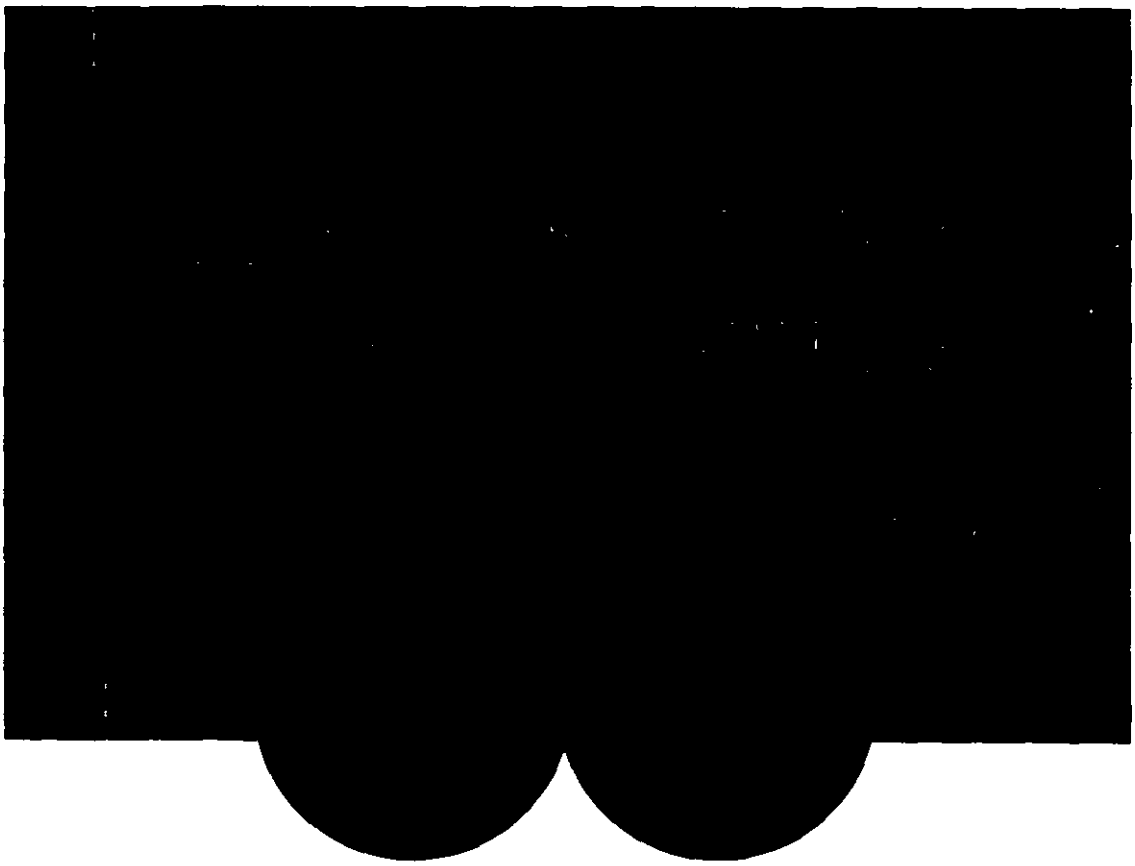


MEASURING THE EFFECT OF OBSERVATIONS USING THE POSTERIOR AND THE INTRINSIC BAYES FACTORS WITH VAGUE PRIOR INFORMATION

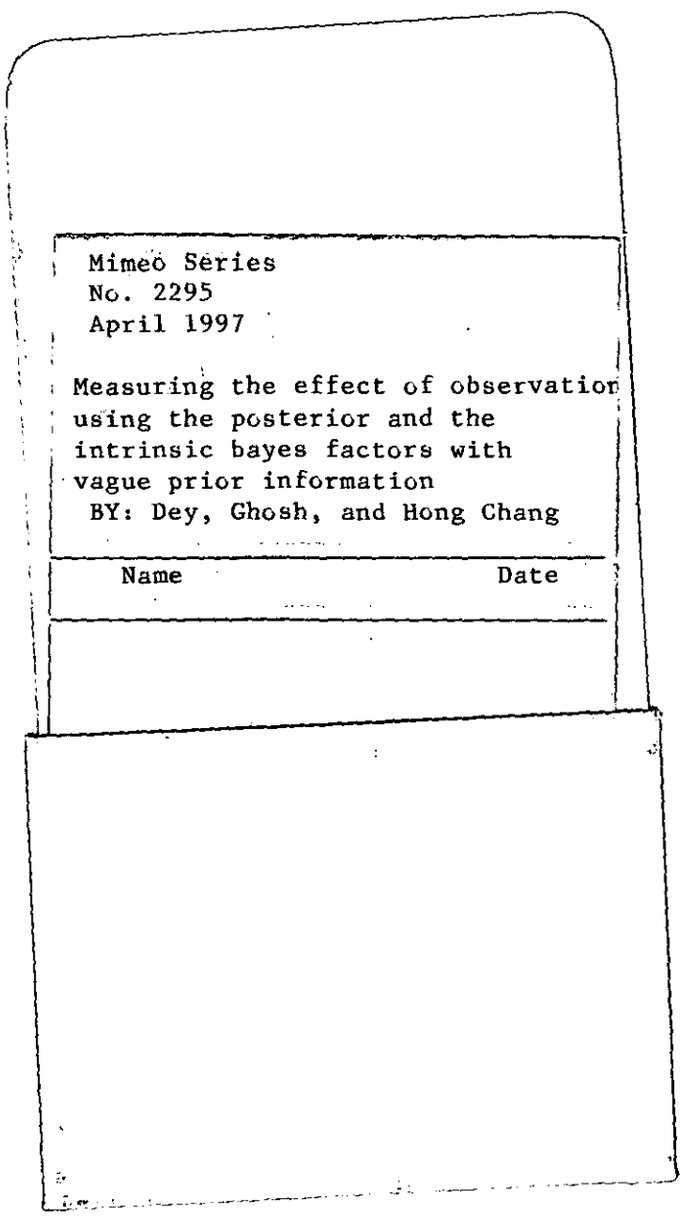
by

Dipak K. Dey, Sujit K. Ghosh, and Hong Chang

Institute of Statistics Mimeograph Series No. 2295

April 1997

NORTH CAROLINA STATE UNIVERSITY
Raleigh, North Carolina



Mimeo Series
No. 2295
April 1997

Measuring the effect of observation using the posterior and the intrinsic bayes factors with vague prior information
BY: Dey, Ghosh, and Hong Chang

Name Date

Measuring the Effect of Observations Using the Posterior and the Intrinsic Bayes Factors with Vague Prior Information

Dipak K. Dey*, Sujit K. Ghosh** and Hong Chang***

Abstract

Model determination is one of the fundamental problems in statistics. In this paper we consider model selection amongst a finite set of models along with model adequacy which are two integrated parts of model determination. We consider a measure of the effect of individual observations on the posterior and the intrinsic Bayes factors by studying the change in it after deleting an observation. The results are extremely useful in those applications where we consider a vague prior information. Several standard examples are provided where the measure can be expressed in closed form. Nonstandard examples which include nonlinear models and double exponential models are considered where sampling based methods are utilized to carry out required computations.

Key words: Bayes factor, double exponential model, influential observation, intrinsic Bayes factor, Jeffreys' prior, Laplace method, noninformative prior, posterior Bayes factor.

