The Role of the Support in Semi-Parametric Hypothesis Testing

Besançon

David Harris and Brendan McCabe

University of Melbourne and University of Liverpool

July 2018

The Model

- Consider the $INAR(1)$ model for count data $\{y_t, t = 1, \ldots, T\}$, which has the form

$$y_t = \beta \circ y_{t-1} + u_t, \tag{1}$$

  where the arrivals process (disturbance) u_t is i.i.d. with some distribution Π on a support $\mathcal{U} \subseteq \mathbb{N}$ ($\mathbb{N}$ is the set of positive integers and zero). Here $\beta\circ$ is the usual binomial thinning operator whereby $\Pr(\beta \circ n = k) = \mathrm{Bi}(n, k; \beta)$ for $k = 0, 1, \ldots, n$ and $0 \leq \beta < 1$.
- This model has a well defined physical interpretation as it may be thought of as a queue, a birth and death process or even a branching process with immigration. Thus, it is a natural contender to model any series of low counts that may be thought of as a "stock" variable.
- We wish to test $\beta = 0$ i.e. y_t is iid against $\beta > 0$ when the support of the disturbances may be $\mathcal{U} \subseteq \mathbb{N}$, the emphasis being on the subset!

The Observation Support

- In many practical situations the support of the observations is restricted and known to be so.
- We could have Censoring e.g. Q: how many times did.... A: $0, 1, 2, 3$ or more
- We could have Truncation. Say, we take a survey on a bus... Q: how many bus journeys....But now we never observe an individual with 0 count. Or say we count patient days spent in hospital $1, 2, 3$ but we don't see patients discharged on the same day.
- We might even have both. We also have ZIPs etc.
- As usual you can't ignore these issues or you will end up with inconsistent estimates. But we can derive the likelihood etc. so estimation and testing possible.

What about the Non Observable Support?

- Typically, one would assume $\mathcal{U} = \mathbb{N}$ and it would generally be considered a pretty innocuous modelling strategy.
- However, when we focus on testing for independence in a series of low counts and there seems to be no compelling reason why $\mathcal{U}$ should be considered devoid of restrictions.
- It may be that u_t is supported on a finite set, having a binomial distribution, say, or even that the arrivals are composed of several unobserved streams whose probability structure may be thought of as a mixture distribution. The point is we don't know.
- If restrictions on $\mathcal{U}$ hold, the consequences of (incorrectly) assuming $\mathcal{U} = \mathbb{N}$ can be severe when testing the hypothesis $\beta = 0$ via likelihood based methods. Standard likelihood theory is invalid as the support of the observations is now parameter dependent.
- Thus we wish to construct an optimal semi-parametric test of independence when $\mathcal{U} = \mathbb{N}$, but one which is robust should the unknown support of the disturbances be finite or contain gaps.

LAN

The *LAN assumption* says that the likelihod ratio may be expressed as

$$L_n(\gamma + n^{-1/2}h, \gamma) = h'S_n - \frac{1}{2}h'Bh + o_p(1) \qquad (2)$$

where $S_n = S_n(x_n, \gamma)$, $B = B(\gamma)$ and $o_p(1)$ is wrt $P_n(\gamma)$. Note B is a pd symmetric matrix that is non random (would exclude unit root models) and does not depend on h. S_n is a random vector which does not depend on h and

$$S_n \rightarrow^{P_n(\gamma)} N(0, B).$$

It says that the likelihood ratio of a "local point" to the fixed point in the direction h can be approximated by the RHS above. The h only appears in the manner indicated and the only random thing is the S_n. Note too that the convergence is done under the measure $P_n(\gamma)$ and not $P_n(\gamma + n^{-1/2}h)$

LAN II

An immediate implication is that

$$L_n(\gamma + n^{-1/2}h, \gamma) \to^{P_n(\gamma)} N\left[-\frac{1}{2}h'Bh, h'Bh\right]$$

in other words the likelihood ratio between the two "points close together" can be approximated by a *random variable* which is normal even though the original likelihood is pretty arbitrary. This means that in the LAN class all problems can be treated as if the likelihood was normal (as an approximation)

We call S_n *scores* and B the *information matrix.* Often, S_n is $n^{-1/2}d\log f_n/d\gamma$ and B the variance of the score but this is not necessary in this generic setup.

