Revista de la Facultad de Ciencias Universidad Nacional de Colombia, Sede Medellín
V 1 N°1 julio 2012 • ISSN 0121-747X • Paper Invited
Páginas 7 a 22

# NEW CRITERIA FOR THE CHOICE OF TRAINING SAMPLE SIZE FOR MODEL SELECTION AND PREDICTION: THE CUBIC ROOT RULE

# UN NUEVO CRITERIO PARA LA ELECCIÓN DEL TAMAÑO DE LA MUESTRA DE ENTRENAMIENTO PARA SELECCIÓN DE MODELOS Y PREDICCIÓN: LA REGLA DE LA RAÍZ CÚBICA

*ISRAEL ALMODOVAR*[1], *LUIS PERICCHI*[1,2]

Paper Invited

**ABSTRACT**: The size of a training sample in Objective Bayesian Testing and Model Selection is a central problem in the theory and in the practice. We concentrate here in simulated training samples and in simple hypothesis. The striking result is that even in the simplest of situations, the optimal training sample $M$, can be *minimal* (for the identification of the *sampling* model) or *maximal* (for optimal prediction of future data). We suggest a compromise that seems to work well whatever the purpose of the analysis: the *5% cubic root rule*: $M = \min[0.05 * n, \sqrt[3]{n}]$.

We proceed to define a comprehensive loss function that combines identification errors and prediction errors, appropriately standardized. We find that the very simple cubic root rule is extremely close to an over- all optimum for a wide selection of sample sizes and cutting points that define the decision rules. The first time that the cubic root has been proposed is in Pericchi(2010). This article propose to generalize the rule and to take full statistical advantage for realistic situations. Another way to look at the rule, is as a synthesis of the rationale that justify both AIC and BIC.

**KEYWORDS**: 5% Cubic root rule, Intrinsic priors, Objective bayesian hypothesis testing, Training sample size.

**RESUMEN**: El tamaño de una muestra de entrenamiento en la selección y prueba en Bayesiana objetiva es un problema central en la teoría y en la práctica. Nos concentraremos en muestras de entrenamiento simuladas y en pruebas de hipótesis simples. El resultado impactante es que, aun en las situaciones más simples, la muestra de entrenamiento $M$ óptima puede ser *minimal* (para la identificación del modelo muestral) o *maximal* (para la predicción óptima de datos futuros). Se sugiere un compromiso que parece funcionar bien para cualquier propósito del análisis: la *regla de la raíz cúbica del* 5%: $M = min[0.05*n, \sqrt[3]{n}]$. Se procede a definir una función de pérdida comprehensiva que combina los errores de identificación y los errores de predicción, estandarizados apropiadamente. Se halla que la regla de la raíz cúbica simple es cercana en extremo a un óptimo general para una amplia selección de tamaños muestrales y puntos de corte que definen las reglas de decisión. La primera vez que se ha propuesto la raíz cúbica fue en Pericchi(2010). Este artículo propone generalizar la regla y tomar una ventaja estadística completa para situaciones reales. Otra forma de ver la regla es como una síntesis de la racionalidad que justifica tanto el AIC como el BIC.

**PALABRAS CLAVE**: Regla de la raíz cúbica 5%, Apriori intrínseca, Pruebas de hipótesis bayesianas objetivas, Tamaño de muestra de entrenamiento.

[1] Department of Statistics, Iowa State University, Ames, IA 50011.
*almodova@iastate.edu*
[2] Department of Mathematics, University of Puerto Rico, Río Piedras Campus, PR 00936-8377.
*luarpr@gmail.com.*

# 1 INTRODUCTION

## 1.1 How large should be the training sample size $M$ in *Wonderland*?

The original approach taken by Berger and Pericchi (1995, 1996) the "*Empirical Intrinsic Bayes Factor*", was based on taken (several) real minimal (in size) training samples, and then forming the arithmetic average of Bayes Factors. Minimal training samples were the most natural, since each training sample used for training was lost for discrimination of models, and furthermore for several points of view *minimality* of training samples encapsulates the concept of *Unit Information Priors*, Kass and Wasserman (1995), and *Matching Predictive Priors* Berger and Pericchi (2001). The empirical intrinsic approach besides solving a problem, that of finding a proper scaling of a Bayes Factor, opened up a host of new possibilities for Objective Hypothesis Testing and Model Selection. Among them:

1. The possibility of using longer than minimal real training samples.

2. The possibility of using intrinsic priors directly, without any real training sample (when intrinsic priors can be calculated).

3. The possibility of using simulated training samples $M$, of virtually *any* size smaller than the sample size $M \leq n$.

The third possibility offers a *"wonderland"* or *"free lunch"* situation , on which the use of training samples does not *"waste"* real samples. The usual situation in *Cross-Validation* or in the framework of Chakrabarti and Ghosh (2007), is not in Wonderland, since they use real training samples, but Casella and Moreno (2009) for example, are in situation 3, that is in Wonderland. Now the question we address is: How can we decide an *optimal* training sample size $M^*$ in Wonderland? For different purposes of the analysis, does the optimal changes?
For simple null hypotheses, we explore two perspectives:

1. The first is the **identification perspective** when the correct identification of the hypothesis is the central focus. We explore with rules like "$T$he Type I Error Rule" for selecting the (simulated) training sample size, or *minimizing the sum of Type I and Type Error II.* We conclude in favor for **extremely** small training samples if not minimal. This validates the original suggestion by Berger and Pericchi (1996) of using *minimal training samples.*

2. On the other hand if **prediction of future observations** with square loss by model averaging is the primary goal, that is the *prediction perspective*, then very large training samples may reduce the error of prediction. For purely prediction purposes, this opens up a window of exciting new methods.

These opposing answers leave us in a drama when the purpose of analysis is not absolutely either identification or prediction. We suggest then a compromise solution on which the errors of identification are kept small, assuring large sample consistency and the error of prediction is drastically diminished. We call it the *5% cubic root* rule on which the compromise training sample $M^*$ is taken as a simple function of the sample size $n : M^* = \min[0.05 \times n, \sqrt[3]{n}]$. More general null hypothesis will be addressed elsewhere.