* *Dipak K. Dey is Professor of Statistics at the Department of Statistics, University of Connecticut, Storrs, CT 06269-3120.*

** *Sujit K. Ghosh is Assistant Professor at the Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203.*

*** *H. Chang is a consultant at Coopers & Lybrand L.L.P., Boston, MA 02110.*

1 Introduction

Model determination is a fundamental task in statistics, which divides into two components: model assessment or checking and model choice or selection. The literature on model assessment and model choice is considerable by now. It begins with the formal Bayes approach which, in the case of two models, results in the Bayes factor. Subsequent work has proposed several modifications of Bayes factors which were needed in the case of vague prior information. Geisser and Eddy (1979) took a predictive approach based on cross validation method to obtain pseudo Bayes factor. Aitkin (1991) defined the posterior Bayes factor and Berger and Pericchi (1996a) defined the intrinsic Bayes factor. Smith and Spiegelhalter (1980) and Spiegelhalter and Smith (1982) introduced the concept of an imaginary training sample and defined a local Bayes factor in this context.

The main objective of this paper is to develop diagnostic measures for model checking and also model selection using the posterior and the intrinsic Bayes factor when the prior information is vague and to demonstrate the calculation for nonlinear and double exponential models where the existing method is hard to implement. Johnson and Geisser (1982, 1983), Pettit and Smith (1983), Chaloner and Brant (1988) and Pettit (1992) have studied the problem of finding outliers and influential observations on posterior or predictive distributions. A current, reasonably thorough review appears in Gelfand, Dey and Chang (1992) and its attendant discussion.

We consider for simplicity a choice between two parametric models denoted interchangeably by joint density $f(\mathbf{y}|\boldsymbol{\theta}_i; M_i)$ or likelihood $L(\boldsymbol{\theta}_i; \mathbf{y}, M_i)$, i=1,2, where $\mathbf{Y}$ is $n \times 1$ and $\boldsymbol{\theta}_i$ is $p_i \times 1$. Following Neyman-Pearson theory, suppose we create the hypothesis H_i: data $\mathbf{y}$ arises from model M_i, i=1,2 with H_1 as null hypothesis. The formulation of a likelihood ratio test requires an unambiguous specification of a null and alternative hypothesis such as the nested models case where M_1 is the reduced model and M_2 is the full model i.e., $M_1 \subset M_2$ and $p_1 < p_2$. The likelihood ratio test then takes the form: reject H_1 if $\lambda_n < c < 1$ where

$$\lambda_n = \frac{L(\hat{\boldsymbol{\theta}}_1; \mathbf{y}, M_1)}{L(\hat{\boldsymbol{\theta}}_2; \mathbf{y}, M_2)}, \tag{1}$$

with $\hat{\boldsymbol{\theta}}_i$ is the maximum likelihood estimate under model M_i, $i = 1, 2$. The Bayesian model adds a prior specification $\pi(\boldsymbol{\theta} \mid M_i)$ to the likelihood specification for the model M_i, $i = 1, 2$. The formal Bayesian model choice procedure goes as follows. Suppose w_i is the prior probability of selecting the model M_i, i=1,2 and $f(\mathbf{y}|M_i)$ is

the predictive distribution for model M_i, i.e,

$$f(\mathbf{y}|M_i) = \int f(\mathbf{y}|\boldsymbol{\theta}_i, M_i)\pi_i(\boldsymbol{\theta}_i|M_i)d\boldsymbol{\theta}_i,$$

where $\pi_i(\boldsymbol{\theta}_i|M_i)$ is the prior under model M_i. If $\mathbf{y}_0$ denotes the observed data, then we choose the model yielding the larger $w_i f(\mathbf{y}_0|M_i)$. Often we set $w_i = .5$, i=1,2 and compute the Bayes factor of M_1 with respect to M_2 as

$$B_{12} = \frac{f(\mathbf{y}_0|M_1)}{f(\mathbf{y}_0|M_2)}. \tag{2}$$

In fact, we would not seek any alternative method if (??) could always be interpreted. Unfortunately, if $\pi(\boldsymbol{\theta} \mid M_i)$ is improper (as it usually will be under a noninformative specification) then $f(\mathbf{y} \mid M_i)$ is also. Therefore, we can not interpret $f(\mathbf{y}|M_i)$ as the probabilities of these models nor can we interpret the ratio. Even under proper priors the Bayes factor tends to attach too little weight to the simpler model even with arbitrarily large sample sizes. An illustration is the well known Lindley paradox dating at least to Barlett (1957). In view of this, Box (1980) and others have encouraged a less formal view with regard to Bayesian model choice.

To avoid the Lindley paradox (Lindley, 1957), Aitkin (1991) proposed the PBF, posterior Bayes factor which is defined as

$$A_{12} = \frac{\bar{L}_1^A}{\bar{L}_2^A} \tag{3}$$

where

$$\begin{aligned} \bar{L}_i^A &= \int f(\mathbf{y}|\boldsymbol{\theta}_i, \mathbf{M_i})\pi_i(\boldsymbol{\theta}_i|\mathbf{y})d\boldsymbol{\theta}_i \\ &= \frac{\int f^2(\mathbf{y}|\boldsymbol{\theta}_i, M_i)\pi_i(\boldsymbol{\theta}_i \mid M_i)d\boldsymbol{\theta}_i}{\int f(\mathbf{y}|\boldsymbol{\theta}_i, M_i)\pi_i(\boldsymbol{\theta}_i \mid M_i)d\boldsymbol{\theta}_i}, \qquad \text{i=1,2.} \end{aligned} \tag{4}$$

Aitkin (1991) also showed that the posterior Bayes factor exists under noninformative prior specification and he suggested the use of Jeffreys (1961) interpretive ranges for the Bayes factor.

Berger and Pericchi (1996a) suggested to use the Intrinsic Bayes Factor (IBF) for comparing models. The idea behind the IBF is to use part of the data as a "training sample" to obtain proper posterior densities for the parameters under each model and then to use these proper posteriors and the remaining data to compute the Bayes factor.

Suppose the entire sample $\mathbf{Y}$ can be divided into two parts: the training sample $\mathbf{Y}(l)$, and the remaining observations used for discrimination $\mathbf{Y}(-l)$. Then the posterior distribution based on the training sample is:

$$\pi(\boldsymbol{\theta}|\mathbf{Y}(l)) = \frac{f(\mathbf{Y}(l)|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int f(\mathbf{Y}(l)|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}.$$

Then the Bayes factor is computed with the remainder of the data $\mathbf{Y}(-l)$, using $\pi(\boldsymbol{\theta}|\mathbf{Y}(l))$ as prior. That is,

$$\begin{aligned} B_{12}(l) &= \frac{\int f(\mathbf{y}(-l)|\boldsymbol{\theta}, \mathbf{y}(l), M_1)\pi(\boldsymbol{\theta}|\mathbf{Y}(l), M_1)d\boldsymbol{\theta}}{\int f(\mathbf{y}(-l)|\boldsymbol{\theta}, \mathbf{y}(l), M_2)\pi(\boldsymbol{\theta}|\mathbf{Y}(l), M_2)d\boldsymbol{\theta}} \\ &= B_{12} \cdot B_{21}(\mathbf{y}(l)). \end{aligned}$$

Clearly, the multiplication of B_{12} and $B_{21}(\mathbf{y}(l))$ leads to the cancelation of the ratio of unspecified constants in the noninformative prior $\pi(\boldsymbol{\theta})$. To obtain the IBF, one takes the arithmetic or geometric average of the Bayes factors computed over a set of different (minimum) training samples of size l. We denote the arithmetic IBF as B_{12}^{AI} and geometric IBF as B_{12}^{GI}. Some ramifications of the intrinsic Bayes factors are given by Berger and Pericchi (1996a). In section 5 we restrict our calculations to geometric IBF, because it produces closed form expression for the normal linear models.

