Estimation of Single-Gaussian and Gaussian Mixture Models for Pattern Recognition

Jan Vaněk, Lukáš Machlica, Josef Psutka

University of West Bohemia in Pilsen, Univerzitní 22, 306 14 Pilsen
Faculty of Applied Sciences, Department of Cybernetics
{vanekyj,machlica,psutka}@kky.zcu.cz

Abstract. Single-Gaussian and Gaussian-Mixture Models are utilized in various pattern recognition tasks. The model parameters are estimated usually via Maximum Likelihood Estimation (MLE) with respect to available training data. However, if only small amount of training data is available, the resulting model will not generalize well. Loosely speaking, classification performance given an unseen test set may be poor. In this paper, we propose a novel estimation technique of the model variances. Once the variances were estimated using MLE, they are multiplied by a scaling factor, which reflects the amount of uncertainty present in the limited sample set. The optimal value of the scaling factor is based on the Kullback-Leibler criterion and on the assumption that the training and test sets are sampled from the same source distribution. In addition, in the case of GMM, the proper number of components can be determined.

Keywords: Maximum Likelihood Estimation, Gaussian Mixture Model, Kullback-Leibler Divergence, Variance, Scaling

1 Introduction

In this article the estimation of parameters of a single Gaussian and Gaussian Mixture Models (GMMs) is investigated. Gaussian models are often used in pattern recognition in order to classify or represent the data. An input training set is given and the task is to extract relevant information in a form of a statistical model. The training set is often limited, thus it is difficult, sometimes even impossible, to capture the true/source data distribution with high accuracy. Moreover, in extreme cases the estimation can produce numerically unstable estimates of unknown model parameters. In order to estimate the model parameters often Maximum Likelihood Estimation (MLE) is used. MLE focuses just on the training set [1], not respecting the representativeness of the true/source distribution from which the given data were sampled. However, in the pattern recognition, the performance of a system on unseen data is crucial.

Methods proposed in this article are based on a reasonable assumption that the source distribution of the training and test set are the same. Therefore, the proposed criterion focuses on the similarity of the true data distribution and estimated model parameters. For this purpose we use the Kullback-Leibler Divergence (KLD) [2] and we integrate over the entire parameter space. We investigate the case where at first the

model parameters are estimated via MLE, and subsequently only the variance parameters are modified. Indeed, the variance does reflect the uncertainty of the model.

At first, the situation with single Gaussian models is examined. Further, the conclusions are extended to the case of Gaussian mixture models. The proposed method is able to determine a proper number of GMM components, which is often set empirically (several data-driven approaches were already studied, see [3–5]).

We demonstrate on a sequence of experiments that the log-likelihood of the modified model given an unseen test set increases, mainly in situations when the number of training data is low.

2 Estimation of Parameters of a Single-Gaussian Model

Assume a random data set $X = \{x_1, x_2, \ldots, x_n\}$, which is iid (independent and identically distributed), and sampled from univariate normal distribution $\mathcal{N}(0, 1)$. The sample mean $\hat{\mu}$ and sample variance $\hat{\sigma}^2$ are given by the formulas:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2. \tag{1}$$

From Central Limit Theorem, it can be derived that the estimate of the sample mean $\hat{\mu}$ has normal distribution $\mathcal{N}(0, \frac{1}{n})$, and the estimate of the sample variance $(n-1)\hat{\sigma}^2$ has a Chi-square distribution $\chi^2(n-1)$ with $n-1$ degrees of freedom and variance equal to $2n-2$ [6]. Note that both the distributions of sample mean and sample variance depend only on the number of samples n. Estimates (1) give the best log-likelihood on the training set, but since MLE does not involve any relation to the source distribution of the data, these estimates do not achieve the highest value of the log-likelihood for unseen data generated from the source distribution $\mathcal{N}(0, 1)$.

