Test Statistics for Participation in Assemblies

Christian Borgelt, European Centre for Soft Computing

Background Rate Estimation (BRE)

The probability that an item i occurs in a given time bin may be decomposed into two constituents: the background occurrence probability, denoted by $\theta_i = \lambda_{i,b} \cdot w$, and the coincidence probability, denoted by $\xi_i = \lambda_{i,c} \cdot w$, which captures the joint influence of *all* assemblies item i participates in. Here $\lambda_{i,b}$ and $\lambda_{i,c}$ are the background and coincidence occurrence rates of item i, respectively, and w is the length of a time bin. Since we assume that the hidden processes that generate the background and the coincident item occurrences are independent, the effective occurrence probability of item i is $\eta_i = \theta_i + \xi_i - \theta_i\xi_i$.

We want to determine whether item i participates in an assembly or not, that is, whether $\xi_i > 0$ (alternative hypothesis) or $\xi_i = 0$ (null hypothesis). Due to the above equation we have $\xi_i = \frac{\eta_i - \theta_i}{1-\theta_i}$. Note that η_i can easily be estimated from the data, namely as $\hat{\eta}_i = T_i/T$, where T_i is the number of time bins in which item i occurs and T is the total number of time bins. As a consequence we can derive a statistical test if we can estimate the background probability θ_i.

To derive an estimator for θ_i, we consider a set of N independent items without assemblies (i.e. $\forall j; 1 \leq j \leq N : \xi_j = 0$). The probability p_0 that *no* item occurs in a time bin is $p_0 = \prod_{j=1}^{N}(1-\theta_j)$ and the probability p_{i0} that *only* item i occurs is $p_{i0} = \theta_i \prod_{j=1,j\neq i}^{N}(1-\theta_j)$. It follows $\frac{p_{i0}}{p_0} = \frac{\theta_i}{1-\theta_i}$ and therefore $\theta_i = \frac{p_{i0}}{p_{i0}+p_0}$. Obviously, p_0 and p_{i0} can be estimated easily, namely as $\hat{p}_0 = T_0/T$ and $\hat{p}_i = T_{i0}/T$, respectively, where T_0 is the number of time bins in which *no* item occurs and T_{i0} the number of bins in which *only* item i occurs.

The crucial insight is now that the probabilities p_0 and p_{i0} remain unaffected if items participate in assemblies, because both refer to time bins with *at most one* event. However, coincident item occurrences, by their very definition, mean that *more than one* item occurs in the same time bin. As a consequence, θ_i can be estimated, even in the presence of assemblies, as $\hat{\theta}_i = T_{i0}/(T_{i0} + T_0)$. The actual test whether item i participates in an assembly checks whether $\hat{\xi}_i$ is sufficiently large so that the null hypothesis $\xi_i = 0$ can be rejected. A natural test statistic, estimating the fraction of coincidence events, is

$$t^{\mathrm{BRE}}(i) = \frac{\hat{\xi}_i}{\hat{\eta}_i} = \frac{\hat{\eta}_i - \hat{\theta}_i}{\hat{\eta}_i(1-\hat{\theta}_i)}.$$

As a generalization of this approach one may consider to estimate the background occurrence rate of item i not only from the time bins in which at most item i occurs, but also from those time bins, in which a maximum of r, $r \geq 0$, other items occur ($r = 0$ yields the case discussed above). This provides an indication whether item i participates in assemblies with more than $r+1$ items. However, one should be aware that for $r > 0$ a possible participation in assemblies of smaller size (at most $r+1$ items) can obscure the participation in larger assemblies (more than $r+1$ items), because in this case we are not estimating the true background occurrence probability, but the probability resulting from background occurrences and participation in small assemblies.

Conditional Pattern Cardinality/Complexity (CPC)

A plausible approach to identify items participating in assemblies is based on the idea that such items should have, on average, more items occurring together with them in the original data than in the surrogates. In other words, if some item i participates in one or more (large) assemblies, there should be several time bins in which it occurs together with a larger number of other items. Hence the average complexity (cardinality/size) of patterns (time bin contents) involving item i should be larger than can be expected by chance. Formally, we use

$$\bar{\mu}(i) = \frac{1}{T} \sum_{l=1}^{T} |I_l - \{i\}| \qquad \text{and} \qquad \mu(i) = \frac{1}{T_i} \sum_{l=1}^{T} \mathbf{1}_{I_l}(i)\ |I_l - \{i\}|,$$