In the context of comparison of two models M_1 and M_2, Pettit and Young (1990) proposed a quantity

$$k_d = \log_{10} B_{12} - \log_{10} B_{12}^d$$

to measure the effect on the Bayes factor of observation d, where B_{12}^d is the Bayes factor excluding the observation d. They pointed out that large values of $|k_d|$ indicate large influence of observation d on the Bayes factor. Here we develop similar measures for the posterior Bayes factors, i.e, we define the quantity

$$c_d = \log_{10} A_{12} - \log_{10} A_{12}^d$$

to measure the effect on the posterior Bayes factor of observation d. Even if the use of the posterior Bayes factor has some controversy, we still recommend the use of c_d because of its simplicity in computation under vague prior information and asymptotic equivalence with k_d . In the same spirit we define the diagnostic measure

$$b_d = \log_{10} B_{12}^{GI} - \log_{10} B_{12}^{dGI}$$

to study the effect on the intrinsic Bayes factor after deleting observation d. Applying the same scale of evidence for assessing the Bayes factor (Jeffreys, 1961), we use the same benchmark for c_d and b_d as that used for k_d (suggested by Pettit and Young, 1990). Namely, an observation with $|c_d| > 0.5$ might be thought of as influential. Also if $c_d < 0$, or $b_d < 0$, the posterior Bayes factor or the intrinsic Bayes factor becomes larger after deleting the observation d, i.e, there is an increase of evidence for M_1. Consequently observation d is in favor of model M_2. Similarly if $c_d > 0$ or $b_d > 0$, observation d is in favor of model M_1.

We can also consider a small set deletion (say two or three components) which will be enable us to study the multiple features of the data, for instance, sets of similar points that are different from the rest of the data which cannot be identified by univariate analysis because of masking (Peña and Tiao, 1992). Gelfand and Dey (1994) also give a complete description of the properties of such multiple components deletion. Young (1992) also extended the idea of k_d to groups of observations. Nevertheless, the single component deletion is attractive in permitting us to work with univariate distribution.

The format of the paper is thus as follows. In section 2, we develop asymptotic results concerning equivalence of c_d with k_d. In section 3, we develop the results concerning normal distributions including normal linear models. In section 4, we study a nonlinear model where the existing method of outliers detection using the Bayes factor is difficult to implement and explain the computational approach using a sampling based method. In section 5, we consider the double exponential family as discussed in Efron (1986) and develop the computations of c_d for a double Poisson family. Here also the existing methods are difficult to implement. Finally, in section 6, we extend some of our results to Intrinsic Bayes factors.

2 Asymptotic Results

We now develop the asymptotic approximation of B_{12} and A_{12} using Laplace method as described in Tierney and Kadane (1986). The basic Laplace approximation is given by the following result.

Theorem 2.1 *(Tierney and Kadane) Suppose h is a strictly concave function of $\boldsymbol{\theta}$ having unique mode at $\hat{\boldsymbol{\theta}}$. Then*

$$\int e^{nh(\boldsymbol{\theta})} d\boldsymbol{\theta} = e^{nh(\hat{\boldsymbol{\theta}})} (2\pi)^{p/2} n^{-p/2} |-H^{-1}(\hat{\boldsymbol{\theta}})|^{1/2} + O(n^{-1}) \tag{5}$$

where $\boldsymbol{\theta}$ *is* $p \times 1$ *and* $H(\boldsymbol{\theta})$ *is a* $p \times p$ *positive definite matrix with the* $(j,k)th$ *element given by* $(H(\boldsymbol{\theta}))_{jk} = \dfrac{\partial^2 h(\boldsymbol{\theta})}{\partial\theta_j\theta_k}$.

The next theorem gives the asymptotic equivalence of k_d and c_d.

Theorem 2.2 *Under regularity conditions, as stated in Theorem 2.1,* c_d *and* k_d *are asymptotically equivalent in the sense that*

$$c_d = k_d + o(1).$$

Proof : First observe that for $g(\boldsymbol{\theta}) > 0$, using (??), it follows that

$$\frac{\int g(\boldsymbol{\theta})e^{nh(\boldsymbol{\theta})}d\boldsymbol{\theta}}{\int e^{nh(\boldsymbol{\theta})}d\boldsymbol{\theta}} = e^{n(h^*(\boldsymbol{\theta}^*)-h(\hat{\boldsymbol{\theta}}))}\left[\frac{|-H^{*-1}(\boldsymbol{\theta}^*)|}{|-H^{-1}(\hat{\boldsymbol{\theta}})|}\right]^{1/2} + O(n^{-2}), \tag{6}$$

where $nh^*(\boldsymbol{\theta}) = nh(\boldsymbol{\theta}) + \log g(\boldsymbol{\theta})$, with h^* having unique mode $\hat{\boldsymbol{\theta}}^*$ and $H^*(\boldsymbol{\theta})$ is a $p \times p$ positive definite matrix with the $(j,k)th$ element $(H^*(\boldsymbol{\theta}))_{jk} = \dfrac{\partial^2 h^*(\boldsymbol{\theta})}{\partial\theta_j\theta_k}$. Now under usual regularity conditions on the likelihood, suppose $\hat{\boldsymbol{\theta}}_{i,n}$ is the maximum likelihood estimate of $\boldsymbol{\theta}_i$ based on a sample size of n, then $\hat{\boldsymbol{\theta}}_{i,n} \xrightarrow{P} \boldsymbol{\theta}_{i,0}$ for some $\boldsymbol{\theta}_{i,0}$ and

$$n^{-1}\left(-\frac{\partial^2 \log L(\boldsymbol{\theta}_i, \mathbf{y}, M_i)}{\partial\theta_j\partial\theta_k}\right) \xrightarrow{P} (I(\boldsymbol{\theta}_i))_{jk}, \qquad i = 1,2,$$

where $I(\boldsymbol{\theta})$ denotes Fisher's information matrix. Now using Gelfand and Dey(1994), it follows that $-H(\boldsymbol{\theta}_i) \xrightarrow{P} I(\boldsymbol{\theta}_i)$ and $-H^*(\hat{\boldsymbol{\theta}}_{i,n}) \xrightarrow{P} I(\boldsymbol{\theta}_{i,0})$. Similarly, $-H^*(\boldsymbol{\theta}_i) \xrightarrow{P} I(\boldsymbol{\theta}_i)$ and $-H(\hat{\boldsymbol{\theta}}_{i,n}) \xrightarrow{P} I(\boldsymbol{\theta}_{i,0})$. These asymptotics reveal the obvious fact that the specification of priors π_i, as long as it is free of n, is asymptotically irrelevant. Thus we obtain

$$B_{12} = \frac{f(\mathbf{y}|\hat{\boldsymbol{\theta}}_{1,n}, M_1)\pi_1(\hat{\boldsymbol{\theta}}_{1,n})|-H_1^{-1}(\hat{\boldsymbol{\theta}}_{1,n})|^{1/2}}{f(\mathbf{y}|\hat{\boldsymbol{\theta}}_{2,n}, M_2)\pi_2(\hat{\boldsymbol{\theta}}_{2,n})|-H_2^{-1}(\hat{\boldsymbol{\theta}}_{2,n})|^{1/2}}\left(\frac{n}{2\pi}\right)^{(p_2-p_1)/2} + O_p(1)$$

and

$$\log_{10} B_{12} = \log_{10}\lambda_n + \frac{p_2 - p_1}{2}\log_{10} n + O_p(1), \tag{7}$$

where λ_n is the likelihood ratio statistic and (??) is the Schwarz's (1978) BIC adjustment of the likelihood ratio test.

