Testing in Additive and Projection Pursuit Models

Arun Kumar Kuchibhotla

Indian Statistical Institute, Kolkata

28th July 2015

Acknowledgements

I would like to thank the committee for selecting and inviting me. I would also like to thank the International Statistical Institute (ISI) for sponsoring my trip to Brazil.

Outline

Measures of Association

- One of the widely studied problems in statistics is to understand whether there is any association between a given set of random variables. And if there is an association (of some kind), how can one use this information to infer about one variable using the others?

- To measure association, Rényi (1959) proposed a set of conditions that he thought should be satisfied by any measure of association.

- His conditions are very restrictive and leads to the measure of association between random variables X and Y called the maximal correlation coefficient given by

$$S = \sup_{f,g} \mathrm{Corr}(f(X), g(Y)),$$

  where the supremum is taken over all function f and g such that the correlation is defined.

Some Easy Remarks

- It is easy to note that the measure S is symmetric in the random variables and $0 \leq S \leq 1$.
- $S = 0$ if and only if the variables involved are independent. $S = 1$ if and only if there exists a perfect functional relation between the variables involved.
- S is invariant to invertible transformations of variables, in the sense that, $S(X, Y) = S(\alpha(X), \beta(Y))$ for any two invertible transformations α, β.
- S is closely related to the canonical correlation coefficient where one restricts the functions, over which the supremum is taken, to be linear.

The Model

- It is well-known that the canonical correlation vectors can be found by using alternating regressions of $u^\top Y$ on $v^\top X$ fixing one of u and v alternatively.
- Define optimal transformations f_0 and g_0 for X and Y as

$$(f_0, g_0) = \arg\sup_{f,g} \operatorname{Corr}(f(X), g(Y)).$$

- One can find "optimal" transformations f_0 and g_0 for X and Y respectively, by using alternating conditional expectations (ACE) or regressions.
- The model in this case can be written as

$$g_0(Y) = f_0(X) + \varepsilon, \qquad E[\varepsilon|X] = 0.$$

The Additive Model

- In case X is a random vector, one can still use the model $g_0(Y) = f_0(X) + \varepsilon$. But this model is not very much useful because of the well-known curse of dimensionality.
- Instead once can use the model

$$g_0(Y) = \sum_{i=1}^{p} f_{i0}(X_i) + \varepsilon, \qquad E[\varepsilon|X] = 0.$$

- It is easy to find the optimal transformations by minimizing

$$E\left[g(Y) - \sum_{i=1}^{p} f_i(X_i)\right]^2,$$

  over all functions $g, f_1, \ldots, f_p$ subject to the constraint $\mathrm{Var}(g(Y)) = 1$.
- This model is known as the additive model or the alternating conditional expectation (ACE) model.

ACE Algorithm

- Start with any mean zero transformation of random variables and variance one transformation for Y. A simple choice is to use $\beta^{(0)}(Y) = \frac{Y-E[Y]}{\sqrt{\mathrm{Var}(Y)}}$, $\alpha_1^{(0)}(X_1) = \ldots = \alpha_p^{(0)}(X_p) = 0$.
- for $k = 1$ to $k = p$ use,

$$E\left[\beta^{(j-1)}(Y) - \sum_{i<k} \alpha_i^{(j)}(X_i) - \sum_{i>k} \alpha_i^{(j-1)}(X_i)) \middle| X_k\right] = \alpha_k^{(j)}(X_k),$$

  to find $\alpha_k^{(j)}(X_k)$.
- Now use the obtained functions $\alpha_i^{(j)}$ to find new $\beta(Y)$, using,

$$E\left[\sum_{i=1}^{p} \alpha_i^{(j)}(X_i)|Y\right] \Big/ \mathrm{SD}\left(E\left[\sum_{i=1}^{p} \alpha_i^{(j)}(X_i)|Y\right]\right) = \beta^{(j)}(Y)$$

  .
- Use steps 2 and 3 until convergence.

