sample from. Next we sample the real-valued observable variables given the factors:

\[ x = Wh + b + \text{noise} \]  

where the noise is typically Gaussian and diagonal (independent across dimensions).

This is illustrated in figure 13.1.

\[ x_1 = Wh + b + \text{noise} \]

**Figure 13.1**
Probabilistic PCA and Factor Analysis

- Linear factor model
- Gaussian prior
- Extends PCA
  - Given an input, yields a distribution over codes, rather than a single code
  - Estimates a probability density function
  - Can generate samples
Independent Components Analysis

• Factorial but non-Gaussian prior

• Learns components that are closer to statistically independent than the raw features

• Can be used to separate voices of $n$ speakers recorded by $n$ microphones, or to separate multiple EEG signals

• Many variants, some more probabilistic than others
Slow Feature Analysis

• Learn features that change gradually over time

• SFA algorithm does so in closed form for a linear model

• Deep SFA by composing many models with fixed feature expansions, like quadratic feature expansion
Sparse Coding

\[ p(x \mid h) = \mathcal{N}(x; Wh + b, \frac{1}{\beta} I). \]  \hfill (13.12)

\[ p(h_i) = \text{Laplace}(h_i; 0, \frac{2}{\lambda}) = \frac{\lambda}{4} e^{-\frac{1}{2} \lambda |h_i|} \]  \hfill (13.13)

\[ \arg \min_{h} \lambda \|h\|_1 + \beta \|x - Wh\|_2^2, \]  \hfill (13.18)
Sparse Coding

Figure 13.2: Example samples and weights from a spike and slab sparse coding model trained on the MNIST dataset. (Left) The samples from the model do not resemble the training examples. At first glance, one might assume the model is poorly fit. (Right) The weight vectors of the model have learned to represent penstrokes and sometimes complete digits. The model has thus learned useful features. The problem is that the factorial prior over features results in random subsets of features being combined. Few such subsets are appropriate to form a recognizable MNIST digit. This motivates the development of generative models that have more powerful distributions over their latent codes. Figure reproduced with permission from Goodfellow et al. (2013d).

13.5 Manifold Interpretation of PCA

Linear factor models including PCA and factor analysis can be interpreted as learning a manifold (Hinton et al., 1997). We can view probabilistic PCA as defining a thin pancake-shaped region of high probability—a Gaussian distribution that is very narrow along some axes, just as a pancake is very flat along its vertical axis, but is elongated along other axes, just as a pancake is wide along its horizontal axes. This is illustrated in figure 13.3. PCA can be interpreted as aligning this pancake with a linear manifold in a higher-dimensional space. This interpretation applies not just to traditional PCA but also to any linear autoencoder that learns matrices $W$ and $V$ with the goal of making the reconstruction of $x$ lie as close to $x$ as possible.

Let the encoder be $h = f(x) = W(x - \mu)$. (13.19)
Manifold Interpretation of PCA

The encoder computes a low-dimensional representation of $h$. With the autoencoder view, we have a decoder computing the reconstruction $\hat{x} = g(h) = b + Vh$.

$\text{Figure 13.3: Flat Gaussian capturing probability concentration near a low-dimensional manifold. The figure shows the upper half of the “pancake” above the “manifold plane” which goes through its middle. The variance in the direction orthogonal to the manifold is very small (arrow pointing out of plane) and can be considered like “noise,” while the other variances are large (arrows in the plane) and correspond to “signal,” and a coordinate system for the reduced-dimension data.}$

The choices of linear encoder and decoder that minimize reconstruction error $E[||x - \hat{x}||^2]$ correspond to $V = W$, $\mu = b = E[x]$ and the columns of $W$ form an orthonormal basis which spans the same subspace as the principal eigenvectors of the covariance matrix $C = E[(x - \mu)(x - \mu)^T]$.

In the case of PCA, the columns of $W$ are these eigenvectors, ordered by the magnitude of the corresponding eigenvalues (which are all real and non-negative).

One can also show that eigenvalue $\lambda_i$ of $C$ corresponds to the variance of $x$ in the direction of eigenvector $v_i$.

$\text{Figure 13.3}$