The Rate Distortion Test

Second Entropy Workshop at EPFL

Peter Harremoës and Nisheeth Sriravastava

CWI, Amsterdam

September 9, 2008

Outline of talk

- Classical theory of testing goodness-of-fit
- The idea
- The rate distortion test
- Some results on why this is not a bad idea
- Conclusion

The Neyman Pearson Lemma

Consider the hypotheses $Q = P_1$ vs. $Q = P_0$. For $r \in [0;1]$ let $Ac(r)$ be the acceptance region defined by

$$Ac(r) = \left\{ \omega \mid \frac{d(P_1^n)}{d(P_0^n)}(\omega) \geq r \right\}$$
$$= \left\{ \omega \mid E_{Emp_n(\omega)}\left[\log \frac{d(P_1)}{d(P_0)}\right] \geq \frac{1}{n} \log r \right\}.$$

Theorem (Neyman-Pearson Lemma)

Let $X_1, X_2, \ldots, X_n$ be independent distributed according to Q. Let the error probabilities be defined by

$$\alpha_0^* = P_0^n(Ac(r));$$
$$\alpha_1^* = P_1^n(\complement Ac(r)).$$

Let B be another decision region with error probabilities α_0 and α_1. Then if $\alpha_0 \leq \alpha_0^*$ then $\alpha_1 \geq \alpha_1^*$.

Testing goodness-of-fit

- Consider a random variable with an unknown continuous distribution function F

Testing goodness-of-fit

- Consider a random variable with an unknown continuous distribution function F
- Based on a sample of size n we want to to test the hypothesis that $F = G$ for some known distribution function G.

Testing goodness-of-fit

- Consider a random variable with an unknown continuous distribution function F
- Based on a sample of size n we want to to test the hypothesis that $F = G$ for some known distribution function G.
- Divide $\mathbb{R}$ into k bins.

Testing goodness-of-fit

- Consider a random variable with an unknown continuous distribution function F
- Based on a sample of size n we want to to test the hypothesis that $F = G$ for some known distribution function G.
- Divide $\mathbb{R}$ into k bins.
- Popular choice is k interval with equal probability according to G, i.e.

$$\left[G^{-1}\left(\frac{j-1}{k}\right); G^{-1}\left(\frac{j}{k}\right) \right]$$

where $j = 1, 2, ..., k$.

Testing goodness-of-fit

- Consider a random variable with an unknown continuous distribution function F
- Based on a sample of size n we want to to test the hypothesis that $F = G$ for some known distribution function G.
- Divide $\mathbb{R}$ into k bins.
- Popular choice is k interval with equal probability according to G, i.e.

$$\left[G^{-1}\left(\frac{j-1}{k}\right); G^{-1}\left(\frac{j}{k}\right)\right]$$

  where $j = 1, 2, ..., k$.
- Test if the empirical distribution on the bins is uniform.

Testing goodness-of-fit

- Consider a random variable with an unknown continuous distribution function F
- Based on a sample of size n we want to to test the hypothesis that $F = G$ for some known distribution function G.
- Divide $\mathbb{R}$ into k bins.
- Popular choice is k interval with equal probability according to G, i.e.

$$\left[G^{-1}\left(\frac{j-1}{k}\right); G^{-1}\left(\frac{j}{k}\right)\right]$$

  where $j = 1, 2, ..., k$.
- Test if the empirical distribution on the bins is uniform.
- Let k grow with the sample size n.

Power divergence statistics

The goodness-of-fit statistic is usually one of the *power divergence statistics* defined by

$$D_\alpha(P,Q)=\sum_{j=1}^{k} q_j\,\phi_\alpha\left(\frac{p_j}{q_j}\right),\quad \alpha\in\mathbb{R},$$

for the power function ϕ_α of order $\alpha\in\mathbb{R}$ given in the domain $t>0$ by the formula

$$\phi_\alpha(t)=\begin{cases}\frac{t^\alpha-\alpha(t-1)-1}{\alpha(\alpha-1)} & \text{when} \quad \alpha(\alpha-1)\neq 0\\ -\ln t+t-1 & \text{when} \quad \alpha=0\\ t\ln t-t+1 & \text{when} \quad \alpha=1\end{cases}.$$

Important examples

- The Pearson statistic ($\alpha = 2$),
- The Neyman statistic ($\alpha = -1$),
- The log-likelihood ratio ($\alpha = 1$),
- The reversed log-likelihood ratio ($\alpha = 0$)
- The Freeman-Tukey statistic ($\alpha = 1/2$).
- Note that

$$D_\alpha(P, U) = \frac{k^{\alpha-1}\mathrm{IC}_\alpha(P) - 1}{\alpha(\alpha - 1)}$$

where IC_α is the *index of coincidence*

$$\mathrm{IC}_\alpha(P) = \sum_{j=1}^{k} p_j^\alpha = e^{(1-\alpha)H_\alpha(P)} .$$

and $H_\alpha(P)$ is the *Rényi entropy* of order α

Efficiency

- The test is efficient in the sense of Lehman and hodge.

