Algorithms for Gaussian Bandwidth Selection in Kernel Density Estimators

José Miguel Leiva Murillo and Antonio Artés Rodríguez
Department of Signal Theory and Communications,
Universidad Carlos III de Madrid
E-mail: {leiva,antonio}@ieee.org.

Abstract

In this paper we study the classical statistical problem of choosing an appropriate bandwidth for Kernel Density Estimators. For the special case of Gaussian kernel, two algorithms are proposed for the spherical covariance matrix and for the general case, respectively. These methods avoid the unsatisfactory procedure of tuning the bandwidth while evaluating the likelihood, which is impractical with multivariate data in the general case. The convergence conditions are provided together with the algorithms proposed. We measure the accuracy of the models obtained by a set of classification experiments.

1 Introduction

A Kernel Density Estimator (KDE) is a non-parametric Probability Density Function (PDF) model that consists of a linear combination of kernel functions centered on the training data $\{\mathbf{x}_i\}_{i=1,\ldots,N}$, i.e.:

$$\hat{p}_\theta(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} k_\theta(\mathbf{x}-\mathbf{x}_i) \tag{1}$$

where $k_\theta(\mathbf{x})$ is the kernel function, which must be unitary, i.e.: $\int k_\theta(\mathbf{x})d\mathbf{x} = 1$ and $\mathbf{x} \in \mathcal{R}^D$. Although the KDEs are commonly considered as non-parametric models, the kernel function is characterized by a bandwidth that determines the accuracy of the model: $\hat{p}_\theta(\mathbf{x}) = \hat{p}(\mathbf{x}|\theta)$. Kernels too narrow of wide lead to overfitted or underfitted models, respectively.

Classical bandwidth selection methods have mainly focused on the unidimensional case. In [1], some first and second generation methods are compilated. Some examples of first generation criteria are the Mean Square Error (MSE), the Mean Integrated Squared Error (MISE), and the asymptotical MISE (AMISE) [1] [2]. Second generation methods include plug-in techniques and bootstrap methods. Kullback-Leibler divergence has also been considered [3].

We are interested in the Maximum-Likelihood (ML) criterion. Cross-validation allows us to apply the ML criterion so that a model built from $N-1$ samples is evaluated on the point left. The model evaluated on each training sample has the form:

$$\hat{p}_\theta(\mathbf{x}_i) = \frac{1}{N-1}\sum_{\substack{j=1\\ j\neq i}}^{N} G(\mathbf{x}_i - \mathbf{x}_j|\boldsymbol{\theta}) \tag{2}$$

where we make explicit the use of a Gaussian kernel. This framework was first proposed in [4] and later studied by other authors [2], [5]. However, these studies lack a closed optimization procedure, so that the bandwidth σ^2 is obtained by a greedy tuning along its possible values. Besides, the multivariate case is only considered in these previous works under a spherical kernel assumption. In this paper, proposed two algorithms that overcome these difficulties.

In a multidimensional Gaussian kernel, the set of parameters consists of the covariance matrix of the Gaussian. In the following, we consider two different degrees of complexity assumed for this matrix: a spherical shape, so that $\mathbf{C} = \sigma^2\mathbf{I}_D$ -only one parameter to adjust-, and an unconstrained kernel, in which a general form is considered for $\mathbf{C}$ with $D(D+1)/2$ parameters.

Sections 2 and 3 describe the bandwidth optimization for the both cases mentioned as presented in [6] and establish their convergence conditions. Some classification experiments are presented in Section 4 to measure the accuracy of the models. Section 5 closes the paper with the most important conclusions.

