A tale of two transforms

Leigh Roberts
Victoria University of Wellington, New Zealand

5 December 2017

Aims

Aims:

- Briefly to compare Khmaladze transforms 1 and 2
- to give some intuition underlying KT2
- to apply KT2 to simple simulated Bernoulli trials:
- to compare practicability of KT1 and KT2 in testing gof in distn free manner

Both KTs purport to test gof by distn free methods

- It is the best of times
  - Elegant and powerful transforms of stochastic processes are available
  - ostensibly very useful methodologies

Both KTs purport to test gof by distn free methods

- It is the best of times
    - Elegant and powerful transforms of stochastic processes are available
    - ostensibly very useful methodologies
- It is the worst of times
    - Almost no-one is applying these transforms

KT1

- The basic empirical process

$$v_n(t) = \frac{1}{\sqrt{n}} \left[F_n(x) - F_\theta(x) \right]$$

  is in the nature of a Brownian bridge.
- Usual projection is not invertible

$$u_n(x) = v_n(x) - g(x, \theta) \int h(y, \theta) \, dv_n(y) + o_P(1)$$

- But this transform from BB to Brownian motion is invertible

$$dv(t) = -\frac{v(t)}{1-t} \, dt + dw(t)$$

$$v(t) = (1-t) \int_0^t \frac{1}{1-s} \, dw(s)$$

This gives rise to the KT1

$$dw_{n,\theta}(x) = \sqrt{n}\left[dF_n(x) - h_\theta(x)dF_\theta(x)\Gamma_{x,\theta}^{-1}\int_{y=x}^{\infty} h_\theta(y)dF_n(y)\right]$$

This is a regression of $dF_n(x)$ on $\int_{y=x}^{\infty} h_\theta(y)dF_n(y)$
The limiting distn of $w_{n,\theta}(x)$ is indeed a BM, so we have achieved a distn free test. But implementation is awkward:

- hard to program: two interlocking integrals
  - two inter-related integrals, one over 'future' values
  - coding is very dependent on distributional family chosen
- underlying maths is technical, and not so intuitive
- cannot be simulated in real time; assume ML estimates known from whole sample
- modern tendency to eschew theory in favour of numerical methods, applied blindly through statistical packages

D = data, F = data generating process;
v = empirical process; w = BM

KT1: distn free testing of gof
D1 + F1 $\rightarrow$ v1 $\rightarrow$ w1
D2 + F2 $\rightarrow$ v2 $\rightarrow$ w2
D3 + F3 $\rightarrow$ v3 $\rightarrow$ w3
D1 + F2 $\rightarrow$ v4 $\rightarrow$ w4 ??

KT2: invertible transforms of empirical processes between 'any' stochastic processes
v1 $\leftrightarrow$ v2   v1 $\leftrightarrow$ v3   v2 $\leftrightarrow$ v3
So choose v3 to be simple, generic
Is $v1 = v2$ in distn? Is v1 generated by F1?
How effective to test these after transforming into v3?

Tale of two operators

Consider the multinomially distd

$$Y_{in} = \frac{\nu_{in} - n\,p_i}{\sqrt{n\,p_i}} \qquad \text{where} \qquad \sum_{i=1}^{m} p_i = 1$$

For a suitable limiting distribution, consider $X \sim \mathcal{N}(0, I)$ and the projection orthogonal to X (Khmaladze 2013)

$$Y = (I - \sqrt{p}\,\sqrt{p}^T)X \qquad \text{where} \qquad \sqrt{p}^T = (\sqrt{p}_1, \sqrt{p}_2, \ldots, \sqrt{p}_m)$$

Then

$$\text{Corr}(Y_{in}) = \text{Corr}(Y) = I - \sqrt{p}\,\sqrt{p}^T$$

A related operator is $I - 2\sqrt{p}\,\sqrt{p}^T$, provided $\| \sqrt{p} \| = 1$, reflecting in an axis perpendicular to $\sqrt{p}$.

