Sample Size Determination: Likelihood Bootstrapping

Nikita Kiselev[1] and Andrey Grabovoy[2]

Moscow Institute of Physics and Technology, Moscow, Russia;
[1]kiselev.ns@phystech.edu;
[2]grabovoy.av@phystech.edu.

Abstract. The problem of determining an appropriate sample size is essential for constructing an efficient machine learning model. However, current techniques are either not rigorously proven or are specific to a particular statistical hypothesis regarding the distribution of model parameters. In this paper we propose two methods based on the likelihood values on resampled subsets. We demonstrate the validity of one of these methods in a linear regression model. Computational experiments show the convergence of the proposed functions as the sample size increases.

Keywords: Sufficient sample size · Likelihood bootstrapping · Linear regression

1 Introduction

The process of supervised machine learning entails the selection of a predictive model from a family of parametric models, a decision often based on specific statistical assumptions, such as optimizing a particular quality function. A model that aligns with these statistical assumptions is termed as a *adequate* model [3,7].

In the planning phase of a computational experiment, it is crucial to estimate the minimum required sample size, which refers to the number of objects necessary to construct a suitable model. The sample size needed to develop an adequate predictive model is called *sufficient* [5,8,2].

This study focuses on the determination of a sufficient sample size, a topic that has been extensively researched with methods categorized into statistical, Bayesian, and heuristic approaches.

Early investigations on this subject, such as [1,12], establish 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 [18], the Likelihood ratio test [19], the Wald statistic [20]. Statistical methods have certain limitations associated with their practical application. They enable the estimation of the sample size based on assumptions about data distribution and information regarding the consistency of observed values with the assumptions of the null hypothesis.

The Bayesian method is another approach to this problem. In the study [13], the sufficient sample size is ascertained by maximizing the expected utility function, which may explicitly incorporate parameter distribution functions and penalties for increasing the sample size. This study also explores alternative methods based on restricting a certain quality criterion for model parameter estimation. Notable among these criteria are the Average Posterior Variance Criterion (APVC), Average Coverage Criterion (ACC), Average Length Criterion (ALC), and Effective Sample Size Criterion (ESC). These criteria have been further refined in subsequent research, such as [17] and [9]. Eventually, the authors of [6] conducted a theoretical and practical comparison of methods from [1,12,13].

Researchers like [4] and [16] delve into the distinctions between Bayesian and frequentist approaches in determining sample size, proposing robust methods for the Bayesian approach and providing illustrative examples for certain probabilistic models.

The paper [10] examines various methods for sample size estimation in generalized linear models, encompassing statistical, heuristic, and Bayesian methods. Techniques 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 highlight the potential of combining Bayesian and statistical approaches to estimate sample size when available sample sizes are insufficient.

In [15], a method for determining sample size in logistic regression is presented, utilizing cross-validation and Kullback-Leibler divergence between posterior distributions of model parameters on similar subsamples. Similar subsamples refer to those that can be derived from each other by adding, removing, or replacing one object.

This paper discusses several approaches to determining a sufficient sample size. It is proposed to estimate the mathematical expectation and variance of the likelihood function on bootstrapped subsamples. A small change in these values when adding another object indicates that a sufficient number of objects in the sample has been reached. The correctness of the definition in the linear regression model is proved. The presented method is easy to use in practice. To do this, it is proposed to calculate the value of the loss function instead of the likelihood.

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 likelihood function of a sample $\mathfrak{D}_m = \{(\mathbf{x}_i, y_i)\}, i \in \mathcal{I} = \{1, \ldots, m\}$ of size m, where $\mathbf{x}_1, \ldots, \mathbf{x}_m$ are i.i.d. together, is defined as

$$L(\mathfrak{D}_m, \mathbf{w}) = p(\mathbf{y}_m|\mathbf{X}_m, \mathbf{w}) = \prod_{i=1}^{m} p(y_i|\mathbf{x}_i, \mathbf{w}).$$

Its logarithm

$$l(\mathfrak{D}_m, \mathbf{w}) = \sum_{i=1}^{m} \log p(y_i|\mathbf{x}_i, \mathbf{w})$$

is called the logarithmic likelihood function. Unless stated otherwise, we consider samples to be i.i.d.

