Generalized Likelihood Ratio Statistics and Wilks Phenomenon

Introduction

There are several general applicable principles for parametric estimation and inferences.

Parametric Estimation:

- Maximum likelihood
- Least Squares
- Generalized method of moments

Parametric Inference (hypothesis test and confidence regions) :

- Likelihood ratio tests
- Jackknife
- bootstrap

Likelihood principle as a general principle of inference for parametric models.

Introduction

For nonparametric models, there are also generally applicable methods for nonparametric estimation and modeling.

- Local polynomial
- Spline
- Orthogonal series methods

On the other hand, while there are many customized methods for constructing confidence intervals and conducting hypothesis testing (Hart, 1997), there are few generally applicable principles for nonparametric inferences.

Brief review of likelihood ratio principle

Suppose that the data generating process is governed by the underlying density $f(\boldsymbol{x};\theta)$, with unknown θ in parametric space Θ. The statistical interest lies in testing:

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta - \Theta_0$$

based on a random sample $\{\boldsymbol{x}_i\}_{i=1}^n$, where Θ_0 is a subspace of Θ.

For a well formulated question, one can use the well-known maximum likelihood ratio statistic

$$\lambda_n = 2\{\max_{\theta\in\Theta} \ell(\theta) - \max_{\theta\in\Theta_0} \ell(\theta)\},$$

where $\ell(\theta)$ is the log-likelihood function for getting the given sample under the model $f(\boldsymbol{x};\theta)$.

Wilks Phenomenon

- An important fundamental property of the likelihood ratio tests is that their asymptotic null distributions are independent of nuisance parameters in the null hypothesis.
- With this property, one can simulate the null distribution by fixing the nuisance parameters at a reasonable value or estimate.
- This property is referred as *Wilks Phenomenon* and is fundamental to all hypothesis testing problems.
- However, in many examples,even though the null hypothesis is well formulated, the alternative hypotheses are not.

A simple example

To see this more clearly, consider the question if variable X (e.g. age) and Y (salary) are related. If we embed the problem in the linear model

$$Y = \alpha + \beta X + \varepsilon,$$

then the problem becomes testing $H_0 : \beta = 0$. If unknown to us, the data generating process is govern by

$$Y = 60 - 0.1(X - 45)^2 + \varepsilon, \quad \varepsilon \sim N\left(0, 5^2\right),$$

with X uniformly distributed on the interval $[25, 65]$, the null hypothesis will be accepted very often since slope β is not statistically significant.

Nonparametric alternatives

- The above discussion reveals that the family of alternative models should be large enough in order to make sensible inferences.
- In many hypothesis testing problems, while the null hypothesis is well formulated, the alternative one is vague.
- These two considerations make nonparametric models as attractive alternative hypothesis.
- without knowing the data generating process, a nature alternative model for the above example is

  $$Y = m(X) + \varepsilon,$$

  where $m(\cdot)$ is smooth, while the null hypothesis is $Y = \mu + \varepsilon$.
- With such a flexible alternative family of models, the aforementioned pitfall is avoided.

Nonparametric null against nonparametric alternative

The problems of testing nonparametric null against nonparametric alternative hypotheses arise also frequently in statistical inferences. For example, the additive model

$$Y = \alpha + \sum_{d=1}^{D} m_d(X_d) + \varepsilon.$$

The question such as if the covariates X_1 and X_2 are related to the response Y arise naturally, which amounts to testing

$$H_0 : m_1(\cdot) = m_2(\cdot) = 0.$$

This is a nonparametric null versus nonparametric alternative hypothesis testing problem.

Nonparametric null against nonparametric alternative

- There are many techniques designed to solve this kind of problems. Many of them focused on an intuitive approach using discrepancy measures (such as the L_2 and L_∞ distance).
- They are generalizations of the Kolmogorov-Smirnov and Cramer-von Mises types of statistics.
- The test statistic based on discrepancy method is $T = \sum_{d=1}^{2} c_d ||\hat{m}_d||$.
- One has to choose not only the norm $|| \cdot ||$, but the weights c_d.
- Second, the null distribution of the test statistic T is unknown and depends critically on the nuisance functions $m_3, \dots, m_d$.

