Feature network models for proximity data : statistical inference, model selection, network representations and links with related models

Frank, L.E.

Chapter 6

Epilogue: General Conclusion and Discussion

6.1 Reviewing statistical inference in Feature Network Models

In this monograph, statistical inference in FNM has been accomplished using the multiple regression framework. It provided the basis for the estimation of standard errors, confidence intervals , model test and features subset selection. This framework has been helpful in solving some problems, but there remain problems unsolved. The following sections review the results.

Constrained estimation

Considering features in FNM as predictor variables leads to the univariate linear regression model with positivity constraints on the feature discriminability parameters. Due to these positivity constraints, the ordinary least squares estimator becomes the inequality constrained least squares estimator (ICLS). One of the key problems has been to assess the variability of the feature discriminability parameters estimated with the ICLS estimator, which has a truncated (normal) distribution. Statistical inference in inequality constrained least squares problems is far from straightforward. While there is abundant literature on the computational aspects of the ICLS estimators (*cf.* Golub & Loan, 1989; Lawson & Hanson, 1995; Wollan & Dykstra, 1987), a recent review on statistical inference in inequality constrained least squares problems (Sen & Silvapulle, 2002) showed that optimal estimators or tests of significance generally do not exist for such nonstandard methods. In this context, there is only one author (Liew, 1976) who proposed a direct method to obtain theoretical standard errors for the ICLS estimator. The combination of Liew's theory on standard errors for the ICLS estimator with an efficient algorithm to compute ICLS estimates (Algorithm AS 225, Wollan & Dykstra, 1987) written in Matlab code, made it possible to obtain feature discriminability values and their associated standard errors. This method can be used in any inequality constrained least squares situation.

Imposing positivity constraints results from a priori ideas about the true model or properties of the population. In the context of FNM the prior belief would be that there exists a representation of the data in terms of a network where the edge lengths are all positive. If one believes that the true model is positive, then, negative values are non existing effects (in the sense that they only result from sampling error) and should lead to parameter values equal to zero. The theory about standard errors of inequality constrained least squares estimators proposed by Liew (1976) is based on the assumption that the true model parameters are positive and, subsequently, yields zero values for standard errors related to parameters where the constraints are activated. Negative effects, and therefore, non existing effects, yield standard errors equal to zero, which expresses the idea that non existing effects are not allowed to have variability. There are differences in opinion about this idea. Tibshirani (1996), for example, uses a ridge regression approximation for the estimation of standard errors for the Lasso parameters that also leads to zero values for the Lasso parameters that are shrunken to zero, and finds this an inconvenience. Shrinking parameters is a rather smooth procedure that eventually leads to parameter values equal to zero. Inequality constrained least squares is not smooth: a parameter is zero or positive. In the context of FNM, negative parameter values are assumed to result from sampling error (irreducible measurement error) and, consequently, should not be part of the model.

The theoretical standard errors for the feature discriminability parameters have been used in two different settings in this monograph. Chapter 2 provided results for the theoretical standard errors for an a priori known feature structure in FNM. Chapter 3 showed an application of the theoretical standard errors for the feature structure of additive trees, a special case of FNM. Standard errors that yielded adequate 95% *t*-confidence intervals were obtained in two situations: a priori known tree topologies and estimated tree topologies. The results on estimated tree topologies, which were based on the possibility to split the data in a training and a test sample, are expected to hold for the general FNM framework as well.

In both aforementioned studies, the performance of the theoretical standard errors was evaluated with Monte Carlo simulation techniques and compared with bootstrap standard deviations of the sampling distribution of the ICLS estimator. The performance criterion was the coverage probability of 95% *t*-confidence intervals. The coverage results for a priori known feature structure for the FNM were different for the theoretical standard errors compared to the bootstrap standard deviations, with a tendency to undercover for the bootstrap standard deviations and a tendency to overcover for the theoretical standard deviations. The tendency to undercover for the bootstrap confidence intervals was more prominent in the simulations with the additive tree models.