Using (??), we can now easily derive the asymptotic approximation of k_d as defined in Pettit and Young (1990). After removing observation d, it follows that

$$\log_{10} B_{12}^d = \log_{10} \lambda_{n-1}^d + \frac{p_2 - p_1}{2} \log_{10}(n-1) + O_p(1). \tag{8}$$

From (??) and (??), it follows that

$$k_d = \log_{10}\left(\frac{\lambda_n}{\lambda_{n-1}^d}\right) + \frac{p_2 - p_1}{2} \log_{10}\left(\frac{n}{n-1}\right). \tag{9}$$

Now to develop the asymptotic approximation of the posterior Bayes factor and hence c_d, we again use Theorem 2.1 and obtain first

$$\bar{L}_i^A \approx \frac{f^2(\mathbf{y}|\hat{\boldsymbol{\theta}}_i^*)\pi_1(\hat{\boldsymbol{\theta}}_i^*)|-H_1^{*-1}(\hat{\boldsymbol{\theta}}_i^*)|^{1/2}}{f(\mathbf{y}|\hat{\boldsymbol{\theta}}_i)\pi_1(\hat{\boldsymbol{\theta}}_i)|-H_1^{-1}(\hat{\boldsymbol{\theta}}_i)|^{1/2}} \quad , i = 1, 2.$$

where $\hat{\boldsymbol{\theta}}_i^*$ and $\hat{\boldsymbol{\theta}}_i$ are the maximizer of integrals in numerator and denominator of (??) respectively. Now using regularity conditions and the fact that $-H(\hat{\boldsymbol{\theta}}_i) \xrightarrow{P} I(\boldsymbol{\theta}_i)$ and $-H^*(\hat{\boldsymbol{\theta}}_i) \xrightarrow{P} 2I(\boldsymbol{\theta}_i)$, it follows that

$$A_{12} \approx \lambda_n 2^{(p_2-p_1)/2} \tag{10}$$

and hence

$$\log_{10} A_{12} \approx \log_{10} \lambda_n + \frac{p_2 - p_1}{2} \log_{10} 2. \tag{11}$$

Result (??) along with some additional calculation are provided in Aitkin(1991). Using (??), it follows that the effect on the posterior Bayes factor after removing observation d is

$$c_d = \log_{10} A_{12} - \log_{10} A_{12}^d \approx \log_{10}\left(\frac{\lambda_n}{\lambda_{n-1}^d}\right). \tag{12}$$

Combining equations (??) and (??) it follows that k_d and c_d are asymptotically equivalent. ■

Asymptotic equivalence of k_d with b_d follows from Theorem 2.1 and Berger and Perrichi (1996a). In practical examples, we will see even for moderate sample sizes, k_d and c_d are sufficiently close.

2.1 PBF for Normal Linear Models

In this section, we develop formulas for posterior Bayes factor and c_d for the normal linear models. As a special case we also consider simple linear regression model

and compare our c_d values with k_d values of Pettit and Young through a numerical example. Consider two models M_1 and M_2, where $M_1 \subset M_2$, i.e., M_1 is nested within M_2, as defined by

$$Y|M_i \sim N_n(A_i\theta_i, \sigma^2\mathbf{I}), \tag{13}$$

where A_i is of full rank p_i, $i = 1, 2$. If the prior is taken to be $\pi(\theta_i, \sigma^2) = (1/\sigma)^r$, where r is the hyperparameter, it can be shown that the posterior Bayes factor for M_1 to M_2 is

$$A_{12} = \frac{\Gamma[\frac{2n+r-p_1-1}{2}]\Gamma[\frac{n+r-p_2-2}{2}]}{\Gamma[\frac{n+r-p_1-2}{2}]\Gamma[\frac{2n+r-p_2-1}{2}]}\left(\frac{RSS_1}{RSS_2}\right)^{-n/2} \tag{14}$$

where RSS_i is the residual sum of squares for model M_i, i=1,2. Similarly, A_{12}^d can obtained from (14) by replacing n with (n-1) and RSS_i by $RSS_{i(d)}$, where $RSS_{i(d)}$ is the residual sum of squares for the model M_i excluding the observation d. Thus c_d is obtained as $log_{10}[A_{12}/A_{12}^d]$. Notice that this quantity is invariant to scaling. It is observed that the choice of r has very little effect on c_d for moderate to large sample sizes.

2.1.1 Simple Linear Regression Model

A special case of a linear model is the simple linear regression model which is given as

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2).$$

Suppose we want to test $H_0(M_1)$: $\beta_1 = 0$ against $H_1(M_2)$: $\beta_1 \neq 0$. The improper prior distributions are given as

$$\pi(\beta_0, \sigma^2) = \frac{1}{\sigma^2}, \quad \text{under } M_1$$

and

$$\pi(\beta_0, \beta_1, \sigma^2) = \frac{1}{\sigma^2}, \quad \text{under } M_2.$$

If we denote the correlation coefficient between y and x as $\hat{\rho}$, then it can be shown that the posterior Bayes factor

$$A_{12} = \frac{\Gamma[n]\Gamma[\frac{n-2}{2}]}{\Gamma[\frac{n-1}{2}]\Gamma[\frac{2n-1}{2}]}(1-\hat{\rho}^2)^{n/2}.$$

and

$$A_{12}^d = \frac{\Gamma[n-1]\Gamma[\frac{n-3}{2}]}{\Gamma[\frac{n-2}{2}]\Gamma[\frac{2n-3}{2}]},$$

where $\hat{\rho}_{(d)}$ is the correlation coefficient between y and x after the dth observation has been deleted. Hence c_d can be obtained easily from above expressions. It is obvious that c_d will be large only if $\hat{\rho}$ and $\hat{\rho}_{(d)}$ are very different, that is when the dth observation is away from the main group of the points.

We use the example from Pettit and Young (1990) to make the comparison of k_d and c_d. The data is given in Table 1 and the values of k_d and c_d are given in Table 2.

Comparison of k_d and c_d shows that the magnitudes of c_d and k_d are almost the same, and for each observation, c_d and k_d shows the same preference to either model. A plot of c_d and k_d against observation number is provided in Figure 1. It is clear from figure 1 that the use of c_d and k_d will produce the same conclusion for the model diagnostics. Namely, both c_d and k_d indicate that the observation 5 and 19 are influential points and the data do indeed support M_2 because of the more negative c_d or k_d values.

