Sample Size for Maximum-Likelihood Estimates of Gaussian Model Depending on Dimensionality of Pattern Space

Josef V. Psutka[a,*], Josef Psutka[a]

[a] *Department of Cybernetics, University of West Bohemia, Pilsen, Czech Republic*

Abstract

The significant properties of the maximum likelihood (ML) estimate are consistency, normality, and efficiency. While it has been proven that these properties are valid when the sample size approaches infinity, the behavior of an ML estimator when working with small sample sizes is largely unknown. However, in real tasks, we usually do not have sufficient data to completely fulfill the conditions of an optimal ML estimate. The question arises as to what amount of data is required to be able to estimate a Gaussian model that provides sufficiently accurate likelihood estimates. This issue is addressed with respect to the number of dimensions of the pattern space.

Keywords: Maximum-likelihood estimate, likelihood function, Gaussian model, Gaussian mixture model, sample size, dimensionality, pattern space, heteroscedastic data.

1. Introduction

The maximum likelihood (ML) method is widely used in many applications for estimating the parameters of statistical models. The accuracy of the estimation directly corresponds to the amount of data utilized in the estimation process. The theory of probability and mathematical statistics indicates that

*Corresponding author

Email addresses: psutka_j@kky.zcu.cz (Josef V. Psutka), psutka@kky.zcu.cz (Josef Psutka)

the desired properties of the ML estimation, namely, consistency, normality, and efficiency, can be achieved only if the sample size N tends to infinity. This is impossible in practical tasks, where we usually only have a limited sample size available. Thus, the question arises as to how the precision of the ML estimate is related to the sample size. However, it is difficult to obtain an exact answer to this question. Users of statistical models have argued that this question cannot be properly answered because the appropriate sample size for an ML estimate is heavily influenced by the possible ill-conditionality of data, correlation of features in multidimensional samples, etc.

Most studies addressing the influence of sample size on ML estimation have focused on the accuracy of the estimated model parameters [1], [2], others have addressed behavior of the determinant of the sample covariance matrix of Gaussian distribution depending on the sample size [3], [4]. A relatively large group of statistical analysis studies addresses the estimation of eigenvalues of covariance matrices in tasks wherein the number of samples N is of the same order of magnitude as the space dimensions d. Typically, they are large numbers (i.e., space dimensions $d \gg 100$), and the goal of these tasks in particular is the principal component analysis (PCA) of the covariance matrix for the subsequent reduction of the space dimensions [5], [6].

The question we attempt to answer in this paper is: what is the indicatively sample size required to estimate the parameters of the Gaussian model that can provide previously defined accuracy values of the likelihood (or log-likelihood) function for the various dimensions of pattern space? The accuracy of the values of the log-likelihood function provided by the Gaussian model is of interest to us because they are key attributes for classification in many pattern recognition tasks. Therefore, we do not address the aspect of the accuracy of the estimation of each model parameter, but only the correspondence among the sample size, accuracy of the log-likelihood function, and the dimension of the pattern space. Our research focuses on covariance matrices with a maximum dimension $d = 100$, and in the vast majority of cases (results), it will be fulfilled that $N \gg d$.

Very few studies have focused on the abovementioned question. In [7], [8]

and in various forums on the web, many recommendations have been made, according to which the sample size (e.g., for models with covariance structure) should be at least 5 (preferably 10) per parameter. Furthermore, it has also been reported that in the case of a small number of space dimensions, the sample size for estimation should be significantly increased. For instance, a sample size of at least 100 samples forms the recommendation for estimating a Gaussian model for dimension $d = 1$ (we note that in this case only 2 parameters are estimated). Most such studies report fairly subjective experiences of researchers, and no study has presented a rigorous approach to determine the sample size needed for various dimensions, particularly for statistical modeling in pattern recognition tasks.

In our previous study [9], we addressed this question only for small dimensions of pattern space ($d \leq 10$) and we did not examine the dependence between the likelihood accuracy and the sample size for individual dimensions of pattern space. Here, we address all significant aspects of these questions (see Sections 2, 3, 4, and 5).

The results obtained in solving this task can find significance in statistical modeling, wherein Gaussian models are widely used. They can also be useful for constructing Gaussian mixture models (GMMs), which is a widely accepted technique for powerful and flexible modeling of heterogeneous data (e.g., data of a non-Gaussian nature) coming from various populations. Just remind that any continuous distribution can be approximated sufficiently well by a weighted sum of the Gaussian distribution [10]. The accuracy of a GMM approximation is closely related to the number of components in the mixture and of course to the accuracy of individual components in the GMM. The parameters of individual components of the GMM (i.e., means, covariance matrices, and weights) are usually estimated by the expectation-maximization (EM) algorithm from the training data (training samples). To determine an optimal number of components in the GMM, methods based on the information criterion [11], [12], or the cross validation technique [13] are usually applied. However, the convergence of the objective function of these methods does not reflect the accuracy (ro-

bustness) of the estimated components and therefore the accuracy of the final model. Nevertheless, such accuracy may be approximately determined from the "effective" amount of data involved in modeling individual components.

Gaussian modeling is the subject of constant research by statisticians [14], [15], [16]. It is typically used in tasks of pattern recognition [17], [18], [19], [20], [21], [22],computer vision [23], [24], [25], speech and language recognition [26], [27], [28], machine learning [29], [30]; however, it also finds use in very different areas such as medical or technical diagnostics [31], [32], [33], [34], [35], [36].

2. Determining accuracy of likelihood estimate

2.1. Mean value of Gaussian log-likelihood function

Our goal is to determine the statistical dependency of the accuracy of the likelihood function values on the data size from which parameters of this function are estimated. We primarily focus on the family of Gaussian functions in this study. For this purpose, we construct in a space of d dimensions a generator G of random numbers (vectors) $\boldsymbol{x}_i$, which generator is directed by a Gaussian model with randomly determined mean $\boldsymbol{\mu}^{\mathrm{G}}$ and covariance matrix $\boldsymbol{C}^{\mathrm{G}}$, i.e., $\boldsymbol{\Theta}^{\mathrm{G}} = (\boldsymbol{\mu}^{\mathrm{G}}, \boldsymbol{C}^{\mathrm{G}})$.

For a Gaussian source with parameters $\boldsymbol{\Theta}^{\mathrm{G}}$, we first determine the mean value of the log-likelihood function $\mathrm{E}\{\ln \mathrm{L}(\boldsymbol{\Theta}^{\mathrm{G}}|X)\}$, i.e.,

$$\mathrm{E}\{\ \ln\ L(\boldsymbol{\Theta}^{\mathrm{G}}\,|X)\} = \lim_{N\to\infty} \left\{ \frac{1}{N} \sum_{i=1}^{N} \ln\ p(\boldsymbol{x}_i \,\Big|\, \boldsymbol{\Theta}^{\mathrm{G}}) \right\}\ , \tag{1}$$

where $X = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_N\}$ are independent observations, which are randomly generated by the Gaussian distribution model with parameters $\boldsymbol{\Theta}^{\mathrm{G}}$. From information theory it is well-known that the relation (1) corresponds to the differential entropy $H(X)$ that is for the multivariate normal distribution given by [9] [37]

$$\mathrm{E}\,\{\ln L(\boldsymbol{\Theta}|X)\} = -H(X) = -\tfrac{d}{2}\ln 2\pi - \tfrac{1}{2}\ln\,\det \boldsymbol{C} - \tfrac{1}{2}d\ . \tag{2}$$

where $\boldsymbol{C}$ denotes the covariance matrix and d the number of dimensions of the pattern space.

