Feature Network Models for Proximity Data

Frank, Laurence Emmanuelle,
Feature Network Models for Proximity Data. Statistical inference, model selection, network representations and links with related models.
Dissertation Leiden University - With ref. - With summary in Dutch.
Subject headings: additive tree; city-block models; distinctive features models; feature models; feature network models; feature selection; Monte Carlo simulation; statistical inference under inequality constraints.

ISBN 90-8559-179-1

Printed by Optima, Rotterdam

Manuscript prepared in LaTeX (pdftex) with the TeX previewer TeXShop (v1.40), using the memoir document class (developed by P. Wilson) and the apacite package for APA style bibliography (developed by E. Meijer).

Feature Network Models for Proximity Data

Statistical inference, model selection, network representations and links with related models

PROEFSCHRIFT

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van de Rector Magnificus Dr. D.D. Breimer,
hoogleraar in de faculteit der Wiskunde en
Natuurwetenschappen en die der Geneeskunde,
volgens besluit van het College voor Promoties
te verdedigen op donderdag 21 september 2006
klokke 13.45 uur

door

Laurence Emmanuelle Frank

geboren te Delft
in 1969

PROMOTIECOMMISSIE

Promotor:	Prof. Dr. W. J. Heiser
Referent:	Prof. J. E. Corter, Ph.D., Columbia University, New York, USA
Overige leden:	Prof. Dr. I. Van Mechelen, K.U. Leuven, België Prof. Dr. V. J. J. P. van Heuven Prof. Dr. J. J. Meulman Prof. Dr. P. M. Kroonenberg

To my parents

”On ne peut se flatter d’avoir le dernier mot d’une théorie, tant qu’on ne peut pas l’expliquer en peu de paroles à un passant dans la rue.”

[It is not possible to feel satisfied at having said the last word about some theory as long as it cannot be explained in a few words to any passer-by encountered in the street.]

Joseph Diaz Gergonne, French mathematician (Chasles, 1875, p. 115).

Acknowledgements

A large number of persons contributed in several ways to this dissertation and I am indebted to them for their support.

I learned a lot about research in psychometrics from being a member of the Interuniversity Graduate School for Psychometrics and Sociometrics (IOPS). I would like to thank the IOPS-students for the agreeable time and a special thank to Marieke Timmerman, for her interest and the pleasant conversations. A special thank also to Susaññа Verdel for the enjoyable way we prepared the IOPS meetings, and for the wonderful way she organizes all practical IOPS-issues, always trying to offer the best possible conditions for staff and students.

I am very grateful for the opportunities given to me to attend conferences of the Psychometric Society and the International Federation of Classification Societies, which introduced me to the scientific community of our field of research and has been very inspiring for my own research.

I am greatly indebted to Prof. L.J. Hubert, Ph.D. (University of Illinois at Urbana-Champaign, USA) for helping me to implement the Dykstra algorithm in Matlab during his stay in Leiden, and for his useful comments on earlier versions of the second and third chapter of this dissertation.

For support on a more daily basis I would like to thank my colleagues of Psychometrics and Research Methodology and Data Theory Group (Universiteit Leiden): Bart Jan van Os for helping me with technical issues at crucial points during this research project, but also for his interest and the pleasant coffee breaks; Mark de Rooij for showing me how to do research in our field and how to accomplish a Ph.D. project; Mariëlle Linting for sharing a lot of nice conference experiences and hotel rooms during the whole project; Matthijs Warrens for the inspiring conversations about research; Marike Polak for the daily pleasant, encouraging conversations with lots of coffee and tea, and her sincere interest, which also holds for Rien van der Leeden.

The accomplishment of this dissertation would not have been possible without the love and support of my family and friends. To all who supported me during these years: many thanks for your friendship and all the joyous moments shared. It helped me to place this work in the right perspective.

