Chapter 6

NONPARAMETRIC DENSITY ESTIMATION

So far we have been discussing the estimation of parameters. Thus, if we can assume we have a density function that can be characterized by a set of parameters, we can design a classifier using estimates of the parameters. Unfortunately, we often cannot assume a parametric form for the density function, and in order to apply the likelihood ratio test we somehow have to estimate the density functions using an unstructured approach. This type of approach is called *nonparametric estimation*, while the former is called *parametric estimation*. Since, in nonparametric approaches, the density function is estimated locally by a small number of neighboring samples, the estimate is far less reliable with larger bias and variance than the parametric counterpart.

There are two kinds of nonparametric estimation techniques available: one is called the *Parzen density estimate* and the other is the *k-nearest neighbor density estimate*. They are fundamentally very similar, but exhibit some different statistical properties. Both are discussed in this chapter.

It is extremely difficult to obtain an accurate density estimate nonparametrically, particularly in high-dimensional spaces. However, our goal here is not to get an accurate estimate. Our goal is, by using these estimates, to design a classifier and evaluate its performance. For this reason, the accuracy of the estimate is not necessarily a crucial issue. Classification and performance evaluation will be discussed in Chapter 7. The intention of this chapter is to make the reader familiar with the fundamental mathematical properties related to nonparametric density estimation in preparation for the material presented in Chapter 7.

6.1 Parzen Density Estimate

Parzen Density Estimate

In order to estimate the value of a density function at a point X, we may set up a small *local region* around X, $L(X)$. Then, the *probability coverage* (or *probability mass*) of $L(X)$ may be approximated by $p(X)v$ where v is the volume of $L(X)$. This probability may be estimated by drawing a large number of samples, N, from $p(X)$, counting the number of samples, k, falling in $L(X)$, and computing k/N. Equating these two probabilities, we may obtain an estimate of the density function as

$$\hat{\mathbf{p}}(X)v = \frac{\mathbf{k}(X)}{N} \quad or \quad \hat{\mathbf{p}}(X) = \frac{\mathbf{k}(X)}{Nv} . \tag{6.1}$$

Note that, with a fixed v, $\mathbf{k}$ is a random variable and is dependent on X. A fixed v does not imply the same v throughout the entire space, and v could still vary with X. However, v is a preset value and is not a random variable.

Kernel expression: The estimate of (6.1) has another interpretation. Suppose that 3 samples, X_3, X_4, and X_5, are found in $L(X)$ as shown in Fig. 6-1. With v and N given, $\hat{p}(X)$ becomes $3/Nv$. On the other hand, if we set up a uniform *kernel function*, $\kappa(\cdot)$, with volume v and height $1/v$ around all existing samples, the average of the values of these kernel functions at X is also $3/Nv$. That is, [1-4]

$$\hat{\mathbf{p}}(X) = \frac{1}{N} \sum_{i=1}^{N} \kappa(X - \mathbf{X}_i) . \tag{6.2}$$

As seen in Fig. 6-1, only the kernel functions around the 3 samples, X_3, X_4, and X_5, contribute to the summation of (6.2).

Once (6.2) is adopted, the shape of the kernel function could be selected more freely, under the condition $\int \kappa(X)\, dX = 1$. For one-dimensional cases, we may seek optimality and select a complex shape. However, in a high-dimensional space, because of its complexity, the practical selection of the ker-

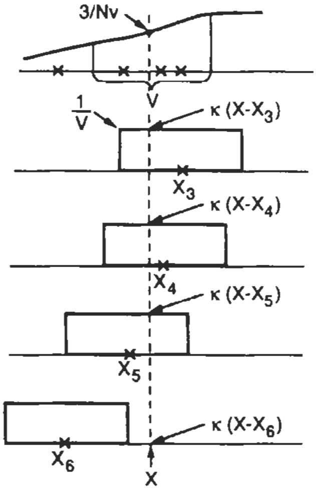

Fig. 6-1 Parzen kernel density estimate.

nel function is very limited to either a normal or uniform kernel. In this book, we will use the following kernel which includes both normal and uniform kernels as special cases:

$$\kappa(X) = \frac{m\Gamma(\frac{n}{2})\Gamma^{n/2}(\frac{n+2}{2m})}{(n\pi)^{n/2}\Gamma^{n/2+1}(\frac{n}{2m})} \times \frac{1}{r^n |A|^{1/2}}$$

$$\times \exp\left[-\left\{\frac{\Gamma(\frac{n+2}{2m})}{n\Gamma(\frac{n}{2m})} X^T(r^2A)^{-1}X\right\}^m\right], \qquad (6.3)$$

where $\Gamma(\cdot)$ is the gamma function, and m is a parameter determining the shape of the kernel. It may be verified that, for any value of m, the covariance matrix of the kernel density (6.3) is r^2A. The parameter m determines the rate at which the kernel function drops off. For $m = 1$, (6.3) reduces to a simple normal kernel. As m becomes large, (6.3) approaches a uniform (hyperelliptical) kernel, always with a smooth roll-off. The matrix A determines the shape of the hyperellipsoid, and r controls the size or volume of the kernel. Other coefficients are selected to satisfy the two conditions mentioned previously:

$\int \kappa(X)dX = 1$ and $\Sigma_\kappa = r^2 A$ where Σ_κ is the covariance matrix of $\kappa(X)$.

Convolution expression: Equation (6.2) can be rewritten in convolution form as

$$\hat{\mathbf{p}}(X) = \hat{\mathbf{p}}_s(X) * \kappa(X) \triangleq \int \hat{\mathbf{p}}_s(Y)\kappa(X-Y)dY \ , \tag{6.4}$$

where $\hat{\mathbf{p}}_s$ is an impulsive density function with impulses at the locations of existing N samples.

$$\hat{\mathbf{p}}_s(Y) = \frac{1}{N}\sum_{i=1}^{N}\delta(Y-\mathbf{X}_i) \ . \tag{6.5}$$

That is, the estimated density $\hat{\mathbf{p}}(X)$ is obtained by feeding $\hat{\mathbf{p}}_s(X)$ through a linear (noncausal) filter whose impulse response is given by $\kappa(X)$. Therefore, $\hat{\mathbf{p}}(X)$ is a smoothed version of $\hat{\mathbf{p}}_s(X)$.

