ENVIRONMETRICS
Environmetrics 2001; **12**: 559–568 (DOI: 10.1002/env.482)

Model selection: Full Bayesian approach

Carlos Alberto de Bragança Pereira[*,†] and Julio Michael Stern[‡]

BIOINFO and IME-USP – University of Sao Paulo, Brazil

SUMMARY

We show how the Full Bayesian Significance Test (FBST) can be used as a model selection criterion. The FBST was presented in Pereira and Stern as a coherent Bayesian significance test. Copyright © 2001 John Wiley & Sons, Ltd.

KEY WORDS: evidence; global optimization; information; model selection; numerical integration; precise hypothesis testing; regularization; soil contamination; toxic waste

1. INTRODUCTION

The Full Bayesian Significance Test (FBST) was presented in Pereira and Stern (1999b, 2000) 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 'Full' Bayesian and consists in the analysis of credible sets. By 'Full' 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. Given a model, we can use the FBST to test if one of its parameters is null. Such a test is used as the basis for a model selection procedure.

2. MOTIVATION FOR THE FBST

In order to better illustrate the FBST we discus a well known problem. Given a sample from a normal distribution with unknown parameters, we want to test if the standard deviation is equal to a constant. The hypothesis $\sigma = c$ is a straight line. We have a precise hypothesis since it is defined by a manifold (surface) of dimension (one) strictly smaller than the dimension of the parameter space (two).

It can be shown that the conjugate family for the normal distribution is a family of bivariate distributions, where the conditional distribution of the mean, μ, for a fixed precision; $\rho = 1/\sigma^2$, is normal, and the marginal distribution of the precision, ρ, is gamma (DeGroot, 1970). We use the

[*] Correspondence to: C. A. B. Pereira, BIOINFO and IME-USP, University of Sao Paulo, Sao Paulo, Brazil.
E-mail: [†]cpereira@ime.usp.br
E-mail: [‡]jstern@ime.usp.br

Received 15 September 2000
Revised 3 January 2001

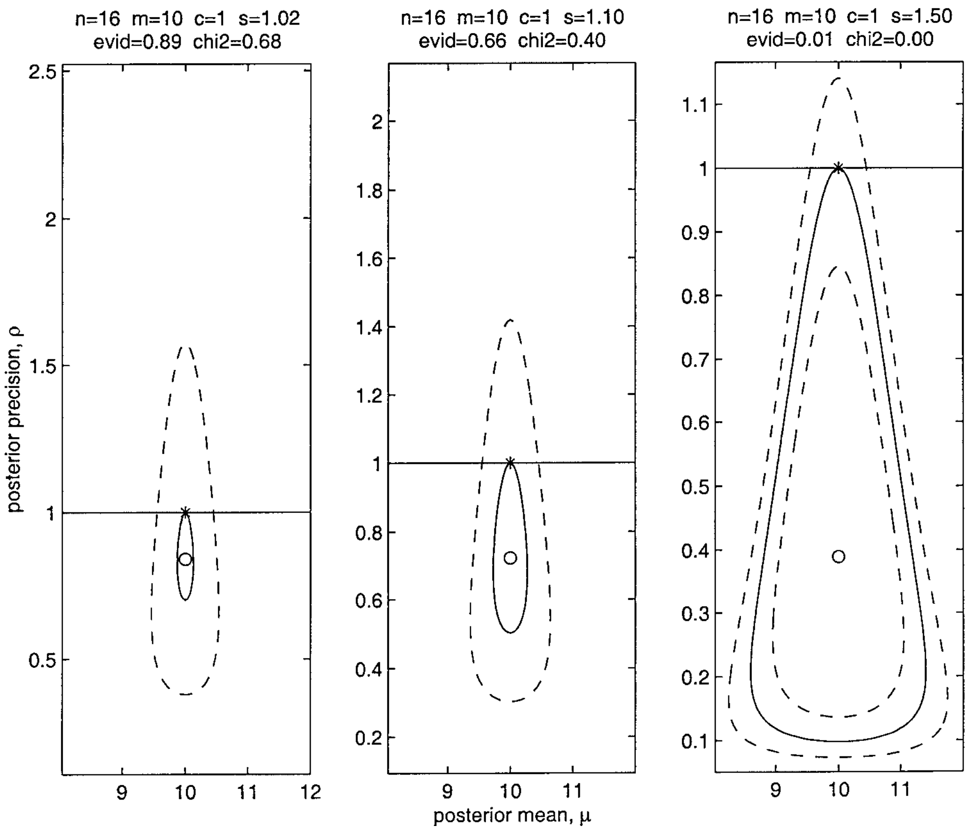

Figure 1. Tangent and other highest probability density sets

standard improper priors, uniform on $]-\infty, +\infty[$ for μ, and $1/\rho$ on $]0, +\infty[$for ρ, in order to get a fair comparison with p-values (DeGroot, 1970). Hence we have the parameter space, hypothesis and posterior joint distribution:

$$\Theta = \{(\mu, \rho) \in R \times R_+\}, \quad \Theta_0 = \{(\mu, \rho) \in \Theta \mid \rho = c\}$$

$$f(\mu, \rho \mid x) \propto \sqrt{\rho}\exp(-n\rho(\mu - u)^2/2)\exp(-b\rho)\rho^{a-1}$$

$$x = [x_1 \ldots x_n], \; a = \frac{n-1}{2}, \; u = \frac{1}{n}\sum_{i=1}^{n} x_i, \; b = \frac{n}{2}\sum_{i=1}^{n}(x_i - u)^2.$$

In Figure 1 we plot some level curves of the posterior density function, including the level curve tangent to the hypothesis manifold. At the tangency point, θ^*, the posterior density attains its maximum, f^*, on the hypothesis. The interior of the tangent level curve, T^*, includes all points with posterior density greater than f^*, i.e. it is the highest probability density set tangent to the hypothesis.

Also in Figure 1 we test $c = 1$ with $n = 16$ observations of mean $m = 10$ and standard deviation $s = 1.02$, 1.1, and 1.5. We give the FBST evidence, *Ev*, and the standard χ^2-test, *chi*2.

The posterior probability of T*, κ^*, gives an indication of inconsistency between the posterior and the hypothesis: small values of κ^* indicate that the hypothesis traverses high density regions, favoring the hypothesis. Therefore we define $\mathrm{Ev(H)} = 1 - \kappa^*$ as the measure of evidence (for the precise hypothesis).

Of course this example is a mere illustration: there is no need of new methods to test the standard deviation of a normal distribution. However, efficient numerical optimization and integration computer programs make it straightforward to extend the FBST to more complex structures. A formal definition of the FBST, several applications, and some philosophical considerations can be found in Pereira and Stern (1999b, 2000) and Irony *et al.* (2000). In the next section we give an operational construction of the FBST.

3. 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:

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

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:

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

Numerical integration step:

$$\kappa^* = \int_{\Theta} f_{\varphi}(\theta \mid d) \mathrm{d}\theta,$$

where $f_{\varphi}(x) = f(x)$ if $f(\mathrm{x}) \geq \varphi$ 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 (Pereira and Stern, 1999b, 2000; Irony *et al.*, 2000).

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, $\mathrm{Ev(H)} = 1 - \kappa^*$, 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.

4. 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 white noise, i.e. a Gaussian noise with $\mathrm{E}(u) = 0$ and $\mathrm{Cov}(u) = \sigma^2 I$ (Zellner, 1971). 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:

$$\begin{aligned}
f(\beta, \sigma \mid y, X) &= \frac{1}{\sigma^{n+1}} \exp\left(-\frac{1}{2\sigma^2} ((n-k)s^2 + (\beta - \hat{\beta})' X' X (\beta - \hat{\beta})) \right) \\
\hat{\beta} &= (X'X)^{-1} X y' \\
\hat{y} &= X\hat{\beta} \\
s^2 &= (y - \hat{y})'(y - \hat{y})/(n-k).
\end{aligned}$$

The log-likelihood and its gradients are given by:

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

A selection of interesting environmetric statistical studies using regression and related models can be found in Pratt (1974) and El-Shaarawi (1991).

5. 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 of 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. A regression model is estimated by minimizing the 2-norm empirical error. However, we cannot 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 models selection criteria define a penalized (or prediction) error $R = r(d, n) * R$, using a regularization (or penalty) 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: FPE $= (1 + p)/(1 - p)$
- Schwarz' Bayesian criterion: SBC $= 1 + \ln(n)p/(2 - 2p)$

- Generalized cross-validation: GCV $= (1-p)^{-2}$
- Shibata model selector: 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, Akaike (1970 and 1974), Barron (1984), Breiman *et al.* (1984), Cherkassky and Mulier (1998), Craven and Wahba (1979), Michie *et al.* (1994), Mueller and Wysotzki (1994), Shibata (1981), Stern *et al.* (1998), Rissanen (1978), Schwarz (1978), Unger and Wysotzki (1981), Vidyasagar (1997) and Vapnik (1995, 1998).

