A very brief guide to using MXM

Michail Tsagris, Vincenzo Lagani, Ioannis Tsamardinos

1 Introduction

MXM is an R package which contains functions for **feature selection**, **cross-validation** and **Bayesian Networks**. The main functionalities focus on feature selection for different types of data. We highlight the option for parallel computing and the fact that some of the functions have been either partially or fully implemented in C++. As for the other ones, we always try to make them faster.

2 Feature selection related functions

MXM offers many feature selection algorithms, namely MMPC, SES, MMMB, FBED, forward and backward regression. The target set of variables to be selected, ideally what we want to discover, is called Markov Blanket and it consists of the parents, children and parents of children (spouses) of the variable of interest assuming a Bayesian Network for all variables.

MMPC stands for Max-Min Parents and Children. The idea is to use the Max-Min heuristic when choosing variables to put in the selected variables set and proceed in this way. Parents and Children comes from the fact that the algorithm will identify the parents and children of the variable of interest assuming a Bayesian Network. What it will not recover is the spouses of the children of the variable of interest. For more information the reader is addressed to (Tsamardinos et al., 2006).

MMMB (Max-Min Markov Blanket) extends the MMPC to discovering the spouses of the variable of interest (Tsamardinos et al., 2003a). SES (Statistically Equivalent Signatures) on the other hand extends MMPC to discovering statistically equivalent sets of the selected variables (Tsamardinos et al., 2012; Lagani et al., 2017). Forward and Backward selection are the two classical procedures.

The functionalities or the flexibility offered by all these algorithms is their ability to handle many types of dependent variables, such as continuous, survival, categorical (ordinal, nominal, binary), longitudinal. Let us now see all of them one by one. The relevant functions are

1. **MMPC and SES**. SES uses MMPC to return multiple statistically equivalent sets of variables. MMPC returns only one set of variables.

In all cases, the log-likelihood ratio test is used to assess the significance of a variable. These algorithms accept categorical only, continuous only or mixed data in the predictor variables side.

2. **wald.mmpc and wald.ses**. SES uses MMPC using the Wald test. These two algorithms accept continuous predictor variables only.

3. **perm.mmpc and perm.ses**. SES uses MMPC where the p-value is obtained using permutations. Similarly to the Wald versions, these two algorithms accept continuous predictor variables only.

4. **ma.mmpc and ma.ses**. MMPC and SES for multiple datasets measuring the same variables (dependent and predictors).

5. **MMPC.temporal and SES.temporal**. Both of these algorithms are the usual SES and MMPC modified for correlated data, such as clustered or longitudinal. The predictor variables can only be continuous.

6. **fbed.reg**. The FBED feature selection method (Borboudakis and Tsamardinos, 2017). The log-likelihood ratio test or the eBIC (BIC is a special case) can be used.

7. **ebic.bsreg**. Backward selection method using the eBIC.

8. **fs.reg**. Forward regression method for all types of predictor variables and for most of the available tests below.

9. **glm.fsreg** Forward regression method for logistic and Poisson regression in specific. The user can call this directly if he knows his data.

10. **lm.fsreg**. Forward regression method for normal linear regression. The user can call this directly if he knows his data.

11. **bic.fsreg**. Forward regression using BIC only to add a new variable. No statistical test is performed.

12. **bic.glm.fsreg**. The same as before but for linear, logistic and Poisson regression (GLMs).

13. **bs.reg**. Backward regression method for all types of predictor variables and for most of the available tests below.

14. **glm.bsreg**. Backward regression method for linear, logistic and Poisson regression (GLMs).

15. **iamb**. The IAMB algorithm Tsamardinos et al. (2003b) which stands for Incremental Association Markov Blanket. The algorithm performs a forward regression at first, followed by a backward regression offering

two options. Either the usual backward regression is performed or a faster variation, but perhaps less correct variation. In the usual backward regression, at every step the least significant variable is removed. In the IAMB original version all non significant variables are removed at every step.

16. **mmmb**. This algorithm works for continuous or categorical data only. After applying the MMPC algorithm one can go to the selected variables and perform MMPC on each of them.

A list with the available options for this argument is given below. Make sure you include the test name within "" when you supply it. Most of these tests come in their Wald and perm (permutation based) versions. In their Wald or perm versions, they may have slightly different acronyms, for example **waldBinary** or **WaldOrdinal** denote the logistic and ordinal regression respectively. Another example is **waldMMreg** which denotes the MM (robust) regression. This test is available in **testIndReg** with the option **robust = TRUE**.

1. **testIndFisher**. This is a standard test of independence when both the target and the set of predictor variables are continuous (continuous-continuous).

2. **testIndSpearman**. This is a non-parametric alternative to **testIndFisher** test (Fieller and Pearson, 1961).