Remarks

- In case of finite samples, use estimates of conditional expectations instead of the conditional expectations in the algorithm.
- The algorithm works even when some of the covariates $X_1, X_2, \ldots, X_p$ are discrete or periodic or of mixed type.
- The ACE model can be made semi-parametric by restricting transformations of a particular variable(s) to a parametric family.
- One has to take into consideration the number of parameters (to be estimated) and the sample size.
- Use of linear smoothers, particularly kernel or k-NN type estimators lead to outlier insensitive estimators. Robustification of estimators can be attained by using robust regression estimators.

Simulated Data

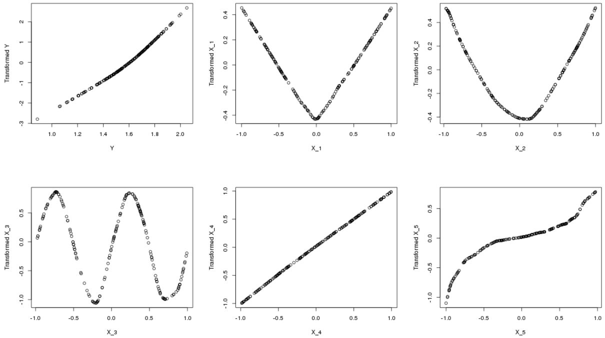

Figure 1: Model is $Y = \log(4 + |X_1| + X_2^2 + \sin(2\pi X_3) + X_4 + X_5^3 + \epsilon)$

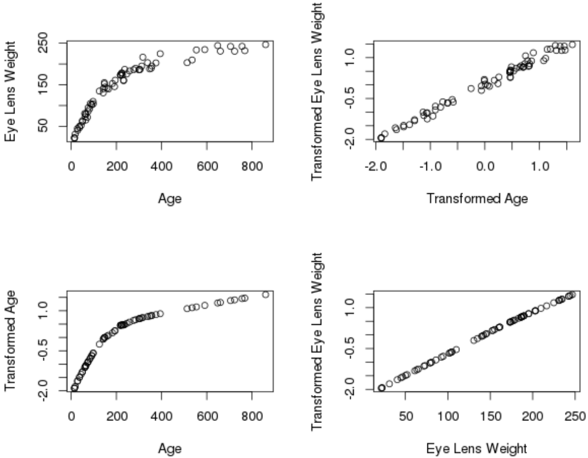

Figure 2: Eye Lens Weight versus Age of rabbit

Prediction in ACE model

- Prediction of response variable is fairly straight forward in ACE model although not as easy as in additive models.
- From the estimated optimal transformations $\hat{g}$ and $\hat{f}_1$, …, f_p of the random variables Y and $X_1, \ldots, X_p$, we consider the prediction at the covariate $x = (x_1, \ldots, x_p)$

$$\hat{y} = \hat{g}^{-1}\left(\sum_{i=1}^{p} \hat{f}_i(x_i)\right).$$

- It is important that the optimal transformation for Y has to be invertible in order to get a meaningful prediction for Y.
- Based on this prediction procedure, we can define a measure of predictive ability as

$$Q := 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

where $\hat{y}_i$ is obtained by fitting the model on the remaining $n-1$ observations. This is the usual PRESS statistic.

The Hypothesis

- As mentioned previously, it is easy to implement ACE with parametric transformations for some variables.
- In this respect, for a better interpretation it might be of interest to test if the optimal transformation is actually from the parametric family.
- In particular, how do we test that the optimal transformation for a particular random variable, say X_1, is linear? or how do we test if ACE model coincides with the multiple linear regression model?
- Since we do not make any distributional assumptions, our aim is to use the technique of bootstrap in order to perform the test.

The Hypothesis Contd.

- The hypothesis we want to test can be formally expressed as

$$\begin{aligned} H_0 &: \alpha_1(X_1) = a_1 + b_1 X_1 \text{ for some } a_1, b_1, \\ H_1 &: \alpha_1(X_1) \text{ is not linear.} \end{aligned}$$

- This hypothesis can also be written in the form of models as,

$$\begin{aligned} H_0 &: \beta(Y) = a_1 + b_1 X_1 + \sum_{i=2}^{p} \alpha_i(X_i) + \epsilon, \\ H_1 &: \beta(Y) = \sum_{i=1}^{p} \alpha_i(X_i) + \epsilon. \end{aligned}$$

- We can use any measure of goodness-of-fit which can discriminate between the null and the alternative. We propose to use either the ratio of the residual sum of squares or the ratio of the prediction sum of squares.