LeCam's 3rd Lemma

This lemma says that that.for any other (asy Normal) statistic Y_n

$$\begin{bmatrix} Y_n \\ L_n \end{bmatrix} \rightarrow^{P_n\left(\gamma+n^{-1/2}h\right)} N\left[\begin{pmatrix} \sigma_{YL} \\ \frac{1}{2}h'Bh \end{pmatrix}, \begin{pmatrix} \Sigma_Y & \sigma_{YL} \\ \sigma_{YL} & h'Bh \end{pmatrix}\right]$$

Note the convergence is under the local alternative measure and we get "local alternative" distributions. The point is that knowing the $P_n\left(\gamma\right)$ result automatically delivers the local alternative result via the Lemma. This is a free lunch. The proof is essentially the change of measure theorem beloved by finance people. Note the distribution is normal and the covariance structure is also the same as under $P_n\left(\gamma\right)$. What we really get are the means.

Thus in testing applications, for example, local power follows directly from the null distribution; you just need to work out a covariance under the null. We can get the local power of the score test directly since $S_n'B^{-1}S_n \rightarrow^{P_n\left(\gamma+n^{-1/2}h\right)} \chi^2\left(h'Bh\right)$, a non-central chisquare by the CMT

First consider $\mathcal{U} = \mathbb{N}$

- To construct optimal tests we prove an infinite dimensional *LAN* property for the model from which it follows that the effective score $(C(\alpha))$ is an asymptotically uniformly most powerful test at a given Π
- This infeasible test is given by

$$S^{*}_{T,\beta} = T^{-1/2} \sum_{t=1}^{T} (y_{t-1} - \mu_u) \left(\frac{\pi_{y_t - 1}}{\pi_{y_t}} - 1 \right)$$

and

$$C_T = S^{*}_{T,\beta} / \omega \rightsquigarrow N(\omega h_\beta, 1)$$

under θ_T, a sequence of $T^{-1/2}$ local alternatives and suitable variance ω^2.

Feasible Tests

- To get a feasible test we need to estimate the parameters (apply ML under the null) and use

$$\hat{S}^*_{T,\beta} = T^{-1/2} \sum_{t=1}^{T} (y_{t-1} - \bar{y}) \left(\frac{\hat{\pi}_{y_t - 1}}{\hat{\pi}_{y_t}} - 1 \right),$$

in conjunction with an $\hat{\omega}^2$. This gives

$$\hat{\xi}_T = \hat{S}^*_{T,\beta} / \hat{\omega}$$

It turns out $\hat{\xi}_T$ is adaptive and hence optimal too.

- The classical Wald, W_T, LR, Λ_T, and one-sided score tests, Ψ^+_T, are also efficient with modified distributions to deal with the 0 lower bound for β
- So far so good.

Null behaviour under Gaps in the Support

- The definition of a "gap" in the support for u_t is that there exists at least one integer $k \geq 1$ such that $k-1 \in \mathcal{U}$ but $k \notin \mathcal{U}$. We also assume there exists at least one integer $k \geq 1$ such that $k-1 \in \mathcal{U}$ and $k \in \mathcal{U}$.
- A leading case is $\mathcal{U} = \{0, 1, 2, ..., M\}$. M finite but unknown.

Theorem

Under "gaps" and $\beta = 0$

(i) $\hat{\tilde{\xi}}_T \rightsquigarrow N(0,1)$,

(ii) $\Psi_T \xrightarrow{p} +\infty$, $\Psi_T^+ \xrightarrow{p} -\infty$,

(iii) $\Lambda_T, W_T \xrightarrow{p} 0$.

Only the effective score test has a non degenerate null distribution. The reason for the degenerate behaviour is that the ordinary score does not have zero mean under the null. Thus, the effective score test doesn't need to know about "gaps" but the others do!

How the Support Affects Power

- Say, under some hypothesis observations should always be positive. Then if you ever see a negative observation it's game over regardless of models and parameterisations!
- So, what if the support (under the null hypothesis) is like $\mathcal{U} = \{0, 1, 2, ..., M\}$ with M unknown? Can this generate a power advantage?
- The answer is yes as long as there is a positive probability of seeing observations outside $\mathcal{U}$ under the alternative.
- This is true even though M and the probability are unknown to you!
- The next step is to formalise the probability of seeing observations outside the null support under alternatives.

