TECHNICAL SUMMARY OF MGPS

For an arbitrary itemset, it is desired to estimate the expectation $\lambda = E[N/E]$, where N is the observed frequency of the itemset, and E is a baseline (null hypothesis) count; e.g., a count predicted from the assumption that items are independent. An itemset is defined by its members *i, j, k, ...*, which occur as subscripts to *N, E,* and other variables, so that, for example, *Nij* is the number of reports involving both items *i* and *j*, *Eijk* is the baseline prediction for the number of reports including the itemset triple *(i, j, k)*, etc.

A common model for computing baseline counts is the assumption of within-stratum independence; when E is computed under this assumption we shall often denote it by $E0$. Assume that all reports are assigned to strata denoted by s = 1, 2, ..., S. Let:

P_{is} = proportion of reports in stratum s that contain item i

n_s = total number of reports in stratum s

Baseline frequencies for pairs and triples are defined under independence as:

$$E0_{ij} = \Sigma_s\, n_s\, P_{is}\, P_{js} \qquad E0_{ijk} = \Sigma_s\, n_s\, P_{is}\, P_{js}\, P_{ks}$$

For itemsets of size 3 or more, an "all-2-factor" loglinear model can be defined as the frequencies E2 for the itemsets that match all the estimated pairwise two-way marginal frequencies but contain no higher-order dependencies. For triples, E2ijk agree with the estimates for the three pairs:

$$\lambda_{ij}E0_{ij} \qquad \lambda_{ik}E0_{ik} \qquad \lambda_{jk}E0_{jk}$$

For "four-tuples," $E2_{ijkl}$ agrees with 6 such pairs, etc.

Analysis of higher-dimensional associations

WebVDME version 5 introduced new definitions of baseline frequencies for more than two dimensions and new definitions of some other output columns. This change allowed the analysis to focus more naturally on combinations in which, for example, there is a strong association between the occurrence of a pair of drugs in a report and some event, irrespective of whether the two drugs are taken together more frequently than chance or not. The change also simplified the analysis and, in particular, no longer required the use of the all-two-factor log-linear model, which did not seem to contribute much, if at all, to substantive understanding in analyses of spontaneous adverse event databases.

For reference, a separate section at the end of this document describes the analysis of higher-dimensional associations made in WebVDME version 4 and earlier.

Introduction

WebVDME allows all items to be partitioned into item types. The prototypical division is into two types, drugs and events, but in general there can be one or more types for a particular data mining run. We conceptually divide the counting and estimation that MGPS does into separate itemset types. That is, if there are two types of items, labeled D and E, we similarly define itemset types such as DD, DE, DDE, DDEE and so forth. This last four dimensional example (DDEE) will be used here to make it plain how the calculations go in general and the corresponding definitions for the other itemset types should be obvious. The individual items will be named D1, D2, E1, E2. The calculation of counts, baseline frequencies, and hyperparameter estimates is separate for every different itemset type. The calculations for a given dimension will sometimes depend on the results from previous calculations for a lower dimension.

Review of Two-Dimensional Calculations

For every pair of items, say j and k, P_{js} and P_{ks} are the respective proportions of the n_s reports in stratum s that contain items j and k. There are n reports total and N_{jk} of the reports contain both items j and k. The 2-dimensional baseline frequencies are:

$$E_{jk} = \Sigma_s\, n_s\, P_{js}\, P_{ks}$$

The values of $RR_{jk} = N_{jk}/E_{jk}$ are smoothed using the MGPS model (separate estimates for each type of pair) to get corresponding values of $EBGM_{jk}$, $EB05_{jk}$ and $EB95_{jk}$.

Handling Homogeneous Itemset Types in Three or More Dimensions

In the descriptions that follow, it will be assumed that the itemset has at least two different item types. Homogeneous itemset types such as DDD or EEEE require the following special convention: *Use the methods described below as if every item being considered were of a different type.* That is, treat homogeneous itemsets like (D1, D2, D3) as if they were (D1, E1, F1), one item from each of three item types, rather than three items of the same type. This is just for the purpose of following the formulas below. We do not actually pool the data of type DDD with other combinations.

