Data Driven Smooth Tests

Przemysław Biecek and Teresa Ledwina

1 General Description

Smooth test was introduced by Neyman (1937) to verify simple null hypothesis asserting that observations obey completely known continuous distribution function F. Smooth test statistic (with k components) can be interpreted as score statistic in an appropriate class of auxiliary models indexed by a vector of parameters $\theta \in R^k,\ k \geq 1$. Pertaining auxilary null hypothesis asserts $\theta = \theta_0 = 0$. Therefore, in this case, the smooth test statistic based on n i.i.d. observations $Z_1, ..., Z_n$ has the form

$$W_k = \left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \ell(Z_i)\right] \mathcal{I}^{-1} \left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \ell(Z_i)\right]',$$

where $\ell(Z_i),\ i = 1, ..., n$, is k-dimensional (row) score vector, the symbol $'$ denotes transposition, while $\mathcal{I} = Cov_{\theta_0}[\ell(Z_1)]'[\ell(Z_1)]$. Following Neyman's idea of modelling underlying distributions one gets $\ell(Z_i) = (\phi_1(F(Z_i)), ..., \phi_k(F(Z_i)))$ and $\mathcal{I}$ being the identity matrix, where ϕ_j's, $j \geq 1$, are zero mean orthonormal functions on [0,1], while F is the completely specified null distribution function.

In case of composite null hypothesis there is also unspecified vector of nuisance parameters γ defining the distribution of observations. Smooth statistic (with k components) in such applications is understood as efficient score statistic for some class of models indexed by an auxiliary parmeter $\theta \in R^k,\ k \geq 1$. Pertaining efficient score vector $\ell^*(Z_i; \gamma)$ is defined as the residual from projection the score vector for θ onto the space spanned by score vector for γ. As such, smooth test is alternative name for $C(\alpha)$ Neyman's test. See Neyman (1959), Bühler and Puri (1966) as well as Javitz (1975) for details. Hence, smooth test, based on n i.i.d. variables $Z_1, ..., Z_n$ rejects hypothesis $\theta = \theta_0 = 0$ for large values of

$$W_k^*(\tilde{\gamma}) = \left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \ell^*(Z_i; \tilde{\gamma})\right] [\mathcal{I}^*(\tilde{\gamma})]^{-1} \left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \ell^*(Z_i; \tilde{\gamma})\right]',$$

where $\tilde{\gamma}$ is an appropriate estimator of γ while $\mathcal{I}^*(\gamma) = Cov_{\theta_0}[\ell^*(Z_1; \gamma)]'[\ell^*(Z_1; \gamma)]$. More details can be found in Janic and Ledwina (2008), Kallenberg and Ledwina (1997 a,b) as well as Inglot and Ledwina (2006 a,b).

Auxiliary models, mentioned above, aim to mimic the unknown underlying model for the data at hand. To choose the dimension k of the auxilary model we apply some model selection criteria. Among several solutions already considered, we decided to implement two following ones, pertaining to the two above described problems and resulting W_k and $W_k^*(\tilde{\gamma})$.

The selection rules in the two cases are briefly denoted by T and T^*, respectively, and given by

$$T = \min\{1 \le k \le d :\ W_k - \pi(k,n,c) \ge W_j - \pi(j,n,c),\ j = 1,...,d\}$$

and

$$T^* = \min\{1 \le k \le d :\ W_k^*(\tilde{\gamma}) - \pi^*(k,n,c) \ge W_j^*(\tilde{\gamma}) - \pi^*(j,n,c),\ j = 1,...,d\}.$$

Both criteria are based on approximations of penalized loglikelihoods, where loglikelihoods are replaced by W_k and $W_k^*(\tilde{\gamma})$, respectively. The penalties for the dimension j in case of simple and composite null hypothesis are defined as follows

$$\pi(j,n,c) = \begin{cases} j\log n, & \text{if } \max_{1\le k\le d}|\mathcal{Y}_k| \le \sqrt{c\log n}, \\ 2j, & \text{if } \max_{1\le k\le d}|\mathcal{Y}_k| > \sqrt{c\log n} \end{cases}$$

and

$$\pi^*(j,n,c) = \begin{cases} j\log n, & \text{if } \max_{1\le k\le d}|\mathcal{Y}_k^*| \le \sqrt{c\log n}, \\ 2j, & \text{if } \max_{1\le k\le d}|\mathcal{Y}_k^*| > \sqrt{c\log n}, \end{cases}$$

