Non-parametric estimation of ordinal factor models

Steffen Grønneberg[1]

[1] *BI Norwegian Business School, Norway*

Factor models were originally developed for continuous data, and later extended to ordinal data through adding a discretization process to the model. With continuous data, factor models can be estimated without strong distributional assumptions such as normality. This is not the case for ordinal models, and a normality assumption is usually made for the underlying continuous variable. Tests of underlying normality indicate that the normality assumption is rarely fulfilled, yet is assumed in the much used standard methodology even though this methodology is not robust towards non-normality. In the talk, I will introduce an estimator that works under non-parametric assumptions, as introduced in the recent paper Foldnes & Grønneberg (2021, Psych Methods).

On the asymptotic behaviour of the variance estimator of a U-statistic

Riccardo De Bin[1], based on a joint work with **Mathias Fuchs**[2], **Roman Hornung**[3] **and Anne-Laure Boulesteix**[3]

[1] *University of Oslo, Norway, debin@math.uio.no*
[2] *mathiasfuchs.de consulting*
[3] *Ludwig-Maximilians-Universität of Munich, Germany*

In supervised learning, including supervised classification as an important special case, the prediction error is often estimated by means of resampling-based procedures such as cross-validation. In methodological studies, the prediction error is used to contrast the performances of several prediction algorithms. A crucial but challenging question is whether the observed differences between the estimates are statistically significant or not, i.e., whether they are compatible with the null-hypothesis of no true difference. To answer this question, a good understanding of the error estimates' distribution is required. In the case of resampling-based procedures, however, the estimation of the variance is difficult: the learning and test sets considered in the successive resampling iterations overlap and, therefore, the iteration-specific error estimates computed in the resampling iterations are dependent. Their covariance structure is complex, thus making the estimation of the variance of their average very arduous in general. An unbiased variance estimator, suggested in the literature, can be recast as a U-statistics variance. However, its kernel size depends on the sample size, preventing asymptotic statements. Here, we solve this issue by decomposing the variance estimator into a linear combination of U-statistics with fixed kernel size, and consequently obtaining the desired asymptotic. We show that it is possible to construct a confidence interval for the true error and derive a statistical test which compares the error estimates of two classification algorithms. The confidence interval's coverage probability and the test are illustrated by means of both a simulation study and real data application.

Empirical analysis of average treatment effects: The minimum squared error IV-estimator

Øyvind Hoveid[1]

[1] *Norwegian Institute of Bioeconomy Research, Ås, Norway, oyvind.hoveid@nibio.no*

In the absence of randomised experiments can average treatment effects of some x on some y in general not be uncovered with regression due to the problem of confounding. Unobserved variables may affect both x and y and the regression coefficient turns biased. Theoretical causal analysis (Pearl, 2009) points to solutions based heavily on the directed acyclic graph (DAG) over observed and unobserved variables. Sometimes an observed instrument z and the IV-estimator solves the problem *as if* a randomised experiment had been conducted. Necessary and sufficient algebraic conditions for the consistency of the IV-estimator are proved. The empirical catch is that these conditions are non-testable and need be established with subject matter insights (Angrist, Imbens & Rubin, 1996). Moreover, at least in social sciences, it is most often unclear which graph the observed and unobserved variables stem from. These problems make valid instruments rare.

For a wider scope of empirical causal analysis, the minimum square error IV-estimator (MSEIV) is introduced. It does not guarantee consistency in the usual interpretation, but subject matter insight is less demanding.

The estimator relies on $x|v$ as instrument where v is a block of observed possibly confounding variables. Widening the block v may gradually reduce the confounding effect. In this sense will estimates be informative. It is proved that using all available v ensures a least bias of the IV-estimator.

The MSEIV-estimator is in turn the result of a model selection procedure where columns of v bringing more variance than reduction of squared bias are removed. Effectively, the least important confounding variables are detected and removed. It is recognised that selection takes place according to a certain focused information criterion (FIC) (Claeskens & Hiort, 2003) for the regression model $y|x, v$. The IV-interpretation is nevertheless preferred here as it points directly to the relevant focus.

The subject matter insights are related to a DAG over nodes (u, v, x, y) with u being the unobserved and v the observed confounding variables. In the current case it need be decided whether a certain observed variable is a possible confounder of x and y or not. In the world of Pearl (2009) where graphs are facts, u and v should be parents of both x and y and x should be a parent of y. In a social science context where few criteria for giving an edge a direction exists beyond separation in time, graphs are merely hypotheses of the data generating process, can weaker requirements be stated: *u or v should not be descendants of x nor y, and x should not be descendant of y.* After all, even with simultaneous variables it make sense to ask how y is changed with x, keeping "everything else" fixed.

The MSEIV-estimator is illustrated with estimation of the effect of income on farmer labour in a cross-section of data.