Naive extension of maximum likelihood ratio tests

- Although likelihood ratio theory contributes tremendous success to parametric inference, there are few general applicable approaches for nonparametric inferences based on function estimation.
- A naive extension is the nonparametric maximum likelihood ratio test. However, nonparametric maximum likelihood estimation usually does not exist. Even if it exists, it is hard to compute. Furthermore, the resulting maximum likelihood ratio tests are not optimal.

A simple example

Consider the following white noise model:

$$Y_i = \theta_i + n^{-1/2}\varepsilon_i, \quad \varepsilon_i \overset{i.i.d.}{\sim} N(0,1), i = 1, 2, ...$$

Now, let us consider testing the simple hypothesis:

$$H_0 : \theta_1 = \theta_2 = \cdots = 0.$$

Consider the parametric space $\mathcal{F}_k^* = \{\theta : \sum_{j=1}^{\infty} j^{2k}\theta_j^2 \leq 1\}$ where $k \geq 0$. Then under the parametric space $\mathcal{F}_k^*$, the MLE is

$$\hat{\theta}_j = (1 + \hat{\xi} j^{2k})^{-1} Y_j,$$

where $\hat{\xi}$ is the Lagrange multiplier satisfying

$$\sum_{j=1}^{\infty} j^{2k}\hat{\theta}_j^2 = 1$$

A simple Proof

The log likelihood is

$$logL \propto \sum_{i=1}^{n}(Y_i - \theta_i)^2.$$

Under the constraint $\sum_{j=1}^{\infty} j^{2k}\theta_j^2 \leq 1$, we can solve the Lagrange problem

$$\max_{\theta_i,\xi} \sum_{i=1}^{n}(Y_i - \theta_i)^2 + \xi(\sum_{j=1}^{\infty} j^{2k}\theta_j^2 - 1).$$

Lemma 2.1

Under the null hypothesis,

$$\hat{\xi} = n^{-2k/(2k+1)} \{ \int_0^\infty \frac{y^{2k}}{(1+y^{2k})^2} dy \}^{2k/(2k+1)} \{1 + o_p(1)\}.$$

The maximum likelihood ratio statistic for the problem is given by

$$\lambda_n^* = \frac{n}{2} \sum_{j=1}^{n} \left(1 - \frac{j^{4k}\hat{\xi}^2}{(1+j^{2k}\hat{\xi})^2} \right) Y_j^2.$$

Theorem 1

Under the null hypothesis (2.3), the normalized maximum likelihood ratio test statistic has an asymptotic χ^2 distribution with degrees of freedom a_n written as $r_k \lambda_n^* \sim_a \chi^2_{a_n}$, where

$$r_k = \frac{4k+2}{2k-1}, \quad a_n = \frac{(2k+1)^2}{2k-1} \left[\frac{\pi}{4k^2 \sin(\pi/(2k))} \right]^{2k/(2k+1)} n^{1/(2k+1)}$$

Proofs of Lemma 2.1

Define

$$F(\xi) = \sum_{j=1}^{\infty} j^{2k}(1 + \xi j^{2k})^{-2} Y_j^2.$$

For each given $\xi_{n,c} = cn^{-2k/(2k+1)} (c > 0)$, under the null hypothesis, by using the mean-variance decomposition, we have

$$F(\xi_{n,c}) = n^{-1} \sum j^{2k}(1 + j^{2k}\xi_{n,c})^{-2} + O_p\left[n^{-1}\{ \sum j^{4k}(1 + j^{2k}\xi_{n,c})^{-4} \} \right].$$

Note that $g_n(x) = x^{2k}/(1 + x^{2k}\xi_{n,c})^2$ is increasing for $0 \leq x \leq \xi_{n,c}^{-1/(2k)}$ and decreasing for $x \geq \xi_{n,c}^{-1/(2k)}$. Then,

$$\begin{aligned} & n^{-1} \sum j^{2k}(1 + j^{2k}\xi_{n,c})^{-2} \\ & = n^{-1} \int_0^\infty \frac{x^{2k}}{(1 + x^{2k}\xi_{n,c})^2} dx + O(n^{-1} g(\xi_{n,c}^{-1/(2k)})) \\ & = c^{-(2k+1)/(2k)} \int_0^\infty \frac{y^{2k}}{(1 + y^{2k})^2} dy + O(n^{-1/(2k+1)}). \end{aligned}$$

