technische universität
dortmund

fakultät
statistik

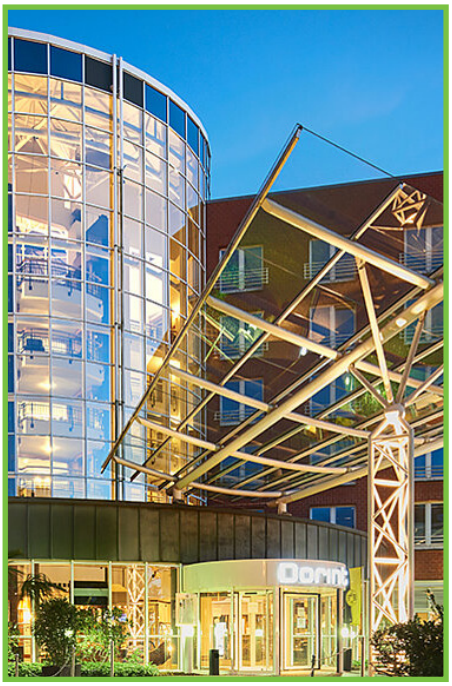
IWSM 2023

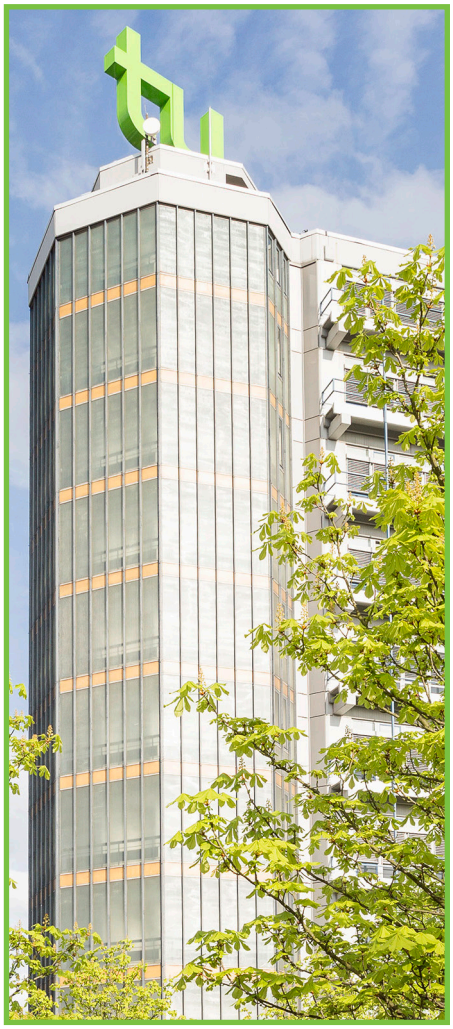
37th International Workshop on Statistical Modelling

16.07. – 21.07.2023

Dortmund

Proceedings book





Proceedings of the 37th International Workshop on Statistical Modelling

July 17-21, 2023 - Dortmund, Germany

Editors
Elisabeth Bergherr
Andreas Groll
Andreas Mayr

TU Dortmund University

International Workshop on Statistical Modelling (37°. 2023. Dortmund)

Proceedings of the 37th International Workshop on Statistical Modelling : July 17-21, 2022
Dortmund, Germany / Elisabeth Bergherr, Andreas Groll, Andreas Mayr (editors). – Dortmund : TU Dortmund University, 2023. – 1 copy online : PDF (693 S. : ill.)

