FBST Regularization and Model Selection

Carlos A. B. Pereira cpereira@ime.usp.br
Julio M. Stern jstern@ime.usp.br
BIOINFO, NOPEF and IME-USP, University of Sao Paulo

ABSTRACT

We show how the Full Bayesian Significance Test (FBST) can be used as a model selection criterion. The FBST was presented by Pereira and Stern [38-42] as a coherent Bayesian significance test.

Key Words: Bayesian test; Evidence; Global optimization; Information; Model selection; Numerical integration; Posterior density; Precise hypothesis; Regularization. AMS: 62A15; 62F15; 62H15.

1 INTRODUCTION

The Full Bayesian Significance Test (FBST) is presented in [38-42] as a coherent Bayesian significance test. The FBST is intuitive and has a geometric characterization. It can be easily implemented using modern numerical optimization and integration techniques. The method is "Fully" Bayesian and consists in the analysis of credible sets. By Fully we mean that we need only the knowledge of the parameter space represented by its posterior distribution. The FBST needs no additional assumption, like a positive probability for the precise hypothesis, that generates the Lindley's paradox effect. The FBST regards likelihoods as the proper means for representing statistical information, a principle stated by Royall in [49] to simplify and unify statistical analysis.

Another important aspect of the FBST is its consistency with the "benefit of the doubt" juridical principle, [18]. This kind of principle establishes that there is no liability as long as there is a reasonable basis for belief, effectively placing the burden of proof on the plaintiff, who, in a lawsuit, must prove false a defendant's misstatement. Such a rule also prevents the plaintiff of making any assumption not explicitly stated by the defendant, or tacitly implied by existing law or regulation. The use of an a priori point mass on the null hypothesis, as on standard Bayesian tests, can be regarded as such an ad hoc assumption.

In the application presented in this paper, as well as in those in [23], [38-42], [55] it is desirable or necessary to use a test with the following characteristics:

- Be formulated directly in the original (natural) parameter space.
- Take into account the full geometry of the null hypothesis as a manifold (surface) imbedded in the whole parameter space.
- Have an intrinsically geometric definition, independent of any non-geometric aspect, like the particular parameterization of the (manifold representing the) null hypothesis being tested.
- Be consistent with the benefit of the doubt juridical principle (or safe harbor liability rule), i.e. consider in the "most favorable way" the claim stated by the hypothesis.
- Considering only the observed sample, allowing no ad hoc artifice (that could lead to judicial contention), like a positive prior probability distribution on the precise hypothesis.
- Consider the alternative hypothesis in equal standing with the null hypothesis, in the sense that increasing sample size should make the test converge to the right (accept/reject) decision.
- Give an intuitive and simple measure of significance for the null hypothesis, ideally, a probability in the parameter space.

FBST has all these theoretical characteristics and can be efficiently implemented with the appropriate computational tools. Moreover, as shown in [32], the FBST is also in perfect harmony with Bayesian decision theory of Rubin [50], in the sense that there are specific loss functions which render the FBST. Although we do can cast the FBST in a decision theoretic framework, we must stress this is optional. Actually, the FBST was originally defined in a pure operational form [39], based only on the benefit of the doubt juridical principle.

Interesting connections of some of the characteristics stated above, with ethics, epistemology, law, psychology and statistics can be found in [8], [10], [14], [16], [18], [21], [27], [28], [29-30], [33], [37-42], [43], [44], [49], [50], [52], [57]. Perhaps the most important characteristics concern the FBST symmetry. By one hand, as the sample size grows, the FBST converges to the Boolean indicator of the hypothesis truth, so achieving perfect symmetry. This is in sharp contrast with statements like “increase sample size to reject (accept) the hypothesis” made by many users of frequentist (standard Bayesian) tests. By the other hand, for small samples, the FBST exhibits an offset. This offset may counterbalance the intrinsically asymmetric formulation of a sharp hypothesis test, or, with the appropriate prior, may also be seen as a precaution or protection for the benefit of the doubt (or context equivalent) principle, keeping us “within the line of the law” (lifnim mishurat hadin).