3 PBF for Nonlinear Models

In this section we consider a nonlinear model and an associated reduced model. Analytical expression for c_d in this case can not be represented in closed form. But the posterior Bayes factor could be calculated using sampling based approach and hence the c_d.

The data concerns the steady state absorption of o-xylene as a function of oxygen concentration (x_1) inlet o-xylene concentration (x_2) and temperature (x_3). The sample of 57 points is presented along with the full model in Bates and Watts (1988, p. 306-309). The full model is in fact

$$y_j = \frac{f_1 f_2}{f_1 + 2.22788 f_2} + \varepsilon_j; \tag{15}$$

where ε_j is assumed independent $N(0, \sigma^2)$ and f_1 and f_2 are given as

$$f_1 = \theta_1 x_{1j} e^{\theta_3 / x_{3j}}$$

$$f_2 = \theta_2 x_{2j} e^{\theta_4 / x_{3j}}.$$

A convenient reduced model of (??) is obtained by taking $\theta_3 = \theta_4$ which is given as:

$$y_j = \frac{\theta_1 \theta_2 x_{1j} x_{2j} e^{-\theta_3 / x_{3j}}}{\theta_1 x_{1j} + 2.22788 \theta_2 x_{2j}} + \varepsilon_j; \tag{16}$$

where ε_j is again assumed to be independent $N(0, \sigma^2)$.

We call this full model by M_2 and the reduced model by M_1, so that $M_1 \subset M_2$. In both cases, we reparametrize $\boldsymbol{\theta}$ to R^p, where $p = 4$ in M_2 and $p = 3$ in M_1, and then consider the noninformative prior $\pi(\boldsymbol{\theta}, \sigma^2) = (\sigma^2)^{-1}$. The analytic evaluation of the required integrals is not available for the form of mean function in both M_1 and M_2. So the Monte Carlo techniques using sampling-based methods are needed to accomplish the required integrations, hence the posterior Bayes Factor. For the detailed implementation of sampling-based methods, refer to Gelfand, Dey and Chang (1992) and Gelfand and Dey (1994). The simulation result shows that $\log_{10} A_{12} = -2.2866$ and $\log_{10} \lambda_{12} = -2.3979$ where λ_{12} represents the likelihood ratio statistic. Both measures indicate that the full model is preferable. The values of c_d are given in table 3.

The identification of influential observations based on c_d indicates that the observation 11, 43 and 44 are influential points. Observations 43 and 44 both carry the positive sign in their c_d statistic. Namely, observations 43 and 44 are in favor of model 2 in this sense. After removing them from the data set, we observe the increase in the posterior Bayes factors. A plot of c_d against d is provided in the figure 2. This c_d plot suggests which observation is in support of which model. This plot can also be used to detect influential observations. If a c_d value is large (say greater than .5), then that observations is influential in distinguishing between the models.

4 PBF for Double Exponential Family Model

In this section we investigate the double exponential family model as introduced by Efron (1986). These models are useful to incorporate over and under dispersion in building a statistical model. In this scenario a one parameter exponential family model is nested within a double exponential family model. We obtain the posterior Bayes factor and required c_d calculations in connection with model selection and model checking. In particular, we consider a double Poisson family and apply our method to a data set to determine the appropriate model and influential observations.

4.1 Double Poisson Family

Suppose we have the sample $y_1, y_2, \ldots, y_n \sim f_{\mu,\theta}(y)$ where

$$f_{\mu,\theta}(y) = (\theta^{1/2} e^{-\theta\mu}) \left(\frac{e^{-y} y^y}{y!}\right) \left(\frac{e\mu}{y}\right)^{\theta y}, \qquad y = 0, 1, 2, \ldots. \tag{17}$$

This is the so called double Poisson family as introduced in Efron (1986). We take M_1 as Poisson(μ) and M_2 as double Poisson model. Notice that M_1 is a sub-model of M_2 by taking $\theta = 1$, where θ can be considered as an overdispersion or underdispersion parameter according as $\theta > 1$ or $\theta < 1$. Now the prior distributions for model parameters are taken as the standard Jeffreys's prior, which after some calculation reduces to

$$\pi(\mu) \propto (\mu)^{-1/2}, \qquad \text{under } M_1$$

and

$$\pi(\mu, \theta) \propto (\mu\theta)^{-1/2}, \qquad \text{under } M_2.$$

The posterior mean of likelihood function of the data $\mathbf{y}$ under M_1 can be easily calculated as

$$\bar{L}_1^A = \frac{1}{t_2}(1/2)^{2t_1+1/2}(1/n)^{t_1}\frac{\Gamma(2t_1+1/2)}{\Gamma(t_1+1/2)} \tag{18}$$

where $t_1 = \sum_{i=1}^{n} y_i$ and $t_2 = \prod_{i=1}^{n} y_i!$.

A closed form expression of posterior mean of likelihood function of the data $\mathbf{y}$ under M_2 can not be obtained. In order to get the posterior mean of the likelihood function of M_2 which is needed in calculation of the posterior Bayes Factor, we use the following approach. It is straightforward to verify that $(t_1, t_3 = \sum_{i=1}^{n} y_i \log(y_i))$ is jointly sufficient for (μ, θ) and the log-likelihood function is given by

$$l(\mu, \theta; \mathbf{y}) = \frac{n}{2}\log(\theta) - n\theta\mu + \theta(t_1 - t_3) + t_1\theta\log(\mu) + (t_3 - t_1 - \log(t_2)).$$

By direct integration it follows that,

$$\begin{aligned} &\int_0^\infty \int_0^\infty f(\mathbf{y} \mid \mu, \theta)\pi(\mu, \theta \mid M_2)\, d\theta d\mu \\ &= \Gamma[\frac{n+1}{2}]\exp\{t_3 - t_1 - \log(t_2)\}\int_0^\infty \frac{1}{\sqrt{\mu}}(n\mu + t_1 - t_3 + t_1\log(\mu))^{-\frac{n+1}{2}}\, d\mu \end{aligned} \tag{19}$$

and

$$\begin{aligned} &\int_0^\infty \int_0^\infty f^2(\mathbf{y} \mid \mu, \theta)\pi(\mu, \theta \mid M_2)\, d\theta d\mu \\ &= \Gamma[\frac{2n+1}{2}]\exp\{2t_3 - 2t_1 - 2\log(t_2)\}\int_0^\infty \frac{1}{\sqrt{\mu}}(2n\mu + 2t_1 - 2t_3 - 2t_1\log(\mu))^{-\frac{2n+1}{2}}\, d\mu \end{aligned} \tag{20}$$

By combining (19) and (20) it follows that,

$$\bar{L}_2^A = \frac{2^{-\frac{2n+1}{2}}\Gamma[\frac{2n+1}{2}]}{\Gamma[\frac{n+1}{2}]}\exp\{t_3 - t_1 - \log(t_2)\}E_\mu\left[(n\mu + t_1 - t_3 + t_1\log(\mu))^{-\frac{n}{2}}\right]$$

where E_μ denotes expectation with respect to the density $g(\mu)$ such that

$$g(\mu) \propto \mu^{-\frac{n}{2}}\left[1 + \frac{t_1 - t_3}{n} + \frac{t_1\log(\mu)}{n\mu}\right]^{-\frac{n+1}{2}}.$$

The above expectation can be evaluated numerically using any statistical subroutine such as IMSL etc. Alternatively, we can use Monte Carlo Integration using a rejection method or Metropolis method.