ISBN: 978-3-947323-42-5

Authors:

Bergherr, Elisabeth
Groll, Andreas
Mayr, Andreas

Topics:

1. Statistics congress. 2. Econometrics models congress
330.015195 = Mathematical statistics

Editors:

ELISABETH BERGHERR
University of Göttingen, Chair of Spatial Data Science and Statistical Learning

ANDREAS GROLL
TU Dortmund University, Department of Statistics

ANDREAS MAYR
University of Bonn, Department of Medical Biometry, Informatics and Epidemiology

ISBN 978-3-947323-42-5 (online)

TU Dortmund University
Department of Statistics
Vogelpothsweg 78
44227 Dortmund
Germany

https://ub.tu-dortmund.de/
https://statistik.tu-dortmund.de/

Scientific Committee

Ruggero Bellio

University of Udine (Italy)

Elisabeth Bergherr (Co-Chair)

University of Göttingen (Germany)

Fernanda De Bastiani

University of Pernambuco (Brazil)

María L. Durbán Reguera

University of Madrid (Spain)

Jan Gertheiss

Helmut Schmidt University, Hamburg (Germany)

Andreas Groll (Chair)

TU Dortmund (Germany)

Thomas Kneib

University of Göttingen (Germany)

Dae-Jin Lee

IE University, School of Science and Technology, Madrid (Spain)

Andreas Mayr (Co-Chair)

University of Bonn (Germany)

Fulvia Pennoni

University of Milano-Bicocca (Italy)

María Xosé Rodríguez Álvarez

University of Vigo (Spain)

Gunther Schauberger

TU München (Germany)

Nicola Torelli

University of Trieste (Italy)

Lola Ugarte

University of Navarra (Spain)

Nikolaus Umlauf

University of Innsbruck, (Austria)

Helga Wagner

University of Linz, (Austria)

Local Organising Committee

Chiara Balestra

TU Dortmund University

Elisabeth Bergherr (Co-Host)

University of Göttingen

Guillermo B. Sánchez

TU Dortmund University

Jennifer Engel

TU Dortmund University

Alexander Gerharz

TU Dortmund University

Colin Griesbach

University of Göttingen

Andreas Groll (Host)

TU Dortmund University

Tobias Hepp

University Erlangen-Nürnberg

Hannah Klinkhammer

University of Bonn

Andreas Mayr (Co-Host)

University of Bonn

Hendrik van der Wurp

TU Dortmund University

Evolutionary algorithm for the estimation of discrete latent variables models

Luca Brusa[1], Fulvia Pennoni[1], Francesco Bartolucci[2]

[1] Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Italy
[2] Department of Economics, University of Perugia, Italy

E-mail for correspondence: luca.brusa@unimib.it

Abstract: The expectation-maximization (EM) algorithm is the most common iterative method employed for maximum likelihood estimation of discrete latent variable models. A common drawback of this estimation method, along with its variant named variational EM (VEM), is that it may be trapped into one of the multiple local maxima of the log-likelihood function. We propose a version of the algorithm based on the evolutionary approach, which allows us to explore the parameter space accurately. The proposal is validated through a Monte Carlo simulation study aimed at comparing its performance with the EM and VEM algorithms by estimating latent class, hidden Markov, and stochastic block models. Results show a significant increase in the chance of reaching a global maximum for the proposed evolutionary EM. The efficacy of the proposal is also validated by applications using longitudinal data on countries' energy production and interactions between karate club members.

Keywords: Expectation-maximization algorithm; Global optimization; Local maxima; Maximum likelihood estimation.

1 Introduction

Discrete latent variable (DLV) models have attracted much attention in statistical literature since they are formulated according to latent variables having a discrete distribution left unspecified. Among others, they ensure a high degree of flexibility in modelling complex dependence data structures (Bartolucci et al., 2022). Maximum likelihood estimation of DLV models is usually performed through the expectation-maximization (EM) algorithm (Dempster et al., 1977). When the latter approach is computationally unfeasible, a variational modification, namely the variational EM

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

(VEM) algorithm (Jordan et al., 1999), represents a popular alternative. A well-known drawback of both estimation methods is related to the multimodality of the likelihood function, resulting in a potential convergence of the algorithm to a local maximum. We propose an extension of the EM algorithm named evolutionary EM (EEM), defined according to the evolutionary algorithm (EA) approach (Ashlock, 2004). At each step of the EEM algorithm, multiple sets of parameters are evaluated according to a quality measure, while evolutionary operators, such as crossover and mutation, ensure an accurate parameter space exploration.