The maximum likelihood estimate of a set of parameters $\mathbf{w} \in \mathbb{W}$ based on the subsample $\mathfrak{D}_k$ of size k is given by

$$\hat{\mathbf{w}}_k = \arg\max_{\mathbf{w} \in \mathbb{W}} L(\mathfrak{D}_k, \mathbf{w}).$$

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 $k \geqslant m^*$.*

3 Proposed sample size determination methods

In this section, we will assume that $m^* \leqslant m$ is valid. This means that we just need to formalize which sample size can be considered sufficient. To determine sufficiency, we will use the likelihood function. When there are enough objects available, it is quite natural to expect that the resulting parameter estimate will not change much from one sample realization to another [11,12]. The same can be said about the likelihood function. Thus, we formulate which sample size can be considered sufficient.

Definition 2. *Let's fix some positive number $\varepsilon > 0$. The sample size m^* is called* ***D-sufficient*** *if for all $k \geqslant m^*$*

$$D(k) = \mathbb{D}_{\hat{\mathbf{w}}_k} L(\mathfrak{D}_m, \hat{\mathbf{w}}_k) \leqslant \varepsilon.$$

On the other hand, when there are enough objects available, it is also quite natural that when adding another object to consideration, the resulting parameter estimate will not change much. Let's formulate another definition.

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

$$M(k) = \left|\mathbb{E}_{\hat{\mathbf{w}}_{k+1}} L(\mathfrak{D}_m, \hat{\mathbf{w}}_{k+1}) - \mathbb{E}_{\hat{\mathbf{w}}_k} L(\mathfrak{D}_m, \hat{\mathbf{w}}_k)\right| \leqslant \varepsilon.$$

In the definitions above instead of the likelihood function $L(\mathfrak{D}_m, \hat{\mathbf{w}}_k)$ we can consider its logarithm $l(\mathfrak{D}_m, \hat{\mathbf{w}}_k)$.

Suppose that $\mathbb{W} = \mathbb{R}^n$. Recall that the Fisher information is called the matrix

$$[\mathcal{I}(\mathbf{w})]_{ij} = -\mathbb{E}\left[\frac{\partial^2 \log p(\mathbf{y}|\mathbf{x}, \mathbf{w})}{\partial w_i \partial w_j}\right].$$

A known result is the asymptotic normality of the maximum likelihood estimate, that is, $\sqrt{k}\,(\hat{\mathbf{w}}_k - \mathbf{w}) \xrightarrow{d} \mathcal{N}\left(0, \mathcal{I}^{-1}(\mathbf{w})\right)$. Convergence in the distribution generally does not imply convergence of the moments of a random vector. Nevertheless, if we assume the latter, then in some models it is possible to prove the correctness of our proposed definition of M-sufficient sample size.

For convenience, we denote the distribution parameters $\hat{\mathbf{w}}_k$ as follows: mathematical expectation $\mathbb{E}\hat{\mathbf{w}}_k = \mathbf{m}_k$ and the covariance matrix $\mathbb{D}\hat{\mathbf{w}}_k = \mathbf{\Sigma}_k$. Then the following theorem holds, the proof of which is given in the appendix.

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 the linear regression model, the definition of M-sufficient sample size is correct. Namely, for any $\varepsilon > 0$, there is such a m^* that for all $k \geqslant m^*$ $M(k) \leqslant \varepsilon$ is satisfied.*

Corollary 1. *Let $\|\mathbf{m}_k - \mathbf{w}\|_2 \to 0$ and $\|\mathbf{\Sigma}_k - [k\mathcal{I}(\mathbf{w})]^{-1}\|_F \to 0$ for $k \to \infty$. Then, in the linear regression model, the definition of an M-sufficient sample size is correct.*

By condition, one sample is given. Therefore, in the experiment it is not possible to calculate the mathematical expectation and variance specified in the definitions. To evaluate them, we will use the bootstrap technique. Namely, we will generate from the given $\mathfrak{D}_m$ a number of B subsamples of size k with a return. For each of them, we get an estimate of the parameters $\hat{\mathbf{w}}_k$ and calculate the value of $L(\mathfrak{D}_m, \hat{\mathbf{w}}_k)$. For the estimation, we will use a sample mean and an unbiased sample variance (for bootstrap samples).

