Analytical Cellular Pathology 25 (2003) 187–191
IOS Press

Application of multilevel models to morphometric data. Part 2. Correlations

O. Tsybrovskyy [a,*] and A. Berghold [b]
[a] *Department of Pathology, School of Medicine, University of Graz, Austria*
[b] *Institute for Medical Informatics, Statistics and Documentation, University of Graz, Austria*

Received 17 October 2002
Accepted 28 April 2003

Abstract. Multilevel organization of morphometric data (cells are "nested" within patients) requires special methods for studying correlations between karyometric features. The most distinct feature of these methods is that separate correlation (covariance) matrices are produced for every level in the hierarchy. In karyometric research, the cell-level (i.e., within-tumor) correlations seem to be of major interest. Beside their biological importance, these correlation coefficients (CC) are compulsory when dimensionality reduction is required. Using MLwiN, a dedicated program for multilevel modeling, we show how to use multivariate multilevel models (MMM) to obtain and interpret CC in each of the levels. A comparison with two usual, "single-level" statistics shows that MMM represent the only way to obtain correct cell-level correlation coefficients. The summary statistics method (take average values across each patient) produces patient-level CC only, and the "pooling" method (merge all cells together and ignore patients as units of analysis) yields incorrect CC at all. We conclude that multilevel modeling is an indispensable tool for studying correlations between morphometric variables.

1. Introduction

In Part 1, we considered methods of testing hypothesis with the help of multilevel models, using an example of exploring nature and significance of differences between benign and malignant follicular tumors of the thyroid. Another type of research questions that can arise in the morphometry is concerned with correlation structure of karyometric features. First, correlation structure can be interesting and important by itself; that is, we can group the measured features on the basis of correlations and thus judge what aspects of nuclear morphology they reflect. Second, correlation (or covariance) matrix can be used to perform dimensionality reduction, i.e., to eliminate redundant variables, which do not convey any more information in addition to others. Dimensionality reduction usually is accomplished by factor or cluster analysis and is sometimes very important before fitting other types of models (e.g., classificatory) [5]. In the present article, we report our first experience of using multilevel models for studying correlations between karyometric variables. In analogy to Part 1, a comparison with the "single-level" statistics is made in order to demonstrate the major benefits of the new technique.

*Corresponding author: Dr. Oleksiy Tsybrovskyy, Department of Pathology, University of Graz, Auenbruggerplatz 25, 8036 Graz, Austria. Tel.: +43 316 385 80 461, +43 316 380 44 11; Fax: +43 316 385 34 32, +43 316 38 43 29; E-mail: oleksiy.tsybrovskyy@kfunigraz.ac.at.

2. Materials and methods

Materials and methods used were described in Part 1. The same data set containing 15378 cells nested within 78 tumors was used. However, only 7 out of 8 measured nuclear features (see Table 1 in Part 1) were used in the present study. We experienced convergence problems with the full model, and thus decided to exclude IOD, as a secondary feature deriving directly form NA and indirectly from MGV.

3. Statistical analysis and results

3.1. Some theoretical considerations

As discussed in Part 1, the most important difference of multilevel models against "single-level" statistics is

Table 1
Pearson correlation coefficients between karyometric features. In each cell, uppermost left CC is for cell-level (MLwiN), lowermost left CC for tumor-level (MLwiN), and right CC is that obtained in the "pooling" method

Variable	NA		NC		MGV		SDGV		SkewGV		KurtGV	
SAD	−0.145	0.100	−0.018	0.196	−0.005	0.261	0.499	0.486	−0.053	−0.346	−0.183	−0.374
	0.357		0.523		0.490		0.470		−0.618		−0.551	
KurtGV	0.019	−0.216	−0.036	−0.245	−0.191	−0.342	−0.198	−0.371	0.319	0.443		
	−0.503		−0.681		−0.562		−0.558		0.660			
SkewGV	−0.394	−0.471	−0.208	−0.328	−0.721	−0.808	0.292	0.178				
	−0.578		−0.588		−0.884		−0.015					
SDGV	−0.219	−0.054	−0.104	0.106	−0.353	−0.247						
	0.186		0.520		−0.070							
MGV	0.553	0.625	0.221	0.370								
	0.721		0.678									
NC	0.209	0.373										
	0.737											

that the variance of the dependent variable(s) is split into corresponding levels of the data hierarchy. In the context of studying correlations, this means that we obtain *two* different covariance matrices and *two* sets of correlation coefficients (CC) – one for each level. And again, we are confronted with the *level-related* interpretation of these CC.