2 Discrete latent variable framework

The key idea of DLV models is to associate observed responses to latent variables according to a joint probability model. Denoting by $\mathbf{Y}$ and $\mathbf{U}$ the sets of observed responses and latent variables, respectively, a DLV model is characterized by the conditional distribution of the responses given the latent variables, and by the distribution of the latent variables.
The EM algorithm maximizes the observed-data log-likelihood function $\ell(\boldsymbol{\theta})$, expressed in terms of model parameters $\boldsymbol{\theta}$, relying on the complete-data log-likelihood function $\ell^*(\boldsymbol{\theta})$. Once the model parameters have been initialized, the algorithm alternates two steps until convergence: (i) an expectation step, where the conditional expected value of $\ell^*(\boldsymbol{\theta})$ is computed given the value of the parameters at the previous step and the observed data, and (ii) a maximization step, where the model parameters are updated by maximizing the expected value of $\ell^*(\boldsymbol{\theta})$.
The VEM algorithm defines instead a lower bound $\mathcal{J}(\boldsymbol{\theta})$ for the observed-data log-likelihood function, to be maximized instead of $\ell(\boldsymbol{\theta})$. To explore the parameter space, the choice of multiple sets of starting values for the model parameters is crucial. The maximum is then taken as the solution corresponding to the largest likelihood value at convergence. Drawbacks of this strategy are the high computational time and the fact that the convergence may be to one of local maxima different from the global one.

3 Evolutionary expectation-maximization algorithm

Following the EA approach, the proposed EEM algorithm is inspired by the Darwinian theory of evolution principles. According to Pernkopf and Bouchaffra (2005), it takes into account an initial "population" P_0 of N_P potential solutions for the optimization problem at issue. Each element of P_0 is a different candidate array of posterior probabilities. The following steps are then alternated until convergence:

1. $P_1 \leftarrow \mathbf{Update}(P_0)$: population P_0 is updated by performing a small number of cycles of the standard EM algorithm with random initialization on each individual, resulting in a new population P_1.

2. $P_2 \leftarrow$ **Crossover**(P_1): pairs of individuals from population P_1 are randomly selected and recombined by swapping corresponding blocks of their arrays. We obtain the N_O offspring of the new population P_2.

3. $P_3 \leftarrow$ **Update**(P_2): population P_2 is updated by performing a small number of cycles of the standard EM algorithm with random initialization on each individual, resulting in the new population P_3.

4. $P_4 \leftarrow$ **Selection**$(P_1 \cup P_3)$: individuals from populations P_1 and P_3 are considered jointly, and the N_P with the highest value of the log-likelihood function are selected for the next generation P_4.

5. $P_5 \leftarrow$ **Mutation**(P_4): variation is introduced to each individual of population P_4 (apart from the best one): given a row of the corresponding array of posterior probabilities, mutation operator swaps the highest value with a random one.

Convergence of the EEM algorithm is measured focusing only on the best solution of population P_4 and analyzing both the relative difference of the log-likelihood of two consecutive steps and that between the corresponding parameter vectors.

4 Simulation studies

To evaluate the performance of the EEM algorithm, we rely on a Monte Carlo simulation study considering latent class (LC), hidden Markov (HMcat and HMcont for categorical and continuous response variables, respectively), and stochastic block (SB) models. This study is based on different scenarios for each model, depending on several features: sample size ($n = 500, 1000$), number of response variables ($r = 6, 12$), response categories ($c = 3, 6$), time occasions ($T = 5, 10$), and latent components ($k = 3, 6$). Concerning the SB model we also distinguish two different behaviors: one defined as assortative with high intra-group and low inter-group connection probabilities and the other as disassortative with low intra-group and high inter-group probabilities. For each scenario the corresponding model is applied 100 times to 50 samples using the EM and EEM algorithms. Both correctly specified and misspecified latent structures are estimated in order to compare the performance of the algorithms through the following criteria.

