Modified Predictive Validation Test for Gaussian Mixture Modelling

Mohammad Sadeghi and Josef Kittler

Centre for Vision, Speech and Signal Processing
School of Electronics, Computing and Mathematics, University of Surrey
Guildford GU2 7XH, UK
{M.Sadeghi,J.Kittler}@surrey.ac.uk
http://www.ee.surrey.ac.uk/CVSSP/

Abstract. This paper is concerned with the problem of probability density function estimation using mixture modelling. In [7] and [3], we proposed the `Predictive Validation`, PV , technique as a reliable tool for the Gaussian mixture model architecture selection. We propose a modified form of the PV method to eliminate underlying problems of the validation test for a large number of test points or very complex models.

1 Introduction

Consider a finite set of data points $\mathbf{X} = \mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_N$, where $\mathbf{x}_i \in \Re^d$ and $1 \leq i \leq N$, that are identically distributed samples of the random variable $\mathbf{x}$. We wish to find the function that describes the data, i.e. its *pdf* , $p(\mathbf{x})$. Building such a model has many potential applications in pattern classification, clustering and image segmentation.

There are basically two major approaches to density estimation: parametric and non-parametric. The parametric approach involves assuming a specific functional form for the data distribution and estimating its parameters from the data set with a likelihood procedure. If the selected form is correct, it leads to an accurate model. In contrast, non-parametric methods attempt to perform an estimation without constraints on the global structure of the density function. The problem with this approach is that the number of parameters in the model quickly grows with the size of the data set. This leads to a huge computational burden, even with today's most capable processors. Semi-parametric techniques offer a successful compromise between parametric and non parametric methods. A finite mixture of functions is assumed as the functional form but the number of free parameters are allowed to vary which motivates a more complex and adaptable model. The number of free parameters does not depend upon the size of data set. The most widely used class of density functions for mixture modelling are Gaussian functions, which are attractive because of their isotropic and unimodal nature, along with their capability to represent distribution by a mean vector and covariance matrices.

An important problem of Gaussian mixture modelling approaches is selection of the model structure, i.e. the number of components. In [7] and [3], we

T. Caelli et al. (Eds.): SSPR&SPR 2002, LNCS 2396, pp. 414–423, 2002.
© Springer-Verlag Berlin Heidelberg 2002

proposed the `Predictive Validation`, PV , technique as a reliable solution to this problem. The PV method provides an absolute measure of goodness of the model which is based on the `calibration` concept: a density function is calibrated if, for the set of events they try to predict, the predicted frequencies match the empirical frequencies derived from the data set. The agreement between the predicted and empirical frequencies is checked using the `chi-squared` statistic. The main problem with this goodness of fit test is that it usually rejects almost everything for a large number of test points. As more accurate models can be built only with a large number of samples, we face a fundamental contradiction which could be resolved only by accepting compromise solutions. Furthermore, in some applications, the data distribution is not exactly a mixture of Gaussian functions and it is not possible to model the data accurately with a finite mixture of such functions. However, a model with a reasonable goodness of fit works practically well.

In this article, we revisit the PV test and eliminate the underlying problem of model validation for a large number of test samples or a very complex model. We show that with a modified test, we can obtain a well behaving measure of goodness of fit which identifies the best structure of the mixture. If the data set can truly be modelled by a finite mixture of Gaussian functions, the method succeeds in finding it. Otherwise, it tries to find the `best` estimation. By best, we mean the simplest model which describes the data distribution well. Also, an important problem in *pdf* modelling approaches is model initialisation. We demonstrate that the PV technique is also quite useful for dealing with this problem.

The rest of this paper is organised as follows. In the next section we define Gaussian mixture models and review the PV technique used to obtain the mixture structure. The problem of the goodness of fit test and our solution to the problem is detailed in Section 3. In Section 4, the use of the validation test to aid the model initialisation is shown. The experimental results are given in Section 5. Finally, some conclusions are drawn and possible directions for future research are suggested in Section 6.