To calculate covariance matrices, we must create a *multivariate* multilevel model (MMM). In MMM, all correlated variables are considered "multivariate response" and thus, specified as dependent variables, in the left part of the model equation. In the right part, we can specify constant term only ("empty model"), or add some additional covariates – e.g., DIAGNOSIS. In the former case we obtain overall CC for the entire population of follicular thyroid tumors, whereas in the latter case, separate CC for adenomas and carcinomas are computed.

As with all linear models, MMM require assumptions of normality, linearity and homoscedasticity to be met. These assumptions are checked in the same way as in the univariate models (see Part 1). After log transform of NA, there were no serious violations of these assumptions in our study.

Note that the current version of MLwiN [7] does not include direct facilities for studying correlations and performing dimensionality reduction. Therefore, correlation matrices produced by MLwiN were imported into SPSS 10.0 and submitted to factor and cluster analyses using SPSS syntax language [6].

3.2. Correlation coefficients

The CC produced by MLwiN for both levels in our study are given in Table 1. Note that for the same pairs of variables, CC are often rather different, sometimes even quite opposite. This emphasizes the importance of the level-related nature of these CC. Level-1 CC are those we would expect to get *within a given single tumor*. For example, CC of −0.145 between NA and SAD at level-1 means that, *within a single given tumor*, larger nuclei tend to have more evenly distributed chromatin. By contrast, level-2 CC refer to the tumors – in essence, to the *average values* of karyometric features across each tumor. Thus, CC of 0.357 at level-2 indicates that tumors with larger nuclei (i.e., with larger mean nuclear size) have usually a higher proportion of nuclei with coarser chromatin. This by no means implies, however, that these are *the same* nuclei that both are larger and have coarser chromatin, because these CC does not refer to cells; the cell-level CC indicates just the opposite tendency.

For comparison, we computed also CC using two "single-level" approaches: summary statistics method (taking average values of karyometric features within each tumor), and "pooling" method (pooling all cells together and ignore patients as units of analysis). CC in the summary statistics method (not shown) were perfectly the same as the level-2 CC produced by MMM. As for the "pooling" method, CC in this approach were always somewhere in between the separate-level CC (see Table 1). This is explained by the fact that the "pooled" CC are computed on the basis of a "mixture" of variances from both levels. It should be stressed that neither of the "single-level" approaches allowed obtaining correct CC for the *cell* level. This is only possible by means of MMM.

As mentioned above, separate CC can be computed for adenomas and carcinomas. For this, the model

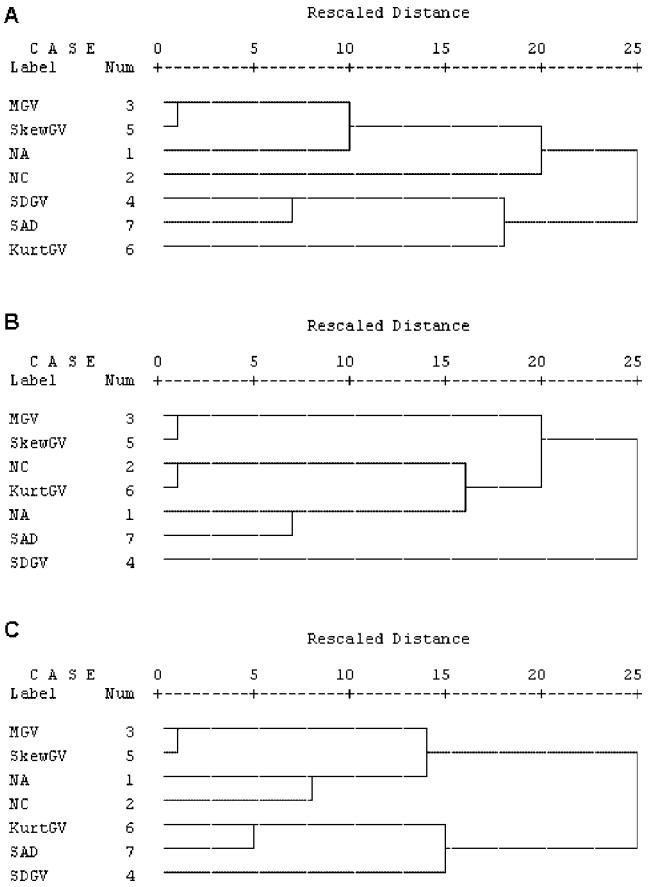

Fig. 1. Horizontal dendrogram representing steps in hierarchical clustering (centroid method) of karyometric variables: A – using level-1 CC obtained in MMM; B – using level-2 CC obtained in MMM (or, equally, using summary statistics method); C – using "pooling" method.