Moments of $\hat{\mathbf{p}}(X)$: The first and second order moments of (6.4) can be easily computed. First, let us compute the expected value of $\hat{\mathbf{p}}_s(X)$ as

$$E\{\hat{\mathbf{p}}_s(X)\} = \frac{1}{N}\sum_{i=1}^{N}\int\delta(X-Z)p(Z)dZ = \frac{1}{N}\sum_{i=1}^{N}p(X) = p(X) \ . \tag{6.6}$$

That is, $\hat{\mathbf{p}}_s(X)$ is an unbiased estimate of $p(X)$. Then, the expected value of $\hat{\mathbf{p}}(X)$ of (6.4) may be computed as

$$\begin{aligned} E\{\hat{\mathbf{p}}(X)\} &= \int E\{\hat{\mathbf{p}}_s(Y)\}\kappa(X-Y)dY \\ &= \int p(Y)\kappa(X-Y)dY = p(X)*\kappa(X) \ . \end{aligned} \tag{6.7}$$

Also,

$$\begin{aligned} E\{\hat{\mathbf{p}}^2(X)\} = \frac{1}{N^2}\Bigg[&\sum_{i=1}^{N}\int\kappa^2(X-Z)p(Z)dZ \\ &+ \sum_{\substack{i=1\\ i\neq j}}^{N}\sum_{j=1}^{N}\iint\kappa(X-Y)\kappa(X-Z)p(Y)p(Z)dYdZ\Bigg] \end{aligned}$$

$$= \frac{1}{N} p(X) * \kappa^2(X) + (1 - \frac{1}{N})[p(X) * \kappa(X)]^2 . \tag{6.8}$$

Therefore, the variance of $\hat{\mathbf{p}}(X)$ is

$$\text{Var}\{\hat{\mathbf{p}}(X)\} = \frac{1}{N}[p(X) * \kappa^2(X) - [p(X) * \kappa(X)]^2] . \tag{6.9}$$

Approximations of moments: In order to approximate the moments of $\hat{\mathbf{p}}(X)$, let us expand $p(Y)$ around X by a Taylor series up to the second order terms as

$$p(Y) \cong p(X) + \nabla p^T(X)(Y-X) + \frac{1}{2}\text{tr}\{\nabla^2 p(X)(Y-X)(Y-X)^T\} . \tag{6.10}$$

Then, $p(X) * \kappa(X)$ may be approximated by

$$\begin{aligned} p(X) * \kappa(X) &= \int p(Y)\kappa(Y-X)dY \\ &\cong p(X)\int \kappa(Y-X)dY \\ &\quad + \frac{1}{2}\text{tr}\{\nabla^2 p(X)\int (Y-X)(Y-X)^T \kappa(Y-X)dY\} , \end{aligned} \tag{6.11}$$

where the first order term disappears because $\kappa(\cdot)$ is a symmetric function. Since $\int \kappa(Y-X)dY = 1$ and $\int (Y-X)(Y-X)^T \kappa(Y-X)dY = r^2 A$ for $\kappa(\cdot)$ of (6.3), (6.11) can be expressed by

$$p(X) * \kappa(X) \cong p(X)[1 + \frac{1}{2}\alpha(X) r^2] , \tag{6.12}$$

where

$$\alpha(X) = \text{tr}\left\{\frac{\nabla^2 p(X)}{p(X)} A\right\} . \tag{6.13}$$

Similarly,

$$\begin{aligned} p(X) * \kappa^2(X) &\cong p(X)\int \kappa^2(Y-X)dY \\ &\quad + \frac{1}{2}\text{tr}\{\nabla^2 p(X)\int (Y-X)(Y-X)\kappa^2(Y-X)dY\} . \end{aligned} \tag{6.14}$$

Although $\kappa(\cdot)$ is a density function, $\kappa^2(\cdot)$ is not. Therefore, $\int \kappa^2(Y)dY$ has a value not equal to 1. Let

$$w = \int \kappa^2(Y)dY \; . \tag{6.15}$$

Then, $\kappa^2(\cdot)/w$ becomes a density function. Therefore, (6.14) becomes

$$p(X)*\kappa^2(X) \cong wp(X) + \frac{w}{2}\mathrm{tr}\{\nabla^2 p(X)\int (Y-X)(Y-X)^T \frac{\kappa^2(Y-X)}{w} dY\}$$

$$= wp(X)[1 + \frac{1}{2}\beta(X)r^2] \; , \tag{6.16}$$

where

$$\beta(X) = \mathrm{tr}\left\{\frac{\nabla^2 p(X)}{p(X)} B\right\} \tag{6.17}$$

and r^2B is the covariance matrix of $\kappa^2(X)/w$.

Substituting (6.12) and (6.16) into (6.7) and (6.9), the moments of $\hat{\mathbf{p}}(X)$ are approximated by

$$E\{\hat{\mathbf{p}}(X)\} \cong p(X)[1 + \frac{1}{2}\alpha(X)r^2] \qquad 2nd \text{ order approximation}$$

$$\cong p(X) \qquad 1st \text{ order approximation} \; , \tag{6.18}$$

$$\mathrm{Var}\{\hat{\mathbf{p}}(X)\} \cong \frac{1}{N}[wp(X)\{1 + \frac{1}{2}\beta(X)r^2\} - p^2(X)\{1 + \frac{1}{2}\alpha(X)r^2\}^2]$$

$$2nd \text{ order approximation}$$

$$\cong \frac{1}{N}[wp(X) - p^2(X)] \qquad 1st \text{ order approximation} \; . \tag{6.19}$$

Note that the variance is proportional to $1/N$ and thus can be reduced by increasing the sample size. On the other hand, the bias is independent of N, and is determined by $\nabla^2 p(X)$, A, and r^2.