The First Idea

- For any particular test statistic T, let the rejection region be specified as $T > \tau$, for some τ based on the level of the test α.
- Compute the test statistics T and also the unknowns in the model satisfying the null hypothesis.
- Take B sub-samples each of size R with replacement and then calculate the value of T as calculated in step 1.
- Now we have B values from the distribution of T under the null. For large enough B, we can find the estimate of critical value, $\hat{\tau}$, based on the empirical distribution of T. Calculation of the p-value can also be done using the empirical distribution, $P(T > T_{obs.})$.

The First Idea Contd.

- The two measures which can be used as test statistics for performing these tests are given by

$$T_1 = SSE_{H_0}/SSE, \qquad T_2 = PSS_{H_0}/PSS,$$

  where $SSE_{H_0} = \sum_{j=1}^{p}(\hat{\beta}(Y_i) - \sum_{i=1}^{p} \hat{\alpha}_i(X_{ij}))^2$ for estimated functions under null and $PSS_{H_0} = \sum_{i=1}^{n}(Y_i - \hat{Y}_{iH_0})^2$.
- Note here that when we take sub-samples, they satisfy the hypothesis that the actual sample satisfies.
- Therefore, when the test statistic T is calculated in this way for each sub-sample, we are estimating the distribution of T under the true hypothesis which need not be the null
- In light of this, the estimate we get from the bootstrap in this way may not converge to the actual cut-off value.

The Refined Idea

- We follow the approach of Davidson and MacKinnon (2004) and Martin (2007) to generate the value of the test statistics under the null irrespectively of the true hypothesis.
- Fit the unrestricted full model and find the residuals, $e_1, \ldots, e_n$. Also fit the null model and find the estimates of parameters of line and estimates of other functions.
- Construct the samples, $(y_1^*, \mathbf{x}_1), (y_2^*, \mathbf{x}_2), \ldots, (y_n^*, \mathbf{x}_n)$ under null hypothesis, where y_i^* is calculated from the model $\hat{\beta}(y_i) = \hat{a}_1 + \hat{b}_1 X_{1i} + \sum_{j=1}^{p} \hat{\alpha}_j(X_{ji}) + e_i^*$ for sub-sampled $e_1^*, \ldots, e_n^*$ obtained from $e_1, \ldots, e_n$ using hat functions obtained from null model.
- Find the value of the test statistic T for each of these samples by fitting the models under null and unrestricted for the newly constructed observations.

- This method can be used to obtain a consistent test for any two nested regression models, in particular to test ACE versus multiple linear regression.
- This test can also be used for variable selection or to test parametric transformations for more than one variable.

Table 1: Tests of linearity: $Y = \log(4 + |X_1| + X_2^2 + \sin(2\pi X_3) + X_4 + X_5^3 + \epsilon)$

Variable	Statistic	Est. 5% Cut-Off	Est. p-value
Y	2.738687	1.032586	< 0.0001
1	8.688747	1.026901	< 0.0001
2	9.720563	1.027773	< 0.0001
3	37.07122	1.022618	< 0.0001
4	0.986757	1.015667	0.597
5	2.626931	1.022109	< 0.0001

Table 2: Tests of linearity for Rabbit Data back

Variable	Statistic	Est. 5% Cut-Off	Est. p-value
Y	1.114493	1.180262	0.129
1	4.531953	1.276061	< 0.0001

Remarks

- For the refined idea, prediction based on the function estimates plays a central role and so the performance of the test depends heavily on the nature of optimal transformation of the response variable under the null.

- The ACE model/algorithm allows for monotone shape-restricted transformation for any variable which can be used for response variable in order to get an invertible transformation.

- In general, one can use the test described for testing the hypothesis related to shape restricted inference such as testing the hypothesis that the optimal transformation for a variable is monotone or convex.

