted in the right panel with the black dotted line.

### 3.3.4.2 Gamma scale mixtures

To explore the performance of empirical Bayes methods for gamma scale mixtures we illustrate a couple of similar cases to those appearing in the previous subsection. We first consider the longitudinal model,

\[
y_{it} = \alpha_i + \sqrt{\theta_i} u_{it}, \quad t = 1, \cdots, m, \quad i = 1, \cdots, n,
\]

with iid \( u_{it} \sim N(0,1) \), and \( \theta_i \sim F \). We take \( m = 11 \) and \( n = 400 \). We will provisionally ignore the heterogeneity in the \( \alpha_i \), or to be more explicit, adopt the naive practice of estimating them by \( \bar{y}_i \). Denoting the individual specific variance estimates by \( x_i = (m-1)^{-1} \sum_t (y_{it} - \bar{y}_i)^2 \), the \( \{x_i\} \) are then distributed as Gamma with shape parameter, \( r = (m-1)/2 \), scale parameter \( \theta_i/r \), and density,

\[
\gamma(x_i|\theta_i) = \frac{1}{\Gamma(r)(\theta_i/r)^r} x_i^{r-1} \exp(-x_i r/\theta_i).
\]

Thus, the marginal density of the sample variances is,

\[
g(x) = \int \gamma(x|\theta) dF(\theta).
\]
The Bayes rule under squared error loss for $\theta_i$ given $x_i$, originally derived by Robbins (1982), is given in the following proposition. Again, it should be stressed that the Bayes rule depends only on the mixture density, $g$, and not directly on the mixing distribution, $F$. Of course, indirectly the Bayes rule does depend on $F$ and in particular the flat portions of the Bayes rule in the third panel of Figure 3.3 representing the points of attraction of Bayes shrinkage are essentially determined by the location of the estimated mass points of $F$. This is particularly crucial in the tails where even small mass points of $\hat{F}$ can exert a large influence on the shrinkage. Whether this sensitivity can be lessened by replacing the Gaussian mixture assumption by something heavier tailed constitutes an intriguing question for future research. The next proposition describes the Bayes rule for estimating the $\theta_i$'s under $L_2$ loss for Gamma mixtures. Note that $\theta$ is not the natural parameter of the exponential family in this case, so the monotonicity of the Bayes rule requires a brief additional argument included in Section 3.6.

**Proposition 3.3.2.** For $X_i \sim \Gamma(r, \theta_i/r)$ and $\{\theta_i\}$ iid $F$, the Bayes rule under $L_2$ loss is:

$$\delta(x) = rx^{r-1}\int_{x}^{\infty} y^{1-r} g(y) dy / g(x)$$

(3.3.3)

and $\delta(x)$ is non-decreasing in $x$.

In Figure 3.3 we illustrate a typical outcome in a format like that of the previous figures. In this example we take $F$ to be the two point distribution: $\frac{2}{3}\delta_{1.5} + \frac{1}{3}\delta_3$. The two point mixing distribution is quite well estimated by the Kiefer-Wolfowitz procedure, and the mixture density appears to be quite accurate as well. The empirical Bayes rule slightly overestimates the variances in the upper tail since it slightly overestimates the location of the upper mass point. But as for the previous examples, there is an enormous improvement over the naive decision rule represented by the black dotted line. The brown dot-dashed line represents the linearized empirical Bayes rule proposed in Robbins (1982).

### 3.3.4.3 Gaussian location scale mixtures

We now would like to consider joint estimation of location and scale mixtures in the context of our longitudinal model. We will maintain the assumption that $\alpha_i$'s and $\theta_i$'s are drawn independently, so we only have to estimate two univariate mixing densities rather than a general bivariate density. We illustrate the procedure with an example that combines a three point distribution for $\alpha$ and a three point distribution for $\theta$: $y_{i1} = \alpha_i + \sqrt{\theta_i}u_{i1}$ with iid $u_{i1} \sim N(0, 1)$, iid $\alpha_i \sim \frac{1}{3}\delta_{-0.5} + \frac{1}{3}\delta_1 + \frac{1}{3}\delta_3$, and iid $\theta_i \sim \frac{1}{3}\delta_{0.5} + \frac{1}{3}\delta_4 + \frac{1}{3}\delta_4$, Maximizing the likelihood of Section 3.2, we obtain the estimates appearing in the first two panels of Figure 3.4 for the location and scale parameters respectively. As in the previous figures, estimates appear in red solid curves, and the true mixing distribution is represented by the blue dashed lines. Focusing on the location parameter, the third panel of the figure depicts the histogram of the observed $\bar{y}_i$, with the estimated marginal density, by integrating out $\alpha$ and $\theta$ with respect to $\hat{F}_\alpha$ and $\hat{F}_\theta$, superimposed with the red solid curve, and the true
Figure 3.3: Empirical Bayes estimation for Gamma mixture example: A sample of \( n = 400 \) and \( m = 11 \) observations from the model \( y_{i1} = \sqrt{\theta_{11}}u_{i1} \) with iid \( u_{i1} \sim N(0, 1) \) and iid \( \theta_{11} \sim \frac{2}{3}\delta_{1.5} + \frac{1}{3}\delta_{3} \) is illustrated by the histogram in the middle panel and the “rug plots” in the adjacent panels. The Kiefer-Wolfowitz estimate of the mixing distribution is illustrated in the left panel with the red solid line, with the target distribution in the blue dashed line. The corresponding estimate of the mixture density and Bayes rule appear in the other panels contrasted to their blue dashed target functions. The unbiased, naive decision rule is depicted in the right panel with a black dotted line. The brown dot-dashed line represents the linearized empirical Bayes rule proposed in Robbins (1982). 