2 FBST OPERATIONAL DEFINITION

We restrict the parameter space, Θ, to be always a subset of R^n, and the hypothesis is defined as a further restricted subset $\Theta_0 \subset \Theta \subseteq R^n$. Usually, Θ_0 is defined by vector valued inequality and equality constraints:

$$\Theta_0 = \{\theta \in \Theta \,|\, g(\theta) \leq 0 \wedge h(\theta) = 0\}.$$

We are interested in precise hypotheses, so we have at least one equality constraint, hence $dim(\Theta_0) < dim(\Theta)$. $f(\theta)$ is the posterior probability density function.

The computation of the evidence measure used on the FBST is performed in two steps, a numerical optimization step, and a numerical integration step. The numerical optimization step consists of finding an argument θ^* that maximizes the posterior density $f(\theta)$ under the null hypothesis. The numerical integration step consists of integrating the posterior density over the region where it is greater than $f(\theta^*)$. That is,

- Numerical Optimization step:

  $$\theta^* \in arg \max_{\theta \in \Theta_0} f(\theta) \ , \quad \varphi = f^* = f(\theta^*)$$

- Numerical Integration step:

  $$\kappa^* = \int_\Theta f_\varphi(\theta \,|\, d) d\theta$$

  where $f_\varphi(x) = f(x)$ if $f(x) \geq \varphi$ and zero otherwise.

Efficient computational algorithms are available, for local and global optimization as well as for numerical integration, and they can be implemented in very user friendly environments, [13], [17], [19], [20], [24], [26], [31], [34], [48].

If the probability of the set T^* is “large”, it means that the null set is in a region of low probability and the evidence in the data, $Ev(H) = 1 - \kappa^*$, is against the null hypothesis. On the other hand, if the probability of T^* is “small”, then the null hypothesis is in a region of high probability and the evidence in the data is in its favor.

3 MULTIPLE LINEAR REGRESSION MODEL

In the standard normal multiple linear regression model we have $y = X\beta + u$, X $n \times k$, where n is the number of observations, k the number of independent variables, β the regression coefficients, and u is a Gaussian white noise, so $E(u) = 0$ and $\mathrm{Cov}(u) = \sigma^2 I$, [7], [12], [15], [22], [62]. Using the diffuse prior $p(\beta, \sigma) = 1/\sigma$, the joint posterior probability density for the parameters and $\sigma \in [0, \infty[$ and $\beta \in]-\infty, \infty[^k$ is given by:

$$f(\beta, \sigma \,|\, y, X) = \tfrac{1}{\sigma^{n+1}}$$
$$. \exp\left(-\tfrac{1}{2\sigma^2}(\,(n-k)s^2 + (\beta - \hat{\beta})'X'X(\beta - \hat{\beta})\,)\right),$$

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1}X'y \ , \\ \hat{y} &= X\hat{\beta} \ , \\ s^2 &= (y - \hat{y})'(y - \hat{y})/(n-k) \ . \end{aligned}$$

The log-likelihood and its gradients are given by:

$$\begin{aligned} fl(\beta, \sigma) &= -(n+1)\log(\sigma) - \frac{1}{2\sigma^2}(\,(n-k)s^2 \\ &\quad +(\beta - \hat{\beta})'X'X(\beta - \hat{\beta})\,) \ , \\ \frac{\partial fl}{\partial \beta}(\beta, \sigma) &= -\frac{1}{\sigma^2}(\beta - \hat{\beta})' \, X'X \ , \\ \frac{\partial fl}{\partial \sigma}(\beta, \sigma) &= -\frac{n+1}{\sigma} + \frac{1}{\sigma^3}(\,(n-k)s^2 \\ &\quad +(\beta - \hat{\beta})'X'X(\beta - \hat{\beta})\,) \ . \end{aligned}$$

In the polynomial multiple linear regression model of order k, the dependent variable y is explained by powers 0 to k of the independent variable x, i.e., matrix $X_{i,j} = (x_i)^j$, $i = 1 \ldots n$, $j = 0 \ldots k$. Note the model of order k has dimension $d = k + 2$, with parameters $\beta_0, \beta_1, \ldots \beta_k$, and σ.