where I_l is the set of items that occur in the lth time bin, $\mathbf{1}_{I_l}(i)$ is the indicator function of the set I_l (which is 1 if $i \in I_l$ and 0 otherwise), T is the total number of time bins, and $T_i = \sum_{l=1}^{T} \mathbf{1}_{I_l}(i)$ is the number of time bins in which item i occurs. Thus, $\bar{\mu}_i$ is simply the overall average pattern cardinality (with events/occurrences of item i removed), while μ_i is the average pattern cardinality/complexity in time bins in which item i occurs (again with events of item i removed), which we may also call the conditional average pattern cardinality/complexity (conditional on events of item i). A natural test statistic is

$$t^{\mathrm{CPC}}(i) = \frac{\mu(i) - \bar{\mu}(i)}{\bar{\mu}(i)}.$$

An obvious way to improve this statistic is to weight large cardinalities more strongly than smaller ones, because large cardinalities are, intuitively, more indicative of assembly activity. A simple technical means to achieve such weighting is to raise the cardinalities to a user-specified power α:

$$\bar{\mu}_\alpha(i) = \frac{1}{T} \sum_{l=1}^{T} |I_l - \{i\}|^\alpha \qquad \text{and} \qquad \mu_\alpha(i) = \frac{1}{T_i} \sum_{l=1}^{T} \mathbf{1}_{I_l}(i)\ |I_l - \{i\}|^\alpha.$$

In other words, instead of a simple mean of the pattern cardinalities, we employ higher moments. The resulting test statistic is

$$t_\alpha^{\mathrm{CPC}}(i) = \frac{\mu_\alpha(i) - \bar{\mu}_\alpha(i)}{\bar{\mu}_\alpha(i)}.$$

Conditional Excess Cardinality/Complexity (CXC)

The test statistic t^{CPC} considers all cardinalities, and only the extended form places higher emphasis on larger cardinalities, since these tell us about possibly existing correlations. This emphasis may be increased by considering only those cardinalities that exceed the average pattern cardinalitiy $\bar{\mu}(i)$ (as defined above), or formally (with the user-specified power α already added):

$$t_\alpha^{\mathrm{CXC}}(i) = \sum_{l=1}^{T} \mathbf{1}_{I_l}(i)\ \zeta(|I_l - \{i\}| > \bar{\mu}(i))\ (|I_l - \{i\}| - \bar{\mu}(i))^\alpha,$$

where $\zeta(\varphi)$ is 1 if φ is true and 0 otherwise. Note that in this case the weighting by the user-specified power α is "shifted" in its influence, because it acts on the difference of the pattern cardinality to the average pattern cardinality and not on the pattern cardinality directly (as it is the case for t^{CPC}).

Conditional Cardinality/Complexity Frequency (CCF)

An alternative way to exploit the influence that correlations have on the pattern cardinalities is to consider the frequencies of the different possible pattern cardinalities. If large pattern cardinalities occur more frequently in those time bins, in which the considered item i occurs, than in all time bins, then item i is likely involved in an assembly. To capture this idea formally, let

$$\bar{\phi}_{ix} = \sum_{l=1}^{T} \zeta(|I_l - \{i\}| = x) \qquad \text{and} \qquad \phi_{ix} = \sum_{l=1}^{T} \mathbf{1}_{I_l}(i)\, \zeta(|I_l - \{i\}| = x)$$

be the overall frequency of the pattern cardinality x and its frequency in those time bins in which item i occurs, respectively (in both cases with events of item i removed). Furthermore, let $\hat{\eta}_i = T_i/T$ be the fraction of time bins in which item i occurs (denoted as $\hat{\eta}_i$, because it is an estimate of the occurrence rate η_i of item i). Thus, $\bar{\phi}_{ix}\hat{\eta}_i$ is the expected frequency of pattern cardinality x in those cases in which item i fires, while ϕ_{ix} is its actual frequency. With these definitions we can define the test statistic

$$t_\alpha^{\mathrm{CCF}}(i) = \sum_{x=\lceil \bar{\mu}(i) \rceil}^{N-1} \zeta(\phi_{ix} > \bar{\phi}_{ix}\hat{\eta}_i)\, (\phi_{ix} - \bar{\phi}_{ix}\hat{\eta}_i)\, (x - \bar{\mu}(i))^\alpha,$$

where $\bar{\mu}(i)$ is the average pattern cardinality (with events of item i removed, see above). That is, we sum, for conditional pattern cardinalities that meet or exceed the average pattern cardinality, the excess over their expected frequency. Since excess frequencies for larger complexities provide stronger evidence, the excess frequencies are weighted with the pattern cardinality (factor $x - \bar{\mu}(i)$), and since larger pattern cardinalities are more strongly indicative of assembly membership, these are additionally weighted with a user-specified power α.