2.2. Accuracy of log-likelihood values in dependence on sample size for Gaussian model estimation

In order to simplify writing the mean of the log-likelihood function of the Gaussian source with parameters $\boldsymbol{\Theta}^{\mathrm{G}}$, we further use the following notation in the form $L_{\mathrm{E-ln}}(\mathrm{G})$ using (1) and (2):

$$L_{\mathrm{E-ln}}(\mathrm{G}) = \mathrm{E}\left\{\ln L(\boldsymbol{\Theta}^{\mathrm{G}}|X)\right\} = 1.4189\,d - 0.5\ln\det\boldsymbol{C}^{\mathrm{G}}, \tag{3}$$

where d denotes the number of dimensions of the pattern space and $\boldsymbol{C}^{\mathrm{G}}$ the covariance matrix of the source model.

With this source G, we randomly generate a set of training data $Y = \{\boldsymbol{y}_1, ..., \boldsymbol{y}_I\}$ and a set of test data $Z = \{\boldsymbol{z}_1, ..., \boldsymbol{z}_J\}$. In general, a set of samples Y does not contain samples from the set Z and vice versa. Next, we use the set Y of the training data, and for an increasing number of samples, we gradually estimate (using the ML approach) the parameters of the Gaussian model. If a subset of training data $Y_i = \{\boldsymbol{y}_1, ..., \boldsymbol{y}_i\}$ has been used for estimating parameters, we denote the estimated parameters as $\boldsymbol{\Theta}_i = (\boldsymbol{\mu}_i, \boldsymbol{C}_i)$ (note: $i \geq d+1$ must be satisfied here). We next define the average value $L_{\mathrm{Av-ln}}(\boldsymbol{\Theta}_i|Z)$ of log-likelihoods on the Gaussian model with parameters $\boldsymbol{\Theta}_i$ and a set of test data $Z = \{\boldsymbol{z}_1, ..., \boldsymbol{z}_J\}$ as

$$L_{\mathrm{Av-ln}}(\boldsymbol{\Theta}_i\,|Z)\} = \tfrac{1}{J}\sum_{j=1}^{J} \ln\, p(\boldsymbol{z}_j|\,\boldsymbol{\Theta}_i)\,. \tag{4}$$

We determine the minimum value i^* of training data $Y_{i*} = \{\boldsymbol{y}_1, ..., \boldsymbol{y}_{i*}\}$, for which the difference between the mean value of the log-likelihood of the generator $L_{\mathrm{E-ln}}(\mathrm{G})$ (see [38]) and $L_{\mathrm{Av-ln}}(\boldsymbol{\Theta}_i|Z)$ is smaller than the established error Δ

$$i^* = \min_i\ \{\,[\,L_{\mathrm{E-ln}}(\mathrm{G}) - L_{\mathrm{Av-ln}}(\boldsymbol{\Theta}_i|Z)\,]\,\leq \Delta\}\,. \tag{5}$$

Note that if J in (4) approaching infinity, then the expression in brackets (5) refers to Kullback-Leibler divergence, for which a closed-form solution can be applied [39].

The question arises as to how the error Δ can reasonably be set. Since (5) includes log-likelihoods, we can form a good argument for the derivation of Δ on the percentage basis of the log-likelihood characteristics. However, the lower bound of the difference of both log-likelihood characteristics in (5) may approach negative infinity (e.g., for models estimated from small-sized training data and a larger number of dimensions of the pattern space). Therefore, it is difficult to justify the basis for calculating the percentage. Instead of deriving the error Δ based on the log-likelihood characteristics $L_{\mathrm{Av-ln}}(\boldsymbol{\Theta}_i|Z)$ and $L_{\mathrm{E-ln}}(\mathrm{G})$, we infer the error by using the likelihood (not log-likelihood) characteristics. For this purpose, we rewrite (1) in the form

$$L_{\mathrm{E-ln}}(\mathrm{G}) = \mathrm{E}\left\{ \ln L(\boldsymbol{\Theta}^{\mathrm{G}}\,|X\,)\right\} = \\ = \ln\left\{ \lim_{N\to\infty} \left[\prod_{i=1}^{N} p(\boldsymbol{x}_i \,\Big|\, \boldsymbol{\Theta}^{\mathrm{G}})\right]^{\frac{1}{N}}\right\} \;=\; \ln\left\{L_{\mathrm{E}}(\mathrm{G})\right\}\;, \qquad (6)$$

where $L_{\mathrm{E}}(\mathrm{G})$ denotes the "mean value" or more precisely the geometric mean of the likelihood function of the generator G. Similarly, it is possible to rewrite (4) as

$$L_{\mathrm{Av-ln}}(\boldsymbol{\Theta}_i|Z) \;=\; \ln\left\{\left[\prod_{j=1}^{J} p(\boldsymbol{z}_j|\boldsymbol{\Theta}_i)\right]^{\frac{1}{J}}\right\} \;=\; \ln\left\{L_{\mathrm{Av}}(\boldsymbol{\Theta}_i|Z)\right\}, \qquad (7)$$

where $L_{\mathrm{Av}}(\boldsymbol{\Theta}_i|Z)$ again denotes the geometric average of likelihood values computed on the Gaussian model with parameters $\boldsymbol{\Theta}_i$ and a set of test data Z.

We now consider the ratio $L_{\mathrm{Av}}(\boldsymbol{\Theta}_i|Z)\,/\,L_{\mathrm{E}}(\mathrm{G})$. With increasing i for estimation of the model $\boldsymbol{\Theta}_i$, this ratio varies from 0 to 1, with the value reaching 1 as i approaches infinity. For increasing i, we now attempt to determine a model $\boldsymbol{\Theta}_i$ such that the ratio $L_{\mathrm{Av}}(\boldsymbol{\Theta}_i|Z)$ and $L_{\mathrm{E}}(\mathrm{G})$ just exceeds the predetermined value β, i.e., $L_{\mathrm{Av}}(\boldsymbol{\Theta}_i|Z)\,/\,L_{\mathrm{E}}(\mathrm{G}) \geq \beta$, where β can be set in terms of estimation accuracy in the range from 0 to 1 (alternatively, from 0% to 100%). To the given inequality, we now apply logarithms, and after manipulation, we obtain

$$\ln L_{\mathrm{E}}(\mathrm{G}) - \ln L_{\mathrm{Av}}(\boldsymbol{\Theta}_i|Z) \;=\; L_{\mathrm{E-ln}}(\mathrm{G}) - L_{\mathrm{Av-ln}}(\boldsymbol{\Theta}_i|Z) \leq -\ln\beta\,. \qquad (8)$$

Table 1: Corresponding values Δ_β for selected values of likelihood accuracy β

β	0.5	0.8	0.9	0.95	0.98	0.99
$\Delta_\beta = -\ln\beta$	0.693	0.223	0.105	0.051	0.020	0.010

This inequality can now be used to to manipulate (5)

$$i^*_\beta = \min_i \ \{ [\, L_{\mathrm{E}-\ln}(\mathrm{G}) - L_{\mathrm{Av}-\ln}(\boldsymbol{\Theta}_i \,|Z\,) \,] \leq -\ln\beta = \Delta_\beta \}\,. \tag{9}$$

The interpretation of this inequality is such that we attempt to determine the least value of training data i^*_β with which we are able to estimate model parameters $\boldsymbol{\Theta}_i$ that will provide the (geometric) mean of the likelihood values for test data set Z with accuracy higher than $\beta\, L_{\mathrm{E}}(\mathrm{G})$, where $L_{\mathrm{E}}(\mathrm{G})$ denotes the (geometric) mean of likelihood function of generator G. At the same time, we always assume a sufficiently large size of test data J.

If we now perform in an d-dimensional space a sufficient number of experiments (this number K should be at least 100) with randomly generated model parameters $\boldsymbol{\Theta}^{\mathrm{G}}$ and randomly generated sets of training and test data, for selected β values, we can obtain the resulting set $i^*_\beta(k)(k = 1, ..., K)$ and arrive at certain interesting statistical features.

