

This is in comparison to other learning algorithms where the cost function depends highly on specific pixel values, making it much more difficult to determine what features the model will learn.

Deep SFA has also been used to learn features for object recognition and pose estimation (Franzius *et al.*, 2008). So far, the slowness principle has not become the basis for any state of the art applications. It is unclear what factor has limited its performance. We speculate that perhaps the slowness prior is too strong, and that, rather than imposing a prior that features should be approximately constant, it would be better to impose a prior that features should be easy to predict from one time step to the next. The position of an object is a useful feature regardless of whether the object’s velocity is high or low, but the slowness principle encourages the model to ignore the position of objects that have high velocity.

## 13.4 Sparse Coding

**Sparse coding** (Olshausen and Field, 1996) is a linear factor model that has been heavily studied as an unsupervised feature learning and feature extraction mechanism. Strictly speaking, the term “sparse coding” refers to the process of inferring the value of  $\mathbf{h}$  in this model, while “sparse modeling” refers to the process of designing and learning the model, but the term “sparse coding” is often used to refer to both.

Like most other linear factor models, it uses a linear decoder plus noise to obtain reconstructions of  $\mathbf{x}$ , as specified in equation 13.2. More specifically, sparse coding models typically assume that the linear factors have Gaussian noise with isotropic precision  $\beta$ :

$$p(\mathbf{x} \mid \mathbf{h}) = \mathcal{N}(\mathbf{x}; \mathbf{W}\mathbf{h} + \mathbf{b}, \frac{1}{\beta}\mathbf{I}). \quad (13.12)$$

The distribution  $p(\mathbf{h})$  is chosen to be one with sharp peaks near 0 (Olshausen and Field, 1996). Common choices include factorized Laplace, Cauchy or factorized Student- $t$  distributions. For example, the Laplace prior parametrized in terms of the sparsity penalty coefficient  $\lambda$  is given by

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

and the Student- $t$  prior by

$$p(h_i) \propto \frac{1}{(1 + \frac{h_i^2}{\nu})^{\frac{\nu+1}{2}}}. \quad (13.14)$$