Another Possible Approach

- Using the approach proposed in Zheng (1996), one can develop a new testing procedure as follows.
- Underlying fact useful in this respect is that if the ACE model is the truth, then $E[u_i|\mathbf{x}_i] = 0$ with $u_i = g_0(y_i) - \sum_{j=1}^{p} f_{0j}(x_{ji})$.
- Also, observe that, $E[u_i E[u_i|\mathbf{x}_i] p(\mathbf{x}_i)] = E[E^2[u_i|\mathbf{x}_i] p(\mathbf{x}_i)] \geq 0$ and equals zero only under actual model.
- Hence we can use a sample analogue of this expectation using the non-parametric estimate of the conditional expectation as a test statistic for selecting a model among ACE and the multiple linear regression.

Projection Pursuit Model

- As mentioned previously, in case of a random vector X, we can try to find functions (optimal transformations) g and f for Y and X instead of for each of the variables individually.
- In order to avoid the curse of dimensionality, the PPR model only considers functions of the form

$$f(X) = \sum_{i=1}^{M} f_i(\alpha_i^\top X),$$

  for some integer M, unit vectors α_i and functions f_i.
- Results of Diaconis and Shahshahani (1984) show that almost any multivariate function can be approximated as closely as needed by the representation above.
- The PPR model (also referred to as multiple index model) formally can be written as

$$Y = \sum_{i=1}^{M} f_i(\alpha_i^\top X) + \varepsilon, \qquad E[\varepsilon|X] = 0.$$

Generalized ACE/PPR

- Both the ACE and the PPR models can be generalized to the model which we call GACE.

$$g(Y) = \sum_{i=1}^{M} f_i(\alpha_i^\top X) + \varepsilon, \qquad E[\varepsilon|X] = 0.$$

- Estimation of the PPR model uses the back-fitting technique. One can use an algorithm similar to that of the ACE algorithm for estimation of the GACE model where we replace the step of finding transformation for each covariate, we use PPR estimation for getting covariate approximation.

- This model generalizes the ACE on the right and the PPR on the left.

Testing for More Models

- The hierarchy of the models discussed (linear regression $\rightarrow$ ACE $\rightarrow$ PPR $\rightarrow$ GACE) leads to models which are more and more flexible and so can give better fits for almost any data set.
- The more flexible the model the more is the chance for over-fitting and more is the loss in simplicity and interpretation.
- In order to avoid this problem of over-fitting, we propose to use the test described above for testing linear regression versus ACE first and then ACE versus GACE to settle on the lighter model for better estimation.

Post-Selection Inference

- A word of caution at the end. We are performing tests on finding which model to use and so we get a random model at the end.
- In this respect, the asymptotics or inference regarding the model thus obtained are a part of post-selection inference.
- Recent results of Berk (2013) show that the results for the model alone may not be directly applicable to the post-selection models.
- Certain care has to be practised in interpreting the results of the tests and the model obtained.

References

- Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. J. Amer. Statist. Assoc., 80(391):580–619.
- Davidson, R. and MacKinnon, J. G. (2004). Econometric theory and methods, volume 5. Oxford University Press New York.
- Diaconis, P. and Shahshahani, M. (1984). On nonlinear functions of linear combinations. SIAM J. Sci. Statist. Comput., 5(1):175–191.
- Dudzinski, M. L. and Mykytowycz, R. (1961). The eye lens as an indicator of age in the wild rabbit in australia. CSIRO Wildlife Research, 6:156–159.

References

- Martin, M. A. (2007). Bootstrap hypothesis testing for some common statistical problems: a critical evaluation of size and power properties. Comput. Statist. Data Anal., 51(12):6321–6342.
- Rényi, A. (1959). On measures of dependence. Acta Math. Acad. Sci. Hungar., 10:44–451.
- Zheng, J. X. (1996). A consistent test of functional form via nonparametric estimation techniques. Journal of Econometrics, 75(2):263 – 289.
- Berk, R., Brown, L. D., Buja, A., Zhang, K. and Zhao, L. (2013) "Valid Post-Selection Inference", Annals of Statistics, 41(2), 802–837.

Thank You!