Global maximum achievement: considering the highest of the maximized log-likelihood values as the global maximum $\hat{\ell}_{\text{MAX}}$, we denote a generic log-likelihood value at convergence as $\hat{\ell}$ and compute the percentage of $\hat{\ell}$ such that $(\hat{\ell}_{MAX} - \hat{\ell})/|\hat{\ell}_{MAX}| < \tilde{\varepsilon}$, where $\tilde{\varepsilon}$ is a suitable threshold. The EEM algorithm performs better in each simulated scenario, significantly increasing the chance to reach the global maximum. Some results of 2 of

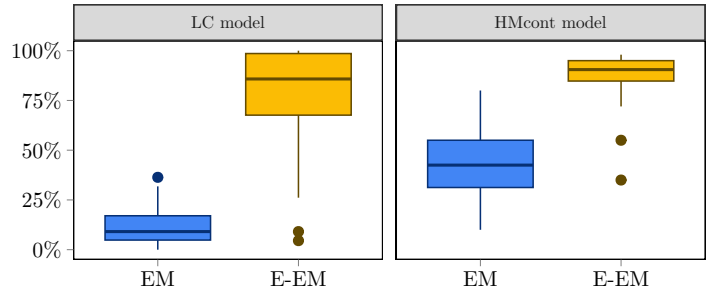


FIGURE 1. Percentages of global maxima reached using EM and EEM algorithms for (i) a correctly specified LC model with $n = 500$, $r = 6$, $c = 3$, and $k = 6$, and (ii) a misspecified HMcont model with $n = 500$, $r = 12$, $T = 5$, and $k = 3$.

the 22 simulated scenarios are depicted in Figure 1. In particular, regarding the estimation of models whose latent structure is correctly specified, the frequency of convergence to the global mode is usually very close to 100%, highlighting that it generally tends to avoid convergence to a local maximum of $\ell(\boldsymbol{\theta})$. The results of the extensive simulation study highlight that the proposal always outperforms the EM algorithm; the improvement is especially evident with many latent components and under scenarios related to the SB model. Its performance is even more remarkable considering models with misspecified latent structures. In this case, while the standard EM algorithm sometimes proves unable to locate the global maximum, the evolutionary approach is always able to correctly detect it, improving the value itself of the global mode, in addition to the chance to reach it.

Average distance from the global maximum: using the EEM algorithm, the distance between each maximum and the global one is quite low for all the examined scenarios. The average distance obtained through the EM algorithm is usually considerably higher. We mention for instance one scenario of the LC model in which the average distance decreases from $4.7 \cdot 10^{-7}$ using the EM algorithm to $2.5 \cdot 10^{-18}$ using the EEM algorithm. In scenarios related to the HMcat model the distance is still reduced by half with the EEM algorithm, dropping, for example, from $1.2 \cdot 10^{-3}$ to $6.8 \cdot 10^{-4}$.

Accurate parameters estimation: dealing with correctly specified models, we also provide the root mean square error (RMSE) between the true and estimated model parameters. Results show the RMSEs obtained with the EEM algorithm are very close to zero under all the simulated scenarios; on the contrary, values obtained with the EM algorithm are always larger, approaching one in some cases. This shows that the evolutionary approach entails a significantly greater accuracy. In particular, the improvement is especially evident when the HMcont and SB models are estimated.

5 Applications

The EEM algorithm is also evaluated to estimate LC, HMcat, HMcont, and SB models with cross-sectional, longitudinal, and network data.
In the following, as a first application, we use longitudinal data measuring the sources of electricity generation in 27 European Union countries (data are available at the link https://ourworldindata.org/energy). A multivariate time homogeneous HMcont model is considered for response variables collected yearly from 2011 to 2020 and referred to the share of electricity deriving from biofuel, coal, natural gas, hydroelectric, nuclear, oil, solar, and wind. Logit and Box-Cox transformations are applied to all the variables. The model is estimated for a number of states ranging from 1 to 12 with both the EM and EEM 100 times. A model with 8 latent states representing sub-populations of countries with similar energetic behaviour is selected according to the Bayesian information criterion. The EEM algorithm ensures convergence to the global maximum, corresponding to a value of the log-likelihood function equal to $-5,452$. The EM algorithm never detects such a maximum, providing $-5,574$ as the highest value for the log-likelihood function at convergence. The estimation with the EEM also provides a reasonable posterior dynamic classification of the countries into groups, while EM does not. Table 1 reports the estimated conditional means of the responses given the latent state. Groups are ordered from the lowest to the highest average value of wind power. Countries in the 1st group are using mainly nuclear power, in the 2nd are predominantly coal-dependent, in the 3rd heavily rely on oil, in the 4th they use a mix of coal, oil and gas, along with the highest average of solar energy. Countries in the 5th state are using mainly gas, in the 6th they use gas and a quota of biofuel over all the other groups, in the 7th they excels in hydroelectric power, and in the 8th they use mainly wind energy along with nuclear power.
As a second application, we estimate the SB model with network data on 34 karate club members (data are available in the `R` package `igraphdata`).