2 Gaussian Mixture Modelling

A mixture model is defined by equation (1). The mixture components, $p(\mathbf{x}|j)$, satisfy the axiomatic property of probability density functions, $\int p(\mathbf{x}|j)d\mathbf{x} = 1$ and the coefficients P_j, the mixing parameters, are chosen such that $\sum_{j=1}^{M} P_j = 1$ and $0 \le P_j \le 1$.

$$p(\mathbf{x}) = \sum_{j=1}^{M} p(\mathbf{x}|j)P_j \qquad (1)$$

A well known group of mixture models is Gaussian mixture in which

$$p(\mathbf{x}|j) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}_j|}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)\right\} \qquad (2)$$

where $\boldsymbol{\mu}_j$ is the d-dimensional mean vector of component j and $\boldsymbol{\Sigma}_j$ the covariance matrix. For a given number of components, P_j, $\boldsymbol{\mu}_j$, and $\boldsymbol{\Sigma}_j$ are estimated via a standard maximum likelihood procedure using the *EM* algorithm [2,3]. An initial guess of the Gaussian mixture parameters is made first. The parameter values are then updated so that they locally maximise the log-likelihood of the samples. Unfortunately, the EM algorithm does not guarantee to find a global maximum. It can easily get confined to a local maximum or saddle point. For this reason, different initialisations of the algorithm have to be considered which may give rise to different models being obtained. Also, the most important problem which is examined under model selection is that prior knowledge of the number of components is rarely available.

2.1 Model Selection

There are several methods for selecting the architecture, i.e. the number of components, M. The simplest approach is to select the model which optimises the likelihood of the data given model [8]. However, this method requires a very large data set which is rarely available. Moreover, this model building process is biased to selecting more complex models than actually required, with the risk of over-fitting the data. Information criteria attempt to remove this bias using an auxiliary term which penalises the log-likelihood by the number of parameters required to define the model (AIC) [1] or by a factor related to the sample size (BIC) [8]. The main advantage of information criteria is their simplicity. The downside is that the chosen penalty term depends on the problem analysed. If the function is complex and the penalty is too strong the model will be underfitted. We advocated the use of the `predictive validation` method. The goal is to find the least complex model that gives a satisfactory fit to the data. The model selection algorithm using the PV technique is a bottom up procedure which starts from the simplest model, a one component model and keeps adding components until the model is validated [3].

The basis of the validation test is that a good model can predict the data. Suppose that a model M_j with j components has been computed for data set $\mathbf{X}$. The validation test is performed by placing hyper-cubic random size windows in random places of the observation space and comparing the empirical and predicted probability. The former is defined as $p_{emp}(\mathbf{x}) = \frac{N_W}{N}$, where N_W is the number of training points falling within window W, and the latter $p_{pred}(\mathbf{x}) = \int_W p(\mathbf{x})d\mathbf{x}$. Although the window size is selected randomly, more stable results can be achieved by controlling it so that p_{emp} falls within a limited range[3]. The agreement between the empirical and predicted frequency is checked by a weighted linear least square fit of p_{emp} against p_{pred} .

3 Weighted Least Squares Fit

If the estimated *pdf* model is good the empirical and predicted frequencies should be approximately equal. Making repeated observations of p_{emp} and p_{pred} permits a weighted linear least square fit between p_{emp} and p_{pred} to be formed

$$p_{emp} = a + b \cdot p_{pred} \tag{3}$$

where a is the intercept and b is the gradient. If the model is good then it should be possible to fit a linear model to the data points. Furthermore, the fitted line should lie close to the line $y = x$. To fit the straight line to the set of points and to check whether the fitted line is close to the desired line the chi-square statistic is used. In these statistical procedures, measurement error plays a crucial role. The chi-square statistics is defined as

$$\chi^2 = \sum_i \left[\frac{1}{\sigma^{(i)}} (y^{(i)} - a - bx^{(i)}) \right]^2 \tag{4}$$

where $\sigma^{(i)}$ is the standard deviation of the measurement error in the y coordinate of the ith point. If the measurement errors are normally distributed then this function will give the maximum likelihood parameter estimation of a and b. To determine a and b, equation (4) is minimised [3].