A Useful Set

- First define $\mathcal{U}^{(0)} = \mathcal{U}$. Now define a set for those integers that differ from $\mathcal{U}^{(0)}$ by an amount 1 i.e. those i that satisfy

$$\mathcal{U}^{(1)} = \left\{ i \notin \mathcal{U}^{(0)} : i - 1 \in \mathcal{U}^{(0)} \right\}$$

  So the members of $\mathcal{U}^{(1)}$ are elements of the gap that are 1 greater than elements of the null support.
- For example, if $\mathcal{U} = \{0, 1, \ldots, M\}$ then $\mathcal{U}^{(1)} = \{M + 1\}$.
- The constants $\pi^{(0)} = \sum_{j \in \mathcal{U}^{(0)}} \pi_{j-1}$ and $\pi^{(1)} = \sum_{j \in \mathcal{U}^{(1)}} \pi_{j-1}$ are of particular relevance to the asymptotic analysis and satisfy $\pi^{(0)} + \pi^{(1)} = 1$. So, for $\mathcal{U} = \{0, 1, \ldots, M\}$, $\pi^{(0)} = 1 - \pi_M$ and $\pi^{(1)} = \pi_M$.

Law of Small Numbers

Define the counting process $N_{T,i} = \sum_{t=1}^{T} 1_i(y_t)$ in T for any i. The following lemma gives a law of small numbers for these processes under local alternatives.

Lemma

Under "gaps" and $\beta_T = T^{-1}h_\beta,\ \ h_\beta > 0$, as $T \to \infty$ then $\{N_{T,i}\}_{i\in\mathcal{U}^{(1)}} \rightsquigarrow \{N_i\}_{i\in\mathcal{U}^{(1)}}$,where $\{N_i\}_{i\in\mathcal{U}^{(1)}}$ is an independent sequence of Poisson random variables with parameters $h_\beta \mu_u \pi_{i-1}$.

Consider, for example, $\mathcal{U}^{(0)} = \{1, 2, ..., M\}$ and let $N_{T,M+1}$ be the number of y_t's in the sample of size T which equal $M+1$. Then, under these local alternatives, $\Pr[N_{T,M+1} = k]$, $k = 0, 1, 2...$ may be computed asymptotically from the Poisson distribution $e^{-\lambda}\lambda^k/k!$ with mean $\lambda = h_\beta \mu_u \pi_M$.

Occupancy of the Null and Alternative Supports

In addition,

$$\Pr\left(y_t \in \mathcal{U}^{(0)} \text{ all } t = 1, \ldots, T\right) \quad \to \quad \exp\left(-h_\beta \mu_u \pi^{(1)}\right)$$
$$\Pr\left(y_t \in \mathcal{U}^{(1)} \text{ any } t = 1, \ldots, T\right) \quad \to \quad 1 - \exp\left(-h_\beta \mu_u \pi^{(1)}\right).$$

These equations basically say that, under T^{-1} local alternatives, asymptotically either the sample $\{y_1, \ldots, y_T\}$ is restricted to $\mathcal{U}^{(0)}$ or the sample contains at least one element of $\mathcal{U}^{(1)}$. No other outcome is possible. The probabilities of these outcomes are given.

Local Power

Theorem

Under "gaps" and $\beta_T = T^{-1} h_{\beta}$,, as $T \to \infty$

$$\hat{\xi}_T = T^{-1/2} \sum_{t=1}^{T} (y_{t-1} - \bar{y}) (\hat{g}_t - 1) / \hat{\omega} \rightsquigarrow \begin{cases} Z, & \text{if all obs in } \mathcal{U}^{(0)} \\ X, & \text{if some obs in } \mathcal{U}^{(1)} \end{cases}$$

where $Z \sim N(0,1)$ and X is a complicated distribution depending on the π_i.

The theorem shows that if the sample is asymptotically restricted to $\mathcal{U}^{(0)}$ which occurs with probability $\exp\left(-h_{\beta}\mu_u \pi^{(1)}\right)$, then $\hat{\xi}_T \rightsquigarrow Z$ and the effective score test has asymptotic power equal to size. If some observations occur in $\mathcal{U}^{(1)}$ asymptotically, the asymptotic distribution of $\hat{\xi}_T$ is non-standard.