Contents

List of Figures

List of Tables

Notation and Symbols

Notation conventions

matrices:	bold capital
vectors:	bold lowercase
scalars, integers:	lowercase

Symbols

Symbol	*Description*
O	an object or stimulus
m	the number of objects, stimuli
i	index $i = 1, \cdots, m$
j	index $j = 1, \cdots, m$
k	index $k = 1, \cdots, m$
n	the number of object pairs $= \frac{1}{2}m(m-1)$
l	index $l = 1, \cdots, n$
N	the number of replications of samples of size $n \times 1$
ℓ	index $\ell = 1, \cdots, N$
f	a frequency value associated with an object pair
δ	a dissimilarity value associated with an object pair
$\hat{\delta}$	an estimated dissimilarity value associated with an object pair
$\boldsymbol{\delta}$	an $n \times 1$ vector with dissimilarities between all object pairs
$\hat{\boldsymbol{\delta}}$	an $n \times 1$ vector with estimated dissimilarities between all object pairs
$\boldsymbol{\Delta}$	an $m \times m$ matrix with dissimilarities
$\widetilde{\Delta}_{l\ell}$	a random variable producing realisations $\tilde{\delta}_{l\ell}$
$\tilde{\delta}_{l\ell}$	a realisation of random variable $\widetilde{\Delta}_{l\ell}$
$\widetilde{\boldsymbol{\Delta}}$	an $n \times N$ matrix of random variables $\widetilde{\Delta}_{l\ell}$
$\overline{\Delta}_l$	mean of a row l of $\widetilde{\boldsymbol{\Delta}}$
ς	a similarity value associated with an object pair
$\boldsymbol{\varsigma}$	an $n \times 1$ vector with similarities between all object pairs
$\boldsymbol{\Sigma}_{\varsigma}$	an $m \times m$ matrix with similarities
F	a feature, which is a binary $(0,1)$ vector of size $m \times 1$
F_{C}	a *cluster* feature, which is a binary $(0,1)$ vector of size $m \times 1$
F_{U}	a *unique* feature, which is a binary $(0,1)$ vector of size $m \times 1$
T	the number of features