Since maximization of the log-likelihood of the model given data sampled from the source distribution is strongly related to the minimization of a KLD [7], we propose a new criterion based on KLD:

$$J(\alpha, n) = E_{\hat{\mu},\hat{\sigma}^2}\left\{D_{\mathrm{KL}}(\mathcal{N}(0,1)\|\mathcal{N}(\hat{\mu}, \alpha\hat{\sigma}^2))\right\}, \tag{2}$$
$$\hat{\mu} \sim \mathcal{N}(0, 1/n), \;\; (n-1)\hat{\sigma}^2 \sim \chi^2(n-1)$$
$$J(\alpha, n) = \iint D_{\mathrm{KL}}(\mathcal{N}(0,1)\|\mathcal{N}(\hat{\mu}, \alpha\hat{\sigma}^2)) p_{\hat{\mu}} p_{\hat{\sigma}^2} \mathrm{d}\hat{\mu}\mathrm{d}\hat{\sigma}^2, \tag{3}$$

where $E_{\hat{\mu},\hat{\sigma}^2}\{\}$ denotes the expectation computed over parameters $\hat{\mu}$, $\hat{\sigma}^2$; α is the unknown scaling factor of the sample variance, and $p_{\hat{\mu}}$, $p_{\hat{\sigma}^2}$ are the prior distributions (normal and scaled χ^2) of sample mean and sample variance, respectively. Thus, we measure how much information is lost when the source distribution $\mathcal{N}(0, 1)$ is approximated by the estimated model $\mathcal{N}(\hat{\mu}, \alpha\hat{\sigma}^2)$. The task is to find an optimal scaling factor α, which depends on the number of samples n and provides the best match of the sample model and the source distribution.

Given the assumptions above the KLD is equal to:

$$D_{\mathrm{KL}}(\mathcal{N}(0,1)\|\mathcal{N}(\hat{\mu}, \alpha\hat{\sigma}^2)) = \frac{1}{2}\left(\frac{\hat{\mu}^2}{\alpha\hat{\sigma}^2} + \frac{1}{\alpha\hat{\sigma}^2} + \ln\alpha + \ln\hat{\sigma}^2 - 1\right) \tag{4}$$

Before the derivation of the solution of (3), let us define:

$$Q(n) = \int_0^\infty \frac{1}{\hat{\sigma}^2} p_{\hat{\sigma}^2} \mathrm{d}\hat{\sigma}^2 = G(n) \int_0^\infty \frac{1}{\hat{\sigma}^2} (\hat{\sigma}^2)^{n/2-1} \exp\left(-\frac{1}{2}\hat{\sigma}^2\right) \mathrm{d}\hat{\sigma}^2, \tag{5}$$

$$G(n) = (2^{n/2}\Gamma(n/2))^{-1}, \tag{6}$$

where $G(n)$ is the normalization term guaranteeing that the χ^2 probability distribution function integrates to one. In order to get an analytical solution for $Q(n)$ let us use the integration by substitution, where the substitution $\delta = 1/\hat{\sigma}^2$ is used. Then, it is easy to show that [6]:

$$\begin{aligned} Q(n) &= G(n) \int_0^\infty \delta \left[\delta^{-n/2-1} \exp\left(-\frac{1}{2\delta}\right)\right] \mathrm{d}\delta \\ &= \int_0^\infty \delta \, p_\delta \, \mathrm{d}\delta = \frac{1}{n-2}, \;\; n > 2, \end{aligned} \tag{7}$$

where p_δ is the Inv-$\chi^2(n)$ distribution with n degrees of freedom, therefore (7) is in fact the mean of this distribution.

Now, substituting for KLD in (3) from (4) and utilizing (7) we get:

$$\begin{aligned} J(\alpha, n) &= const + \frac{1}{2}\left(\frac{1}{\alpha}\int_{-\infty}^\infty \hat{\mu}^2 p_{\hat{\mu}} \mathrm{d}\hat{\mu} \int_0^\infty \frac{1}{\hat{\sigma}^2} p_{\hat{\sigma}^2} \mathrm{d}\hat{\sigma}^2 + \frac{1}{\alpha}\int_0^\infty \frac{1}{\hat{\sigma}^2} p_{\hat{\sigma}^2} \mathrm{d}\hat{\sigma}^2 + \ln\alpha\right) \\ &= const + \frac{1}{2}\left(\frac{n-1}{n\alpha} Q(n-1) + \frac{n-1}{\alpha} Q(n-1) + \ln\alpha\right) \\ &= const + \frac{(n+1)(n-1)}{2n\alpha} Q(n-1) + \frac{1}{2}\ln\alpha, \end{aligned} \tag{8}$$