## 2 GENERAL METHODOLOGY

In this paper, for simplicity and ease of exposition, we assume that we have a point simple null hypothesis, that is:

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \text{ against } H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0.$$

More general situations, is subject of current research, see Pericchi (2009).
Our starting point is: **the intrinsic prior equation** which for exchangeable observations (for the sake of simplicity) is (Berger and Pericchi, 1996a,b):

$$\pi^I(\boldsymbol{\theta}) = \pi^N(\boldsymbol{\theta}) \cdot \int f(\mathbf{x}(l)|\boldsymbol{\theta}) \cdot \frac{f(\mathbf{x}(l)|\boldsymbol{\theta}_0)}{m_1^N(\mathbf{x}(l))} \cdot d\mathbf{x}(l), \tag{1}$$

where $\mathbf{x}(l)$ is a random training sample of size (length) $M$ and the marginal density is define as,

$$m_1^N(\mathbf{x}(l)) = \int f(\mathbf{x}(l)|\boldsymbol{\theta})\pi^N(\boldsymbol{\theta})d\boldsymbol{\theta}. \tag{2}$$

**Example 1. Exponential Distribution:** Assume the data $X_1, \ldots, X_n$ comes from a exponential likelihood with parameter $\beta$:

$$f(x_i|\beta) = \beta \cdot \exp(-\beta \cdot x_i), \quad x_i > 0, \beta > 0 \text{ and } i = 1, \ldots, n.$$

In this illustration the hypotheses are:

$$H_0 : \beta = \beta_0 = 2 \text{ against } H_1 : \beta \neq \beta_0.$$

It turns out, using the moment generating function of an exponential distribution, that for a training sample of length $M$, in terms of the sufficient statistics $\bar{X}_M = \frac{1}{M}\sum_{i=1}^{M} X_i$. The density of $\bar{X}_M$ is,

$$f(\bar{X}_M|\beta) = \frac{(M \cdot \beta)^M}{\Gamma(M)} \cdot \bar{X}_M^{(M-1)} \cdot \exp(-M\beta\bar{X}_M). \tag{3}$$

The Jeffreys prior for $\beta$ is: $\pi^N(\beta) = 1/\beta$, and it follows that $m_1^N(\bar{X}_M) = 1/\bar{X}_M$.

Therefore the intrinsic prior (IPrior) for a training sample of length $M$ is using expressions (1) and (2),

$$\pi^I(\boldsymbol{\theta}) = \pi^N(\boldsymbol{\theta}) \cdot \int f(\bar{X}_M|\beta)[M \cdot \beta_0 \cdot \bar{X}_M]^M \exp[-M \cdot \beta_0 \cdot \bar{X}_M]d\bar{X}_M. \tag{4}$$

It turns out that that this integral is:

$$\pi^I(\boldsymbol{\theta}) = \frac{\Gamma(2 \cdot M)}{\Gamma(M)^2} \cdot \beta_0^M \cdot \frac{\beta^{(M-1)}}{(\beta_0 + \beta)^{(2 \cdot M)}}. \tag{5}$$

This is a very interesting distribution. In fact this is known as a beta distribution of the second kind,

$$\beta/\beta_0 \sim Beta_2(M, M).$$

**Definition**: *Beta Distribution of the Second Kind*

$$f(y|p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} \cdot \frac{y^{p-1}}{(1+y)^{(p+q)}}, y > 0,$$

denoted by $Y \sim Beta_2(p,q)$. To generate samples from a beta distribution of second kind we can use the following algorithm:
**Generating a** $Y \sim Beta_2(p,q)$:

1. Generate $z \sim Beta(p,q)$.
2. Then let $y = \frac{z}{1-z}$.

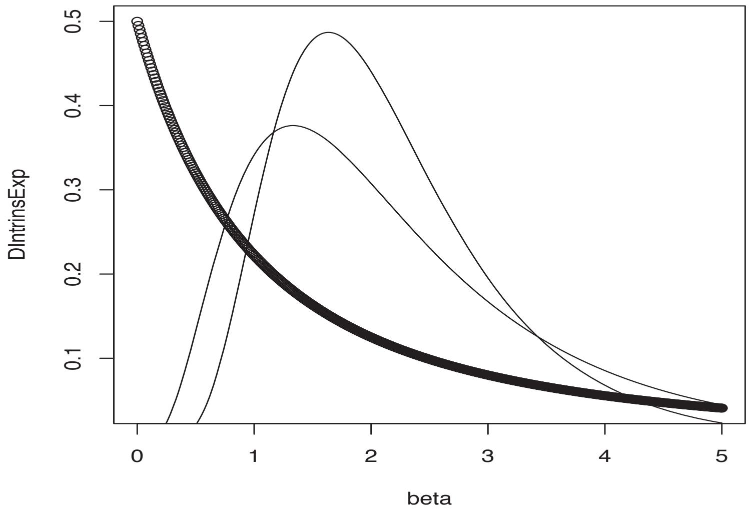

Figure 1: Intrinsic priors for exponential example with minimal training sample $M = 1$ (points) and $M = 5$, $M = 10$ and $H_0 = \beta_0 = 2$.

