CHAPTER

EXAMPLES OF MULTIDIMENSIONAL PROBABILITY DISTRIBUTIONS

6

CHAPTER CONTENTS

In this chapter, popular multidimensional probability distributions and their properties such as the expectation, the variance, and the moment-generating functions are illustrated.

6.1 MULTINOMIAL DISTRIBUTION

The binomial distribution explained in Section 3.2 is the probability distribution of the number x of successful trials in n Bernoulli trials with the probability of success p. The *multinomial distribution* is an extension of the binomial distribution to multidimensional cases.

Let us consider a d-sided dice with the probability of obtaining each side $\boldsymbol{p} = (p_1, \ldots, p_d)^\top$, where

$$p_1, \ldots, p_d \geq 0 \quad \text{and} \quad \sum_{j=1}^{d} p_j = 1.$$

Let $\boldsymbol{x} = (x^{(1)}, \ldots, x^{(d)})^\top$ be the number of times each side appears when the dice is thrown n times, where

$$\boldsymbol{x} \in \Delta_{d,n} = \left\{ \boldsymbol{x} \;\middle|\; x^{(1)}, \ldots, x^{(d)} \geq 0,\ x^{(1)} + \cdots + x^{(d)} = n \right\}.$$

The probability distribution that $\boldsymbol{x}$ follows is the multinomial distribution and is denoted by $\mathrm{Mult}(n, \boldsymbol{p})$.

The probability that the jth side appears $x^{(j)}$ times is $(p_j)^{x^{(j)}}$, and the number of combinations each of the d sides appears $x^{(1)}, \ldots, x^{(d)}$ times for n trials is given by

An Introduction to Statistical Machine Learning. DOI: 10.1016/B978-0-12-802121-7.00017-0

$\frac{n!}{x^{(1)}!\cdots x^{(d)}!}$. Putting them together, the probability mass function of Mult$(n, \boldsymbol{p})$ is given by

$$f(\boldsymbol{x}) = \frac{n!}{x^{(1)}! \dots x^{(d)}!}(p_1)^{x^{(1)}} \cdots (p_d)^{x^{(d)}}.$$

When $d = 2$, Mult$(n, \boldsymbol{p})$ is reduced to Bi(n, p_1).

The binomial theorem can also be extended to *multinomial theorem*:

$$(p_1 + \cdots + p_d)^n = \sum_{\boldsymbol{x} \in \Delta_{d,n}} \frac{n!}{x^{(1)}! \dots x^{(d)}!}(p_1)^{x^{(1)}} \cdots (p_d)^{x^{(d)}},$$

with which the moment-generating function of Mult$(n, \boldsymbol{p})$ can be computed as

$$\begin{aligned} M_{\boldsymbol{x}}(\boldsymbol{t}) = E[e^{\boldsymbol{t}^\top \boldsymbol{x}}] &= \sum_{\boldsymbol{x} \in \Delta_{d,n}} e^{t_1 x^{(1)}} \cdots e^{t_d x^{(d)}} \frac{n!}{x^{(1)}! \dots x^{(d)}!}(p_1)^{x^{(1)}} \cdots (p_d)^{x^{(d)}} \\ &= \sum_{\boldsymbol{x} \in \Delta_{d,n}} \frac{n!}{x^{(1)}! \dots x^{(d)}!}(p_1 e^{t_1})^{x^{(1)}} \cdots (p_d e^{t_d})^{x^{(d)}} \\ &= (p_1 e^{t_1} + \cdots + p_d e^{t_d})^n. \end{aligned}$$

From this, the expectation of and (co)variance of Mult$(n, \boldsymbol{p})$ can be computed as

$$E[x^{(j)}] = np_j \quad \text{and} \quad \mathrm{Cov}[x^{(j)}, x^{(j')}] = \begin{cases} np_j(1 - p_j) & (j = j'), \\ -np_j p_{j'} & (j \neq j'). \end{cases}$$

6.2 MULTIVARIATE NORMAL DISTRIBUTION

When random variables $y^{(1)}, \dots, y^{(d)}$ independently follow the standard normal distribution $N(0, 1)$, the joint probability density function of $\boldsymbol{y} = (y^{(1)}, \dots, y^{(d)})^\top$ is given by

$$\begin{aligned} g(\boldsymbol{y}) &= \prod_{j=1}^{d} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y^{(j)})^2}{2}\right) \\ &= \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}\boldsymbol{y}^\top \boldsymbol{y}\right). \end{aligned}$$

The expectation and variance-covariance matrices of $\boldsymbol{y}$ are given by

$$E[\boldsymbol{y}] = \boldsymbol{0} \quad \text{and} \quad V[\boldsymbol{y}] = \boldsymbol{I},$$

where $\boldsymbol{0}$ denotes the zero vector and $\boldsymbol{I}$ denotes the identity matrix.

