Sample Size Determination: Posterior Distributions Proximity

Nikita Kiselev[1*] and Andrey Grabovoy[1]

[1]Moscow Institute of Physics and Technology, Dolgoprudny, Russia.

*Corresponding author(s). E-mail(s): kiselev.ns@phystech.edu;
Contributing authors: grabovoy.av@phystech.edu;

Abstract

The issue of sample size determination is crucial for constructing an effective machine learning model. However, the existing methods for determining a sufficient sample size are either not strictly proven, or relate to the specific statistical hypothesis about the distribution of model parameters. In this paper we present two approaches based on the proximity of posterior distributions of model parameters on similar subsamples. We show that these two methods are valid for the model with normal posterior distribution of parameters. Computational experiments demonstrate the convergence of the proposed functions as the sample size increases. We also compare the proposed methods with other approaches on different datasets.

Keywords: Sufficient sample size, Posterior distributions proximity, Normal posterior distribution, Linear regression

1 Introduction

The task of supervised machine learning involves selecting a predictive model from a parametric family. This choice is usually based on certain statistical hypotheses, such as maximizing a quality functional. A model that satisfies these statistical hypotheses is called an *adequate* model [1–3].

When planning a computational experiment, it is necessary to estimate the minimum sample size — the number of objects required to build an adequate model. The sample size required to build an adequate predictive model is called *sufficient* [4–6].

This work addresses the issue of determining the sufficient sample size. There are numerous studies dedicated to this topic, with approaches classified into statistical, Bayesian, and heuristic methods.

Some of the early researches on this topic [7, 8] formulate a specific statistical criterion, where the sample size estimation method associated with this criterion guarantees achieving a fixed statistical power with a Type I error not exceeding a specified value. Statistical methods include the Lagrange multipliers test [9], the Likelihood ratio test [10], the Wald statistic [11]. Statistical methods have certain limitations associated with their practical application. They allow for estimating the sample size based on assumptions about the data distribution and information about the agreement of observed values with the assumptions of the null hypothesis.

The Bayesian approach also has a place in this problem. In the work [12] the sufficient sample size is determined based on maximizing the expected utility function. This may explicitly include parameter distribution functions and penalties for increasing the sample size. This work also considers alternative approaches based on constraining a certain quality criterion for estimating model parameters. Among these criteria, the Average Posterior Variance Criterion (APVC), Average Coverage Criterion (ACC), Average Length Criterion (ALC), and Effective Sample Size Criterion (ESC) stand out. These criteria have been further developed in other works, for example, [13] and [14]. Over time, the authors of [15] conducted a theoretical and practical comparison of methods from [7, 8, 12].

Authors like [16], as well as [17], discuss the differences between Bayesian and frequentist approaches in determining sample size. They also propose robust methods for the Bayesian approach and provide illustrative examples for some probabilistic models.

In the paper [18], various methods for estimating sample size in generalized linear models are considered, including statistical, heuristic, and Bayesian methods. Methods such as Lagrange Multiplier Test, Likelihood Ratio Test, Wald Test, Cross-Validation, Bootstrap, Kullback-Leibler Criterion, Average Posterior Variance Criterion, Average Coverage Criterion, Average Length Criterion, and Utility Maximization are analyzed. The authors point out the potential development of combining Bayesian and statistical approaches to estimate sample size for insufficient available sample sizes, which is a pre hoc sample size estimation.

In [19] a method for determining sample size in logistic regression is discussed, using cross-validation and Kullback-Leibler divergence between posterior distributions of model parameters on similar subsamples. Similar subsamples are those that can be obtained from each other by adding, removing, or replacing one object.

In this paper, two approaches based on the distance between the posterior distributions are presented. It is proposed to consider two similar subsamples. The posterior distributions of the model parameters over these subsamples turn out to be close if the sample size is sufficient. It is proposed to use the Kullback-Leibler divergence [19] as a measure of the proximity of distributions, as well as the s-score model comparison function [20].

The novelty of this work lies in proving the correctness of the proposed methods. The correctness of the definition for a given model implies that the corresponding

function converges to zero as the sample size increases. This establishes a foundation for selecting a specific sample size based on the descending of the respective function. Correctness is proved in a probabilistic model with a normal posterior distribution of parameters. For the linear regression model, the theorem on the moments of the limit posterior distribution of parameters is proved.

Through the detailed empirical study, we compare our methods with others. This comparison allows to conclude the main features of the methods we propose. Namely, that KL-sufficient sample size is typically the largest, while S-sufficient, on the contrary, is the smallest. The reason for this lies in the functions used to compare probability distributions. We encourage the reader to familiarize with the experiment code at the following link: https://github.com/kisnikser/Posterior-Distributions-Proximity.

2 Problem statement

An object is defined as a pair $(\mathbf{x}, y)$, where $\mathbf{x} \in \mathbb{X} \subseteq \mathbb{R}^n$ is the feature vector, and $y \in \mathbb{Y}$ is the target variable. In regression problems $\mathbb{Y} = \mathbb{R}$, and in K-class classification problems $\mathbb{Y} = \{1, \ldots, K\}$.

The feature-object matrix for a sample $\mathfrak{D}_m = \{(\mathbf{x}_i, y_i)\}, i \in \mathcal{I} = \{1, \ldots, m\}$ of size m is called the matrix $\mathbf{X}_m = [\mathbf{x}_1, \ldots, \mathbf{x}_m]^\top \in \mathbb{R}^{m \times n}$.

The target variable vector for a sample $\mathfrak{D}_m = \{(\mathbf{x}_i, y_i)\}, i \in \mathcal{I} = \{1, \ldots, m\}$ of size m is denoted by $\mathbf{y}_m = [y_1, \ldots, y_m]^\top \in \mathbb{Y}^m$.

A model is a parametric family of functions f, mapping the Cartesian product of the set of feature vector values $\mathbb{X}$ and the set of parameter values $\mathbb{W}$ to the set of target variable values $\mathbb{Y}$:

$$f : \mathbb{X} \times \mathbb{W} \to \mathbb{Y}.$$

A probabilistic model is a joint distribution

$$p(y, \mathbf{w}|\mathbf{x}) = p(y|\mathbf{x}, \mathbf{w})p(\mathbf{w}) : \mathbb{Y} \times \mathbb{W} \times \mathbb{X} \to \mathbb{R}^+,$$

where $\mathbf{w} \in \mathbb{W}$ is the set of model parameters, $p(y|\mathbf{x}, \mathbf{w})$ specifies the likelihood of an object, and $p(\mathbf{w})$ represents the prior distribution of parameters.