Conditional Cardinality Frequency Ratio (CCR)

In the preceding statistic (t^{CCF}), the excess frequencies of larger pattern cardinalities were weighted higher by multiplying them with the pattern cardinality itself (factor $x - \bar{\mu}(i)$). In addition, we may exploit that usually smaller pattern cardinalities are more frequent than larger ones. Consequently, we can achieve a weighting by forming the ratio to the expected frequency:

$$t_\alpha^{\mathrm{CCR}}(i) = \sum_{x=\lceil \bar{\mu}(i) \rceil}^{N-1} \zeta(\phi_{ix} > \bar{\phi}_{ix}\hat{\eta}_i) \left(\frac{\phi_{ix} - \bar{\phi}_{ix}\hat{\eta}_i}{\bar{\phi}_{ix}\hat{\eta}_i + 1} \right) (x - \bar{\mu}(i))^\alpha.$$

The +1 in the denominator helps handling vanishing expected frequencies.

Conditional Fano Factor (CFF)

The Fano factor—also known as the dispersion index, the coefficient of dispersion or the variance-to-mean ratio—is a measure of the dispersion of a probability distribution. It is simply defined as the variance of a probability distribution divided by its mean value. In order to apply the Fano factor to the pattern cardinality distribution, we define (in addition to the mean values $\bar{\mu}(i)$ and $\mu(i)$

as defined above in the context of the conditional pattern cardinality statistic) the overall and the conditional variance (that is, for those time bins, in which item i occurs), respectively, of the pattern cardinality as

$$\bar{\sigma}^2(i) = \frac{1}{T-1} \sum_{l=1}^{T} (|I_l - \{i\}| - \bar{\mu}(i))^2$$

and

$$\sigma^2(i) = \frac{1}{T_i - 1} \sum_{l=1}^{T} \mathbf{1}_{I_l}(i) \; (|I_l - \{i\}| - \mu(i))^2.$$

With these variances we can define the overall and the conditional (on events of item i) Fano factor for the pattern cardinality as

$$\bar{F}(i) = \frac{\bar{\sigma}^2(i)}{\bar{\mu}(i)} \qquad \text{and} \qquad F(i) = \frac{\sigma^2(i)}{\mu(i)}.$$

A natural test statistic based on these quantities is

$$t^{\mathrm{CFF}}(i) = \frac{F(i) - \bar{F}(i)}{\bar{F}(i)}.$$

A user-specified power can be introduced in a similar fashion as for the conditional pattern cardinality, namely by using higher moments (about the mean),

$$\bar{m}_\alpha(i) = \frac{1}{T-1} \sum_{l=1}^{T} (|I_l - \{i\}| - \bar{\mu}(i))^{\alpha+1}$$

and

$$m_\alpha^2(i) = \frac{1}{T_i - 1} \sum_{l=1}^{T} \mathbf{1}_{I_l}(i) \; (|I_l - \{i\}| - \mu(i))^{\alpha+1},$$

and using these quantities to define the generalized Fano factors

$$\bar{F}_\alpha(i) = \frac{\bar{m}_\alpha(i)}{\bar{\mu}(i)} \qquad \text{and} \qquad F_\alpha(i) = \frac{m_\alpha(i)}{\mu(i)},$$

thus arriving at the test statistic

$$t_\alpha^{\mathrm{CFF}}(i) = \frac{F_\alpha(i) - \bar{F}_\alpha(i)}{\bar{F}_\alpha(i)}.$$

Conditional Item Frequency (CIF)

In a second line of approaches we take into account how often other individual items occur together with item i. The idea is that if item i participates in one or more assemblies, it should occur more often together with certain other items (namely those also in the assemblies) than can be expected by chance. In order to be less sensitive to differing occurrence rates, we use the excess occurrence rates to form a test statistic: we compute for each item j, $j \neq i$, the difference between the conditional occurrence rate T_{ij}/T_i and the expected (or global) rate of such events, estimated as T_j/T, where $T_{ij} = \sum_{l=1}^{T} \mathbf{1}_{I_l}(i)\mathbf{1}_{I_l}(j)$ is the number

of joint occurrences of the items i and j, $T_i = \sum_{l=1}^{T} \mathbf{1}_{I_l}(i)$ and $T_j = \sum_{l=1}^{T} \mathbf{1}_{I_l}(j)$ are the numbers of occurrences of items i and j, respectively, and T is the total number of time bins. Since only excess rates tell us about possible correlations, negative differences are ignored. Formally, the test statistic is

$$t^{\mathrm{CIF}}(i) = \frac{1}{N-1} \sum_{j=1, j \neq i}^{N} \zeta(T_{ij}/T_i > T_j/T)\ (T_{ij}/T_i - T_j/T),$$

where N is the total number of items and, as above, $\zeta(\varphi)$ is 1 if φ is true and 0 otherwise. In analogy to previous modifications we may consider weighting a large excess rate more strongly than a small excess rate, as a large excess rate is certainly more indicative of assembly activity. In order to achieve this, we once again introduce a user-specified power α to which the excess rate is raised:

$$t_\alpha^{\mathrm{CIF}}(i) = \frac{1}{N-1} \sum_{j=1, j \neq i}^{N} \zeta(T_{ij}/T_i > T_j/T)\ (T_{ij}/T_i - T_j/T)^\alpha.$$