In Figure , the IPrior is plotted for $M = 1$, minimal training sample, and also for $M = 5$ and 10. All three priors are sensible since apart form being proper, their medians are all equal to $\beta_0 = 2$, so it is balanced and the null hypothesis is in the ”center” of the prior. However, it can be argued, that for $M > 1$, the prior is more concentrated around the null hypothesis, which is ”more pleasant to the mind”. Fine, but now we have opened a ”Pandora Box”, how to assign $M$? A possibility is to consider the class of IPriors for $1 < M < n$, where $n$ is the sample size, this is the approach taken by Casella and Moreno (2009). This is fine if there is robustness with respect to the decision. But this is the exception rather than the rule, since there is enormous variation in that class of priors. Decision robustness can only be obtain when there is a overwhelming support for one of the hypothesis. It is the intermediate cases which are the most interesting, particularly when the data tend to favor the null. This can not be addressed by the whole class of IPriors. As a case in point consider a normal location likelihood, and assume that the $p$-value is equal to one half with sample

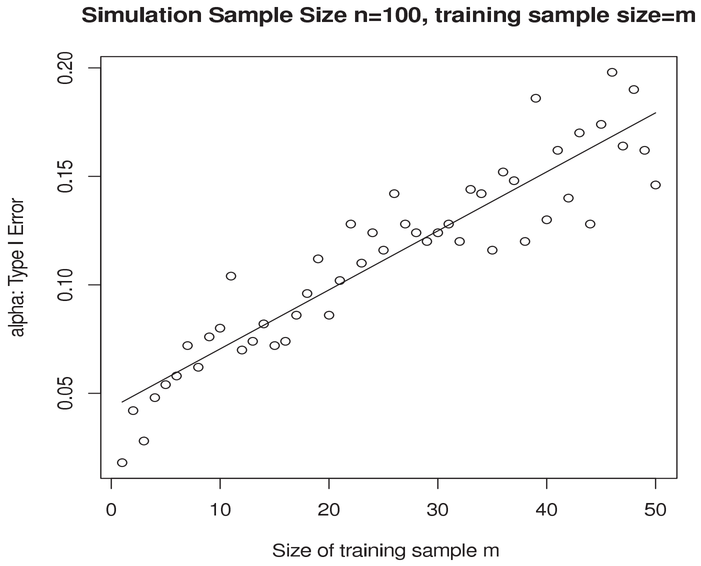

Figure 2: Intrinsic priors for exponential example with minimal training sample $M = 1$ (points) and $M = 5$, $M = 10$ and $H_0 = \beta_0 = 2$.

size one hundred. The evidence in favor of the null ought to be strong or more. The Bayes Factor starting from a value of 120, for a minimal training sample goes down as the training sample size $M$ grows to $n$, to a value close to one, that is null and alternative equally likely, Pericchi (2010). So the class of IPriors for all possible $M$ is too wide, for sensitivity analysis, at least when the data are reasonably supporting the null hypothesis.
We explore different ways to assign $M$ on a non trivial set of training samples $1 < M < M_1$.
In order to compute the Bayes Factor in this example, we need to calculate:

$$m^I(\bar{y}) = \int \beta^n \exp(-\beta \cdot n \cdot \bar{y}) \times \pi^I(\beta). \tag{6}$$

Since this is an involved integral we resort to the following general approximation (Berger and Pericchi, 1996)

$$m^I(\bar{y}) = m^N(\bar{y}) \times [\pi^I(\hat{\beta})/\pi^N(\hat{\beta}) + o(1/n)]. \tag{7}$$

Using this approximation it follows that the Bayes Factor satisfies approximately the following expression.
The Bayes Factor for the exponential example is,

$$B_{01}^M = \beta_0^n \exp(-\beta_0 \cdot n\bar{y}_n)(\bar{y}_n)^{(n+M)}(\beta_0 + \frac{1}{\bar{y}_n})^{(2\cdot M)} \times \frac{n^n \Gamma(M)^2}{\beta_0^M \Gamma(n)\Gamma(2\cdot M)}. \tag{8}$$

We try to explore here a rule like: **Type I Error Rule:** Compute:

$$Pr(B_{01}^M \leq 1 | H_0 : \beta = \beta_0) = \alpha(M, n), \tag{9}$$

by simulation from $H_0$. The range to be considered for $M$ is such that $\alpha(M, n) \leq 0.05$.
We perform extensive simulations in Figure 2, and we conclude that Type I Error is increasing with

$M$, and thus the optimal is the minimal training sample $M_0 = 1$, although small values of $M$ above $M = 1$ lead to small type I error. An interesting *curiosity* is that a 5% type I error corresponds to a $M = 0.05 \times n$. This *curiosity* will be confirmed in our next important example.
Tentatively we consider the class of IPriors:

$$\text{Five Percent class: } \{\pi^I(\beta|M), 1 \leq M \leq \min[0.05 \times n, H]\}.$$

**Remark:** It is advisable to avoid letting $M$ be a proportion of the sample size $n$, since then consistency will be lost when sampling from the null hypothesis, as will be clearly seen in the next example. By bounding the class by $H$ as above, consistency, under the null hypothesis, is guaranteed. Note that in the exponential example $H = 5$. To continue the exploration of different routes to choose $M$ in a non-trivial manner, we move to a normal example.

# 3 AN EXACT CALCULATION

We now explore in depth a case amenable to all calculations, for a normal mean with known variance.

$$Y_i \sim N(\mu, \sigma_0^2), \text{ with hypotheses: } H_0 : \mu = \mu_0 \text{ against } H_1 : \mu \neq \mu_0.$$

Using the intrinsic prior equation (1), for a training sample of size $M$ we find

$$\pi^I(\mu) = N(\mu_0, 2\sigma_0^2/M). \tag{10}$$

Using this prior we get the marginal under $H_1$:

$$m_1^I(\bar{y}|M) = N\left(\mu_0, \sigma_0^2\left(\frac{M+2n}{nM}\right)\right) \tag{11}$$

using equation (10), the Bayes Factor is,

$$2\log(B_{01}) = \log(1 + 2n/M) - \frac{n(\bar{y}-\mu_0)^2}{\sigma_0^2} \cdot \frac{2n}{M+2n}. \tag{12}$$

Then under $H_0$

$$\frac{n(\bar{y}-\mu_0)^2}{\sigma_0^2} \sim \chi_1^2, \tag{13}$$

and under $H_1$ according to the marginal density (11),

$$\frac{(\bar{y}-\mu_0)^2}{\sigma_0^2(\frac{M+2n}{n \cdot M})} \sim \chi_1^2. \tag{14}$$

## 3.1 The identification route

We begin setting the problem as an *identification problem*, and we compare procedures by (Bayesian versions) of Type I and Type II errors.