The task is to determine the sufficient sample size m^*. Let a criterion T be given. E.g. it can be constructed based on heuristics regarding the behaviour of model parameters.

Definition 1. *The sample size m^* is called **sufficient** according to the criterion T, if T holds for all $k \geqslant m^*$.*

Specifically, the criterion may include information about the quality of the resulting model. For instance, the Accuracy metric might take a value greater than 80%, or the criterion could be based on a certain degree of stability in the set of optimal parameters obtained. This aspect is explored in our subsequent work, as stability is understood here in terms of a minor change in the posterior distribution of model parameters.

3 Proposed sample size determination methods

In [19], it is suggested to use the Kullback-Leibler divergence to estimate a sufficient sample size in a binary classification problem. The idea is based on the fact that if

two subsamples differ from each other by one object, then the posterior distributions obtained from them should be close. This proximity is determined by the Kullback-Leibler divergence.

In this paper, the question of the correctness of this approach is considered. The method is studied in an arbitrary probabilistic model. As a measure of proximity, it is proposed to use not only the Kullback-Leibler divergence, but also the s-score similarity function from [20].

Consider two subsamples $\mathfrak{D}^1 \subseteq \mathfrak{D}_m$ and $\mathfrak{D}^2 \subseteq \mathfrak{D}_m$. Let $\mathcal{I}_1 \subseteq \mathcal{I} = \{1, \ldots, m\}$ and $\mathcal{I}_2 \subseteq \mathcal{I} = \{1, \ldots, m\}$ — corresponding to them subsets of indexes.

Definition 2. *Subsamples $\mathfrak{D}^1$ and $\mathfrak{D}^2$ are called **similar** if $\mathcal{I}_2$ can be obtained from $\mathcal{I}_1$ by deleting, replacing or adding one element, that is*

$$|\mathcal{I}_1 \triangle \mathcal{I}_2| = |(\mathcal{I}_1 \setminus \mathcal{I}_2) \cup (\mathcal{I}_2 \setminus \mathcal{I}_1)| = 1.$$

Consider two similar subsamples $\mathfrak{D}_k = (\mathbf{X}_k, \mathbf{y}_k)$ and $\mathfrak{D}_{k+1} = (\mathbf{X}_{k+1}, \mathbf{y}_{k+1})$ of sizes k and $k+1$, respectively. This means that the larger one is obtained by adding one element to the smaller one. Let's find the posterior distribution of the model parameters over these subsamples:

$$p_j(\mathbf{w}) = p(\mathbf{w}|\mathfrak{D}_j) = \frac{p(\mathfrak{D}_j|\mathbf{w})p(\mathbf{w})}{p(\mathfrak{D}_j)} \propto p(\mathfrak{D}_j|\mathbf{w})p(\mathbf{w}), \quad j = k, k+1.$$

Definition 3. *Let's fix some positive number $\varepsilon > 0$. The sample size m^* is called **KL-sufficient** if for all $k \geqslant m^*$*

$$KL(k) = D_{KL}(p_k \| p_{k+1}) = \int p_k(\mathbf{w}) \log \frac{p_k(\mathbf{w})}{p_{k+1}(\mathbf{w})} d\mathbf{w} \leqslant \varepsilon.$$

For a pair of normal distributions, the Kullback-Leibler divergence has a fairly simple form. Assume that the posterior distribution is normal, that is, $p_k(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_k, \mathbf{\Sigma}_k)$. Guided by the heuristic that the convergence of the moments of such a distribution should entail the proximity of posterior distributions on similar subsamples, the following statement can be formulated.

Theorem 1. *Let $\|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 \to 0$ and $\|\mathbf{\Sigma}_{k+1} - \mathbf{\Sigma}_k\|_F \to 0$ as $k \to \infty$. Then, in a model with a normal posterior distribution of parameters, the definition of a KL-sufficient sample size is correct. Namely, for any $\varepsilon > 0$, there is such a m^* that for all $k \geqslant m^*$ $KL(k) \leqslant \varepsilon$ is satisfied.*

This statement implies that the distance between two gaussian distributions approaches zero as their mean vectors and covariance matrices converge. This allows us to delve deeper into the issue of the convergence of the Kullback–Leibler divergence by examining analytical expressions for expectations and variances.

In this paper, it is proposed to use the s-score similarity function from [20] as a measure of proximity of distributions:

$$\text{s-score}(g_1, g_2) = \frac{\int_{\mathbf{w}} g_1(\mathbf{w}) g_2(\mathbf{w}) d\mathbf{w}}{\max_{\mathbf{b}} \int_{\mathbf{w}} g_1(\mathbf{w} - \mathbf{b}) g_2(\mathbf{w}) d\mathbf{w}}.$$

Definition 4. *Let's fix some positive number $\varepsilon > 0$. The sample size m^* is called* ***S-sufficient*** *if for all $k \geqslant m^*$*

$$S(k) = s\text{-}score(p_k, p_{k+1}) \geqslant 1 - \varepsilon.$$

As in the case of a KL-sufficient sample size, in a model with a normal posterior distribution, it is possible to write an expression for the criterion used. Thus, one more statement can be formulated.

Theorem 2. *Let $\|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 \to 0$ as $k \to \infty$. Then, in a model with a normal posterior distribution of parameters, the definition of an S-sufficient sample size is correct. Namely, for any $\varepsilon > 0$, there is such a m^* that for all $k \geqslant m^*$ $S(k) \geqslant 1 - \varepsilon$ is satisfied.*

The significance of this theorem is similar to that of Theorem 1. In essence, the closeness of normal distributions in terms of the s-score similarity function reduces to the convergence of their means. Notably, unlike Theorem 1, convergence of the covariance matrices is not necessary here.

Let the linear regression model have a normal prior distribution of parameters. By the conjugacy property of the prior distribution and likelihood, the posterior distribution is also normal. Thus, we come to one of the simplest examples of a model for which the theorems presented above are valid. In fact, simpler statements can be formulated for linear regression.