Reflection

To some extent, distns can be related via a common 'base' distn:

$$Y = (I - \sqrt{p}\sqrt{p}^T)X \qquad Z = (I - \sqrt{r}\sqrt{r}^T)X$$

These projections involve only the second moment. But we are mainly concerned with zero mean Gaussian processes.

More generally, to move from one distn to another, define the reflection

$$U = I - \frac{2}{\| \sqrt{p} - \sqrt{r} \|^2}(\sqrt{p} - \sqrt{r})(\sqrt{p} - \sqrt{r})^T$$

whence $U\sqrt{p} = \sqrt{r}$ and $U\sqrt{r} = \sqrt{p}$.

For simple null, U maps from one empirical process to another.

For composite hypotheses with single parameter θ needing to be estimated from the data, need U to simultaneously map $\sqrt{p} \to \sqrt{r}$ and the score function $h_p \to h_r$. For $\theta = \theta_1, \theta_2, \ldots, \theta_K$, need U simultaneously to map $\sqrt{p} \to \sqrt{r}$ and the K derivatives in the (gradient) score function from one distn to the other.

For sufficiently large K, the vector spaces for discrete distns may not have sufficient dimensions to accommodate all constraints on the operator U. So we group the data artificially in order to increase the available dimensionality.
A single data 'point' is then a subgroup of the original sample, which subgroup is 'rotated' into a simpler 'empirical' process.

Simulation of Bernoulli Trials

Bernoulli trials
mn independent trials, but with common distributional family of success and failure depending on a covariate X_{ij}, for $1 \leq i \leq m, 1 \leq j \leq n$.
Y_{ij} = outcome = 1 for success, 0 for failure.
For the jth group of trials, $X_j = (X_{1j}, X_{2j}, \ldots, X_{mj})$ and $Z_j = (Y_{1j}, Y_{2j}, \ldots, Y_{mj})$

Order the results lexicographically.
Eg: when $m = 1$,
$y = 0, 1$ corresponds to $z = 1, 2 = 2^m$.
Eg: when $m = 2$,
$(y_1, y_2) = 00, 01, 10, 11$ corresponds to $z = 1, 2, \ldots, 4 = 2^m$.

Underlying joint distn given by

$$P_\theta(dx, dz) = P_{x,\theta}(dz)H(dx) \tag{1}$$

where θ is a vector of K parameters, and the distn H is unspecified. Denoting Borel sets in the respective spaces by A and B, we define the following centred empirical process:

$$\alpha_n^P(A, B, \theta) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \left[\mathbb{1}_{\{Z_j \in B\}} - P_\theta(B \middle| X_j) \right] \mathbb{1}_{\{X_j \in A\}} \tag{2}$$

Considering functions $\phi(x, z)$ in $L_2(P_\theta)$, the space of square integrable functions wrt the measure P_θ, we rewrite (2) in the form

$$\alpha_n^P(\phi, \theta) = \int \int \phi(x, z) \alpha_n^P(dx, dz, \theta) \tag{3}$$

Outline

We wish to transform processes defined on $L_2(P_\theta)$ to processes defined on say $L_2(Q_\theta)$, where by analogy to (1) we have

$$Q_\theta(dx, dz) = Q_{x,\theta}(dz)H(dx)$$

Consider functions $\phi \in L_2(P_\theta)$ and $\psi \in L_2(Q_\theta)$. Provided that P_θ and Q_θ are equivalent, define

$$\ell_{x,\theta}(dz) = \sqrt{\frac{dQ_{x,\theta}(dz)}{dP_{x,\theta}(dz)}} \tag{4}$$

so that $\psi \in L_2(Q_\theta)$ if and only if $\ell\psi \in L_2(P_\theta)$.

We simplify the conditional $Q_{x,\theta}(y)$ measure so that the numerator of the quantity under the square root in (4) depends on neither x nor θ: we may then write

$$\ell_{x,\theta}(z) = \sqrt{\frac{\prod_{i=1}^{m} q^{1-y_i}(1-q)^{y_i}}{\prod_{i=1}^{m} p_{x,\theta}^{1-y_i}(1-p_{x,\theta})^{y_i}}} \tag{5}$$

with q and $p_{x,\theta}$ denoting probabilities of failure, and $y = 1$ indicating success.