2 The spherical case

The expression for the kernel function is, for the spherical case:

$$G_{ij}(\sigma^2) = G(\mathbf{x}_i - \mathbf{x}_j|\sigma^2) = (2\pi)^{-D/2}\sigma^{-D}\exp\left(-\frac{1}{2\sigma^2}\|\mathbf{x}_i - \mathbf{x}_j\|^2\right)$$

We want to find the σ that maximizes the log-likelihood $\log L(\mathbf{X}|\sigma^2) = \sum_i \log \hat{p}_\theta(\mathbf{x}_i)$. The derivative of this likelihood is:

$$\nabla_\sigma \log L(\mathbf{X}|\sigma^2) = \frac{1}{N-1}\sum_i \frac{1}{\hat{p}(\mathbf{x}_i)}\sum_{j\neq i}\left(\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma^3} - \frac{D}{\sigma}\right)G_{ij}(\sigma^2)$$

We now search for the point that makes the derivative null:

$$\sum_i \frac{1}{\hat{p}(\mathbf{x}_i)}\sum_{j\neq i}\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma^3}G_{ij}(\sigma^2) = \sum_i \frac{1}{\hat{p}(\mathbf{x}_i)}\frac{D}{\sigma}\sum_{j\neq i}G_{ij}(\sigma^2) = \frac{N(N-1)D}{\sigma}$$

The second equality has been obtained by the fact that, by definition, $\sum_{j\neq i} G_{ij} = (N-1)\hat{p}(\mathbf{x}_i)$. Then we obtain the following fixed-point algorithm:

$$\sigma^2_{t+1} = \frac{1}{N(N-1)D}\sum_i \frac{1}{\hat{p}_t(\mathbf{x}_i)}\sum_{j\neq i}\|\mathbf{x}_i - \mathbf{x}_j\|^2 G_{ij}(\sigma^2_t) \tag{3}$$

where $\hat{p}_t$ denotes the KDE obtained in iteration t, i.e. the one that makes use of the width σ^2_t.

We prove the convergence of the algorithm in (3) by means of the following convergence theorem:

Theorem 1 *There is a fixed point in the interval* $(\frac{\overline{d^2_{NN}}}{D}, \frac{2\,\mathrm{tr}\{\boldsymbol{\Sigma}_{\boldsymbol{x}}\}}{D})$, *being* $\overline{d^2_{NN}}$ *the mean quadratic distance to the nearest neighbor and* $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ *the covariance matrix of* $\mathbf{x}$. *Besides, the fixed point is unique and the algorithm converges to it in the mentioned interval if the following condition holds:*

$$\frac{1}{2\sigma^4 N(N-1)^2 D}\sum_i \frac{1}{\hat{p}_l(\mathbf{x}_i)^2}\sum_{j\neq i}\sum_{k\neq i,j}(d_{ij}^2-d_{ik}^2)^2\exp(-\frac{d_{ij}^2+d_{ik}^2}{2\sigma^2})<1 \tag{4}$$

Proof 1 *Let* $\sigma^2 = g(\sigma^2)$ *the function in* (3), *whose fixed point is to be obtained. The proof of the fixed point existence is based on the search of an interval* (a, b) *such that* $a < g(\sigma^2) < b$ *if* $\sigma^2 \in (a, b)$.

In order to demonstrate that the interval $(\frac{\overline{d^2_{NN}}}{D}, \frac{2\,\mathrm{tr}\{\boldsymbol{\Sigma}_{\boldsymbol{x}}\}}{D})$ *holds that condition, we need to prove these three facts:*

1. $g(\frac{\overline{d^2_{NN}}}{D}) > \frac{\overline{d^2_{NN}}}{D}$
2. $g(\frac{2\,\mathrm{tr}\{\boldsymbol{\Sigma}_{\boldsymbol{x}}\}}{D}) < \frac{2\,\mathrm{tr}\{\boldsymbol{\Sigma}_{\boldsymbol{x}}\}}{D}$
3. $g(\sigma^2)$ *is monotonic in the interval.*

This way we are guaranteed that there is at least one crossing point between the function $g(\sigma^2)$ *and the line* $g(\sigma^2) = \sigma^2$.