As an illustrative example, we consider the data in Table 5, originally collected by Thyrian (1961), and presented in Seal (1969) and have been analyzed by Lindsay (1986) and Gelfand and Dalal (1990). The data consists of observed counts of accidents in a year for 9461 Belgian drivers.

Table 1: The c_d values for the accident data (Seal, 1969).

y	observed counts	c_d
0	7840	-0.1026
1	1317	0.0297
2	239	0.2608
3	42	0.5233
4	14	1.0039
5	4	1.2057
6	4	1.5432
7	1	2.0173

The $\log_{10}(A_{12})$ is -50.2289 and the $\log_{10}(\lambda_n)$ is -50.2867. Both indicate that the full model (double Poisson) overwhelmingly dominates the reduced model (Poisson). The c_d statistic reveals that 7840 observations out of 9461 support the full model.

5 IBF and Effect of Observations on IBF

In this section, we develop formulae for the intrinsic Bayes factors and b_d for the normal linear models and nonlinear models.

5.1 Effect of Observations on IBF for Normal Linear Models

Consider the linear model given in (13) in section 2.1. Now the Berger-Bernardo reference prior has the form

$$\pi(\boldsymbol{\theta}_i, \sigma^2) = (1/\sigma)$$

Following Berger and Pericchi (1996b) it can be shown that $y(l)$ = minimum training sample with corresponding design matrix $X_i(l)$ of size $m = max\{p_i\} + 1$. Here we consider the geometric intrinsic Bayes factor (because it has a closed form expression) which is given as

$$B_{12}^{GI} = C_{n,m} \frac{|X_1^T X_1|^{1/2}}{|X_2^T X_2|^{1/2}} \frac{RSS_1^{(n-p_1)/2}}{RSS_2^{(n-p_2)/2}} \left\{ \prod_{l=1}^{L} \frac{|X_1(l)^T X_1(l)|^{1/2}}{|X_2(l)^T X_2(l)|^{1/2}} \frac{RSS_1(l)^{(m-p_1)/2}}{RSS_2(l)^{(m-p_2)/2}} \right\}^{1/L},$$

where $c_{n,m} = \frac{\Gamma(\frac{n-p_1}{2})\Gamma(\frac{m-p_2}{2})}{\Gamma(\frac{n-p_2}{2})\Gamma(\frac{m-p_1}{2})}$ and L is the number of minimal training samples. Deleting d^{th} observation leads to

$$\begin{aligned} B_{12(d)}^{GI} &= C_{n-1,m-1} \frac{|X_1^{(d)^T} X_1^{(d)}|^{1/2}}{|X_2^{(d)^T} X_2^{(d)}|^{1/2}} \frac{RSS_{1(d)}^{(n-p_1-1)/2}}{RSS_{2(d)}^{(n-p_2-1)/2}} \\ &\quad \left\{ \prod_{l=1}^{L} \frac{|X_1^{(d)^T}(l) X_1^{(d)}(l)|^{1/2}}{|X_2^{(d)}(l)^T X_2^{(d)}(l)|^{1/2}} \frac{RSS_1(l)^{(m-p_1-1)/2}}{RSS_2(l)^{(m-p_2-1)/2}} \right\}^{1/L}. \end{aligned}$$

Recall Andrews and Pregibon (1978) diagnostic as

$$AP_i^{(d)} = \frac{|X_i^{(d)^T} X_i^{(d)}|}{|X_i^T X_i|} \frac{RSS_{i(d)}}{RSS_i}, \text{ i=1,2,}$$

then, effect of the d^{th} observation on IBF is given by

$$\begin{aligned} b_d &= \log B_{12}^{GI} - \log B_{12}^{GI(d)} \\ &= \frac{1}{2}\left(\frac{AP_1^{(d)}}{AP_2^{(d)}}\right) + \frac{n-p_1-1}{2}\log RSS_1 - \frac{n-p_2-1}{2}\log RSS_2 \\ &\quad + \frac{n-p_2-2}{2}\log RSS_{2(d)} - \frac{n-p_1-2}{2}\log RSS_{1(d)} \\ &\quad + \frac{1}{L}\sum_{l=1}^{L}\left[\frac{1}{2}\log\frac{AP_1^{(d)}(l)}{AP_1^{(d)}(l)} - \frac{m-p_1-1}{2}\log RSS_1(l)\right. \\ &\quad + \frac{m-p_2-1}{2}\log RSS_2(l) - \frac{m-p_2-2}{2}\log RSS_{2(d)}(l) \\ &\quad \left. + \frac{m-p_1-2}{2}\log RSS_{1(d)}\right]\log\frac{c_{n,m}}{c_{n-1,m-1}}. \end{aligned}$$

We give following example for the purpose of comparing previously defined k_d and c_d with b_d. We again use the rat data. The last column in Table 2 gives the b_d values. Again, a plot of all three influential statistics against observation number is provided in Figure 3.

5.2 Effect of Observation on the IBF for Nonlinear Models

In general, $B_{12}^{GI} = B_{12}^{N}\left\{\prod_{l=1}^{L} B_{12}^{N}(l)\right\}^{1/L}$, so deleting d^{th} observation resulting $B_{12}^{GI(d)} = B_{12}^{N(d)}\left\{\prod_{l=1}^{L} B_{12}^{N(d)}(l)\right\}^{1/L}$. Thus

$$\frac{B_{12}^{GI}}{B_{12}^{N(d)}} = \frac{m_1^N(y)}{m_2^N(y)}\frac{m_1^N(y(d))}{m_2^N(y(d))} = \frac{f_1(y_d|y_{(d)})}{f_2(y_d|y_{(d)})} = \frac{CPO^1}{CPO^2},$$

where $m_i^N(y) = \int f(y|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$ and $m_i^N(y(d)) = \int f(y(d)|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$. The quantity CPO^i denotes the Conditional Predictive Ordinate (Geisser and Eddy, 1979) values for model M_i, $i = 1, 2$. Therefore,

$$\frac{B_{12}^{GI}}{B_{12}^{N(d)}} = \frac{CPO^1}{CPO^2}\left\{\prod_{l=1}^{L}\frac{CPO^1(l)}{CPO^2(l)}\right\}^{1/L}$$

so

$$b_d = \log\frac{CPO^1}{CPO^2} + \frac{1}{L}\sum_{l=1}^{L}\log\frac{CPO^1(l)}{CPO^2(l)}.$$

The calculation of CPO^i, and $CPO^i(l)$ can be obtained by Monte Carlo estimates. See Gelfand, Dey and Chang (1992). Calculation of b_d for the nonlinear models can be performed easily using sampling based approach. This however is not pursued in this article.

Acknowledgments

The authors wish to thank a referee for many helpful suggestions. The authors also wish to thank Prof. Alan E. Gelfand for valuable discussions.

References

[1] Aitkin, M. (1991). *Posterior Bayes factors* J. Roy. Statist. Soc. B. 111-142.

[2] Bartlett, M. (1957). *A comment on D. V. Lindley's statistical paradox.* Biometrika **44**, 533-534.

[3] Bates, D.M. and Watts, D.G. (1988). *Nonlinear Regression Analysis and Its Applications.* John Wiley & Sons, Inc.