where $const$ represents the part of the criterion independent of α. To find the minimum of (8), the partial derivative is taken with respect to the unknown parameter α. Setting the derivative to zero yields:

$$\frac{\partial J}{\partial \alpha} = 0 \Longrightarrow \frac{1}{2\alpha} - \frac{(n^2-1)}{2n\alpha^2} Q(n-1) = 0, \tag{9}$$

$$\alpha_n = \frac{n^2-1}{n} Q(n-1) = \frac{n^2-1}{n(n-3)}. \tag{10}$$

It should be stated that $Q(n-1)$ given in (7) has no solution for $n < 4$. However, sometimes also models for a low amount of samples may be requested (such situation may occur quite often when estimating GMM parameters, see Section 3). Therefore, we extrapolated the α values in order to get the solution for $n > 1$. The function used for extrapolation was a rational one, what is in agreement with the solution given in (10). Moreover, we request that the first derivative and the value at the point $n = 3.5$ (this point was taken to match the experimental values for $n < 4$ reported below) of the extrapolation function and function given by equation (10) are equal. The form of the extrapolation function is:

$$\alpha_n = \frac{66.83}{n-1} - 20.31, \tag{11}$$

which goes to infinity at the point $n = 1$.

To support the analytically derived values we performed several experiments. At first we draw a large amount of n-tuples for a specific value of n, and computed sample mean and sample variance of samples in each tuple. Next, we took each sample mean and sample variance computed in the previous step, multiplied the sample variance by one specific value of α, evaluated the KLD (4) for each sample mean and scaled sample variance, and computed the mean $m_{\alpha,n}^{\mathrm{KLD}}$ across all the obtained KLDs. This was repeated for various values of α. Finally, the optimal value α^* was the one which gave minimal $m_{\alpha,n}^{\mathrm{KLD}}$, thus $\alpha^* = \arg\min_\alpha m_{\alpha,n}^{\mathrm{KLD}}$. The process was repeated several times, hence the optimal value of α was a random variable. The graph of optimal variance scaling factors α^* obtained analytically and experimentally is depicted in Figure 1, note that for increasing n the value of α^* converges to 1.

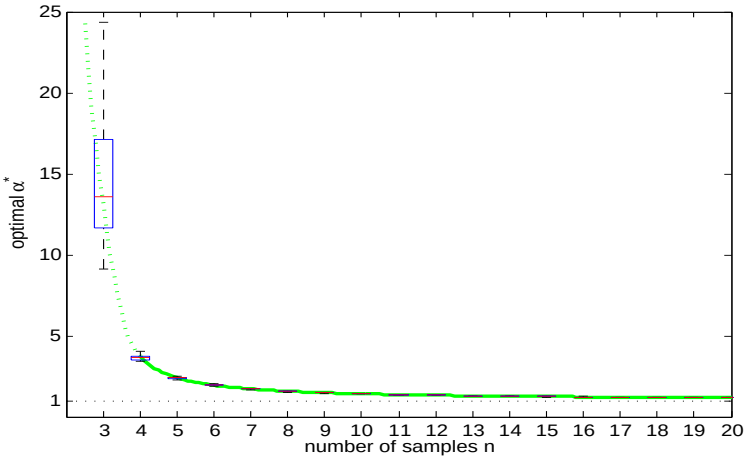

Fig. 1. *Dependence of the optimal value of variance scaling factor α on the number of samples. The solid line represents the optimal values given by the analytical solution* (10)*, the dotted line represents the extrapolation* (11)*. The edges of the boxes represent the 25th and 75th percentile of the optimal α^* computed using the Monte Carlo simulations described in the text, and the line inside the box is the median value.*

2.1 Additional Notes

- When deriving the multiplication factor α, for simplicity the source distribution was assumed standard normal $\mathcal{N}(0, 1)$. Without any loss of generality the solution is valid also for the more general case of the source distribution $\mathcal{N}(\mu, \sigma^2)$, but the derivations would involve additional shifting and scaling.
- The solutions (10) and (11) can be used also for non-integer values, e.g. in the estimation process of GMM discussed below.