How much can be gained by using an individual specific estimate of variance? The Bayes rules appearing in the last panel of Figure 3.4 are conditional only on the observed \( \bar{y}_{i} \) with the variance effect integrated out. Thus, when we see a value of \( \bar{y} \) near one of the mass points in \( \{-0.5, 1, 3\} \), the Bayes rule shrinks aggressively toward the corresponding \( \alpha \). Between these values, the predicted \( \alpha \), being a conditional mean, takes intermediate values. Extreme values of \( \bar{y} \) in either tail again get aggressively shrunk toward the extreme points of the estimated prior. The situation we have just described is artificial in the sense that we effectively are assuming that we have observed \( \bar{y}_{i} \) for each cross sectional unit, but apparently have forgotten to compute the associated variance estimate. If we now rectify this oversight, we can consider a two dimensional Bayes rule that maps \( (\bar{y}_{i}, S_{i}) \) pairs into predictions of the \( \alpha_{i} \)’s. Using the same data and the estimates underlying Figure 3.4
Figure 3.4: Empirical Bayes estimation for Gaussian location-scale mixture: A sample of \( n = 800 \) and \( m = 11 \) observations from the model \( y_{it} = \alpha_i + \sqrt{\theta_i}u_{it} \) with iid \( u_{it} \sim N(0,1) \), iid \( \alpha_i \sim \frac{1}{3}\delta_{-0.5} + \frac{1}{3}\delta_1 + \frac{1}{3}\delta_3 \), and iid \( \theta_i \sim \frac{1}{3}\delta_{0.5} + \frac{1}{3}\delta_2 + \frac{1}{3}\delta_4 \), is illustrated by the histogram in the middle panel. The Kiefer-Wolfowitz estimate of the mixing distributions is illustrated in the two left panels with red solid lines, with the target distribution in blue dashed lines. The corresponding estimate of the mixture density and the Bayes rule appear in the other panels contrasted to their blue dashed target functions.

we illustrate a contour plot of this two dimensional Bayes rule (Proposition 3.3.1) in Figure 3.5. We see that for central values of \( \bar{y}_i \) the contours are essentially vertical indicating the variance is uninformative about the mean, however for outlying values of \( \bar{y}_i \) the nonlinearity of the Bayes rule is apparent with large observed variances making us more uncertain about the \( \alpha_i \)'s.

When \( \bar{y} \) is in the extremes, the Bayes rule should shrink its estimate of \( \alpha \) to the extreme mass points at -0.5 and 3, but since the estimated prior has smaller mass points nearby very extreme observations are attracted to these values. In both tails one can see the effect of the variance estimate on this shrinkage effect; when the estimated variance is small then there is more shrinkage to the nearest mass point of the \( \alpha \) distribution. When the observed variance is large, then the posterior for \( \alpha \) is more evenly divided among several mass points and consequently the posterior mean is more central. For example, when \( \bar{y} = 1.5 \) and the estimated variance is low, then we can be quite confident that the observation comes from the \( \alpha = 1 \) population. Similarly when \( \bar{y} = -1.5 \) and the estimated variance is low, we can be confident that this is a \( \alpha = -0.5 \) observation. However, in either of these cases as the variance increases our confidence ebbs, and the Bayes rule assigns more probability to the other nearby mass points. For central values of \( \bar{y} \) the contours are nearly vertical indicating that the observed variance is not informative in this region. The observed pairs \((\bar{y}_i, S_i)\) are superimposed on the contours to give some sense of their dispersion.

This form of the Bayes rule clearly illustrates that variances are informative about the means in such circumstances, but the fact that we’ve imposed independence between \( \alpha \) and \( \theta \) may sacrifice valuable information in many applications. If we allow for dependence and estimate their joint distribution as in our empirical application, we will see that the sample variances provide crucial...
information for estimating $\alpha_i$.

### 3.4 Heterogeneous income dynamics

The vast literature on longitudinal models of income dynamics can be conveniently decomposed into two strands: one focusing on a permanent-transitory time-series structure that eschews individual specific sources of heterogeneity, exemplified by MaCurdy (1982), and going back at least to Friedman (1957), and another that relies on heterogeneity to account for observed persistence, as for example in Lillard and Weiss (1979), Baker (1997), Haider (2001), Guvenen (2009), Browning, Ejrnæs, and Alvarez (2010) and Hospido (2012). Considerable flexibility can be introduced into the former approach with the aid of age specific deterministic trends in mean and variances, as for example in Blundell, Graber, and Mogstad (2014), or stochastic specifications of the variance process, as in Meghir and Pistaferri (2004). While most of the foregoing work relies on first and second order moment information and therefore, at least implicitly adopts a Gaussian framework, there is evidence that such assumptions may distort important features of the earnings process. Mixture models of individual heterogeneity introduce further flexibility: Horowitz and Markatou (1996) and Bonhomme and Robin (2010) explore semiparametric deconvolution, while Geweke and Keane (2000) and Hirano (2002) propose Bayesian MCMC methods for estimating semiparametric mixture models. Our nonparametric empirical Bayes approach maintains the mixture model formulation, but expands the nature of the heterogeneity to encompass both location and scale effects.
In terms of estimation methods our approach is closest to that of Hirano since the KWMLE can be viewed as a limiting form of his Dirichlet process prior for the scale mixture setting. See Gu and Koenker (2013) for further details on this relationship, illustrated with an application to Gaussian location mixtures.