To prove the first point, we rewrite (3) *as:*

$$g(\sigma^2) = \frac{1}{ND}\sum_i\sum_{j\neq i} d_{ij}^2 \frac{1}{1+\sum_{k\neq i,j}\exp(\frac{d_{ij}^2-d_{ik}^2}{2\sigma^2})} \tag{5}$$

The limit at 0 *is given by:*

$$\lim_{\sigma^2\to 0} f(\sigma^2) = \frac{1}{ND}\sum_i \min_{j\neq i} d_{ij}^2 = \frac{\overline{d^2_{NN}}}{D}$$

because the elements in the denominator of (5) *are null with the exception of the cases in which* $d_{ij}^2 < d_{ik}^2$, $\forall k \neq j$. *The first point is already proved because* $\frac{\overline{d^2_{NN}}}{D}$ *is the minimum value that* $g(\sigma^2)$ *can reach.*

To prove the second point, we take the limit at infinite:

$$\lim_{\sigma^2\to\infty} g(\sigma^2) = \lim_{\sigma^2\to\infty}\frac{1}{ND}\sum_i\sum_{j\neq i} d_{ij}^2 \frac{\exp(-\frac{d_{ij}^2}{2\sigma^2})}{\sum_{k\neq i}\exp(-\frac{d_{ik}^2}{2\sigma^2})} = \frac{1}{ND}\sum_i\sum_{j\neq i} d_{ij}^2\frac{1}{N-1}$$

The sum of distances may be expressed in terms of expectation, as:

$$\frac{1}{N(N-1)}\sum_i\sum_{j\neq i} d_{ij}^2 = E_{i,j}\{(\mathbf{x}_i-\mathbf{x}_j)^T(\mathbf{x}_i-\mathbf{x}_j)\} = 2E_i\{\mathbf{x}_i^T\mathbf{x}_i\} - 2\boldsymbol{\mu}_{\boldsymbol{x}}{}^T\boldsymbol{\mu}_{\boldsymbol{x}}$$

According to a property of linear algebra, if $\boldsymbol{\mu}_{\boldsymbol{x}} = E_{\mathbf{x}}\{\mathbf{x}\}$ *and* $\boldsymbol{\Sigma}_{\boldsymbol{x}} = E_{\mathbf{x}}\{\mathbf{x}\mathbf{x}^T\} - \boldsymbol{\mu}_{\boldsymbol{x}}\boldsymbol{\mu}_{\boldsymbol{x}}{}^T$, *then* $E\{\mathbf{x}^T\mathbf{x}\} = \mathrm{tr}\{\boldsymbol{\Sigma}_{\boldsymbol{x}}\} + \boldsymbol{\mu}_{\boldsymbol{x}}{}^T\boldsymbol{\mu}_{\boldsymbol{x}}$.

$$2E\{\mathbf{x}^T\mathbf{x}\} - 2\boldsymbol{\mu}_{\boldsymbol{x}}{}^T\boldsymbol{\mu}_{\boldsymbol{x}} = 2\,\mathrm{tr}\{\boldsymbol{\Sigma}_{\boldsymbol{x}}\}$$

We obtain:

$$\lim_{\sigma^2\to\infty} g(\sigma^2) = \frac{2\,\mathrm{tr}\{\boldsymbol{\Sigma}_{\boldsymbol{x}}\}}{D}$$

The second point is then proved since the maximum value of $g(\sigma^2)$ is reached at the infinite.

To demonstrate the last point, we compute the derivative of $g'(\sigma^2)$ and check out that it is positive:

$$\begin{aligned}\frac{dg(\sigma^2)}{d\sigma^2} &= \frac{1}{2\sigma^4 ND}\sum_i\sum_{j\neq i} d_{ij}^2 \frac{\sum_k (d_{ij}^2 - d_{ik}^2)\exp(-\frac{d_{ij}^2+d_{ik}^2}{2\sigma^2})}{(\sum_{l\neq i}\exp(-\frac{d_{il}^2}{2\sigma^2}))^2} \\ &= \frac{1}{2\sigma^4 N(N-1)^2 D}\sum_i \frac{1}{\hat{p}_l(\mathbf{x}_i)^2}\sum_{j\neq i}\sum_{k\neq i,j}(d_{ij}^2 - d_{ik}^2)^2\exp(-\frac{d_{ij}^2+d_{ik}^2}{2\sigma^2}) \geq 0\end{aligned} \tag{6}$$

The existence of a unique fixed point is then proved. To demonstrate the convergence of the algorithm in such interval, we need to check out the condition $|g'(\sigma^2)| < 1$ [7]. In that case, we are guaranteed that only a crossing point between $g(\sigma^2)$ and the line $g(\sigma^2) = \sigma^2$ exists. The convergence condition (4) *means that the value of* (6) *is lesser than* 1.