Theorem 3. *Let the sets of values of the features and the target variable be bounded, that is, $\exists M \in \mathbb{R} : \|\mathbf{x}\|_2 \leqslant M$ and $|y| \leqslant M$. If $\lambda_{\min}(\mathbf{X}_k^{\intercal}\mathbf{X}_k) = \omega(\sqrt{k})$ for $k \to \infty$, then in a linear regression model with a normal prior distribution of parameters $\|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 \to 0$ and $\|\mathbf{\Sigma}_{k+1} - \mathbf{\Sigma}_k\|_F \to 0$ as $k \to \infty$.*

This theorem is of primary importance in this paper. Its distinctive feature is that, under mild and straightforward assumptions, it implies the convergence of the moments of the posterior distribution of parameters.

The first assumption in Theorem 3 relates to the restriction on the range of values for the features and target variable. This is common in practical applications, so it serves primarily for theoretical analysis purposes.

The second condition of Theorem 3 is of greater interest, as it delves into the behavior of the minimum eigenvalue of the sample covariance matrix of features. Unfortunately, this paper does not provide theoretical guarantees for this convergence, although it is verified through experiments. This can be regarded as a limitation, so we will try to reveal this in the future work.

4 Computational experiment

This section provides an extensive empirical study of the proposed methods. Experiments consist of several parts. In the first one, we verify the convergences obtained trough the theoretical analysis. Further, we make size estimations for various sample sets using different approaches. Finally, we study the dependence of the sufficient sample size on available sample set.

4.1 Convergence Verification

Here we study whether the theoretical convergences we have obtained are actually observed in practice. Namely, first we look at the behavior of the minimum eigenvalue of the matrix $\mathbf{X}_k^\top \mathbf{X}_k$ as the sample size increases. Then we investigate the convergence of the proposed functions $KL(k)$ and $S(k)$. Finally, the dependence of sufficient sample size of the threshold parameters is studied. The experiment is conducted on two datasets: synthetic regression and Liver Disorders.

Synthetic data is generated from a linear regression model. The number of objects is 500, the number of features is 10. To generate a synthetic regression dataset, we have sampled original features, model parameters, and noise residuals from the standard normal distribution. The prior distribution of parameters was set to standard normal too, both for synthetic regression and Liver Disorders datasets, which has 345 objects and 5 features. We have preprocessed the input features using standard scaler.

One object is sequentially removed from the given sample until the number of objects in the subsample is equal to the number of features. For each sample size k we calculate the minimum eigenvalue of the matrix $\mathbf{X}_k^\top \mathbf{X}_k$. Also, the values of $KL(k)$ and $S(k)$ are calculated. This process is repeated $B = 100$ times.

Fig. 1 shows the asymptotic behavior of the minimum eigenvalue of the matrix $\mathbf{X}_k^\top \mathbf{X}_k$. We see that when the sample size tends to infinity, the minimum eigenvalue also tends to infinity. Meanwhile, as is necessary for the Theorem 3, the graph is higher than $\sqrt{k}$.

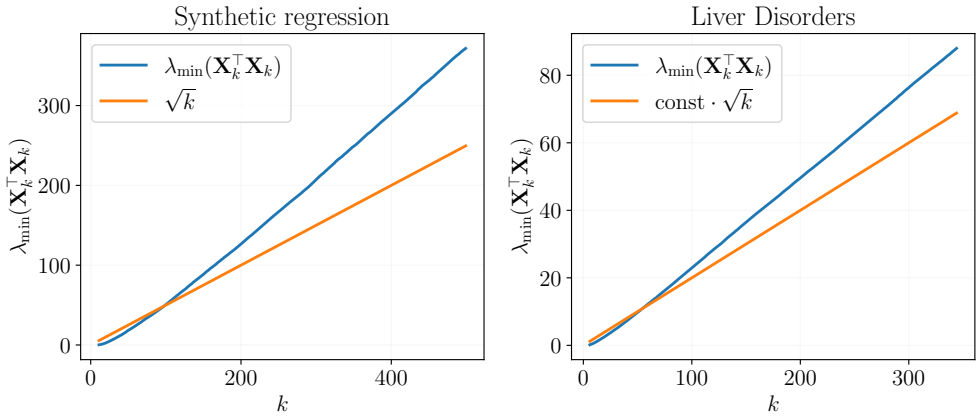

Fig. 1 Minimum eigenvalue vs available sample size

In the Fig. 2, we can observe the obtained dependencies between the available sample size k and the proposed functions $KL(k)$ and $S(k)$ for the synthetic regression dataset. At the same time, in the Fig. 3, we see the same plots for the Liver Disorders dataset. It can be seen that in both cases, the value of $KL(k)$ approaches zero as the sample size increases, and $S(k)$ tends towards one. These empirical results confirm the theoretical ones obtained earlier.