Our empirical analysis is based on the PSID sample used in Meghir and Pistaferri (2004), Browning, Ejrnæs, and Alvarez (2010) and Hospido (2012). The initial data consists of log real earnings of 2069 individuals between the ages of 25 and 55, with at least 9 consecutive records between 1968 and 1993. We further reduce the sample to 938 individuals who have continuous records from age 25 onwards.

We consider the model,

\[ y_{it} = \alpha_i + \beta_i x_{it} + \nu_{it}, \]
\[ \nu_{it} = \rho \nu_{i,t-1} + \sqrt{\theta_i} \epsilon_{it}, \quad \epsilon_{it} \sim N(0, \sigma^2) \]

Following standard practice in the literature, \( y_{it} \) denotes residuals from distinct annual regressions of log real earnings on a quadratic in age, and indicators for race, educational attainment, region and marital status. Heterogeneity around the mean earnings profile is captured by the random intercept and slope parameters; experience, \( x_{it} \), is defined as age minus \( \max\{\text{years of schooling}, 12\} - 6 \). Heterogeneity in the variance of earnings is captured by the \( \theta_i \)'s. More complex short-run dynamics could obviously be introduced, but our strategy is to proceed parsimoniously trying to understand at each stage the consequences of expanding the flexibility of the model.

### 3.4.1 Homogeneous trend and variance

Under the restrictions that \( \beta_i \equiv 0 \) and \( \theta_i \equiv 1 \) we can rewrite the model as,

\[ y_{it} = \rho y_{it-1} + (1 - \rho) \alpha_i + \epsilon_{it}. \]

This is a textbook dynamic panel model; in such models of earnings dynamics estimates of \( \rho \) are typically very close to one. These findings have led to considerable controversy over whether individual earnings processes “have a unit root.” In contrast to Meghir and Pistaferri (2004), who postulate a permanent component of earnings with a unit-root, Browning, Ejrnæs, and Alvarez (2010) – using the same data – find no unit root after introducing further heterogeneity in covariance structure of the model. In Figure 3.6, we present some preliminary evidence that helps to explain why the persistence of innovations may be reduced by introducing heterogeneity, for example, by relaxing the restrictions of a homogeneous variance.

The QQ plots of Figure 3.6 confirm earlier evidence of Horowitz and Markatou (1996) and Guvenen, Karahan, Ozkan, and Song (2014) based on more extensive CPS and Social Security data respectively that earnings innovations are considerably heavier tailed than our usual Gaussian assumptions would imply. There are a variety of possible treatments for this disease: one option would be to abandon the Gaussian assumption entirely, but this would lead us into realm of choosing
Figure 3.6: Normal QQ plots of partial differenced earnings for various $\rho$: For $\rho \in \{0.2, 0.3, \ldots, 1.0\}$ we plot empirical quantiles of the partial differences $y_{it} - \rho y_{i,t-1}$ standardized by their empirical standard deviation, against the corresponding Gaussian quantiles. The $\rho$’s are indicated in the thin strip at the top of each panel, the solid line in each plot is the 45 degree line indicating conformity to the Gaussian hypothesis. It is apparent from the plot that the observed quantiles are far too leptokurtic, that is much too peaked near the median and exhibiting much heavier tails than the Gaussian. For small $\rho$ there is also some left skewness in innovations that becomes less apparent for larger $\rho$. 

89
Figure 3.7: Scatterplot of individual specific mean and log variance effects for various $\rho$: For $\rho \in \{0.4, 0.6, 0.8, 1.0\}$ we plot sample means, $\bar{y}$, and log variances, $S$, of the partial differences $y_{it} - \rho y_{it-1}$. The $\rho$'s are indicated in the thin strip at the top of each panel. The more elliptical shape of the scatter for smaller $\rho$ may suggest that it could be more parsimoniously fit by our Gaussian/Gamma location-scale mixture model.

a non-Gaussian likelihood model that would, inevitably, be rather arbitrary. It is well known that heavy tailed distributions can be very flexibly modeled as scale mixtures of Gaussians, see for example the extensive discussion in Andrews, Bickel, Hampel, Huber, Rogers, and Tukey (1974), and we have already seen that it is relatively straightforward to estimate these mixture models; so this is the approach we will adopt.

To provide a further visual impression of the degree of individual heterogeneity we present in Figure 3.7 scatter plots of the individual specific sample means, $\bar{y}$, and log variances, $S$, for the partial differenced $y_{it}$ data for several $\rho$'s. In addition to confirming that there is substantial heterogeneity in these quantities the Figure also reveals that more moderate values of $\rho$ yield a more elliptical scatter that seems to be favored by our Gaussian/Gamma location-scale mixture likelihood as we will see in the next subsection.