3. **testIndReg**. In the case of target-predictors being continuous-mixed or continuous-categorical, the suggested test is via the standard linear regression. If the robust option is selected, M estimators (Maronna et al., 2006) are used. If the target variable consists of proportions or percentages (within the (0, 1) interval), the logit transformation is applied beforehand.

4. **testIndRQ**. Another robust alternative to **testIndReg** for the case of continuous-mixed (or continuous-continuous) variables is the **testIndRQ**. If the target variable consists of proportions or percentages (within the (0, 1) interval), the logit transformation is applied beforehand.

5. **testIndBeta**. When the target is proportion (or percentage, i.e., between 0 and 1, not inclusive) the user can fit a regression model assuming a beta distribution (Ferrari and Cribari-Neto, 2004). The predictor variables can be either continuous, categorical or mixed.

6. **testIndPois**. When the target is discrete, and in specific count data, the default test is via the Poisson regression. The predictor variables can be either continuous, categorical or mixed.

7. **testIndNB**. As an alternative to the Poisson regression, we have included the Negative binomial regression to capture cases of overdispersion (Hilbe, 2011). The predictor variables can be either continuous, categorical or mixed.

8. **testIndZIP**. When the number of zeros is more than expected under a Poisson model, the zero inflated poisson regression is to be employed (Lambert, 1992). The predictor variables can be either continuous, categorical or mixed.

9. **testIndLogistic** (Binomial). When the target is categorical with only two outcomes, success or failure for example, then a binary logistic regression is to be used. Whether regression or classification is the task of interest, this method is applicable. The advantage of this over a linear or quadratic discriminant analysis is that it allows for categorical predictor variables as well and for mixed types of predictors.

10. **testIndLogistic** (Un-ordered multinomial). If the target has more than two outcomes, but it is of nominal type (political party, nationality, preferred basketball team), there is no ordering of the outcomes, multinomial logistic regression will be employed. Again, this regression is suitable for classification purposes as well and it to allows for categorical predictor variables. The predictor variables can be either continuous, categorical or mixed.

11. **testIndLogistic** (Ordered multinomial). This is a special case of multinomial regression, in which case the outcomes have an ordering, such as **not satisfied**, **neutral**, **satisfied**. The appropriate method is ordinal logistic regression. The predictor variables can be either continuous, categorical or mixed.

12. **testIndTobit** (Tobit regression for left censored data). Suppose you have measurements for which values below some value were not recorded. These are left censored values and by using a normal distribution we can by pass this difficulty. The predictor variables can be either continuous, categorical or mixed.

13. **testIndBinom**. When the target variable is a matrix of two columns, where the first one is the number of successes and the second one is the number of trials, binomial regression is to be used. The predictor variables can be either continuous, categorical or mixed.

14. **testIndSpeedglm**. If you have a few tens of thousands of observations, the default functions for linear, binary logistic and poisson regression will be slow causing the computer to jam. For this reason, memory efficient handling regressions should be used. The predictor variables can be either continuous, categorical or mixed.

15. **gSquare**. If all variables, both the target and predictors are categorical the default test is the G^2 test of independence. An alternative to the **gSquare** test is the **testIndLogistic**. With the latter, depending on the nature of the target, binary, un-ordered multinomial or ordered multinomial the appropriate regression model is fitted. The predictor variables can be either continuous, categorical or mixed.

16. **censIndCR**. For the case of time-to-event data, a Cox regression model (Cox, 1972) is employed. The predictor variables can be either continuous, categorical or mixed.

17. **censIndWR**. A second model for the case of time-to-event data, a Weibull regression model is employed (Smith, 1991; Scholz, 1996). Unlike the semi-parametric Cox model, the Weibull model is fully parametric. The predictor variables can be either continuous, categorical or mixed.

18. **censIndER**. A third model for the case of time-to-event data, an exponential regression model is employed. The predictor variables can be either continuous, categorical or mixed. This is a special case of the Weibull model.

19. **testIndIGreg**. When you have non negative data, i.e. the target variable takes positive values (including 0), a suggested regression is based on the the inverse Gaussian distribution. The link function is not the inverse of the square root as expected, but the logarithm. This is to ensure that the fitted values will be always be non negative. An alternative model is the Weibull regression (**censIndWR**). The predictor variables can be either continuous, categorical or mixed.

20. **testIndGamma** (Gamma regression). Gamma distribution is designed for strictly positive data (greater than zero). It is used in reliability analysis, as an alternative to the Weibull regression. This test however does not accept censored data, just the usual numeric data. The predictor variables can be either continuous, categorical or mixed.

21. **testIndNormLog** (Gaussian regression with a log link). Gaussian regression using the log link (instead of the identity) allows non negative data to be handled naturally. Unlike the gamma or the inverse gaussian regression zeros are allowed. The predictor variables can be either continuous, categorical or mixed.

