Stat210A: Theoretical Statistics **Lecture Date: November 13, 2014**

Composite-Composite Testing II / The Bootstrap

Lecturer: Michael I. Jordan *Scribe: Maxim Rabinovich*

1 Beyond Generalized Likelihood Ratios

1.1 Recap of generalized likelihood ratios

In the previous lecture, we introduced **generalized likelihood ratio (GLR) tests** for distinguishing between a full model class $\mathcal{P} = \{P_\theta \colon \theta \in \Omega\}$ and a subclass $\mathcal{P}_0 = \{P_\theta \colon \theta \in \Omega_0\}$. The key ingredient in this test was the generalized likelihood ratio

$$\lambda(X) = \frac{\sup_{\theta \in \Omega} L(\theta)}{\sup_{\theta \in \Omega_0} L(\theta)},$$

where L is the likelihood function. When the suprema are attained at some points $\hat{\theta} \in \Omega$ and $\tilde{\theta} \in \Omega_0$, we can write

$$\lambda(X) = \frac{L(\hat{\theta})}{L(\tilde{\theta})}.$$

As before, we assume that this is the case, in which case $\hat{\theta}$ is the MLE under the full model and $\tilde{\theta}$ is the MLE under the restricted model.

Example 1. Suppose we have iid random variables from a bivariate normal

$$(X_i,\ Y_i) \sim \mathcal{N}\left(\begin{pmatrix}\mu_x \\ \mu_y\end{pmatrix},\ \begin{pmatrix}\sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2\end{pmatrix}\right),$$

where the free parameters are $\mu_x, \mu_y, \rho \in \mathbf{R}$ and $\sigma_x, \sigma_y \in \mathbf{R}_+$. One question we might ask is whether they are independent—that is, whether $\rho = 0$. In this case, we have

$$H_0 \colon \rho = 0 \quad \text{and} \quad H_1 \colon \rho \neq 0$$

and we can show

$$\log \lambda_n = -\frac{n}{2} \log\left(1 - \hat{\rho}^2\right) \qquad \left[\hat{\rho} = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (X_i - \bar{X})^2 \cdot \sum_i (Y_i - \bar{Y})^2}}\right].$$

Tests based on λ_n can in fact be shown to be equivalent to those based on

$$T_n = \frac{\hat{\rho}\sqrt{n-2}}{\sqrt{1 - \hat{\rho}^2}},$$

which follows a Student-t distribution on $n - 2$ degrees of freedom under H_0.

In the remainder of this section, we sketch the derivation of the limiting distribution of $2 \log \lambda_n(X)$ as the number of data points $n \to \infty$ and to introduce two tests which are asymptotically equivalent to GLR tests but which require different kinds of information. Further details on everything discussed can be found in Chapter 17 (Keener (2010)).

1.2 Asymptotics of $2\log\lambda_n$

To compute the limiting distribution of $2\log\lambda_n$, we assume enough regularity so that

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \Rightarrow \mathcal{N}(0,\ I^{-1}(\theta)),$$

where I denotes the Fisher information. We assume further that Ω is a smooth submanifold of $\mathbf{R}^r$ of dimension r, that $\Omega_0 \subset \Omega$ is a smooth q dimensional submanifold of Ω with $q < r$, and that both $\hat{\theta}_n$ and $\tilde{\theta}_n$ (that is, the MLEs under the full and restricted models, respectively) are consistent for θ. Our goal is then to compute the power function $\beta_n(\theta)$ both when $\theta \in \Omega_0$ and when θ is "near" Ω_0 in some sense. In the former case, we can prove **Wilks's theorem**, which states that

$$2\log\lambda_n \Rightarrow \chi^2_{r-q}.$$

The general case is more delicate and requires in particular the right notion of asymptotic "nearness." Specifically, if we suppose

$$\theta_n = \theta_0 + \frac{\Delta}{\sqrt{n}}, \quad \theta_0 \in \Omega_0,\ \Delta \in \mathbf{R}^r,$$

we can use the theory of contiguity to prove

$$2\log\lambda_n \Rightarrow \chi^2_{r-q}\left(c(\Delta)^2\right),$$

where $\chi^2_{r-q}(c^2)$ denotes a noncentral χ^2 on $r-q$ degrees of freedom and with noncentrality parameter c^2. As we have not yet encountered this distribution, we recall its definition.

Definition 2 (Definition 14.7, Keener (2010))**.** Suppose $Z_1, \ldots, Z_p$ are independent with

$$Z_1 \sim \mathcal{N}(\delta,\ 1) \ \text{ and } \ Z_j \sim \mathcal{N}(0,\ 1),\ 2 \le j \le p,$$

with $\delta \ge 0$. The distribution of

$$W = \sum_{j=1}^{p} Z_j^2$$

is called a **noncentral χ^2 distribution** on p **degrees of freedom** with **noncentrality parameter** δ^2, denoted $\boldsymbol{\chi^2_p(\delta^2)}$.

1.3 Asymptotically equivalent tests

In this section, we introduce two tests which complement GLR tests and which are asymptotically equivalent to GLR. The first of these is the **Wald test**, which applies to the testing situation

$$H_0\colon g(\theta) = 0 \ \text{ and } \ H_1\colon g(\theta) \neq 0,$$

where $g\colon \Omega \to \mathbf{R}^{r-q}$ is continuously differentiable with Jacobian $Dg \in \mathbf{R}^{r\times r-q}$ having full rank everywhere. With these assumptions, we can form Wald's test statistic

$$T_{W,n} = ng(\hat{\theta}_n)^T \left[Dg(\hat{\theta}_n)^T I(\hat{\theta}_n) Dg(\hat{\theta}_n)\right]^{-1} g(\hat{\theta}^n).$$

$T_{W,n}$ follows a χ^2 distribution, as the following lemma helps us show.