### 3.4.2 Homogeneous trend with heterogeneous variances

If we fix $\rho$ and $\sigma^2_e$, and let $z_{it} = y_{it} - \rho y_{it-1}$, we can rewrite our model as,

$$z_{it} = (1 - \rho) \alpha_i + \sqrt{\theta_i} \epsilon_{it}.$$
As in Section 3.3, under Gaussian conditions, sufficient statistics for \( \alpha \) and \( \theta \) are respectively the sample mean and sample variance:

\[
\bar{y}_i = \frac{1}{T_i} \sum_{t=1}^{T_i} z_{it}, \\
S_i = \frac{1}{T_i} \sum_{t=1}^{T_i} (z_{it} - \bar{y}_i)^2. 
\]

Furthermore, we have, \( \bar{y}_i \mid \alpha_i, \theta_i \sim N((1 - \rho)\alpha_i, \theta_i \sigma^2_{\epsilon}/T_i) \) and \( (T_i - 1)S_i/\theta_i \mid \theta_i \sim \chi^2_{T_i-1} \). Assuming the pairs \( (\alpha_i, \theta_i) \) are iid with distribution function \( H \), we can discretize \( H \) on a two dimensional grid and write the likelihood of observing \( (z_{i1}, \ldots, z_{iT_i}) \) as a function of \( H, \rho \) and \( \sigma^2_{\epsilon} \), and apply the KWMLE.

Various special case of this model has been considered in the literature, for example the random effects model of Alvarez and Arellano (2003) assumes \( \theta_i \) to be degenerate taking value 1 while \( \alpha_i \sim N(\psi y_{i0}, \sigma^2_H) \). This leads to a marginal density for the \( \bar{y}_i \) conditional on \( y_{i0} \) as \( \bar{y}_i \sim N(\psi y_{i0}, \sigma^2_{\alpha}) \) with \( \sigma^2_{\alpha} = \sigma^2_{\epsilon}/T_i + \sigma^2_H \) as a free parameter. The parameters \( (\rho, \psi, \sigma^2_H, \sigma^2_{\alpha}) \) can then be estimated by maximizing the likelihood conditional on \( y_{i0} \). The Gaussian assumption on the \( \alpha_i \) is very convenient and very commonly employed, notably in Chamberlain (1980), Chamberlain and Hirano (1999) among many others. However, the normality assumption on the \( \alpha_i \) may be hard to justify. As we have seen in Figure 3.7, there is also considerable heterogeneity in the \( \theta_i \), and it seems plausible that there may be some dependence between \( \alpha \) and \( \theta \). These considerations motivate us to consider a non-parametric maximum likelihood framework allowing us to estimate the non-parametric mixing distribution \( H(\alpha, \theta) \) conditional on some structural parameters like \( \rho \), that can, in turn, be estimated by maximizing a profile likelihood.

Without loss of generality, we can set \( \sigma^2_{\epsilon} = 1 \), since it is not identified once we allow individual specific \( \theta_i \) unless we make further moment restrictions on \( \theta_i \). We have the following KWMLE problem:

\[
\hat{H}_\rho := \operatorname{argmax}_{H \in \mathcal{H}} \prod_{i=1}^{n} \int f(\bar{y}_i \mid \alpha, \theta)g(S_i \mid \theta) dH(\alpha, \theta)
\]

where \( \mathcal{H} \) is the space of all two dimensional distribution functions on the domain of \( \mathbb{R} \times \mathbb{R}_+ \). Here, \( f \) is the conditional normal density of \( \bar{y}_i \) and \( g \) is the conditional gamma density for \( S_i \). The KWMLE for \( H \) is indexed by \( \rho \) because both \( \bar{y}_i \) and \( S_i \) involve \( \rho \), which we have suppressed in the notation, but can be estimated by maximizing the profile log likelihood,

\[
l(\rho, \hat{H}_\rho) = \sum_{i=1}^{n} \mathcal{K}(\bar{y}_i, S_i) + \log \int f(\bar{y}_i \mid \alpha, \theta)g(S_i \mid \theta) d\hat{H}_\rho(\alpha, \theta).
\]

Allowing heterogeneous individual variances in earnings innovations is not new. Geweke and Keane (2000) contend that variance heterogeneity is crucial to account for non-Gaussian features of innovation distribution and use a three-component mixture formulation. Hirano (2002) adopts a more flexible Dirichlet prior specification for similar reasons. Browning, Ejrnæs, and Alvarez (2010) also find significant evidence that the variance of innovations varies across individuals. Their model
Figure 3.8: Profile likelihood for the $\rho$ parameter and heterogeneity distribution $H(\alpha, \theta)$: In the left panel we plot the Kiefer-Wolfowitz profile likelihood as a function of $\rho$. The shaded region represents a 0.95 confidence interval for $\rho$ based on the usual Wilks inversion procedure. In the right panel we plot the estimated joint heterogeneity distribution, evaluated at the optimal $\hat{\rho}$, $\hat{H}_\hat{\rho}(\alpha, \theta)$. Darker hexagons indicate greater mass, lighter ones less mass and white regions contain no mass.

posits eight latent factors all of which are constrained to obey parametric marginals. They comment “Nowhere in the literature is there any indication of how to specify a general joint distribution for these parameters, nor is there any hope of identifying the joint distribution non-parametrically.” In contrast, our approach allows only two latent factors, but has the advantage that it permits non-parametric estimation of their joint distributions.