3. Results of analytical and experimental studies

To verify the proposed method for estimating the training data size to determine the Gaussian model that can provide likelihood values with the previously defined properties, we carried out a number of experiments using $d = 1, 2, \ldots, 10, 15, 20, 25, \ldots, 50, 60, \ldots, 100$ and likelihood accuracy β =0.5, 0.8, 0.9, 0.95, 0.98, and 0.99. For these selected values of β, we determined the corresponding values of Δ_β (Table 1). For each combination of d and β, we carried out at least 100 experiments (several hundred for d = 1, 2, and 3), each with randomly generated parameters and randomly generated training Y and test Z data sets. We mention here that in each experimental trial, the training

and test data sets contained up to 2 million new data values provided by the source generator G along with the corresponding parameters.

In addition to the mean value $i^*_\beta(k)$ calculated for each d and β, we attempted to define a suitable interval or rather a kind of upper limit for the number of required samples for which it is possible to achieve the desired accuracy of likelihood values. Our analysis of the procedure of obtaining the values $i^*_\beta(k)$ (i.e., procedures of randomly generated parameters $\boldsymbol{\mu}^{\mathrm{G}}$ and $\boldsymbol{C}^{\mathrm{G}}$, randomly generated training Y and test Z data, averaging operations etc.) led us to the assumption (also confirmed practically) that $i^*_\beta(k)$ follows a normal distribution. In this case, the confidence interval can be defined for $i^*_\beta(k)$. In our case, we are interested in the upper limit $i^*_{\beta-\mathrm{up}}$ of this interval, for which the following expression holds [38]

$$i^*_{\beta-\mathrm{up}} = i^*_{\beta-\mathrm{mean}} + 1.64\frac{\sigma^*_\beta}{\sqrt{K}}, \tag{10}$$

where $i^*_{\beta-\mathrm{mean}} = \frac{1}{K}\sum_{k=1}^{K} i^*_\beta(k)$, σ^*_β represents the standard deviation computed from values $i^*_\beta(k)$, k=1, ... , K and K the number of samples in the experiment (as mentioned previously K was generally equal to 100). It should be noted that (10) is valid for a confidence level of 95%. The results of our experiments are listed in Table 2 and illustrated in Figure 1.

We mention here that the number of experiments was greater than 2500. Computational problems occurred only when we estimated values of $i^*_{0.99-\mathrm{up}}$ for dimensions of d=60, 70, 80, 90, and 100; many experiments were not completed because their execution exceeded the allocated computing time per experiment, which was 720 hours. Because missing values of these incomplete experiments clearly disrupt the correctness of the final estimations, the results of $i^*_{0.99-\mathrm{up}}$ for d=60, 70, 80, 90, and 100 are not listed in Table 2 and illustrated in Figure 1.

Finally, we briefly discuss the influence of ill-conditionality of the task on the results. We mention here that for all experiments with randomly generated Gaussian distribution parameters, we always determined the condition number of matrix $\kappa(\boldsymbol{C}^{\mathrm{G}})$, particularly the Frobenius norm condition number defined as

Table 2: Results of experiments for β=0.5, 0.8, 0.9, 0.95, 0.98, and 0.99; $d(j)$ denotes the number of dimensions of the pattern space for the j-th experimental data set, and σ^*_β the standard deviation.

j	$d(j)$	β=0.50		β=0.80		β=0.90		β=0.95		β=0.98		β=0.99	
		$i^*_{\beta-\mathrm{up}}$	σ^*_β	$i^*_{\beta-\mathrm{up}}$	σ^*_β	$i^*_{\beta-\mathrm{up}}$	σ^*_β	$i^*_{\beta-\mathrm{up}}$	σ^*_β	$i^*_{\beta-\mathrm{up}}$	σ^*_β	$i^*_{\beta-\mathrm{up}}$	σ^*_β
1	1	4.6	3.0	7.9	5.6	12.6	9.6	20.6	16.2	39.0	31.5	67.1	64.4
2	2	8.8	3.9	16.2	8.6	28.6	15.7	49.6	25.7	108.5	74.4	208.9	169
3	3	13.1	5.7	27.4	12.2	53.3	25.6	97.6	48.2	217.3	103	413.7	179
4	4	19.1	5.8	40.3	13.5	77.2	24.6	142.0	53.7	354.7	147	750.6	315
5	5	24.5	6.3	55.9	15.8	109.1	35.6	199.2	68.0	473.0	163	977.7	331
6	6	32.4	9.2	74.1	20.0	150.3	44.9	282.5	84.5	692.7	200	1 395	412
7	7	39.7	9.9	94.6	21.0	197.9	46.7	377.1	94.2	918.4	235	1 786	502
8	8	47.9	10.7	117.9	24.4	240.4	48.4	469.6	99.0	1 121	258	2 357	673
9	9	56.9	11.7	140.1	25.1	289.2	53.1	555.6	110	1 400	293	2 836	746
10	10	68.2	12.1	167.3	26.3	359.0	55.7	696.1	125	1 674	311	3 459	926
11	15	124.7	14.3	332.1	39.1	690.6	87.2	1 371	159	3 353	518	7 315	2 783
12	20	208.3	22.5	582.4	60.4	1 242	133	2 293	341	6 352	1 389	15 770	9 262
13	25	301.8	22.9	849.7	73.8	1 848	174	3 655	430	9 450	1 992	22 470	11 735
14	30	416.8	28.8	1 169	79.1	2 530	195	5 041	450	12 728	2 166	28 538	15 890
15	35	552.7	35.1	1 574	84.2	3 466	249	6 832	667	17 579	3 223	41 718	19 388
16	40	689.3	35.8	2 017	109	4 393	260	8 731	793	22 488	4 412	53 882	27 883
17	45	854.6	36.5	2 526	131	5 485	332	10 848	1 107	27 805	6 397	67 753	41 860
18	50	1 053	41.1	3 093	141	6 796	361	13 483	1 842	34 067	6 680	78 176	45 624
19	60	1 498	49.0	4 447	186	9 811	684	19 848	2 403	56 472	19 402	-	-
20	70	1 978	55.4	5 919	207	13 133	761	26 706	3 166	76 921	28 404	-	-
21	80	2 534	79.8	7 552	302	17 134	1 255	34 440	3 721	91 743	37 542	-	-
22	90	3 184	83.8	9 606	323	21 331	1 397	43 418	5 651	123 996	45 581	-	-
23	100	3 905	99.6	11 707	467	25 776	1 808	52 081	6 369	136 858	51 032	-	-

below [40]

$$\kappa(C^{\mathrm{G}}) = \left\| (\boldsymbol{C}^{\mathrm{G}})^{-1} \right\|_{\mathrm{F}} \left\| \boldsymbol{C}^{\mathrm{G}} \right\|_{\mathrm{F}} . \tag{11}$$

A matrix with condition number close to 1 is said to be well-conditioned, whereas when the condition number $\kappa(.)$ becomes large, the problem is regarded as being ill-conditioned. The question subsequently is, how large must $\kappa(.)$ be for a problem to be classified as ill-conditioned? Again, there is no clear answer to this question. Analytical studies indicate that for a system with condition number $\kappa(.)$, we can expect a reduction of roughly $\log_{10}\kappa(.)$ decimal places in the accuracy of the solution. The standard double-precision number format

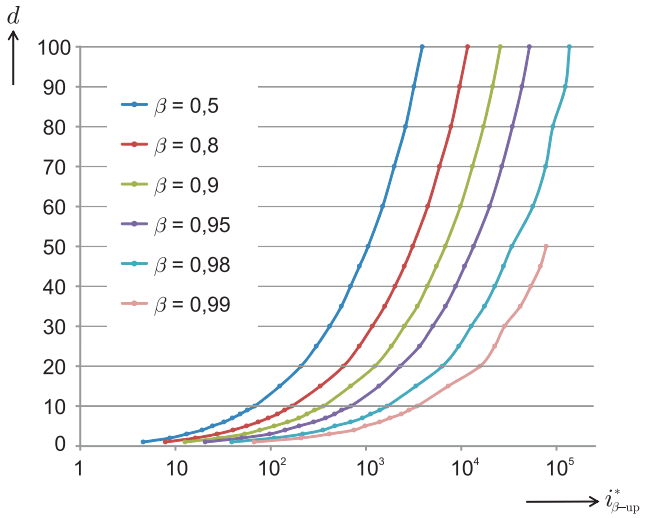

Figure 1: Illustration of experimental results for β=0.5, 0.8, 0.9, 0.95, 0.98, and 0.99.

in computation works with about 16 decimal digits of accuracy. Thus, these form the limits for the computation of ill-conditioned covariance matrices. In our case, in all experiments, most of the values $\kappa(\boldsymbol{C}^{\mathrm{G}})$ were in the interval $< 1; 10^6)$. However, even for several tens of values for which $\kappa(\boldsymbol{C}^{\mathrm{G}}) > 10^6$, we did not encounter any problems in computation, and the values of $i^*_{\beta-\mathrm{up}}$ were within the standard specified limits.

