1. (i) After calculating the mean vector and the two covariance matricies, I found that

$$\hat{\mu} = \begin{bmatrix} 5.843 \\ 3.057 \\ 3.758 \\ 1.199 \end{bmatrix}, \quad S_b = \begin{bmatrix} 31.606 & -9.976 & 82.624 & 35.640 \\ -9.976 & 5.672 & -28.620 & -11.466 \\ 82.624 & -28.620 & 218.551 & 93.387 \\ 35.640 & -11.466 & 93.387 & 40.207 \end{bmatrix}$$

and

$$S_w = \begin{bmatrix} 0.265 & 0.093 & 0.168 & 0.038 \\ 0.093 & 0.115 & 0.055 & 0.033 \\ 0.168 & 0.055 & 0.185 & 0.043 \\ 0.038 & 0.033 & 0.043 & 0.042 \end{bmatrix}$$

(ii) The eigenvalues of $S_w^{-1}S_b$ are

$$\lambda = 2366.107,\ 20.976,\ 0.000,\ 0.000$$

This tells us that variation among blocks largely lies on a 1 or 2 dimmensional manifold. In particular, it can overwheminly be captured as occuring on a line. That $S_w^{-1}S_b$ follows from S_w being positive definite. Due to this, we can say $S_w = BB$ For some positive definite matrix B. Then,

$$\begin{aligned} S_w^{-1}S_b &= BBS_b \\ B^{-1}S_w^{-1}S_bB &= BS_bB \end{aligned}$$

BS_bB must be symmetric and thus have real eigenvalues. Since I have established $S_w^{-1}S_b$ is similar to a BS_bB, it too must have real eigenvalues. Since the eigenvalues are real, so are the eigenvectors, H. However, the $S_w^{-1}S_b$ need not be symmetric, so nothing is forcing H to be orthogonal. Let h_1 be the first eigenvector. The portion of the variance due to the clustering in Yh_1 is smaller than for any other linear combination.

(iii) To show that $\mathrm{E}(S_w) = \Sigma$, we observe

$$\begin{aligned} \Sigma &= \mathrm{E}\left[\left(Y_i - \mathrm{E}[Y_i \mid b_i]\right)' \left(Y_i - \mathrm{E}[Y_i \mid b_i]\right)\right] \\ &= \mathrm{E}\left[\frac{1}{n-|B|}\sum_{i=1}^{n}\left(Y_i - \frac{1}{n_b}\sum_{j\in b}Y_j\right)'\left(Y_i - \frac{1}{n_b}\sum_{j\in b}Y_j\right)\right] \\ &= \mathrm{E}\left[Y'Y + Y'\bar{B}Y\right] \\ &= \mathrm{E}\left[Y'(I+\bar{B})Y\right] \\ &= \mathrm{E}\left[S_w\right] \end{aligned}$$

Now, to show the result for $\mathrm{E}(S_b)$

$$\begin{aligned} (|B|-1)S_b &= \sum(Y_iY_i') - n\bar{Y}\bar{Y}' - Y'(I-J)Y \\ (|B|-1)S_b &= \sum(Y_iY_i' - \bar{Y}\bar{Y}') - Y'(I-J)Y \\ (|B|-1)S_b &= \sum(Y_iY_i' - \frac{1}{n}\sum_b n_b\bar{Y}_b\bar{Y}_b') - Y'(I-J)Y \\ (|B|-1)S_b &= \sum_b\left(\frac{1}{n}\sum_{i\in b}Y_iY_i' - \bar{Y}_b\bar{Y}_b'\right) - Y'(I-J)Y \\ \mathrm{E}(S_b) &= \Sigma + \theta\Sigma\frac{n^2 - \sum n_b^2}{n(|B|-1)} \end{aligned}$$

Using the result above, we can estimate the variance ratio based on

$$\mathrm{E}\left[S_b\right] = \mathrm{E}\left[S_w\right]\left(1 + \theta\frac{n^2 - \sum n_b^2}{n(|B|-1}\right)$$

Allowing us to derive an estimator

$$\begin{aligned}
tr(S_b) &= tr(S_w)\left(1 + \hat{\theta}\frac{n^2 - \sum n_b^2}{n(|B|-1)}\right) \\
\frac{tr(S_b)}{tr(S_w)} &= 1 + \hat{\theta}\frac{n^2 - \sum n_b^2}{n(|B|-1)} \\
\frac{tr(S_b)}{tr(S_w)} - 1 &= \hat{\theta}\frac{n^2 - \sum n_b^2}{n(|B|-1)} \\
\hat{\theta} &= \frac{n(|B|-1)}{n^2 - \sum n_b^2}\left(\frac{tr(S_b)}{tr(S_w)} - 1\right)
\end{aligned}$$

Using this estimator, I get an estimate of $\hat{\theta} = 9.727$ on the original scale, and $\hat{\theta} = 18.210$ when the log is taken of the data.