**Definition of Bayesian Type I Error:**

$$\alpha_B(M, n) = Pr[2\log(B_{01}) < 0|H_0]. \tag{15}$$

In the example of this section, since under $H_0$, $\frac{n(\bar{y}-\mu_0)^2}{\sigma_0^2} \sim \chi_1^2$ it turns out that:

$$\alpha_B(M,n) = Pr(\chi_1^2 > (1+f/2))\log(1+2/f)), \tag{16}$$

where $f = M/n$ is the fraction of the sample, which is the training sample, e.g $f = 0.05$ is that the training sample is 5% of the sample $n$.

**Definition: Bayesian type Error II**:

$$\beta_B(M,n) = Pr(2\log(B_{01}) > 0|H_1), \tag{17}$$

where the probability is calculated from $m_1(\mathbf{y}|M)$. In the normal mean example it turns out that

$$\beta_B(M,n) = Pr(\chi_1^2 < f/2 \cdot \log(1+2/f)), \tag{18}$$

since under $m_1$ it is the case that: $\frac{nM(\bar{y}-\mu_0)^2}{\sigma_0^2(M+2n)} \sim \chi_1^2$.
The surprising fact is that *b*oth Type I and Type II Bayesian errors are increasing with $f$, see Figures 3 -7. In fact is very interesting that 5 percent as training sample leads to a near 5 percent of Bayesian type I error, a *curiosity* already noted in the previous example.

**Conclusion :** To reduce both types of error, we should reduce training samples, that is take minimal training samples, in this case $M = 1$ is optimal *in that sense*. If a compromise has to be reached still the training sample sizes should be rather smaller than large.
Alternatively, instead of a fraction we may like to use a power of sample size. The fact is that if we use any fraction $f > 0$ of sample size, then under $H_0$ the $2\log(B_{01})$ is bounded for all sample sizes, an so it is not consistent under $H_0$. This fact can be easily deduced for example from equation (11), for $M/n = f = constant$.
For $f = 0.05$, Figure 4 tells us that a power of 0.7, approximates for a long range, so if this power is assumed the 5 percent rule is approximately obeyed. Sample size to the power of 0.7, will yield a consistent procedure, but with a high Bayesian Ttype II error, of about $\beta_B = 0.25$, see Figure 6, and with a non-negligible Type I error of 0.05 even for very large sample sizes. This seems to be unacceptably large. Thus if we are going to envisage a *"power rule"*, that power should be much smaller than 0.7. Next, we turn to a radically different criterion.

## 3.2 The prediction route

We now take a completely different route: We focus on the Bayesian risk of predicting the observations, but **NOT** by one single model but by the **fully Bayesian route of model averaging**, which is the most widely justifiable procedure under a Bayesian point of view. For which training sample size $M$ will be the Bayesian risk minimized? First of all notice that typically the variance of the log of the Bayes Factor decreases with $M$.

**Normal Example Continued**: From expression (11) we get that under $H_0$ the $Var(2\log(B_{01}))$ is $2\cdot(\frac{2n}{M+2n})^2$ or $2\cdot(\frac{2}{2+f})^2$. Under $m_1$ the factor turns to be $(\frac{2}{f})^2$ both decreasing with $M$ as seen in Figure 6. This variance reduction may anticipate an effect in the reduction of error of prediction.

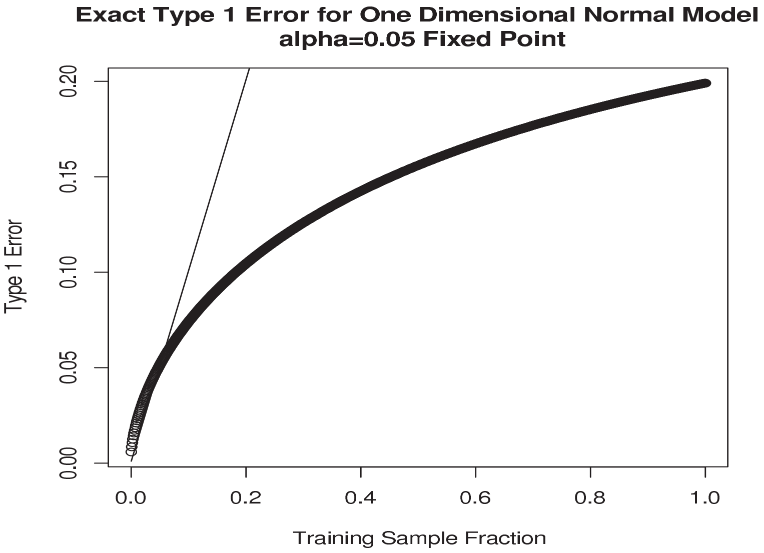

Figure 3: Bayesian type I error as function of $M$, 5% proportion produces 5% type I error.

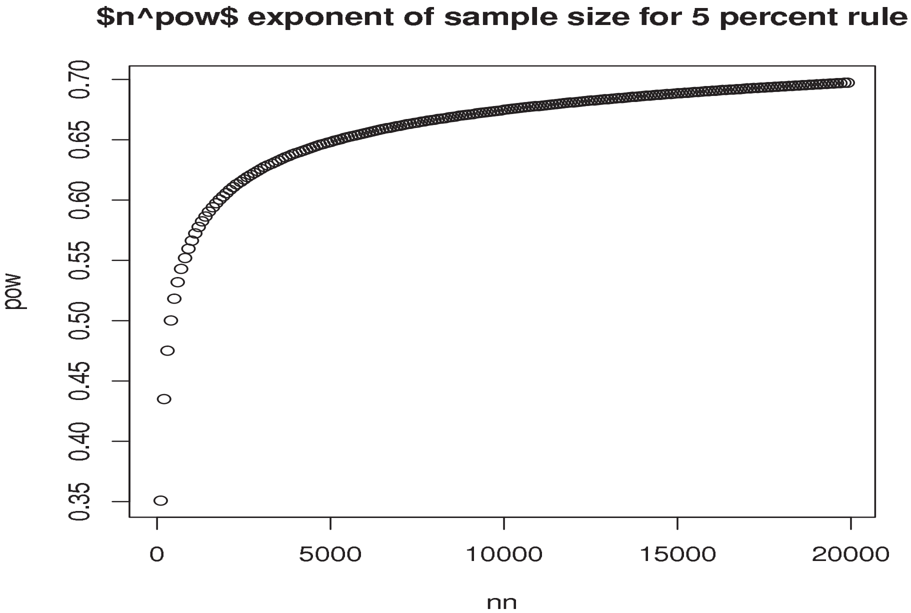

Figure 4: Power of $M$ equivalent to 5%.