4. Determination of required data size based on dimensionality of pattern space and value of likelihood accuracy β

Our next concern is to determine whether it would be possible to describe the dependency between $i^*_{\beta-\mathrm{up}}$ and the dimension size (d) of the pattern space by a suitable functional relationship. If such a functional relationship with sufficient accuracy of approximation existed, it would not be necessary for a "fast" determination of the sample size of training data (with the specified accuracy β) to determine the recommended values in Table 2; they could instead be determined from a simple mathematical relation. Based on the method of determining the values of $i^*_{\beta-\mathrm{up}}$ (which is a statistical variable with variance σ^*_β)

and the shape of the curves reflecting dependency of $i^*_{\beta-\mathrm{up}}$ onto the number of dimensions of a pattern space, we decided to use polynomial regression. Here, we note that polynomial regression is a form of linear regression in which the relationship between the independent and the dependent variables is modeled as an I-th degree polynomial. It is expected that the values of the independent variables are burdened with a random error ε. Therefore, it is assumed (for a fixed β) that the values of $i^*_{\mathrm{up}}(j)$ can be approximated by a polynomial

$$i^*_{\mathrm{up}}(j) = q_0 + q_1 d^1(j) + q_2 d^2(j) + \ ... \ + q_I d^I(j) + \varepsilon_j \,, \tag{12}$$

where $j = 1, \ ..., \ J$, and J denotes the number of observations (in our case, the number of sets of experiments), I the degree of the polynomial used to approximate the observed dependence, and q_i the i-th coefficient of the polynomial approximation. Further, $d^i(j)$ denotes the i-th power of the j-th element of the input sequence of measurement (e.g., $d^i(17) = 45^i$, see Table 2), $i^*_{\mathrm{up}}(j)$ the j-th element of the output sequence of measurement, and ε_j the disturbance (or error) term.

A standard approach in regression analysis to approximate overdetermined systems is the method of least squares. The polynomial least-squares method falls into category of linear or ordinary least squares (OLS), which has a closed-form solution. The approach describes the variance in a prediction of the dependent variable $i^*_{\mathrm{up}}(j)$ as a function of the independent variable (in our case the dimensionality d) and the deviations from the fitted curve. Here, we remark that the Best Linear Unbiased Estimator (BLUE) of the coefficients is given by the OLS estimator; however, in this case, from the GaussMarkov theorem [41], the errors in linear regression model must have an expectation of zero, have no correlation, and have equal variances. From the manner of obtaining the source data, we can consider that the first two conditions are fulfilled in our case. However, the third condition is not met, i.e., errors ε_j for individual dimensions $d(j)$ have different variances (see Table 2); the data are heteroscedastic. However, if the data are uncorrelated but have different variances (heteroscedastic data), a modified approach based on weighted least squares can be used. Aitken [42]

showed that when a weighted sum of squared residuals is minimized, the estimation of q_i (i=0,1,..., I) is the BLUE if each weight is equal to the reciprocal of the variance of the measurements. We denote the weighted sum of squared residuals as $\Delta^2(\boldsymbol{q})$, and it can be expressed in the form

$$\Delta^2(\boldsymbol{q}) = \sum_{j=1}^{J} w_j \left[\sum_{i=0}^{I} q_i \, d^{\,i}(j\,) - i^*_{\rm up}(j\,) \right]^2 , \tag{13}$$

where J represents the number of observations (in our case the number of sets of measures), I the degree of the polynomial used in the approximation, and q_i the i-th coefficient of the polynomial approximation. Further, $d^{\,i}(j\,)$ represents the i-th power of the j-th element of the input sequence of measurement, $i^*_{\rm up}(j)$ the j-th element of the output sequence of measurement, and w_j the weight for a couple of values ($\sum_{i=0}^{I} q_i \, d^{\,i}(j\,)$, $i^*_{\rm up}(j\,)$), which according to Aitken [42] must equal $w_j = 1/(\sigma^*_\beta(j\,))^2$. Equation (13) can also be transcribed into a matrix notation as

$$\Delta^2(\boldsymbol{q}) = \left(\boldsymbol{D}\,\boldsymbol{q} - \boldsymbol{i}^*_{\rm up}\right)^{\rm T} \boldsymbol{W} \, \left(\boldsymbol{D}\,\boldsymbol{q} - \boldsymbol{i}^*_{\rm up}\right) , \tag{14}$$

where matrices $\boldsymbol{D}_{J\times(I+1)}$ and $\boldsymbol{W}_{J\times J}$ are respectively in the form

$$\boldsymbol{D} = \begin{pmatrix} 1 & d^1(1) & d^2(1) & .... & d^I(1) \\ 1 & d^1(2) & d^2(2) & .... & d^I(2) \\ : & : & : & : & : \\ 1 & d^1(J) & d^2(J) & .... & d^I(J) \end{pmatrix} \text{ and } \boldsymbol{W} = \begin{pmatrix} w_1 & 0 & ... & 0 \\ 0 & w_2 & ... & 0 \\ : & : & : & : \\ 0 & 0 & ... & w_J \end{pmatrix} \tag{15}$$

and vectors $\boldsymbol{q}$ and $\boldsymbol{i}^*_{\rm up}$ in the form $\boldsymbol{q} = [q_0, q_1, \ldots, q_I]^{\rm T}$ and $\boldsymbol{i}^*_{\rm up} = [i^*_{\rm up}(1), \ldots, i^*_{\rm up}(J)]^{\rm T}$, respectively.

To determine the values of q_i, i=0, 1, ... , I that minimize $\Delta^2(\boldsymbol{q})$, we differentiate (14) with respect to $\boldsymbol{q}$, and we set the result equal to 0

$$\frac{\partial}{\partial \boldsymbol{q}} \Delta^2(\boldsymbol{q}) = 2(\boldsymbol{D}^{\rm T}\boldsymbol{W}\,\boldsymbol{D}\,\boldsymbol{q} - \boldsymbol{D}^{\rm T}\boldsymbol{W}\,\boldsymbol{i}^*_{\rm up}) = 0 \, . \tag{16}$$

The vector $\boldsymbol{q}$ of coefficients of the polynomial approximation of the I-th degree, which minimizes the weighted sum of squared residuals, can be expressed as

$$\boldsymbol{q} = (\boldsymbol{D}^{\rm T}\boldsymbol{W}\,\boldsymbol{D})^{-1}\boldsymbol{D}^{\rm T}\boldsymbol{W}\,\boldsymbol{i}^*_{\rm up} \, . \tag{17}$$

We tested the functional dependence $i^*_{\mathrm{up}}(j)$ on the dimensionality d using a polynomial of the first, second, and third degrees. To evaluate the goodness of fit, we used a measure based on the root mean squared error ($RMSE$), which is also known as the standard error of the regression, and is defined as

$$RMSE = \sqrt{\frac{1}{v}\,\Delta^2(\boldsymbol{q})}\,, \tag{18}$$

where $\Delta^2(\boldsymbol{q})$ represents the weighted sum of squared residuals as defined in (13) or (14) and v the residual degrees of freedom, which is defined as the number of response values minus the number of fitted coefficients estimated from the response values. In our case, $v = J - (I + 1)$. The results of our experiments for the polynomials with degrees $I = 1$, 2, and 3 are listed in Table 3. From the $RMSE$ values, it is apparent that polynomials of the second degree approximate the relationship between $\boldsymbol{i}^*_{\mathrm{up}}(j)$ and the number of dimensions of the pattern space with sufficient accuracy, i.e., $RMSE$ for $I = 2$ and $I = 3$ are nearly identical.