3 The unconstrained case

The general expression for a Gaussian kernel is:

$$G_{ij}(\mathbf{C}) = |2\pi\mathbf{C}|^{-1/2}\exp\left(-\frac{1}{2}(\mathbf{x}_i - \mathbf{x}_j)^T\mathbf{C}^{-1}(\mathbf{x}_i - \mathbf{x}_j)\right)$$

and its derivative w.r.t. $\mathbf{C}$:

$$\nabla_{\mathbf{C}} G_{ij}(\mathbf{C}) = \frac{1}{2}\left(\mathbf{C}^{-1}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T - \mathbf{I}\right)\mathbf{C}^{-1}G_{ij}(\mathbf{C})$$

As in the previous cases, we take the derivative of the log-likelihood and make it equal to zero:

$$\sum_i \frac{1}{\hat{p}(\mathbf{x}_i)}\frac{1}{N-1}\sum_{j\neq i}\frac{1}{2}\mathbf{C}^{-1}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T\mathbf{C}^{-1}G_{ij} = \sum_i \frac{1}{\hat{p}(\mathbf{x}_i)}\frac{1}{N-1}\sum_{j\neq i}\frac{1}{2}\mathbf{C}^{-1}G_{ij}$$

By multiplying both members by $\mathbf{C}$, both at the right and the left, we obtain:

$$\sum_i \frac{1}{\hat{p}(\mathbf{x}_i)}\sum_{j\neq i}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T G_{ij} = \mathbf{C}\sum_i \frac{1}{\hat{p}(\mathbf{x}_i)}\sum_{j\neq i} G_{ij}$$

After some simplifications as in the spherical case, we reach the following fixed-point algorithm:

$$\mathbf{C}_{t+1} = \frac{1}{N(N-1)}\sum_i \frac{1}{\hat{p}_t(\mathbf{x}_i)}\sum_{j\neq i}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T G_{ij}(\mathbf{C}_t) \tag{7}$$

The expression in (7) suggests a relationship with the Expectation-Maximization result for Gaussian Mixture Models (GMM). A GMM is a PDF estimator given by the expression $\hat{p}(\mathbf{x}) = \sum_{k=1}^{K} \alpha_k G(x|\boldsymbol{\mu}_k, \mathbf{C}_k)$. The weights of the K components of the mixture are given by the α_k, and each Gaussian is characterized by its mean vector $\boldsymbol{\mu}_k$ and its covariance matrix $\mathbf{C}_k$. The solution provided by the EM algorithm consists of an iterative procedure where the parameters at step t are obtained by the ones at step $t-1$. To do so, a matrix of auxiliary variables is used, $r_{ki} = p(k|\mathbf{x}_i)$, expressing the likelihood of the sample to belong to the k-th component of the mixture. These probabilities must hold $\sum_k r_{ki} = 1$. The EM solution establishes the following updating rule for the covariance matrix at step t:

$$\mathbf{C}_k^t = \sum_k \sum_i r_{ki}^t \frac{(\mathbf{x}_i - \boldsymbol{\mu}_k^t)(\mathbf{x}_i - \boldsymbol{\mu}_k^t)^T}{N} \tag{8}$$

where the r_{ki}^t and $\boldsymbol{\mu}_k^t$ are also iteratively updated. Note that our KDE model can be considered as a special case of GMM where i) there are as many mixtures as samples ($K = N$) with the same weights $\alpha_k = 1/N$; ii) mean vectors are fixed: $\boldsymbol{\mu}_k = \mathbf{x}_k$; iii) the covariance matrix is the same for each of the components, and iv) $r_{ki} = 0$ if $k = i$ and $r_{ki} = 1/(N-1)$ if $k \neq i$.

With these particularizations, the updating rule in (8) becomes equal to the one given by the iteration in (7).