Normal kernel: When the kernel function is normal with zero expected vector and covariance matrix r^2A, $N_X(0,r^2A)$, $\kappa^2(X)$ becomes normal as $cN_X(0,r^2A/2)$ where $c = 2^{-n/2}(2\pi)^{-n/2}|A|^{-1/2}r^{-n}$. Therefore,

$$w = \frac{1}{2^{n/2}(2\pi)^{n/2}|A|^{1/2}r^n} , \tag{6.20}$$

$$\beta(X) = \frac{1}{2}\alpha(X) . \tag{6.21}$$

Uniform kernel: For a uniform kernel with the covariance matrix r^2A,

$$\kappa(Y) = \begin{cases} 1/v & \textit{inside } L(X) \\ 0 & \textit{outside } L(X) . \end{cases} \tag{6.22}$$

where

$$L(X) = \{Y: d(Y,X) \le r\sqrt{n+2}\} , \tag{6.23}$$

$$d^2(Y,X) = (Y-X)^T A^{-1}(Y-X) , \tag{6.24}$$

and

$$v = \int_{L(X)} dY = \frac{\pi^{n/2}}{\Gamma(\frac{n+2}{2})} |A|^{1/2}(r\sqrt{n+2})^n . \tag{6.25}$$

Then, $\kappa^2(X)$ is also uniform in $L(X)$ with the height $1/v^2$. Therefore,

$$w = \int_{L(X)} \kappa^2(Y)dY = \frac{1}{v} . \tag{6.26}$$

Also, since the covariance matrix of $\kappa(X)$ is r^2A, the covariance matrix of $\kappa^2(X)/w$ is also r^2A as

$$\int_{L(X)} (Y-X)(Y-X)^T \frac{1}{v} dY = r^2 A . \tag{6.27}$$

Therefore, for the uniform distribution of (6.22),

$$B = A \quad \text{and} \quad \beta(X) = \alpha(X) . \tag{6.28}$$

Note that w's for both normal and uniform kernels are proportional to r^{-n} or v^{-1}. In particular, $w = 1/v$ for the uniform kernel from (6.26). Using this relation, the first order approximation of the variance can be simplified further as follows:

$$\operatorname{Var}\{\hat{\mathbf{p}}(X)\} \cong \frac{1}{N}\left[\frac{p(X)}{v} - p^2(X)\right]$$

$$= p^2(X)\left[\frac{1}{Nvp(X)} - \frac{1}{N}\right] \cong p^2(X)\left[\frac{1}{k} - \frac{1}{N}\right]$$

$$\cong \frac{p^2(X)}{k} , \tag{6.29}$$

where $p \cong k/Nv$ and $N \gg k$ are used. This suggests that the second term of (6.19) is much smaller than the first term, and can be ignored. Also, (6.29) indicates that $k \rightarrow \infty$ is required along with $N \rightarrow \infty$ for the Parzen density estimate to be consistent. These are the known conditions for asymptotic unbiasness and consistency [2].

Convolution of normal distributions: If $p(X)$ is assumed to be normal and a normal kernel is selected for $\kappa(X)$, (6.7) and (6.9) become trivial to evaluate. When two normal densities $N_X(0,A)$ and $N_X(0,B)$ are convolved, the result is also a normal density of $N_X(0,K)$, where

$$K^{-1} = B^{-1} - B^{-1}(B^{-1} + A^{-1})^{-1}B^{-1}$$

$$= A^{-1} - A^{-1}(A^{-1} + B^{-1})^{-1}A^{-1} . \tag{6.30}$$

In particular, if $A = \Sigma$ and $B = r^2\Sigma$

$$K = (1 + r^2)\Sigma . \tag{6.31}$$

Optimal Kernel Size

Mean-square error criterion: In order to apply the density estimate of (6.1) (or (6.2) with the kernel function of (6.3)), we need to select a value for r [5-11]. The optimal value of r may be determined by minimizing the *mean-square error* between $\hat{\mathbf{p}}(X)$ and $p(X)$ with respect to r.

$$MSE\{\hat{\mathbf{p}}(X)\} = E\{[\hat{\mathbf{p}}(X) - p(X)]^2\} . \tag{6.32}$$

This criterion is a function of X, and thus the optimal r also must be a function of X. In order to make the optimal r independent of X, we may use the *integral mean-square error*

$$IMSE = \int MSE\{\hat{\mathbf{p}}(X)\}dX \ . \tag{6.33}$$

Another possible criterion to obtain the globally optimal r is $E_X\{MSE\{\hat{\mathbf{p}}(\mathbf{X})\}\} = \int MSE\{\hat{\mathbf{p}}(X)\}p(X)dX$. The optimization of this criterion can be carried out in a similar way as the *IMSE*, and produces a similar but a slightly smaller r than the *IMSE*. This criterion places more weight on the *MSE* in high density areas, where the locally optimal r's tend to be smaller.

Since we have computed the bias and variance of $\hat{\mathbf{p}}(X)$ in (6.18) and (6.19), $MSE\{\hat{\mathbf{p}}(X)\}$ may be expressed as

$$MSE\{\hat{\mathbf{p}}(X)\} = [E\{\hat{\mathbf{p}}(X)\} - p(X)]^2 + \mathrm{Var}\{\hat{\mathbf{p}}(X)\} \ . \tag{6.34}$$

In this section, only the **uniform kernel function** is considered. This is because the Parzen density estimate with the uniform kernel is more directly related to the k nearest neighbor density estimate, and the comparison of these two is easier. Since both normal and uniform kernels share similar first and second order moments of $\hat{\mathbf{p}}(X)$, the normal kernel function may be treated in the same way as the uniform kernel, and both produce similar results.

When the first order approximation is used, $\hat{\mathbf{p}}(X)$ is unbiased as in (6.18), and therefore $MSE = \mathrm{Var} = p/Nv - p^2/N$ as in (6.29). This criterion value is minimized by selecting $v = \infty$ for a given N and p. That is, as long as the density function is linear in $L(X)$, the variance dominates the *MSE* of the density estimate, and can be reduced by selecting larger v. However, as soon as $L(X)$ is expanded and picks up the second order term of (6.10), the bias starts to appear in the *MSE* and it grows with r^2 (or $v^{2/n}$) as in (6.18). Therefore, in minimizing the *MSE*, we select the best compromise between the bias and the variance. In order to include the effect of the bias in our discussion, we have no choice but to select the second order approximation in (6.18). Otherwise, the *MSE* criterion does not depend on the bias term. On the other hand, the variance term is included in the *MSE* no matter which approximation of (6.19) is used, the first or second order. If the second order approximation is used, the accuracy of the variance may be improved. However, the degree of improvement may not warrant the extra complexity which the second order approximation brings in. Furthermore, it should be remembered that the optimal r will be a function of $p(X)$. Since we never know the true value of

$p(X)$ accurately, it is futile to seek the more accurate but more complex expression for the variance. After all, what we can hope for is to get a rough estimate of r to be used.