Two-sample Empirical Likelihood for quantile inference of weakly dependent processes

Reinis Alksnis[1] **and Janis Valeinis**[2]

[1] *University of Latvia, Latvia, reinis.alksnis@lu.lv*
[2] *University of Latvia, Latvia, janis.valeinis@lu.lv*

Keywords: Empirical Likelihood, weakly dependent processes, two-sample problems, smoothing.

Statistical inference related to quantile estimation is ubiquitous in applied sciences with various applications, such as risk management in finance, change-point analysis in anomaly detection and group comparisons in biomedicine to name a few. The estimation of quantiles is especially useful when underlying distributions tend to be skewed, which is often the case for real world data. Also, depending on the sample generation process, it is often necessary to take into account the associated dependence structure.

In a nonparametric context Chen and Wong [3] approached such problem by developing the so called smoothed block Empirical Likelihood. Based on the Empirical Likelihood (EL) method, introduced by Owen [1], it combines ideas of smoothing [2] and blocking [5] to reduce the length of confidence intervals and at the same time to handle the dependence that is present in the data [5].

We are interested to make an inference of the difference of two population quantiles in the presence of some weak dependence and propose to use EL for this task. Analogously to [3], we formulate the two-sample smoothed block EL to construct corresponding confidence intervals. Moreover, we discuss other methods found in EL literature ([4], [5], [6]) that can be applied for the same task. Finally, we carry out a simulation study to analyze the performance of proposed methods.

References

[1] Owen, A. (1988). Empirical likelihood Ratio Confidence Intervals for a Single Functional. *Biometrika* **75(2)**, 237–249.

[2] Chen, S., Hall, P. (1993). Smoothed Empirical Likelihood Confidence Intervals for Quantiles. *The Annals of Statistics* **21(3)**, 1166–1181.

[3] Chen, S.Xi, Wong, C.M. (2009). Smoothed Block Empirical Likelihood for Quantiles of Weakly Dependent Processes. *Statistica Sinica* **19(1)**, 1166–1181.

[4] Nordman, D., Lahiri, S. (2014). A review of empirical likelihood methods for time series. *Journal of Statistical Planning and Inference* **155**, 1–18.

[5] Kitamura, Y. (1997). Empirical likelihood methods with weakly dependent processes. *The Annals of Statistics* **25(5)**, 2084–2102.

[6] Lopez, E., Van Keilegom, I. and Veraverbeke, N. (2009). Empirical Likelihood for Non-Smooth Criterion Functions. *Scandinavian Journal of Statistics* **36(3)**, 413–432.

A generalized smoothly trimmed mean estimator of location

Jānis Valeinis[1] **and Elīna Kresse** [2]

[1] *University of Latvia, Latvia, janis.valeinis@lu.lv*
[2] *University of Latvia, Latvia, elina.kresse@lu.lv*

The median, the trimmed mean and the Winsorized mean are the classical L-estimators used for the robust inference of the central tendency. The origins of the trimmed and the Winsorized mean can be traced back to 1962, when they were introduced by John W. Tukey and Donald H. McLaughlin. Although the trimmed mean is still a commonly used and very popular robust location estimator, Stigler showed in 1973 [1] that it has a significant disadvantage – the limiting distribution of the trimmed mean may not be asymptotically normal if the data have a discrete or continuous distribution with gaps.

Although the smoothly trimmed mean has been mentioned in the literature since 1967 [2], it has been very little studied. One of the reasons for it could be the difficulty to find its variance and the optimal size of trimming. In the case of smoothly trimmed mean, only the triangular and trapezoidal weight functions have been considered. We introduce a new estimator which is a more general version of the smoothly trimmed mean. It covers both versions of smoothly trimmed means mentioned before and also some classical estimators as the limiting cases.

Until now, the jackknifing has been used to estimate the variance of smoothly trimmed mean [3]. Although the technique is effective and simple, it is time-consuming in computational terms, so we establish the asymptotic variance and its estimator using the influence function. Moreover, based on Qin and Tsao results [4] the empirical likelihood method is established and implemented for the new version of smoothly trimmed mean. Finally, the simulation study has been provided confirming the theoretical results.

References

[1] Stigler. S. M. The asymptotic distribution of the trimmed mean. (1973). *The Annals of Statistics*, 472–477.

[2] Crow, E. L. and Siddiqui M. M. (1967). Robust estimation of location. *Journal of the American Statistical Association*, **62**(318), 353 – 389.

[3] Parr, W. C. and Schucany W. R. (1982). Jackknifing L-statistics with smooth weight functions. *Journal of the American Statistical Association*, **77**(379), 629 – 638.

[4] Qin, G. and Tsao M. (2002). Empirical likelihood ratio confidence interval for the trimmed mean. *Communications in statistics-theory and methods* **31**(12), 2197 – 2208.