(iv) Calculating the predictive probabilities with $\lambda = 1$ and $\theta = 10$, I got this result:

	setosa	versicolor	virginica	other
(6.3, 2.9, 4.9, 1.7)	0.000	0.471	0.529	0.000
(5.5, 3.1, 2.9, 0.8)	0.071	0.167	0.000	0.762
(5.8, 3.2, 3.5, 1.1)	0.000	0.986	0.000	0.014
(5.6, 3.2, 3.2, 1)	0.000	0.868	0.000	0.132

(v) Now, with $\lambda = 1$ and $\theta = 5$, I got

	setosa	versicolor	virginica	other
(6.3, 2.9, 4.9, 1.7)	0.000	0.464	0.535	0.001
(5.5, 3.1, 2.9, 0.8)	0.035	0.079	0.000	0.886
(5.8, 3.2, 3.5, 1.1)	0.000	0.956	0.000	0.044
(5.6, 3.2, 3.2, 1)	0.000	0.692	0.000	0.308

And with $\lambda = 1$ and $\theta = 50$, I got

	setosa	versicolor	virginica	other
(6.3, 2.9, 4.9, 1.7)	0.000	0.477	0.523	0.000
(5.5, 3.1, 2.9, 0.8)	0.240	0.592	0.000	0.169
(5.8, 3.2, 3.5, 1.1)	0.000	0.999	0.000	0.001
(5.6, 3.2, 3.2, 1)	0.000	0.992	0.000	0.008

The predictions do seem to be somewhat sensitive to θ. In general, θ effects the relative probabilities of the existing clusters only slightly, but has a large effect on the probability of a new cluster.

(vi) As $\theta \to 0$

$$n_b\,\phi_4\!\left(y(u') - \frac{\mu + n_b\theta\bar{y}_b}{1 + n_b\theta}; \Sigma\left(1 + \frac{\theta}{1 + n_b\theta}\right)\right) \to n_b\,\phi_4\big(y(u') - \mu; \Sigma\big)$$

and

$$\lambda\,\phi_4\big(y(u') - \mu; \Sigma\,(1+\theta)\big) \to \lambda\,\phi_4\big(y(u') - \mu; \Sigma\big)$$

This means that, since $\phi_4(y(u') - \mu; \Sigma)$ is a common factor,

$$\mathrm{P}(u' \mapsto b \mid ...) \to \begin{cases} \frac{n_b}{\lambda+n} & b \in B \\ \frac{\lambda}{\lambda+n} & b = \emptyset \end{cases}$$

Which is the Ewens process and independent of the feature vector. On the other hand, if $\theta \to \infty$, we see that

$$n_b\,\phi_4\left(y(u') - \frac{\mu + n_b\theta\bar{y}_b}{1+n_b\theta};\Sigma\left(1+\frac{\theta}{1+n_b\theta}\right)\right) \to n_b\,\phi_4\left(y(u') - \bar{y}_b;\Sigma\left(1+\frac{1}{n_b}\right)\right)$$