Therefore, using the second order approximation of (6.18) and the first order approximation of (6.29) for simplicity,

$$MSE\{\hat{\mathbf{p}}(X)\} \cong \frac{p(X)}{Nv} + \frac{1}{4}\alpha^2(X)p^2(X)r^4 \ . \tag{6.35}$$

Note that the first and second terms correspond to the variance and squared bias of $\hat{\mathbf{p}}(X)$, respectively.

Minimization of MSE: Solving $\partial MSE/\partial r = 0$ [5], the resulting optimal r, r^*, is

$$\begin{aligned} r^*(X) &= \left[\frac{n}{c\alpha^2 p}\right]^{\frac{1}{n+4}} \times N^{-\frac{1}{n+4}} \\ &= \left[\frac{n\Gamma(\frac{n+2}{2})}{\pi^{1/2}(n+2)^{n/2}p\,|A|^{1/2}\alpha^2}\right]^{\frac{1}{n+4}} \times N^{-\frac{1}{n+4}} \ , \end{aligned} \tag{6.36}$$

where $v = cr^n$ and

$$c = \frac{\pi^{n/2}(n+2)^{n/2}\,|A|^{1/2}}{\Gamma(\frac{n+2}{2})} \ . \tag{6.37}$$

The resulting mean-square error is obtained by substituting (6.36) into (6.35).

$$MSE^*\{\hat{\mathbf{p}}(X)\} = \frac{n+4}{4}\left[\frac{\Gamma^{4/n}(\frac{n+2}{2})p^{2+4/n}\alpha^2}{n(n+2)^2\pi^2\,|A|^{2/n}}\right]^{\frac{n}{n+4}} \times N^{-\frac{4}{n+4}} \ . \tag{6.38}$$

When the integral mean-square error of (6.33) is computed, v and r are supposed to be constant, being independent of X. Therefore, from (6.35)

$$IMSE = \frac{1}{Nv}\int p(X)dX + \frac{1}{4}r^4\int\alpha^2(X)p^2(X)dX$$

$$= \frac{1}{Nv} + \frac{1}{4} r^4 \int \alpha^2(X) p^2(X) dX \; . \tag{6.39}$$

Again, by solving $\partial IMSE / \partial r = 0$ [5],

$$r^* = \left[\frac{n}{c \int \alpha^2(X) p^2(X) dX} \right]^{\frac{1}{n+4}} \times N^{-\frac{1}{n+4}}$$

$$= \left[\frac{n \Gamma(\frac{n+2}{2})}{\pi^{n/2} (n+2)^{n/2} |A|^{1/2} \int \alpha^2(X) p^2(X) dX} \right]^{\frac{1}{n+4}} \times N^{-\frac{1}{n+4}} \; . \tag{6.40}$$

The resulting criterion value is obtained by substituting (6.40) into (6.39),

$$IMSE^* = \frac{n+4}{4} \left[\frac{\Gamma^{4/n}(\frac{n+2}{2}) \int \alpha^2(X) p^2(X) dX}{n (n+2)^2 \pi^2 |A|^{2/n}} \right]^{\frac{n}{n+4}} \times N^{-\frac{4}{n+4}} \; . \tag{6.41}$$

Optimal Metric

Another important question in obtaining a good density estimate is how to select the metric, A of (6.3). The discussion of the optimal A is very complex unless the matrix is diagonalized. Therefore, we first need to study the effect of linear transformations on the various functions used in the previous sections.

Linear transformation: Let Φ be a non-singular matrix used to define a linear transformation. This transformation consists of a rotation and a scale change of the coordinate system. Under the transformation, a vector and metric become

$$Z = \Phi^T X \; , \tag{6.42}$$

$$A_Z = \Phi^T A_X \Phi \; . \tag{6.43}$$

The distance of (6.24) is invariant since

$$(Y-X)^T A_X^{-1} (Y-X) = (W-Z)^T A_Z^{-1} (W-Z) \; , \tag{6.44}$$

where $W = \Phi^T Y$. The following is the list of effects of this transformation on various functions. Proofs are not given but can be easily obtained by the reader.

(1) $p_Z(Z) = |\Phi|^{-1} p_X(X)$ [*Jacobian*] , (6.45)

(2) $\nabla^2 p_Z(Z) = |\Phi|^{-1} \Phi^{-1} \nabla^2 p_X(X) \Phi^{T^{-1}}$

[from (6.10),(6.42), and (6.45)] , (6.46)

(3) $r(Z) = r(X)$ [from (6.44)] , (6.47)

(4) $v(Z) = |\Phi| v(X)$ [from (6.25),(6.43), and (6.47)] , (6.48)

(5) $MSE\{\hat{\mathbf{p}}_Z(Z)\} = |\Phi|^{-2} MSE\{\hat{\mathbf{p}}_X(X)\}$ [from (6.32) and (6.45)] , (6.49)

(6) $IMSE_Z = |\Phi|^{-1} IMSE_X$ [from (6.33) and (6.42)] . (6.50)

Note that both *MSE* and *IMSE* depend on Φ. The mean-square error is a coordinate dependent criterion.

Minimization of IMSE: We will now use the above results to optimize the integral mean-square error criterion with respect to the matrix A. However, it is impossible to discuss the optimization for a general $p(X)$. We need to limit the functional form of $p(X)$. Here, we choose the following form for $p(X)$:

$$p(X) = |B|^{-1/2} g((X-M)^T B^{-1} (X-M)) , \tag{6.51}$$

where $g(\cdot)$ does not involve B or M. The $p(X)$ of (6.51) covers a large family of density functions including the ones in (6.3). The expected vector, M, can be assumed to be zero, since all results should be independent of a mean shift. Now, we still have the freedom to choose the matrix A in some optimum manner. We will manipulate the two matrices B and A to simultaneously diagonalize each, thus making the analysis easier. That is,

$$\Phi^T B \Phi = I \quad \text{and} \quad \Phi^T A \Phi = \Lambda \tag{6.52}$$

and

$$p(Z) = g(Z^T Z) , \tag{6.53}$$

where Λ is a diagonal matrix with components $\lambda_1, \ldots, \lambda_n$.