To check whether a linear model can be applied to the data correctly, a goodness-of-fit measure, $Q(\chi^2|\nu)$, is computed. This is done via the incomplete gamma function, Γ [5]. If the goodness-of-fit test fails, the validation test also fails and it proceeds no further. Since, the line parameters are estimated by minimising equation (4), from this equation, we can see that the relative sizes of $\sigma^{(i)}$ do not affect the placement of the fitted line. They do affect the value of the χ^2 statistic which we use to test the linear model's validity. This is why it is imperative to calculate $\sigma^{(i)}$ correctly.

After the best fit line has been found we need to check whether this line is statistically close to the $y = x$ line. This can be done again by making use of the chi-squared statistic. For the data set we have found a minimum value of χ^2_{min} for our estimated parameters, a and b. If these values are perturbed then the value of χ^2 increases. The change in the chi-squared value, $\Delta\chi^2 = \chi^2 - \chi^2_{min}$, defines an elliptical confidence region around the point $[a, b]^T$.

$$\Delta\chi^2 = \begin{bmatrix} \delta a \\ \delta b \end{bmatrix}^T \cdot \begin{bmatrix} \sigma_a^2 & \sigma_{ab}^2 \\ \sigma_{ab}^2 & \sigma_b^2 \end{bmatrix}^{-1} \cdot \begin{bmatrix} \delta a \\ \delta b \end{bmatrix} \tag{5}$$

where δa and δb are the changes in the line parameters, σ_a^2 and σ_b^2 are the variances in the estimates of a and b respectively and σ_{ab}^2 is the covariance of a and b [5,3]. In the original PV test, to accept a model we computed the 99.0% confidence region around $[a, b]^T$ and checked whether our true parameter value vector $[0, 1]^T$ is encompassed within this elliptical region [3], i.e. the model is accepted if for $[\delta a \;\; \delta b] = [0 - a \;\; 1 - b]$

$$\Delta\chi^2 \leq \Delta\chi^2_\nu(p) \tag{6}$$

where ν is the degree of freedom in $\Delta\chi^2$, i.e. the number of parameters and p is the desired confidence interval. For a confidence level of 99.0% with two degrees of freedom, the value of $\Delta\chi^2_\nu(p)$ is 9.21.

Our further investigations showed that this test is very hard to pass when it is performed using too many test points. In [3] we checked experimentally the assumption of the un-correlatedness between measurement errors which affects the value of the degree of freedom and we found that this assumption is justifiable. The choice of the standard deviation of the measurement errors, $\sigma^{(i)}$, is an important issue in the test. An overestimated value helps the test to pass, but it may lead to an under-fitted model and very small error make the χ^2 test difficult to pass. So, to deal with the problem of modelling using a large data set, the choice of the measurement error is studied more accurately.

3.1 Measurement Uncertainty

In the PV method [7,3], standard deviation of the measurement error, $\sigma^{(i)}$, is estimated using a binomial distribution. Considering the random sized window, W, in the feature space, the probability of finding a point within the window is p. The probability of finding it outside W is $q = 1 - p$. In other words the number of points falling inside W, N_W, is a stochastic variable which is binomially distributed. The standard deviation of a binomial distribution is given by

$$\sigma^{(i)}_{binom} = \sqrt{\frac{p^{(i)}(1-p^{(i)})}{N}} \tag{7}$$

$p^{(i)}$ is estimated by the empirical probability value within the window, $p^{(i)}_{emp}$.

By considering p_{pred} as the measurement without uncertainty (x coordinate), equation (7) makes a good approximation of the measurement error on the empirical probability value, p_{emp} . However, the effect of some other error sources like the effects of sampling and the integration error on p_{pred} need to be studied. If such errors are important a bias term has to be added to the measurement error.

To investigate the effect of the bias experimentally, we built a single component Gaussian model. This model was then used to generate 500 samples. The empirical and predicted probabilities were then calculated within randomly placed windows using the data and the true model. Finally, the mean and variance of $p_{diff} = p_{emp} - p_{pred}$ were calculated. Obviously, in the ideal conditions these values should be zero. This experiment was repeated for the different number of Gaussian components, data samples and window placements. The experimental results showed that the variance is almost independent of the number of components and the number of window placements and highly dependent on the data set size. Moreover, the experiments showed that σ_{diff} changes in a very similar manner to σ_{binom} when the number of test points changes. Therefore, σ_{binom} describe the sampling error well and integration error is negligible and no additional term as the bias error needs to be taken into account.