For practical use of the obtained results, i.e., for easily memorizing the values of coefficients q_i of the polynomial approximation, we attempted to express the resulting polynomials in an even simpler form at the cost of only a slight increase in the corresponding $RMSE$. Table 4 lists the resulting coefficients $\hat{q}_i$ for individual values of β along with the resulting polynomial representing a simplified relationship between the required sample size $\hat{S}_\beta$ and the dimensions of the pattern space d. For a better insight into the results, we depict in Figures 2a-2f the sample size as a function of d for individual values of β dependencies expressed both by polynomials with coefficients q_i and simplified polynomials with coefficients $\hat{q}_i$.

A very interesting feature emerges upon analyzing the simplified approximation polynomial for $\beta = 0.95$. We can write the value of the sample size $\hat{S}_{0.95}$ in the form

$$\hat{S}_{0.95} = 5d(d+3) = 10 \times \left[\frac{d(d+3)}{2}\right], \tag{19}$$

where the term in brackets, i.e., $d(d+3)/2$, is equal to the number of parameters of the Gaussian model, which are estimated using training data (d parameters

Table 3: The resulting coefficients q_i of the approximation polynomials including corresponding root mean squared error ($RMSE$) computed for polynomial degrees of $I = 1$, 2, and 3 and for $\beta = 0.5$, 0.8, 0.9, 0.95, 0.98, and 0.99.

β	I=1			I=2			
	q_1	q_0	$RMSE$	q_2	q_1	q_0	$RMSE$
0.50	20.8	-50.5	8.32	0.36	2.98	1.23	0.15
0.80	50.1	-107.8	8.07	1.12	5.59	0.70	0.15
0.90	84.0	-147.7	6.23	2.52	9.82	-0.11	0.16
0.95	121.8	-179.0	4.25	5.11	16.2	-0.98	0.14
0.98	204.4	-228.7	2.26	13.4	33.4	-7.9	0.17
0.99	304.0	-290.0	1.08	31.2	43.7	-5.4	0.13

β	I=3				
	q_3	q_2	q_1	q_0	$RMSE$
0.50	$-2.0 \cdot 10^{-5}$	0.36	2.93	1.33	0.15
0.80	$-2.7 \cdot 10^{-4}$	1.15	5.19	1.46	0.14
0.90	$-1.3 \cdot 10^{-4}$	2.53	9.68	0.12	0.16
0.95	$1.6 \cdot 10^{-3}$	5.00	17.35	-2.62	0.13
0.98	$1.5 \cdot 10^{-2}$	12.6	40.00	-14.8	0.16
0.99	$6.5 \cdot 10^{-2}$	28.8	58.8	-20.1	0.13

for the mean and $d(d+1)/2$ for the covariance matrix). This result indicates that if we want to achieve a β value $\geq$0.95, we must use at least 10 times more training data compared to the number of estimated parameters of the Gaussian model. Here, we recall that the values of the sample size $\hat{S}_\beta$ are determined from the upper limits $i^*_{\beta-\mathrm{up}}$ of confidence intervals with confidence level of 95%. In a similar manner, we can also interpret the relation for $\hat{S}_{0.98}$, which leads to the recommendation that we use a data size that is at least 26 times the number of model parameters to be estimated for the Gaussian model.

Table 4: Simplified approximation polynomials representing relationships between the sample size $\hat{S}_\beta$ and the dimension d of the pattern space.

β	I=2				
	$\hat{q}_2$	$\hat{q}_1$	$\hat{q}_0$	$RMSE$	Sample size $\hat{S}_\beta$
0.50	0.35	3.5	0	0.30	$\hat{S}_{0.50} = 0.35\,d(d+10)$
0.80	1.10	6.6	0	0.26	$\hat{S}_{0.80} = 1.1\,d(d+6)$
0.90	2.50	10.0	0	0.17	$\hat{S}_{0.90} = 2.5\,d(d+4)$
0.95	5.00	15.0	0	0.24	$\hat{S}_{0.95} = 5\,d(d+3)$
0.98	13.00	39.0	0	0.22	$\hat{S}_{0.98} = 13\,d(d+3)$
0.99	30.00	45.0	0	0.15	$\hat{S}_{0.99} = 30\,d(d+1.5)$

5. Determination of value of likelihood accuracy β based on data size

In some applications, it is useful to know the correspondence between the likelihood accuracy β and the size n of the training data set. To derive dependencies between the likelihood accuracy β and the number n of data for modeling Gaussian density functions, we use results of previous experiments and techniques described especially in Section 2. Remind that for each dimension d of the pattern space we performed at least one hundred independent experiments each with randomly generated parameters and randomly generated training Y and test Z data sets. Results of experiments are given in Table 5. Table 5 lists the indicative number of data to construct a Gaussian density function which provides the likelihood accuracy $\beta = 0.1, 0.2, \ldots, 0.9, 0.95, 0.98,$ and 0.99 ; these results are presented for dimensions of the pattern space $d = 1, 2, \ldots, 10, 15, 20, 25, \ldots, 50, 60, \ldots, 100$ (e.g. for modeling the Gaussian density function in the space of dimension $d = 10$ and the likelihood accuracy $\beta = 0.9$, we will need approximately 360 data).

For a better insight into the results presented in Table 5, we depict in Figure 3 the likelihood accuracy β as a function of the sample size n for individual

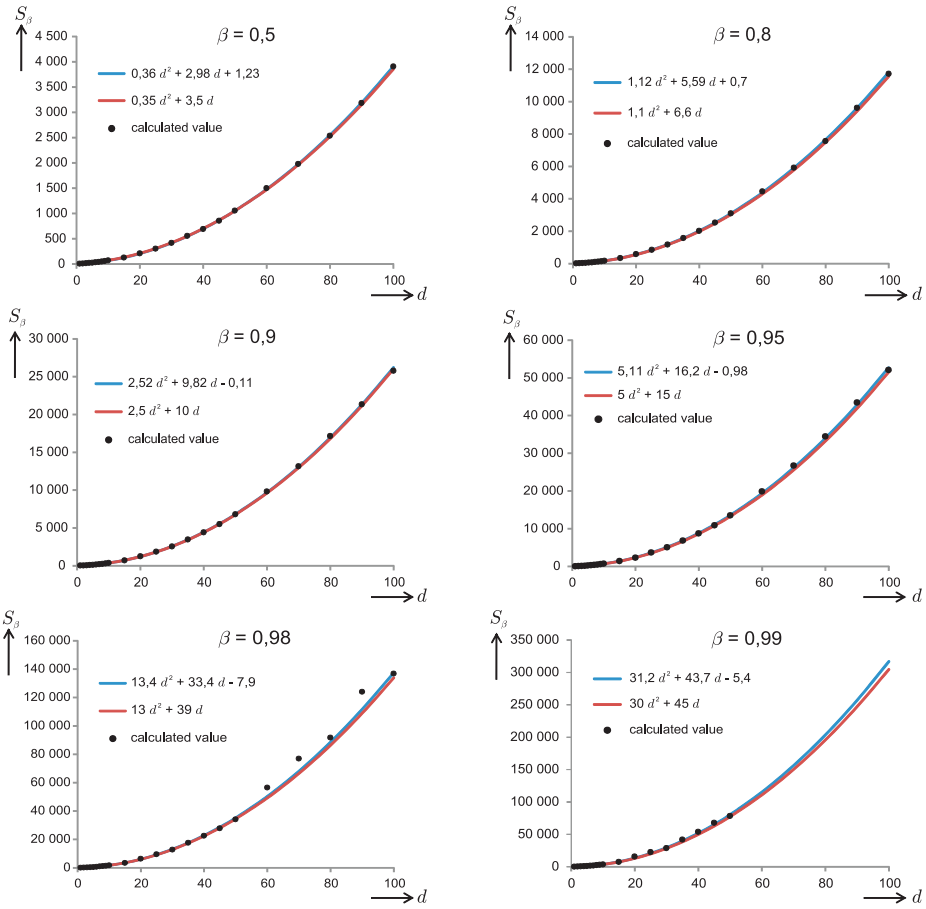

Figure 2: Illustrations of dependencies between the sample size S_β and the dimension d of the pattern space depicted both for approximation and simplified approximation polynomials and expressed for $\beta = 0.5$, 0.8, 0.9, 0.95, 0.98, and 0.99.