The definitions proposed above can also be applied in those problems where an arbitrary loss function is minimized rather than the likelihood function is maximized. We do not provide any theoretical justification for this, but in practice such a heuristic turns out to be quite successful.

4 Computational experiment

This section provides an empirical study of the proposed methods. Experiments were conducted on synthetic data and dataset Liver Disorders from [14]. The code is available in open source in the GitHub repository[1].

Synthetic data is generated from linear regression and logistic regression models. The number of objects is 1000, the number of features is 20. $B = 1000$ bootstrapped subsamples are used. The values of $D(k)$ and $M(k)$ are calculated. Regression dataset Liver Disorders has 345 objects and 5 features. We also use $B = 1000$ subsamples made by bootstrapping to estimate mean and variance of the loss function.

In the Fig. 1, we can observe the obtained dependencies between the available sample size k and the proposed functions $D(k)$ and $M(k)$ for the synthetic regression dataset. Results for the synthetic classification dataset are in the Fig. 2. At the same time, in the Fig. 3, we see the same plots for the Liver Disorders dataset. It can be seen that in all cases, the values of $D(k)$ and $M(k)$ approaches zero as the sample size increases. These empirical results confirm the theoretical ones obtained earlier.

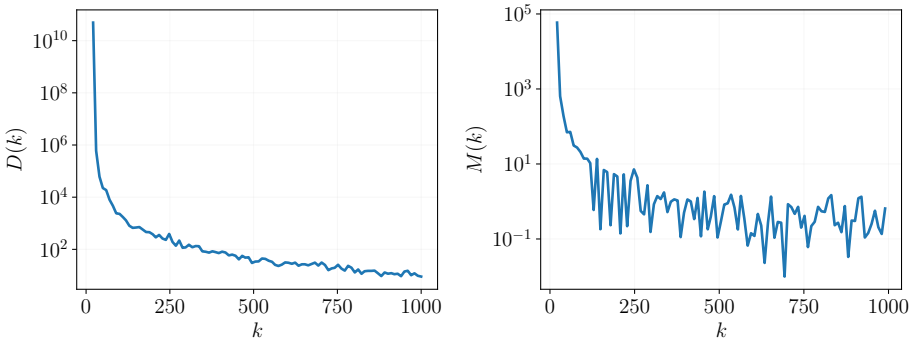

Fig. 1. Convergence of the proposed functions $D(k)$ and $M(k)$ for the synthetic regression dataset, i.e. linear regression model. Both functions tends to zero as the sample size increases.

For the definitions of D-sufficiency and M-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.

To compare the performance of the proposed methods on different datasets, samples have been chosen from the open repository [14]. The detailed information about each dataset, the number of observations and number of features, are provided in Table 1. For demonstration purposes, a value of the hyperparameter

[1] `https://github.com/kisnikser/Likelihood-Bootstrapping`

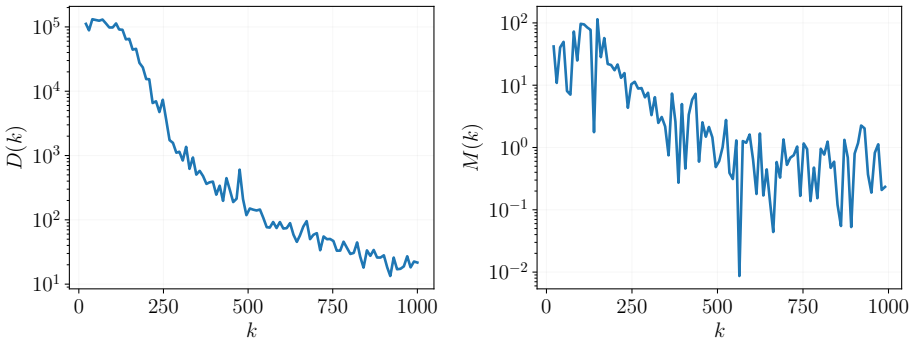

Fig. 2. Convergence of the proposed functions $D(k)$ and $M(k)$ for the synthetic classification dataset, i.e. logistic regression model. Both functions tends to zero as the sample size increases.