Let us transform $\boldsymbol{y}$ by $d \times d$ *invertible matrix* $\boldsymbol{T}$ and d-dimensional vector $\boldsymbol{\mu}$ as

$$\boldsymbol{x} = \boldsymbol{T}\boldsymbol{y} + \boldsymbol{\mu}.$$

Then the joint probability density function of $\boldsymbol{x}$ is given by

$$\begin{aligned} f(\boldsymbol{x}) &= g(\boldsymbol{y})|\det(\boldsymbol{T})|^{-1} \\ &= \frac{1}{(2\pi)^{d/2}\det(\boldsymbol{\Sigma})^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right), \end{aligned}$$

where $\det(\boldsymbol{T})$ is the Jacobian (Fig. 4.2), and

$$\boldsymbol{\Sigma} = \boldsymbol{T}\boldsymbol{T}^\top.$$

This is the general form of the *multivariate normal distribution* and is denoted by $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The expectation and variance-covariance matrices of $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are given by

$$\begin{aligned} E[\boldsymbol{x}] &= \boldsymbol{T}E[\boldsymbol{y}] + \boldsymbol{\mu} = \boldsymbol{\mu}, \\ V[\boldsymbol{x}] &= V[\boldsymbol{T}\boldsymbol{y} + \boldsymbol{\mu}] = \boldsymbol{T}V[\boldsymbol{y}]\boldsymbol{T}^\top = \boldsymbol{\Sigma}. \end{aligned}$$

Probability density functions of two-dimensional normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are illustrated in Fig. 6.1. The contour lines of two-dimensional normal distributions are *elliptic*, and their *principal axes* agree with the coordinate axes if the variance-covariance matrix is diagonal. Furthermore, the elliptic contour lines become spherical if all diagonal elements are equal (i.e., the variance-covariance matrix is proportional to the identity matrix).

Eigendecomposition of variance-covariance matrix $\boldsymbol{\Sigma}$ (see Fig. 6.2) shows that the principal axes of the ellipse are parallel to the eigenvectors of $\boldsymbol{\Sigma}$, and their length is proportional to the square root of the eigenvalues (Fig. 6.3).

6.3 DIRICHLET DISTRIBUTION

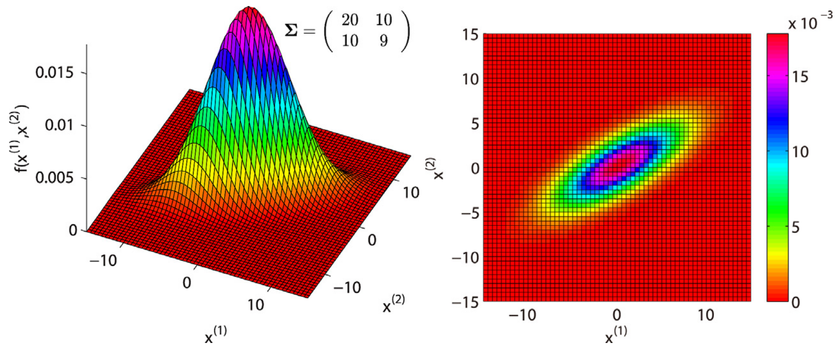
Let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)^\top$ be a d-dimensional vector with positive entries and let $y^{(1)}, \ldots, y^{(d)}$ be random variables that independently follow the gamma distribution $\mathrm{Ga}(\alpha_j, \lambda)$. For $V = \sum_{j=1}^d y^{(j)}$, let

$$\boldsymbol{x} = (x^{(1)}, \ldots, x^{(d)})^\top = \left(\frac{y^{(1)}}{V}, \ldots, \frac{y^{(d)}}{V}\right)^\top.$$