dimension d; such a function we denote as $\beta_d(n)$. Figure 3 shows the dependence $\beta_{10}(n)$, i.e. for $d = 10$. Let us mention that the shape of this dependence is very similar for the other examined dimensions d.

The functional dependence of this process is appropriate to model[1] by the

[1]For this purpose we used the Curve Fitting Toolbox in Matlab.

Table 5: The correspondence among the indicative number of training data n, the accuracy β of the Gaussian likelihood function and the dimension of the pattern space d.

d	# of	likelihood accuracy β											
	samples	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95	0.98	0.99
1	n	3.5	3.6	3.8	4.0	4.3	4.8	5.5	7	13	21	39	67
2	n	6.4	6.8	7.3	8.0	8.8	10	12	16	29	50	110	210
3	n	10.5	11.4	11.9	13.1	14.5	17	21	27	53	97	220	420
4	n	13.5	14.5	15.6	17.0	19	23	28	40	77	142	355	750
5	n	16.3	17.6	19.2	21.5	25	29	37	56	109	200	473	980
6	n	18	20	23	27	32	40	53	78	150	285	695	1 400
7	n	22	25	28	33	40	49	68	95	200	380	920	1 800
8	n	25	29	33	39	48	60	82	118	240	470	1 150	2 400
9	n	30	35	40	47	57	72	97	140	290	560	1 400	2 850
10	n	33	38	45	54	68	85	120	168	360	700	1 700	3 500
15	n	55	66	80	98	125	165	230	335	695	1 400	3 400	7 400
20	n	84	103	130	160	210	280	400	585	1 250	2 300	6 400	15 800
25	n	125	155	190	235	300	400	560	850	1 850	3 700	9 500	22 500
30	n	170	210	255	320	420	550	780	1 200	2 550	5 100	12 800	28 500
35	n	225	275	340	430	550	730	1 050	1 600	3 500	6 900	17 600	42 000
40	n	275	340	420	530	690	930	1 300	2 100	4 400	8 800	22 500	54 000
45	n	335	415	520	660	850	1 140	1 630	2 500	5 500	11 000	28 000	68 000
50	n	400	500	630	800	1 050	1 400	2 000	3 100	6 800	13 500	34 000	78 000
60	n	540	700	880	1 140	1 500	2 020	2 900	4 500	9 800	20 000	57 000	-
70	n	740	930	1 180	1 500	1 980	2 660	3 830	5 900	13 200	26 700	77 000	-
80	n	920	1 170	1 500	1 930	2 540	3 450	5 000	7 600	17 200	34 500	92 000	-
90	n	1 160	1 480	1 900	2 400	3 200	4 300	6 200	9 600	21 400	43 500	124 000	-
100	n	1 420	1 800	2 300	3 000	3 900	5 300	7 600	11 700	25 800	52 000	137 000	-

rational function, which can be written as

$$\beta_d(n) = P(n)/Q(n) \tag{20}$$

where P a Q are polynomials. Based on a careful analysis of the data in Table 5 we found that the functional relationship (20) can be approximated (with sufficient accuracy for all investigated dimensions d) by a simple rational func-

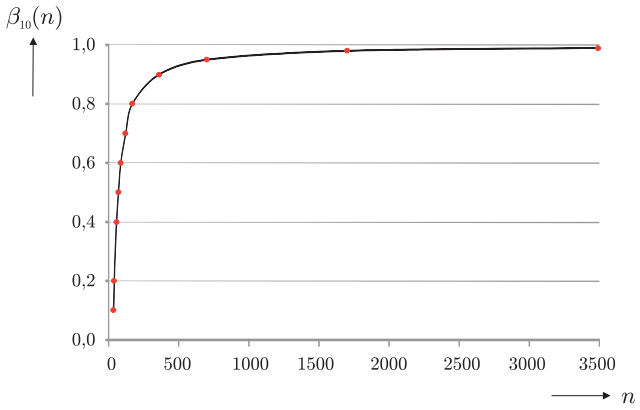

Figure 3: Illustration the likelihood accuracy $\beta_{10}(n)$ as a function of the sample size n for dimension $d = 10$.

tion where $P(n)$ and $Q(n)$ are polynomials of the first degree. The functional relationship (20) can therefore be indicated in the shape

$$\beta_d(n) = a_1(n + a_2)/(n + a_3)\,. \tag{21}$$

Values of the coefficients a_1, a_2, and a_3 are specified for each dimension d of the pattern space in Table 6. Since the values of a_1 are for all dimensions very close to 1, we tried to rewrite the equation (21) into a simplified form

$$\hat{\beta}_d(n) \approx (n + \hat{a}_2)/(n + \hat{a}_3)\,. \tag{22}$$

Approximate coefficients $\hat{a}_1$, $\hat{a}_2$, and $\hat{a}_3$ are listed in the second (upright) part of Table 6 including intervals of values of n for which relationships $\hat{\beta}_d(n)$ can be used. To evaluate the goodness of fit, we again (as in Section 4) used the root mean squared error (RMSE). The RMSE, which expresses here the error of approximation of β, is for each dimension less than 0.01 with the exception $d = 1$ and 2, where it is less than 0.02. Figure 4 summarizes all the results obtained in Section 5.

6. Conclusion

Our paper attempts to contribute towards solving problems that frequently appear in practical applications involving statistical Gaussian modeling. A fre-

Table 6: Rational functions and corresponding coefficients for modeling dependencies $\beta_d(n)$ and $\hat{\beta}_d(n)$.

d	$\beta_d(n) = a_1(n-a_2)/(n-a_3)$			$\hat{\beta}_d(n) \approx (n-\hat{a}_2)/(n-\hat{a}_3)$			
	a_1	a_2	a_3	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	n
1	1.001	-3.39	-2.47	1	-3.4	-2.4	$\langle 4;\infty)$
2	1.008	-6.16	-3.39	1	-6.1	-3.4	$\langle 7;\infty)$
3	1.001	-10.01	-5.47	1	-10.0	-5.5	$\langle 11;\infty)$
4	0.997	-12.82	-6.38	1	-12.8	-6.4	$\langle 14;\infty)$
5	0.996	-15.16	-5.82	1	-15.2	-5.8	$\langle 16;\infty)$
6	1.005	-16.62	-0.73	1	-16.3	-0.7	$\langle 18;\infty)$
7	1.002	-19.42	0.48	1	-19.4	0.5	$\langle 22;\infty)$
8	1.004	-22.24	3.46	1	-22.0	3.5	$\langle 25;\infty)$
9	1.003	-27.12	3.01	1	-27.0	3.0	$\langle 30;\infty)$
10	1.005	-29.11	9.35	1	-28.8	9.2	$\langle 33;\infty)$
15	1.006	-46.98	32.14	1	-46	32	$\langle 55;\infty)$
20	1.006	-69.81	71.24	1	-68	72	$\langle 84;\infty)$
25	0.998	-104.3	91.70	1	-104	92	$\langle 120;\infty)$
30	1.003	-140.5	134.1	1	-140	132	$\langle 170;\infty)$
35	1.001	-185.3	176.8	1	-185	177	$\langle 220;\infty)$
40	1.001	-218.9	250.1	1	-220	250	$\langle 270;\infty)$
45	1.001	-270.6	310.3	1	-270	310	$\langle 330;\infty)$
50	1.002	-319.9	412.2	1	-320	400	$\langle 400;\infty)$
60	1.003	-425.9	648.0	1	-420	640	$\langle 540;\infty)$
70	1.001	-582.3	808.7	1	-580	810	$\langle 740;\infty)$
80	1.002	-723.2	1 084	1	-720	1 100	$\langle 900;\infty)$
90	1.001	-901.4	1 368	1	-910	1 370	$\langle 1\,100;\infty)$
100	1.002	-1 108	1 681	1	-1 100	1 700	$\langle 1\,400;\infty)$