In the transformed Z-space, $IMSE_Z^*$ of (6.41) becomes

$$IMSE_Z^* = c_1 \left[c_2 \int \operatorname{tr}^2 \{ \nabla^2 p_Z(Z) \frac{\Lambda}{|\Lambda|^{1/n}} \} dZ \right]^{\frac{n}{n+4}} , \tag{6.54}$$

where c_1 and c_2 are positive constants. $IMSE_Z^*$ can be minimized by minimizing $\operatorname{tr}^2\{\cdot\}$ with respect to Λ. Since Λ is normalized by $|\Lambda|^{1/n}$ such that

$$\left| \frac{\Lambda}{|\Lambda|^{1/n}} \right| = 1 , \tag{6.55}$$

the scale of the matrix has no effect. Thus, we will minimize $\operatorname{tr}^2\{\cdot\}$ with respect to λ_i's with the constraint

$$|\Lambda| = \prod_{i=1}^{n} \lambda_i = 1 . \tag{6.56}$$

Now, $\operatorname{tr}\{\cdot\}$ can be evaluated as

$$\operatorname{tr}\{\nabla^2 p_Z(Z)\Lambda\} = \sum_{i=1}^{n} \lambda_i \frac{\partial^2 p_Z(Z)}{\partial z_i^2} = \theta \sum_{i=1}^{n} \lambda_i , \tag{6.57}$$

where

$$\theta = \frac{\partial^2 p_Z(Z)}{\partial z_i^2} = \frac{\partial}{\partial z_i} \left[\frac{dg(Z^TZ)}{d(Z^TZ)} \frac{\partial (Z^TZ)}{\partial z_i} \right] = 2 \frac{dg(Z^TZ)}{d(Z^TZ)} . \tag{6.58}$$

Thus, the criterion to be optimized is

$$\begin{aligned} J &= \operatorname{tr}^2\{\nabla^2 p_Z(Z)\Lambda\} - \mu(\prod_{i=1}^{n} \lambda_i - 1) \\ &= \theta^2 \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j - \mu(\prod_{i=1}^{n} \lambda_i - 1) , \end{aligned} \tag{6.59}$$

where μ is a Lagrange multiplier. Taking the derivative of J with respect to λ_k and setting the result equal to zero,

$$\frac{\partial J}{\partial \lambda_k} = \theta^2(\lambda_k + \sum_{i=1}^{n} \lambda_i) - \frac{\mu}{\lambda_k} = 0 \qquad (6.60)$$

or

$$\lambda_k^2 + \lambda_k(\sum_{i=1}^{n} \lambda_i) = \frac{\mu}{\theta^2} \qquad (k = 1, \ldots, n) . \qquad (6.61)$$

In order to satisfy (6.61), all λ_i's must be equal. Since $|\Lambda| = 1$, the solution of (6.61) must be

$$\Lambda = I . \qquad (6.62)$$

That is, in the transformed Z-space, the optimal matrix A_Z is I for $B_Z = I$. Therefore, the optimal matrix A to use in the original X-space is identical to B of (6.51) [5]. The neighborhoods should take the same ellipsoidal shape as the underlying distribution. For the normal distribution we see that the covariance matrix $B = \Sigma$ is indeed optimal for A.

It is important to notice that (6.62) is the locally optimal metric regardless of the location, because $IMSE^*$ of (6.54) is minimized not after but before taking the integration. The same result can be obtained by minimizing MSE^* of (6.38).

Normal Case

In order to get an idea of what kind of numbers should be used for r, in this section let us compute the optimal r for a normal distribution. The partial derivatives $\nabla p(X)$ and $\nabla^2 p(X)$ for $N_X(M,\Sigma)$ are

$$\nabla p(X) = - p(X)\Sigma^{-1}(X-M) , \qquad (6.63)$$

$$\nabla^2 p(X) = p(X)[\Sigma^{-1}(X-M)(X-M)^T\Sigma^{-1} - \Sigma^{-1}] . \qquad (6.64)$$

For the simplest case in which $M = 0$ and $\Sigma = I$,

$$\operatorname{tr}\{\nabla^2 p(X)\} = p(X)(X^TX - n) = p(X)(\sum_{i=1}^{n} x_i^2 - n) . \qquad (6.65)$$

Note that the optimal A is also I in this case. It is easy to show that, if $p(X) = N_X(0,I)$, then $p^2(X) = 2^{-n/2}(2\pi)^{-n/2}N_X(0,I/2)$. Therefore,

$$\int \text{tr}^2\{\nabla^2 p(X)\}dX = \frac{1}{2^{n/2}(2\pi)^{n/2}} \frac{n(n+2)}{4} . \tag{6.66}$$

Accordingly, from (6.40)

$$r^* = \left[\frac{2^{n+2}\Gamma(\frac{n+2}{2})}{(n+2)^{n/2+1}} \right]^{\frac{1}{n+4}} \times N^{-\frac{1}{n+4}} . \tag{6.67}$$

TABLE 6-1

OPTIMAL r OF THE UNIFORM KERNEL FUNCTION FOR NORMAL DISTRIBUTIONS

n	4	8	16	32	64	128
r^*	$0.94N^{-1/8}$	$0.89N^{-1/12}$	$0.86N^{-1/20}$	$0.85N^{-1/36}$	$0.85N^{-1/68}$	$0.85N^{-1/132}$
$r^*\sqrt{n+2}$	$2.29N^{-1/8}$	$2.81N^{-1/12}$	$3.66N^{-1/20}$	$4.98N^{-1/36}$	$6.92N^{-1/68}$	$9.72N^{-1/132}$

Table 6-1 shows these r^*'s for various values of n. Remember that the above discussion is for the uniform kernel, and that the radius of the hyperellipsoidal region is $r\sqrt{n+2}$ according to (6.23). Therefore, $r^*\sqrt{n+2}$'s are also presented to demonstrate how large the local regions are.

6.2 *k* Nearest Neighbor Density Estimate

Statistical Properties

***kNN* density estimate:** In the Parzen density estimate of (6.1), we fix v and let **k** be a random variable. Another possibility is to fix k and let **v** be a random variable [12-16]. That is, we extend the local region around X until the kth nearest neighbor is found. The local region, then, becomes random, $\mathbf{L}(X)$, and the volume becomes random, $\mathbf{v}(X)$. Also, both are now functions of X. This approach is called the *k nearest neighbor* (*kNN*) density estimate. The *kNN* approach can be interpreted as the Parzen approach with a uniform kernel function whose size is adjusted automatically, depending on the location. That is, with k fixed throughout the entire space, v becomes larger in low density areas and smaller in high density areas. The kNN density estimate may be rewritten from (6.1) as [12-14]

$$\hat{\mathbf{p}}(X) = \frac{k-1}{N\mathbf{v}(X)} \ . \tag{6.68}$$

The reason why $(k-1)$ is used instead of k will be discussed later.