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 it is not null.'

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

Consider, as a simple example, the d-dimensional vector x with normal distribution, (β, I), and suppose we want to use the FBST to test the hypothesis: $\beta_1 = 0$. Consider the normal distribution with parameter $(0, I)$ on a prior for β. The posterior distribution of β is $(x/2,\ (1/2)\ I)$, DeGroot (1970). The probability of the H.P.D. region tangent to the null hypothesis manifold,: $\beta_1 = 0$, is $\kappa^* = \{\chi_d^2 \leq x_1^2/2\}$.

The chi-square density with d degrees of freedom is

$$f_d(x) = \frac{x^{(d/2-1)}(-x/2)}{2^{d/2}\Gamma(d/2)}, \quad \text{with } (\, f_d(x)) = d, \quad \mathrm{Var}\,(\, f_d(x)) = 2d,$$

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

$$\kappa^* = \Pr\left\{ \frac{\chi_d^2}{\sqrt{2d}} - \sqrt{\frac{d}{2}} \leq \frac{x_1^2}{2\sqrt{2d}} - \sqrt{\frac{d}{2}} \right\}.$$

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

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

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

The intrinsic regularization of the example above is partially explained by simple geometry related to symmetry properties of the model density function (Fang *et al.*, 1990). 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 $\mathrm{Vol}(S_d) = (2/d)\pi^{d/2}/\Gamma(d/2)$ and $\mathrm{Vol}(C_d) = 2^d$. These volume ratios make it easy to see that the model invariant sphere has comparatively 'small volume inhigh dimension',

$$\frac{\mathrm{Vol}(S_d)}{\mathrm{Vol}(C_d)} = \frac{2}{d}\left(\frac{\pi}{4}\right)^{d/2} \Big/ \Gamma\left(\frac{d}{2}\right).$$

Table 1. Hildebrand and Liu's production data

i	Value	Labor	Capital	i	Value	Labor	Capital
1	657.29	162.31	279.99	15	1917.55	536.73	2109.34
2	935.93	214.43	542.50	16	9849.17	1564.83	13989.55
3	1110.65	186.44	721.51	17	1088.27	214.62	884.24
4	1200.89	245.83	1167.68	18	8095.63	1083.10	9119.70
5	1052.68	211.40	811.77	19	3175.39	521.74	5686.99
6	3406.02	690.61	4558.02	20	1653.38	304.85	1701.06
7	2427.89	452.79	3069.91	21	5159.31	835.69	5206.36
8	4257.46	714.20	5585.01	22	3378.40	284.00	3288.72
9	1625.19	320.54	1618.75	23	592.85	150.77	357.32
10	1272.05	253.17	1562.08	24	1601.98	259.91	2031.93
11	1004.45	236.44	662.04	25	2065.85	497.60	2492.98
12	598.87	140.73	875.37	26	2293.87	275.20	1711.74
13	853.10	145.04	1696.98	27	745.67	137.00	768.59
14	1165.63	240.27	1078.79				

6. NUMERICAL EXAMPLES

The classical data set in Table 1, by Hildebrand and Liu (1957), presents the value added, v, labor input, p_1, and capital (gross value of plant and equipment), p_2, for several industrial establishments in the primary metals sector. This data set has subsequently been used by a number of authors, like Aigner *et al.* (1977) and Greene (1991).

The classical econometric Cobb–Douglas production function assumes constant elasticity, i.e. a constant relative growth in value for a given production input. Defining $y_i = \ln(v_i)$ and $x_{i,j} = \ln(p_{i,j})$, $i = 1 \ldots n$, $j = 1 \ldots 2$,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + (0, \sigma I),$$
$$\frac{\partial v / v}{\partial p_j / p_j} = \frac{\partial \ln(v)}{\partial \ln(p_j)} = \beta_j.$$

The log-linear regression model has great importance in econometrics as well as in environmetric modeling (El-Shaarawi and Viveros, 1997).