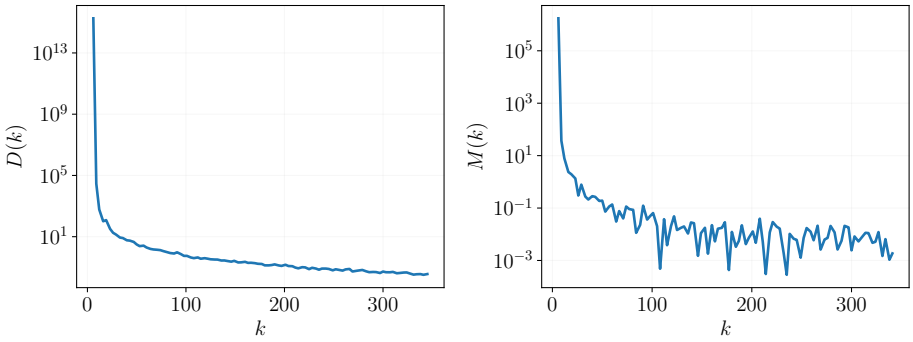

Fig. 3. Convergence of the proposed functions $D(k)$ and $M(k)$ for the Liver Disorders dataset. Both functions tends to zero as the sample size increases.

ε has been selected, at which the value of the target function, either $D(k)$ or $M(k)$, decreases by half. The corresponding results are in Table 1. Omissions mean that the initial sample size is not sufficient.

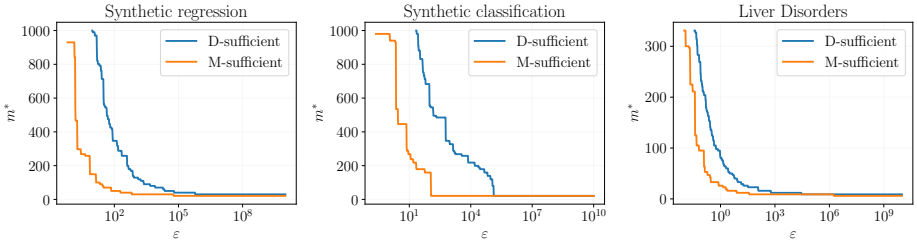

Fig. 4. Sufficient sample size versus threshold for three datasets: synthetic regression, synthetic classification, and Liver Disorders. As the threshold value ε increases, the sufficient sample size decreases. It means, that one can choose fewer objects to satisfy the desirable values of the proposed functions $D(k)$ and $M(k)$.

Table 1. Comparison of the proposed methods of sample size determination: based on $D(k)$ and $M(k)$. For each of the proposed functions the threshold value ε was chosen in such way that the initial function value decreases by half. The results were obtained for a variety of datasets with regression task. Omissions in the table mean that the initial sample size is not sufficient.

Dataset name	Objects m	Features n	D	M
Abalone	4177	8	96	96
Auto MPG	392	8	15	15
Automobile	159	25	70	156
Liver Disorders	345	6	12	19
Servo	167	4	41	—
Forest fires	517	12	208	—
Wine Quality	6497	12	144	144
Energy Efficiency	768	9	24	442
Student Performance	649	32	129	177
Facebook Metrics	495	18	31	388
Real Estate Valuation	414	7	15	23
Heart Failure Clinical Records	299	12	63	224
Bone marrow transplant: children	142	36	—	—

5 Discussion

In this paper, we have proposed two novel methods for determining a sufficient sample size based on the likelihood values on resampled subsets. The first method, referred to as D-sufficiency, relies on the variance of the likelihood function, while the second method, M-sufficiency, focuses on the difference in the expected likelihood function when adding an additional object to the sample. We have demonstrated the validity of the M-sufficient sample size definition in a linear regression model, under certain conditions on the model parameters.

The computational experiments conducted on synthetic and real-world datasets have shown that the proposed functions, $D(k)$ and $M(k)$, converge to zero as the sample size increases. The experiments also highlight the practicality of the methods, as they can be easily applied to various datasets.

The proposed methods have the potential to be applied to a wide range of models and datasets, beyond linear regression. Although we have only proved the correctness of the M-sufficient sample size definition for linear regression, the empirical results suggest that the methods may be effective for other models as well. Future work should focus on extending the theoretical analysis to other models, including probably neural networks.