What if $\hat{\rho} \approx 1$? Our joint distribution for $(\alpha_i, \theta_i)$ would then be meaningless, since the $\alpha_i$’s would be annihilated. The left panel of Figure 3.8 plots the profile likelihood for $\rho$, which (fortunately) peaks at 0.48. The shaded region indicates a 0.95 confidence interval for $\rho$ as determined by the classical Wilks inversion procedure, see e.g. Murphy and Van der Vaart (2000), and Fan, Zhang, and Zhang (2001). Our estimate of $\rho$ is close to the estimate of Hospido (2012) who also allows an individual specific variance component in a ARCH effect variance. She adopts a fixed effect specification for $(\alpha_i, \theta_i)$ and uses a bias corrected estimator for $\rho$ to account for the asymptotic bias introduced by estimating all the incidental parameters $(\alpha_i, \theta_i)$, $i = 1, \cdots, n$. A plausible explanation for why estimates of $\rho$ tend to be close to one in models without heterogeneity in variances is that individual specific persistence is mistaken for AR persistence in innovations.

The right panel of Figure 3.8 plots the two-dimensional non-parametric estimate of $\hat{H}_\hat{\rho}(\alpha, \theta)$ on a $60 \times 60$ grid. Mass points of the estimated distribution are indicated by shaded hexagons with darker shading indicating more mass. The support of $\hat{H}$ is determined by the support of the observed $(\hat{y}_1, S_1)$. The mixing distribution shows some negative dependence between $\alpha$ and $\theta$,
especially for $\alpha < 0$. So a low draw for $\alpha$ is more likely to be accompanied by a more risky (higher) $\theta$. Most of the mass of $\hat{H}$ is concentrated at very low levels of $\theta$, but it is not at all obvious how one might represent this estimated heterogeneity by a conventional parametric model.

### 3.4.3 Heterogeneous trends and variances

Reintroducing trend heterogeneity to our model of earnings dynamics gives us,

$$y_{it} - \rho y_{it-1} = (1 - \rho) \alpha_i + \beta_i \rho + (1 - \rho) \beta_i x_{it} + \sqrt{\theta_i} \epsilon_{it},$$

and obviously brings a new layer of complexity to the estimation problem. Our framework is capable of incorporating this third dimension of heterogeneity and we have made some tentative estimation efforts for the full model. However, this is challenging not only due to the jump from 2d to 3d grids, but because the trend term invalidates our sufficient statistic dimension reduction device. Some preliminary testing for trend heterogeneity using the LM test recently proposed in Juhl and Lugovskyy (2014) produced very weak evidence against homogeneity. We have also considered a variety of other, more elaborate, modeling strategies for the variance effect including ARCH effects, and deterministic trends in the variance. These can be estimated by adding new parameters to the profile likelihood problem, but again we saw no compelling evidence that they were needed. However, further study, particularly with larger datasets like that of Guvenen, Karahan, Ozkan, and Song (2014) may reveal something different within our framework.

### 3.4.4 Prediction

We now return to our original objective: we would like to adapt the well-known univariate empirical Bayes rules described earlier to compound decision problems for longitudinal data models. This objective is closely aligned with the objectives of Chamberlain and Hirano (1999), although our computational methods, and perhaps our philosophical outlook, are quite distinct. Given an initial trajectory for an individual’s earnings we would like to predict the remainder of the trajectory based not only on the prior history for the given individual, but also on the observed experience of a large sample of similar individuals. Chamberlain and Hirano motivate this prediction problem as one facing a typical financial advisor; similar problems present themselves in biomedical settings where diagnosis is based on reference growth charts.

Given a trajectory $Y_0 = \{y_t : t = 1, \ldots, T_0\}$ for a hypothetical individual we can easily determine a posterior, $p(\alpha, \theta | Y_0)$, based on our estimated mixture model. This KWMLE posterior is necessarily discrete, but one may feel entitled to draw uniformly from the grid rectangles of the estimated model for simulation purposes. In any case, the following simulation strategy can be employed to construct an ensemble of completed trajectories:

1. Draw $(\alpha, \theta)$ from $p(\alpha, \theta | Y_0)$,
2. Simulate \( Y_1 = \{ y_t : t = T_0 + 1, \cdots, T \} \) as,

\[
y_{T_0+s} = \alpha + \hat{\rho} y_{T_0+s-1} + \sqrt{\hat{\theta}} u_s, \quad s = 1, \cdots, T - T_0, \text{ and } u_s \sim N(0, 1),
\]

to obtain \( m \) paths, \( Y_1 \), then

3. Repeat steps 1 and 2 \( M \) times.

This procedure yields \( mM \) trajectories from which it is easy to construct pointwise and/or uniform prediction bands.

From a formal Bayesian perspective the foregoing procedure is rather heretical. We began with a perfectly legitimate likelihood formulation: data was assumed to be generated from a very conventional Gaussian model, but individuals had idiosyncratic \((\alpha, \theta)\) parameters whose distribution, \( H \), could be viewed as a prior. If this \( H \) were delivered on a silver platter by some local oracle we could proceed just as we have described. Bayes rule would allow us to update \( H \) in the light of the observed initial trajectory, \( Y_0 \) for each individual, and we would use these updated, individual specific, \( \hat{H}_i \)’s to construct an ensemble of forecast paths. Various functionals of these forecast paths could then be presented. Lacking a local oracle, we have relied instead on the KWMLE and the largess of the PSID to produce an \( \hat{H} \). Not only \( H \), but also \( \rho \) and potentially other model parameters are estimated by maximum likelihood. Remarkably, no further regularization is required, and profile likelihood delivers an asymptotically efficient estimator of these “homogeneous” parameters. Admittedly, we have “sinned” – we’ve peeked when we shouldn’t have peeked, but our peering has revealed a much more plausible \( H \) than we could have otherwise expected to produce by pure introspection. This is the charm of the empirical Bayes approach.