In the following, we will compare the Cobb–Douglas first order log-linear model with the more general second order Translog model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2/2 + \beta_4 x_2^2/2 + \beta_5 x_1 x_2 + (0, \sigma I).$$

Table 2 presents the empirical error, EMP $= ||y - \hat{y}||_2^2/n$, for several sub-models of the Translog model. The first column presents the index list of non-null parameters, $\beta_0 \ldots \beta_5$, in the sub-model. Table 2 also presents several regularizations defined in Section 5, and the FBST for the hypothesis stating that the last parameter in the sub-model index list is equal to zero. The FBST is computed with an absolute numerical error of 1 per cent or less.

We see that all the penalized errors are minimized for the sub-model 0, 1, 2. The FBST gives strong evidence against $\beta_k = 0$, $k = 0, 1, 2$, and weak evidence for non-null parameters of higher order. So all

Table 2. Selection criteria, $\times 10^{-2}$, and FBST for Trabslog model

Sub-model	EMP	FPE	SBC	GCV	SMS	FBST
0,1	5.701	7.126	6.876	7.216	6.968	0.00
0,1,2	3.154	4.251	4.058	4.347	4.089	0.00
0,1,2,4	3.004	4.369	4.129	4.524	4.116	0.99
0,1,2,5	3.122	4.542	4.292	4.703	4.279	1.00
0,1,2,3	3.149	4.581	4.329	4.743	4.316	1.00

selection criteria elect the Cobb–Douglas sub-model as the better adapted to the example at hand. The situation is illustrated in Figure 2, presenting for some of the sub-models the data points, (+), the fitted maximum posterior density model with free parameters in the list, (*), and the fitted maximum posterior model with all but the last free parameters in the list, (O).

As a second example we model the chemical concentration of a toxic industrial contaminant, in a deactivated solid waste land fill in the city of Cubatao, Brazil. The old land fill is now located inside a 'preservation' area, and the model is being used to weight the relative adverse impacts of processing (cleaning by incineration) the contaminated soil, versus leaving it in its present condition.

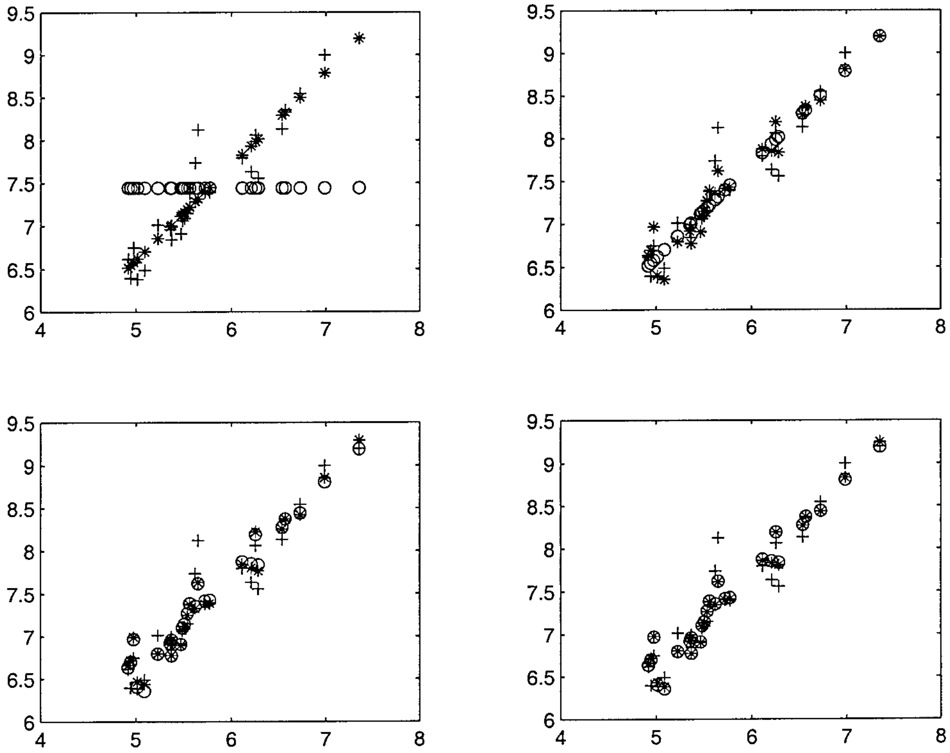

Figure 2. Models 0,1; 0,1,2; 0,1,2,4; and 0,1,2,5

Table 3. Selection criteria, and FBST for contamination model

Sub-model	EMP	FPE	SBC	GCV	SMS	FBST
0,4	38.924	41.918	42.213	41.975	41.807	0.00
0,4,3	10.450	11.535	11.642	11.563	11.482	0.00
0,4,3,1	10.384	11.750	11.885	11.795	11.666	0.99
0,4,3,5	10.431	11.804	11.939	11.849	11.719	1.00
0,4,3,2	10.436	11.809	11.944	11.854	11.724	1.00