TABLE 1. Estimated means of the HMcont model with $k = 8$ latent states for the European Union countries electricity data.

	Latent states							
Source	1	2	3	4	5	6	7	8
Coal	5.58	41.91	4.52	38.52	14.86	0.00	20.57	15.10
Oil	3.62	2.85	51.80	10.77	6.62	3.97	3.55	3.04
Gas	5.34	12.24	7.20	24.53	57.43	44.52	21.06	19.87
Nuclear	50.21	19.96	0.00	0.00	1.42	0.00	13.20	31.94
Biofuel	7.68	4.55	4.77	0.53	3.17	12.35	2.92	9.65
Hydro	22.12	9.85	24.50	8.96	1.12	21.16	19.05	0.39
Solar	0.90	3.41	1.70	6.62	2.51	3.34	2.89	2.48
Wind	4.52	4.87	5.48	10.08	12.87	14.66	16.63	17.53

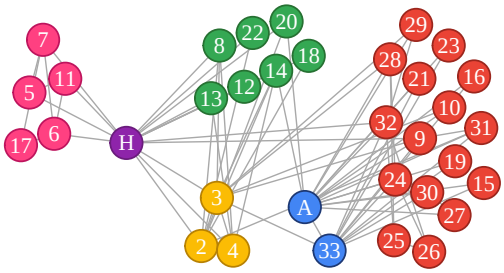


FIGURE 2. Graph visualization with nodes colored by estimated partition for the SB model with $k = 6$ latent blocks for the karate club data.

Relationships among members are measured by a 34×34 adjacency binary matrix. Using the EEM algorithm, an SB model with $k = 6$ latent blocks is selected according to the integrated classification likelihood criterion. The EEM algorithm consistently converges to a log-likelihood function value equal to -277.91; if the model is estimated with the EM algorithm its highest value is -316.46. Figure 2 shows the network with nodes colored by the estimated partition. The model correctly identifies positions taken for president (**A**, in blue) or instructor (**H**, in violet). The faction led by the president consists of a single additional latent block (in red), presenting a high connection probability with its leader (equal to 0.75). The remaining three latent blocks (depicted in pink, green, and yellow) constitute the faction led by the instructor; each of these blocks has a high connection probability with their leader (equal to 0.81, 1.00, and 1.00, respectively). Connection probabilities between blocks of different factions are very low (0.17 at most).

References

Ashlock, D. (2004). *Evolutionary Computation for Modeling and Optimization*. New York: Springer.

Bartolucci, F., Pandolfi, S. and Pennoni, F. (2022). Discrete Latent Variable Models. *Annual Review of Statistics and its Application*, **6**, 1 – 31.

Jordan, M.I., Ghahramani, Z., Jaakkola, T.S., and Saul, L.K. (1999). An Introduction to Variational Methods for Graphical Models. *Machine Learning*, **37**, 183 – 233.

Dempster, A., Laird, N. and Rubin, D. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm (with Discussion). *Journal of the Royal Statistical Society, Series B*, **39**, 1 – 38.

Pernkopf, F. and Bouchaffra, D. (2005). Genetic-Based EM Algorithm for Learning Gaussian Mixture Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **27**, 1344 – 1348.