and

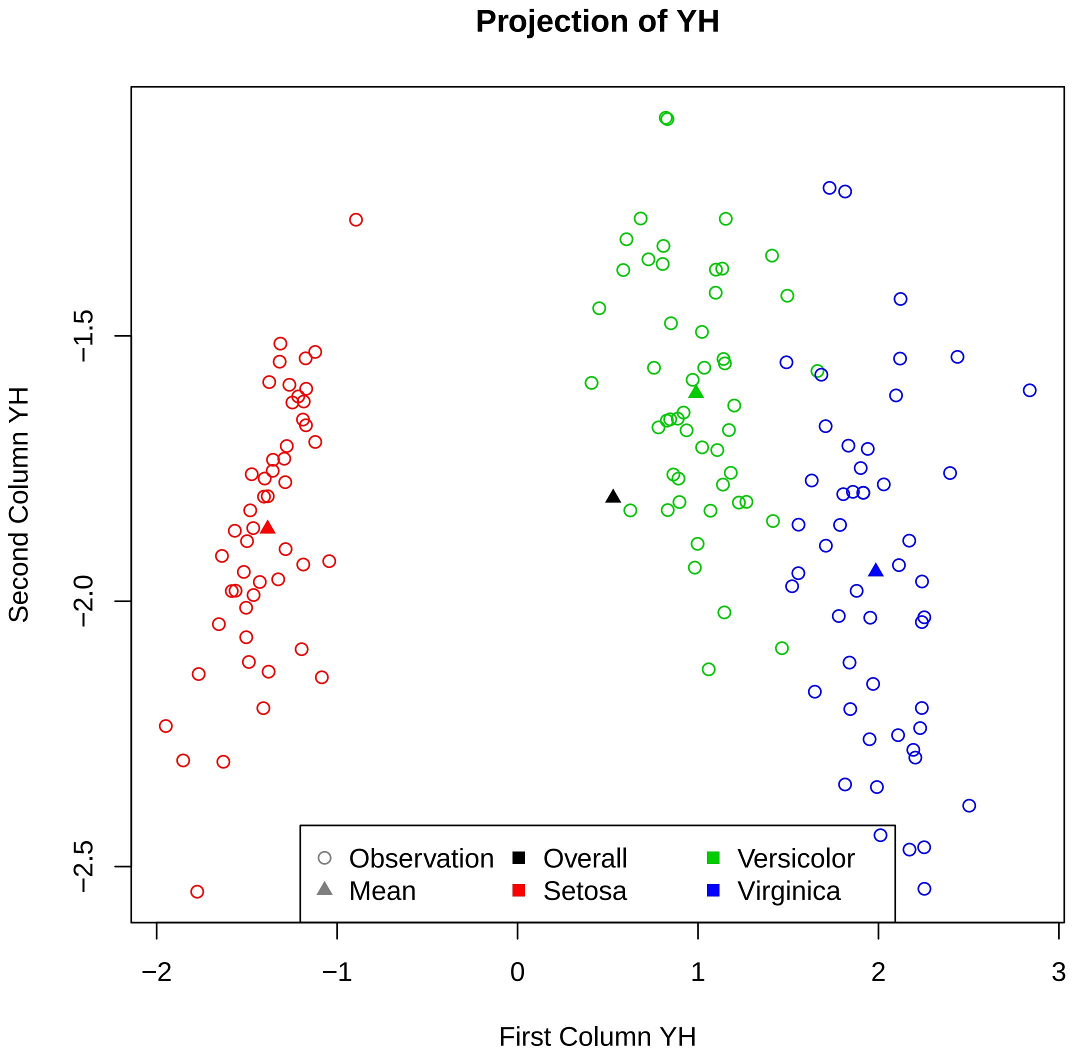
$$\lambda\,\phi_4\big(y(u') - \mu;\Sigma\,(1+\theta)\big) \to 0$$

so the predictive probability for the new class tends to zero and

$$\mathrm{P}(u' \mapsto b \mid b \in B\ldots) \propto n_b\phi_4\big(y(u') - \bar{y}_b;\Sigma\big)$$

which is the Bayes optimal solution based on the Fisher discriminant and class frequencies based on sample frequencies.

(vii) Making the scatter plot of the projection:

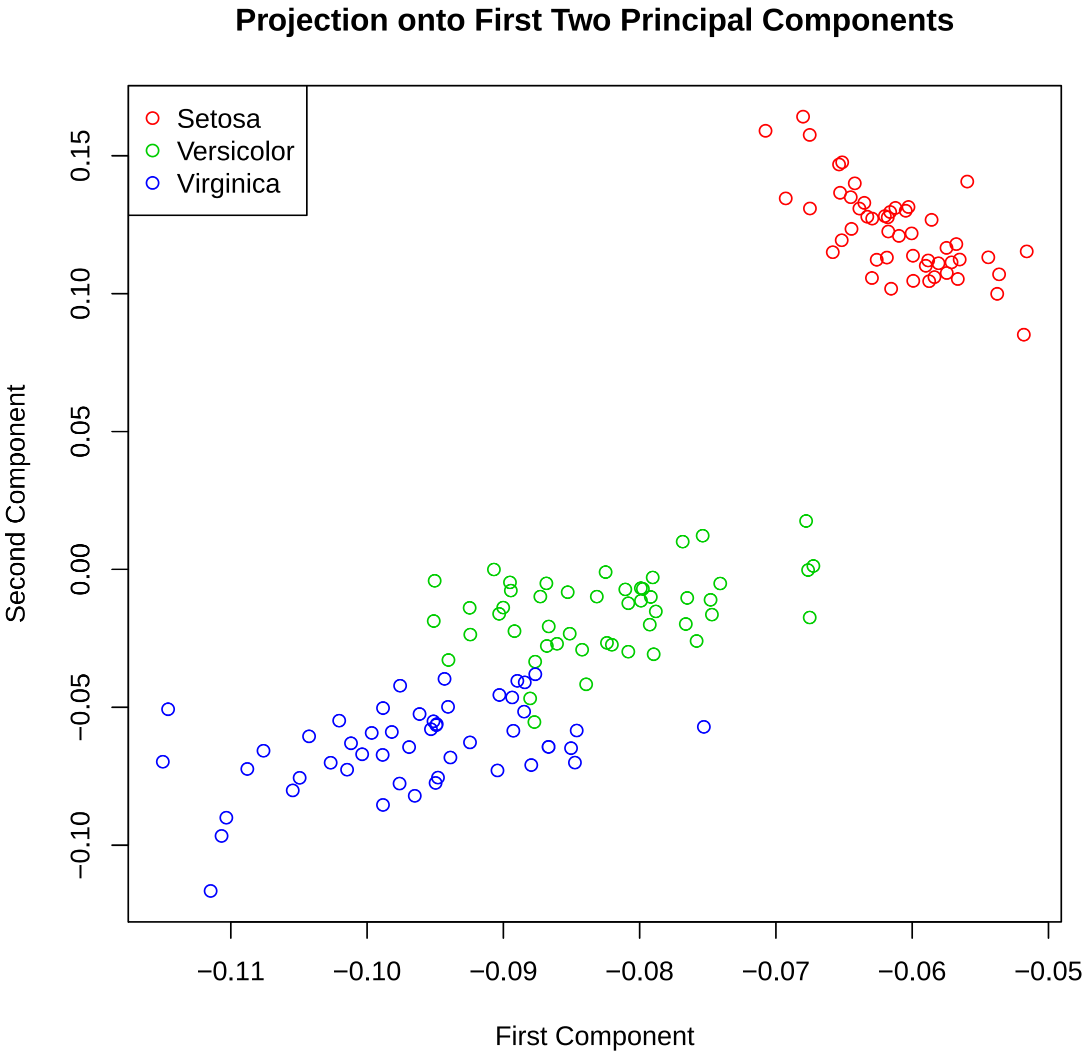


Looking at this plot, its unsurprising that none of the values we predicted had high probabilities for both Setosa and another existing cluster; it is far more distinct from the other two than they are from each other. If its probability is of similar magnitude to one of the other clusters, then that means that the point is in the big gap in the middle, and the most likely prediction is probably a new cluster. Now, predicting $\bar{y}$, I got

	setosa	versicolor	virginica	other
Mean y	0.000	0.999	0.000	0.001

The overall mean is much closer to the Versicolor mean than the other means, so these probabilities are very skewed towards Versicolor.

(viii) Here is the alternate plot of the data



This is a plot of the data's projection into the first two principal components of the data matrix (though without subtracting out the mean or normalizing). We are still capturing the majority of the variation in the data; we're just using a different basis.

(ix) One might want to use a log scale since all the measurements by their nature must be positive, which defies the normality assumption. When one makes predictions based on the log scale, these are the new probabilities.

	setosa	versicolor	virginica	other
(6.3, 2.9, 4.9, 1.7)	0.000	0.415	0.585	0.000
(5.5, 3.1, 2.9, 0.8)	0.000	0.543	0.000	0.457
(5.8, 3.2, 3.5, 1.1)	0.000	0.997	0.000	0.002
(5.6, 3.2, 3.2, 1)	0.000	0.972	0.000	0.028

Besides the second point, the predictions are quite similar as before.