Conditional Item Frequency Ratio (CIR)

As a straightforward variant of the previous statistic, one may relate the excess occurrence rate to the expected (or global) ocurrence rate, to achieve a normalization of the terms of the statistic. In this case, the statistic reads (with the user-specified power α already added):

$$t_\alpha^{\mathrm{CIR}}(i) = \frac{1}{N-1} \sum_{j=1, j \neq i}^{N} \zeta(T_{ij}/T_i > T_j/T) \left(\frac{T_{ij}/T_i - T_j/T}{T_j/T} \right)^\alpha.$$

Conditional Item Weight (CIW)

For the test statistic of conditional item frequencies it is only considered whether another item j occurs together with the considered item i. The cardinality of the pattern in which this coincidence occurs is neglected. However, it is plausible that coincident events of two items i and j in a pattern of high cardinality are more indicative of possible correlations than coincident events in a pattern of low cardinality. In particular, patterns that contain only the events of the two items i and j, but no other events, do not tell us much. This idea can be exploited by not simply counting the number of co-occurrences of items, but to weight these co-occurrences with the cardinality of the containing pattern. Formally, this idea can be captured as follows: let

$$w_j^{(i)} = \sum_{l=1}^{T} \mathbf{1}_{I_l}(j)\ |I_l - \{i\}| \qquad \text{and} \qquad w_{ij}^{(i)} = \sum_{l=1}^{T} \mathbf{1}_{I_l}(i)\ \mathbf{1}_{I_l}(j)\ |I_l - \{i\}|$$

be the overall weight of events of item j and the weight of joint events of item i and item j, respectively (in both cases events of item i are removed). Then we define the test statistic (with a user-specified power α already added)

$$t_\alpha^{\mathrm{CIW}}(i) = \frac{1}{N-1} \sum_{j=1, j \neq i}^{N} \zeta\big(w_{ij}^{(i)}/T_i > w_j^{(i)}/T\big)\ \big(w_{ij}^{(i)}/T_i - w_j^{(i)}/T\big)^\alpha.$$

Conditional Item Weight Ratio (CWR)

In analogy to the conditional item frequency ratio (t^{CIR}, which is a normalized form of t^{CIF}), we may define a conditional item weight ratio as

$$t_\alpha^{\mathrm{CIW}}(i) = \frac{1}{N-1} \sum_{j=1, j\neq i}^{N} \zeta\big(w_{ij}^{(i)}/T_i > w_j^{(i)}/T\big) \left(\frac{w_{ij}^{(i)}/T_i - w_j^{(i)}/T}{w_j^{(i)}/T} \right)^\alpha .$$

Conditional Pattern Overlap (CPO)

All preceding statistics consider only individual time bins in order to compute a test statistic. However, in order to find higher order correlations, *pairs* of time bins provide much better information. In particular, if we can find many pairs of time bins in which a considered item i occurs together with the same *set* of other items, this strongly suggests that item i is involved in an assembly. In order to capture this idea formally, we define the test statistic

$$t^{\mathrm{CPO}}(i) = \sum_{a=2}^{T} \sum_{b=1}^{a-1} \mathbf{1}_{I_a \cap I_b}(i)\, \zeta(|I_a \cap I_b - \{i\}| > 1)\, |I_a \cap I_b - \{i\}|,$$

where the factor $\zeta(|I_a \cap I_b - \{i\}| > 1)$ excludes patterns that overlap only in one other item. Such overlaps are fairly likely to happen by chance and thus would deteriorate the sensitivity of the statistic. In addition, we may introduce a user-specified power α, so that large overlaps, which are clearly more indicative of assembly activity, are we weighted more strongly. This leads to

$$t_\alpha^{\mathrm{CPO}}(i) = \sum_{a=2}^{T} \sum_{b=1}^{a-1} \mathbf{1}_{I_a \cap I_b}(i)\, \zeta(|I_a \cap I_b - \{i\}| > 1)\, |I_a \cap I_b - \{i\}|^\alpha.$$

It should be noted that, due to the double sum, this statistic is computationally more demanding than all other statistics: it is quadratic in the number of events of the considered item i, while all other statistics are linear in this number.