Efficiency

- The test is efficient in the sense of Lehman and hodge.
- If n/k tend to a constant then the Pearson statistic is most Pitman efficient.

Efficiency

- The test is efficient in the sense of Lehman and hodge.
- If n/k tend to a constant then the Pearson statistic is most Pitman efficient.
- We focus on the typical situation where $k = k_n$ satisfies a conditions of the type
$$\frac{n}{k \log k} \to \infty \text{ for } n \to \infty.$$

Efficiency

- The test is efficient in the sense of Lehman and hodge.
- If n/k tend to a constant then the Pearson statistic is most Pitman efficient.
- We focus on the typical situation where $k = k_n$ satisfies a conditions of the type
$$\frac{n}{k \log k} \to \infty \text{ for } n \to \infty.$$
- In this case the *Pitman asymptotic relative efficiencies* of all statistics D_α, $\alpha \in \mathbb{R}$ coincide.

Efficiency

- The test is efficient in the sense of Lehman and hodge.
- If n/k tend to a constant then the Pearson statistic is most Pitman efficient.
- We focus on the typical situation where $k = k_n$ satisfies a conditions of the type
$$\frac{n}{k \log k} \to \infty \text{ for } n \to \infty.$$
- In this case the *Pitman asymptotic relative efficiencies* of all statistics D_α, $\alpha \in \mathbb{R}$ coincide.
- In this situation preferences between these statistics must be based on the *Bahadur efficiencies* $\mathrm{BE}(D_{\alpha_1} \mid D_{\alpha_2})$.

Efficiency

- The test is efficient in the sense of Lehman and hodge.
- If n/k tend to a constant then the Pearson statistic is most Pitman efficient.
- We focus on the typical situation where $k = k_n$ satisfies a conditions of the type
$$\frac{n}{k \log k} \to \infty \text{ for } n \to \infty.$$
- In this case the *Pitman asymptotic relative efficiencies* of all statistics D_α, $\alpha \in \mathbb{R}$ coincide.
- In this situation preferences between these statistics must be based on the *Bahadur efficiencies* $\mathrm{BE}(D_{\alpha_1} \mid D_{\alpha_2})$.
- $\mathrm{BE}(D_1 \mid D_2) = \infty$ (Quine and Robinson, 1985).

Consistency

For $\alpha \in \mathbb{R}$ and a sequence of alternatives P_n we say that the model satisfies the *Bahadur condition* if there exists a constatnt $\Delta_\alpha > 0$ such that

$$D_\alpha(P_n, U) = \Delta_\alpha \ .$$

The statistic $D_\alpha(\hat{P}_n, U)$ is *consistent* if the Bahadur condition holds and

$$D_\alpha(\hat{P}_n, U) \longrightarrow 0 \quad \text{under } U \text{ in probability}$$
$$D_\alpha(\hat{P}_n, U) \longrightarrow \Delta_\alpha \quad \text{under } P_n \text{ in probability.}$$

Theorem

The divergence $D_\alpha(\hat{P}_n, U)$ is consistent if

$$\lim_{n\to\infty} \frac{k}{n} = 0 \text{ for } \alpha \in [0; 2] \ ,$$
$$\lim_{n\to\infty} \frac{k \log k}{n} = 0 \text{ for } \alpha > 2.$$

Consistency holds for all f-divergences that are uniformly continuous.