T_{C}	the number of *cluster* features
T_{U}	the number of *unique* features
$T_{\mathcal{D}}$	the total number of distinctive features $= \frac{1}{2}(2^m) - 1$
t	index for the features: $t = 1, \cdots, T$
t_{C}	index for the *cluster* features: $t_{\mathrm{C}} = 1, \cdots, T_{\mathrm{C}}$
t_{U}	index for the *unique* features: $t_{\mathrm{U}} = 1, \cdots, T_{\mathrm{U}}$
$\mathcal{S}_i$	the set of features that represents object O_i
$\mathbf{E}$	an $m \times T$ matrix with columns representing features
$\mathbf{e}$	a row vector from the matrix $\mathbf{E}$
e	an element of the matrix $\mathbf{E}$
$\mathbf{E}_T$	an $\mathbf{E}$ matrix with special feature structure that yields a tree representation
$\mathbf{E}_C$	the part of $\mathbf{E}_T$ (size $m \times T_{\mathrm{C}}$) that represents the set of *cluster* features
$\mathbf{E}_U$	the part of $\mathbf{E}_T$ (size $m \times T_{\mathrm{U}}$) that represents the set of *unique* features
$\mathbf{X}$	an $n \times T$ matrix with featurewise distances obtained with $\mathbf{x}' = \lvert \mathbf{e}_{it} - \mathbf{e}_{jt} \rvert$
$\mathbf{x}'$	a row vector from the matrix $\mathbf{X}$
$\mathbf{x}$	a column vector from the matrix $\mathbf{X}$
$\mathbf{X}_T$	an $n \times T_{\mathrm{C}} + T_{\mathrm{U}}$ matrix with featurewise distances obtained with $\mathbf{E}_T$
$\mathcal{D}$	the complete set of featurewise distances
d	a distance between an object pair
$\mathbf{d}$	an $n \times 1$ vector of distances between all object pairs
$\hat{\mathbf{d}}$	an $n \times 1$ vector of estimated distances between all object pairs
$\hat{\mathbf{d}}_{\mathrm{T}}$	an $n \times 1$ vector of estimated distances between all object pairs for a tree structure
η	feature discriminability parameter
η_{OLS}	true value of ordinary least squares feature discriminability parameter
η_{ICLS}	true value of inequality constrained least squares feature discriminability parameter
η_{L}	true value of Lasso feature discriminability parameter
η_{PL}	true value of Positive Lasso feature discriminability parameter
$\boldsymbol{\eta}$	an $T \times 1$ vector of feature discriminability parameters
$\boldsymbol{\eta}_{\mathrm{OLS}}$	an $T \times 1$ vector of true values η_{OLS}
$\boldsymbol{\eta}_{\mathrm{ICLS}}$	an $T \times 1$ vector of true values η_{ICLS}
$\boldsymbol{\eta}_{\mathrm{L}}$	an $T \times 1$ vector of true values η_{L}
$\boldsymbol{\eta}_{\mathrm{PL}}$	an $T \times 1$ vector of true values η_{PL}
$\hat{\eta}$, $\hat{\eta}_{\mathrm{OLS}}$	estimated values of η, η_{OLS}, η_{ICLS}, η_{L}, η_{PL}
C	the number of constraints necessary to obtain $\hat{\boldsymbol{\eta}}_{\mathrm{ICLS}}$
c	index $c = 1, \cdots, C$
$\mathbf{r}$	a $C \times 1$ vector with constraints
$\mathbf{A}$	a $C \times T$ matrix of constraints of rank c
$\boldsymbol{\lambda}_{\mathrm{KT}}$	a $m \times 1$ vector with Kuhn-Tucker mutipliers
$\boldsymbol{\epsilon}$	a $n \times 1$ vector with error values ($\boldsymbol{\epsilon} = \boldsymbol{\delta} - \mathbf{X}\boldsymbol{\eta}$)
$\hat{\boldsymbol{\epsilon}}$	a $n \times 1$ vector with estimated error values ($\hat{\boldsymbol{\epsilon}} = \boldsymbol{\delta} - \mathbf{X}\hat{\boldsymbol{\eta}}$)
$\hat{\epsilon}$	an element from the vector $\hat{\boldsymbol{\epsilon}}$
σ^2, σ	true variance and standard deviation of $\boldsymbol{\epsilon}$
$\hat{\sigma}^2$, $\hat{\sigma}$	estimated variance and standard deviation of $\hat{\boldsymbol{\epsilon}}$
σ^2_{η}, σ_{η}	true variance and standard error of η
$\hat{\sigma}^2_{\eta}$, $\hat{\sigma}_{\eta}$	estimated nominal variance and estimated nominal standard error of η
$\hat{\sigma}^2_{\hat{\eta}}$, $\hat{\sigma}_{\hat{\eta}}$	estimated nominal variance and nominal standard error of $\hat{\eta}$
σ^2_{OLS}, σ_{OLS}	true variance and standard error of $\hat{\eta}_{\mathrm{OLS}}$

$\hat{\sigma}^2_{\mathrm{OLS}}, \hat{\sigma}_{\mathrm{OLS}}$	estimated variance and standard error of $\hat{\eta}_{\mathrm{OLS}}$
$\sigma^2_{\mathrm{ICLS}}, \sigma_{\mathrm{ICLS}}$	true variance and standard error of $\hat{\eta}_{\mathrm{ICLS}}$
$\hat{\sigma}^2_{\mathrm{ICLS}}, \hat{\sigma}_{\mathrm{ICLS}}$	estimated variance and standard error of $\hat{\eta}_{\mathrm{ICLS}}$
B	number of bootstrap samples
b	index $b = 1, \cdots, B$
$\mathbf{b}_b$	a bootstrap sample ($n \times 1$ vector)
$\mathbf{b}^*_b$	a bootstrap sample, multivariate
$\tilde{\mathbf{b}}_b$	a bootstrap sample, with sampled residuals
sd_B	standard deviation of B bootstrap samples
S	number of simulation samples
a	index $a = 1, \cdots, S$
$\mathbf{s}^*$	a simulation sample ($n \times 1$ vector)
κ, p	parameters binomial distribution
GCV	generalized cross-validation statistic
GCV_{FNM}	GCV using inequality constrained least squares estimation