quent issue in such cases involves the sample size needed for maximum-likelihood estimates of such a Gaussian model. Our results provide useful recommendations for researchers working on applications wherein it is necessary to construct statistical models from experimental data with a given accuracy. The recommended values of the sample size resulting from our simplified approximation of the polynomial model (for accuracies of $\beta \geq 0.95$) are easy to remember (10 times the number of estimated parameters of the Gaussian model), and they surprisingly almost exactly coincide with recommendations for sample size as indicated by "statistical practitioners", but without the theoretical justification. Further, the derived dependencies of likelihood accuracy β on size n of the data sets for a given dimension d of the pattern space are very simple and can be useful in many application tasks.

Our experiments were performed for space dimensions of $d \leq 100$. Here, we note that some statistical modeling tasks (e.g., in the area of computer vision) work with considerably higher dimensions. Therefore, in our further research, we plan to address tasks wherein $d > 100$.

It must be mentioned that the results presented in the paper should be seen as a recommendation for estimation of the accuracy of the model; it is always useful to seek additional training data for increasing the model accuracy and robustness (particularly for tasks with a small number of pattern-space dimensions). We believe that the presented results could aid in applications wherein statistical modeling and statistical pattern recognition are involved.

Acknowledgment

This paper was supported by the project no. P103/12/G084 of the Grant Agency of the Czech Republic. Computational resources were provided by the CESNET LM2015042 and the CERIT Scientific Cloud LM2015085, "Projects of Large Research, Development, and Innovations Infrastructures".

[1] J. L. Fleiss, B. Levin, M. C. Paik, Statistical Methods for Rates and Properties, John Wiley & Sohns, New York, 2003.

[2] K. Kelley, S. E. Maxwell, Sample size for multiple regression: Obtaining regression coefficients that are accurate, not simply significant, Psychological Methods 8 (3) (2003) 305 – 321. doi:10.1037/1082-989X.8.3.305.

[3] T. T. Cai, T. Liang, H. H. Zhou, Law of log determinant of sample covariance matrix and optimal estimation of differential entropy for high-dimensional Gaussian distributions, Journal of Multivariate Analysis 137 (2015) 161 – 172. doi:10.1016/j.jmva.2015.02.003.

[4] H. H. Nguyen, V. Vu, Random matrices: Law of the determinant, The Annals of Probability 42 (1) (2014) 146 – 167. doi:10.1214/12-AOP791.

[5] N. El Karoui, Spectrum estimation for large dimensional covariance matrices using random matrix theory, Ann. Statist. 36 (6) (2008) 2757–2790. doi:10.1214/07-AOS581.

[6] P. J. Bickel, E. Levina, Covariance regularization by thresholding, The Annals of Statistics 36 (6) (2008) 2577–2604. doi:10.1214/08-AOS600.

[7] J. S. Long, Regression Models for Categorical and Limited Dependent Variables, SAGE Publications Inc., Thousand Oaks, London, New Delhi, 1997.

[8] J. S. Long, J. Freese, Regression Models for Categorical Dependent Variables Using Stata, Stata Press, College Station, Texas, 2014.

[9] J. V. Psutka, J. Psutka, Sample size for maximum likelihood estimated of Gaussian model, Lecture Notes on Computer Science: Computer Analysis of Images and Patterns 9257 (2015) 462 – 469. doi:10.1007/978-3-319-23117-4_40.

[10] C. Xie, J. Chang, Y. Liu, Estimating the number of components in Gaussian mixture models adaptively, Journal of Information and Computational Science 10 (14) (2013) 4453. doi:10.12733/jics20102195.

[11] H. Akaike, A new look at the statistical model identification, IEEE Transactions on Automatic Control 19 (6) (1974) 716–723. doi:10.1109/TAC.1974.1100705.

[12] G. Schwarz, Estimating the dimension of a model, Ann. Statist. 6 (2) (1978) 461 – 464. doi:10.1214/aos/1176344136.

[13] J. V. Psutka, Gaussian mixture model selection using multiple random subsampling with initialization (2015) 678 – 689doi:10.1007/978-3-319-23192-1_57.

[14] J. Zhao, L. Jin, L. Shi, Mixture model selection via hierarchical BIC, Computational Statistics & Data Analysis 88 (2015) 139 – 153. doi:10.1016/j.csda.2015.01.019.

[15] D. Kim, B. Seo, Assessment of the number of components in Gaussian mixture models in the presence of multiple local maximizers, Journal of Multivariate Analysis 125 (2014) 100 – 120. doi:10.1016/j.jmva.2013.11.018.

[16] S. H. Lin, I. M. Wu, On the common mean of several inverse Gaussian distributions based on a higher order likelihood method, Applied Mathematics and Computation 217 (12) (2011) 5480 – 5490. doi:10.1016/j.amc.2010.12.019.

[17] M.-S. Yang, C.-Y. Lai, C.-Y. Lin, A robust EM clustering algorithm for Gaussian mixture models, Pattern Recognition 45 (11) (2012) 3950 – 3961. doi:10.1016/j.patcog.2012.04.031.

[18] A. Mehrjou, R. Hosseini, B. N. Araabi, Improved Bayesian information criterion for mixture model selection, Pattern Recognition Letters 69 (2016) 22 – 27. doi:10.1016/j.patrec.2015.10.004.

[19] H. Jin, M.-L. Wong, K. S. Leung, Scalable model-based clustering for large databases based on data summarization, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (11) (2005) 1710–1719. doi:10.1109/TPAMI.2005.226.

[20] C. Ari, S. Aksoy, O. Arikan, Maximum likelihood estimation of Gaussian mixture models using stochastic search, Pattern Recognition 45 (7) (2012) 2804 – 2816. doi:10.1016/j.patcog.2011.12.023.

[21] L. Scrucca, Identifying connected components in Gaussian finite mixture models for clustering, Computational Statistics & Data Analysis 93 (2016) 5 – 17. doi:10.1016/j.csda.2015.01.006.

[22] J. Yu, C. Chaomurilige, M.-S. Yang, On convergence and parameter selection of the EM and DA-EM algorithms for gaussian mixtures, Pattern Recognition 77 (2018) 188 – 203. doi:10.1016/j.patcog.2017.12.014.

[23] B. Jian, B. C. Vemuri, Robust point set registration using Gaussian mixture models, IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (8) (2011) 1633–1645. doi:10.1109/TPAMI.2010.223.

[24] K. Kayabol, S. Kutluk, Bayesian classification of hyperspectral images using spatially-varying Gaussian mixture model, Digital Signal Processing 59 (2016) 106 – 114. doi:10.1016/j.dsp.2016.08.010.

[25] Y. Liu, F. Perronnin, A similarity measure between unordered vector sets with application to image categorization, in: 2008 IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8. doi:10.1109/CVPR.2008.4587600.