Rates of Convergence

But the remarkable thing is that the local alternatives converge at a rate T^{-1} while the statistic uses $T^{-1/2}$ as a scaling. It can be shown that $\hat{\omega}$ (which depends on $N_{T,i}$) is of order $T^{1/2}$ under regime $\mathcal{U}^{(1)}$ and converges under $\mathcal{U}^{(0)}$. So

$$T^{-1/2}\sum_{t=1}^{T}\left(y_{t-1}-\bar{y}\right)\left(\hat{g}_t-1\right)/\hat{\omega} \approx$$

$$T^{-1/2}\sum_{t=1}^{T}\left(y_{t-1}-\mu_u\right)\sum_{i\in\mathcal{U}^{(0)}}1_i\left(y_t\right)\left(\frac{\hat{\pi}_{i-1}}{\hat{\pi}_i}-1\right)/\hat{\omega}$$

$$+T^{1/2}\sum_{i\in\mathcal{U}^{(1)}:N_{T,i}>0}\sum_{t=1}^{T}\left(y_{t-1}-\mu_u\right)1_i\left(y_t\right)\frac{\pi_{i-1}}{N_{T,i}}/\hat{\omega}.$$

converges as usual to a normal under $\mathcal{U}^{(0)}$ since the second term is not present. But under $\mathcal{U}^{(1)}$ the convergence rate of $\hat{\omega}$ forces the first term to zero while the second self normalises and converges to a rv. All of this convergence leverages off the T^{-1} local rate of the $N_{T,i}$.

A Hybrid

- While it is true that the NP test is asymptotically optimal, this gives no guarantees about finite sample behaviour.
- In fact, if the counts start to get even slightly high (15 say) a lot of parameters (probabilities) have to be estimated which can impact on the power of the NP test.
- On the other hand, the simple lag 1 correlation coefficient (optimal with u_t Poisson) has no such problems.
- This suggests a hybrid test which combines both might have merit.

The general form of the combined test, for some critical value c_α, is

$$\text{reject } H_0 \text{ if } \hat{\xi}_T > c_\alpha \text{ and/or } \hat{\rho}_T > c_\alpha,$$

Under H_0 and $\mathcal{U} \subseteq \mathbb{N}$ these statistics can easily be seen to be jointly asymptotically distributed as

$$\begin{pmatrix} \hat{\xi}_T \\ \hat{\rho}_T \end{pmatrix} \rightsquigarrow \begin{pmatrix} Z_\xi \\ Z_\zeta \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \lambda \\ \lambda & 1 \end{pmatrix} \right),$$

where in practice λ can be consistently estimated by

$$\hat{\lambda} = \frac{T^{-1} \sum_{t=1}^{T} \left(\hat{g}_t - \bar{g} \right) \left(y_t - \bar{y} \right)}{\hat{\omega}},$$

and the joint critical value obtained by solving

$$1 - \Phi_2 \left(\hat{c}_\alpha, \hat{c}_\alpha; \hat{\lambda} \right) = \alpha,$$

for $\hat{c}_\alpha$. Since $\hat{g}_t$ and y_t are generally found to be positively correlated, this procedure of estimating λ and the subsequent critical value will provide some power gain relative to a standard Bonferonni inequality.

A Simulation

Table: Simulated size and power, Equal mixture of Poisson(1) and Negative Binomial (10,0.75) arrivals

$T = 100$				$T = 800$			
β	$\hat{\tilde{\xi}}_T$	$\hat{\rho}_T$	Hybrid	β	$\hat{\tilde{\xi}}_T$	$\hat{\rho}_T$	Hybrid
0.000	0.059	0.047	0.060	0.000	0.051	0.059	0.059
0.080	0.230	0.160	0.236	0.015	0.185	0.118	0.169
0.160	0.445	0.406	0.531	0.030	0.410	0.208	0.383
0.240	0.575	0.704	0.792	0.045	0.664	0.340	0.626
0.320	**0.652**	**0.918**	**0.936**	**0.060**	**0.846**	**0.499**	**0.820**
0.400	0.692	0.986	0.989	0.075	0.947	0.670	0.931
0.480	0.680	0.998	0.999	0.090	0.983	0.806	0.978
0.560	0.668	1.000	1.000	0.105	0.996	0.898	0.996
0.640	0.643	1.000	1.000	0.120	0.999	0.954	1.000
0.720	0.559	1.000	1.000	0.135	1.000	0.985	1.000

Concluding Remarks

- Extensions to the multivariate case are more complicated but the message is clear. Restrictions on the support of the disturbances matter.
- Using robust methods in an uncertain envoronment makes sense.
- Allowing for (even unknown) support restrictions can make a big difference to power
- There is a paper with references etc
- Special issue on Count Data in the journal *Econometrics*, Editor-in-Chief Prof. Marc S.Paolella.
- http://www.mdpi.com/journal/econometrics/editors
- Brendan.McCabe@Liverpool.ac.uk