Density of coverage: Although the density function of $\mathbf{v}$ is not available, the density function of the coverage (the probability mass in the local region), $\mathbf{u}$, may be obtained as follows [17].

Let $L(X)$ and $\Delta L(X)$ be defined by

$$L(X) = \{Y : d(Y,X) \leq \ell\} \text{ and } \Delta L(X) = \{Y : \ell < d(Y,X) \leq \ell + \Delta\ell\} \tag{6.69}$$

and

$$u = \int_{L(X)} p(Y)dY \quad \text{and} \quad \Delta u = \int_{\Delta L(X)} p(Y)dY \ , \tag{6.70}$$

where $d^2(Y,X) = (Y-X)^T A^{-1}(Y-X)$. Also, let two events G and H be defined as

$$G = \{(k-1) \text{ samples in } L(X)\} \ , \tag{6.71}$$

$$H = \{1 \text{ sample in } \Delta L(X)\} \ . \tag{6.72}$$

Then, the probability of the kth NN in $\Delta L(X)$ is

$$Pr\{G \text{ and } H\} = Pr\{G\}Pr\{H \mid G\} \ , \tag{6.73}$$

where

$$Pr\{G\} = \begin{bmatrix} N \\ k-1 \end{bmatrix} u^{k-1}(1-u)^{N-k+1} \ , \tag{6.74}$$

$$Pr\{H \mid G\} = \begin{bmatrix} N-k+1 \\ 1 \end{bmatrix} \left[\frac{\Delta u}{1-u} \right] \left[1 - \frac{\Delta u}{1-u} \right]^{N-k} \ . \tag{6.75}$$

Note that the coverage of $\Delta L(X)$ in the complementary domain of $L(X)$ is $\Delta u/(1-u)$. Substituting (6.74) and (6.75) into (6.73) and using $\{1-\Delta u/(1-u)\} \to 1$ as $\Delta u \to 0$, the probability of (6.73) becomes the product of Δu and a function of u, $p_u(u)$. Therefore, $p_u(u)$ should be the density function of $\mathbf{u}$, where $\mathbf{u}$ is the coverage of $\mathbf{L}(X)$ whose boundary is determined by the kth NN.

$$p_u(u) = \frac{N!}{(k-1)!(N-k)!}u^{k-1}(1-u)^{N-k} \quad 0 \le u \le 1 \ . \tag{6.76}$$

That is, $p_u(u)$ is a *Beta distribution* $Be(k,N-k+1)$. Also, note that the distribution of $\mathbf{u}$ is independent of the underlying distribution, $p(X)$.

More generally, the joint density function of $\mathbf{u}_1, \ldots, \mathbf{u}_k$ may be obtained as [17]

$$p(u_1, \ldots, u_k) = \frac{N!}{(N-k)!}(1-u_k)^{N-k} \ , \tag{6.77}$$

where $\mathbf{u}_i$ is the coverage of $\mathbf{L}_i(X)$, the region extended until the ith NN is found. Note that the joint density depends on u_k only. The marginal density of $\mathbf{u}_k$ can be obtained by integrating (6.77) with respect to $u_1, \ldots, u_{k-1}$ as

$$\int_0^{u_k} \ldots \int_0^{u_2} p(u_1, \ldots, u_k)du_1 \ldots du_{k-1} = \frac{N!}{(k-1)!(N-k)!}u_k^{k-1}(1-u_k)^{N-k} \ . \tag{6.78}$$

Equation (6.78) is the same as (6.76).

The relationship between u and v may be obtained by integrating (6.10) over $L(X)$ with respect to Y. That is,

$$\begin{aligned} u(X) &\cong p(X)v(X) + \frac{1}{2}\mathrm{tr}\{\nabla^2 p(X)\int_{L(X)}(Y-X)(Y-X)^T dY\} \\ &= p(X)v(X)[1 + \frac{1}{2}\alpha(X)r^2(X)] \ , \end{aligned} \tag{6.79}$$

where α is given in (6.13). Note that $\int(Y-X)(Y-X)^T dY = vr^2 A$ from (6.27). The term $[1+\alpha r^2/2]$ of (6.79) appeared in (6.18) in the Parzen case. Again, $u = pv$ gives the first order approximation, and (6.79) is the second order approximation of u in terms of v.

Moments of $\hat{\mathbf{p}}(X)$: When the first order approximation of $u = pv$ is used, from (6.68) and (6.76)

$$E\{\hat{\mathbf{p}}(X)\} \cong \int_0^1 \frac{(k-1)p}{Nu} p_u(u)du = p(X) \ , \tag{6.80}$$

where the following formula is used

$$\int_0^1 x^b(1-x)^c dx = \frac{\Gamma(b+1)\Gamma(c+1)}{\Gamma(b+c+2)} . \tag{6.81}$$

Equation (6.80) indicates that $\hat{\mathbf{p}} = (k-1)/N\mathbf{v}$ is unbiased as long as $u = pv$ holds. If $k/N\mathbf{v}$ is used instead, the estimate becomes biased. This is the reason why $(k-1)$ is used in (6.68) instead of k. The variance of $\hat{\mathbf{p}}(X)$ also can be computed under the approximation of $u = pv$ as

$$\begin{aligned}\text{Var}\{\hat{\mathbf{p}}(X)\} &\cong \int_0^1 \frac{(k-1)^2p^2}{N^2u^2}p_u(u)du - p^2 \\ &= p^2(X)[\frac{1}{k-2}(1-\frac{1}{N}) - \frac{1}{N}] \cong \frac{p^2(X)}{k-2} .\end{aligned} \tag{6.82}$$

Comparison of (6.29) and (6.82) shows that the variance of the kNN density estimate is larger than the one for the Parzen density estimate. Also, (6.82) indicates that, in the kNN density estimate, k must be selected larger than 2. Otherwise, a large variance may result.