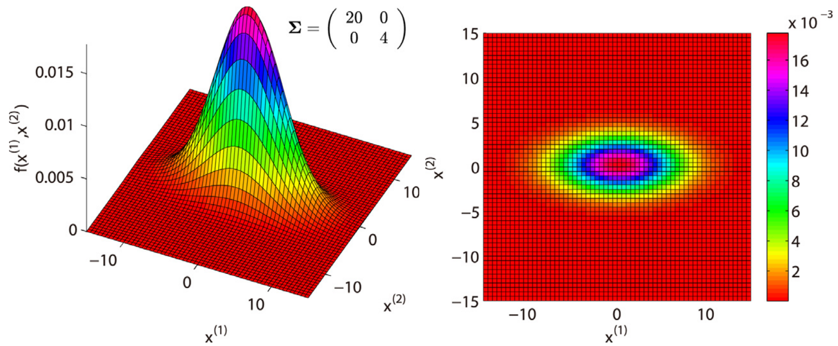
Then the distribution that the above d-dimensional vector $\boldsymbol{x}$ follows is called the *Dirichlet distribution* and is denoted by $\mathrm{Dir}(\boldsymbol{\alpha})$. The domain of $\boldsymbol{x}$ is given by

$$\Delta_d = \left\{\boldsymbol{x} \;\middle|\; x^{(1)}, \ldots, x^{(d)} \geq 0,\ x^{(1)} + \cdots + x^{(d)} = 1\right\}.$$

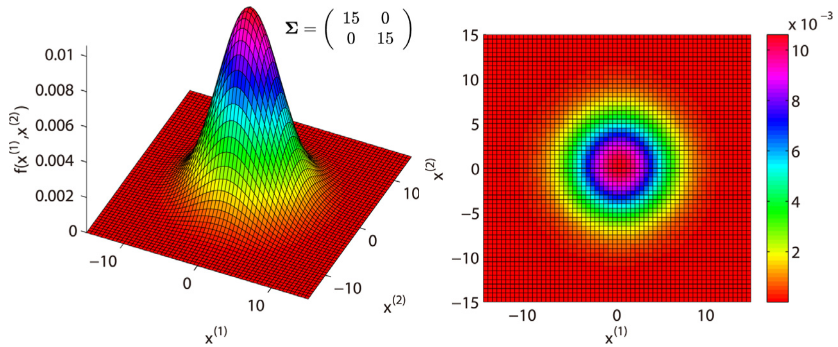
Drawing a value from a Dirichlet distribution corresponds to generating a (unfair) d-sided die.



(a) When Σ is generic, the contour lines are elliptic



(b) When Σ is diagonal, the principal axes of the elliptic contour lines agree with the coordinate axes



(c) When Σ is proportional to identity, the contour lines are spherical

FIGURE 6.1

Probability density functions of two-dimensional normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = (0,0)^\top$.

For $d \times d$ matrix $\boldsymbol{A}$, a nonzero vector $\boldsymbol{\phi}$ and a scalar λ that satisfy

$$\boldsymbol{A}\boldsymbol{\phi} = \lambda\boldsymbol{\phi}$$

are called an *eigenvector* and an *eigenvalue* of $\boldsymbol{A}$, respectively. Generally, there exist d eigenvalues $\lambda_1, \ldots, \lambda_d$ and they are all real when $\boldsymbol{A}$ is symmetric. A matrix whose eigenvalues are all positive is called a *positive definite matrix*, and a matrix whose eigenvalues are all non-negative is called a *positive semidefinite matrix*. Eigenvectors $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_d$ corresponding to eigenvalues $\lambda_1, \ldots, \lambda_d$ are orthogonal and are usually normalized, i.e., they are *orthonormal* as

$$\boldsymbol{\phi}_j^\top \boldsymbol{\phi}_{j'} = \begin{cases} 1 & (j = j'), \\ 0 & (j \neq j'). \end{cases}$$

A matrix $\boldsymbol{A}$ can be expressed by using its eigenvectors and eigenvalues as

$$\boldsymbol{A} = \sum_{j=1}^{d} \lambda_j \boldsymbol{\phi}_j \boldsymbol{\phi}_j^\top = \boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{\Phi}^\top,$$

where $\boldsymbol{\Phi} = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_d)$ and $\boldsymbol{\Lambda}$ is the diagonal matrix with diagonal elements $\lambda_1, \ldots, \lambda_d$. This is called *eigenvalue decomposition* of $\boldsymbol{A}$. In MATLAB, eigenvalue decomposition can be performed by the `eig` function. When all eigenvalues are nonzero, the inverse of $\boldsymbol{A}$ can be expressed as

$$\boldsymbol{A}^{-1} = \sum_{j=1}^{d} \lambda_j^{-1} \boldsymbol{\phi}_j \boldsymbol{\phi}_j^\top = \boldsymbol{\Phi}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Phi}^\top.$$