Although the bootstrap is renowned to be applicable in a wide range of situations, the inequality constrained least squares framework poses some limitations to the use of the bootstrap to assess the variance of a statistic. The consequence of imposing positivity constraints is that the empirical distribution is no longer centered around the true parameter value, which threatens the consistency of the bootstrap distribution and this might explain the tendency to undercover for the bootstrap confidence intervals . Especially when more constraints are activated, the distributions of the parameters are affected although it is not precisely known what the consequences are. Self and Liang (1987) studied several cases of constrained parameters when true parameter values are on the boundary of parameter space. The authors approximated the distributions of the constrained parameters by a mixture of χ^2 distributions when a small number of parameters are constrained. With more activated constraints, as is often the case in additive trees, the distribution of the constrained parameters can no longer be approximated by a mixture of χ^2 distributions. The reason that the theoretical standard errors proposed by Liew (1976) work is that they are based on the standard errors of the unconstrained (ordinary) least squares estimator. According to Self and Liang (1987) in the presence of boundary parameters one could always reflect the parametric distributions across the boundary to create a larger problem where the boundary points become interior points. The ordinary least squares solution represents the larger problem in this case.

Bootstrap standard deviation

In this monograph, the theoretical standard errors were compared to the true values and to the bootstrap standard deviations. Unlike the computation of the theoretical standard deviations, which is straightforward once the correct formula is known, the calculation of the bootstrap standard deviations is preceded by several decisions and limitations. Considering the FNM as univariate regression models influenced the choice of a resampling method, but also the presence of dissimilarity values in FNM. In the context of regression models there are different ways of resampling with different outcomes (Efron & Tibshirani, 1998; Freedman, 1981; Freedman & Peters, 1984): sampling residuals or sampling pairs of observations (value of the dependent variable and corresponding row or the matrix of predictor variables). Bootstrapping pairs was imposed by the FNM setting as a way to avoid negative dissimilarities that could result from sampling residuals. It is also one of the resampling methods in the linear regression model context that is robust in presence of heterogeneous error variance (Liu & Singh, 1992b).

In addition to the choices that have to be made about the method of resampling, the properties of the bootstrap standard deviations are not completely understood yet. Even in the "simple" linear regression case, without constraints, the consistency of the variance of the bootstrap distribution has not received much attention, while the consistency of the OLS estimator is well established in the literature; Gonçalves and White (2005) say on this topic:

> "The consistency of the bootstrap distribution, however, does not guarantee the consistency of the variance of the bootstrap distribution (the bootstrap variance) as an estimator of the asymptotic variance, because it is well known that convergence in distribution of a random sequence does not imply convergence of moments".

A better evaluation of the bootstrap standard deviations themselves could be achieved by performing a double bootstrap, as suggested by Gonçalves and White (2005) where the bootstrap is used to simulate the distribution of the *t* statistic which is based on a standard deviation that in turn has been estimated by the bootstrap. However, the authors did not actually perform the double bootstrap, which means bootstrapping the bootstrap, because the implementation is extremely computationally intensive. There is room for further improvement of bootstrap standard deviations in the case of inequality constraints.

Assumptions and limitations

The application of the theoretical standard errors and associated 95% t-confidence intervals is limited to the assumption of normally distributed error terms. It is well known that the assumption of normality does not always hold in psychological research (Micceri, 1989). In the context of FNM, each dissimilarity value typically represents the mean of the dissimilarity values obtained from N subjects, and by the central limit theorem it is to be expected that the mean dissimilarity values become approximately normally distributed fairly rapidly. A more challenging problem is the issue of correlated data: ”the future of linear models research lies primarily in developing methods for correlated data” (Christensen, 2002). The problem of dependency in the data has not been addressed in this monograph, but is likely to occur in the FNM because error terms associate with dissimilarity values that share the same objects are possibly correlated. The data used in the simulation studies were generated under the assumptions of normality, independence and homogeneity.

For the special case of FNM, the specification of the error structure might be a difficult task to accomplish in practice. There are specific experimental settings that inevitably produce data that yield correlated residuals, also called unobserved heterogeneity, as in longitudinal or multilevel data (*cf.* Skrondal & Rabe-Hesket, 2004). When individuals are clustered, for example students in classes, or when the same individuals are measured several times in a longitudinal setting, the residuals become correlated and the error structure is reasonably predictable. In the context of FNM, or more generally, dissimilarity matrices, the error structure is not precisely known. In addition, it is not clear how to assess the amount of dependency present in the data and the available tests are limited to specific settings, not comparable to the situation in FNM. The Durbin-Watson test (Durbin & Watson, 1950) is intended for autoregressive residuals and other tests like the Box test (Box, 1949) and the intraclass correlation (*cf.* Stevens, 1992) are useful to test the independence assumptions in the presence of several groups of individuals.