Handling Heterogeneous Itemset Types in Three or More Dimensions

Notation for the four dimensional example

As mentioned above, we will assume for presentation purposes that we are calculating the MGPS model for the itemset type of form DDEE. As a shorthand, we use the abbreviations D1 = 1, D2 = 2, E1 = 3, E2 = 4, especially when using subscripts. Thus, for example, the number of reports overall that contain both E1 and E2 will be denoted N_{34}, and the number of reports that contain the triple (D1, D2, E1) is denoted N_{123}. The proportion of reports in stratum s that contain item E2 is P_{4s}, and so forth.

Defining baseline frequencies

Let E0 be the expected frequency when all four items are independent, namely:

$$E0 = \Sigma_s\, n_s\, P_{1s}\, P_{2s}\, P_{3s}\, P_{4s}$$

However, E0 is modified by multiplying by terms corresponding to each item type with more than one item, in this case both the D and E items.

$$E = E0 \times [N_{12}/M_{12}] \times [N_{34}/M_{34}] \qquad (1)$$

Where $M_{12} = \Sigma_s\, n_s\, P_{1s}\, P_{2s}$ and $M_{34} = \Sigma_s\, n_s\, P_{3s}\, P_{4s}$

E represents the baseline frequency under the assumption that each of the item types is independent from other item types, but within an item type complete dependence is assumed. It is as if the pair (D1, D2) was thought of as a compound drug whose count in

the database is N_{12}, and that is why E1 is multiplied by the correction factor N_{12}/M_{12}. An analogous argument applies to the pair (E1, E2), which could be thought of as a syndrome treated as a single event.

If the itemset being considered has no duplicate item types, then E = E0, as would happen if the original itemset was completely homogeneous and we were following the prescription in the previous section to treat such itemsets as if every item were a different type.

Technically, $N = N_{1234}$ and $E = E_{1234}$, but we will just define N, E RR = N/E, EBGM, etc. for the 4-tuple without subscripts.

Defining the interaction signal score

Let EB95max be the highest two-factor EB95 for pairs that are NOT of the same type. In this four-dimensional example, there are six possible pairs, but only four of the pairs are of heterogeneous type, and EBmax is the largest of those. That is,

$$EB95max = \max\{EB95_{13}, EB95_{23}, EB95_{14}, EB95_{24}\}$$

The EB95 values above would come from the previously performed pair-wise analyses of DE itemset types. The value EB95max represents the largest estimated upper 95% confidence limit association found among the heterogeneous PAIRS of items being considered. Remember that if the itemset was originally all of one type, then we pretend they are all different and so EB95max would be the largest of all the included 2 dimensional EB95s.

We declare an "interaction alert" if the smoothed value of RR = N/E for the sextuple, is significantly greater than EB95max. Namely,

INTSS (interaction signal score) = EB05/EB95max

INTSS > 1 is the threshold for an alert for a 3D or higher-way association that cannot be explained by any single pairwise association. When INTSS > 1, the confidence interval around EBGM for the 4-tuple does not overlap any of the confidence intervals for the heterogeneous pairwise EBGMs that are part of the 4-tuple.

How the EB model is set up

Based on the two-factor analyses, we do not expect the baseline expected counts E to fit the observed counts N well for 3D or higher dimensions if many pairwise EBGMs are large, since the baseline E is the prediction one might make if it were known that all pairwise EBGMs = 1. Therefore, we prefer to shrink N towards the product E × EBmax rather than just toward E, where EBmax is the largest of the pairwise EBGMs for included heterogeneous pairs of items. Analogous to the definition of EB95max, we define EBmax (for our D1-D2-E1-E2 example) as

$$EBmax = \max\{EBGM_{13}, EBGM_{23}, EBGM_{14}, EBGM_{24}\}$$

Note that the pair of items that define EBmax might not be the same pair that define EB95max as defined above. The shrinkage model is set up as follows:

Let E* = E × EBmax

Note that E* is the product of terms for absolute independence (E0) times terms for within item type dependence, as in (1), times a measure of heterogeneous item type dependence, EBmax.