For $d \times d$ matrix $\boldsymbol{A}$ and $d \times d$ positive symmetric matrix $\boldsymbol{B}$, a nonzero vector $\boldsymbol{\phi}$ and a scalar λ that satisfy

$$\boldsymbol{A}\boldsymbol{\phi} = \lambda\boldsymbol{B}\boldsymbol{\phi}$$

are called a *generalized eigenvector* and a *generalized eigenvalue* of $\boldsymbol{A}$, respectively.

FIGURE 6.2

Eigenvalue decomposition.

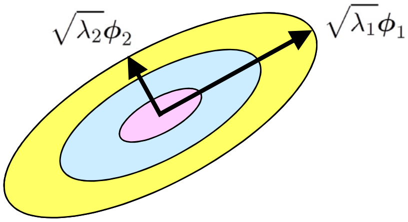


FIGURE 6.3

Contour lines of the normal density. The principal axes of the ellipse are parallel to the eigenvectors of variance-covariance matrix $\boldsymbol{\Sigma}$, and their length is proportional to the square root of the eigenvalues.

The probability density function of $\mathrm{Dir}(\boldsymbol{\alpha})$ is given by

$$f(\boldsymbol{x}) = \frac{\prod_{j=1}^{d}(x^{(j)})^{\alpha_j-1}}{B_d(\boldsymbol{\alpha})},$$

where

$$B_d(\boldsymbol{\alpha}) = \int_{\Delta_d}\prod_{j=1}^{d}(x^{(j)})^{\alpha_j-1}\mathrm{d}\boldsymbol{x} \tag{6.1}$$

is the d-dimensional *beta function*.

For $x = (\sqrt{p}\sin\theta)^2$ with $0 \le p \le 1$, Eq. (4.6) implies

$$\begin{aligned}
&\int_0^p x^{\alpha-1}(p-x)^{\beta-1}\mathrm{d}x \\
&= \int_0^{\frac{\pi}{2}} (\sqrt{p}\sin\theta)^{2(\alpha-1)}\left(p-(\sqrt{p}\sin\theta)^2\right)^{\beta-1}\frac{\mathrm{d}x}{\mathrm{d}\theta}\mathrm{d}\theta \\
&= \int_0^{\frac{\pi}{2}} p^{\alpha-1}(\sin\theta)^{2(\alpha-1)}p^{\beta-1}(\cos\theta)^{2(\beta-1)}\cdot 2p\sin\theta\cos\theta\mathrm{d}\theta \\
&= p^{\alpha+\beta-1}\left(2\int_0^{\frac{\pi}{2}}(\sin\theta)^{2\alpha-1}(\cos\theta)^{2\beta-1}\mathrm{d}\theta\right) \\
&= p^{\alpha+\beta-1}B(\alpha,\beta).
\end{aligned}$$

Letting $p = 1 - \sum_{j=1}^{d-2} x^{(j)}$, the integration with respect to $x^{(d-1)}$ in

$$\begin{aligned}B_d(\boldsymbol{\alpha}) = &\int_0^1 (x^{(1)})^{\alpha_1-1} \times \int_0^{1-x^{(1)}} (x^{(2)})^{\alpha_2-1} \times \\ &\cdots \times \int_0^{1-\sum_{j=1}^{d-2} x^{(j)}} (x^{(d-1)})^{\alpha_{d-1}-1} \\ &\times \left(1 - \sum_{j=1}^{d-1} x^{(j)}\right)^{\alpha_d-1} \mathrm{d}x^{(1)}\mathrm{d}x^{(2)} \cdots \mathrm{d}x^{(d-1)}\end{aligned}$$

can be computed as

$$\begin{aligned}&\int_0^{1-\sum_{j=1}^{d-2} x^{(j)}} (x^{(d-1)})^{\alpha_{d-1}-1} \left(1 - \sum_{j=1}^{d-1} x^{(j)}\right)^{\alpha_d-1} \mathrm{d}x^{(d-1)} \\ &= \left(1 - \sum_{j=1}^{d-2} x^{(j)}\right)^{\sum_{j=d-1}^{d} \alpha_j - 1} B(\alpha_{d-1}, \alpha_d).\end{aligned}$$