Given the difficulty to specify the error structure and the lack of adequate tests, the question comes up whether it is useful to adjust for correlated residuals a posteriori. Unlike the longitudinal studies, where dependency is inevitable, the FNM setting offers possibilities to reduce the occurrence of correlated error terms. For example, taking the mean of the dissimilarity values from a substantial number of subjects already reduces the correlation. Or one could use permutation techniques on the collected data matrices to disentangle the correlation structure between dissimilarity values that share the same object. More research is necessary to specify the error structure and to find adequate methods to prevent dependency in dissimilarity data.

If it is not possible to prevent dependency during the data collection step, one could use generalized least squares to take into account error correlation, but it would not solve the constrained estimation problem. The challenge is to combine inequality constrained least squares with generalized least squares. Very few attempts have been made to obtain estimates for the generalized inequality constrained least squares estimator, GICLS (Werner, 1990; Werner & Yapar, 1996). Assessing the variability of the GICLS estimator is still a far way to go, although Gulliksson and Wedin (2000) obtained some results in the perturbation theory for GICLS that are useful to assess the stability of the solution.

It should be noted that violations of the independence assumption not only affect parametric statistical inference (theoretical standard errors), but also affect the bootstrap. This means that the bootstrap method needs to be adjusted. Künsch (1989) and Liu and Singh (1992a) introduced the moving blocks bootstrap for use with dependent data. Gonçalves and White (2005) have further refined this method for the linear models with dependent data by establishing conditions for the consistency of the moving blocks bootstrap estimators of the variance of the least squares estimator. Using this bootstrap method might improve the assessment of bootstrap standard deviations for the feature discriminability parameters in FNM, although the results might not be the same for the ICLS estimator, and needs further research.

Another issue that has not been addressed in this monograph, is the problem of multiple confidence intervals. If confidence intervals are used to decide which features are important (especially the additive tree models have a large number of features), it eventually leads to the problem of multiple testing. A way out could be to use the Positive Lasso to select the best subset of features. In several applications in this monograph, the feature set selected by the Positive Lasso corresponded to the set of features with appropriate confidence intervals. There obviously exists a link between the two methods, although it has not been demonstrated yet. Furthermore, there are promising results available for the extension of the LARS algorithm to generalized linear models (see the discussion of Efron et al., 2004), which might be a solution to possible correlated error terms in the FNM. For use with the additive trees models, the Positive Lasso needs further adjustments because the feature structure is more restricted than in the general FNM. The Positive Lasso and the Lasso both have the additional advantage of being robust to correlated predictors, or multicollinearity.

6.2 Features and graphical representation

The set of distinctive features

The representation of features with Gray codes proved to be useful in several aspects. In a practical sense, the representation of features by the Gray code considerably simplifies computer manipulations of feature sets. The convenient attribute of Gray codes to represent features by a rank number (a simple integer) saves computer time and memory because the original feature set can be retrieved by simply keeping track of the corresponding rank number. Another advantage of the representation of features by a unique rank number is the possibility to get back the original features after transformations to featurewise distances, which is the transformation from the objects × features matrix $\mathbf{E}$ to the pairs of objects × features matrix $\mathbf{X}$. This transformation is not reversible because the results are not unique. The practical properties of the Gray code rank numbers are particularly useful in the generation of the feature sets, for efficient storage during Monte Carlo simulations, but also for efficient comparison of different tree topologies. In a more conceptual sense, the representation in Gray codes allows for defining a finite solution space, which can be further reduced to distinctive features only, through the transformation from **E** to **X**. This transformation limits the search for predictors to the set of truly distinctive features and increases the gain of using Gray coding.

Given that it is possible to generate the complete set of distinctive features for up to 22 objects in an efficient way, it is tempting to search for the optimal set of features for a given data set. Instead, the Positive Lasso selects a suboptimal solution that has better generalizability properties. This combination of the Positive Lasso algorithm and the complete set of distinctive features, has several additional advantages over the existing algorithms. It selects the best subset regardless of the number of features and avoids deciding on the number of features a priori. The selected subset is not biased toward a certain graphical representation because all possible feature structures are allowed. However, the method needs to be improved for data sets that have more than 22 objects or stimuli.