4 MODEL SELECTION AND REGULARIZATION

The multiple linear regression model family presented in the last section is typical, in the sense that it offers a class o models of increasing dimension, or complexity. This poses the problem of deciding, among all models in the family, the “better” adapted to our data. It is natural to look for a model that accomplishes a small empirical error, the estimated model error in the training data, R_{emp}. A regression model is estimated by minimizing the 2-norm empirical error. However, we can not select the “best” model based only on the empirical error, because we would usually select a model of high complexity. In general, when the dimensionality of the model is high enough, the empirical error can be made equal to zero by simple interpolation. It is a well known fact in statistics (or learning theory), that the prediction (or generalization) power of such high dimension models is poor. Therefore the selection criterion also has to penalize the model dimension. This is known as a regularization mechanism. Some model selection criteria define a penalized (or prediction) error $R_{pen} = r(d,n) * R_{emp}$, using a regularization (or penalization) function, $r(d,n)$, where d is the model dimension and n the number of training data. Common regularization functions, using $p = (d/n)$, are:

- Akaike's final prediction error:
  $\text{fpe} = (1+p)/(1-p)$,
- Schwartz' Bayesian criterion:
  $\text{sbc} = 1 + \ln(n)p/(2-2p)$,
- Generalized cross validation:
  $\text{gcv} = (1-p)^{-2}$,
- Shibata model selector:
  $\text{sms} = 1 + 2p$.

All those regularization functions are supported by theoretical arguments as well as by empirical performance; other regularization methods are model dependent, like Akaike information criterion (AIC), and Vapnik-Chervonenkis (VC) prediction error, [1], [2-4], [5], [6], [9], [11], [25], [35], [36], [45], [46-47], [51], [53], [54], [55], [58], [59], [60-61].

We can use the FBST as a model selection criterion, testing the hypothesis of some of its parameters being null, and using the following version of the "Ockham razor: Do not include in the model a new parameter unless there is strong evidence that it is not null."

The FBST selection criterion has a intrinsic regularization mechanism, under some general circumstances discussed later.

Consider, as a simple example, the d-dimensional vector $x \,|\, \beta$ with normal distribution, $N(\beta, I)$, and suppose we want to use the FBST to test the hypothesis $H : \beta_1 = 0$. Consider the priori for β as $N(0,I)$. The posterir distribution of β is $N(x/2, (1/2)I)$, DeGroot (1990). The probability of the H.P.D. region tangent to the null hypothesis manifold, $H : \beta_1 = 0$, is $\kappa^* = Pr\{\chi^2_d \leq x_1^2/2\}$.

The chi-square density with d degrees of freedom is

$$f_d(x) = \frac{x^{(d/2-1)}\, exp(-x/2)}{2^{d/2}\;,\;(d/2)}\;,$$

with $E(f_d(x)) = d$, $\text{Var}(f_d(x)) = 2d$.

So, subtracting the mean and dividing by the standard deviation,

$$\kappa^* = \Pr\left\{ \frac{\chi^2_d}{\sqrt{2d}} - \sqrt{\frac{d}{2}} \;\leq\; \frac{x_1^2}{2\sqrt{2d}} - \sqrt{\frac{d}{2}} \right\}$$

Using the central limit theorem, as $d \to \infty$,

$$\Pr\left\{ \frac{\chi^2_d}{\sqrt{2d}} - \sqrt{\frac{d}{2}} \;\leq\; t \right\} \approx \Phi\,(t)$$

making it is easy to see that, as $d \to \infty$,
$evid(H) = 1 - \kappa^* \to 1$.