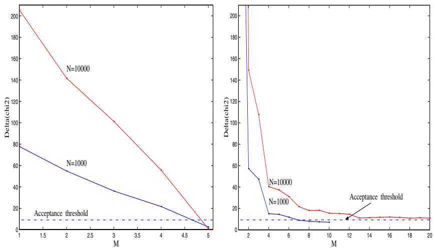
3.2 F Test

Consider a specified number of Gaussian components, M. As the number of data samples increases, the variance of the binomial distribution, equation (7), and therefore, the value of the elements of the covariance matrix in equation (5) reduce. At the same time, if the Gaussian model has not been improved significantly, the change in the chi-squared value, $\varDelta\chi^2$, increases which makes the test more difficult to pass. In fact, in the modelling process, if the data distribution is a perfect mixture of normal functions, a model with the same number of Gaussian components would become more accurate as the number of samples increases. Eventually, its parameters would become identical with the true distribution parameters. So, in the validation test, although σ_{binom} and the variances of the line parameters decrease, the difference between the estimated and the true line parameters also decrease, so $\varDelta\chi^2$ will not increase noticeably.

However, in a number of practical applications, the distribution is not an exact Gaussian mixture model. In such a conditions, when the number of data samples increases, although a more accurate model is achieved, the resulting effect on the improvements of the line parameters is not as significant as the effect caused by the reduction of σ_{binom}. Now, even when the number of components is increased, the improvement in $\varDelta\chi^2$ is not significant enough to meet the condition 6.

Figure 1(a) shows the value of $\varDelta\chi^2$ versus the number of components, M, when the method is used to model 1000 and 10000 samples generated by a mixture of 5 Gaussian components while figure 1(b) shows the results of the same experiments for a face image data set. Figure 2 also shows the logarithm of $\varDelta\chi^2$ versus the number of data points for different number of Gaussian components. As we expect, when the size of the `class5` data set increases, although $\varDelta\chi^2$ for the incorrect structures ($M < 5$) becomes larger, for the correct one ($M = 5$), it is even smaller than the value for the smaller size data set which emphasises that using more data points, more accurate model is achieved. For the face data set, the problem is not the same. Using about 1 percent of the image samples, 1000 samples, an eight components model is validated. When 10000 samples are used to train and validate the model, as the number of components is increased, a better model is built and $\varDelta\chi^2$ is reduced accordingly. For the models with more than 13 Gaussian components, although $\varDelta\chi^2$ is very close to the acceptance threshold, it is not reduced noticeably. In the PV method, we are seeking the simplest model which predicts the data well. So, it seems that, the model selection process has to be controlled using a more intelligent test.

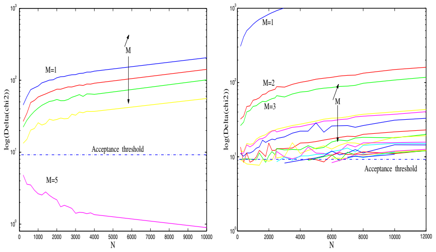
The simplest solution to this problem is to avoid using large data sets and instead use a few samples, especially in the model validation stage. Such a solution may lead to an inaccurate model. The other solution is to define an acceptable structural error and add a term as the residual error to equation 7. However, selection of such an error is an important and difficult problem. In different applications and for different data sets, this term has to be selected carefully. As plot 1(b) suggests, the best solution is to check whether adding more components to the model improves the prediction ability of the model or not. Since, very high structural error is also not desirable, the absolute value of $\varDelta\chi^2$ also has to be taken into account. As we mentioned earlier in the validation test the 99.0% confidence region around the estimated line parameters is considered as the re-



(a) `class5` data set

(b) A face data set

Fig. 1. $\Delta\chi^2$ versus the number of components using 1000 and 10000 samples



(a) `class5` data set

(b) A face data set

Fig. 2. $log(\Delta\chi^2)$ versus the number of samples for various M

quired confidence limit. Our experiments demonstrate that if the condition 6 is satisfied with such a confidence limit, the Gaussian model is absolutely reliable. Now, we propose that if the true parameters are within the 99.9% confidence area and more complex models do not improve $\Delta\chi^2$ significantly, the model is acceptable. In order to check the $\Delta\chi^2$ value variations, we apply the `F-test`.