The transformation of the empirical process on $L_2(P_{x,\theta})$ to that on $L_2(Q)$ is effected as follows. From Khmaladze (2017) and (3),

$$\alpha_n^Q(\psi, \theta) = \alpha_n^Q(\psi) = \alpha_n^P(U\ell\psi, \theta) \tag{6}$$

in which U is a unitary operator.

We wish to carry out goodness of fit tests as to whether or not the data in (2) could reasonably have been generated by P_θ, for some value of θ. A standard means of doing this would be to employ the KS statistic, defined from (3) as

$$\max_{\phi \in \Phi} |\alpha_n^P(\phi, \theta)| \tag{7}$$

where the maximum is taken over all functions ϕ in a class Φ of functions in $L_2(P_\theta)$.

Kolmogorov Smirnov test on rotated distn

The point of transforming to the Q space, however, is that we may instead employ the KS statistic in the Q space:

$$\max_{\psi \in \Psi} |\alpha_n^Q(\psi)| \tag{8}$$

and test goodness of fit by an empirical process defined on a simpler measure Q depending on neither x nor θ.
A suitable class Ψ is generated by indicator functions in x and z:

$$\psi_{x_0, z_0}(x, z) = \mathbb{1}_{\{x \leq x_0\}} \left[\mathbb{1}_{\{z \leq z_0\}} - 2^{-m} z_0 \right] \tag{9}$$

in which ψ is regarded as a vector of length 2^m, so that the operator U assumes the form of a $2^m \times 2^m$ orthogonal matrix.

Target distn

The original process is generated by P, with normalised score functions $a_1, a_2, \ldots, a_K$: the vectors $1, a_1, a_2, \ldots, a_K$ are mutually orthogonal (wrt the P measure). The target process has $1, b_1, b_2, \ldots, b_K$ mutually orthogonal wrt the Q measure.
As the target process for the Bernoulli trials we choose the logistic distribution with parameter set to zero, and covariates to unity.
For $m = 1$, the target score function becomes $b_1 = (-1, 1)$; while for $m = 2, 3$ and 4, b_1 becomes respectively

$$\sqrt{2}(-1, 0, 0, 1) \qquad \frac{1}{\sqrt{3}}(-3, -1, -1, 1, -1, 1, 1, 3)$$

$$\frac{1}{\sqrt{4}}(-4, -2, -2, 0; -2, 0, 0, 2; -2, 0, 0, 2; 0, 2, 2, 4)$$

More generally, however, the target score functions can be arbitrary, save for the orthogonality constraints?

Choosing covariates

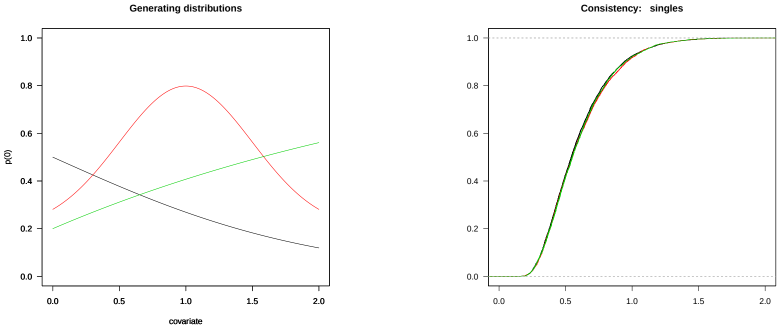

Figure:
Choosing covariates for Bernoulli trials
consistency: known θ, 96 trials, singles, 5000 replications