The intrinsic regularization of the example above is partially explained by simple geometry related to symmetry properties of the model density function, [15]. The Normal distribution is spherically (or elliptically) symmetric, i.e., the (scaled) distribution is invariant under action of the orthogonal group, whose invariant metric is the 2-norm, whereas the unitary volume in R^d is defined by a cube, a sphere in the infinite-norm. The volumes of the unitary radius d-dimensional (2-norm) sphere and cube are, $\text{Vol}(S_d) = (2/d)\pi^{d/2}/, (d/2)$ and $\text{Vol}(C_d) = 2^d$. These volumes ratio make it easy to see that the model invariant sphere has comparatively "small volume in high dimension".

$$\frac{\text{Vol}(S_d)}{\text{Vol}(C_d)} \;=\; \frac{2}{d}\left(\frac{\pi}{4}\right)^{d/2} \Big/ \;,\; \left(\frac{d}{2}\right)$$

5 NUMERICAL EXAMPLES

In the classical example presented in Sakamoto [51], we want to fit a linear regression polynomial model of order k,

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 \ldots + \beta_k x^k + N(0, \sigma I)$$

through $n = 21$ points, (x_i, y_i). This example was produced by Sakamoto simulating the i.i.d. stochastic process

$$y_i = \exp((x_i - 0.3)^2) - 1 + 0.1 * N(0,1)\;,$$

a structure that can not be expressed exactly as a finite order linear regression polynomial model.

Table 1 presents the empirical error, $||y - \hat{y}||_2^2/n$ for models of order $k = 0, \ldots 5$, several regularizations defined in section 5, and the Akaike information criterion (AIC) computed by Sakamoto. Table 1 also presents the FBST for the hypothesis $H : \beta_k = 0$. The FBST is computed with an absolute numerical error of less than 1%.

We see that all the penalized errors, as well as AIC criterion, are minimized for $k = 2$. The FBST gives strong evidence against $\beta_k = 0$, $k = 0, 1\,2$, and week evidence for non-null parameters of higher order. So all selection criteria elect the second order model as the better adapted to the example at hand. The situation is illustrated by Figure 2, presenting for models of order $k = 1 \ldots 4$, the data points, $(+)$, the fitted maximum posterior density model of order k, $(*)$, and the fitted maximum posterior model of order k making $\beta_k = 0$, (O).

As a second example, also taken from Sakamoto [51], we want to fit the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + N(0, \sigma I)$$

where y is the average daily minimum January temperature for 20 cities in Japan, and the explanatory variables are, respectively, the cities' latitude, altitude and longitude. Sakamoto gives a detailed analysis of the model. The independent variables have already been ordered in decreasing explanatory power. Table 2 presents the selection criteria for the models $\beta_j = 0$, $j > k$, and the FBST for the hypothesis $H : \beta_k = 0$.

It is interesting to follow the discussion in Sakamoto about the "instability" of the parameter β_2, due to the few observations in higher altitude, only 6 above 100m. This explains why the FBST stays at 5% when testing $\beta_2 = 0$. Skamoto continues the discussion adding supplementary data of 19 cities, including 8 cities above 100m. The FBST for $H : \beta_2 = 0$, using all data, is less than 1%.

REFERENCES

[1] Anthony,M. Biggs,N. (1992). *Computational Learning Theory.* Cambridge Univ. Press.

[2] Akaike,H. (1969). Fitting Autoregressive Models for Prediction. *Ann. Inst. Stat. Math,* 21, 243–247.

[3] Akaike,H. (1970). Statistical Prediction Identification. *Ann. Inst. Stat. Math,* 22, 203–217.

[4] Akaike,H. (1974). A New Look at the Statistical Model Identification. *IEEE Trans. Autom. Control.* 19, 716–723.

[5] Barron,A.R. (1984) Predicted Squared Error: A Criterion for Automatic Model Selection. in Farlow,S.J. *Self-Organizing Methods in Modeling: GMDH-type Algorithms.* Basel: Marcel Dekker.

[6] Breiman,L. Friedman,J.H. Stone,C.J. (1984). *Classification and Regression Trees.* London: Chapman and Hall.

[7] Box, G.E.P., Tiao, G.C. (1973). *Bayesian Inference in Statistical Analysis..* N.Y: Wiley.

[8] Carnap, R. (1962). *Logical Foundations of Probability.* Univ. Chicago Press.