[4] Berger, J. O. and Pericchi, L. R.(1996a). *The Intrinsic Bayes factor for Model Selection and Prediction.* J. Amer. Statist. Assoc. **91**, 109-122.

[5] Berger, J. O. and Pericchi, L. R.(1996b). *The intrinsic Bayes factor for Linear Models* Bayesian Statistics **5**, (J.M. Bernardo et al. eds.), Oxford University Press, London 23-42.

[6] Bernardo, J. (1974). *Expected Information as Expected Utility.* Ann. Statist.**7**, 686-690.

[7] Box, G. (1980). *Sampling and Bayes' inference in scientific modeling and robustness.* J. Roy Statist. Soc. A **143** 382-430, (with discussion).

[8] Chaloner, K and Brant, R. (1988). *A Bayesian approach to outlier detection and residual analysis.* Biometrika **75**, 651-659.

[9] Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression.* Chapman and Hall, London.

[10] Efron, B. (1986). *Double exponential Families and Their Use in Generalized Linear Regression.* J. Amer. Statist. Assoc., **81**, 709-720.

[11] Gelfand, A.E., Dalal, S. R. (1990). *A note on overdispersed exponential families.* Biometrika **77**, 55-64.

[12] Gelfand, A.E., Dey, D.K. and Chang, H. (1992). *Model Determination Using Predictive Distributions With Implementation Via Sampling-Based Methods.* Bayesian Statistics **4**, (J. M. Bernardo, et.al. eds), Oxford University Press. Oxford, pp 147-167, (with discussion).

[13] Gelfand, A.E. and Dey, D.K. (1994). *Bayesian Model Choice: Asymptotics and Exact Calculations.* J. Roy. Statist. Soc, B, **56,** 501-514.

[14] Geisser, S. and Eddy, W. (1979). *A predictive approach to model selection.* J. Amer. Statist. Assoc. **74**, 153-160.

[15] Jeffreys, H. (1961). *Theory of Probability.* University Press.

[16] Johnson, W. and Geisser, S. (1982). *Assessing the predictive influence of observations.* Statistics and Probability in Honor of C.R. Rao. (Kallianpur, Krishnaiah and Ghosh, eds.) Amsterdam: North-Holland 343-358.

[17] Johnson, W. and Geisser, S. (1983). *A predictive view of the detection and characterization of influential observations in regression analysis.* J. Amer. Statist. Assoc. **78**, 144-167.

[18] Lindley, D. (1957). *A statistical paradox.* Biometrika, **44**, 187-192

[19] Lindsay, B. (1986). *Exponential family mixture models (with least squares estimators).* Ann. Statist. **14**, 124-137.

[20] Pettit, L.I. and Smith, A. F. M. (1985). *Outliers and influential observations in linear model.* Bayesian Statistics **2**, (J. M. Bernardo, et.al. eds), Amsterdam: North-Holland 473-494, (with discussion).

[21] Pettit, L.I. and Young K.D.S. (1990). *Measuring the effect of observations on Bayes factors.* Biometrika, **77**, 455-466.

[22] Pettit, L.I. (1992). *Bayes Factors for Outlier Models Using the Device of Imaginary Observations.* J. Amer. Statist. Assoc., **87**, 541-545.

[23] Seal, H. L. (1969). *Stochastic theory of a risk business.* New York: Wiley.

[24] Smith, A.F.M.and Spiegelhalter, D.J. (1980). *Bayes Factor and Choice Criteria for Linear Models.* J.Roy. Statist. Soc. B, **42**, 213-220.

[25] Spiegelhalter, D.J. and Smith, A.F.M. (1982). *Bayes factors for linear and log-linear models with vague prior information.* J.Roy. Statist. Soc. B, **44**, 377-387.

[26] Thyrion, P. (1961). *Contribution à l'étude des bonus pour non sinistre en assurance automobile.* Astin Bull **1**, 142-162.

[27] Tierney, L. and Kadane, J. B. (1986). *Accurate Approximations for Posterior Moments and Marginal Densities.* J. Amer. Statist. Assoc., **81**, 82-86.