The EM guarantees the monotonic increase of the likelihood cost and so its convergence to a local minimum, as proved in the literature [8]. The algorithm given in (7) is subject to the same conditions, so that its convergence is also proved. However, in situations in which $N \approx D$, empirical covariance matrices are close to singularity, so that numerical problems may arise as in GMM design.

4 Application to Parzen classification

We have tested the performance of the obtained models on a set of public classification problems from [9]. For doing so, we apply the Parzen classifier, which performs the simple Bayes criterion:

$$\hat{y} = \arg\max_l \hat{p}_{\theta_l}(\mathbf{x}|c_l)$$

with per-class spherical (S-KDE) and unconstrained (U-KDE) models $\hat{p}_{\theta_l}(\mathbf{x}|c_l)$ optimized according to the proposed method, being c_l each of the L classes considered. We have compared these results with the ones obtained by other classification methods such as K-Nearest-Neighbors (KNN, with K=1) and the one-versus-the-rest Support Vector Machine (SVM) with RBF kernel. The results are shown in

Data	Train	Test	L	D	S-KDE	U-KDE	KNN	SVM
Pima	738	-	2	8	71.22	75.13	73.18	76.47
Wine	178	-	3	13	75.84	99.44	76.97	100
Landsat	4435	2000	6	36	89.45	86.10	90.60	90.90
Optdigits	3823	1797	10	64	97.89	93.54	94.38	98.22
Letter	16000	4000	26	16	95.23	92.77	95.20	97.55

Table 1: Classification performance on some public datasets. Leave-one-out accuracy is provided when there are not test data.

Table 1. The most remarkable conclusion from this result is that either S-KDE or U-KDE provides, in each case, a classification performance close to SVM's. The comparison between S-KDE and U-KDE performance is closely related, in each case, to the dimension of the data when compared to the number of samples. In the datasets with higher dimensionality, the performance of S-KDE is higher due to its lower risk of overfitting. Parzen classifiers have not enjoyed the popularity of other methods, mainly due to the difficulty of obtaining a reliable bandwidth for the kernel. However, in this experiment we have shown how the bandwidth chosen by our algorithm provides a classification performance close to a state-of-the-art classifier such as the SVM.

5 Conclusions

We have presented two algorithms for the optimization of the likelihood in the bandwidth selection problem for KDE models. Unlike previous results in the literature, the methods tackle in a natural way the multivariate case, for which we provide solutions based on both spherical and complete (unconstrained) Gaussian kernel. The convergence conditions have been described for both algorithms. By a set of experiments, we have shown that the models obtained are accurate enough to provide good classification results. This demonstrates that the model do not overfit to the data, even in problems involving a high number of variables.

References

[1] M. C. Jones, J. S. Marron, and S. J. Sheather, "A brief survey of bandwidth selection for density estimation," *Journal of the American Statistical Association*, vol. 91, no. 433, pp. 401–407, 1996.

[2] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, Londres, 1986.

[3] A. Bowman, "An alternative method of cross-validation for the smoothing of density estimates," *Biometrika*, vol. 71, pp. 353–360, 1984.

[4] R. Duin, "On the choice of smoothing parameters for Parzen estimators of probability density functions," *IEEE Trans. on Computers*, vol. 25, no. 11, 1976.

[5] P. Hall, "Cross-validation in density estimation," *Biometrika*, vol. 69, no. 2, pp. 383–390, 1982.

[6] J.M. Leiva-Murillo and A. Artés-Rodríguez, "A fixed-point algorithm for finding the optimal covariance matrix in kernel density modeling," in *International Conference on Acoustic, Speech and Signal Processing*, Toulouse, Francia, 2006.

[7] R. Fletcher, *Practical Methods of Optimization (2nd Edition)*, John Wiley & Sons, New York, 1995.

[8] G. McLachlan, *The EM algorithm and extensions*, John Wiley & Sons, New York, 1997.

[9] D.J. Newman, S. Hettich, C.L. Blake, and C.J. Merz, "UCI repository of machine learning databases," Tech. Rep., Univ. of California, Dept. ICS, 1998.