[9] Cherkaasky,V. Mulier,F. (1998). *Learning from Data.* N.Y: Wiley.

[10] Cox, D.R. (1977). The role of significance tests. *Scand J Statist* 4, 49-70

[11] Craven,P. Wahba,G. (1979). Smoothing Noisy Data with Spline Functions. *Numerische Matematik,* 31, 377–403.

[12] DeGroot,M.H. (1970). *Optimal Statistical Decisions.* N.Y: McGraw-Hill.

[13] Deak,I. (1990). *Random Number Generation and Simulation.* Mathematical Methods of Operations Research 4. Budapest: Akademiai Kiado.

[14] d'Espagnat,B. (1976). *Conceptual Fundations of Quantum Mechanics.* London: W.A.Bemjamin.

[15] Fang,K.T. Kotz,S. Ng,K.W. (1990). *Symmetric Multivariate and Related Distributions.* London: Chapman and Hall.

[16] Feyerabend,P. (1996). *Against Method.* Verso Books.

[17] Galassi, M., Davies J., Theiler, J., Gough, B., Priedhorsky, R., Jungman, G., Booth, M. (1999). *GSL - GNU Scientific Library Reference Manual V-0.5.* WWW: lanl.gov.

[18] Gaskins, R.H. (1992). *Burdens of Proof in Modern Discourse.* New Haven: Yale Univ. Press.

[19] Gentle, J.E. (1998). *Random Number Generator and Monte Carlo Methods.* N.Y: Springer.

[20] Gomez, C. (1999). *Engineering and Scientific Computing with Scilab.* Berlin: Birkhäuser.

[21] Good, I.J. (1983). *Good Thinking: The Foundations of Probability and its Applications.* Univ. Minnesota Press.

[22] Hjorth,J.S.U. (1984). *Computer Intensive Statistical Methods.* London: Chapman and Hall.

[23] Irony, T.Z., Lauretto, M., Pereira, C.A.B., Stern, J.M. (2000). *A Weibull Wearout Test: Full Bayesian Approach.* RT-MAC-2000-5, Dept. Comp. Science, Univ. Sao Paulo.

[24] Johnson, M.E. (1980). *Multivariate Statistical Simulation.* N.Y: Wiley.

[25] Kearns,M.J. Vazirani,U.V. (1994). *Computational Learning Theory.* Cambridge: MIT Press.

[26] Kelly,C.T. (1999). *Iterative Methods for Optimization.* Philadelphia: SIAM.

[27] Kuhn,T.S. (1996). *The Structure of Scientific Revolutions.* Univ. Chicago Press.

[28] Kyburg,H.E. Smokler,H.E. (1963). *Studies in Subjective Probability.* N.Y: John Wiley.

[29] Lindley, D.V. (1957). A Statistical Paradox. *Biometrika* 44, 187-192.

[30] Lindley, D.V. (1978). The Bayesian Approach. *Scand J Statist* 5, 1-26.

[31] Luenberger, D.G. (1984). *Linear and Nonlinear Programming.* Reading: Addison-Wesley.

[32] Madruga, M.R. Esteves, L.G., Wechsler, S. (2000). On the Bayesianity of Pereira-Stern Tests. RT-MAE-2000-10, Dept. Statistics, Univ. Sao Paulo. To appear in *Test.*

[33] Maimon, M. (1992). Shemona Perakim, Hakdama Lemassechet Avot. Sao Paulo: Maayanot.

[34] McDonald, R.P., Swaminathan. H. (1973). A Simple Matrix Calculus with Applications to Multivariate Analysis. *General Systems*, 18, 37-54

[35] Michie,D. Spiegelhalter,D.J. TaylorC.C (1994). *Machine Learning, Neural and Statistical Classification.* N.Y: Ellis Horwood.

[36] Mueller,W. Wysotzki,F. (1994). Automatic Construction of Decision Trees for Classification. *Annals of Operations Research,* 52, 231-247.

[37] Pereira, C.A.B, Wechsler, S. (1993). On the Concept of *p-value. Braz J Prob Statist* 7, 159-177.