Proof of Lemma 2.1

Using the similar argument, we have

$$n^{-1}\{\sum j^{4k}(1+j^{2k}\xi_{n,c})^{-4}\} = O\{n^{-1/(2(2k+1))}\}.$$

Then

$$F(\xi_{n,c}) = (c_0/c)^{(2k+1)/(2k)} + O_p(n^{-1/(2(2k+1))}),$$

where $c_0 = (\int_0^\infty y^{2k}(1+y^{2k})^{-2}dy)^{2k/(2k+1)}$.

For any $\varepsilon > 0$, since the function $F(x)$ is strictly decreasing,

$$P(|n^{2k/(2k+1)}(\hat{\xi}-\xi_{n,c_0})| > \varepsilon) = P(F(\hat{\xi}) < F(\xi_{n,c_0+\varepsilon}))+P(F(\hat{\xi}) > F(\xi_{n,c_0-\varepsilon}))$$

which implies $\hat{\xi} - \xi_{n,c_0} = o_p(n^{-2k/(2k+1)})$.

Note $F(\hat{\xi}) = 1$.

Proof of Theorem 1

Define the j -th coefficient in $F(\xi)$ and λ_n^* as

$$F(j;\xi) = \frac{j^{2k}}{(1+j^{2k}\xi)^2}, \quad \lambda(j;\xi) = \frac{1+2j^{2k}\xi}{(1+j^{2k}\xi)^2}.$$

Then

$$F^{'}(j;\xi) = -\frac{2j^{4k}}{(1+j^{2k}\xi)^3}, \lambda^{'}(j;\xi) = -\xi F^{'}(j;\xi).$$

For any $\eta_{n,j}$ and $\zeta_{n,j}$between $\hat{\xi}$ and ξ_{n,c_0}, it can easily shown that

$$\sup_{j\geq 1} |\frac{F^{'}(j;\eta_{n,j}) - F^{'}(j;\xi_{n,c_0})}{F^{'}(j;\xi_{n,c_0})}| = o_p(1), \quad \sup_{j\geq 1} |\frac{\lambda^{'}(j;\zeta_{n,j}) - \lambda^{'}(j;\xi_{n,c_0})}{\lambda^{'}(j;\xi_{n,c_0})}| = o_p(1)$$

Proof of Theorem 1

Let $\lambda_n = \frac{1}{2}\sum_{j=1}^{\infty}(1+2j^{2k}\xi)/(1+j^{2k}\xi)^2\varepsilon_j^2$. By using Taylor expansion,

$$\begin{aligned}
\lambda_n^* &= \frac{1}{2}\sum_{j=1}^{\infty}[\lambda(j,\eta_{n,c_0}) + (\hat{\xi} - \xi_{n,c_0)})\lambda^{'}(j;\zeta_{n,j})]\varepsilon_j^2 \\
&= \lambda_n(\xi_{n,c_0}) + [F(\hat{\xi}) - F(\xi_{n,c_0})]\frac{\frac{1}{2}\sum_{j=1}^{\infty}\lambda^{'}(j;\xi_{n,c_0})\varepsilon_j^2}{1/n\sum_{j=1}^{\infty}F^{'}(j;\xi_{n,c_0})\varepsilon_j^2}(1+o_p(1)) \\
&= \lambda_n(\xi_{n,c_0}) + [1 - F(\xi_{n,c_0})]\frac{n}{2}\xi_{n,c_0} + o_p(n^{1/(2(2k+1))}) \\
&= \frac{1}{2}\sum_{j=1}^{\infty}\frac{1}{(1+j^{2k}\xi_{n,c_0})}\varepsilon_j + \frac{1}{2}c_0 n^{1/(2k+1)} + o_p(n^{1/(2(2k+1))}).
\end{aligned}$$

Then prove the asymptotical normality of $\frac{1}{2}\sum_{j=1}^{\infty}\frac{1}{(1+j^{2k}\xi_{n,c_0})}\varepsilon_j$.