Now use the pairs (N, E*) to estimate hyperparameters and get smoothed values of N/E*, which would be summarized by the estimates and confidence limits EBGM*, EB05* and EB95*. These computations are performed just as in the previous section with the substitution of E* for E. Finally, we convert back to smoothed estimates of N/E by defining:

EBGM = EBGM* × EBmax;

EB05 = EB05* × EBmax;

EB95 = EB95* × EBmax

Thus EBGM as a measure of deviation from independence of item types, which is more directly comparable to EBGMs of lower dimensional heterogeneous item types.

The value of INTSS = EB05/EB95max is a multiplicative relative measure of how much excess association is present in the four-tuple that cannot be explained by any single DE association. INTSS > 1 will identify those 3D and higher-way combinations in which there are large cross-item associations that cannot be explained by any single 2D cross-item association. Because INTSS is defined as the ratio of a lower confidence limit divided by an upper confidence limit it will tend to be conservative alerting threshold when the relevant counts are small and the corresponding confidence intervals are wide.

The value INTSS is a relative measure and does not provide information as to the absolute number of reports in the database that are in excess of the number that might be explained by a single cross-item association. The quantity Excess2 focuses on the absolute number of such reports, defined as

$$\begin{aligned} Excess2 &= E \times EB95max \times (INTSS - 1) \\ &= E \times (EB05 - EB95max) \end{aligned}$$

It is a conservative estimate of the number of reports of that itemset in the database that cannot be explained by any single cross-type association.

Correspondence between notation here and variable labels in WebVDME

Mathematical Notation	WebVDME and Help
E0	E_IND
λ	True Relative Ratio
E	E
E2	E2_IND
Excess02	EXCESS2_IND
Excess2	EXCESS2
EB95max	EB05/INTSS
EBmax	EBMAX
Pair where EBGM = EBmax	MAXITEM1, MAXITEM2
E*	E * EBMAX

Analysis of higher-dimensional associations in earlier WebVDME versions

WebVDME version 4 and earlier analyzed associations among itemsets of size 3 or more by comparing the estimated frequency to the all-2-factor prediction by simple subtraction. For example, in case of triples:

$$Excess02_{ijk} = \lambda_{ijk} E0_{ijk} - E2_{ijk}$$

The parameters λ above are estimated by their geometric means, denoted EBGM, of their empirical Bayes posterior distributions, using E = E0 in the formulas below.

For simplicity, the formulas below use just two subscripts, for itemsets of size 2, such as the occurrence of drug i and symptom j in a medical report. Estimates for other itemset sizes are computed analogously. Let:

N_{ij} = the observed counts

E_{ij} = the expected (baseline) counts

$RR_{ij} = N_{ij} / E_{ij}$ = ratio of observed to baseline

We wish to estimate $\lambda_{ij} = \mu_{ij} / E_{ij}$, where $N_{ij} \sim \text{Poisson}(\mu_{ij})$. Assume a superpopulation model for λ_{ij} (prior distribution) based on a mixture of two gamma distributions (a convenient 5-parameter family of distributions that can fit almost any empirical distribution):

$$\pi(\lambda; \alpha_1, \beta_1, \alpha_2, \beta_2, P) = P\, g(\lambda; \alpha_1, \beta_1) + (1 - P)\, g(\lambda; \alpha_2, \beta_2)$$

$$g(\lambda; \alpha, \beta) = \beta^{\alpha} \lambda^{\alpha-1} e^{-\beta\lambda} / \Gamma(\alpha)$$

Estimate the prior distribution from all the *(Nij, Eij)* pairs.