Lemma 3 (Lemma 14.9, Keener (2010))**.** *Let $\Sigma \in \mathbf{R}^{p\times p}$ be positive definite and let $Z \sim \mathcal{N}_p(\mu,\ \Sigma)$. Then*

$$Z^T\Sigma^{-1}Z \sim \chi^2_p(\mu^T\Sigma^{-1}\mu).$$

Using the fact that

$$Z_n = \sqrt{n}\left(\hat{\theta}_n - \theta\right) \Rightarrow \mathcal{N}(0,\ I(\theta)^{-1})$$

and the delta method applied to g, we deduce that

$$G_n = \sqrt{n}\left(g(\hat{\theta}_n) - g(\theta)\right) \Rightarrow \mathcal{N}\left(0,\ \left[Dg(\hat{\theta}_n)^T I(\hat{\theta}_n) Dg(\hat{\theta}_n)\right]^{-1}\right)$$

and, therefore, that when $\theta \in \Omega_0$,

$$T_{W,n} = G_n^T \left[Dg(\hat{\theta}_n)^T I(\hat{\theta}_n) Dg(\hat{\theta}_n)\right]^{-1} G_n \Rightarrow \chi^2_{r-q}.$$

By Wilks's theorem, we therefore see that $T_{W,n}$ has the same asymptotic distribution as $2\log\lambda_n$, so that Wald's test is asymptotically equivalent to a GLR.

The second test we consider is **Rao's score test**, which makes use of the test statistic

$$T_{S,n} = \frac{1}{n} \nabla \ell_n(\tilde{\theta}_n)^T I(\tilde{\theta}_n)^{-1} \nabla \ell_n(\tilde{\theta}_n),$$

where ℓ_n denotes the likelihood function of the data (after n observations). As before, we can show

$$T_{S,n} \Rightarrow \chi^2_{r-q}.$$

While these tests are all asymptotically equivalent in most cases, they do have some substantial differences. Where deploying a GLR requires computing MLEs under both the full and restricted model, Wald's test requires only the ability to compute the MLE under the full model and Rao's score test requires only the ability to compute the MLE under the restricted model. This can sometimes be advantageous from a computational perspective when computing the MLE for one or the other model poses a problem. On the other hand, in some instances where the strong regularity assumptions we have leaned on fails, asymptotic equivalence fails, and GLR can turn out to be more robust. Finally, even when the three tests are asymptotically equivalent overall, they can have different asymptotic power in different directions (corresponding to the choice of Δ above). In practice, all three are deployed in statistical computing software.

2 The Bootstrap

We now turn to **the bootstrap**, a very general way of estimating expectations under a model. Suppose to begin with that we have a dataset $X_{1:n}$ drawn iid from some unknown distribution P. The boostrap is based on sampling a new iid dataset $X^*_{1:n}$ based on X with

$$X^*_i \sim \hat{P}(X) = \frac{1}{n}\sum_i \delta_{X_i}.$$

These samples can then be used to form Monte Carlo estimates of expectations under P.

The bootstrap also has a powerful and illuminating interpretation as a **plug-in estimator** of the unknown distribution P. Indeed, by the Glivenko-Cantelli theorem,

$$||P - \hat{P}|| \to 0.$$

The idea of the bootstrap is to take advantage of this by plugging in $\hat{P}$ into any expression that depends on P (that is, into any functional of the distribution P). For example, we can write the quadratic risk for an

estimator $\hat{\theta}$ as

$$\begin{aligned} R(P) &= E_P\left[(\hat{\theta} - \theta(P))^2\right] \\ &= \left(E_P\left[\hat{\theta}\right] - \theta(P)\right)^2 + E_P\left[(\hat{\theta} - E_P[\hat{\theta}])^2\right], \end{aligned}$$

in which the variance of the estimator is clearly a functional of P. If we let $\hat{\theta}^*$ denote the value of the estimator computed using $X_i^* \sim \hat{P}$, we can therefore estimate the variance by

$$E_{\hat{P}}\left[(\hat{\theta}^* - E_{\hat{P}}[\hat{\theta}^*])^2\right].$$

In practice, it may be difficult or impossible to compute this expectation analytically, in which case we can form a Monte Carlo sample by drawing B datasets X_b^* from $\hat{P}$ (more precisely, the distribution on data sets of size n induced by iid sampling from $\hat{P}$, so $\hat{P}^n$) and using the approximation

$$\frac{1}{B}\sum_b \left(\hat{\theta}(X_b^*) - \frac{1}{B}\sum_{b'}\hat{\theta}(X_{b'}^*)\right)^2 \approx E_{\hat{P}}\left[(\hat{\theta}^* - E_{\hat{P}}[\hat{\theta}^*])^2\right].$$

This gives us the original view of the bootstrap as a sampling procedure.

A particularly remarkable feature of the bootstrap is its ability to estimate quantities that might initially seem totally out of reach, among these the bias of an estimator $\hat{\theta}$. Although we generally view P as derived from a parameter θ, the correspondence is generally one-to-one, so that we can equally write $\theta = \theta(P)$. If this functional extends to discrete distributions, so that it can be applied to $\hat{\theta}$, we can actually estimate the bias by

$$\hat{b} = E_{\hat{P}}\left[\hat{\theta}^*\right] - \theta(\hat{P}).$$

Plugging this into the expression for the risk, and combining with the variance estimator above, we thereby obtain a generic way to estimate the risk under an unknown distribution P.

Further details on the bootstrap can be found in Chapter 19 (Keener (2010)).

References

Keener, R. (2010). *Theoretical Statistics: Topics for a Core Course*. Springer, New York, NY.