Theorem 2

There exists a $\theta \in \mathcal{F}_k$ satisfying $||\theta|| = n^{-(k+d)/(2k+1)}$ with $d > 1/8$ such that the power function of the maximum likelihood ratio test at the point θ is bounded by α, namely,

$$\limsup P_\theta\{r_k\lambda_n^* > a_n + z_\alpha(2a_n)^{1/2}\} \leq \alpha,$$

where z_α is the upper a quantile of the standard normal distribution.

- Thus, the maximum likelihood ratio test A* can detect alternatives with a rate no faster than $n^{-(k+d)/(2k+1)}$.
- When $k > 1/4$, by taking d sufficiently close to $1/8$, the rate $n^{-(k+d)/(2k+1)}$ is slower than the optimal rate $n^{-2k/(4k+1)}$ given in Ingster(1993).

Proof of Theorem 2

Take $j_n^{-k} = n^{-(k+d)(2k+1)}$. Let θ be a vector whose j_n th position is j_n^{-k} and the rest are zero.

Then $\theta \in \mathcal{F}_k$ and $||\theta|| = n^{-(k+d)/(2k+1)}$. For $\xi_{n,c} = cn^{-2k/(2k+1)}$,we have

$$j_n^{2k}\xi_{n,c} = cn^{2d/(2k+1)}.$$

Under this specific alternative, we have

$$F(\xi_{n,c}) = n^{-1}\sum_{j=1}^{\infty}\frac{j^{2k}}{(1+j_n^{2k}\xi_{n,c})^2}\varepsilon_j^2 + o_p(n^{-1/(2(2k+1))}).$$

By arguments as in the proof of Lemma 2.1, one can see that

$$\hat{\xi} = \xi_{n,c_0}(1+o_p(1)).$$

Proof of Theorem 2

Let

$$\lambda_{n,0} = \frac{1}{2}\sum_j (1 - \frac{j^{4k}\hat{\xi}^2}{(1+j^{2k}\hat{\xi})^2})\varepsilon_j^2.$$

Then

$$\lambda_n^* = \lambda_{n,0} + o_p(n^{1/(2(2k+1))}).$$

By a similar proof to Theorem1, $r_k\lambda_{n,0} \sim \chi^2_{a_n}$, which entails that

$$P_\theta(r_k\lambda_n^* > a_n + z_\alpha(2a_n)^{1/2}) = \alpha + o(1).$$

Generalized likelihood ratio tests

Take the generalized likelihood ratio test as

$$\lambda_n = \frac{n}{2} \sum_{j=1}^{n} \left(1 - \frac{j^{4k} \xi_n^2}{(1 + j^{2k} \xi_n)^2} \right) Y_j^2.$$

with $\xi_n = cn^{-4k/(4k+1)}$ for some $c > 0$.

This ameliorated procedure achieves optimal rate of convergence for hypothesis testing.

Theorem3

Under the null hypothesis, $r_k^{'}\lambda_n \sim_a \chi^2_{a_n^{'}}$, where

$$r_k^{'} = \frac{2k+1}{2k-1}\frac{48k^2}{24k^2+14k+1},$$

$$a_n^{'} = \frac{(2k+1)^2}{2k-1}\frac{24k^2c^{-1/(2k)}}{24k^2+14k+1}\left[\frac{\pi}{4k^2\sin(\pi/(2k))}\right]n^{2/(4k+1)}.$$

Furthermore, for any sequence $c_n \to \infty$, the power function of generalized likelihood ratio test is asymptotically one,

$$\inf_{\theta\in\mathcal{F}_k:||\theta||\geq c_n n^{-2k/(4k+1)}} P_\theta\{\frac{r_k^{'}\lambda_n - a_n^{'}}{\sqrt{2a_n^{'}}} > z_\alpha\} \to 1.$$

Another GLR test

- The GLR test allows one to use any reasonable nonparametric estimator to construct the test.
- For the Sobolev class $\mathcal{F}_k^*$, another popular class of nonparametric estimator is the truncation estimator

  $$\hat{\theta}_j = Y_j, \quad \text{for} \quad j = 1, \cdots, m, \quad \text{for} \quad j > m,$$

  for a given m.
- Then the GLR is the Neyman(1937) test

  $$T_N = \sum_{i=1}^{m} nY_i^2$$

  With choice of $m = cn^{2/4k+1}$, the Neyman test, can achieve the optimal rate $n^{-2k/(4k+1)}$.