Figure 1: Plot of c_d vs. observation number for o-xylene Data

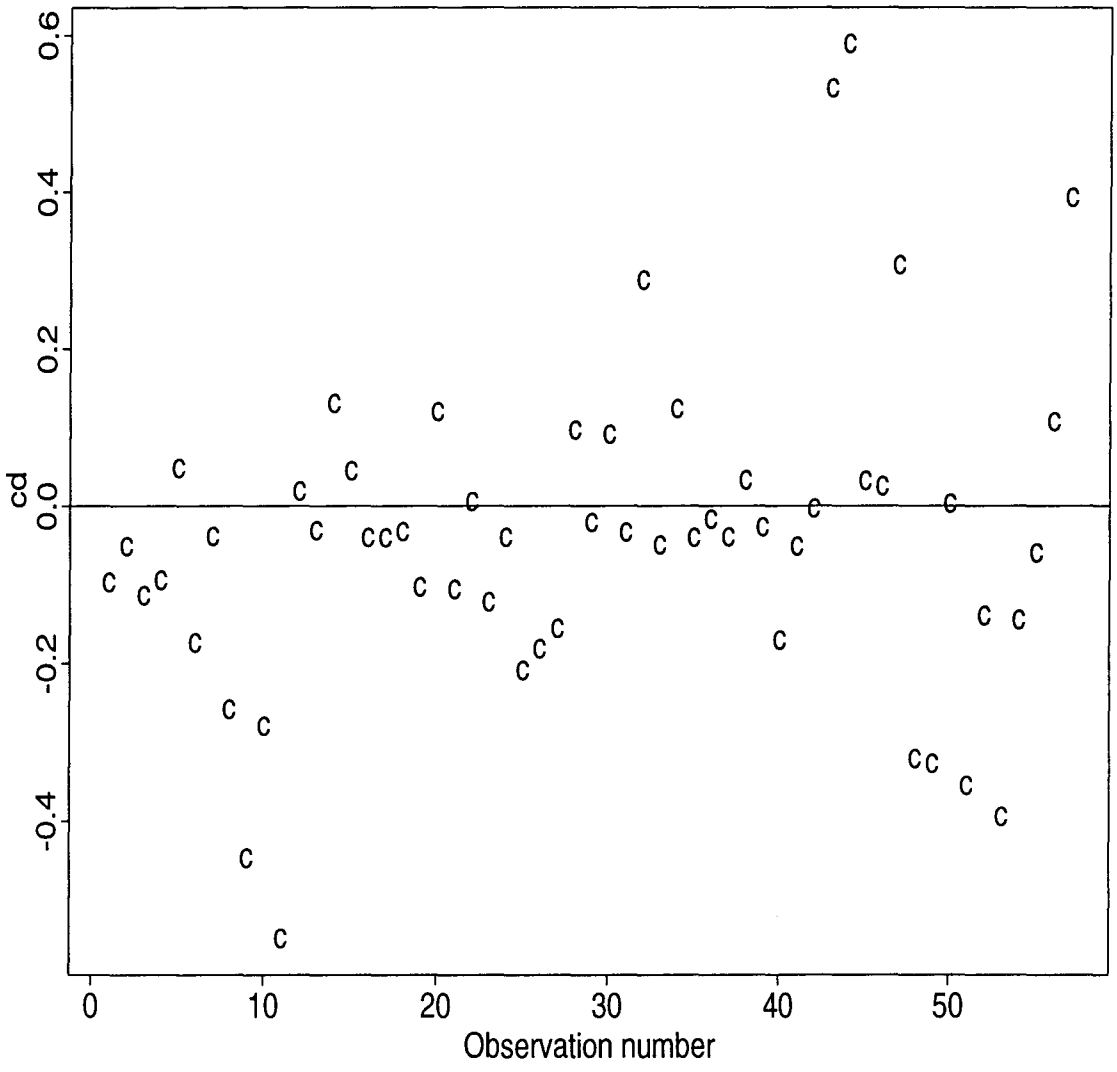


Figure 2: Plot of c_d, k_d and b_d vs observation number for Rat Data

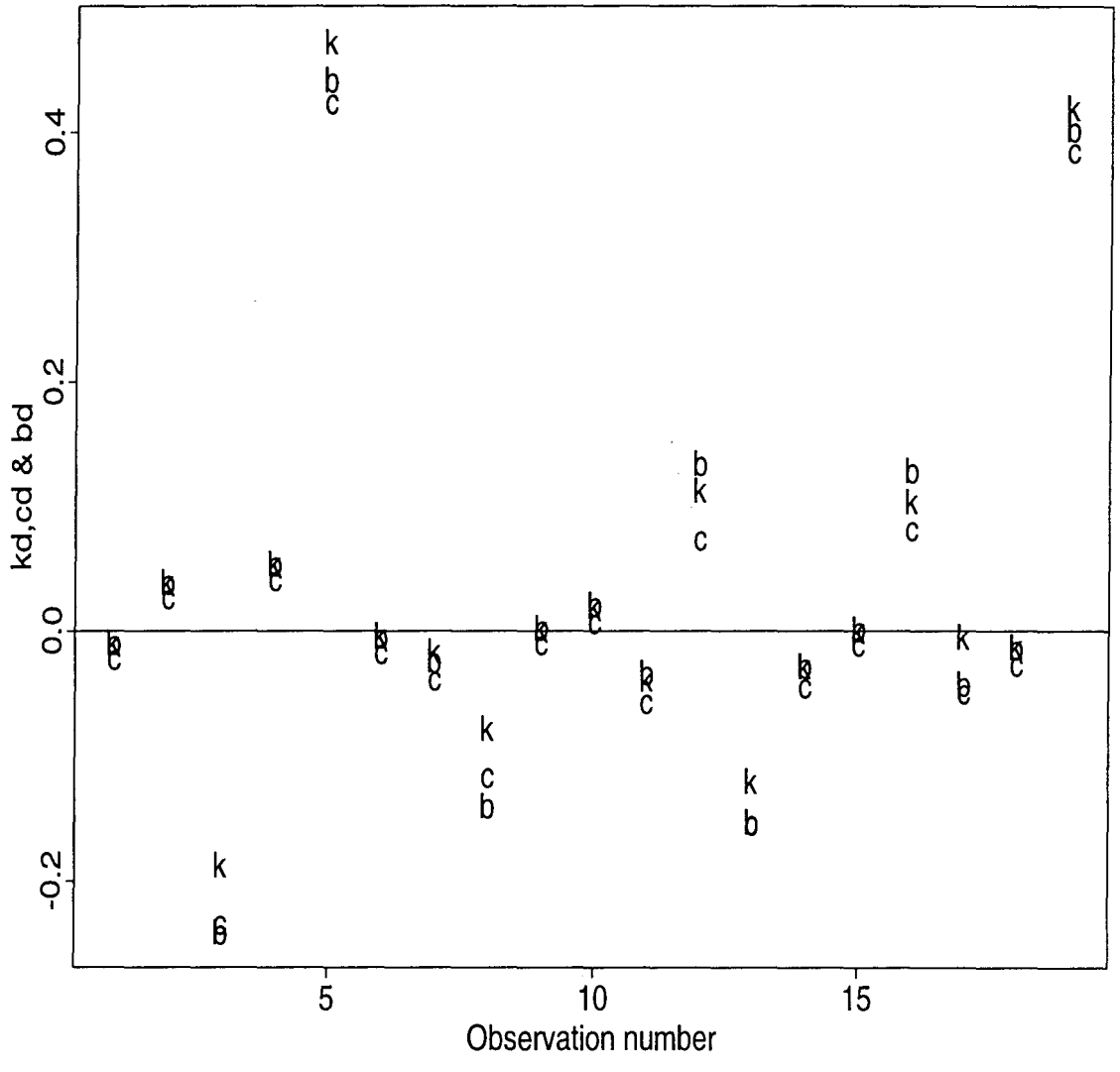


Table 2: Rat Data: k_d, c_d and b_d values

obsn.#	y	x	kd	cd	bd
1	0.42	0.110	-0.0092	-0.0235	-0.0092
2	0.25	0.110	0.0393	0.0251	0.0393
3	0.56	1.000	-0.1865	-0.2360	-0.2413
4	0.23	0.110	0.0539	0.0397	0.0539
5	0.23	1.000	0.4721	0.4226	0.4417
6	0.32	-0.260	-0.0035	-0.0190	-0.0042
7	0.37	0.555	-0.0158	-0.0403	-0.0233
8	0.42	0.850	-0.0779	-0.1170	-0.1397
9	0.33	0.110	0.0027	-0.0115	0.0027
10	0.38	-0.185	0.0203	0.0057	0.0218
11	0.27	-0.480	-0.0386	-0.0590	-0.0334
12	0.36	-0.925	0.1131	0.0720	0.1346
13	0.21	-0.850	-0.1194	-0.1559	-0.1528
14	0.28	-0.410	-0.0280	-0.0465	-0.0279
15	0.34	-0.110	0.0013	-0.0126	0.0014
16	0.28	0.555	0.1040	0.0796	0.1292
17	0.30	-1.000	-0.0043	-0.0505	-0.0423
18	0.37	0.260	-0.0122	-0.0284	-0.0133
19	0.46	-0.850	0.4204	0.3839	0.4034

Table 3: The values of c_d for the o-xylene absorption model

Obs	c_d	Obs	c_d	Obs	c_d
1	-0.0978601	21	-0.1060611	41	-0.0513641
2	-0.0513441	22	0.0053279	42	-0.0037391
3	-0.1144521	23	-0.1217571	43	0.5335919
4	-0.0945431	24	-0.0399331	44	0.5891729
5	0.0482409	25	-0.2082121	45	0.0318799
6	-0.1734141	26	-0.1812531	46	0.0248609
7	-0.0392001	27	-0.1556261	47	0.3072919
8	-0.2581321	28	0.0966279	48	-0.3217671
9	-0.4474931	29	-0.0209861	49	-0.3277311
10	-0.2798411	30	0.0910789	50	0.0027809
11	-0.5493301	31	-0.0334471	51	-0.3548081
12	0.0188069	32	0.2875969	52	-0.1393461
13	-0.0317931	33	-0.0503461	53	-0.3949711
14	0.1306049	34	0.1235489	54	-0.1442651
15	0.0440679	35	-0.0405351	55	-0.0604241
16	-0.0406861	36	-0.0173911	56	0.1074579
17	-0.0406661	37	-0.0400381	57	0.3928759
18	-0.0330041	38	0.0322469		
19	-0.1034341	39	-0.0277681		
20	0.1207539	40	-0.1702411		