Our prediction exercise takes \( T_0 = 9 \) so the first nine years of observed earnings have been used as \( Y_0 \) to construct individual specific \( \hat{H}_i \) that are then used to construct pointwise confidence bands for earnings in subsequent years. We have selected a few pairs of individuals to illustrate the variety of earnings predictions generated by our model. In Figure 3.9 we contrast predictions for an individual with relatively large mean, high \( \alpha \), and large variance, high \( \theta \), with an individual with large variance, but lower mean. The “fan plot” depicts pointwise quantile prediction bands from 0.05 to 0.95 based on the simulated trajectories described above. For the high mean individual, the bands are relatively narrow reflecting the fact that the “prior” assigns little mass to high \( \theta \) individuals. In contrast, for the lower mean individual the bands are much wider, indeed the upper portion of the band overlaps with the lower portion of the band for the higher \( \alpha \) individual. Nevertheless, we see that the lower 0.05 quantile of the prediction band is exceeded. Our uniform band (not shown) for this individual just barely covers this excursion.

In Figure 3.10 we contrast high mean, low variance individual with low mean, high variance one. The prediction band is very narrow for the former individual, and much wider for the latter. Other features are also apparent from these figures: individuals who begin the forecast period below their pre-forecast mean, like PSID 59, are predicted to come back to their mean, and some asymmetry is visible, for example in PSID 44, whose lower tail is somewhat wider than the upper one. Note
Figure 3.9: Fan plot of earnings forecasts for two individuals: Based on the initial 9 years earnings, pointwise prediction bands are shown with graduated shading indicating bands from the 0.05 to 0.95 quantiles.

Figure 3.10: Fan plot of earnings forecasts for two individuals: Based on the initial 9 years earnings, pointwise prediction bands are shown with graduated shading indicating bands from the 0.05 to 0.95 quantiles.
that asymmetry requires some asymmetry in the location component of the mixture distribution \( \hat{H} \), since pure scale mixtures of Gaussians are necessarily symmetric.

### 3.4.5 Estimation of random effects

To conclude our discussion of earning dynamics we will briefly consider the problem of estimating random effects. Such problems have a long history; in econometrics they can be traced back to the seminal work of Goldberger (1962) on best linear unbiased prediction (BLUP). For a comprehensive survey of the early literature, see Robinson (1991). It may seem odd to consider estimation of random effects, but in many applications including our earning dynamics setting it is natural to ask: How would we estimate \( \alpha_i \)'s? The BLUP approach has a long history in animal breeding where \( \alpha_i \)'s are interpreted as a latent productivity variable. Our approach is considerably more flexible than earlier methods that assumed conjugate parametric priors for the mixing distributions.

In this section we illustrate the Bayes rule for estimating \( \alpha_i \)'s given the observed pair \((\bar{y}_i, S_i)\) for a given individual, and interpret the resulting shrinkage rules. Because of the general bivariate structure of estimated prior these shrinkage strategies can be considerably more complicated than those illustrated in the independent prior setting of the previous section. Figure 3.11 plots contours of the Bayes rule, \( \hat{\alpha}_i = E(\alpha | \bar{y}_i, S_i) \) in Proposition 3.3.1. This figure is analogous to Figure 3.5 except that the nature of the shrinkage for moderate \( S_i \) is more severe. If we first focus on the right side of the plot for positive \( \bar{y}_i \)'s we see that observations with moderate variances are shrunken quite substantially toward zero. So, for example, if we saw an observation with \( \bar{y} = 0.5 \) and \( S = 0.25 \) the Bayes rule estimates \( \alpha = 0 \). Why? The first thing to say is that we never saw points like this, the observed \((\bar{y}, S)\) pairs are depicted as the grey dots, so an \( S \) as big as 0.25 is much more likely to come from a low \( \alpha \) individual and this accentuates the shrinkage. We should stress that the empirical distribution of the points appearing in the plot although they are a key ingredient in the construction of the estimated prior \( \hat{H} \) illustrated in Figure 3.7, is only a starting point for building the Bayes rule underlying the contour plot. The Bayes rule requires updating individuals posterior for \((\alpha, \theta)\) in the light of \( \hat{H} \) and the observed \((\bar{y}, S)\) and then computing expectations as in Proposition 3.3.1. On the left side of the plot, for \( \bar{y} < 0 \) the situation is somewhat similar, but the shrinkage is less severe.

In Figure 3.12 we illustrate the Bayes rule for \( \alpha \) as a function of \( \bar{y} \) for several fixed values of \( S \), essential plotting our contour values for horizontal cross-sections. The naive estimator, \( \hat{\alpha} = \bar{y} \) is shown as the 45 degree line. For both low and high values of \( S \) we have monotone Bayes rules, so larger \( \bar{y} \) implies larger \( \hat{\alpha} \), however for the intermediate \( S = 0.272 \) value we see that the Bayes rule is clearly non-monotone. Similar calculations could be employed to estimate the variability parameter, \( \theta \), as a function of “observed” \((\bar{y}, S)\). (Recall that \((\bar{y}, S)\) implicitly depends upon an estimated \( \rho \) parameter.) Of course, there is nothing sacred about \( L_2 \) loss, and it is entirely reasonable to consider other loss functions that would lead to alternative Bayes rules: posterior medians, posterior modes, etc.
Figure 3.11: Contour plot of the Bayes rule $\mathbb{E}(\alpha|\bar{y}, S)$: The plot illustrates pairs $(\bar{y}_i, S_i)$ that produce the same posterior mean of $\alpha$.

