Statistical Methods III: Spring 2013

Jonathan Wand

Stanford University

KLIC + model select

Outline

Testing shapes based on functional relationships

Challenges

1. Finding best model
   - imposing theoretically motivated constraints
   - minimizing assumptions made for convenience, which may mislead
   - unconstrained curves are useful as specification test
2. Often there are more than one possible model.
   - goal: to find *set* of “best” models, equal good fits
   - provide: level of confidence in this set, prespecified test size
3. Comparisons of shape constrained models are non-standard.
   - within inequality constraints, dimensionality is stochastic
   - requires formultating least favorable nulls
   - complicated...

KLIC

Definition (Kullback-Leibler information criterion (KLIC))

Let

- $y = y_1, ...y_n$ be a random sample with density $f(y) = \prod f(y_i)$.
- $g(y) = \prod g(y_i)$, which we will call the model density.

KLIC provides a summary of the fit of g as an approximation to f,

$$KLIC = \int f(y) \log \left(\frac{f(y)}{g(y)} \right) dy = E \log \frac{f(y)}{g(y)}$$

Note:

- what is the value of KLIC if $g = f$?

KLIC

$$\begin{aligned} KLIC &= \int f(y) \log \left(\frac{f(y)}{g(y)} \right) dy \\ &= \int f(y) \log f(y) dy - \int f(y) \log g(y) dy \\ &= C_f - \int f(y) \log g(y) dy \\ &= C_f - E \log g(y) \end{aligned}$$

Notes:

- for a given y, C_f does not depend on g (it is a constant)
- we often let density g depend on unknown parameters, $g(y, \theta)$
- the $\hat{\theta}$ that minimizes $-\log g(y, \theta)$ is the (quasi-)MLE maximizing:

$$L(\theta) = \sum \log g(y_i, \theta)$$

- this *also* minimizes KLIC

KLIC

What do we know? what do we not know? what do we want to know?

- if we knew f, we would not need g or KLIC analysis that follows!
- in general, we do not know f (exception: bootstrap where we generate data)
- we choose g (e.g., likelihood)
- often we choose a parametric and functional form with unknown parameters, e.g,. a logit

$$g(y) = g(y; x, \beta) = \sum y_i \log \Lambda(x_i\beta) + (1 - y_i) \log \Lambda(x_i\beta)$$

  In this approach we know *impose* the logit form and additive aggregator function, but treate β as unknown.
- KLIC is of interest with respect a particular g, not a family of distributions with unspecified parameters...
- We might first ask: what is distance between f and $\hat{g}$, where $\hat{g}$ is the density conditional on filling in parameters θ at a particular value $\hat{\theta}$.

KLIC

- We need to keep track of data used to estimate $\hat{\theta}$ versus data used in expectation!
- Assume you have one sample $\tilde{y}$, which you use to estimate $\hat{\theta}(\tilde{y})$ conditional on choice of g;
- we write $\hat{\theta}$ as a function of $\tilde{y}$ in order to emphasize that some draw from y gives us MLE!
- Other stuff remains unchanged, we are going to integrate over distribution of all possible draws of y:

$$
\begin{aligned}
KLIC &= C_f - \int f(y) \log g(y, \hat{\theta}(\tilde{y}))dy \\
&= C_f - E_y \log g(y, \hat{\theta}(\tilde{y}))
\end{aligned}
$$

Expected KLIC

- $\tilde{y}$ produces a single $\hat{\theta}(\tilde{y})$
- we are next interested in the expected difference between f and g, where we will condition on an estimation method for picking $\hat{\theta}$...
- this gives us the Expected KLIC:

$$\begin{aligned} E(KLIC) &= E_{\tilde{y}} C_f - E_{\tilde{y}} E_y \log g(y, \hat{\theta}(\tilde{y})) \\ &= C_f - E_{\tilde{y}} E_y \log g(y, \hat{\theta}(\tilde{y})) \end{aligned}$$

because C_f is a constant.

KLIC

Omitting C_f, can describe (sort of) E(KLIC) in terms of

$$\begin{aligned} T &= -E_y E_{\tilde{y}} \log g(\tilde{y}, \hat{\theta}(y)) \\ &= -L(\hat{\theta}) + h(k, n) \end{aligned}$$

- k is dimensionality of model,
- n is sample size
- h is a function
- we will see that different model criterion have different h, see AIC ($h = k$) and BIC ($1/2k \log n$) in following slides
- $h(k, n)$ is the amount we need to add (over average) to log-likelihood in order to recover Expected KLIC if g includes correct model!

AIC

Definition (Akaike information criterion (AIC))

Let k be the number of parameters in the model, and L be the maximized value of the log-likelihood, then

$$\mathrm{AIC} = -2 \log L + 2k$$

- operationalizes trade-off between goodness of fit and complexity/dimensionality (why?)
- provides information only relative to other models (why?)
- when used for comparing nested models, is related to LRT (how/why?)
- what happens when comparing non-nested models with same dimensionality?

BIC

Definition (Bayesian information criterion (BIC))

Let k be the number of parameters in the model, n be number of observations, and L be the maximized value of the log-likelihood, then

$$\text{BIC} = -2 \log L + k \log n$$

- again operationalizes trade-off between goodness of fit and complexity/dimensionality (why?)
- again provides information only relative to other models (why?)
- odd assumptions (puts an equal prior on all models, irrespective of k), and lack of optimality on MSE criteria