22. **testIndClogit**. When the data come from a case-control study, the suitable test is via conditional logistic regression (Gail et al., 1981). The predictor variables can be either continuous, categorical or mixed.

23. **testIndMVReg**. In the case of multivariate continuous targets, the suggested test is via a multivariate linear regression. The target variable can be compositional data as well (Aitchison, 1986). These are positive data, whose vectors sum to 1. They can sum to any constant, as long as it the same, but for convenience reasons we assume that they are normalised to sum to 1. In this case the additive log-ratio transformation (multivariate logit transformation) is applied beforehand. The predictor variables can be either continuous, categorical or mixed.

24. **testIndGLMM**. In the case of a longitudinal or clustered targets (continuous, proportions within 0 and 1 (not inclusive), binary or counts), the suggested test is via a (generalised) linear mixed model (Pinheiro and Bates, 2006). The predictor variables can only be continuous. This test is only applicable in SES.temporal and MMPC.temporal.

To avoid any mistakes or wrongly selected test by the algorithms you are advised to select the test you want to use. All of these tests can be used with SES and MMPC, forward and backward regression methods. MMMB accepts only **testIndFisher**, **testIndSpearman** and **gSquare**. The reason for this is that MMMB was designed for variables (dependent and predictors) of the same type. For more info the user should see the help page of each function.

2.1 A more detailed look at some arguments of the feature selection algorithms

SES, MMPC, MMMB, forward and backward regression offer the option for robust tests (the argument *robust*). This is currently supported for the case of Pearson correlation coefficient and linear regression at the moment. We plan to extend this option to binary logistic and Poisson regression as well. These algorithms have an argument *user_test*. In the case that the user wants to use his own test, for example, *mytest*, he can supply it in this argument as is, without "". For all previously mentioned regression based conditional independence tests, the argument works as *test="testIndFisher"*. In the case of the *user_test* it works as *user_test=mytest*. The *max_k* argument must always be at least 1 for SES, MMPC and MMMB, otherwise it is a simple filtering of the variables. The argument *ncores* offers the option for parallel implementation of the first step of the algorithms. The filtering step, where the significance of each predictor is assessed. If you have a few thousands of variables, maybe this option will do no significant improvement. But, if you have more and a "difficult" regression test, such as quantile regression (**testIndRQ**), then with 4 cores this could reduce the computational time of the first step up to nearly 50%. For the Poisson, logistic and normal linear

regression we have included C++ codes to speed up this process, without the use of parallel.

The FBED (Forward Backward Early Dropping) is a variant of the Forward selection is performed in the first phase followed by the usual backward regression. In some, the variation is that every non significant variable is dropped until no mre significant variables are found or there is no variable left.

The forward and backward regression methods have a few different arguments.For example *stopping* which can be either "BIC" or "adjrsq", with the latter being used only in the linear regression case. Every time a variable is significant it is added in the selected variables set. But, it may be the case, that it is actually not necessary and for this reason we also calculate the BIC of the relevant model at each step. If the difference BIC is less than the *tol* (argument) threshold value the variable does not enter the set and the algorithm stops.

The forward and backward regression methods can proceed via the BIC as well. At every step of the algorithm, the BIC of the relevant model is calculated and if the BIC of the model including a candidate variable is reduced by more that the *tol* (argument) threshold value that variable is added. Otherwise the variable is not included and the algorithm stops.

2.2 Other relevant functions

Once SES or MMPC are finished, the user might want to see the model produced. For this reason the functions **ses.model** and **mmpc.model** can be used. If the user wants to get some summarised results with MMPC for many combinations of *max_k* and *treshold* values he can use the **mmpc.path** function. Ridge regression (**ridge.reg** and **ridge.cv**) have been implemented. Note that ridge regression is currently offered only for linear regression with continuous predictor variables. As for some miscellaneous, we have implemented the zero inflated Poisson and beta regression models, should the user want to use them.

2.3 Cross-validation

cv.ses and **cv.mmpc** perform a K-fold cross validation for most of the aforementioned regression models. There are many metric functions to be used, appropriate for each case. The folds can be generated in a stratified fashion when the dependent variable is categorical. The Tibshirani and Tibshirani (Tibshirani and Tibshirani, 2009) bias correction is used. Note that two of its arguments, *metric* and *modeler* require their input values without "".

3 Networks

Currently three algorithms for constructing Bayesian Networks (or their skeleton) are offered.