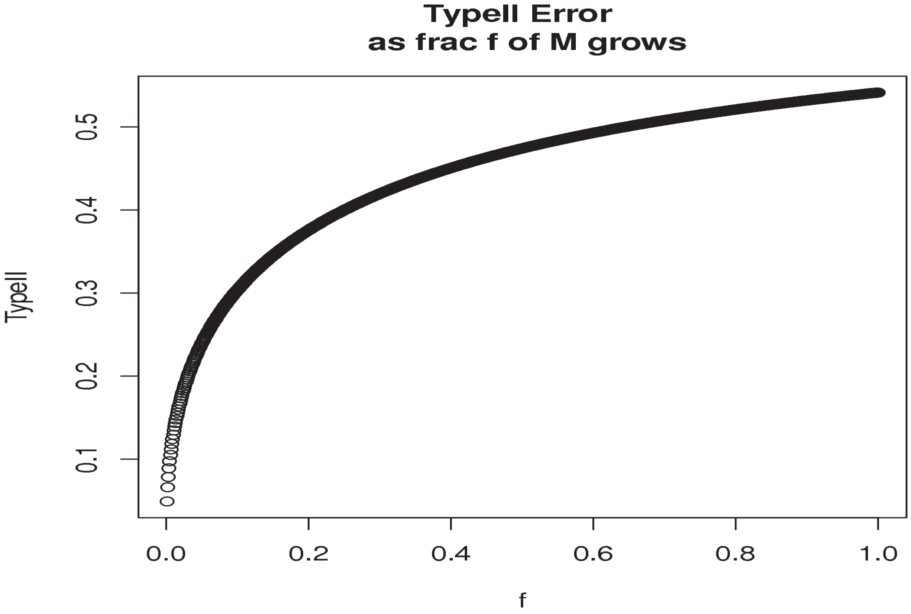

Figure 5: Bayesian type II error as function of $M$, growing steady.

## 4 THE PREDICTION CRITERIA

**Definition: The prediction risk as a function of** $M$
We use the square error loss, the most used loss function. The classical definition of the square loss is,

$$Risk(M) = \int (\hat{y}^* - \bar{y})^2 p(\bar{y}|M) d\bar{y}, \tag{19}$$

where,

$$p(\bar{y}|M) = P(H_0|\mathbf{y}, M) f(\bar{y}|\mu_0) + P(H_1|\mathbf{y}, M) m_1(\bar{y}|M),$$

and

$$\hat{y}^* = P(H_0|\mathbf{y}, M) \int \bar{y} f(\bar{y}|\mu_0) d\bar{y} + P(H_1|\mathbf{y}, M) \int \bar{y} m_1(\bar{y}|M) d\bar{y} = P(H_0|\mathbf{y}, M)\hat{y}_0^* + P(H_0|\mathbf{y}, M)\hat{y}_1^*.$$

Lengthy algebra shows that the risk function can be express as,

$$Risk(M) = (\hat{y}_0^* - \hat{y}_1^*)^2 P(H_0|\mathbf{y}, M) \cdot P(H_1|\mathbf{y}, M) + Var_0(\bar{y}) P(H_0|\mathbf{y}, M) + Var_1(\bar{y}|M) P(H_1|\mathbf{y}, M), \tag{20}$$

where $Var_0(\bar{y}) = \sigma_0^2/n,$ and $Var_1(\bar{y}|M) = \sigma_0^2(1/n + 2/M)$.

Figures 9 show the reduction of the risk as the training sample size grows. Note however an important phenomenon: the decrease in risk is quite fast. So allowing slightly bigger than minimal training samples permits a drastic reduction of the prediction risk. On the other hand we see in Figure that the change in the probability of the null may be size-able with $M$. Taking into consideration all these facts we finally propose the rule:

$$M = \sqrt[3]{n}.$$

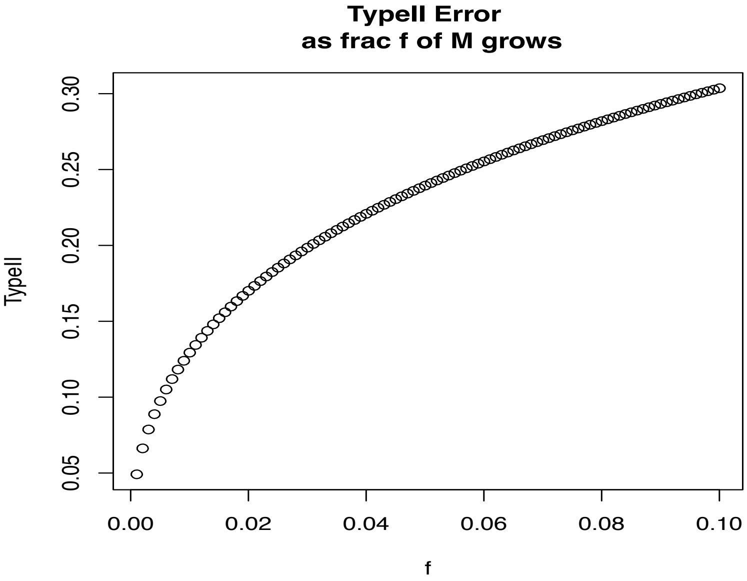

Figure 6: Bayesian type II error as function of $M$ 5% proportion produces 25% type II error.