– As illustrated in Figure 1 and from the fact that for $n < 4$ analytical solution for α is not defined, models estimated from such a low amount of samples are unreliable. Hence, a careful consideration should precede before they are used.
– By now, only a univariate case was assumed. In the multivariate case with a diagonal covariance matrix, individual dimensions are mutually independent. Therefore, the scaling factor α can be applied on each diagonal element of the covariance matrix separately (recall that α depends only on the number of training data).
– Dealing with multivariate normal distributions with full covariance matrices is considerably more difficult. A method based on two multiplicative constants, one for diagonal and one for non-diagonal elements of the covariance matrix, was proposed in [8].

3 Robust Estimation of Parameters of a GMM

In the case of a Gaussian mixture model with diagonal covariance matrix, the conclusions made in the previous section may be used. Thus, variance of individual Gaussians is multiplied by the scaling factor α_n in dependence on the number of samples accounted for this Gaussian. However, rather than an exact number of samples accounted for each Gaussian, a soft count n_m^{s} is given for each Gaussian $m = 1, \ldots, M$:

$$n_m^{\mathrm{s}} = \sum_{t=1}^{n} \gamma_{mt}, \;\; \gamma_{mt} = \frac{\omega_m \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_m, \boldsymbol{C}_m)}{\sum_{i=1}^{M} \omega_i \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_i, \boldsymbol{C}_i)} \tag{12}$$

where γ_{mt} is the a-posterior probability of feature vector $\boldsymbol{x}_t$ occupying m-th Gaussian in the GMM, n is the overall number of samples, ω_m is the weight of the m-th Gaussian. Now, new ML estimates of mean vectors $\hat{\boldsymbol{\mu}}_m$ and diagonal covariance matrices $\hat{\boldsymbol{C}}_m$ of a GMM are computed as:

$$\hat{\boldsymbol{\mu}}_m = \frac{1}{n_m^{\mathrm{s}}} \sum_{t=1}^{n} \gamma_{mt} \boldsymbol{x}_t, \tag{13}$$

$$\hat{\boldsymbol{C}}_m = \mathrm{diag}\left(\frac{1}{n_m^{\mathrm{s}}} \sum_{t=1}^{n} \gamma_{mt} (\boldsymbol{x}_t - \hat{\boldsymbol{\mu}}_m)(\boldsymbol{x}_t - \hat{\boldsymbol{\mu}}_m)^{\mathrm{T}} \right), \tag{14}$$

where the function diag() zeros the non-diagonal elements.

As discussed in Section 2, the distribution of diagonal elements of sample covariance matrix $\hat{\boldsymbol{C}}_m$ is the scaled $\chi^2(n_m^{\mathrm{e}} - 1)$ distribution with variance $n_m^{\mathrm{e}} - 1$, but note that n_m^{e} does not equal n_m^{s}. The value of n_m^{e} will depend on a-posteriors γ_{mt}, and in order to derive the correct value we will proceed as follows.

Given two sample sets X_a of size n_a and X_b of size n_b drawn from $\mathcal{N}(0, 1)$, the variance of the sample mean of each set will be $1/n_a$ and $1/n_b$. Note that the variance of the total sum of sample sets X_a, X_b is:

$$\mathrm{var}\left(\sum_{x \in X_a} x \right) = n_a, \;\; \mathrm{var}\left(\sum_{x \in X_b} x \right) = n_b. \tag{15}$$

Now, let all the samples in the set X_a be weighted by a scalar a and the samples in X_b by a scalar b. The variance of the total sum of sample sets X_a, X_b changes to:

$$\text{var}\left(\sum_{x\in X_a} ax\right) = a^2 n_a, \;\; \text{var}\left(\sum_{x\in X_b} bx\right) = b^2 n_b. \tag{16}$$