The `F-test` is usually applied to check whether two distributions have significantly different variances. It is done by trying to reject the null hypothesis that the variances are consistent. The statistic `F` which is the ratio of the variances indicates very significant differences if either $F >> 1$ or $F << 1$ [5]. If we consider $\Delta\chi^2_M$ and $\Delta\chi^2_{M-1}$ as the change in the chi-squared value of the model M and $M-1$, since they have χ^2 distribution, their ratio obeys the Fisher's F-distribution law with (ν_1, ν_2) degrees of freedom where $\nu_1 = \nu_2 = 2$, i.e. the number of the line parameters. The distribution of `F` in the null case is calculated using equation 8.

$$Q(F|\nu_1, \nu_2) = I_{\frac{\nu_1}{\nu_2 + \nu_1 F}} \left(\frac{\nu_2}{2}, \frac{\nu_1}{2} \right) \tag{8}$$

where $I_x(\alpha, \beta)$ is incomplete beta function [5]. In the validation process, if the true line parameters are between the 99.0% and 99.9% confidence area of the estimated parameters, to check whether $\Delta\chi^2_M$ and $\Delta\chi^2_{M-1}$ are consistent, F is considered as the ratio of the larger value to the smaller one. Then $p = 2 \cdot Q(F|\nu_1, \nu_2)$ is calculated. If the value of p is very close to one ($p > 0.99$), the null hypothesis is accepted [5] and the model is validated.

4 Model Initialisation

As we mentioned earlier, model initialisation is an important problem in the mixture modelling and different initialisations may lead to different models. We adopted our PV technique to select the best initialised model. In the model selection algorithm, for a given number of components, M, different models are built using the EM algorithm with different initialisation. During the validation step, the change in the chi-squared value, $\Delta\chi^2$, is calculated and the model with the minimum $\Delta\chi^2$ value is selected as the best M components model. If this minimum value satisfies the PV tests conditions also, the model is accepted.

5 Experiments

Two groups of experiments are reported here. In the first experiments the performance of the modified PV technique is compared with the original one. Then, the improvement achieved in a specific application, lip tracker initialisation, is shown.

5.1 Comparison of the Model Selection Methods

These experiments were performed on the `class5` data set, the face data set and the lip area of the face data set. The first row of figure 3 contains the experiments results using the information criteria methods, AIC and BIC, while the next row shows the results of the same experiments using the PV methods. In these plots the results using the original validation method (say *M1*), the results when the model initialisation is checked by the PV technique (*M2*) and the results when the modified test is also applied (*M3*) have been shown. Figures 3(a) and (d) contain plots of the number of components accepted versus the sample size considering the `class5` data set. As one can see, the AIC and BIC methods usually select over-fitted models. A five component model is always built using the *M2* and *M3* methods. Apparently, in such a cases, no structural error needs to be taken into account. Figures 3(b, e) and (c, f) show the results when performing the same experiments considering samples generated from the face and lip data sets. Although more stable results are obtained when the model is initialised intelligently, the effect of the `F` test on the model validation is noticeable. The test offers a compromise solution between the model accuracy and the model complexity.

5.2 Lip Tracker Initialisation

In [6], Gaussian Mixture Modelling using the PV technique along with a Gaussian components grouping algorithm was used to aid an un-supervised classification of lip data. The lip pixel classification experiments were performed on 145 colour images taken from the xm2vts database [4]. The first column of figure 4 shows two examples of the rectangular colour mouth region image blocks. The second and the third columns show the associated segmentation results using the original and modified algorithm. The segmentation error was calculated using ground truth images. The average error decreases from 7.12% using the original method to 6.87% after modifying the test.