5. (i) **Show this is a probability distribution:** This follows from this manipulation

$$\begin{aligned}
\mathrm{P}(T = t; m, n) &= \frac{\prod_i n_i! \prod_j m_j!}{n_*! \prod_{ij} t_{ij}!} \\
&= \frac{n_*! \prod_i n_i! \prod_j m_j!}{n_*! n_*! \prod_{ij} t_{ij}!} \\
&= \frac{\prod_i n_i!}{n_*!} \frac{\prod_j m_j}{n_*!} \frac{n_*!}{\prod_{ij} t_{ij}!} \\
&= \frac{\binom{n_*}{t_{11},\ldots,t_{cr}}}{\binom{n_*}{n_1,\ldots,n_r}\binom{n_*}{m_1,\ldots,m_c}}
\end{aligned}$$

The numerator is the number of ways to fill fill each cell with the desired t_{ij} from n_* observations, and the denominator is the number of all possible tables which satisfy the marginal distributions. Thus, it is clear that summing the probability over all possible T which satisfy the marginals will yield 1, making this a proper probability distribution.

(ii) **Show $T = \tilde{A}'P\tilde{B}$ has a hypergeometric distribution:** Since PB is just a permutation of the rows of B, the column sums will remain the same. This means that the marginals for the new factor mapping to PB is the same as for the factor for B. In turn, the resulting table $A'PB$ will have the same marginals in both the rows and columns, so clearly there is the proper conditioning. Further, each permutation is chosen with probability $\frac{1}{n_*!}$. However, most of these permutations result in the same table. In particular, if one permutes the values of B matched with each particular value of A, then the table is the same. There are $\prod_i n_i!$ such permutations. Similarly, permutations of a particular value of B, of which there are $\prod_j m_j$ will also result in the same table. Then, there are $\prod_{ij} t_{ij}!$ permutations that do both of these things at once. It follows that there are

$$\frac{\prod_i n_i! \prod_j m_j!}{\prod_{ij} t_{ij}!}$$

permutations total that result in the same table. Thus, we conclude that the probability under this model of drawing a particular table is

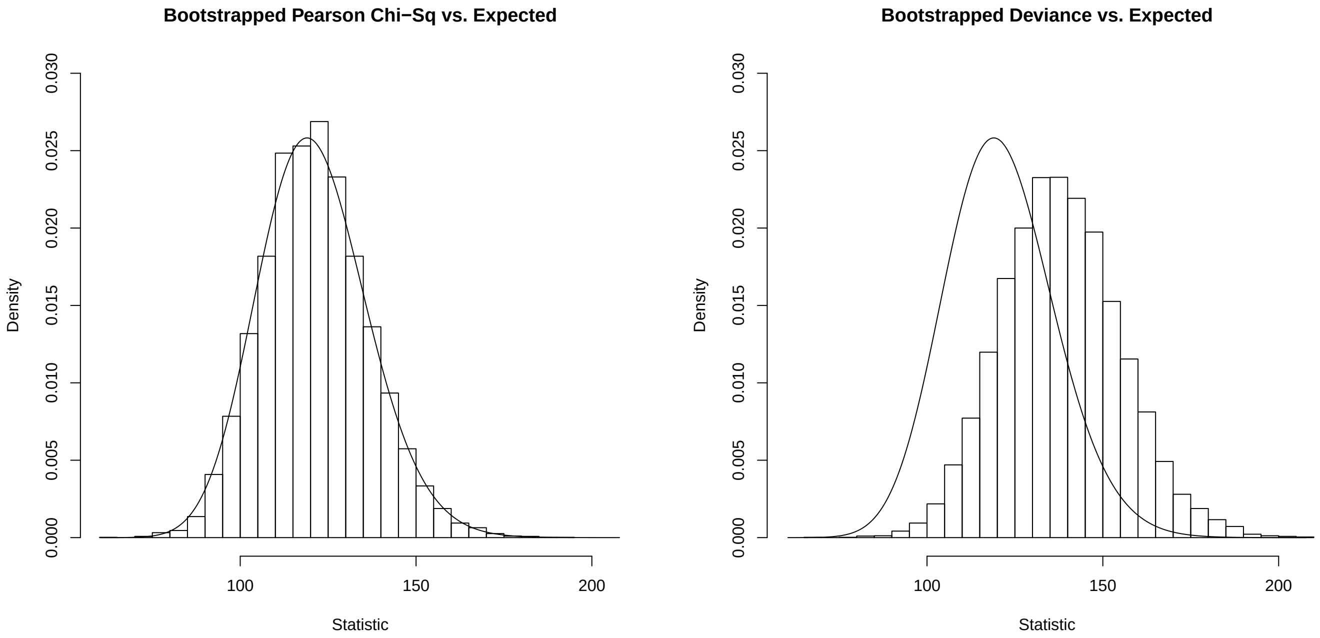
$$\frac{\prod_i n_i! \prod_j m_j!}{n_*! \prod_{ij} t_{ij}!}$$

Making it a hypergeometric distribution.