Bahadur function and Bahadur efficiency

- For $\alpha \in \mathbb{R}$ we say that the Bahadur function for the statistic $D_\alpha(\hat{P}_n, U)$ exists if there exists a sequence $c_{\alpha,n} > 0$ and a continuous function $g_\alpha : (0,\infty) \to (0,\infty)$ such that under $\mathcal{H}$

$$\lim_{n\to\infty} -\frac{c_{\alpha,n}}{n} \ln \mathrm{P}(D_\alpha(\hat{P}_n, U) \geq \Delta) = g_\alpha(\Delta), \quad \Delta > 0.$$

Bahadur function and Bahadur efficiency

- For $\alpha \in \mathbb{R}$ we say that the Bahadur function for the statistic $D_\alpha(\hat{P}_n, U)$ exists if there exists a sequence $c_{\alpha,n} > 0$ and a continuous function $g_\alpha : (0,\infty) \to (0,\infty)$ such that under $\mathcal{H}$

$$\lim_{n\to\infty} -\frac{c_{\alpha,n}}{n} \ln \mathrm{P}(D_\alpha(\hat{P}_n, U) \geq \Delta) = g_\alpha(\Delta), \quad \Delta > 0.$$

- Assume that the statistics $D_{\alpha_1}(\hat{P}_n, U)$ and $D_{\alpha_2}(\hat{P}_n, U)$ are consistent and that the corresponding Bahadur functions g_{α_1} and g_{α_2} exist. The Bahadur efficiency is defined by

$$BE(D_{\alpha_1} \mid D_{\alpha_2}) = \frac{g_{\alpha_1}(\Delta_{\alpha_1})}{g_{\alpha_2}(\Delta_{\alpha_2})} \lim_{n\to\infty} \frac{c_{\alpha_1,n}}{c_{\alpha_2,n}} .$$

Case $\alpha \geq 1$

Theorem

If $k = k_n$ increases so slowly that

$$\frac{n}{k \log k} \to \infty$$

then the Bahadur efficiency of the statistic D_{α_1} with respect to D_{α_2} satisfies the relation

$$\mathrm{BE}(D_{\alpha_1} \mid D_{\alpha_2}) = \infty$$

for all $1 \leq \alpha_1 < \alpha_2$.

Proof.

See Harremoës and Vajda IEEE Trans. Inform. Theory Jan. 2008 and Haremoës and Vajda ISIT 2008. □

Case $0 < \alpha \leq 1$

Theorem

If $k = k_n$ increases so slowly that

$$\frac{n}{k \log k} \to \infty$$

then the Bahadur efficiency of the statistic D_{α_1} with respect to D_{α_2} satisfies the relation

$$\mathrm{BE}(D_{\alpha_1} \mid D_{\alpha_2}) = \frac{\Delta_1}{\Delta_2}$$

for all $0 < \alpha_1 < \alpha_2 \leq 1$.

Proof.

The extreme case is a sequence of alternatives that are uniform on subsets. In this case

$$D_{\alpha_1}(P_n \| U) = D_{\alpha_2}(P_n \| U) = \log \frac{|\text{support of } P_n|}{k}.$$

□

Discussion

A core observation is that

$$\inf_{D_\alpha(P,U)\geq\Delta} D(P,U) \begin{cases} =0 & \text{for } \alpha>1 \\ >0 & \text{for } \alpha\in\left]0;1\right[\end{cases}.$$

Absolute continuity Note that $D(P\|Q)<\infty$ implies that $P\ll Q$.

Contiguity Similarly $D(P_n\|Q_n)\to\Delta$ implies that $P_n\lhd Q_n$.

Assume that $P=\delta_a$ and Q is continuous. Then $P_n\ntriangleleft Q_n$.

Theorem (Informal version)

Information divergence is more Bahadur efficient than any Rényi divergence of order $\alpha\in\left]0,1\right[$ for testing a P against Q when $P\not\ll Q$ except if $P\perp Q$.

In practice D_a is not very efficient if $D(P\|Q)$ is large.

Open questions

- How many bins should be chosen? I.e. How should we choose k as a function of n?
- How should we choose the shapes of the bins?
- Should the bins be chosen with equal probability or are some less uniform choice of bins better?

A different approach to testing

Let $X_1, X_2, \cdots$ denote a sequence of binary random variables. We want to test the null hypothesis that they come from a Bernoulli $(1/2, 1/2)$-source.

If H_0 is true the entropy of the sequence is maximal and it is not possible to compress it.

Choose your favorite data compressor and see how much it is able to compress X_1^n.

If no or only a little compression is obtained accept H_0. Otherwise reject.

A rate distortion version of the previous idea

Let $d : \mathcal{X} \times \hat{\mathcal{X}} \to R$ be a distortion function. At some distortion level d_0 rate distortion theory tells us how optimally to compress data at distortion level d_0 if the distribution of of X is Q. Compress data into a binary sequence and test if the sequence is Bernoulli $(1/2, 1/2)$.

Problems:

- Depends on data compressor.
- Is hard to analyse.

