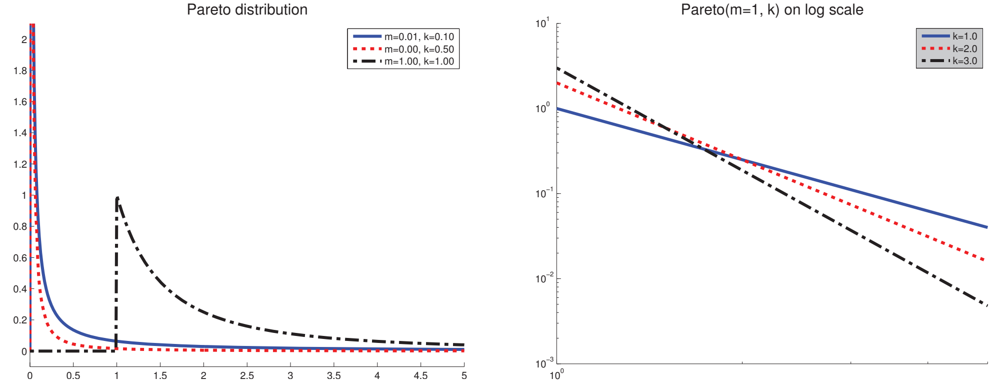

(a) (b)

Figure 2.11 (a) The Pareto distribution Pareto$(x|m, k)$ for $m = 1$. (b) The pdf on a log-log scale. Figure generated by `paretoPlot`.

2.5 Joint probability distributions

So far, we have been mostly focusing on modeling univariate probability distributions. In this section, we start our discussion of the more challenging problem of building joint probability distributions on multiple related random variables; this will be a central topic in this book.

A **joint probability distribution** has the form $p(x_1, \ldots, x_D)$ for a set of $D > 1$ variables, and models the (stochastic) relationships between the variables. If all the variables are discrete, we can represent the joint distribution as a big multi-dimensional array, with one variable per dimension. However, the number of parameters needed to define such a model is $O(K^D)$, where K is the number of states for each variable.

We can define high dimensional joint distributions using fewer parameters by making conditional independence assumptions, as we explain in Chapter 10. In the case of continuous distributions, an alternative approach is to restrict the form of the pdf to certain functional forms, some of which we will examine below.

2.5.1 Covariance and correlation

The **covariance** between two rv's X and Y measures the degree to which X and Y are (linearly) related. Covariance is defined as

$$\text{cov}\left[X, Y\right] \triangleq \mathbb{E}\left[(X - \mathbb{E}\left[X\right])(Y - \mathbb{E}\left[Y\right])\right] = \mathbb{E}\left[XY\right] - \mathbb{E}\left[X\right]\mathbb{E}\left[Y\right] \tag{2.65}$$

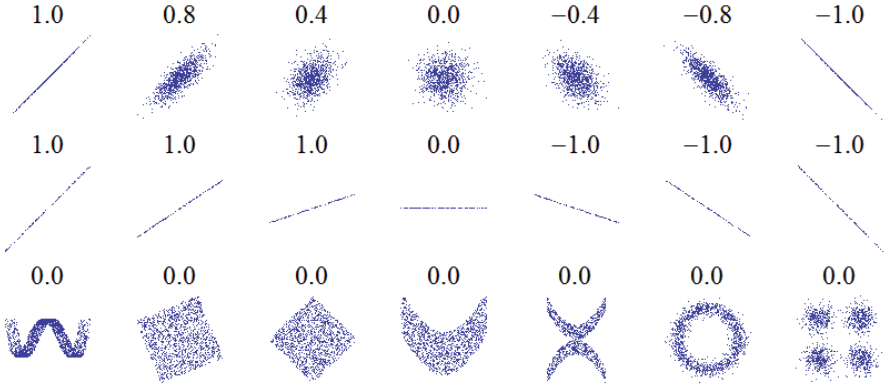

Figure 2.12 Several sets of (x, y) points, with the correlation coefficient of x and y for each set. Note that the correlation reflects the noisiness and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many aspects of nonlinear relationships (bottom). N.B.: the figure in the center has a slope of 0 but in that case the correlation coefficient is undefined because the variance of Y is zero. Source: `http://en.wikipedia.org/wiki/File:Correlation_examples.png`

If $\mathbf{x}$ is a d-dimensional random vector, its **covariance matrix** is defined to be the following symmetric, positive definite matrix:

$$\operatorname{cov}[\mathbf{x}] \triangleq \mathbb{E}\left[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^T\right] \tag{2.66}$$

$$= \begin{pmatrix} \operatorname{var}[X_1] & \operatorname{cov}[X_1, X_2] & \cdots & \operatorname{cov}[X_1, X_d] \\ \operatorname{cov}[X_2, X_1] & \operatorname{var}[X_2] & \cdots & \operatorname{cov}[X_2, X_d] \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{cov}[X_d, X_1] & \operatorname{cov}[X_d, X_2] & \cdots & \operatorname{var}[X_d] \end{pmatrix} \tag{2.67}$$