(iii) **Compute Pearson's chi-squared test and deviance statistic:** I found that the nominal Pearson Chi-square statistic is 109.883, and the deviance statistic is

$$2\sum_{i,j} t_{ij}\left(\log(t_{ij}) - \log\left(n_* \frac{m_j}{n_*}\frac{n_i}{n_*}\right)\right) = 130.599$$

(iv) **Estimate the p-value by simulation:** Over 10000 simulations, I found that the Perason's chi-squared test statistic was exceeded 7723 times, and the deviance statistic was exceeded 6689 times. These correspond to p-values of .772 and .669 respectively. Below I've plotted both sets of bootstraps versus the density of the expected chi-square with 121 degrees of freedom:



The Pearson statistic is extremely close the Chi-sq, as one would hope for. However, the deviance statistic is noticeably shifted to the right. This

(v) **Aggregate the data by birth death month difference:** After aggregating the expected and actual birth and death combinations by the difference in month, this is the resulting table:

	-5	-4	-3	-2	-1	0	1	2	3	4	5	6
Actual	31.000	20.000	23.000	34.000	16.000	26.000	36.000	37.000	41.000	26.000	34.000	24.000
Expected	29.101	29.187	28.733	28.052	28.302	28.773	28.885	28.730	29.767	29.250	29.362	29.859

We seem to see fewer than expected deaths in the months leading up to someone's birthday, and more deaths than expected afterwards. Using the Pearson Chi-squared test to evaluate this, I get a p-value of .027, which is evidence, though not overwhelming, supporting that this is statistically significant.

(vi) **Explain how to extend the hypergeometric simulation to three dimensions where each one dimensional table fixed:** For each cell, the expected count in each cell under the null should still be the product of the marginal probabilities times the total n. Thus, for for the indicator matrices $\tilde{A}$, $\tilde{B}$, $\tilde{C}$, we can simulate a random null matrix fitting the marginals using random permutation matrices P_1 and P_2. A cell of the table is

$$T_{i,j,k} = \sum_{m=1}^{n} \tilde{A}_{mi}(P_1\tilde{B})_{mj}(P_2\tilde{C})_{mk}$$

with expectation, under the null, of

$$\mathrm{E}[T_{i,j,k} \mid H_o] = n\frac{\|\tilde{A}_i\|_1}{n}\frac{\|\tilde{B}_j\|_1}{n}\frac{\|\tilde{C}_k\|_1}{n}$$

From here, the simulation is the same, bootstrapping a test statistic over many simulated tables and comparing it to the observed value.

(vii) **Explain how to extend the hypergeometric simulation to three dimensions with two-dimension tables fixed:** Here, the null is that the cells are conditionally independent. In other words,

$$\begin{aligned}\mathrm{P}(A_m = i, B_m = j, C_m = k) &= \mathrm{P}(A_m = i, B_m = j \mid C_m = k)\mathrm{P}(C_m = k)\\ &= \mathrm{P}(A_m = i \mid C_m = k)\mathrm{P}(B_m = j \mid C_m = k)\mathrm{P}(C_m = k)\end{aligned}$$

This can be tested by the same method as with two way tables, just induced on all the subtables $T_{*,*,k}$ at the same time. For each subset of A_k and B_k, which are the values of A and B for which $C = k$, draw a random permutation P, and then $\tilde{A}_k' P \tilde{B}_k$ should be hypergeometric. A test statistic is then generated on all the two way tables for all k by comparing the seen cells $T_{i,j,k}$ to

$$\mathrm{E}(T_{i,j,k}) = \frac{1}{u_k} \sum_{i'} T_{i',j,k} \sum_{j'} T_{i,j',k}$$

Where u_k is the number of times level k of C occurs.