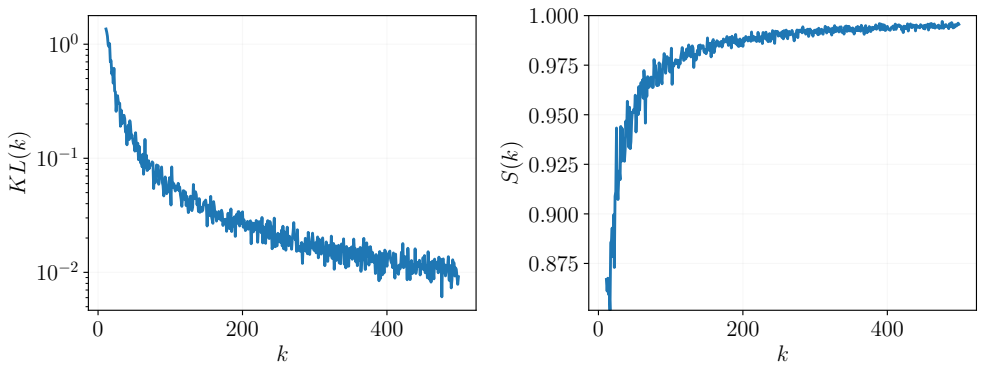

Fig. 2 Synthetic regression dataset

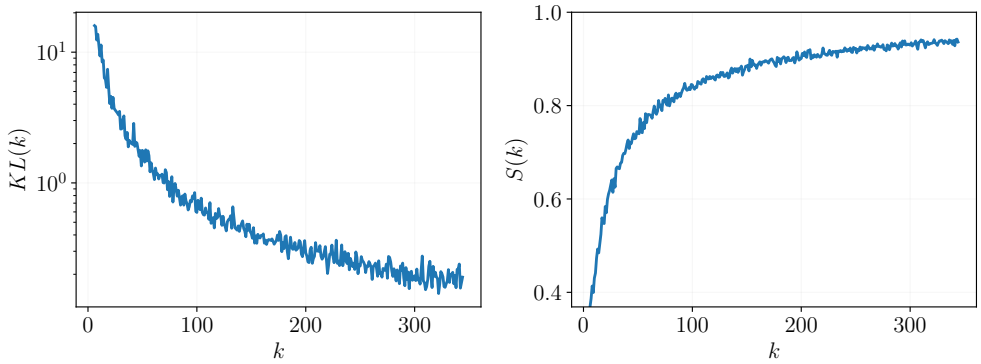

Fig. 3 Liver Disorders dataset

For the definitions of KL-sufficiency and S-sufficiency, there is a hyperparameter ε, which corresponds to the threshold for a sufficient sample size m^*. In order to study the dependence between them, we introduce Fig. 4, which shows what sample sizes can be chosen to provide a certain level of confidence.

4.2 Sample size estimation for various datasets

To compare our proposed methods with baselines, we used the following experiment setup. The machine learning model is linear regression. We have chosen 4 open-source datasets with regression task: Boston, Diabetes, Forestfires, and Servo. Their descriprive statistics are provided in the Table 1. We have applied 9 different baseline methods of sample size estimation on them: Lagrange Multipliers Test, Likelihood Ratio Test, Wald Test, Cross Validation, Bootstrap, Average Posterior Variance Criterion (APVC), Average Coverage Criterion (ACC), Average Length Criterion (ALC), and Utility function. Default parameters values were used for this purpose. All these

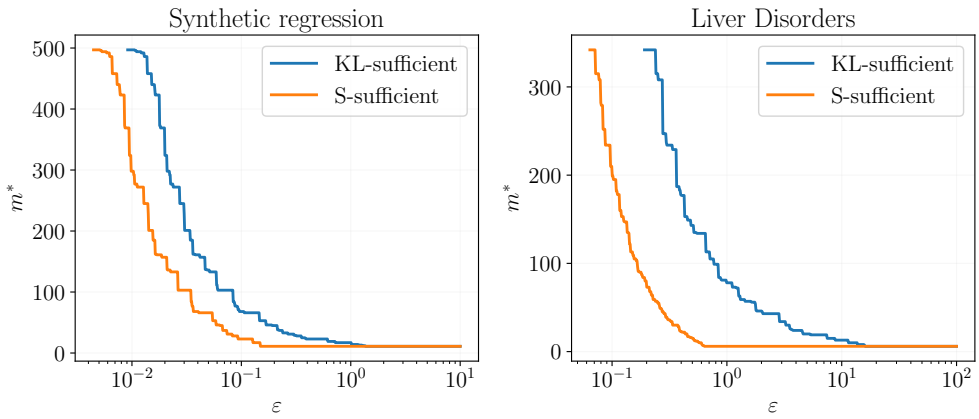

Fig. 4 Sufficient sample size vs threshold

methods were utilized with the help of SampleSizeLib. As for our methods, we slightly changed the definitions of sufficiency, transferring them to terms of relative change. Namely, we consider the sample size to be sufficient if the $KL(k)$ function has a relative deviation from its value in the entire sample of no more than ε. Similarly with the $S(k)$ function. We fixed $\varepsilon = 0.05$ and got the resulting sample sizes. This value was chosen, because other methods, especially statistical, use 0.05 as a value of type I error.

Table 1 Descriptive statistics of the sample sets

Sample set	Number of features, n	Size of sample set, m
Boston Housing	14	506
Diabetes	20	576
Forest Fires	13	517
Servo	4	167

The results in Table 2 indicate that the KL-divergence criterion is more conservative, requiring a larger sample size, whereas the S-sufficiency criterion suggests that a minimal sample size is sufficient. We believe this is a typical outcome for the s-score similarity function, which was designed for comparing different machine learning models, particularly in cases with uninformative distributions. If the distributions have high variance, the proximity function approaches one, leading the criterion to consider even a small sample size as sufficient.

4.3 Dependence of sufficient sample size on available datasets

This part includes the most comprehensive analysis of the various sample size determination methods. We analyse how the sufficient sample size depends on the available sample set. Particularly, we increase the sample size, and calculate the sufficient one,

Table 2 Experimental estimation of sample size for various sample sets

Methods and sample sets	Boston	Diabetes	Forest Fires	Servo
Lagrange Multipliers Test	18	25	44	38
Likelihood Ratio Test	17	25	43	18
Wald Test	66	51	46	76
Cross Validation	178	441	171	120
Bootstrap	113	117	86	60
APVC	98	167	351	20
ACC	228	441	346	65
ALC	98	267	516	25
Utility function	148	172	206	105
KL (ours)	493	437	86	165
S (ours)	28	22	26	10

based on the different methods. Thus, we get the Figure 5, which allows us to compare the above methods in terms of their conservatism.

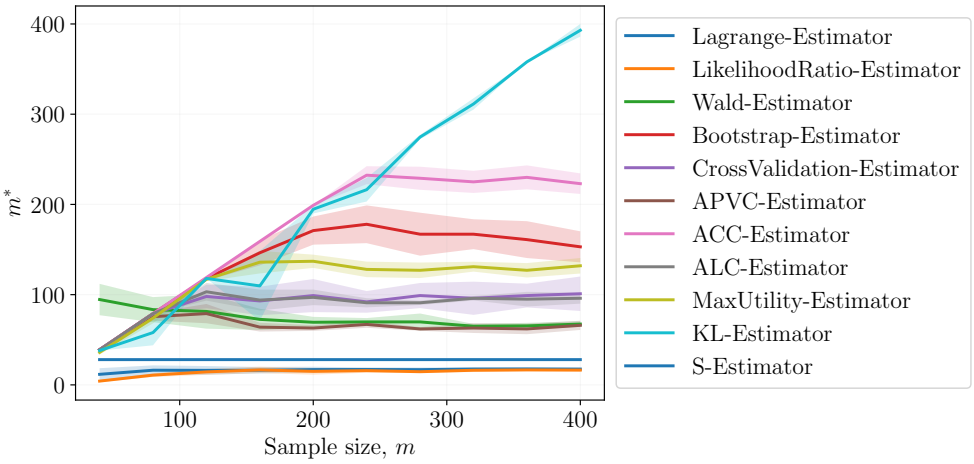

Fig. 5 Estimated sample size m^* versus the available sample size m for each method

One can see that S-sufficient sample size is often the minimum one. We have already discussed the reason of it in the previous subsection. Also, the KL-sufficient sample size tends to require an almost total sample. In our opinion, this is due to the fact that the Kullback-Leibler divergence is extremely sensitive to changes in the mean and variance of the distributions being compared. Thus, the stabilization of the distance between them occurs quite late.