## 3.5 Conclusion

Models of unobserved heterogeneity for longitudinal data are common in applied econometrics. We have argued that empirical Bayes methods based on nonparametric maximum likelihood estimation of mixture models offer a natural formulation of these models. Recent developments in convex optimization greatly facilitate estimation of such models. Semiparametric versions of these models including covariate effects are shown to be effectively analyzed with profile likelihood. A potential criticism of the foregoing approach is that it requires us to assume a parametric form for the base distribution, in our setting the Gaussian. Of course, location-scale mixtures of Gaussians is quite a general class, so from a prediction perspective the normality assumption seems not to be terribly onerous.

Empirical Bayes applications have generally either assumed a parametric form for parameter
Figure 3.12: Bayes Rule $\hat{\alpha} = \mathbb{E}(\alpha|\bar{y}, S)$ for several (fixed) $S$: The plot depicts the posterior mean of $\alpha$ as a function of $\bar{y}$ for several values of $S$.

heterogeneity as in the hierarchical Bayes literature or considered univariate parametric heterogeneity as in the more recent compound decision literature. We are not aware of any prior non-parametric bivariate heterogeneity specifications. Many econometric applications, however, involve mean-variance trade-offs that naturally suggest more flexible bivariate specifications. As we have seen, modern optimization methods linked to the Kiefer-Wolfowitz MLE accommodate such models quite easily. Because the formulation is cast directly in terms of likelihood there are convenient methods of handling estimation and inference for other (global) parametric components via profiling. We would also like to stress that there is nothing crucial about the Gaussian framework that we have employed; other specifications of the base measure for the mixture can be easily accommodated. In addition to the normal-gamma mixtures explored here, we have also considered Weibull, Gompertz, Pareto, Binomial and Poisson mixtures in other work.

There are many possible extensions left to explore. More flexible treatment of the covariates in the initial stage of our procedure would be desirable; in larger datasets this could be easily handled with further stratification of the sample. More flexible treatment of the variance effects would also be desirable, either with deterministic age effects or some form of stochastic ARCH-type effects. Trend heterogeneity is also feasible, but perhaps only with larger scale data sources. We have tried to encourage further exploration of these methods by providing the R package REBayes that implements the methods we have described here as well as a variety of other model specifications.
3.6 Proofs

**Proof of Proposition 3.2.1.** The simplest way to derive Tweedie’s Formula in Proposition 3.2.1 for the Gaussian case seems to be to consider the more general exponential family compound decision problem in which

\[ g(y) = \int \varphi(y, \eta) dF(\eta), \]

where \( \varphi \) is a known exponential family density with natural parameter \( \eta \), so we may write,

\[ \varphi(y, \eta) = m(y)e^{\eta h(\eta)}, \]

and \( F \) is again a mixing distribution over the parameter \( \eta \). Quadratic loss implies that the Bayes rule is the conditional mean:

\[
\delta(y) = \frac{d}{dy} \log(\int e^{\eta h(\eta)} dF / \int e^{\eta h(\eta)} dF) = \frac{d}{dy} \log(g(y)/m(y))
\]

Differentiating again,

\[
\delta'(y) = \frac{d}{dy} \left[ \frac{\int \eta \varphi dF}{\int \varphi dF} \right] = \frac{\int \eta^2 \varphi dF}{\int \varphi dF} - \left( \frac{\int \eta \varphi dF}{\int \varphi dF} \right)^2 = \mathbb{E}[\eta^2|Y = y] - (\mathbb{E}[\eta|Y = y])^2 = \mathbb{V}[\eta|Y = y] \geq 0,
\]

implying that \( \delta \) must be monotone. When \( \varphi \) is Gaussian with known variance \( \theta \) we have natural parameter \( \eta = \alpha/\theta \),

\[
\varphi(y, \alpha/\theta) = \varphi((y - \alpha)/\sqrt{\theta})/\sqrt{\theta} = K \exp(-(y - \alpha)^2/2\theta) = Ke^{-y^2/2\theta} \cdot e^{\alpha^2/2\theta} \cdot e^{-\alpha^2/2\theta},
\]

so \( m(y) = e^{-y^2/2\theta} \) and the logarithmic derivative yields our Bayes rule in Proposition 3.2.1. \( \square \)

**Proof of Proposition 3.3.1.** Suppose we have \( y_{i1} | \alpha_i, \theta_i \sim N(\alpha_i, \theta_i) \). Let \( \bar{y}_i \) and \( S_i \) be defined respectively as the sample mean and sample variance with conditional density \( \phi(\bar{y}_i | \alpha_i, \theta_i) \) and \( \gamma(S_i | \theta_i) \). Denote the marginal density for the vector \( (y_{i1}, \ldots, y_{im}) \) as \( g(\bar{y}_i, S_i) \). Under squared
error loss, we wish to minimize the expected loss,
\[
\min_\alpha \mathbb{E}_{\alpha, \theta}[\|\hat{\alpha} - \alpha\|^2].
\]