The rate distortion test

Let $d : \mathcal{X} \times \hat{\mathcal{X}} \to R$ be a distortion function and Q a probability distribution on $\mathcal{X}$. Choose a sequence of distortion levels d_n such that $d_n \to 0$ for $n \to \infty$. For each n find the Markov kernel Ψ_n such that Ψ_n gives the optimal coupling corresponding to distortion level d_n. Use

$$D\left(\Psi_n\left(Emp_n\left(\omega\right)\right)\|\Psi_n\left(Q\right)\right)$$

as statistic.

Example: test of uniformity

We consider a set A with l elements. The set has no particular structure so we use Hamming distortion as distortion function. Our null hypothesis is $P = U$ where u denotes the uniform distribution on A. In this case the Markov kernel Ψ_{d_0} has the form

$$\Psi_{d_0} : x \longrightarrow \alpha\delta_x + (1-\alpha)\, U$$

for some value $\alpha \in [0;1]$ determined by d_0. The Markov kernel maps the uniform distribution into the uniform distribution. Therefore the statistic of the rate distortion test has the for

$$D\left(\alpha Emp_n(\omega) + (1-\alpha)\, U \| U\right).$$

If α is small the statistic can be approximated by the Pearson statistic that is Pitman efficient.

Example: Test of normality

We consider the real numbers with squared Euclidian distance as distortion function. Our null hypothesis is $P = \Phi$ where Φ denotes the standard Gaussian distribution. The optimal Markov kernel for the rate distortion problem sends x into the distribution of $\alpha x + \left(1-\alpha^2\right)^{1/2} Z$ where Z is a standard Gaussian random variable. We see that the Gaussian distribution is mapped into it self. Thus the statistic of the rate distortion test is

$$D\left(\alpha X + \left(1-\alpha^2\right)^{1/2} Z \| \Phi\right)$$

where we have identified the random variable $\alpha X + \left(1-\alpha^2\right)^{1/2} Z$ with its distribution. This Markov kernel can be rewritten as

$$D\left(\alpha X + \left(1-\alpha^2\right)^{1/2} Z \| \Phi\right) = D\left(X + \left(\frac{1}{\alpha^2} - 1\right)^{1/2} Z \| \Phi\left(0, \alpha^2\right)\right)$$

so the Markov kernels essentially smooth data by adding an independent Gaussian random variable with variance $\alpha^{-2} - 1$. The idea of smoothing data is well-known in statistics.

Example: Angular data

We consider data with values on the circle s_1 that we can identify with $\mathbb{R}/2\pi\mathbb{Z}$. As distortion function we shall use $4\cos^2\left(\frac{\theta_2-\theta_1}{2}\right)$. We shall test the hypothesis $P = U$ where U denotes the uniform distribution on the circle. The optimal Markov kernel is a smoothing by adding a von Mises distribution

$$\frac{\exp\left(\kappa\cos\left(\theta\right)\right)}{2\pi I_0\left(\kappa\right)}$$

where I_0 is the modified Bessel function of order 0 with parameter κ determined by the distortion level. The Markov kernel maps the uniform distribution into the uniform distribution.

Consistency

If data is generated by P then

$$D\left(\Psi_n\left(Emp_n\left(\omega\right)\right)\|\Psi_n\left(Q\right)\right)\to D\left(P\|Q\right)$$

in probability.

Efficiency

- The rate distortion test is efficient in the sense of Hodge and Lehman.
- We have no results on the Pitman efficiency of the rate distortion test but conjecture that it is Pitman efficient under regularity conditions.
- What can be said about Bahadur efficiency?

Exponential families

Let P_μ denote an exponential family in its mean value representation. Define a distortion function by

$$d(\lambda, \mu) = D\left(P_\lambda \| P_\mu\right).$$

Then d is a Bregman divergence and characerizes the exponential family. The rate distortion test is Bahadur efficient against alternatives in the exponential family.

Bahadur efficiency

Theorem

Let G denote a compact group and let d denote a distortion function that is continuous and invariant under the group action. Then the rate distortion test is Bahadur efficient.

This result can be extended to compact sets under regularity conditions for which we have no counterexamples.

Conclusion

- In the rate distortion test the question of the shape and probability of the bins can be replaced with the question of choosing a distortion function.
- The question of the number of bins (or the rate of the rate distortion test) can be discussed without confusion of the other now solved problems.
- All our results on efficiency are positive.

Open questions

- Asymptotic normality.
- More efficiency results.