should include DIAGNOSIS as a covariate with *random* coefficient at both levels.[1] In addition, all redundant covariance parameters must be set to 0, so that, finally, only 3 of them remain: covariance between constant term and DIAGNOSIS coefficient for each of the nuclear features, and covariance between the features themselves (i.e., between constant terms). CC for adenomas computed from such covariance matrix is displayed directly in the "Estimates" window of MLwiN. For carcinomas, however, manual imputation of CC is necessary (for details, see [4,7,9]). In our study, CC for carcinomas were generally somewhat lower than for adenomas, which might be due to higher nuclear atypia in malignant tumors. However, these differences were mostly minor and non-significant (data not shown).

[1]If the coefficient for DIAGNOSIS is specified as a *fixed*, the variances of both dependent variables decrease, but no additional variance–covariance terms appear in the model equation, and the correlation between the variables remains exactly the same.

3.3. Cluster analysis

Figure 1 shows hierarchical grouping of variables repeated for each set of CC given in Table 1. Figure 1A suggests the presence of 2 pairs of correlated variables, having similar biological interpretation: MGV and SkewGV, as well as SDGV and SAD. Taking into account the calculation methods of these features [2], the first pair reflects the amount of lightly stained areas within a nucleus. The second pair describes the heterogeneity, roughness of the chromatin, the contrast between lightly and intensively stained areas in a nucleus [2]. By contrast, Fig. 1B implies the following grouping of variables: MGV and SkewGV; NC and KurtGV; NA and SAD. While the first pair of variables is the same as in Fig. 1A, the other two are completely different and difficult to explain. In any case, such an explanation should refer to tumors, not to cells, since the underlying CC are from the level-2. Finally, Fig. 1C suggests also 3 groups: MGV and SkewGV; NA and NC;

KurtGV and SAD. Again, the last two pairs of variables are different from those in Fig. 1A – certainly because the underlying CC are not cell-related, but rather both cell- and tumor-related, as discussed above. Thus, if a true cell-related classification of variables or dimensionality reduction is required, both "single-level" methods are inappropriate.

4. Discussion

The most specific feature of studying correlations in a multilevel design is that separate CC are generated for each of the levels. This may appear very confusing. What CC should be used to explore interrelations between karyometric features? First, it should be noted that there are no "right" or "wrong" CC generated in MMM. CC from different levels reflect essentially different aspects of association between variables and are complementary to each other. One example of interpretation was given in the previous section. As another example, consider CC between nuclear size (NA) and nuclear form factor (NC). The positive cell-level CC (0.209, see Table 1) indicates that *within a single given tumor*, smaller cells tend to be rounder, whereas larger cells are more irregular in shape; however, this association is weak. It is this CC that we would expect to obtain having measured a cell sample within a given tumor. By contrast, the tumor-level CC is much higher (0.737, see Table 1), and its meaning is distinct: it says that tumors composed of cells with smaller nuclei contain also more cells with round-shaped nuclei, and vice versa. It does not tell, however, that these are the *same cells* that are both large and irregularly shaped; this might well be different cell populations within each tumor.

It is clear from the reasoning above that cell-level CC usually are more interesting for the morphometric research than tumor-level CC. An even more serious argument in favor of estimating cell-level CC is a dimensionality reduction (e.g., by means of factor or cluster analysis). Modern programs for image analysis usually have tens of measurable parameters [1], and the maximum number of karyometric features (including textural) makes up to several hundreds to date [2,3]. Since it is often impossible to take correspondingly larger samples of patients, a low sample–feature ratio ensues. It is very hazardous to develop any classificatory models or search for a "best feature subset" on such data sets [8]. However, many nuclear features are closely correlated with others in the set [2]. It is thus