- MMHC (Max-Min Hill-Climbing) (Tsamardinos et al., 2006), (**mmhc.skel**) which constructs the skeleton of the Bayesian Network (BN).
- PC algorithm (Spirtes et al., 2001) (**pc.skel** for which the orientation rules (**pc.or**) have been implemented as well. Both of these algorithms accept continuous or categorical data only. The skeleton of the PC algorithm has the option for permutation based conditional independence tests (Tsamardinos and Borboudakis, 2010).
- The functions **ci.mm** and **ci.fast** perform a symmetric test with mixed data (continuous, ordinal and binary data) (Tsagris et al., 2017). This is employed by the PC algorithm as well.
- Skeleton of a network with continuous data using forward selection. The command **corfs.network** does a similar to MMHC task. It goes to every variable and instead applying the MMPC algorithm it applies the forward selection regression. All data must be continuous, since the Pearson correlation is used. The algorithm is fast, since the forward regression with the Pearson correlation is very fast.

We also have utility functions, such as

1. **rdag** and **rdag2**. Data simulation assuming a BN (Colombo and Maathuis, 2014).
2. **findDescendants** and **findAncestors**. Descendants and ancestors of a node (variable) in a given Bayesian Network.
3. **dag2eg**. Transforming a DAG into an essential (mixed) graph, its class of equivalent DAGs.
4. **equivdags**. Checking whether two DAGs are equivalent.
5. **is.dag**. In fact this checks whether cycles are present by trying to topologically sort the edges. BNs do not allow for cycles.
6. **mb**. The Markov Blanket of a node (variable) given a Bayesian Network.
7. **nei**. The neighbours of a node (variable) given an undirected graph.
8. **undir.path**. All paths between two nodes in an undirected graph.
9. **transitiveClosure**. The transitive closure of an adjacency matrix, with and without arrowheads.
10. **plotnetwork**. Interactive plot of a graph.

4 Acknowledgments

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n. 617393.

References

Aitchison, J. (1986). *The statistical analysis of compositional data*. Chapman and Hall London.

Borboudakis, G. and Tsamardinos, I. (2017). Forward-Backward Selection with Early Dropping.

Colombo, D. and Maathuis, M. H. (2014). Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, 15(1):3741–3782.

Cox, D. H. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society*, 34(2):187–220.

Ferrari, S. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7):799–815.

Fieller, E. C. and Pearson, E. S. (1961). Tests for rank correlation coefficients: II. *Biometrika*, 48:29–40.

Gail, M. H., Lubin, J. H., and Rubinstein, L. V. (1981). Likelihood calculations for matched case-control studies and survival studies with tied death times. *Biometrika*, 68(3):703–707.

Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge University Press.

Lagani, V., Athineou, G., Farcomeni, A., Tsagris, M., and Tsamardinos, I. (2017). Feature Selection with the R Package MXM: Discovering Statistically-Equivalent Feature Subsets. *Journal of Statistical Software*, 80(7).

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14.

Maronna, R., Martin, D., and Yohai, V. (2006). *Robust statistics*. John Wiley & Sons, Chichester. ISBN.

Pinheiro, J. and Bates, D. (2006). *Mixed-effects models in S and S-PLUS*. Springer Science & Business Media.

Scholz, F. (1996). Maximum likelihood estimation for type I censored Weibull data including covariates.

Smith, R. L. (1991). Weibull regression models for reliability data. *Reliability Engineering & System Safety*, 34(1):55–76.

Spirtes, P., Glymour, C., and Scheines, R. (2001). *Causation, Prediction, and Search*. The MIT Press, second edi edition.

Tibshirani, R. J. and Tibshirani, R. (2009). A bias correction for the minimum error rate in cross-validation. *The Annals of Applied Statistics*, 3(2):822–829.

Tsagris, M., Borboudakis, G., Lagani, V., and Tsamardinos, I. (2017). Constraint-based Causal Discovery with Mixed Data. In *The 2017 ACM SIGKDD Workshop on Causal Discovery, 14/8/2017, Halifax, Nova Scotia, Canada*.

Tsamardinos, I., Aliferis, C. F., and Statnikov, A. (2003a). Time and sample efficient discovery of Markov blankets and direct causal relations. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 673–678. ACM.

Tsamardinos, I., Aliferis, C. F., Statnikov, A. R., and Statnikov, E. (2003b). Algorithms for Large Scale Markov Blanket Discovery. In *FLAIRS conference*, volume 2, pages 376–380.

Tsamardinos, I. and Borboudakis, G. (2010). Permutation testing improves Bayesian network learning. In *ECML PKDD'10 Proceedings of the 2010 European conference on Machine learning and knowledge discovery in databases*, pages 322–337. Springer-Verlag.

Tsamardinos, I., Brown, L. E., and Aliferis, C. F. (2006). The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm. *Machine Learning*, 65(1):31–78.

Tsamardinos, I., Lagani, V., and Pappas, D. (2012). Discovering multiple, equivalent biomarker signatures. In *In Proceedings of the 7th conference of the Hellenic Society for Computational Biology & Bioinformatics, Heraklion, Crete, Greece*.