This chapter discusses latent linear models. One of the issues with mixture models is that they only use a single latent variable to generate the observations, and in particular, each of these observations can only come from one of K prototypes. A mixture model can be thought of as using K hidden binary variables representing a one-hot encoding of the cluster identity, but because these variables are mutually exclusive, the model is still limited in its representational power. An alternative to this is to use a vector of real-valued latent variables, for example, $z_i \in R^L$.

When using a factor loading matrix, given as a D x L matrix W, and covariance matrix Ψ which is taken to be diagonal, the overall model is called factor analysis, also known as FA. The special case where $\Psi = \sigma^2 I$ is called probabilistic principal components analysis, also called PPCA. The generative process where L = 1, D = 2, and Ψ is diagonal, we can take an isotropic Gaussian and slide it along the 1 dimensional line defined by $wz_i + \mu$ which induces an elongated and therefore correlated Gaussian in 2 dimensions.

Factor analysis can be thought of as a way of specifying a joint density model on x using a small number of parameters. This can be seen noting that the induced marginal distribution $p(x_i|\theta)$ is a Gaussian. Although FA can sometimes be thought of as just a way to define a density on x, it is often used because we hope that the latent factors z will reveal something interesting about the data. In order to do this, we need to compute the posterior over the latent factors, which can be done using the Bayes rule for Gaussians. In the FA model, Σ_i is independent of i, so we can denote it by Σ. Computing this matrix takes $O(L^3 + L^2D)$ time, and computing each $m_i = E[z_i|x_i, \theta]$ takes $O(L^2 + LD)$ time. The m_i are sometimes called latent scores or latent factors.