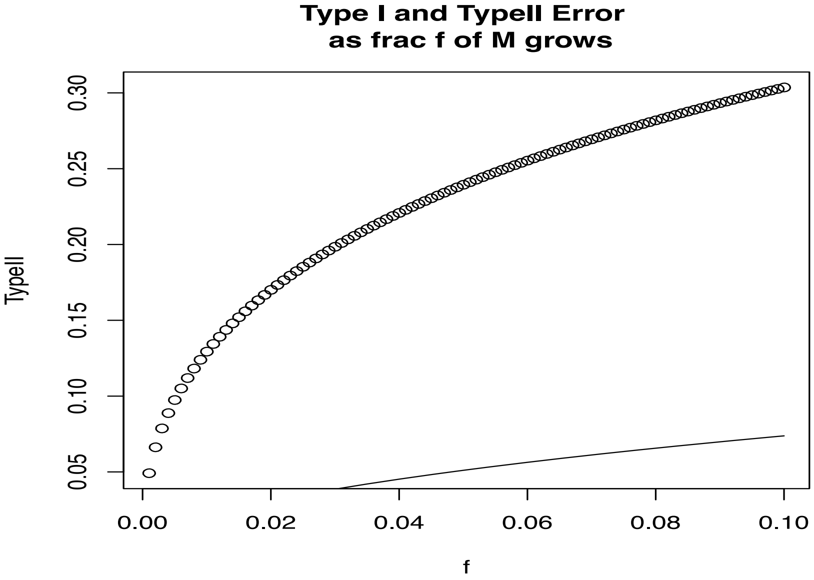

Figure 7: Bayesian type I and type II error as function of $M$. To minimize error minimize $M$.

The cubic root rule is next justified formally by a comprehensive loss function.

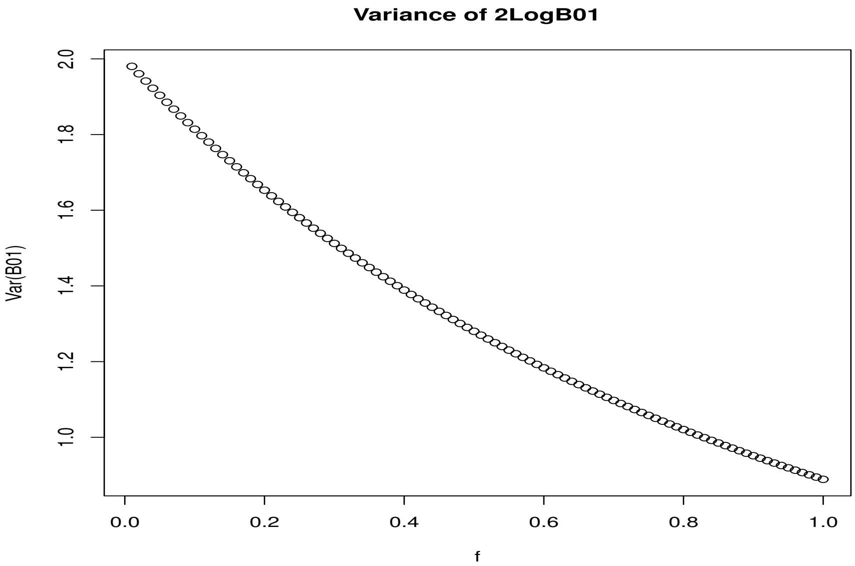

Figure 8: Variance of $2\log B_{01}$ as function of fraction $f = M/n$.

## 5 AN INTEGRATED LOSS FUNCTION AND A GLOBAL SOLUTION

The ”best of the two worlds” solution, good identification and good prediction, may be achieved by minimizing an integrated loss, which involves all Type I, II and prediction errors, as suggested by J.K.Ghosh. We propose to minimize its sum. However care should be taken, so that errors would be comparable. We minimize the following scaled and comprehensive loss function:

$$\tilde{SE}(M,n) = \tilde{\alpha}(M,n) + \tilde{\beta}(M,n) + \tilde{Risk}(M,n); \tag{21}$$

where for each fixed sample size $n$ all the components have been normalized by its maximum value, i. e

$$\tilde{\alpha} = \frac{\alpha(M,n)}{\max_M \alpha(M,n)}, \tilde{\beta} = \frac{\beta(M,n)}{\max_M \beta(M,n)} \text{ and } \tilde{Risk} = \frac{Risk(M,n)}{\max_M Risk(M,n)}.$$

We search to find the following,

$$\tilde{SE}(M_0,n) = \min_M \tilde{SE}(M,n). \tag{22}$$

It should be noted that $M_0 = M_0(n,C)$, that is the solution depends upon the sample size $n$ and the Bayes Factor cutting point $C$. In the following Figures 10 and 11 it is striking that for both $C_1 = 1$ and $C_2 = 0.1$ and for a wide range of sample sizes $n = 10,\ldots,10,000$ is,

$$\tilde{SE}(M_0,n) = \min_M \tilde{SE}(M,n) \approx \tilde{SE}(\sqrt[3]{n},n). \tag{23}$$

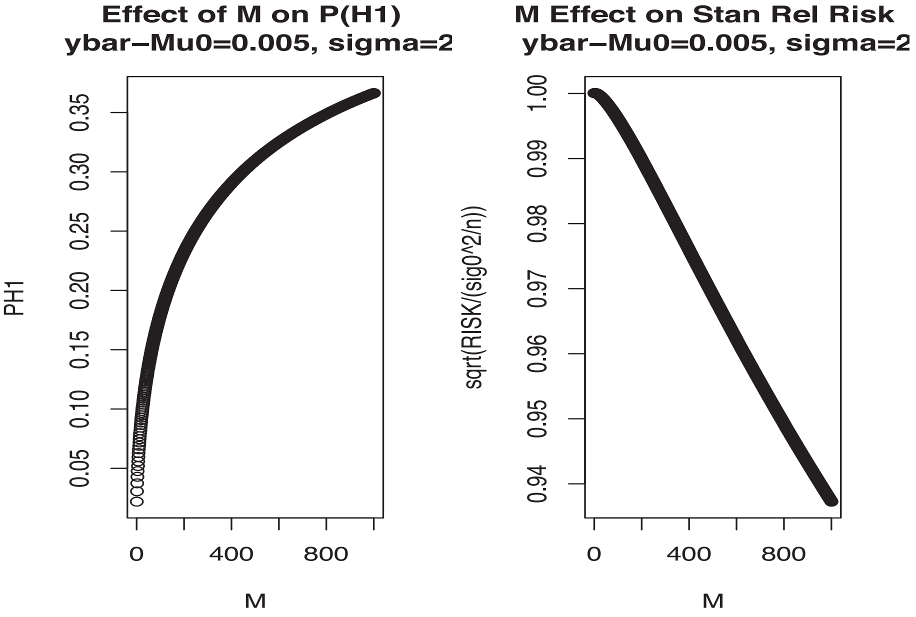

Figure 9: Effect on the probability of the alternative and the relative risk of the training sample size $M$.