Second order approximation: When the second order approximation is needed, (6.79) must be used to relate u and v. However, since r^2 and v are related by $v = cr^n$, it is difficult to solve (6.79) for v and a series of approximations is necessary. Since $\hat{\mathbf{p}} = (k-1)/N\mathbf{v}$, the computation of the first and second order moments of $\hat{\mathbf{p}}(X)$ requires $E\{\mathbf{v}^{-1}\}$ and $E\{\mathbf{v}^{-2}\}$. We start to derive $\mathbf{v}^{-1}$ from (6.79) as

$$\begin{aligned}\mathbf{v}^{-1} &\cong p[\mathbf{u}^{-1} + \frac{1}{2}\alpha c^{-2/n}\mathbf{v}^{2/n}\mathbf{u}^{-1}] \\ &\cong p[\mathbf{u}^{-1} + \frac{1}{2}\alpha(cp)^{-2/n}\mathbf{u}^{2/n-1}] ,\end{aligned} \tag{6.83}$$

where the approximation of $u = pv$ is applied to the second term to obtain the second line from the first. Note that the second term was ignored in the first order approximation and therefore is supposed to be much smaller than the first term. Thus, using $u = pv$ to approximate the second term is justified. From (6.83)

$$\mathbf{v}^{-2} \cong p^2[\mathbf{u}^{-2} + \alpha(cp)^{-2/n}\mathbf{u}^{2/n-2} + \frac{1}{4}\alpha^2(cp)^{-4/n}\mathbf{u}^{4/n-2}] . \tag{6.84}$$

On the other hand, from (6.76) and (6.81),

$$E\{\mathbf{u}^{-m}\} = \frac{\Gamma(k-m)\Gamma(N+1)}{\Gamma(k)\Gamma(N+1-m)} \quad \text{for} \quad k-m > 0 . \tag{6.85}$$

Therefore,

$$E\{\mathbf{u}^{-1}\} = \frac{N}{k-1} \quad \text{and} \quad E\{\mathbf{u}^{-2}\} = \frac{N(N-1)}{(k-1)(k-2)} \tag{6.86}$$

and

$$E\{\mathbf{u}^{\delta-1}\} = \frac{\Gamma(k-1+\delta)\Gamma(N+1)}{\Gamma(k)\Gamma(N+\delta)} = \frac{N}{k-1}\frac{\Gamma(k-1+\delta)}{\Gamma(k-1)}\frac{\Gamma(N)}{\Gamma(N+\delta)} , \tag{6.87}$$

$$E\{\mathbf{u}^{\delta-2}\} = \frac{\Gamma(k-2+\delta)\Gamma(N+1)}{\Gamma(k)\Gamma(N-1+\delta)} = \frac{N(N-1)}{(k-1)(k-2)}\frac{\Gamma(k-2+\delta)}{\Gamma(k-2)}\frac{\Gamma(N-1)}{\Gamma(N-1+\delta)} \tag{6.88}$$

where $\Gamma(x+1) = x\Gamma(x)$ is used. It is known that

$$\frac{\Gamma(x+\delta)}{\Gamma(x)} \cong x^{\delta} \tag{6.89}$$

is a good approximation for large x and small δ. Therefore, applying this approximation,

$$E\{\mathbf{u}^{\delta-1}\} \cong (\frac{k-1}{N})^{\delta-1} \quad \text{and} \quad E\{\mathbf{u}^{\delta-2}\} \cong \frac{N}{k-1}(\frac{k-2}{N-2})^{\delta-1} . \tag{6.90}$$

Combining (6.83), (6.84), (6.86), and (6.90),

$$E\{\hat{\mathbf{p}}(X)\} = \frac{k-1}{N}E\{\mathbf{v}^{-1}\} \cong p(X)[1+\frac{1}{2}\alpha(X)(cp(X))^{-2/n}(\frac{k-1}{N})^{2/n}]$$

$$\cong p(X)[1+\frac{1}{2}\alpha(X)(cp(X))^{-2/n}(\frac{k}{N})^{2/n}] , \tag{6.91}$$

$$E\{\hat{\mathbf{p}}^2(X)\} = (\frac{k-1}{N})^2 E\{\mathbf{v}^{-2}\}$$

$$\cong p^2 \left[\left\{ 1+\frac{1}{k-2}(1-\frac{1}{N})-\frac{1}{N} \right\} + \alpha(cp)^{-2/n}(\frac{k-1}{N})(\frac{k-2}{N-1})^{2/n-1} \right.$$

$$+\ \frac{1}{4}\alpha^2(cp)^{-4/n}(\frac{k-1}{N})(\frac{k-2}{N-1})^{4/n-1}\Bigg]$$

$$\cong p^2\left[(1+\frac{1}{k}) + \alpha(cp)^{-2/n}(\frac{k}{N})^{2/n} + \frac{1}{4}\alpha^2(cp)^{-4/n}(\frac{k}{N})^{4/n}\right], \tag{6.92}$$

where $N \gg k \gg 1$ is assumed. Therefore, the variance and mean-square error of $\hat{\mathbf{p}}(X)$ are

$$\text{Var}\{\hat{\mathbf{p}}(X)\} \cong \frac{p^2(X)}{k}, \tag{6.93}$$

$$MSE\{\hat{\mathbf{p}}(X)\} \cong p^2\left[\frac{1}{k} + \frac{1}{4}\alpha^2(cp)^{-4/n}(\frac{k}{N})^{4/n}\right]. \tag{6.94}$$

Again, in (6.94) the first and second terms are the variance and the squared bias respectively. It must be pointed out that the series of approximations used to obtain (6.91)-(6.94) is valid only for large k. For small k, different and more complex approximations for $E\{\hat{\mathbf{p}}(X)\}$ and $\text{Var}\{\hat{\mathbf{p}}(X)\}$ must be derived by using (6.87) and (6.88) rather than (6.90). As in the Parzen case, the second order approximation for the bias and the first order approximation for the variance may be used for simplicity. Also, note that the MSE of (6.94) becomes zero as $k \rightarrow \infty$ and $k/N \rightarrow 0$. These are the conditions for the kNN density estimate to be asymptotically unbiased and consistent [14].

Optimal Number of Neighbors

Optimal k: In order to apply the kNN density estimate of (6.68), we need to know what value to select for k. The optimal k under the approximation of $u = pv$ is ∞, by minimizing (6.82) with respect to k. That is, when $L(X)$ is small and $u = pv$ holds, the variance dominates the MSE and can be reduced by selecting larger k or larger $L(X)$. As $L(X)$ becomes larger, the second order term produces the bias and the bias increases with $L(X)$. The optimal k is determined by the rate of the variance decrease and the rate of bias increase.