6 Conclusion

This paper introduces two novel methods, D-sufficiency and M-sufficiency, for determining a sufficient sample size based on likelihood values on resampled subsets. The validity of the M-sufficient sample size definition is demonstrated in a linear regression model, and computational experiments on synthetic and real-world datasets show that the proposed functions converge to zero as the sample size increases, highlighting the practicality of the methods. Future work should focus on extending the theoretical analysis to other models, including neural networks.

References

1. Adcock, C.J.: A bayesian approach to calculating sample sizes. The Statistician **37**(4/5), 433 (1988). https://doi.org/10.2307/2348770, http://dx.doi.org/10.2307/2348770
2. 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)
3. 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)
4. Brutti, P., De Santis, F., Gubbiotti, S.: Bayesian-frequentist sample size determination: a game of two priors. METRON **72**(2), 133–151 (May 2014). https://doi.org/10.1007/s40300-014-0043-2, http://dx.doi.org/10.1007/s40300-014-0043-2
5. 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)
6. 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 (Dec 2009). https://doi.org/10.1016/j.jspi.2009.05.041, http://dx.doi.org/10.1016/j.jspi.2009.05.041
7. 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)
8. 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)
9. 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) (May 2002). https://doi.org/10.1214/ss/1030550861, http://dx.doi.org/10.1214/ss/1030550861
10. 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 (Sep 2022). https://doi.org/10.1134/s1995080222120125, http://dx.doi.org/10.1134/S1995080222120125
11. Joseph, L., Berger, R.D., Bélisle, P.: Bayesian and mixed bayesian/likelihood criteria for sample size determination. Statistics in Medicine **16**(7), 769–781 (1997). https://doi.org/https://doi.org/10.1002/(SICI)1097-0258(19970415)16:7<769::AID-SIM495>3.0.CO;2-V, https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-0258%2819970415%2916%3A7%3C769%3A%3AAID-SIM495%3E3.0.CO%3B2-V
12. 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), http://www.jstor.org/stable/2348439
13. Lindley, D.V.: The choice of sample size. Journal of the Royal Statistical Society: Series D (The Statistician) **46**(2), 129–138 (Jul 1997). https://doi.org/10.1111/1467-9884.00068, http://dx.doi.org/10.1111/1467-9884.00068
14. Markelle, K., Rachel, L., Kolby, N.: The uci machine learning repository, https://archive.ics.uci.edu

15. 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/https://doi.org/10.1016/j.cam.2013.06.031, https://www.sciencedirect.com/science/article/pii/S0377042713003294
16. 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 (Apr 2008). https://doi.org/10.1177/0962280208089298, http://dx.doi.org/10.1177/0962280208089298
17. 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 (Jul 1997). https://doi.org/10.1111/1467-9884.00069, http://dx.doi.org/10.1111/1467-9884.00069
18. Self, S.G., Mauritsen, R.H.: Power/sample size calculations for generalized linear models. Biometrics pp. 79–86 (1988)
19. Shieh, G.: On power and sample size calculations for likelihood ratio tests in generalized linear models. Biometrics **56**(4), 1192–1196 (2000)
20. 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)

A Proof of Theorem 1

Proof. Consider the definition of an M-sufficient sample size in terms of the logarithm of the likelihood function. In a linear regression model

$$L\left(\mathfrak{D}_m, \hat{\mathbf{w}}_k\right) = p(\mathbf{y}|\mathbf{X}, \hat{\mathbf{w}}_k) = \prod_{i=1}^{m} p(y_i|\mathbf{x}_i, \hat{\mathbf{w}}_k) = \prod_{i=1}^{m} \mathcal{N}\left(y_i|\hat{\mathbf{w}}_k^\top \mathbf{x}_i, \sigma^2\right) =$$

$$= \left(2\pi\sigma^2\right)^{-m/2} \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}_k\|_2^2\right).$$

Take a logarithm:

$$l\left(\mathfrak{D}_m, \hat{\mathbf{w}}_k\right) = \log p(\mathbf{y}|\mathbf{X}, \hat{\mathbf{w}}_k) = -\frac{m}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}_k\|_2^2.$$