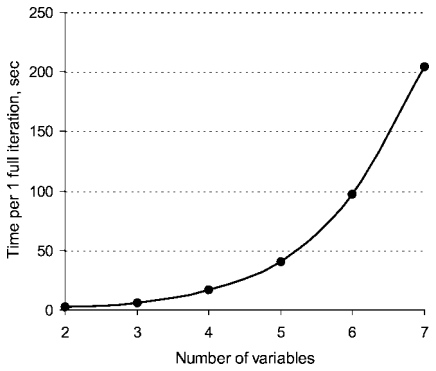

Fig. 2. Time required by MMM to complete one full iteration, in dependence on the number of correlated variables included. Calculations were performed by a PC equipped with AMD Athlon processor 700 MHz and 128 MB memory on a data set containing 15378 cells nested within 78 tumors.

reasonable to perform a reduction of dimensions prior to any other statistical tests and procedures, in order to improve the sample-feature ratio [5]. Note that if the goal is, e.g., to develop a "cell classifier" and a dimensionality reduction is necessary, cell-level CC are *the only* CC suited for this purpose, because reduction of dimensions is performed on the level of cells. And, as noted above, MMM represent the only way to compute correct cell-level CC. Summary statistics method produces level-2 CC only, and the "pooling" method produces incorrect CC at all. Even the recent extension of ANOVA, namely mixed-effect general linear model, is of no use here, because it cannot handle multivariate data [6].

On the other hand, we must again discuss some technical problems related to MMM. As mentioned in Part 1, multivariate models are rather computationally extensive. In MMM, this becomes an extreme problem as the number of dependent variables increases, especially taking into account that the number of level-1 units in the morphometry is commonly also very large. Regarding our data, the length of a full iteration increased with inclusion of each additional variable nearly-exponentially, for 7 variables being about 3 min 24 sec (Fig. 2). In addition, the number of iterations for model to converge increased from 2 (with only 2 variables included) to 44 (with all 7 variables included).[2] Moreover, MMM fail to converge under certain circumstances. This is usually due to the presence of either highly correlated variables (in our experience, when CC at both levels are 0.98 and higher) or

[2] A model containing 10 variables did not converge after 4 days of uninterrupted calculations (47 iterations).

some non-significant parameters with very high (relative) standard errors, which, again, is especially frequent with a large number of variables. Fortunately, the problem can be solved by splitting the whole set of variables into a number of small subsets and computing CC within each subset. We have been convinced of MLwiN generating the same CC regardless of the subset size, down to pairwise (data not shown).

To sum up, multilevel models are superior over conventional, "single level" statistics when applied to morphometric data, in that they produce correct cell-level CC. These CC are the most interesting with regard to their biological interpretation. Moreover, they (and only they) can be used for dimensionality reduction, if required by subsequent statistical procedures involving the level of cells.

References

[1] *Optimas 6.5: User Guide and Technical Reference*, Media Cybernetics, Silver Spring, 1999.

[2] A. Doudkine, C. Macaulay, N. Poulin and B. Palcic, Nuclear texture measurements in image cytometry, *Pathologica* **87**(3) (1995), 286–299.

[3] T. Dreyer, I. Knoblauch, D. Garner, A. Doudkine, C. MacAulay, B. Palcic and C. Popella, Specific changes of chromatin structure in nuclei of normal epithelium adjacent to laryngeal squamous cell carcinoma, *Analyt. Cell. Pathol.* **20**(2/3) (2000), 141–151.

[4] H. Goldstein, *Multilevel Statistical Models. Internet ed.*, Edward Arnold, London, 1999.

[5] F. Harrell, K.L. Lee and D.B. Mark, Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors, *Statistics in Medicine* **15** (1996), 361–387.

[6] M.J. Norusis, *SPSS 10.0 Syntax Reference Guide*, SPSS Inc., Chicago, 1999.

[7] J. Rabash, W. Browne, H. Goldstein, M. Yang, I. Plewis, M. Healy, G. Woodhouse, D. Draper, I. Langford and T. Lewis, *A User's Guide to MLwiN*, Univ. of London, London, 2000.

[8] H. Schulerud, G.B. Kristensen, K. Liestol, L. Vlatkovic, A. Reith, F. Albregtsen and H.E. Danielsen, A review of caveats in statistical nuclear image analysis, *Analyt. Cell. Pathol.* **16**(2) (1998), 63–82.

[9] A.B. Snijders and R.J. Bosker, *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, Sage Publishers, London, 1999.