The optimal k, k^*, may be found by minimizing the mean-square error of (6.94). That is, solving $\partial MSE/\partial k = 0$ for k yields [5]

$$k^*(X) = \left[\frac{n(cp)^{4/n}}{\alpha^2}\right]^{\frac{n}{n+4}} \times N^{\frac{4}{n+4}}$$

$$= \left[\frac{n(n+2)^2\pi^2 p^{4/n} |A|^{2/n}}{\Gamma^{4/n}(\frac{n+2}{2})\alpha^2}\right]^{\frac{n}{n+4}} \times N^{\frac{4}{n+4}} . \qquad (6.95)$$

As in the Parzen case, the optimal k is a function of X. Equation (6.95) indicates that k^* is invariant under any non-singular transformation. That is,

$$k^*(Z) = k^*(X) . \qquad (6.96)$$

Also, k^* and r^* of (6.36) are related by

$$p(X) = \frac{k^*(X)}{Ncr^{*n}(X)} . \qquad (6.97)$$

This indicates that both the Parzen and *kNN* density estimates become optimal in the same local range of $L(X)$. The resulting mean-square error is obtained by substituting (6.95) into (6.94).

$$MSE^*\{\hat{p}(X)\} = \frac{n+4}{4}\left[\frac{\Gamma^{4/n}(\frac{n+2}{2})p^{2+4/n}\alpha^2}{n(n+2)^2\pi^2|A|^{2/n}}\right]^{\frac{n}{n+4}} \times N^{-\frac{4}{n+4}} . \qquad (6.98)$$

Note that (6.98) and (6.38) are identical. That is, both the Parzen (with the uniform kernel) and *kNN* density estimates produce the same optimal *MSE*.

The globally optimal k may be obtained by minimizing the integral mean-square error criterion. From (6.94), with a fixed k,

$$IMSE = \frac{1}{k}\int p^2(X)dX + \frac{1}{4}c^{-4/n}(\frac{k}{N})^{4/n}\int \alpha^2(X)p^{2-4/n}(X)dX . \qquad (6.99)$$

Solving $\partial IMSE/\partial k = 0$ generates [5]

$$k^* = \left[\frac{nc^{4/n} \int p^2(X)dX}{\int \alpha^2(X) p^{2-4/n}(X)dX} \right]^{\frac{n}{n+4}} \times N^{\frac{4}{n+4}}$$

$$= \left[\frac{n(n+2)^2 \pi^2 \int p^2(X)dX}{\Gamma^{4/n}(\frac{n+2}{2}) \int \alpha^2(X) p^{2-4/n}(X)dX} \right]^{\frac{n}{n+4}} \times N^{\frac{4}{n+4}} . \tag{6.100}$$

The resulting *IMSE* is

$$IMSE^* = \frac{n+4}{4} \left[\frac{\Gamma^{4/n}(\frac{n+2}{2}) [\int p^2(X)dX]^{4/n} \int \alpha^2(X) p^{2-4/n}(X)dX}{n(n+2)^2 \pi^2 |A|^{2/n}} \right]^{\frac{n}{n+4}}$$

$$\times N^{-\frac{4}{n+4}} . \tag{6.101}$$

It should be pointed out that $E_X\{MSE\{\hat{\mathbf{p}}(\mathbf{X})\}\}$ can be minimized by a similar procedure to obtain the globally optimal k. The resulting k^* is similar but slightly smaller than k^* of (6.100).

Optimal metric: The optimal metric also can be computed as in the Parzen case. Again, a family of density functions with the form of (6.51) is studied with the metric of (6.24). In order to diagonalize both B and A to I and Λ respectively, X is linearly transformed to Z. In the transformed Z-space, $IMSE_Z^*$ becomes, from (6.101) and (6.13),

$$IMSE_Z^* = c_1 \left[c_2 \int p_Z^{-4/n}(Z) \text{tr}^2 \left\{ \nabla^2 p_Z(Z) \frac{\Lambda}{|\Lambda|^{1/n}} \right\} dZ \right]^{\frac{n}{n+4}} , \tag{6.102}$$

where c_1 and c_2 are positive constants. $IMSE_Z^*$ can be minimized with respect to Λ by minimizing

$$J = \text{tr}^2\{\nabla^2 p_Z(Z)\Lambda\} - \mu(\prod_{i=1}^{n} \lambda_i - 1) , \tag{6.103}$$

which is identical to (6.59).

Therefore, the optimal metric A for the kNN density estimate is identical to B. Also, note that the same optimal metric is obtained by minimizing MSE^* of (6.98), and thus the metric is optimal locally as well as globally.

Normal example: The optimal k for a normal distribution can be computed easily. For a normal distribution with zero expected vector and identity covariance matrix,

$$\int p^2(X)dX = \frac{1}{(2\pi)^{n/2}2^{n/2}} , \tag{6.104}$$

$$\int p^{-4/n}(X)\mathrm{tr}^2\{\nabla^2 p(X)\}dX = \frac{\pi^{2-n/2}n^{2+n/2}(n^2-6n+16)}{2^n(n-2)^{2+n/2}} . \tag{6.105}$$

Substituting (6.104) and (6.105) into (6.100), and noting that the optimal metric A is I in this case,

$$k^* = \left[\frac{(n+2)^2(n-2)^{2+n/2}}{\Gamma^{4/n}(\frac{n+2}{2})n^{1+n/2}(n^2-6n+16)} \right]^{\frac{n}{n+4}} \times N^{\frac{4}{n+4}} . \tag{6.106}$$

TABLE 6-2

OPTIMAL k FOR NORMAL DISTRIBUTIONS

n	4	8	16	32	64	128
k^*	$0.75N^{1/2}$	$0.94N^{1/3}$	$0.62N^{1/5}$	$0.34N^{1/9}$	$0.17N^{1/17}$	$0.09N^{1/33}$
N for $k^*=5$	4.4×10	1.5×10^2	3.4×10^4	3.2×10^{10}	9.2×10^{24}	3.8×10^{57}

Table 6-2 shows k^* for various values of n [5]. Also, Table 6-2 shows how many samples are needed for k^* to be 5. Note that N becomes very large after $n = 16$. This suggests how difficult it is to estimate a density function in a high-dimensional space, unless an extremely large number of samples is available.