We consider a second order response surface, as in Chen and Shao (1999),

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + (0, \sigma I).$$

The independent variables have already been ordered in decreasing explanatory power. Table 3 is similar to Table 1, presenting the selection criteria for several sub-models and the FBST for the hypothesis: $\beta_k = 0$ for the last parameter in the sub-model. The elliptic model is the most adequate to the data at hand, as may be suggested from the contamination level contour plot at Figure 3, for our 81 field samples.

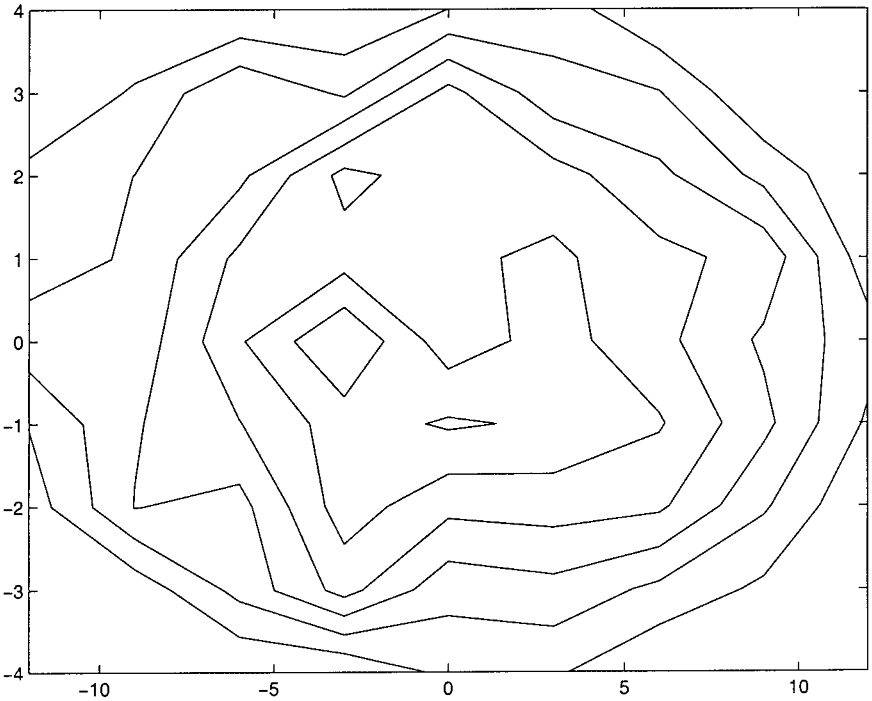

Figure 3. Contamination level contour plot

7. FINAL REMARKS

The objective of this article was to present a model selection criterion based on the Full Bayesian Significance Test (FBST) for the precise hypothesis. The authors are currently working on several applications with models of higher dimension, where the FBST is the only exact test known. In some of these models the FBST is also used for automatic model selection. In these applications, as well as in those in Pereira and Stern (1999a, 1999b, 2000) and Irony *et al.* (2000), it is desirable or necessary to use a test with the following characteristics:

- be formulated directly in the 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 used
- 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 Madruga *et al.* (2000), the FBST is also in perfect harmony with the Bayesian decision theory of Rubin (1987), in the sense that there are specific loss functions which render the FBST. Finally, we notice that statements like 'increase sample size to reject (accept) the hypothesis' made by many users of frequentist (standard Bayesian) tests do not hold for the FBST.

REFERENCES

Aigner D, Lovel K, Schnidt P. 1977. Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics* **6**: 21–37.

Akaike H. 1970. Statistical prediction identification. *Annals of the Institute of Statistical Mathematics* **22**: 203–217.

Akaike H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**: 716–723.