[38] Pereira, C.A.B., Stern, J.M. (1999a). A Dynamic Software Certification and Verification Procedure. *Proc. ISAS/SCI-99.* II, 426-435.

[39] Pereira, C.A.B., Stern, J.M. (1999b). Evidence and Credibility: Full Bayesian Significance Test for Precise Hypotheses. *Entropy* 1, 69-80.

[40] Pereira, C.A.B., Stern, J.M. (2000a). *Full Bayesian Significance Test: the Behrens-Fisher and Coefficients of Variation Problems.* RT-MAC-2000-4, Dept. Comp. Science, Univ. Sao Paulo. Also in ISBA 2000 - Internt. Soc. for Bayesian Anal. 6th World Meeting, Knossos.

[41] Pereira, C.A.B., Stern. J.M. (2000b). *Model Selection: Full Bayesian Approach.* RT-MAE-2000-17, Dept. Statistics, Univ. Sao Paulo. To appear in *Environmetrics.*

[42] Pereira, C.A.B., Stern. J.M. (2000c). *Intrinsic Regularization in Model Selection Using the Full Bayesian Significance Test.* RT-MAC-2000-5, Dept. Comp. Science, Univ. Sao Paulo.

[43] von Plato, J. (1998). *Creating Modern Probability: Its Mathematics, Physics and Philosophy in Historical Perspective.* Cambridge Univ. Press.

[44] Popper, K.R. (1989). *Conjectures and Refutations: The Growth of Scientific Knowledge.* London: Routledge.

[45] Quinlan,J.R. (1986) Induction of Decision Trees. *Machine Learning,* 1, pp.221–234.

[46] Rissanen,J. (1978). Modeling by Shortest Data Description. *Automatica,* 14, 465–471.

[47] Rissanen,J. (1989). *Stochastic Complexity in Statistical Inquiry.* N.Y: World Scientific.

[48] Rogers, G.S. (1980). *Matrix Derivatives.* Lecture Notes in Statistics Vol. 2. Basel: Marcel Dekker.

[49] Royall, R. (1997). *Statistical Evidence: A Likelihood Paradigm.* London: Chapman & Hall.

[50] Rubin, H. (1987). A Weak System of Axioms for "Rational" Behavior and the Non-Separability of Utility from Prior. *Stat. and Decisions* 5, 47-58.

[51] Sakamoto,Y. Ishiguro,M. Kitagawa,G. (1986). *Akaike Information Criterion Statistics.* Dordrecht: Reidel-Kluwer.

[52] Savage, L.J. (1962). *The Foundations of Statistical Inference.* London: Methuen.

[53] Schwartz,G. (1978). Estimating the Dimension of a Model. *Annals of Statistics,* 6, 461–464.

[54] Shibata,R. (1981). An Optimal Selection of Regression Variables. *Biometrika,* 68, 45–54.

[55] Stern,J.M. Ribeiro,C.O. Lauretto,M.S. Nakano,F. (1998). REAL: Real Attribute Learning Algorithm. *Proc. ISAS/SCI-98* 2, 315–321.

[56] Stern,J.M. (2001). *The Full Bayesian Significance Test for the Covariance Structure Problem.* RT-MAC-2001-3, Dept. Comp. Science, Univ. Sao Paulo. *Proc. ISAS/SCI-2001.*

[57] Tomonaga,S.I. (1970). *Quantum Mechanics.* Amsterdam: Elsevier.

[58] Unger,S. Wysotzki,F. (1981). *Lernfaehige Klassifizierungssysteme.* Berlin: Akademie Verlag.

[59] Vidyasagar,M. (1997). *A Theory of Learning and Generalization.* London: Springer.

[60] Vapnik,V.N. (1995). *The Nature of Statistical Learning Theory.* N.Y: Springer.

[61] Vapnik,V.N. (1998). *Statistical Learning Theory: Inference for Small Samples.* N.Y: Wiley.

[62] Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics.* N.Y: Wiley.