This leads to the Bayes rule:
\[
\hat{\alpha}_t = \mathbb{E}[\alpha \mid \bar{y}_i, S_i] = \int \alpha \mathbb{E}[\alpha \mid \bar{y}_i, S_i, \theta] d\theta d\alpha
\]
\[
= \int \mathbb{E}[\alpha \mid \bar{y}_i, S_i, \theta] h(\alpha \mid \theta) d\alpha
\]
\[
= \int \mathbb{E}[\alpha \mid \bar{y}_i, S_i, \theta] h(\alpha \mid \theta) d\theta / g(\bar{y}_i, S_i)
\]
\[
= \int \mathbb{E}[\alpha \mid \bar{y}_i, S_i, \theta] h(\alpha \mid \theta) d\theta / g(\bar{y}_i, S_i)
\]

To check for monotonicity with respect to \(\bar{y}\), we differentiate \(\mathbb{E}[\alpha \mid \bar{y}, S]\) for some fix \(S\), which leads to
\[
\frac{d}{d\bar{y}} \mathbb{E}(\alpha \mid \bar{y}, S) = \int \frac{d}{d\bar{y}} \mathbb{E}(\alpha \mid \bar{y}, S) f(\theta \mid \bar{y}, S) d\theta + \int \mathbb{E}(\alpha \mid \bar{y}, \theta) \frac{d}{d\bar{y}} f(\theta \mid \bar{y}, S) d\theta
\]

The first piece is non-negative due to Proposition 3.2.1, while the sign of the second piece is undetermined.

\[\square\]

**Proof of Proposition 3.3.2.**

For the gamma mixture case there is an analogous formula as the Gaussian case in Proposition 3.2.1, for the natural parameter \(-r/\theta\), proceeding as before,
\[
\tilde{\delta}(x) = \mathbb{E}[-r/\theta \mid X = x] = \int \frac{-x f(x\theta) dG(\theta)}{f(x\theta) dG(\theta)} = \frac{d}{dx} \log \left( \frac{g(x)}{x^{r-1}} \right) = \frac{g'(x)}{g(x)} - \frac{r - 1}{x}
\]

If, on the other hand, we would, quite naturally, like to compute the expectation of the unnatural parameter \(\theta\), then we obtain instead,
\[
\delta(x) = \mathbb{E}[\Theta \mid X = x]
\]
\[
= \int \gamma(\theta) \mathbb{E}[\alpha \mid \theta] dF(\theta)/\gamma(\theta) dF(\theta)
\]
\[
= \int \frac{\theta}{\Gamma(r) \theta^r} x^{r-1} \exp(-x r/\theta) dF(\theta) / g(x)
\]
\[
= x^{r-1} \int \frac{(r/\theta)^r \theta}{\Gamma(r)} \exp(-x r/\theta) dF(\theta) / g(x)
\]
\[
= x^{r-1} \int \frac{(r/\theta)^r \theta}{\Gamma(r)} \int_x^\infty \exp(-x r/\theta) dF(\theta) / g(x)
\]
\[
= x^{r-1} \int_x^\infty y^{r-1} \gamma(y) dF(\theta) / g(x)
\]
\[
= x^{r-1} \int_x^\infty y^{r-1} g(y) dF(\theta) / g(x).
\]
It remains to show that this formulation of the Bayes rule is monotone:

\[
\delta'(x) = \frac{r(r-1)x^{r-2} \int_x^\infty y^{1-r}g(y)dy + rx^{r-1}(-x^{1-r})g(x)}{g(x)} \\
- \frac{rx^{r-1} \int_x^\infty y^{1-r}g(y)dy g'(x)}{g^2(x)} \\
= \frac{rx^{r-1} \int_x^\infty y^{1-r}g(y)dy}{g(x)} \left[ \frac{r-1}{x} - \frac{g'(x)}{g(x)} \right] - r \\
= -\delta(x)\delta(x) - r \\
= \delta(x) \left[ E[\frac{r}{\theta}|X = x] - E[\frac{r}{\theta}|X = x] \right] \\
= \delta(x) \left[ E[\frac{1}{\theta}|X = x] - \frac{1}{E[\theta|X = x]} \right] ,
\]

which is positive by Jensen’s inequality. \(\square\)
References

ABELSON, R., AND J. TUKEY (1963): “Efficient Utilization of Non-numerical Information in Qua-
titative Analysis: General Theory and the Case of Simple Order,” Annals of Mathematical Statis-
tics, 34, 1347–1369.

AKHARIF, A., AND M. HALLIN (2003): “Efficient Detection of Random Coefficients in Autoregres-

ALVAREZ, J., AND M. ARELLANO (2003): “The time series and cross-section asymptotics of dy-
namic panel data estimators,” Econometrica, 71, 1121–1159.


mixture models with or without structural parameter,” ESAIM. Probability and Statistics, 13,
301–327.


BENNALA, N., M. HALLIN, AND D. PAINDAVEINE (2012): “Pseudo-Gaussian and Rank-based Op-
timal Tests for Random Individual Effects in Large n Small T Panels,” Journal of Econometrics,
170, 50–67.

in a prototypical nonregular problem,” in Statistics and Probability: A Raghu Raj Bahadur
New Delhi.

BICKEL, P., C. KLAASSEN, Y. RITOV, AND J. WELLNER (1993): Efficient and Adaptive Estima-

BLUNDELL, R., M. GRABER, AND M. MOGSTAD (2014): “Labor Income Dynamics and the Insur-