Similarly, letting $p = 1 - \sum_{j=1}^{d-3} x^{(j)}$, the integration with respect to $x^{(d-2)}$ in the above equation can be computed as

$$\begin{aligned}&\int_0^{1-\sum_{j=1}^{d-3} x^{(j)}} (x^{(d-2)})^{\alpha_{d-2}-1} \left(1 - \sum_{j=1}^{d-2} x^{(j)}\right)^{\sum_{j=d-1}^{d} \alpha_j - 1} \mathrm{d}x^{(d-2)} \\ &= \left(1 - \sum_{j=1}^{d-3} x^{(j)}\right)^{\sum_{j=d-2}^{d} \alpha_j - 1} B\left(\alpha_{d-2}, \sum_{j=d-1}^{d} \alpha_j\right).\end{aligned}$$

Repeating this computation yields

$$B_d(\boldsymbol{\alpha}) = \prod_{j=1}^{d-1} B\left(\alpha_j, \sum_{j'=j+1}^{d} \alpha_{j'}\right).$$

Applying

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}$$

given in Eq. (4.7) to the above equation, the d-dimensional beta function can be expressed by using the gamma function as

$$B_d(\boldsymbol{\alpha}) = \prod_{j=1}^{d-1} \frac{\Gamma(\alpha_j)\Gamma\left(\sum_{j'=j+1}^{d} \alpha_{j'}\right)}{\Gamma\left(\sum_{j'=j}^{d} \alpha_{j'}\right)} = \frac{\prod_{j=1}^{d} \Gamma(\alpha_j)}{\Gamma(\alpha_0)},$$

where

$$\alpha_0 = \sum_{j=1}^{d} \alpha_j.$$

When $d = 2$, $x^{(2)} = 1 - x^{(1)}$ holds and thus the Dirichlet distribution is reduced to the beta distribution shown in Section 4.4:

$$f(x) = \frac{x^{\alpha_1 - 1}(1-x)^{\alpha_2 - 1}}{B(\alpha_1, \alpha_2)}.$$

Thus, the Dirichlet distribution can be regarded as a multidimensional extension of the beta distribution.

The expectation of $\mathrm{Dir}(\boldsymbol{\alpha})$ is given by

$$\begin{aligned}
E[x^{(j)}] &= \frac{\int x^{(j)} \prod_{j'=1}^{d} (x^{(j')})^{\alpha_{j'}-1} \mathrm{d}x^{(j)}}{B_d(\boldsymbol{\alpha})} \\
&= \frac{B_d(\alpha_1, \ldots, \alpha_{j-1}, \alpha_j + 1, \alpha_{j+1}, \ldots, \alpha_d)}{B_d(\alpha_1, \ldots, \alpha_{j-1}, \alpha_j, \alpha_{j+1}, \ldots, \alpha_d)} \\
&= \frac{\Gamma(\alpha_1) \cdots \Gamma(\alpha_{j-1}) \Gamma(\alpha_j + 1) \Gamma(\alpha_{j+1}) \cdots \Gamma(\alpha_d) \times \Gamma(\alpha_0)}{\Gamma(\alpha_0 + 1) \times \Gamma(\alpha_1) \cdots \Gamma(\alpha_{j-1}) \Gamma(\alpha_j) \Gamma(\alpha_{j+1}) \cdots \Gamma(\alpha_d)} \\
&= \frac{\Gamma(\alpha_j + 1)\Gamma(\alpha_0)}{\Gamma(\alpha_0 + 1)\Gamma(\alpha_j)} = \frac{\alpha_j \Gamma(\alpha_j)\Gamma(\alpha_0)}{\alpha_0 \Gamma(\alpha_0)\Gamma(\alpha_j)} \\
&= \frac{\alpha_j}{\alpha_0}.
\end{aligned}$$