5 Discussion

In this paper, we have assumed that the posterior distribution of model parameters is gaussian. This assumption allowed us to significantly simplify the theoretical calculations. Namely, we were able to write down explicit formulas for the Kullback-Leibler divergence and the s-score similarity function. This can be attributed to one of the limitations of this work. Nevertheless, we believe that getting rid of this condition significantly complicates the theoretical analysis. This is most significant for the s-score similarity function, since an analytical expression without integration is available for it only in the case of the distance between normal distributions. We hope that our results can become the basis for further research in this direction. We will also try to reveal this assumption about the non-normal posterior distribution of parameters in our future work.

In the experiments section, we used proposed methods in the relative way, calculating the relative deviation from the value on the entire dataset. Thus, we suggest doing the same in practice. Note that the choice of the threshold remains with the researcher. However, we recommend choosing small values, for example, 0.05 or 0.1.

The potential challenge consists only in calculating the Kullback-Leibler divergence and the s-score similarity function. Both formulas for the normal distribution involve inversion of the covariance matrix. The greater the number of features in the dataset, the more critical this calculation will be. Thus, we do not recommend using these approaches for very large models, especially neural network models. We plan to develop research in this direction in the future.

Regarding the comparison of KL and S-sufficiency, we discussed this issue in detail in the experiments section. Namely, if there is a desire to get a more conservative estimate, then we recommend using KL-divergence. If a more optimistic estimate is required, then s-score similarity function.

6 Conclusion

Approaches to determining a sufficient sample size based on the proximity of posterior distributions of model parameters on similar subsamples are proposed. The correctness of the proposed approaches is proved under certain restrictions on the model used. The theorem on the moments of the limit posterior distribution of parameters in a linear regression model is proved. The conducted computational experiment makes it possible to analyze the properties of the proposed methods and their effectiveness.

Appendix A Proof of Theorem 1

Proof. The Kullback-Leibler divergence for a pair of normal posterior distributions has the form

$$D_{\mathrm{KL}}\left(p_k \| p_{k+1}\right) = \frac{1}{2}\left(\operatorname{tr}\left(\boldsymbol{\Sigma}_{k+1}^{-1}\boldsymbol{\Sigma}_k\right) + (\mathbf{m}_{k+1}-\mathbf{m}_k)^{\intercal}\boldsymbol{\Sigma}_{k+1}^{-1}(\mathbf{m}_{k+1}-\mathbf{m}_k) - n + \log\left(\frac{\det\boldsymbol{\Sigma}_{k+1}}{\det\boldsymbol{\Sigma}_k}\right)\right).$$

Let's express $\boldsymbol{\Sigma}_{k+1}$ as $\boldsymbol{\Sigma}_{k+1} = \boldsymbol{\Sigma}_k + \Delta\boldsymbol{\Sigma}$. Let's consider each term separately.

$$\operatorname{tr}\left(\boldsymbol{\Sigma}_{k+1}^{-1}\boldsymbol{\Sigma}_k\right) = \operatorname{tr}\left(\left(\boldsymbol{\Sigma}_k + \Delta\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\Sigma}_k\right) \to \operatorname{tr}\mathbf{I}_n = n \text{ as } \|\Delta\boldsymbol{\Sigma}\|_F \to 0,$$

$$\left|(\mathbf{m}_{k+1}-\mathbf{m}_k)^{\intercal}\boldsymbol{\Sigma}_{k+1}^{-1}(\mathbf{m}_{k+1}-\mathbf{m}_k)\right| \leqslant \|\mathbf{m}_{k+1}-\mathbf{m}_k\|_2^2\|\boldsymbol{\Sigma}_{k+1}^{-1}\|_2 \to 0 \text{ as } \|\mathbf{m}_{k+1}-\mathbf{m}_k\|_2 \to 0,$$

$$\log\left(\frac{\det\boldsymbol{\Sigma}_{k+1}}{\det\boldsymbol{\Sigma}_k}\right) = \log\left(\frac{\det\left(\boldsymbol{\Sigma}_k+\Delta\boldsymbol{\Sigma}\right)}{\det\boldsymbol{\Sigma}_k}\right) \to \log\det\mathbf{I}_n = \log 1 = 0 \text{ as } \|\Delta\boldsymbol{\Sigma}\|_F \to 0,$$

from where we have the required. □

Appendix B Proof of Theorem 2

Proof. Let's use the s-score expression for a pair of normal posterior distributions from [20]:

$$\text{s-score}(p_k, p_{k+1}) = \exp\left(-\frac{1}{2}(\mathbf{m}_{k+1}-\mathbf{m}_k)^{\intercal}\left(\boldsymbol{\Sigma}_k+\boldsymbol{\Sigma}_{k+1}\right)^{-1}(\mathbf{m}_{k+1}-\mathbf{m}_k)\right).$$

Because

$$\left|(\mathbf{m}_{k+1}-\mathbf{m}_k)^{\intercal}\left(\boldsymbol{\Sigma}_k+\boldsymbol{\Sigma}_{k+1}\right)^{-1}(\mathbf{m}_{k+1}-\mathbf{m}_k)\right| \leqslant \|\mathbf{m}_{k+1}-\mathbf{m}_k\|_2^2\|\left(\boldsymbol{\Sigma}_k+\boldsymbol{\Sigma}_{k+1}\right)^{-1}\|_2 \to 0$$

if $\|\mathbf{m}_{k+1}-\mathbf{m}_k\|_2 \to 0$, then the value of the quadratic form inside the exponent tends to zero. Therefore, s-score$(p_k, p_{k+1}) \to 1$ as $\|\mathbf{m}_{k+1}-\mathbf{m}_k\|_2 \to 0$. □