FNM is restricted to the use of distinctive features, which has some computational advantages as discussed in the previous paragraphs. In the introduction to this monograph, the distinctive features were presented as opposed to common features. The Contrast Model (Tversky, 1977) combines both types of features, but most of the models based on features that were developed later, mainly concentrate on one type of feature leading to common features models (CF) or distinctive features models (DF). The possibility to transform the CF model into the DF model and vice versa, has already been demonstrated by Sattath and Tversky (1987) and Carroll and Corter (1995). Chapter 5 further refined the transformation from CF to DF by showing that for any fitted CF model it is possible to find an equally well fitting DF model with the same set of shared features (common features) and associated feature weights, while keeping the same number of parameters. In this transformation, the CF model is a special case of the distinctive features model, which is a new result. Within this framework, a model that combines common and distinctive features can be represented as a sum of two separate DF models. However, the opposite transformation, from DF to CF is only possible if the objects are equidistant from the origin.

Network representation of features

Compared to all the models that are based on feature structuress, the graphical representation in terms of a network is unique for FNM. The network representation of features offers an interesting framework for theoretical comparison and practical use of several scaling and clustering methods. Figure 5.10 in Chapter 5 has shown that a whole family of discrete models of similarity are in fact special cases of the distinctive features model. The distinctive features models themselves are special cases of the city-block model and result from the restriction that the coordinate values be binary. Chapter 5 demonstrated that a coordinate system is not always necessary to represent city-block models. The additivity properties of the city-block distances allow for dropping the whole coordinate system and as a consequence, the dissimilarities can equally well be embedded in a network.

Embedding the network

To represent FNM, which are in general high-dimensional structures, as low dimensional feature graphs it is necessary to considerably reduce the dimensionality of the space. In this monograph, all network representations were obtained with multidimensional scaling performed with PROXSCAL[1]. The input distances were Euclidean distances computed with the feature discriminability parameters, and they were usually represented in 2-dimensional space. To represent the features in the same solutions space, PROXSCAL offers the possibility to constrain the solution space by a linear combination of the feature variables. As a result, the features can be represented as vectors leading from the origin through the point with coordinates equal to the correlations of each feature with each dimension. For ease of interpretation, the centroids of the objects that possess a particular feature can be projected onto the vector representing that feature. The same can be done for the objects that are not characterized by that feature. Labeling these projected points with plus and minus signs gives insight in the feature patterns of the objects.

The embedding of feature graphs in a lower dimensional space is for display purposes only because the model is specified by the network structure, the feature discriminability parameters and the model fit (the goodness-of-fit between the data and the reconstructed network distance). The fit between the network distance and the distance in the visual display, the embedding fit, is of secondary importance. The embedding is somewhat arbitrary because there are many possibilities to achieve this goal, depending on the distance transformations used or the type of start configuration used. In addition, the embedding is not restricted to the use of multidimensional scaling on the feature distances. Without using the feature distances, the objects and the features could also be represented in a biplot obtained with correspondance analysis.

Figure 6.1 shows an example of a plot representing the 14 features that characterize the presidents of the United States (the data were described in the introductory chapter), obtained with correspondance analysis, using row principal normalization. In row-principal normalization, a president point is located in the center of gravity of the features that he possesses. By connecting each president point with his own feature points, it is possible to reconstruct the feature graph from the correspondence analysis plot. However, in contrast to the feature network representation of the same data (See Figure 1.1, Chapter 1), the correspondence analysis plot does not allow a direct reconstruction of the distances between presidents. The strong point of the feature network representation results from the possibility to represent the feature structure as well as the distances between the objects by labeling the edges with the corresponding feature distances.

[1] PROXSCAL is a multidimensional scaling program distributed as part of the Categories package by SPSS, Meulman & Heiser, 1999

FNM and tree representations

Chapter 3 showed that given a special, nested, feature structure, formed by a combination of cluster features, unique features and internal nodes, the feature network representation becomes an additive tree representation. Do the results obtained for the additive trees also apply to hierarchical trees? Not directly, because in the hierarchical tree additional constraints are necessary to obtain equal distances to the root. In our context this means that extra constraints should be imposed on the feature discriminability parameters for the unique features. The problem could however be circumvented by using common features. The hierarchical tree can be obtained from a common feature model without unique features, which means that no extra constraints are necessary.