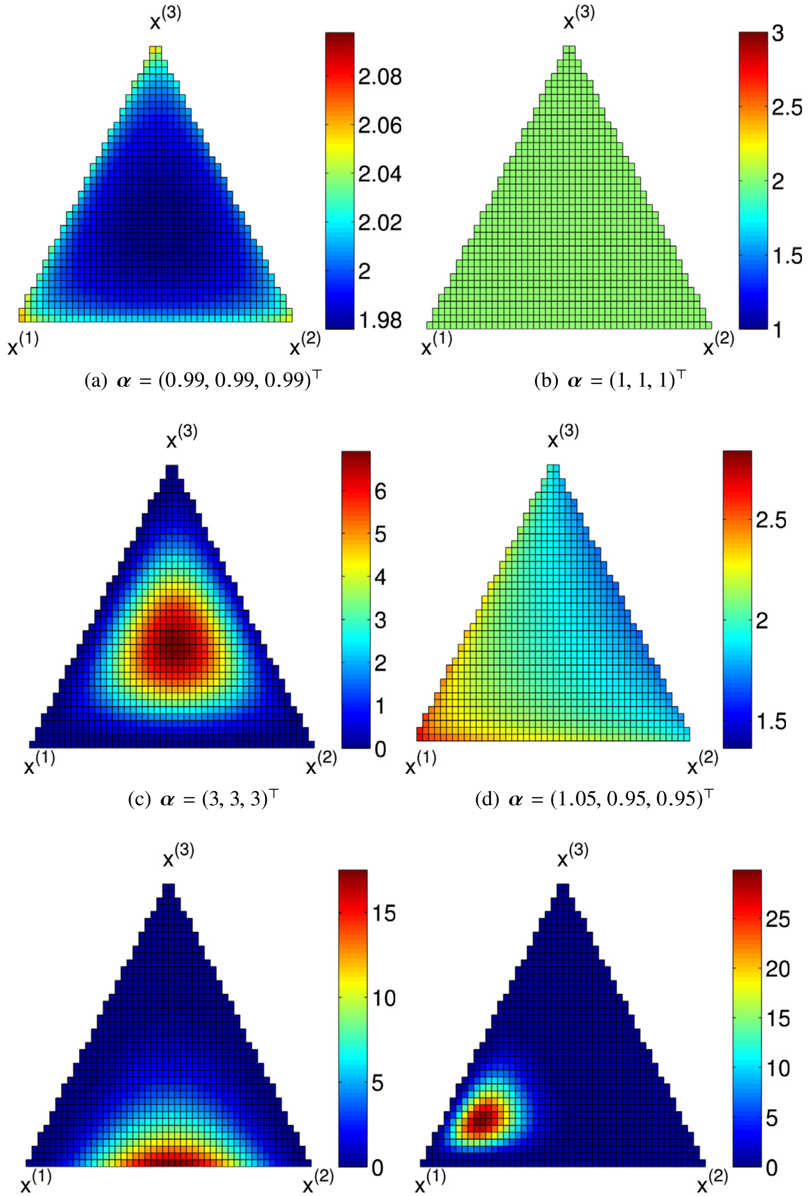
The variance and covariance of $\mathrm{Dir}(\boldsymbol{\alpha})$ can also be obtained similarly as

$$\mathrm{Cov}[x^{(j)}, x^{(j')}] = \begin{cases} \dfrac{\alpha_j(\alpha_0 - \alpha_j)}{\alpha_0^2(\alpha_0 + 1)} & (j = j'), \\ -\dfrac{\alpha_j \alpha_{j'}}{\alpha_0^2(\alpha_0 + 1)} & (j \neq j'). \end{cases}$$

Probability density functions of $\mathrm{Dir}(\boldsymbol{\alpha})$ when $d = 3$ are illustrated in Fig. 6.4. When $\alpha_1 = \alpha_2 = \alpha_3 = \alpha$, the probability density takes larger values around the corners of the triangle if each $\alpha < 1$, the probability density is uniform if each $\alpha = 1$, and the probability density takes larger values around the center if each $\alpha > 1$. When $\alpha_1, \alpha_2, \alpha_3$ are nonuniform, the probability density tends to take large values around the corners of the triangle with large α_j.