Let's take the mathematical expectation of $\mathfrak{D}_k$, given that $\mathbb{E}_{\mathfrak{D}_k}\hat{\mathbf{w}}_k = \mathbf{m}_k$ and $\operatorname{cov}(\hat{\mathbf{w}}_k) = \boldsymbol{\Sigma}_k$:

$$\mathbb{E}_{\mathfrak{D}_k} l\left(\mathfrak{D}_m, \hat{\mathbf{w}}_k\right) = -\frac{m}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\Big(\|\mathbf{y} - \mathbf{X}\mathbf{m}_k\|_2^2 + \operatorname{tr}\left(\mathbf{X}^\top\mathbf{X}\boldsymbol{\Sigma}_k\right)\Big).$$

Let's write down an expression for the difference in mathematical expectations:

$$\mathbb{E}_{\mathfrak{D}_{k+1}} l(\mathfrak{D}_m, \hat{\mathbf{w}}_{k+1}) - \mathbb{E}_{\mathfrak{D}_k} l(\mathfrak{D}_m, \hat{\mathbf{w}}_k) =$$

$$= \frac{1}{2\sigma^2}\Big(\|\mathbf{y} - \mathbf{X}\mathbf{m}_k\|_2^2 - \|\mathbf{y} - \mathbf{X}\mathbf{m}_{k+1}\|_2^2\Big) + \frac{1}{2\sigma^2}\operatorname{tr}\Big(\mathbf{X}^\top\mathbf{X}\left(\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_{k+1}\right)\Big) =$$

$$= \frac{1}{2\sigma^2}\Big(2\mathbf{y}^\top\mathbf{X}(\mathbf{m}_{k+1} - \mathbf{m}_k) + (\mathbf{m}_k - \mathbf{m}_{k+1})^\top\mathbf{X}^\top\mathbf{X}(\mathbf{m}_k + \mathbf{m}_{k+1})\Big) +$$

$$+ \frac{1}{2\sigma^2}\operatorname{tr}\Big(\mathbf{X}^\top\mathbf{X}\left(\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_{k+1}\right)\Big).$$

The value of the function $M(k)$ is a module from the above expression. Let's apply the triangle inequality for the module, and then evaluate each term.
We estimate the first term using the Cauchy-Schwarz inequality:

$$\left|\mathbf{y}^\top\mathbf{X}(\mathbf{m}_{k+1} - \mathbf{m}_k)\right| \leqslant \|\mathbf{X}^\top\mathbf{y}\|_2\|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2.$$

The second term is estimated using the Cauchy-Schwarz inequality, the consistency property of the spectral matrix norm, as well as the limitation of the sequence of vectors $\mathbf{m}_k$, which follows from the presented convergence condition:

$$\left|(\mathbf{m}_k - \mathbf{m}_{k+1})^\top\mathbf{X}^\top\mathbf{X}(\mathbf{m}_k + \mathbf{m}_{k+1})\right| \leqslant \|\mathbf{X}(\mathbf{m}_k - \mathbf{m}_{k+1})\|_2\|\mathbf{X}(\mathbf{m}_k + \mathbf{m}_{k+1})\|_2 \leqslant$$

$$\leqslant \|\mathbf{X}\|_2^2\|\mathbf{m}_k - \mathbf{m}_{k+1}\|_2\|\mathbf{m}_k + \mathbf{m}_{k+1}\|_2 \leqslant C\|\mathbf{X}\|_2^2\|\mathbf{m}_k - \mathbf{m}_{k+1}\|_2.$$

We estimate the last term using the Holder's inequality for the Frobenius norm:

$$\left|\operatorname{tr}\Big(\mathbf{X}^\top\mathbf{X}\left(\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_{k+1}\right)\Big)\right| \leqslant \|\mathbf{X}^\top\mathbf{X}\|_F\|\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_{k+1}\|_F.$$

Finally, since $\|\mathbf{m}_k - \mathbf{m}_{k+1}\|_2 \to 0$ and $\|\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_{k+1}\|_F \to 0$ as $k \to \infty$, then $M(k) \to 0$ as $k \to \infty$, which proves the theorem.

B Proof of Corollary 1

Proof. From the convergence conditions given, it follows that $\|\mathbf{m}_k - \mathbf{m}_{k+1}\|_2 \to 0$ and $\|\mathbf{\Sigma}_k - \mathbf{\Sigma}_{k+1}\|_F \to 0$ for $k \to \infty$. The application of the 1 theorem completes the proof.