[26] V. Hughes, Sample size and the multivariate kernel density likelihood ratio: How many speakers are enough?, Speech Communication 94 (2017) 15 – 29. doi:10.1016/j.specom.2017.08.005.

[27] P. Kumar, H. Gauba, P. P. Roy, D. P. Dogra, Coupled HMM-based multi-sensor data fusion for sign language recognition, Pattern Recognition Letters 86 (2017) 1 – 8. doi:10.1016/j.patrec.2016.12.004.

[28] J. Du, Y. Hu, H. Jiang, Boosted mixture learning of Gaussian mixture hidden markov models based on maximum likelihood for speech recognition,

IEEE Transactions on Audio, Speech, and Language Processing 19 (7) (2011) 2091–2100. doi:10.1109/TASL.2011.2112352.

[29] W. Dong, M. Zhou, Gaussian classifier-based evolutionary strategy for multimodal optimization, IEEE Transactions on Neural Networks and Learning Systems 25 (6) (2014) 1200–1216. doi:10.1109/TNNLS.2014.2298402.

[30] F. Lv, G. Yang, W. Zhu, C. Liu, Generative classification model for categorical data based on latent Gaussian process, Pattern Recognition Letters 92 (2017) 56 – 61. doi:10.1016/j.patrec.2017.03.025.

[31] B. Bayar, N. Bouaynaya, R. Shterenberg, SMURC: high-dimension small-sample multivariate regression with covariance estimation, IEEE Journal of Biomedical and Health Informatics 21 (2) (2017) 573–581. doi:10.1109/JBHI.2016.2515993.

[32] N. Majdi-Nasab, M. Analoui, E. J. Delp, Decomposing parameters of mixture Gaussian model using genetic and maximum likelihood algorithms on dental images, Pattern Recognition Letters 27 (13) (2006) 1522 – 1536. doi:10.1016/j.patrec.2006.03.005.

[33] A. Pags-Zamora, M. Cabrera-Bean, C. Daz-Vilor, Unsupervised online clustering and detection algorithms using crowdsourced data for malaria diagnosis, Pattern Recognition 86 (2019) 209 – 223. doi:10.1016/j.patcog.2018.09.001.

[34] P. R. Gawde, A. K. Bansal, J. A. Nielson, Integrating Markov model, bivariate Gaussian distribution and GPU based parallelization for accurate real-time diagnosis of arrhythmia subclasses, in: K. Arai, R. Bhatia, S. Kapoor (Eds.), Proceedings of the Future Technologies Conference (FTC) 2018, Springer International Publishing, Cham, 2019, pp. 569–588.

[35] W. Peng, Model selection for Gaussian mixture model based on desirability level criterion, Optik - International Journal for Light and Electron Optics 130 (2017) 797 – 805. doi:10.1016/j.ijleo.2016.10.125.

[36] X.-S. Tang, D.-Q. Li, Z.-J. Cao, K.-K. Phoon, Impact of sample size on geotechnical probabilistic model identification, Computers and Geotechnics 87 (2017) 229 – 240. doi:10.1016/j.compgeo.2017.02.019.

[37] T. M. Cover, J. A. Thomas, Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing), Wiley-Interscience, 2006.

[38] M. Smithson, Confidence Intervals, SAGE Publications Inc., Thousand Oaks, London, New Delhi, 2003.

[39] J. Duchi, Derivations for linear algebra and optimizations, http://stanford.edu/~jduchi/projects/general_notes.pdf.

[40] G. H. Golub, C. F. Van Loan, Matrix Computations, The Johns Hopkins University Press, Baltimore, Maryland USA, 1996.

[41] W. D. Berry, Understanding Regression Assumptions, SAGE Publications Inc., Newbury Park, California, 1993.

[42] A. C. Aitken, On least-squares and linear combinations of observations, Proceedings Of the Royal Society of Edinburgh 55 (2014) 42 – 48. doi:10.1017/S0370164600014346.

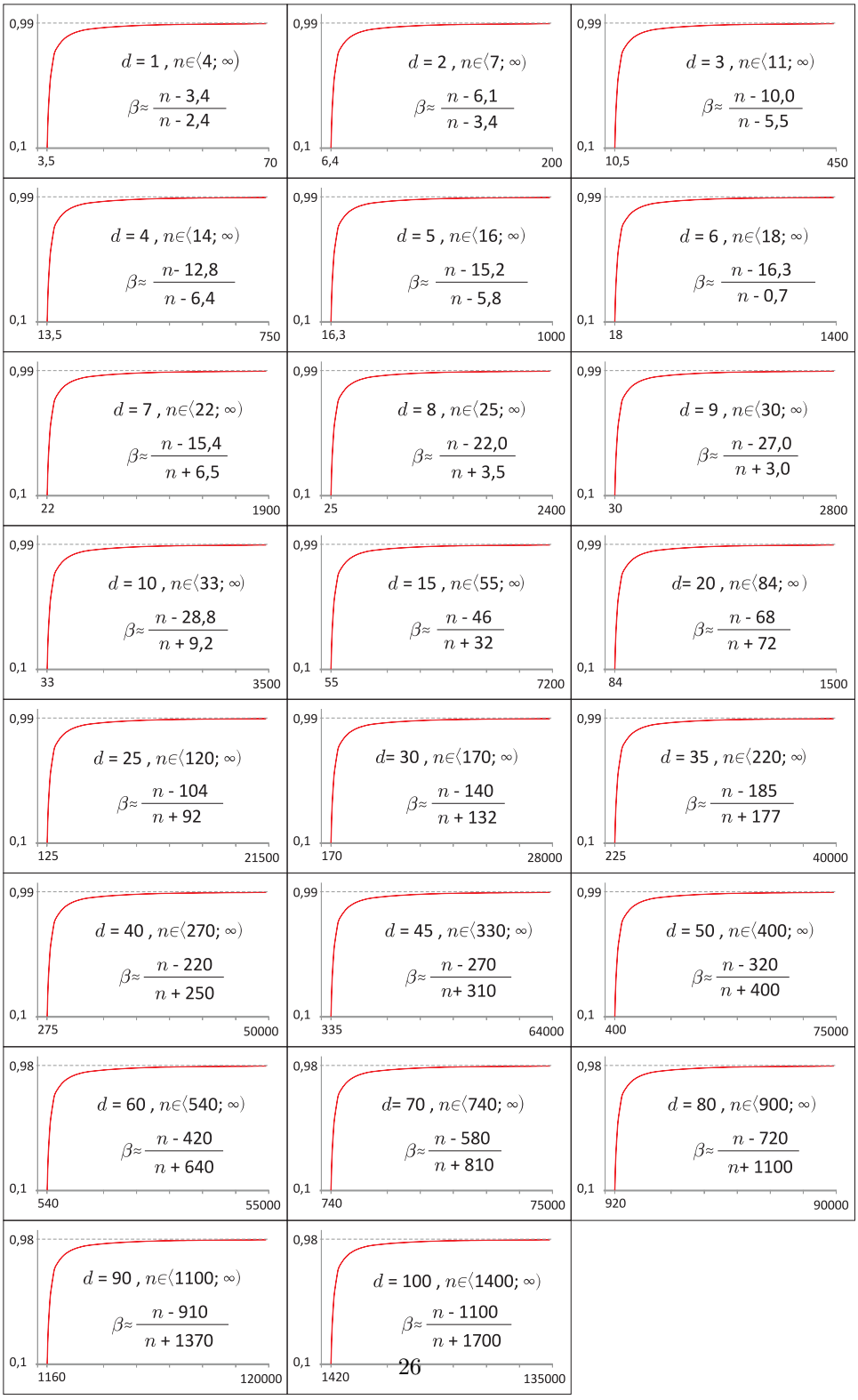

Figure 4: Illustration the likelihood accuracy $\beta_d(n)$ as a function of the sample size n for dimension $d = 1, 2, \ldots, 10, 15, 20, \ldots, 50, 60, \ldots, 100$.