Appendix C Proof of Theorem 3

Proof. Let be a normal prior distribution of parameters $p(\mathbf{w}) = \mathcal{N}\left(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}\right)$. In a linear regression model, likelihood is normal, namely

$$p(\mathbf{y}|\mathbf{X},\mathbf{w}) = \mathcal{N}\left(\mathbf{y}|\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I}\right) = \left(2\pi\sigma^2\right)^{-m/2}\exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|_2^2\right).$$

Using the conjugacy of the prior distribution and likelihood, it is easy to find the parameters of the posterior distribution:

$$p(\mathbf{w}|\mathbf{X},\mathbf{y}) = \mathcal{N}\left(\mathbf{w}|\mathbf{m},\boldsymbol{\Sigma}\right),$$

where

$$\boldsymbol{\Sigma} = \left(\alpha\mathbf{I} + \frac{1}{\sigma^2}\mathbf{X}^\intercal\mathbf{X}\right)^{-1}, \qquad \mathbf{m} = \left(\mathbf{X}^\intercal\mathbf{X} + \alpha\sigma^2\mathbf{I}\right)^{-1}\mathbf{X}^\intercal\mathbf{y}.$$

Consider the expression $\|\boldsymbol{\Sigma}_{k+1} - \boldsymbol{\Sigma}_k\|_2$ norms of difference of covariance matrices for subsamples of size k and $k+1$. Let's introduce the notation $\mathbf{A}_k = \frac{1}{\sigma^2}\mathbf{X}_k^\intercal\mathbf{X}_k$. Given the formulas above, we have

$$\|\boldsymbol{\Sigma}_{k+1} - \boldsymbol{\Sigma}_k\|_2 = \left\|(\alpha\mathbf{I} + \mathbf{A}_{k+1})^{-1} - (\alpha\mathbf{I} + \mathbf{A}_k)^{-1}\right\|_2 =$$

$$= \left\|(\alpha\mathbf{I} + \mathbf{A}_{k+1})^{-1}(\mathbf{A}_{k+1} - \mathbf{A}_k)(\alpha\mathbf{I} + \mathbf{A}_k)^{-1}\right\|_2 \leqslant$$

Let's use the submultiplicativity of the spectral matrix norm.

$$\leqslant \left\|(\alpha\mathbf{I} + \mathbf{A}_{k+1})^{-1}\right\|_2 \left\|(\alpha\mathbf{I} + \mathbf{A}_k)^{-1}\right\|_2 \|\mathbf{A}_{k+1} - \mathbf{A}_k\|_2 =$$

Now let's use the expression of the spectral matrix norm in terms of the maximum eigenvalue.

$$= \frac{1}{\lambda_{\min}(\alpha\mathbf{I} + \mathbf{A}_{k+1})}\frac{1}{\lambda_{\min}(\alpha\mathbf{I} + \mathbf{A}_k)}\|\mathbf{A}_{k+1} - \mathbf{A}_k\|_2 \leqslant$$

$$\leqslant \frac{1}{\lambda_{\min}(\mathbf{A}_{k+1})}\frac{1}{\lambda_{\min}(\mathbf{A}_k)}\|\mathbf{A}_{k+1} - \mathbf{A}_k\|_2 =$$

$$= \sigma^2\frac{1}{\lambda_{\min}\left(\mathbf{X}_{k+1}^\intercal\mathbf{X}_{k+1}\right)}\frac{1}{\lambda_{\min}(\mathbf{X}_k^\intercal\mathbf{X}_k)}\left\|\mathbf{X}_{k+1}^\intercal\mathbf{X}_{k+1} - \mathbf{X}_k^\intercal\mathbf{X}_k\right\|_2.$$

Further, since by the condition $\|\mathbf{x}\|_2 \leqslant M$, then

$$\left\|\mathbf{X}_{k+1}^\intercal\mathbf{X}_{k+1} - \mathbf{X}_k^\intercal\mathbf{X}_k\right\|_2 = \left\|\sum_{i=1}^{k+1}\mathbf{x}_i\mathbf{x}_i^\intercal - \sum_{i=1}^{k}\mathbf{x}_i\mathbf{x}_i^\intercal\right\|_2 = \left\|\mathbf{x}_{k+1}\mathbf{x}_{k+1}^\intercal\right\|_2 = \lambda_{\max}\left(\mathbf{x}_{k+1}\mathbf{x}_{k+1}^\intercal\right) =$$

A matrix of unit rank has a single nonzero eigenvalue.

$$= \mathbf{x}_{k+1}^\intercal\mathbf{x}_{k+1} = \|\mathbf{x}_{k+1}\|_2^2 \leqslant M^2.$$

By condition $\lambda_{\min}(\mathbf{X}_k^\intercal\mathbf{X}_k) = \omega(\sqrt{k})$, then $\|\boldsymbol{\Sigma}_{k+1} - \boldsymbol{\Sigma}_k\|_2 = o(k^{-1})$ as $k \to \infty$. Next, we will use the equivalence of matrix norms, namely

$$\|\boldsymbol{\Sigma}_{k+1} - \boldsymbol{\Sigma}_k\|_F \leqslant \sqrt{k}\|\boldsymbol{\Sigma}_{k+1} - \boldsymbol{\Sigma}_k\|_2 = o(k^{-1/2}) \text{ as } k \to \infty,$$

which was exactly what needed to be proved. Now let's estimate the norm of the difference in mathematical expectations.

$$\|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2 = \left\|\left(\mathbf{X}_{k+1}^\intercal\mathbf{X}_{k+1} + \alpha\sigma^2\mathbf{I}\right)^{-1}\mathbf{X}_{k+1}^\intercal\mathbf{y}_{k+1} - \left(\mathbf{X}_k^\intercal\mathbf{X}_k + \alpha\sigma^2\mathbf{I}\right)^{-1}\mathbf{X}_k^\intercal\mathbf{y}_k\right\|_2 =$$

Consider that $\mathbf{X}_{k+1}^\intercal = [\mathbf{X}_k^\intercal, \mathbf{x}_{k+1}]$ and $\mathbf{y}_{k+1} = [\mathbf{y}_k, y_{k+1}]^\intercal$, then $\mathbf{X}_{k+1}^\intercal \mathbf{X}_{k+1} = \mathbf{X}_k^\intercal \mathbf{X}_k + \mathbf{x}_{k+1}\mathbf{x}_{k+1}^\intercal$ and $\mathbf{X}_{k+1}^\intercal \mathbf{y}_{k+1} = \mathbf{X}_k^\intercal \mathbf{y}_k + \mathbf{x}_{k+1} y_{k+1}$.