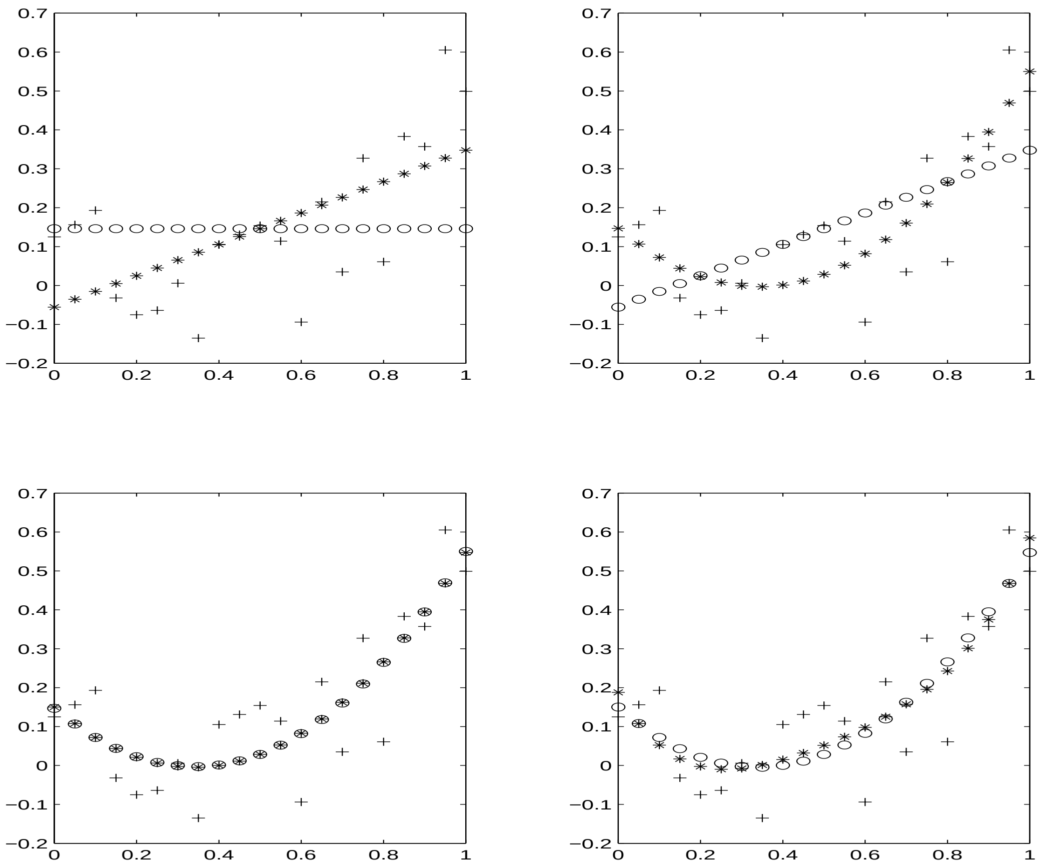

Figure 1: Fitted model for order $k = 1 \ldots 4$, also with $\beta_k = 0$

Order	EMP	FPE	SBC	GCV	SMS	AIC	FBST
0	0.03712	0.04494	0.04307	0.04535	0.04419	-07.25	0.00
1	0.02223	0.02964	0.02787	0.03025	0.02858	-20.35	0.00
2	0.01130	0.01661	0.01534	0.01724	0.01560	-32.13	0.00
3	0.01129	0.01835	0.01667	0.01946	0.01667	-30.80	1.00
4	0.01088	0.01959	0.01751	0.02133	0.01710	-29.79	0.99
5	0.01087	0.02173	0.01913	0.02445	0.01811	-27.86	1.00

Table 1: Selection Criteria for the Polynomial Model

k	EMP	FPE	SBC	GCV	SMS	AIC	FBST
0	32.66	39.92	38.10	40.32	39.19	130.5	0.00
1	8.734	11.82	11.04	12.09	11.35	106.5	0.00
2	2.464	3.695	3.386	3.849	3.449	82.8	0.05
3	2.462	4.103	3.691	4.377	3.693	84.8	1.00

Table 2: Selection Criteria for the Japan Model