In the psychological literature, the relation between features and tree representations exists for a long time, while this relation is unknown in the phylogenetic tree domain. Both research areas might benefit from their mutual results. In particular, the use of features in FNM along with the univariate multiple regression framework led to two results that might be of interest for phylogenies. The first one is the possibility to use the *generalized cross-validation* statistic as an estimate of prediction error. This convenient closed form formula can be used to compare different tree topologies, even if both topologies have the same number of degrees of freedom. Being able to compare tree topologies with the same number of degrees of freedom, is an advantage over the likelihood ratio test that is commonly used to compare tree topologies but is limited to the case of nested topologies. The second result concerns the possibility of using cluster features to test in an easy way, events of speciation, the evolutionary process by which new biological species arise. Further improvements of statistical inference in the additive tree representations of FNM could be obtained by using the Positive Lasso to prune the tree, instead of using confidence intervals to select the relevant set of features. Pruning the tree will necessitate modifications of the present implementation of the Positive Lasso to simplify the tree structure in specific areas in order to keep the representation tree-shaped.

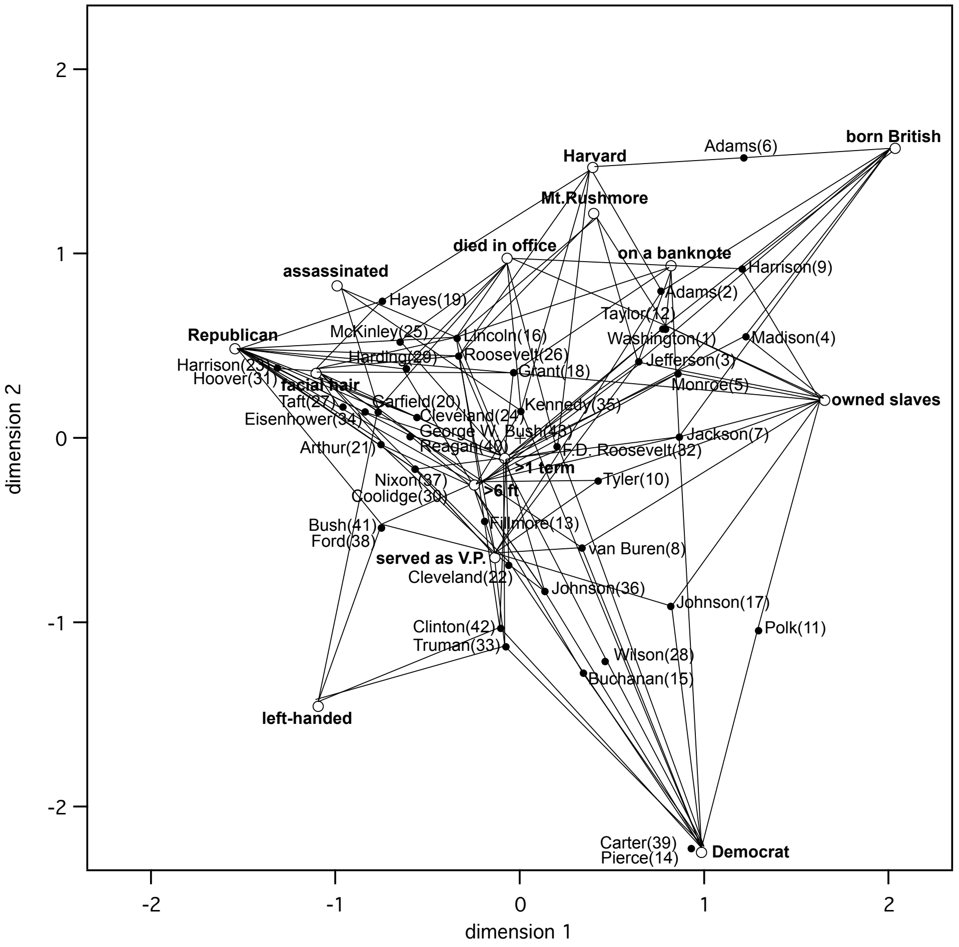

Figure 6.1: Biplot in 2 dimensions obtained with correspondence analysis of the 14 features describing the 43 presidents of the United States. The presidents are linked with the features they possess. (Normalization: row principal).