**Initial Result:** For the Normal model using constants $C_1$ and $C_2$,

$$0.95 \leq \frac{\tilde{SE}(M_0,n)}{\tilde{SE}(\sqrt[3]{n},n)} \leq 1. \tag{24}$$

In the next Figures we see the impressive performance of the cubic root proposal, $M = \sqrt[3]{n}$. Particularly in Figures 12 and 13, for wide values of sample sizes and cutting points, the ratio (24) is extremely close to one. This propose method is based in the fruitful general method of intrinsic priors (originated in Berger and Pericchi (1996)), that has been very influential in Bayesian statistics, see for example in the Journal Bayesian analysis there is a section called intrinsic priors and another "Intrinsic Bayes Factors". The current project, builds on that theory but pushes the theory and application far more. Intrinsic priors were originally devised for minimal training samples, but now the amount of training is left as a decision rule. On the other hand, we introduce a new "comprehensive" loss function that after scaling, puts together errors of identification and errors of prediction, the two of foremost importance goals of model comparison exercises, which are in a sharp contradiction. Here we set up, for the first time for the best of our knowledge, a unifying framework that offers a synthesis of purposes, at a philosophical level. It emerges an eminent simple and powerful solution that $M$ should be a simple function of the cubic root of n, to achieve simultaneously good identification and prediction performances. The new criterion that emerges is a non-trivial and powerful method which is a synthesis of AIC and BIC.

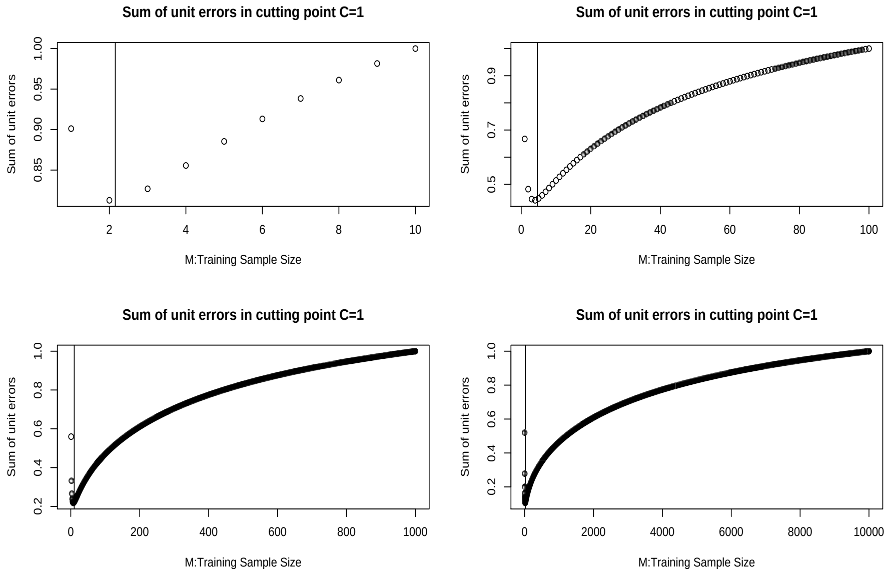

Figure 10: Minimum attained very close to the cubic root denoted by a vertical line with $C = 1$.

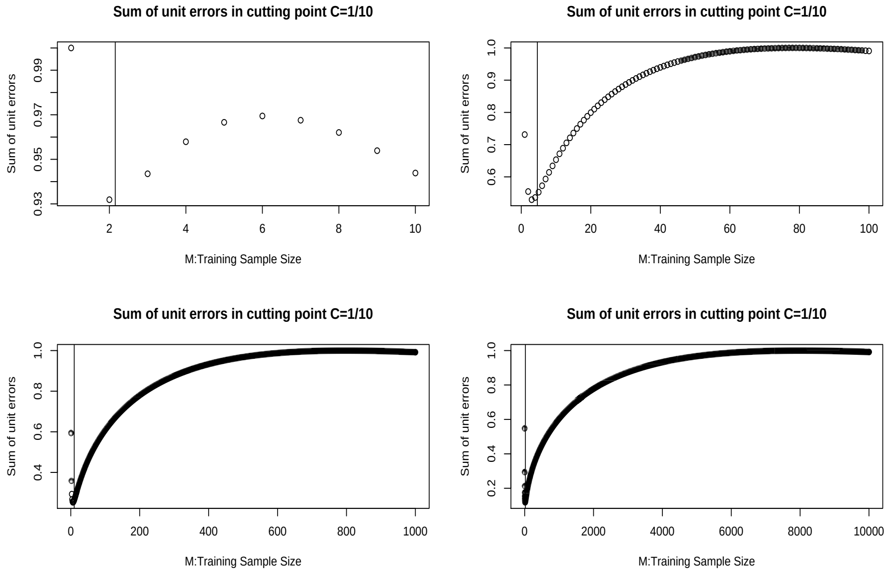

Figure 11: Minimum attained very close to the cubic root denoted by a vertical line with $C = 0.1$.