respectively, where c is some calibrating constant, d is maximal dimension taken into account, $(\mathcal{Y}_1,...,\mathcal{Y}_k) = [\frac{1}{\sqrt{n}}\sum_{i=1}^n \ell(Z_i)]\mathcal{I}^{-1/2}$ while $(\mathcal{Y}_1^*,...,\mathcal{Y}_k^*) = [\frac{1}{\sqrt{n}}\sum_{i=1}^n \ell^*(Z_i;\tilde{\gamma})][\mathcal{I}^*(\tilde{\gamma})]^{-1/2}$. In consequence, data driven smooth tests for the simple and composite null hypothesis reject for large values of W_T and $W_{T^*} = W_{T^*}(\tilde{\gamma})$, respectively. For details see Inglot and Ledwina (2006 a,b,c).

The choice of c in T and T^* is decisive to finite sample behaviour of the selection rules and pertaining statistics W_T and $W_{T^*}(\tilde{\gamma})$. In particular, under large c's the rules behave similarly as Schwarz's (1978) BIC while for $c = 0$ they mimic Akaike's (1973) AIC. For moderate sample sizes, values $c \in (2, 2.5)$ guarantee, under 'smooth' departures, only slightly smaller power as in case BIC were used and simultaneously give much higher power than BIC under multimodal alternatives. In genral, large $c's$ are recommended if changes in location, scale, skewness and kurtosis are in principle aimed to be detected. For evidence and discussion see Inglot and Ledwina (2006 c).

It $c > 0$ then the limiting null distribution of W_T and $W_{T^*}(\tilde{\gamma})$ is central chi-squared with one degree of freedom. In our implementation, for given n, both critical values and p-values are computed by MC method.

Empirical distributions of T and T^* as well as W_T and $W_{T^*}(\tilde{\gamma})$ are not essentially influenced by the choice of reasonably large d's, provided that sample size is at least moderate.

References

Akaike, H. (1973). Information theory and the maximum likelihood principle. In: *2nd International Symposium on Information Theory*, (eds. B. N. Petrov and F. Csàki), 267-281. Akademiai Kiàdo, Budapest.

Bühler, W.J., Puri, P.S. (1966). On optimal asymptotic tests of composite hypotheses with several constraints. *Z. Wahrsch. verw. Geb.* **5**, 71–88.

Inglot, T., Ledwina, T. (2006 a). Data-driven score tests for homoscedastic linear regression model: asymptotic results. *Probab. Math. Statist.* **26**, 41–61.

Inglot, T., Ledwina, T. (2006 b). Data-driven score tests for homoscedastic linear regression model: the construction and simulations. In *Prague Stochastics 2006. Proceedings*, (eds. M. Hušková, M. Janžura), 124–137. Matfyzpress, Prague.

Inglot, T., Ledwina, T. (2006 c). Towards data driven selection of a penalty function for data driven Neyman tests. *Linear Algebra and its Appl.* **417**, 579–590.

Javitz, H.S. (1975). Generalized smooth tests of goodness of fit, independence and equality of distributions. Ph.D. thesis at University of California, Berkeley.

Janic, A. and Ledwina, T. (2008). Data-driven tests for a location-scale family revisited. *J. Statist. Theory. Pract. Special issue on Modern Goodness of Fit Methods. accepted.*

Kallenberg, W.C.M., Ledwina, T. (1997 a). Data driven smooth tests for composite hypotheses: Comparison of powers. *J. Statist. Comput. Simul.* **59**, 101–121.

Kallenberg, W.C.M., Ledwina, T. (1997 b). Data driven smooth tests when the hypothesis is composite. *J. Amer. Statist. Assoc.* **92**, 1094–1104.

Neyman, J. (1937). 'Smooth test' for goodness of fit. *Skand. Aktuarietidskr.* **20**, 149-199.

Neyman, J. (1959). Optimal asymptotic tests of composite statistical hypotheses. In *Probability and Statistics*, (ed. U. Grenander), Harald Cramér Volume, 212–234. Wiley, New York.

2 Data Driven Smooth Test for Uniformity

Embeding null model into the original exponential family introduced by Neyman (1937) leads to the information matrix $\mathcal{I}$ being identity and smooth test statistic with k components

$$W_k = \frac{1}{\sqrt{n}} \sum_{j=1}^{k} \sum_{i=1}^{n} [\phi_j(Z_i)]^2 ,$$

where ϕ_j is jth degree normalized Legendre polynomial on [0,1] (default value of parameter base = 'ddst.base.legendre'). Alternatively, in our implementation, cosine system can be selected (base = 'ddst.base.cos'). For details see Ledwina (1994) and Inglot and Ledwina (2006).