$$= \left\| \left(\mathbf{X}_k^\intercal \mathbf{X}_k + \alpha\sigma^2 \mathbf{I} + \mathbf{x}_{k+1}\mathbf{x}_{k+1}^\intercal\right)^{-1} \left(\mathbf{X}_k^\intercal \mathbf{y}_k + \mathbf{x}_{k+1} y_{k+1}\right) - \left(\mathbf{X}_k^\intercal \mathbf{X}_k + \alpha\sigma^2 \mathbf{I}\right)^{-1} \mathbf{X}_k^\intercal \mathbf{y}_k \right\|_2 =$$

Let's take out the multiplier in the first term:

$$\left(\mathbf{X}_k^\intercal \mathbf{X}_k + \alpha\sigma^2 \mathbf{I} + \mathbf{x}_{k+1}\mathbf{x}_{k+1}^\intercal\right)^{-1} = \left(\mathbf{I} + \left(\mathbf{X}_k^\intercal \mathbf{X}_k + \alpha\sigma^2 \mathbf{I}\right)^{-1} \mathbf{x}_{k+1}\mathbf{x}_{k+1}^\intercal\right)^{-1} \left(\mathbf{X}_k^\intercal \mathbf{X}_k + \alpha\sigma^2 \mathbf{I}\right)^{-1}.$$

Next, we will take out the common multiplier for both terms.

$$= \left\| \left[\left(\mathbf{I} + \left(\mathbf{X}_k^\intercal \mathbf{X}_k + \alpha\sigma^2 \mathbf{I}\right)^{-1} \mathbf{x}_{k+1}\mathbf{x}_{k+1}^\intercal\right)^{-1} - \mathbf{I}\right] \left(\mathbf{X}_k^\intercal \mathbf{X}_k + \alpha\sigma^2 \mathbf{I}\right)^{-1} \mathbf{X}_k^\intercal \mathbf{y}_k + \right.$$
$$\left. + \left(\mathbf{X}_{k+1}^\intercal \mathbf{X}_{k+1} + \alpha\sigma^2 \mathbf{I}\right)^{-1} \mathbf{x}_{k+1} y_{k+1} \right\|_2 =$$

Let's use the triangle inequality, as well as the consistency and submultiplicativity property of the spectral norm.

$$\leqslant \left\| \left(\mathbf{I} + \left(\mathbf{X}_k^\intercal \mathbf{X}_k + \alpha\sigma^2 \mathbf{I}\right)^{-1} \mathbf{x}_{k+1}\mathbf{x}_{k+1}^\intercal\right)^{-1} - \mathbf{I} \right\|_2 \left\| \left(\mathbf{X}_k^\intercal \mathbf{X}_k + \alpha\sigma^2 \mathbf{I}\right)^{-1} \right\|_2 \left\| \mathbf{X}_k^\intercal \mathbf{y}_k \right\|_2 +$$
$$+ \left\| \left(\mathbf{X}_{k+1}^\intercal \mathbf{X}_{k+1} + \alpha\sigma^2 \mathbf{I}\right)^{-1} \right\|_2 \left\| \mathbf{x}_{k+1} y_{k+1} \right\|_2$$

Let's evaluate each term separately. In the first multiplier of the first term, we apply the formula for the difference of inverse matrices, as we did with covariance matrices.

$$\left\| \left(\mathbf{I} + \left(\mathbf{X}_k^\intercal \mathbf{X}_k + \alpha\sigma^2 \mathbf{I}\right)^{-1} \mathbf{x}_{k+1}\mathbf{x}_{k+1}^\intercal\right)^{-1} - \mathbf{I} \right\|_2 \leqslant$$

$$\leqslant \left\| \left(\mathbf{I} + \left(\mathbf{X}_k^\intercal \mathbf{X}_k + \alpha\sigma^2 \mathbf{I}\right)^{-1} \mathbf{x}_{k+1}\mathbf{x}_{k+1}^\intercal\right)^{-1} \right\|_2 \cdot \|\mathbf{I}\|_2 \cdot \left\| \left(\mathbf{X}_k^\intercal \mathbf{X}_k + \alpha\sigma^2 \mathbf{I}\right)^{-1} \mathbf{x}_{k+1}\mathbf{x}_{k+1}^\intercal \right\|_2 \leqslant$$

Again, we use submultiplicativity, as well as an expression for the norm of a matrix of unit rank.

$$\leqslant \frac{1}{\lambda_{\min}\left(\mathbf{I} + \left(\mathbf{X}_k^\intercal \mathbf{X}_k + \alpha\sigma^2 \mathbf{I}\right)^{-1} \mathbf{x}_{k+1}\mathbf{x}_{k+1}^\intercal\right)} \frac{\|\mathbf{x}_{k+1}\|_2^2}{\lambda_{\min}\left(\mathbf{X}_k^\intercal \mathbf{X}_k + \alpha\sigma^2 \mathbf{I}\right)} \leqslant$$

$$\leqslant \frac{1}{1 + \lambda_{\min}\left(\left(\mathbf{X}_k^\intercal \mathbf{X}_k + \alpha\sigma^2 \mathbf{I}\right)^{-1} \mathbf{x}_{k+1}\mathbf{x}_{k+1}^\intercal\right)} \frac{M^2}{\lambda_{\min}\left(\mathbf{X}_k^\intercal \mathbf{X}_k\right)} \leqslant$$

The minimum eigenvalue of the product of matrices is estimated by the product of their minimum eigenvalues. In addition, the minimum eigenvalue of the matrix of unit rank $\mathbf{x}_{k+1}\mathbf{x}_{k+1}^{\intercal}$ is zero.

$$\leqslant \frac{1}{1+\lambda_{\max}\left(\mathbf{X}_k^{\intercal}\mathbf{X}_k+\alpha\sigma^2\mathbf{I}\right)\lambda_{\min}\left(\mathbf{x}_{k+1}\mathbf{x}_{k+1}^{\intercal}\right)}\frac{M^2}{\lambda_{\min}\left(\mathbf{X}_k^{\intercal}\mathbf{X}_k\right)}=\frac{M^2}{\lambda_{\min}\left(\mathbf{X}_k^{\intercal}\mathbf{X}_k\right)}.$$

The second and third multipliers of the first term are evaluated as follows.

