In this case, Φ maps each input to a Gaussian centered on this point (see Figure 2.2). We already know from Theorem 2.18, however, that no Gaussian can be written as a linear combination of Gaussians centered at *different* points. Therefore, in the Gaussian case, none of the expansions (18.2), excluding trivial cases with only one term, has an exact pre-image.

18.1.2 Approximate Pre-Images

The problem we initially set out to solve has turned out to be insolvable in the general case. Consequently, rather than trying to find exact pre-images, we attempt to obtain approximate pre-images. We call $z \in \mathbb{R}^N$ an *approximate pre-image* of Ψ if

$$\rho(z) = \|\Psi - \Phi(z)\|^2 \tag{18.9}$$

is small.[1]

Are there vectors Ψ for which good approximate pre-images exist? As we shall see, this is indeed the case. As described in Chapter 14, for $n = 1, 2, \ldots, p$, Kernel PCA provides projections

$$P_n\Phi(x) := \sum_{j=1}^{n} \left\langle \Phi(x), \mathbf{v}^j \right\rangle \mathbf{v}^j \tag{18.10}$$

with the following optimal approximation property (Proposition 14.1): Assume that the $\mathbf{v}^j$ are sorted according to nonincreasing eigenvalues λ_j, with λ_n being the smallest nonzero eigenvalue. Then P_n is the n-dimensional projection minimizing

$$\sum_{i=1}^{m} \|P_n\Phi(x_i) - \Phi(x_i)\|^2. \tag{18.11}$$

Therefore, $P_n\Phi(x)$ can be expected to have a good approximate pre-image, provided that x is drawn from the same distribution as the x_i; to give a trivial example, x itself is already a good approximate pre-image. As we shall see in experiments, however, even better pre-images can be found, which makes some interesting applications possible [474, 365]:

Applications of Pre-Images

Denoising. Given a noisy x, map it to $\Phi(x)$, discard components corresponding to the eigenvalues $\lambda_{n+1}, \ldots, \lambda_m$ to obtain $P_n\Phi(x)$, and then compute a pre-image z. The hope here is that the main structure in the data set is captured in the first n directions in feature space, and the remaining components mainly pick up the noise — in this sense, z can be thought of as a denoised version of x.

Compression. Given the Kernel PCA eigenvectors and a small number of features $P_n\Phi(x)$ (cf. (18.10)) of $\Phi(x)$, but not x, compute a pre-image as an approximate reconstruction of x. This is useful if n is smaller than the dimensionality of the input data.

1. Just how small it needs to be in order to form a satisfactory approximation depends on the problem at hand. Therefore, we have refrained from giving a formal definition.