An application of the pertaining selection rule T for choosing k gives related 'ddst.unif.test()' based on statistic W_T.

Similar approach applies to testing goodness-of-fit to any fully specified continuous distribution function F. For this purpose it is enough to apply the above solution to transformed observations $F(z_1), ..., F(z_n)$.

References

Inglot, T., Ledwina, T. (2006). Towards data driven selection of a penalty function for data driven Neyman tests. *Linear Algebra and its Appl.* **417**, 579–590.

Ledwina, T. (1994). Data driven version of Neyman's smooth test of fit. *J. Amer. Statist. Assoc.* **89** 1000-1005.

Neyman, J. (1937). 'Smooth test' for goodness of fit. *Skand. Aktuarietidskr.* **20**, 149-199.

Code examples

keep.source=TRUE

```
library(ddst)

## Loading required package:  orthopolynom
## Loading required package:  polynom
## Loading required package:  evd

# for given vector of 19 numbers
z = c(13.41, 6.04, 1.26, 3.67, -4.54, 2.92, 0.44, 12.93, 6.77, 10.09,
    4.10, 4.04, -1.97, 2.17, -5.38, -7.30, 4.75, 5.63, 8.84)
ddst.uniform.test(z, compute.p=TRUE)

##
##  Data Driven Smooth Test for Uniformity
##
```

```
## data:  z,   base: ddst.base.legendre   c: 2.4
## WT = 1.9112e+33, n. coord = 10, p-value < 2.2e-16

# when H0 is true
z = runif(80)
ddst.uniform.test(z, compute.p=TRUE)

##
##  Data Driven Smooth Test for Uniformity
##
## data:  z,   base: ddst.base.legendre   c: 2.4
## WT = 2.9852, n. coord = 1, p-value = 0.146

# for known fixed alternative, here N(10,16)
ddst.uniform.test(pnorm(z, 10, 16), compute.p=TRUE)

##
##  Data Driven Smooth Test for Uniformity
##
## data:  pnormz1016,   base: ddst.base.legendre   c: 2.4
## WT = 519.07, n. coord = 10, p-value < 2.2e-16

# when H0 is false
z = rbeta(80,4,2)
(t = ddst.uniform.test(z, compute.p=TRUE))

##
##  Data Driven Smooth Test for Uniformity
##
## data:  z,   base: ddst.base.legendre   c: 2.4
## WT = 38.928, n. coord = 3, p-value < 2.2e-16

t$p.value

## [1] 0
```

3 Data Driven Smooth Test for Exponentiality

Null density is given by

$$f(z;\gamma) = \exp\{-z/\gamma\} \quad \text{for} \quad z \geq 0$$

and 0 otherwise. Modelling alternatives similarly as in Kallenberg and Ledwina (1997 a,b), e.g., and estimating γ by $\tilde{\gamma} = \frac{1}{n}\sum_{i=1}^{n} Z_i$ yields the efficient score vector $\ell^*(Z_i;\tilde{\gamma}) = (\phi_1(F(Z_i;\tilde{\gamma})), ..., \phi_k(F(Z_i;\tilde{\gamma})))$, where ϕ_j's are jth degree orthonormal Legendre polynomials on [0,1] or cosine functions $\sqrt{2}\cos(\pi j x)$, $j \geq 1$, while $F(z;\gamma)$ is the distribution function pertaining to $f(z;\gamma)$. The matrix $[\mathcal{I}^*(\tilde{\gamma})]^{-1}$ does not depend on $\tilde{\gamma}$ and is calculated for succeding dimensions k using some recurrent relations for Legendre's polynomials and computed in a numerical way in case of cosine basis. In the implementation the default value of c in T^* is set to be 100. Therefore, T^* practically coincides with S1 considered in Kallenberg and Ledwina (1997 a).

References

Kallenberg, W.C.M., Ledwina, T. (1997 a). Data driven smooth tests for composite hypotheses: Comparison of powers. *J. Statist. Comput. Simul.* 59, 101–121.

Kallenberg, W.C.M., Ledwina, T. (1997 b). Data driven smooth tests when the hypothesis is composite. *J. Amer. Statist. Assoc.* 92, 1094–1104.