What is Wilks' Phenomenon

- By Wilks phenomenon, we mean that the asymptotic null distributions of test statistics are independent of nuisance parameters and functions.
- Typically, the asymptotical null distribution of the GLR statistic λ_n is nearly χ^2 with large degrees of freedom in the sense that

$$r\lambda_n \overset{d}{\simeq} \chi^2_{\mu_n}$$

- The asymptotic null distribution is independent of the nuisance parameters/functions.
- One does not have to derive theoretically the constant μ_n and r in order to use the GLR tests one can simply simulate the null distributions by setting nuisance parameters under the null hypothesis at reasonable values or estimates.

Nonparametric regression

Consider the following nonparametric model

$$Y_i = m(X_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

where ε_i are $i.i.d.$ random variables such that $E(\varepsilon_i) = 0$ and $Var(\varepsilon_i) = \sigma^2$.

We now apply the GLR test to the testing problem

$$H_0 : m(x) = \alpha_0 + \alpha_1 x \quad \text{versus} \quad H_1 : m(x) \neq \alpha_0 + \alpha_1 x,$$

Using the local linear fit with a kernel K and a bandwidth h, one one can obtain the estimator $\hat{m}_h(\cdot)$of the unknown function m under the full model.

The GLR statistic is

$$\lambda_{n,1} = \ell(\hat{m}_h, \hat{\sigma}) - \ell(\hat{m}_0, \hat{\sigma}_0) = \frac{n}{2} \log(RSS_0/RSS_1),$$

where $RSS_0 = \sum_{i=1}^n (Y_i - \hat{\alpha}_0 - \hat{\alpha}_1 X_i)^2$, $RSS_1 = \sum_{i=1}^n (Y_i - \hat{m}_h(X_i))^2$.

Nonparametric Regression

Under the null hypothesis and certain conditions, if $nh^{3/2} \to \infty$, the Wilks type of the result holds,

$$r_K \lambda_n \overset{a}{\sim} \chi^2_{r_K c_K |\Omega|/h},$$

where Ω denotes the support of X and

$$c_K = K(0) - 2^{-1}||K||^2, r_K = c_K/d_K, d_K = ||K - 0.5K * K||^2.$$

This result can be generalized to the testing problem

$$H_0 : m(x) = m(x, \theta) \quad \text{versus} \quad H_1 : m(x) \neq m(x, \theta).$$

Nonparametric Regression

Based on Wilks' phenomenon, the null distribution of the GLR statistic can be estimated by using the following conditional bootstrap method:

1. Obtain the parametric estimates $\hat{\alpha}_0$ and $\hat{\alpha}_1$ and nonparametric estimate $\hat{m}(x)$ under both and null and the alternative models. Fix the bandwidth at its estimated value $\hat{h}$ in the estimated stage.
2. Compute the GLR test statistic $\lambda_{n,1}$ and the residuals $\hat{\varepsilon}_i$ from the nonparametric model.
3. For each X_i, draw a bootstrap residual $\hat{\varepsilon}_i^*$ from the centered empirical distribution of $\hat{\varepsilon}_i$ and compute $Y_i^* = \hat{\alpha}_0 + \hat{\alpha}_1 X_i + \hat{\varepsilon}_i^*$.
4. Use the above bootstrap sample to construct the GLR statistic $\lambda_{n,1}^*$.
5. Repeat Steps 3 and 4 B times (say B=1,000) and obtain B values of the statistic $\lambda_{n,1}^*$.
6. Use the B values in Step 5 to determine the quantiles of the test statistic under H_0. The p-value is simply the percentage of $\lambda_{n,1}^*$ values greater than $\lambda_{n,1}$.

Varying-coefficient models

The varying-coefficient model assumes

$$Y = a_1(U)X_1 + \cdots + a_p(U)X_p + \varepsilon,$$

where ε is independent of covariates $(U, X_1, \ldots, X_p)$ and has mean zero and variance σ^2.