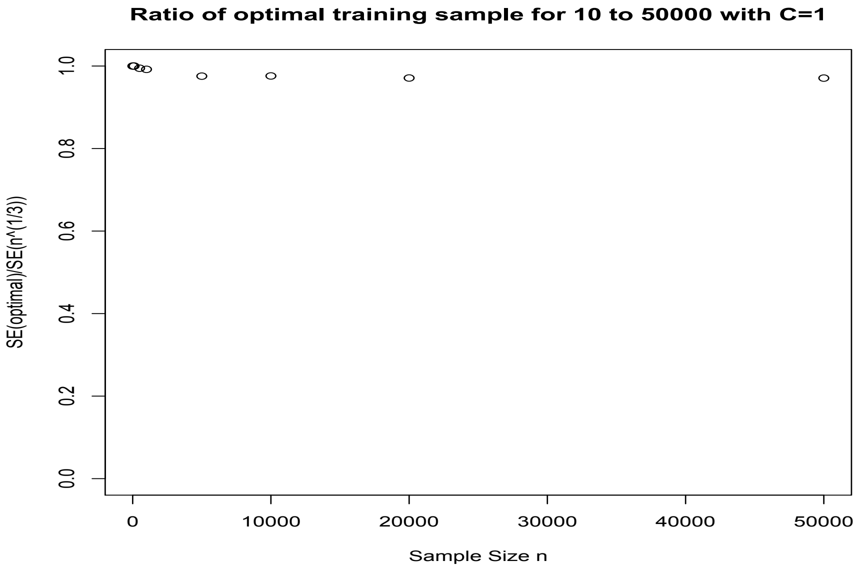

Figure 12: Ratio of the minimal loss function over the loss function evaluated at the cubic root point using $C = 1$.

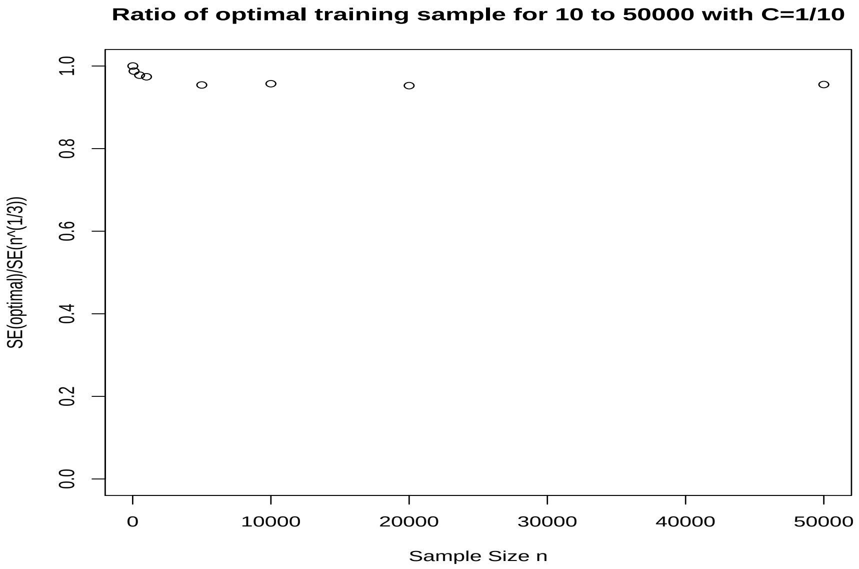

Figure 13: Ratio of the minimal loss function over the loss function evaluated at the cubic root point using $C = 0.1$.

# 6 CONCLUSIONS

## 6.1 Discussion of the FivePercent-Cubic-Root Rule

1. *Limitations:* It should first be noted that we considered situations on which the null hypothesis is simple and the alternative is one-dimensional. Generalizations are being worked out.

2. *Versions:* Apart from the *continuous* version, we may take *integer* versions, by taking the function ”ceiling” in the rule above. There is nothing to prevent the use of fractional $M$ though, like $M = 0.25$ for $n = 5$. The `ceiling` version would lead to $M = 1$. This seems fine, but smaller type I errors seems preferable.

3. `Pure Cubic Root Rule?` Setting the rule simply as: $M = \sqrt[3]{n}$, achieve most of the desiderata of a sensible procedure. Nevertheless the simple cubic root will allow higher values of type I error for very small sample sizes, so our preference for $M = \min[n \times 0.05; \sqrt[3]{n}]$.

4. `Different Thresholds:` It may be argued that the rule depends on the threshold (on the log Bayes Factor) of 0 in the definitions of Bayesian Type I and II Errors. However the value of zero is a natural one, and does not seem essential in our arguments which involve type II error. On the other hand, Type I error seems to be fairly flat with respect to the training sample size for smaller than one thresholds.

5. *Tentative nature of the rule:* Still the rule should be contrasted with practical experiences, for fine tuning. However a compromise between minimal and maximal training samples seems useful to be considered.

# References

Abramowitz, M.; Stegun, I. A. (1972), Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. New York: Dover Publications, Inc.

Berger J.O. and Pericchi L.R. (1996a), The Intrinsic Bayes Factor for Model Selection and Prediction. *Jour, Amer. Statist. Ass.*, **91**, 109-122.

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

Casella G. and Moreno E. (2009), Assessing Robustness of Intrinsic Test of Independence in Two-way Contingency Tables. Tech Report.

Chakrabarti A. and Ghosh J. (2007), Some Aspects of Bayesian Model Selection for Prediction. *Bayesian Statistics* 8, 51-90.

Kass R.E. and Wasserman L. (1995), A Reference Bayesian Test for Nested Hypothesis and its Relationship with Schwarz Criterion. *Jour, Amer. Statist. Ass.* **90**, 928-934.

Pericchi, L.R. (2010), How large should be the training sample? Invited Chapter in the book: *”Frontiers of Decision Making and Bayesian Analysis. In Honor of James O. Berger”*, Chen MH et al editors. Springer. (In press).

Spiegelhalter DJ, Best NG, Carlin BP and Van der Linde A, (2002), Bayesian Measures of Model Complexity and Fit (with Discussion). *Journal of the Royal Statistical Society*, Series B, 64(4):583-616, and in The BUGS Project DIC www.mrc-bsu.cam.ac.uk/bugs/winbugs/dicpage.shtml.