Code examples

```
# for given vector of 19 numbers
z = c(13.41, 6.04, 1.26, 3.67, -4.54, 2.92, 0.44, 12.93, 6.77, 10.09,
   4.10, 4.04, -1.97, 2.17, -5.38, -7.30, 4.75, 5.63, 8.84)
ddst.exp.test(z, compute.p=TRUE)

##
##  Data Driven Smooth Test for Expotentiality
##
## data:  z,   base: ddst.base.legendre,   c: 100
## WT* = 81.353, n. coord = 5, p-value < 2.2e-16

# when H0 is true
z = rexp(80,4)
ddst.exp.test (z, compute.p = TRUE)

##
##  Data Driven Smooth Test for Expotentiality
##
## data:  z,   base: ddst.base.legendre,   c: 100
## WT* = 0.21797, n. coord = 1, p-value = 0.658

# when H0 is false
z = rchisq(80,4)
ddst.exp.test (z, compute.p = TRUE)
```

```
##
##  Data Driven Smooth Test for Expotentiality
##
## data:  z,   base: ddst.base.legendre,   c: 100
## WT* = 9.2498, n. coord = 1, p-value = 0.021
```

4 Data Driven Smooth Test for Normality

Null density is given by

$$f(z;\gamma) = \frac{1}{\sqrt{2\pi}\gamma_2} \exp\left\{-\frac{(z-\gamma_1)^2}{2\gamma_2^2}\right\} \quad \text{for} \quad z \in R.$$

We model alternatives similarly as in Kallenberg and Ledwina (1997 a,b) using Legendre's polynomials or cosine basis. The parameter $\gamma = (\gamma_1, \gamma_2)$ is estimated by $\tilde{\gamma} = (\tilde{\gamma}_1, \tilde{\gamma}_2)$, where $\tilde{\gamma}_1 = \frac{1}{n}\sum_{i=1}^{n} Z_i$ and $\tilde{\gamma}_2 = \frac{1}{n-1}\sum_{i=1}^{n-1} \frac{Z_{n:i+1}-Z_{n:i}}{H_{i+1}-H_i}$, while $Z_{n:1} \leq ... \leq Z_{n:n}$ are ordered values of $Z_1, ..., Z_n$ and $H_i = \Phi^{-1}\left(\frac{i-3/8}{n+1/4}\right)$, cf. Chen and Shapiro (1995). The above yields auxiliary test statistic $W_k^*(\tilde{\gamma})$ described in details in Janic and Ledwina (2008), in case when Legendre's basis is applied. The pertaining matrix $[\mathcal{I}^*(\tilde{\gamma})]^{-1}$ does not depend on $\tilde{\gamma}$ and is calculated for succeding dimensions k using some recurrent relations for Legendre's polynomials and is computed in a numerical way in case of cosine basis. In the implementation of T^* the default value of c is set to be 100. Therefore, in practice, T^* is Schwarz-type criterion. See Inglot and Ledwina (2006) as well as Janic and Ledwina (2008) for comments. The resulting data driven test statistic for normality is $W_{T^*} = W_{T^*}(\tilde{\gamma})$.

References

Chen, L., Shapiro, S.S. (1995). An alternative test for normality based on normalized spacings. *J. Statist. Comput. Simul.* 53, 269–288.

Inglot, T., Ledwina, T. (2006). Towards data driven selection of a penalty function for data driven Neyman tests. *Linear Algebra and its Appl.* **417**, 579–590.

Janic, A. and Ledwina, T. (2008). Data-driven tests for a location-scale family revisited. *J. Statist. Theory. Pract. Special issue on Modern Goodness of Fit Methods. accepted.*

Kallenberg, W.C.M., Ledwina, T. (1997 a). Data driven smooth tests for composite hypotheses: Comparison of powers. *J. Statist. Comput. Simul.* 59, 101–121.

Kallenberg, W.C.M., Ledwina, T. (1997 b). Data driven smooth tests when the hypothesis is composite. *J. Amer. Statist. Assoc.* 92, 1094–1104.

