All these techniques provide nonlinear feature extractors defined on the whole input space. In other words, they can be evaluated on patterns regardless of whether these are elements of the training set or not. Some other methods, such as the *LLE* algorithm [445] and *multidimensional scaling (MDS)* [116], are *restricted* to the training data. They aim to only provide a lower-dimensional representation of the training data, which is useful for instance for data visualization.

MDS

Williams [598] recently pointed out that when considering the special case where we only extract features from the training data, Kernel PCA is actually closely connected to MDS. In a nutshell, MDS is a method for embedding data into $\mathbb{R}^q$, based on pairwise dissimilarities. Consider a situation where the dissimilarities are actually Euclidean distances in $\mathbb{R}^N$ ($N > q$). In the simplest variant of MDS ("classical scaling"), we attempt to embed the training data into $\mathbb{R}^q$ such that the squared distances $\Delta_{ij}^2 := \|x_i - x_j\|^2$ between all pairs of points are (on average) preserved as well as possible. It can be shown from Proposition 14.1 that this is readily achieved by projecting onto the first q principal components.

In *metric MDS*, the dissimilarities Δ_{ij} are transformed by a (nonlinear) function ϕ before the embedding is computed. In this case, the computation of the embedding involves the minimization of a nonlinear "stress" function, which consists of the sum over all mismatches. Usually, this stress function is minimized using nonlinear optimization methods. This can be avoided for a large class of nonlinearities ϕ, however. Williams [598] showed that the metric MDS solution is a by-product of performing kernel PCA with RBF kernels, $k(x_i, x_j) = \phi(\|x_i - x_j\|) = \phi(\Delta_{ij})$.[4] In this case, we thus get away with solving an eigenvalue problem.

Locally Linear Embedding

The second of the aforementioned dimensionality reduction algorithms, LLE, can also be related to kernel PCA. One can show that one obtains the solution of LLE by performing kernel PCA on the Gram matrix computed from what we might call the *locally linear embedding kernel*. This kernel assesses similarity of two patterns based on the similarity of the coefficients required to represent the two patterns in terms of neighboring patterns. For details, see Problem 14.17.

Orthogonal Series Density Estimation

We conclude this section by noting that it has recently been pointed out that one can also connect kernel PCA to *orthogonal series density estimation* [200]. The kernel PCA eigenvalue decomposition provides the coefficients for a truncated density estimator expansion taking the form $p_q(x) = \sum_{n=1}^{q} \lambda_n \left(\frac{1}{m} \sum_{i=1}^{m} \alpha_i^n \langle \mathbf{v}^n, \Phi(x) \rangle \right)$, where q is the number of components taken into account, and α_i^n and $\mathbf{v}$ are defined (and

4. One way of performing metric MDS is to first apply ϕ, and then run classical MDS on the resulting dissimilarity matrix. An interesting class of nonlinearities is the power transformation $\phi(\Delta_{ij}) = \Delta_{ij}^{\mu}$, where $\mu > 0$ ([127], cited after [598]). Provided the original dissimilarities Δ_{ij} arise from Euclidean distances, the power transformation generally leads to a conditionally positive definite matrix $(-\frac{1}{2}\phi(\Delta_{ij})^2)_{ij}$ if and only if $\mu \leq 1$ (cf. (2.81)). The *centered* version of this matrix, which is used in MDS, is thus positive definite if and only if $\mu \leq 1$ (cf. Proposition 2.26). Therefore, it is exactly in these cases that we can run classical MDS after applying ϕ without running into problems. This answers a problem posed by [127], for the case of Euclidean dissimilarities.