Barron AR. 1984. Predicted squared error: a criterion for automatic model selection. *Self-Organizing Methods in Modeling: GMDH-type Algorithms.* Forlow SJ (ed.), Marcel Dekker: Basel.

Breiman L, Friedman JH, Stone CJ. 1984. *Classification and Regression Trees.* Chapman & Hall: London.

Chen L, Shao J. 1999. Bootstrap minimum cost estimation of the average chemical concentration in contaminated soil. *Environmetrics* **10**: 153–161.

Cherkassky V, Mulier F. 1998. *Learning from Data.* Wiley: NY.

Craven P, Wahba G. 1979. Smoothing noisy data with spline functions. *Numerische Matematik* **31**: 377–403.

DeGroot MH. 1970. *Optimal Statistical Decisions.* McGraw-Hill: NY.

El-Shaarawi AH. 1991. *Statistical Methods for Environmental Sciences.* Kluwer: Dordrecht.

El-Shaarawi AH, Viveros R. 1997. Inference about the mean in log-regression with environmental applications. *Environmetrics* **8**: 569–582.

Evans M. 1997. Bayesian inference procedures derived via the concept of relative surprise. *Communications in Statistics* **26**: 1125–1143.

Fang KT, Kotz S, Ng KW. 1990. *Symmetric Multivariate and Related Distributions.* Chapman & Hall: London.

Greene WH. 1991. *Econometric Analysis.* Maxwell McMillan: NY.

Hildebrand G, Liu T. 1957. *Manufacturing Production Functions in the United States.* Cornell University, Ithaca.

Irony TZ, Lauretto M, Pereira CAB, Stern JM. 2000. *A Weibull Wearout Test: Full Bayesian Approach. Technical Report RT-MAC-2000-5*, Department of Computer Science, University of Sao Paulo.

Madruga MR, Esteves LG, Wechsler S. 2000. *On the Bayesianity of Pereira–Stern Tests. Technical Report RT-MAE-2000-10*, Department of Statistics, University of Sao Paulo.

Michie D, Spiegelhalter DJ, Taylor CC. 1994. *Machine Learning, Neural and Statistical Classification*. Ellis Horwood: NY.

Mueller W, Wysotzki F. 1994. Automatic construction of decision trees for classification. *Annals of Operations Research* **52**: 231–247.

Pereira CAB, Stern JM. 1999a. A dynamic software certification and verification procedure. *Procedure ISAS-99-International Conference on Information Systems Analysis and Synthesis*, vol. **2**: 426–435.

Pereira CAB, Stern JM. 1999b. Evidence and credibility: full Bayesian significance test for precise hypotheses. *Entropy* **1**: 69–80.

Pereira CAB, Stern JM. 2000. *Full Bayesian Significance Test: the Behrens–Fisher and Coefficients of Variation Problems. Technical Report RT-MAC-2000-4*, Department of Computer Science, University of Sao Paulo. Also presented at ISBA 2000 – International Society for Bayesian Analysis 6th World Meeting.

Pratt JW. 1974. *Statistical and Mathematical Aspects of Pollution Problems.* Marcel Dekker: NY.

Rissanen J. 1978. Modeling by shortest data description. *Automatica* **14**: 465–471.

Rubin H. 1987. A weak system of axioms for 'rational' behavior and the non-separability of utility from prior. *Statistics and Decisions* **5**: 47–58.

Shibata R. 1981. An optimal selection of regression variables. *Biometrika* **68**: 45–54.

Schwartz G. 1978. Estimating the dimension of a model. *Annals of Statistics* **6**: 461–464.

Stern JM, Ribeiro CO, Lauretto MS, Nakano F. 1998. REAL: real attribute learning algorithm. *Procedings of ISAS-98 – International Conference on Information Systems Analysis and Synthesis* vol. **2**: 315–321.

Unger S, Wysotzki F. 1981. *Lernfaehige Klassifizierungssysteme.* Akademie Verlag: Berlin.

Vidyasagar M. 1997. *A Theory of Learning and Generalization.* Springer: London.

Vapnik VN. 1995. *The Nature of Statistical Learning Theory.* Springer: NY.

Vapnik VN. 1998. *Statistical Learning Theory: Inference for Small Samples.* Wiley: NY.

Zellner A. 1971. *An Introduction to Bayesian Inference in Econometrics.* Wiley: NY.