Suppose we have a random sample $\{(U, X_{i1}, \ldots, X_{i,p}, Y_i)\}$ from the above model. Let

$\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^T$ and $\boldsymbol{A}(U) = (a_1(U), \ldots, a_p(U))^T$. Then the model can be rewritten as

$$Y_i = \boldsymbol{A}(U_i)^T \boldsymbol{X}_i + \varepsilon_i.$$

The unknown coefficient functions $a_j(\cdot)$ can be estimated by using local linear regression techniques. For any given u_0 and u in a neighbourhood of u_0, it follows from the Taylor expansion that

$$a_j(u) \approx a_j(u_0) + a_j^{'}(u_0)(u - u_0) \equiv a_j + b_j(u - u_0).$$

Varying-coefficient models

Using the data with U_i around u_0, one can estimate the coefficient functions and their derivatives by the solutions to the following optimization problem:

$$\min_{a_j, b_j} \sum_{i=1}^{n} [Y_i - \sum_{j=1}^{p} \{a_j + b_j(U_i - u_0)\} X_{ij}]^2 K_h(U_i - u_0).$$

This yields a nonparametric estimator under the full model and the residual sum of squares under the nonparametric model

$$RSS_1 = \sum_{i=1}^{n} (Y_i - \hat{\boldsymbol{A}}(U_i)^T \boldsymbol{X}_i)^2,$$

where $\hat{\boldsymbol{A}}(U) = (\hat{a}_1(U), \ldots, \hat{a}_p(U))^T$.

Varying-coefficient models

One asks naturally if the coefficients in $\boldsymbol{A}(u)$ really vary with u. The former null hypothesis is parametric: $\boldsymbol{A}(u) = \boldsymbol{\beta}$.

Let us consider testing the following parametric null hypothesis:

$$H_0 : \boldsymbol{A}(u) = \boldsymbol{A}(u, \boldsymbol{\beta}).$$

We obtain the GLR test statistic

$$\lambda_{n,2} = \frac{n}{2} \log(RSS_0/RSS_1).$$

Under Certain conditions, if $\boldsymbol{A}(u, \boldsymbol{\beta})$ is linear in u or $nh^{9/2} \to 0$, then as $nh^{3/2} \to \infty$,

$$r_K \lambda_{n,2} \overset{d}{\simeq} \chi^2_{\mu_n},$$

where $u_n = \rho r_K c_K |\Omega|/h$ with $|\Omega|$ being the length of the support of U.

Varying-coefficient models

One asks also if certain covariates in X are related to the response Y . The former null hypothesis is

$$H_0 : a_1(\cdot) = \cdots = a_d(\cdot) = 0.$$

Under the null hypothesis, it is still a varying coefficient model:

$$Y = a_{d+1}(U)X_{d+1} + \cdots + a_p(U)X_p + \epsilon.$$

Denote by RSS_0^*, the resulting sum of the squares, defined similarly to RSS_1. Then following the same derivation as before, the GLR test statistic is

$$\lambda_{n,3} = \frac{n}{2} \log(RSS_0^*/RSS_1).$$

If $nh^{3/2} \to \infty$, then

$$r_K \lambda_{n,3} \overset{d}{\simeq} \chi^2_{dr_K c_K|\Omega|/h}.$$

Additive models

Additive models model a random sample $\{(Y_i, \boldsymbol{X}_i)\}$ by

$$Y_i = \alpha + \sum_{d=1}^{D} m_d(X_{di}) + \varepsilon_i, \quad i = 1, \ldots, n.$$

Fan and Jiang (2005) study the GLR test for the following hypothesis testing problem using the backfitting algorithm with the local polynomial smoothing technique to estimate nonparametric components:

$$H_0 : m_{D-d_0}(x_{D-d_0}) = \cdots = m_D(x_D) = 0,$$

Additive models

The GLR test is

$$\lambda_{n,6} = \ell(H_1) - \ell(H_0) = \frac{n}{2} \log \frac{RSS_0}{RSS_1}.$$

Put

$$u_n = c_K \sum_{d=D-d_0}^{D} \frac{|\Omega_d|}{h_d}, \sigma_n^2 = d_K \frac{|\Omega_d|}{h_d}, r_K = \mu_n / \sigma_n^2,$$

Under regularity conditions, Fan and Jiang (2005) established that

$$r_K \lambda_{n,6} \overset{d}{\simeq} \chi^2_{r_K \mu_n}.$$