Let X_c be the set constructed from all of the weighted samples from both X_a and X_b. The weighted sample mean and the variance of the total sum of samples in X_c are given by formulas:

$$\hat{\mu}_c = \frac{\sum_{x\in X_a} ax + \sum_{x\in X_b} bx}{an_a + bn_b}, \tag{17}$$

$$\text{var}\left(\sum_{x\in X_a} ax + \sum_{x\in X_b} bx\right) = a^2 n_a + b^2 n_b, \tag{18}$$

respectively, and therefore for the variance of the weighted sample mean $\hat{\mu}_c$ we get:

$$\text{var}(\hat{\mu}_c) = \frac{a^2 n_a + b^2 n_b}{(an_a + bn_b)^2}. \tag{19}$$

In the case, where each sample in the set X_c is weighted by a different weight c_i, equation (19) changes to:

$$\text{var}(\hat{\mu}_c) = \frac{\sum_{i=1}^{n_c} c_i^2}{\left(\sum_{i=1}^{n_c} c_i\right)^2}. \tag{20}$$

Comparing the variance of weighted and unweighted sample mean, the equivalent number of unweighted samples n^{e} can be derived:

$$\frac{1}{n^{\text{e}}} = \frac{\sum_{i=1}^{n_c} c_i^2}{\left(\sum_{i=1}^{n_c} c_i\right)^2}, \;\; n^{\text{e}} = \frac{\left(\sum_{i=1}^{n_c} c_i\right)^2}{\sum_{i=1}^{n_c} c_i^2}. \tag{21}$$

Hence, in the case of mth Gaussian in the GMM the value of n_m^{e} is given as:

$$n_m^{\text{e}} = \frac{\left(\sum_{t=1}^{n} \gamma_{mt}\right)^2}{\sum_{t=1}^{n} \gamma_{mt}^2}. \tag{22}$$

Note that the value of n_m^{e} is a real number, but this is not a problem since both (10) and (11) are defined also for non-integer values.

3.1 Robust Update of GMM Variances

According to equations derived above, the robust estimation of GMM consists of steps:

1. Compute new maximum likelihood estimate of means (13) and covariances (14) of the GMM.
2. Evaluate the value of n_m^{e} given in (22) for each $m = 1, \ldots, M$.

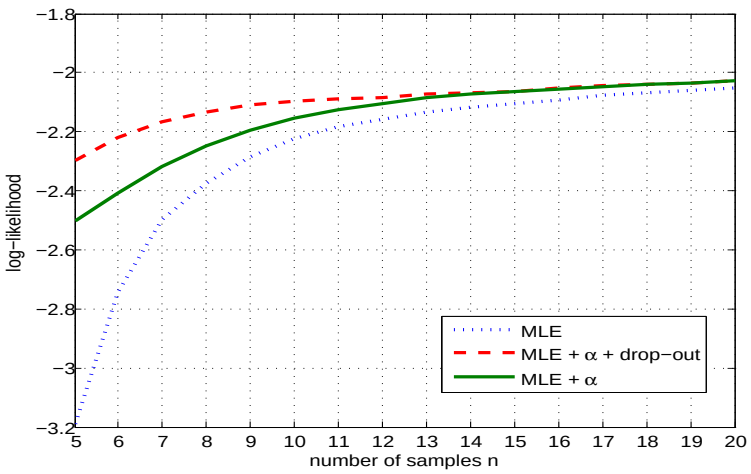

Fig. 2. *Dependence of the log-likelihood of a GMM given a large number of samples generated from the source distrubtion on the number of samples used to train the GMM. The source distribution of samples is represented by a GMM with 2 components, from which limited amount of data is sampled. In common, 3 GMMs with 2 components were trained, but only from the limited number of samples (x-axis) generated from the source distribution. Dotted line represents the baseline (GMM trained via MLE, no variance adjustments); in the case of the solid line MLE estimates of the GMM's variance were multiplied by the optimal scaling factor α; in the case of the dashed line the scaling factor α was used and GMM components with $n_m^e < 4$ were discarded during the estimation process (only a single Gaussian model was used). The experiment was run a large number of times, and for each number of training samples (x-axis) the mean value of log-likelihood, obtained in each run of the experiment, was computed.*

3. Compute the scaling factor $\alpha_{m,n_m^{\mathrm{e}}}$ for each Gaussian $m = 1, \ldots, M$ given the respective n_m^{e}.
4. Multiply diagonal elements of each covariance matrix $\hat{\boldsymbol{C}}_m$ by $\alpha_{m,n_m^{\mathrm{e}}}$.