When the Dirichlet parameters $\boldsymbol{\alpha}$ are all equal, i.e.,

$$\alpha_1 = \cdots = \alpha_d = \alpha,$$



(a) $\boldsymbol{\alpha} = (0.99, 0.99, 0.99)^\top$ (b) $\boldsymbol{\alpha} = (1, 1, 1)^\top$

(c) $\boldsymbol{\alpha} = (3, 3, 3)^\top$ (d) $\boldsymbol{\alpha} = (1.05, 0.95, 0.95)^\top$

(e) $\boldsymbol{\alpha} = (4, 4, 1)^\top$ (f) $\boldsymbol{\alpha} = (16, 4, 4)^\top$

FIGURE 6.4

Probability density functions of Dirichlet distribution $\mathrm{Dir}(\boldsymbol{\alpha})$. The center of gravity of the triangle corresponds to $x^{(1)} = x^{(2)} = x^{(3)} = 1/3$, and each vertex represents the point that the corresponding variable takes one and the others take zeros.

the probability density function is simplified as

$$f(\boldsymbol{x}) = \frac{\Gamma(\alpha d)}{\Gamma(\alpha)^d} \prod_{j=1}^{d} (x^{(j)})^{\alpha-1}.$$

This is called *symmetric Dirichlet distribution* and is denoted by $\mathrm{Dir}(\alpha)$. The common parameter α is often referred to as the *concentration parameter*.

6.4 WISHART DISTRIBUTION

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be d-dimensional random variables independently following the normal distribution $N(\boldsymbol{0}, \boldsymbol{\Sigma})$, and let

$$\boldsymbol{S} = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top$$

be the *scatter matrix* which is assumed to be invertible. The probability distribution that $\boldsymbol{S}$ follows is called the *Wishart distribution* with n degrees of freedom and is denoted by $W(\boldsymbol{\Sigma}, d, n)$. The probability density function of $W(\boldsymbol{\Sigma}, d, n)$ is given by

$$f(\boldsymbol{S}) = \frac{\det(\boldsymbol{S})^{\frac{n-d-1}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}\right)\right)}{\det(2\boldsymbol{\Sigma})^{\frac{n}{2}} \Gamma_d(\frac{n}{2})},$$

where $\det(\cdot)$ denotes the determinant of a matrix, $\mathrm{tr}(\cdot)$ denotes the trace of a matrix, $\Gamma_d(\cdot)$ denotes the d-dimensional *gamma function* defined by

$$\Gamma_d(a) = \int_{\boldsymbol{S} \in \mathbb{S}_d^+} \det(\boldsymbol{S})^{a-\frac{d+1}{2}} \exp\left(-\mathrm{tr}(\boldsymbol{S})\right) \mathrm{d}\boldsymbol{S}, \tag{6.2}$$

and $\mathbb{S}_d^+$ denotes the set of all $d \times d$ positive symmetric matrices.

When $\boldsymbol{\Sigma} = \frac{1}{2}\boldsymbol{I}$, the definition of d-dimensional gamma function immediately yields

$$\int_{\boldsymbol{S} \in \mathbb{S}_d^+} f(\boldsymbol{S}) \mathrm{d}\boldsymbol{S} = 1. \tag{6.3}$$

Since the Jacobian is given by $\det(2\boldsymbol{\Sigma})^{\frac{d+1}{2}}$ when $\boldsymbol{S}$ is transformed to $\frac{1}{2}\boldsymbol{\Sigma}^{-1}\boldsymbol{S}$ (Fig. 4.2), Eq. (6.3) is also satisfied for generic $\boldsymbol{\Sigma} \in \mathbb{S}_d^+$.

d-dimensional gamma function $\Gamma_d(\cdot)$ can be expressed by using two-dimensional gamma function $\Gamma(\cdot)$ as

$$\begin{aligned} \Gamma_d(a) &= \pi^{\frac{d(d-1)}{4}} \prod_{j=1}^{d} \Gamma\left(a + \frac{1-j}{2}\right) \\ &= \pi^{\frac{d-1}{2}} \Gamma_{d-1}(a) \Gamma\left(a + \frac{1-d}{2}\right). \end{aligned}$$

An operator that transforms an $m \times n$ matrix $\boldsymbol{A} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n)$ to the mn-dimensional vector,

$$\mathrm{vec}(\boldsymbol{A}) = (\boldsymbol{a}_1^\top, \ldots, \boldsymbol{a}_n^\top)^\top,$$

is called the *vectorization operator*. The operator that transforms an $m \times n$ matrix $\boldsymbol{A}$ and a $p \times q$ matrix $\boldsymbol{B}$ to an $mp \times nq$ matrix as