$$\left\|\left(\mathbf{X}_k^{\intercal}\mathbf{X}_k+\alpha\sigma^2\mathbf{I}\right)^{-1}\right\|_2\left\|\mathbf{X}_k^{\intercal}\mathbf{y}_k\right\|_2\leqslant\frac{\left\|\mathbf{X}_k^{\intercal}\mathbf{y}_k\right\|_2}{\lambda_{\min}\left(\mathbf{X}_k^{\intercal}\mathbf{X}_k\right)}=\frac{\left\|\sum\limits_{i=1}^{k}\mathbf{x}_i y_i\right\|_2}{\lambda_{\min}\left(\mathbf{X}_k^{\intercal}\mathbf{X}_k\right)}\leqslant\frac{kM^2}{\lambda_{\min}\left(\mathbf{X}_k^{\intercal}\mathbf{X}_k\right)}$$

Finally, let's evaluate the second term.

$$\left\|\left(\mathbf{X}_{k+1}^{\intercal}\mathbf{X}_{k+1}+\alpha\sigma^2\mathbf{I}\right)^{-1}\right\|_2\left\|\mathbf{x}_{k+1}y_{k+1}\right\|_2\leqslant\frac{M^2}{\lambda_{\min}\left(\mathbf{X}_{k+1}^{\intercal}\mathbf{X}_{k+1}\right)}$$

In total, we have the following estimate.

$$\|\mathbf{m}_{k+1}-\mathbf{m}_k\|_2\leqslant\frac{kM^3}{\lambda_{\min}^2\left(\mathbf{X}_k^{\intercal}\mathbf{X}_k\right)}+\frac{M^2}{\lambda_{\min}\left(\mathbf{X}_{k+1}^{\intercal}\mathbf{X}_{k+1}\right)}=k\cdot o(k^{-1})+o(k^{-1/2})=o(1)\text{ as }k\to\infty$$

Thus, we obtained the required convergence. □

References

[1] Bies, R.R., Muldoon, M.F., Pollock, B.G., Manuck, S., Smith, G., Sale, M.E.: A genetic algorithm-based, hybrid machine learning approach to model selection. Journal of pharmacokinetics and pharmacodynamics **33**(2), 195 (2006)

[2] Cawley, G.C., Talbot, N.L.: On over-fitting in model selection and subsequent selection bias in performance evaluation. The Journal of Machine Learning Research **11**, 2079–2107 (2010)

[3] Raschka, S.: Model evaluation, model selection, and algorithm selection in machine learning. arXiv preprint arXiv:1811.12808 (2018)

[4] Byrd, R.H., Chin, G.M., Nocedal, J., Wu, Y.: Sample size selection in optimization methods for machine learning. Mathematical programming **134**(1), 127–155 (2012)

[5] Figueroa, R.L., Zeng-Treitler, Q., Kandula, S., Ngo, L.H.: Predicting sample size required for classification performance. BMC medical informatics and decision making **12**, 1–10 (2012)

[6] Balki, I., Amirabadi, A., Levman, J., Martel, A.L., Emersic, Z., Meden, B., Garcia-Pedrero, A., Ramirez, S.C., Kong, D., Moody, A.R., *et al.*: Sample-size determination methodologies for machine learning in medical imaging research: a systematic review. Canadian Association of Radiologists Journal **70**(4), 344–353 (2019)

[7] Adcock, C.J.: A bayesian approach to calculating sample sizes. The Statistician **37**(4/5), 433 (1988) https://doi.org/10.2307/2348770

[8] Joseph, L., Wolfson, D.B., Berger, R.D.: Sample size calculations for binomial proportions via highest posterior density intervals. Journal of the Royal Statistical Society. Series D (The Statistician) **44**(2), 143–154 (1995). Accessed 2023-12-12

[9] Self, S.G., Mauritsen, R.H.: Power/sample size calculations for generalized linear models. Biometrics, 79–86 (1988)

[10] Shieh, G.: On power and sample size calculations for likelihood ratio tests in generalized linear models. Biometrics **56**(4), 1192–1196 (2000)

[11] Shieh, G.: On power and sample size calculations for wald tests in generalized linear models. Journal of Statistical Planning and Inference **128**(1), 43–59 (2005)

[12] Lindley, D.V.: The choice of sample size. Journal of the Royal Statistical Society: Series D (The Statistician) **46**(2), 129–138 (1997) https://doi.org/10.1111/1467-9884.00068

[13] Pham-Gia, T.: On bayesian analysis, bayesian decision theory and the sample size problem. Journal of the Royal Statistical Society: Series D (The Statistician) **46**(2), 139–144 (1997) https://doi.org/10.1111/1467-9884.00069

[14] Gelfand, A.E., Wang, F.: A simulation-based approach to bayesian sample size determination for performance under a given model and for separating models. Statistical Science **17**(2) (2002) https://doi.org/10.1214/ss/1030550861

[15] Cao, J., Lee, J.J., Alber, S.: Comparison of bayesian sample size criteria: Acc, alc, and woc. Journal of Statistical Planning and Inference **139**(12), 4111–4122 (2009) https://doi.org/10.1016/j.jspi.2009.05.041

[16] Brutti, P., De Santis, F., Gubbiotti, S.: Bayesian-frequentist sample size determination: a game of two priors. METRON **72**(2), 133–151 (2014) https://doi.org/10.1007/s40300-014-0043-2

[17] Pezeshk, H., Nematollahi, N., Maroufy, V., Gittins, J.: The choice of sample size: a mixed bayesian / frequentist approach. Statistical Methods in Medical Research **18**(2), 183–194 (2008) https://doi.org/10.1177/0962280208089298

[18] Grabovoy, A.V., Gadaev, T.T., Motrenko, A.P., Strijov, V.V.: Numerical

methods of sufficient sample size estimation for generalised linear models. Lobachevskii Journal of Mathematics **43**(9), 2453–2462 (2022) https://doi.org/10.1134/s1995080222120125

[19] Motrenko, A., Strijov, V., Weber, G.-W.: Sample size determination for logistic regression. Journal of Computational and Applied Mathematics **255**, 743–752 (2014) https://doi.org/10.1016/j.cam.2013.06.031

[20] Aduenko, A.: Selection of multimodels in classification tasks. PhD thesis, MIPT (2017). https://www.frccsc.ru/diss-council/00207305/diss/list/aduenko_aa