Data-driven goodness-of-fit tests

Mikhail Langovoy

University of Bonn
E-mail: langovoy@wiener.iam.uni-bonn.de

We consider a general class of statistical tests. The class contains Neyman's smooth tests and data-driven score tests as special examples. Our tests are additionally incorporated with model selection rules. The rules are based on the penalization idea. Most of the optimal penalties, derived in statistical literature, can be used in our tests. We prove general consistency theorems for the tests from the class. We show that the tests can be applied for simple and composite parametric, semi- and nonparametric hypotheses. Possible applications are in statistical inverse problems, time series analysis and statistics for stochastic processes.

There are two main approaches to constructing test statistics. In the first approach, roughly speaking, some measure of distance between the theoretical and the corresponding empirical distributions is proposed as the test statistic. Classical examples of this approach are the Cramer-von Mises and the Kolmogorov-Smirnov statistics. More generally, L^p-distance based tests, as well as graphical tests based on confidence bands, usually belong to this type. Although, these tests works and are capable of giving very good results, but each of these tests is asymptotically optimal only in a finite number of directions of alternatives to a null hypothesis (see [12]).

Nowadays, there is an increasing interest to the second approach of constructing test statistics. The idea of this approach is to construct tests in such a way that the tests would be asymptotically optimal in some sense, or most powerful, at least in a reach enough set of directions. Test statistics constructed following this approach are often called score test statistics. The pioneer of this approach was Neyman [11]. See [8], [2], [9] for subsequent developments, and [3], [1] and [10] for recent results in the field. This approach is also closely related to the theory of efficient estimation [4]. Data-driven score tests are, at least in basic situations, asymptotically optimal in an infinite number of directions of alternatives (see [5]).

Classical score tests have been substantially generalized in recent literature: see, for example, generalized likelihood ratio statistics for nonparametric models [3], tailor-made tests [1] and semiparametric generalized likelihood ratio statistics [10].

We propose a new development of the theory of *data-driven* score tests. We introduce the notions of NT- and GNT-tests, generalizing the concept of data-driven score tests, for both simple and composite hypotheses. We propose a unified approach for proving asymptotic consistency of our tests. Moreover, for any NT- or GNT-test, we have an explicit rule to determine whether the test will be consistent against any particular alternative.

Our method is applicable to statistical inverse problems ([7]), dependent data, and inference for stochastic processes ([6]).

References

[1] Bickel, P. J. and Ritov, Y. and Stoker, T. M. Tailor-made tests for goodness of fit to semiparametric hypotheses. *Ann. Statist.*, **34** (2006), no. 2, 721–741.

[2] Cox, D. R. and Hinkley, D. V. Theoretical statistics. Chapman and Hall, London, 1974.

[3] Fan, J. and Zhang, C. and Zhang, J. Generalized likelihood ratio statistics and Wilks phenomenon. *Ann. Stat.*, **29** (2001), no. 1, 153–193.

[4] Ibragimov, I. A. and Has'minskii, R. Z. Statistical estimation. Springer-Verlag, New York, 1981. ISBN: 0-387-90523-5

[5] Inglot, T. and Ledwina, T. Asymptotic optimality of data-driven Neyman's tests for uniformity. *Ann. Statist.*, **24** (1996), no. 5, 1982–2019.

[6] Langovoy, Mikhail. Data-driven goodness-of-fit tests. Ph.D. thesis. University of Göttingen, Göttingen, 2007.

[7] Langovoy, Mikhail. Data-driven efficient score tests for deconvolution hypotheses. *Inverse Problems*, **24** (2008), no. 2, 1–17.

[8] Le Cam, L. On the asymptotic theory of estimation and testing hypotheses. Third Berkeley Symposium on Mathematical Statistics and Probability, vol. I. 1956, 129–156.

[9] Ledwina, T. Data-driven version of Neyman's smooth test of fit. *J. Amer. Statist. Assoc.*, **89** (1994), no. 427, 1000–1005.

[10] Li, R. and Liang, H. Variable selection in semiparametric regression modeling. *Ann. Statist., to appear*, (2007)

[11] Neyman, Jerzy. Smooth test for goodness of fit. *Skand. Aktuarietidskr.*, **20** (1937), 150–199.

[12] Nikitin, Ya'akov. Asymptotic efficiency of nonparametric tests. Cambridge University Press, Cambridge, 1995. ISBN: 0-521-47029-3