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{pmatrix} a_{1,1}\boldsymbol{B} & \cdots & a_{1,n}\boldsymbol{B} \\ \vdots & \ddots & \vdots \\ a_{m,1}\boldsymbol{B} & \cdots & a_{m,n}\boldsymbol{B} \end{pmatrix}$$

is called the *Kronecker product*. The vectorization operator and the Kronecker product satisfy the following properties:

$$\begin{aligned}
\mathrm{vec}(\boldsymbol{ABC}) &= (\boldsymbol{C}^\top \otimes \boldsymbol{A})\mathrm{vec}(\boldsymbol{B}) \\
&= (\boldsymbol{I} \otimes \boldsymbol{AB})\mathrm{vec}(\boldsymbol{C}) \\
&= (\boldsymbol{C}^\top\boldsymbol{B}^\top \otimes \boldsymbol{I})\mathrm{vec}(\boldsymbol{A}), \\
(\boldsymbol{A} \otimes \boldsymbol{B})^{-1} &= \boldsymbol{A}^{-1} \otimes \boldsymbol{B}^{-1}, \\
(\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{C} \otimes \boldsymbol{D}) &= (\boldsymbol{AC} \otimes \boldsymbol{BD}), \\
\mathrm{tr}(\boldsymbol{A} \otimes \boldsymbol{B}) &= \mathrm{tr}(\boldsymbol{A})\mathrm{tr}(\boldsymbol{B}), \\
\mathrm{tr}(\boldsymbol{AB}) &= \mathrm{vec}(\boldsymbol{A}^\top)^\top \mathrm{vec}(\boldsymbol{B}), \\
\mathrm{tr}(\boldsymbol{ABCD}) &= \mathrm{vec}(\boldsymbol{A}^\top)^\top (\boldsymbol{D}^\top \otimes \boldsymbol{B})\mathrm{vec}(\boldsymbol{C}).
\end{aligned}$$

These formulas allow us to compute the product of big matrices efficiently.

FIGURE 6.5

Vectorization operator and Kronecker product.

Thus, the Wishart distribution can be regarded as a multidimensional extension of the gamma distribution. When $d = 1$, $\boldsymbol{S}$ and $\boldsymbol{\Sigma}$ become scalars, and letting $S = x$ and $\Sigma = 1$ yields

$$f(x) = \frac{x^{\frac{n}{2}-1} \exp\left(-\frac{x}{2}\right)}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)}.$$

This is equivalent to the probability density function of the chi-squared distribution with n degrees of freedom explained in Section 4.3.

The moment-generating function of $W(\boldsymbol{\Sigma}, d, n)$ can be obtained by transforming $\boldsymbol{S}$ to $(\frac{1}{2}\boldsymbol{\Sigma}^{-1} - \boldsymbol{T})\boldsymbol{S}$ as

$$\begin{aligned} M_{\boldsymbol{S}}(\boldsymbol{T}) &= E[e^{\mathrm{tr}(\boldsymbol{T}\boldsymbol{S})}] \\ &= \int_{\boldsymbol{S}\in\mathbb{S}_d^+} \frac{\det(\boldsymbol{S})^{\frac{n-d-1}{2}} \exp\left(-\mathrm{tr}\left((\frac{1}{2}\boldsymbol{\Sigma}^{-1} - \boldsymbol{T})\boldsymbol{S}\right)\right)}{2^{\frac{dn}{2}} \det(\boldsymbol{\Sigma})^{\frac{n}{2}} \Gamma_d(\frac{n}{2})} \mathrm{d}\boldsymbol{S} \\ &= \det(\boldsymbol{I} - 2\boldsymbol{T}\boldsymbol{\Sigma})^{-\frac{n}{2}}. \end{aligned}$$

The expectation and variance-covariance matrices of $\boldsymbol{S}$ that follows $W(\boldsymbol{\Sigma}, d, n)$ are given by

$$E[\boldsymbol{S}] = n\boldsymbol{\Sigma} \quad \text{and} \quad V[\mathrm{vec}(\boldsymbol{S})] = 2n\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma},$$

where $\mathrm{vec}(\cdot)$ is the vectorization operator and $\otimes$ denotes the *Kronecker product* (see Fig. 6.5).