Selection consistency of two-step selection method for misspecified logistic model

Mariusz Kubkowski and Jan Mielniczuk

We consider two-stage selection method of predictors when the underlying random binary regression model is misspecified and a response function is erroneously modelled as a logistic function. In this case an aim of selection is to recover the support of Kullback-Leibler projection of the binary model on the parametric family. The proposed procedure consists of screening and ordering predictors by Lasso and then selecting a subset of predictors which minimizes Generalized Information Criterion on the nested family pertaining to them.

More specifically, we consider random variable $(X,Y) \in R^p \times \{0,1\}$ and the corresponding response function $q(x) = P(Y=1|X=x)$, where the distribution of (X,Y) as well as dimension p of X may depend on n. Assume that logistic model $P(Y=1|X) = q_L(\beta^T X)$ is fitted to the data having this distribution, where $q_L(s) = 1/(1+e^{-s})$. When the model is misspecified it is often of interest to estimate parameter β_0^* of Kullback-Leibler (KL) projection of the binary model on the logistic family defined as

$$\beta_0^* = \mathrm{argmin}_{\beta \in R^p} E_X D(q(X)||q_L(X^T\beta)), \tag{1}$$

where D is KL divergence of two probability vectors. Let s_0^* be the support of β_0^*: $s_0^* = \{i : \beta_{0,i}^* \neq 0\}$. Here we discuss the problem of consistent recovery of s_0^*. Assume that we observe n independent copies (X_i, Y_i) of (X,Y) and let

$$\hat{\beta}_L(\lambda) = \mathrm{argmin}_{\beta \in R^p} \{l_n(\beta, Y|X) + \lambda_L \sum_{i=1}^{p} |\beta_i|\}, \tag{2}$$

Jan Mielniczuk
Institute of Computer Sciences Polish Academy of Sciences and Warsaw University of Technology
e-mail: miel@ipipan.waw.pl

Mariusz Kubkowski
Warsaw University of Technology and Institute of Computer Sciences Polish Academy of Sciences
e-mail: M.Kubkowski@mini.pw.edu.pl.

where the conditional loglikelihood $l_n(\beta, Y|X)$ is

$$l_n(\beta, Y|X) = \sum_{i=1}^{n} \{Y_i \log[q_L(X_i^T \beta)] + (1 - Y_i) \log[1 - q_L(X_i^T \beta)]\} \tag{3}$$

We consider two-stage selection procedure being a simplified version of SOS algorithm introduced in [**?**]

- Arrange coordinates of $\hat{\beta}_L(\lambda)$ monotonically: $|\hat{\beta}_{L,j_1}(\lambda)| \geq |\hat{\beta}_{L,j_2}(\lambda)| \geq \ldots \geq |\hat{\beta}_{L,j_{|w|}}(\lambda)|$, where $|w|$ is size of the support of $\hat{\beta}_L(\lambda)$;
- Consider the nested family $\mathscr{M} = \{\{j_1\}, \{j_1, j_2\}, \ldots, \{j_1, j_2, \ldots, j_{|s|}\}\}$ and let

  $$\hat{s}_0^* = \mathrm{argmin}_{s \in \mathscr{M}} GIC(s),$$

  where Generalized Information Criterion GIC is penalized log-likelihood

  $$GIC(s) = -2 l_n(\hat{\beta}_s, Y|X_s) + a_n |s|,$$

  a_n is a chosen penalty and s is a given submodel containing $|s|$ explanatory variables.

In the contribution we discuss sufficient conditions in [**?**] on the parameters of the method and distribution of (X, Y) under which the above procedure is consistent i.e. $P(\hat{s}_0^* = s_0^*) \to 1$. They in particular allow for exponential increase of number of predictors p as a function of sample size n. The derivation relies on showing consistent screening property of the Lasso for random predictors and misspecification case by proving the following separation result: with large probability we have with $\bar{s}_0^*$ being a complement of s_0^*,

$$\min_{j \in s_0^*} |\hat{\beta}_{L,j}(\lambda)| \geq \max_{j \in \bar{s}_0^*} |\hat{\beta}_{L,j}(\lambda)|,$$

which is an analogue of the result for deterministic predictors [**?**]. Secondly, consistency of GIC on supermodels and submodels od s_0^* is proved using empirical processes approach to loglikelihood. Applications of the result include the special case of the misspecification when $q(x) = \tilde{q}(\beta_0^T x)$. Then under assumptions that regressions of X given $\beta_0^T X$ are linear in $\beta_0^T X$ we have that $\beta_0^* = c\beta_0$ (see e.g. [**?**]) and the result is used to recover the support of unknown β_0.
In numerical experiments we discuss performance of several modifications of the above procedure, in particular its net version when the nested family $\mathscr{M}$ is replaced by the sum of such families constructed for a net of λs and the version when the ordering by Lasso is replaced by ordering with respect to squared values of Wald statistics for the fitted logistic model.

References

[Kubkowski and Mielniczuk(2017)] M. Kubkowski and J. Mielniczuk. Active set of predictors for misspecified logistic regression. *Statistics*, 51:1023–1045, 2017.

[Kubkowski and Mielniczuk(2018)] M. Kubkowski and J. Mielniczuk. Selection consistency of two-step selection method for misspecified logistic regression. in preparation, 2018.

[Pokarowski and Mielniczuk(2015)] P. Pokarowski and J. Mielniczuk. Combined ℓ_1 and ℓ_0 penalized least squares. *Journal of Machine Learning Research*, 16:961–992, 2015.

[Ye and Zhang(2010)] F. Ye and C.H. Zhang. Rate minimaxity of the lasso and Dantzig selector for the q loss in r balls. *J. Mach. Learn. Res.*, 11:3519–3540, 2010.