Code examples

```
# for given vector of 19 numbers
z = c(13.41, 6.04, 1.26, 3.67, -4.54, 2.92, 0.44, 12.93, 6.77, 10.09,
   4.10, 4.04, -1.97, 2.17, -5.38, -7.30, 4.75, 5.63, 8.84)
ddst.norm.test(z, compute.p=TRUE)

##
##  Data Driven Smooth Test for Normality
```

```
##
## data:  z,   base: ddst.base.legendre,   c: 100
## WT* = 0.13095, n. coord = 1, p-value = 0.7201

# when H0 is true
z = rnorm(80)
ddst.norm.test(z, compute.p=TRUE)

##
##  Data Driven Smooth Test for Normality
##
## data:  z,   base: ddst.base.legendre,   c: 100
## WT* = 0.0056928, n. coord = 1, p-value = 0.9463

# when H0 is false
z = rexp(80,4)
ddst.norm.test(z, B=5000, compute.p=TRUE)

##
##  Data Driven Smooth Test for Normality
##
## data:  z,   base: ddst.base.legendre,   c: 100
## WT* = 185.8, n. coord = 5, p-value = 5.414e-13
```

5 Data Driven Smooth Test for Extreme Value Distribution

Null density is given by

$$f(z;\gamma) = \frac{1}{\gamma_2} \exp\Big\{\frac{z-\gamma_1}{\gamma_2} - \exp\Big(\frac{z-\gamma_1}{\gamma_2}\Big)\Big\}, \quad z \in R.$$

We model alternatives similarly as in Kallenberg and Ledwina (1997) and Janic-Wróblewska (2004) using Legendre's polynomials or cosines. The parameter $\gamma = (\gamma_1, \gamma_2)$ is estimated by $\tilde{\gamma} = (\tilde{\gamma}_1, \tilde{\gamma}_2)$, where $\tilde{\gamma}_1 = -\frac{1}{n}\sum_{i=1}^{n} Z_i + \epsilon G$, where $\epsilon \approx 0.577216$ is the Euler constant and $G = \tilde{\gamma}_2 = [n(n-1)\ln 2]^{-1} \sum_{1\le j<i\le n}(Z^o_{n:i} - Z^o_{n:j})$ while $Z^o_{n:1} \le ... \le Z^o_{n:n}$ are ordered variables $-Z_1, ..., -Z_n$, cf Hosking et al. (1985). The above yields auxiliary test statistic $W^*_k(\tilde{\gamma})$ described in details in Janic and Ledwina (2008), in case when Legendre's basis is applied. The related matrix $[\mathcal{I}^*(\tilde{\gamma})]^{-1}$ does not depend on $\tilde{\gamma}$ and is calculated for succeding dimensions k using some recurrent relations for Legendre's polynomials and numerical methods for cosine functions. In the implementation the default value of c in T^* was fixed to be 100. Hence, T^* is Schwarz-type model selection rule. The resulting data driven test statistic for extreme value distribution is $W_{T^*} = W_{T^*}(\tilde{\gamma})$.

References

Hosking, J.R.M., Wallis, J.R., Wood, E.F. (1985). Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. *Technometrics* 27, 251–261.

Janic-Wróblewska, A. (2004). Data-driven smooth test for extreme value distribution. *Statistics* 38, 413–426.

Janic, A. and Ledwina, T. (2008). Data-driven tests for a location-scale family revisited. *J. Statist. Theory. Pract. Special issue on Modern Goodness of Fit Methods. accepted.*

Kallenberg, W.C.M., Ledwina, T. (1997). Data driven smooth tests for composite hypotheses: Comparison of powers. *J. Statist. Comput. Simul.* 59, 101–121.

Code examples

```
# for given vector of 19 numbers
z = c(13.41, 6.04, 1.26, 3.67, -4.54, 2.92, 0.44, 12.93, 6.77, 10.09,
    4.10, 4.04, -1.97, 2.17, -5.38, -7.30, 4.75, 5.63, 8.84)
ddst.extr.test(z, compute.p=TRUE)

##
##  Data Driven Smooth Test for Extreme Values
##
## data:  z,   base: ddst.base.legendre,   c: 100
## WT* = 1.9073, n. coord = 1, p-value = 0.594
```

```r
# when H0 is true
library(evd)
z = -qgumbel(runif(100),-1,1)
ddst.extr.test (z, compute.p = TRUE)

##
##  Data Driven Smooth Test for Extreme Values
##
## data:  z,   base: ddst.base.legendre,   c: 100
## WT* = 0.08588, n. coord = 1, p-value = 1

# when H0 is false
z = rexp(80,4)
ddst.extr.test (z, compute.p = TRUE)

##
##  Data Driven Smooth Test for Extreme Values
##
## data:  z,   base: ddst.base.legendre,   c: 100
## WT* = 7762.2, n. coord = 5, p-value = 0.002
```