Estimate the 5 hyperparameters:

$$\theta = (\alpha_1, \beta_1, \alpha_2, \beta_2, P)$$

by maximizing the likelihood function L(θ) in 5 dimensions:

$$L(\theta) = \Pi_{i,j}\{P\, f(N_{ij}; \alpha_1, \beta_1, E_{ij}) + (1 - P)\, f(N_{ij}; \alpha_2, \beta_2, E_{ij})\}$$

$$f(n; \alpha, \beta, E) = (1 + \beta/E)^{-n}(1 + E/\beta)^{-\alpha}\, \Gamma(\alpha + n) / \Gamma(\alpha)\, n!$$

In WebVDME, MGPS requires the specification of a threshold (n^* = minimum count > 1) for the observed counts of all combinations that are analyzed. In accord with this specification, the formula for $f(n; \alpha, \beta, E)$ is modified to incorporate the condition $N_{ij} \geq n^*$:

$$\text{If } n^* = 1,\ f(n; \alpha, \beta, E, n^*=1) = f(n; \alpha, \beta, E) / [1 - (1 + E/\beta)^{-\alpha}]$$

For other n^*, the denominator above is $[1 - \Sigma f(n'; \alpha, \beta, E)]$, where the sum extends over $n' = 0, 1, \ldots, n^*-1$. For simplicity, the formulas below omit the reference to n^*.

Given θ, the posterior distributions of each λ_{ij} are also a mixture of gamma distributions used to create "shrinkage" estimates. Assuming that θ and E are known, then the distribution of N is:

$$\text{Prob}(N = n) = P\, f(n; \alpha_1, \beta_1, E) + (1 - P)\, f(n; \alpha_2, \beta_2, E)$$

Let *Qn* be the posterior probability that λ came from the first component of the mixture, given $N = n$. From Bayes rule, the formula for *Qn* is:

$$Q_n = P\, f(n; \alpha_1, \beta_1, E) / [P\, f(n; \alpha_1, \beta_1, E) + (1 - P)\, f(n; \alpha_2, \beta_2, E)]$$

Then, the posterior distribution of λ after observing $N = n$ can be represented as:

$$\lambda | N = n \sim \pi(\lambda; \alpha_1 + n, \beta_1 + E, \alpha_2 + n, \beta_2 + E, Q_n)$$

where (as above):

$$\pi(\lambda; \alpha_1, \beta_1, \alpha_2, \beta_2, P) = P\, g(\lambda; \alpha_1, \beta_1) + (1 - P)\, g(\lambda; \alpha_2, \beta_2)$$

Because the posterior distribution of λ is often very skewed, we focus on the logarithmic expected value,

$$E[\log(\lambda_{ij}) \mid N_{ij}, \theta]$$

while defining our preferred point estimate of λ_{ij}.

To obtain a quantity on the same scale as *RR*, we define the Empirical Bayes Geometric Mean (EBGM):

$$EBGM_{ij} = e^{E[\log(\lambda_{ij}) | N_{ij}, \theta]}, \text{ where:}$$

$$E[\lambda \mid N = n, \theta] = Q_n (\alpha_1 + n)/(\beta_1 + E) + (1 - Q_n)(\alpha_2 + n)/(\beta_2 + E)$$

$$E[\log(\lambda) \mid N = n, \theta] = Q_n\, [\psi(\alpha_1 + n) - \log(\beta_1 + E)] + (1 - Q_n)\, [\psi(\alpha_2 + n) - \log(\beta_2 + E)]$$

where $\psi(x) = d(\log \Gamma(x))/dx$. In the same way, the cumulative gamma distribution function can be used to obtain percentiles of the posterior distribution of λ. The 5th percentile of λ is denoted:

$$EB05_{ij} = \text{Solution to: } \text{Prob}(\lambda < EB05 \mid N_{ij}, \theta) = 0.05$$

and is interpreted as a lower 1-sided 95% confidence limit. The upper limit $EB95_{ij}$ is defined analogously.