Covariances can be between 0 and infinity. Sometimes it is more convenient to work with a normalized measure, with a finite upper bound. The (Pearson) **correlation coefficient** between X and Y is defined as

$$\operatorname{corr}[X, Y] \triangleq \frac{\operatorname{cov}[X, Y]}{\sqrt{\operatorname{var}[X]\operatorname{var}[Y]}} \tag{2.68}$$

A **correlation matrix** has the form

$$\mathbf{R} = \begin{pmatrix} \operatorname{corr}[X_1, X_1] & \operatorname{corr}[X_1, X_2] & \cdots & \operatorname{corr}[X_1, X_d] \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{corr}[X_d, X_1] & \operatorname{corr}[X_d, X_2] & \cdots & \operatorname{corr}[X_d, X_d] \end{pmatrix} \tag{2.69}$$

One can show (Exercise 4.3) that $-1 \leq \operatorname{corr}[X, Y] \leq 1$. Hence in a correlation matrix, each entry on the diagonal is 1, and the other entries are between -1 and 1.

One can also show that $\operatorname{corr}[X, Y] = 1$ if and only if $Y = aX + b$ for some parameters a and b, i.e., if there is a *linear* relationship between X and Y (see Exercise 4.4). Intuitively one might expect the correlation coefficient to be related to the slope of the regression line, i.e., the coefficient a in the expression $Y = aX + b$. However, as we show in Equation 7.99 later, the regression coefficient is in fact given by $a = \text{cov}[X, Y] / \text{var}[X]$. A better way to think of the correlation coefficient is as a degree of linearity: see Figure 2.12.

If X and Y are independent, meaning $p(X, Y) = p(X)p(Y)$ (see Section 2.2.4), then $\text{cov}[X, Y] = 0$, and hence $\text{corr}[X, Y] = 0$ so they are uncorrelated. However, the converse is not true: *uncorrelated does not imply independent*. For example, let $X \sim U(-1, 1)$ and $Y = X^2$. Clearly Y is dependent on X (in fact, Y is uniquely determined by X), yet one can show (Exercise 4.1) that $\text{corr}[X, Y] = 0$. Some striking examples of this fact are shown in Figure 2.12. This shows several data sets where there is clear dependendence between X and Y, and yet the correlation coefficient is 0. A more general measure of dependence between random variables is mutual information, discussed in Section 2.8.3. This is only zero if the variables truly are independent.

2.5.2 The multivariate Gaussian

The **multivariate Gaussian** or **multivariate normal** (**MVN**) is the most widely used joint probability density function for continuous variables. We discuss MVNs in detail in Chapter 4; here we just give some definitions and plots.

The pdf of the MVN in D dimensions is defined by the following:

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \triangleq \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] \tag{2.70}$$

where $\boldsymbol{\mu} = \mathbb{E}[\mathbf{x}] \in \mathbb{R}^D$ is the mean vector, and $\boldsymbol{\Sigma} = \text{cov}[\mathbf{x}]$ is the $D \times D$ covariance matrix. Sometimes we will work in terms of the **precision matrix** or **concentration matrix** instead. This is just the inverse covariance matrix, $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$. The normalization constant $(2\pi)^{-D/2}|\boldsymbol{\Lambda}|^{1/2}$ just ensures that the pdf integrates to 1 (see Exercise 4.5).

Figure 2.13 plots some MVN densities in 2d for three different kinds of covariance matrices. A full covariance matrix has $D(D+1)/2$ parameters (we divide by 2 since $\boldsymbol{\Sigma}$ is symmetric). A diagonal covariance matrix has D parameters, and has 0s in the off-diagonal terms. A **spherical** or **isotropic** covariance, $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_D$, has one free parameter.

2.5.3 Multivariate Student t distribution

A more robust alternative to the MVN is the **multivariate Student t** distribution, whose pdf is given by

$$\mathcal{T}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \frac{\Gamma(\nu/2 + D/2)}{\Gamma(\nu/2)} \frac{|\boldsymbol{\Sigma}|^{-1/2}}{\nu^{D/2}\pi^{D/2}} \times \left[1 + \frac{1}{\nu}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]^{-(\frac{\nu+D}{2})} \tag{2.71}$$

$$= \frac{\Gamma(\nu/2 + D/2)}{\Gamma(\nu/2)} |\pi\mathbf{V}|^{-1/2} \times \left[1 + (\mathbf{x}-\boldsymbol{\mu})^T\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]^{-(\frac{\nu+D}{2})} \tag{2.72}$$

where $\boldsymbol{\Sigma}$ is called the scale matrix (since it is not exactly the covariance matrix) and $\mathbf{V} = \nu\boldsymbol{\Sigma}$. This has fatter tails than a Gaussian. The smaller ν is, the fatter the tails. As $\nu \to \infty$, the