Consistency

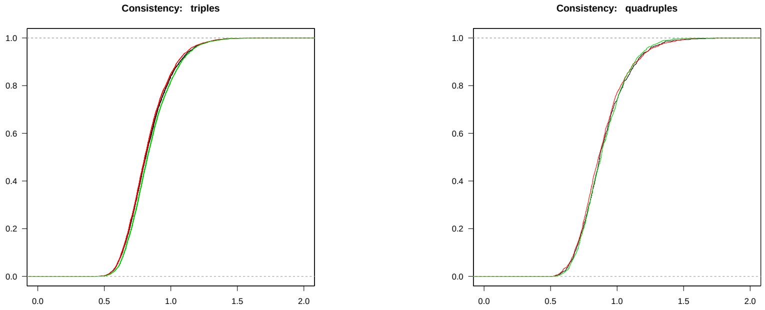

Figure:
consistency: unknown θ, 96 trials, triples, 5000 replications
consistency: unknown θ, 96 trials, quadruples, 1000 replications

Power

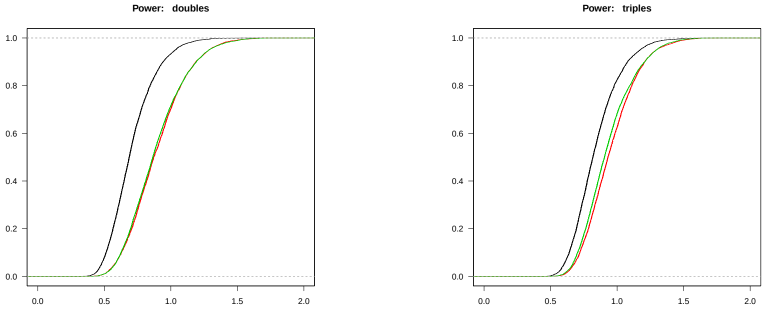

Figure:
Power, unknown θ, 96 trials, doubles, 5000 replications
Power, unknown θ, 96 trials, triples, 5000 replications

Conclusion

advantages of KT2

- intuitive
  - for discrete distns, linear algebra suffices for basic intuition
- easy to code, computing fast

So why has KT2 been so little applied?

Nguyen (2017*a*,*b*) successfully applies KT2 to contingency tables and distributional tails; but she does not have covariates changing over the sample, and nor does she allow for estimating unknown parameters.

For this simple simulation of Bernoulli trials

- We have verified consistent mappings of distinct empirical processes into common base distn
- and shown high power to reject incorrect null hypotheses

- KT = KT1 beautifully elegant way to test gof by distn free methods, but technically difficult and hard to implement
- KT2 reflections or 'rotations' easily implemented
- Testing after 'rotation' of empirical processes to that of a simple process seems effective
  - possible need to artificially group data for rotation
  - but this seems not to make the KT2 'rotation into distn free' methods less effective.
- This paper is apparently the first to apply KT2 when distns depend on exogenous covariates, and the first to investigate empirically the working of the methodology when parameters need to be estimated.

- All sorts of outstanding issues
  - tie in with other mathematical techniques? eg, reflections $\leftrightarrow$ root systems of Lie algebras?
  - mapping between distns with varying number of parameters?
  - Are the score functions of target really so arbitrary?
  - Testing constraints on parameters after rotation?
  - Using other gof tests, eg ω^2, as well as KS statistics after rotation.
  - Large number of parameters?
  - KT2 good for this simple simulation of Bernoulli trials. For more general situations?
  - Clarify minimum group size for given modelling situations and number of parameters. Optimal group size?
    Preferable to have X well dispersed within groups?
- WHAT are we to call KT2?

Khmaladze, E. V. (2013), 'Note on distribution free testing for discrete distributions', *Annals of Statistics* **41**(6), 2979–2993.

Khmaladze, E. V. (2017), 'Distribution free testing for conditional distributions given covariates', *Statistics & Probability Letters* **129**, 348–354.

Nguyen, T. T. M. (2017*a*), Asymptotic methods of testing statistical hypotheses, PhD thesis, School of Mathematics and Statistics, Victoria University, Wellington, New Zealand.

Nguyen, T. T. M. (2017*b*), 'A new approach to distribution free tests in contingency tables', *Metrika* **80**(2), 153–170.

Roberts, L. A. (2016), On distribution free testing in Bernoulli trials with distributions depending on covariates. arXiv:1610.04297v1.

Thank you

and also thanks to Estate Khmaladze

as well as apologies to Charles Dickens