We performed simple experiments, which demonstrate the effect of the proposed procedure. Results are given in Figure 2. Note that when the GMM components with $n_m^{\mathrm{e}} < 4$ are discarded during the estimation process, the log-likelihood of the test (unseen) samples is higher. Since the training of a GMM is an iterative procedure, the number of equivalent samples n_m^{e} is determined in each iteration for each GMM component m. Thus, the number of GMM components is controlled through the entire estimation. Hence, a GMM with a proper number of components is obtained at the end of the estimation.

4 Conclusions

The paper investigated the estimation of parameters of Gaussian models in cases with low amount of training data. It was shown that the model trained via MLE does not generalize well to unseen data. We have demonstrated how to adjust the parameters if

the source distribution of test and training data is identical. The method is based on the Kullback-Leibler divergence, we adjust the variance of the model multiplying it by a scaling factor α, which depends only on the number of samples.

Through the paper a crucial assumption was made that the samples are mutually independent. However, this is often not the case in real applications (e.g. time series of a measurement), where instead of number of given samples one should estimate the number of independent samples. I.e. the information content present in a set of mutually dependent samples is lower than the information content in a sample set of the same size containing independent samples. Therefore, the estimated number of independent samples should be lower. Technique aimed to estimate the independent number of samples was investigated in [8].

The proposed estimation updates were incorporated into the GMM estimation software implemented at the Faculty of Applied Sciences, University of West Bohemia, Czech Republic. The GMM estimator supports both diagonal and full covariance matrices, and it is well suited for processing of large datasets. Moreover, it supports also acceleration provided by GPU [9], [10] and multi-threaded SSE instructions. The license is free for academic use. More information are available at `http://www.kky.zcu.cz/en/sw/gmm-estimator`.

5 Acknowledgments

This research was supported by the Technology Agency of the Czech Republic, project No. TA01011264.

References

[1] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., et al.: Top 10 Algorithms in Data Mining. In: Knowledge and Information Systems, pp. 1-37, 2007.
[2] Kullback, S., Leibler, R.A.: On Information and Sufficiency. In: Annals of Mathematical Statistics 22, pp. 79-86, 1951.
[3] Bell, P.: Full Covariance Modelling for Speech Recognition. Ph.D. Thesis, The University of Edinburgh, 2010.
[4] Figueiredo, M., Leitão, J., Jain, A.: On Fitting Mixture Models. In: Proc. EMMCVPR 1999, Lecture Notes In Computer Science, Springer-Verlag London, pp. 54–69, 1999.
[5] Paclík, P., Novovičová, J.: Number of Components and Initialization in Gaussian Mixture Model for Pattern Recognition. In: Proc. Artificial Neural Nets and Genetic Algorithms, Springer-Verlag Wien, pp. 406–409, 2001.
[6] Taboga, M.: Lectures on Probability Theory and Mathematical Statistics. CreateSpace Independent Publishing Platform, ISBN: 978-1480215238, 2008.
[7] Bishop, C. M.: Pattern Recognition and Machine Learning (1st ed. 20.). Springer, ISBN: 978-0387310732, 2007.
[8] Vanek J., Machlica, L., Psutka, J.V., Psutka, J.: Covariance Matrix Enhancement Approach to Train Robust Gaussian Mixture Models of Speech Data. In: SPECOM, 2013.
[9] Machlica, L., Vanek, J., Zajic, Z.: Fast Estimation of Gaussian Mixture Model Parameters on GPU using CUDA. In: Proc. PDCAT, Gwangju, South Korea, 2011.
[10] Vanek J., Trmal, J., Psutka, J.V., Psutka, J.: Optimized Acoustic Likelihoods Computation for NVIDIA and ATI/AMD Graphics Processors. In: IEEE Transactions on Audio, Speech and Language Processing, Vol. 20, 6, pp. 1818–1828, 2012.