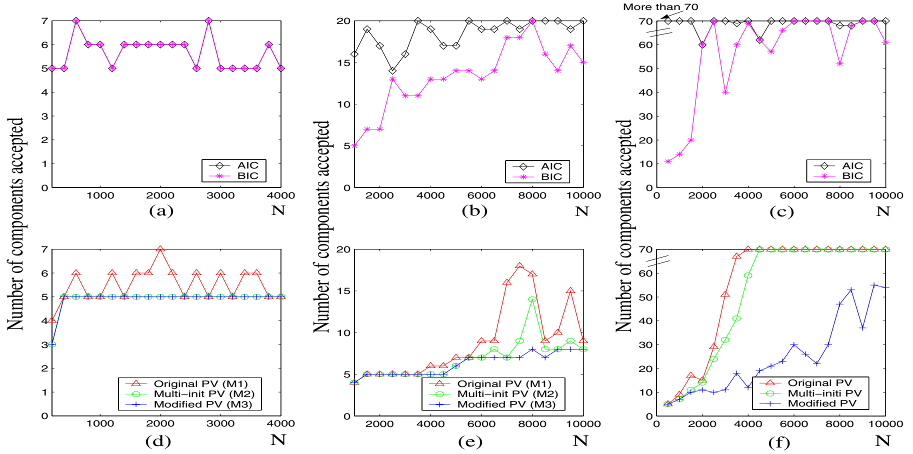
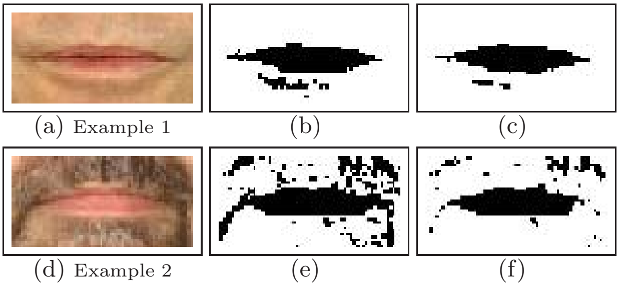


Fig. 3. The number of components accepted versus the number of samples using (top) AIC and BIC, (below) the original, multi-initialised and modified *PV* methods.(left: `class5` data set, middle: A face data set, right: Lip area data set)

6 Conclusions

In this paper we modified our proposed `Predictive Validation` algorithm in order to eliminate underlying problems of the model validation test for a large number of test points or very complex Gaussian mixture model. We demonstrated that `F` test avoids uncontrolled growth of the model complexity when more complex models do not improve the model calibration. It was also demonstrated that the *PV* technique is quite useful for dealing with the problem of model initialisation.

Even using the modified test, when we are dealing with a huge data set to avoid computational complexity of the *PV* test, it is desirable to place the vali-



(a) Example 1 (b) (c)

(d) Example 2 (e) (f)

Fig. 4. *(Left:)*Two examples of the rectangular blocks taken from the xm2vts database images. *(Middle:)* The segmentation results using the original method.*(Right:)* The segmentation results using the modified method dation windows over a sub-samples of the data set. The effective selection of the number of windows is a matter of interest in the future works.

Acknowledgements

The financial support from the EU project Banca and from the Ministry of Science, Research and Technology of Iran is gratefully acknowledged.

References

1. H. Akaike. A new look at the statistical model identification. *IEEE trans. on Automatic Control*, AC-19(6):716–723, 1974. 416
2. A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977. 416
3. J. Kittler, K. Messer, and M. Sadeghi. Model validation for model selection. In S. Singh, N. Murshed, and W. Kropatsch, editors, *Proceedings of International Conference on Advances in Pattern Recognition ICAPR 2001*, pages 240–249, 11-14 March 2001. 414, 416, 417, 418
4. K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. Xm2vtsdb: The extended m2vts database. In *Second International Conference on Audio and Video-based Biometric Person Authentication*, March 1999. 421
5. W. Press, B. Flanney, S. Teukolsky, and W. Vetterling. *Numerical Recipes in C: The Art of Scientific Computing.* Cambridge University Press, Cambridge, 2nd edition, 1992. 417, 420
6. M. Sadeghi, J. Kittler, and K. Messer. Segmentation of lip pixels for lip tracker initialisation. In *Proceedings IEEE International Conference on Image Processing, ICIP2001*, volume I, pages 50–53, 7-10 October 2001. 421
7. L. Sardo and J. Kittler. Model complexity validation for pdf estimation using gaussian mixtures. In S. V. A.K. Jain and B. Lovell, editors, *International Conference on Pattern Recognition*, pages 195–197, 1998. 414, 418
8. G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978. 416