Chapter 12

The Multivariate Normal Distribution

12.1. DEFINITION OF THE MVN

The *multivariate normal* (MVN) distribution is the natural generalization of the Gaussian, or normal distribution (Section 4.4.2) to multivariate, or vector data. The MVN is by no means the only known continuous parametric multivariate distribution (e.g., Johnson and Kotz, 1972; Johnson, 1987), but overwhelmingly it is the most commonly used. Some of the popularity of the MVN follows from its relationship to the multivariate Central Limit Theorem, although it is also used in other settings without strong theoretical justification because it enjoys a number of convenient properties that will be outlined in this section. This convenience is often sufficiently compelling to undertake transformation of non-Gaussian multivariate data to approximate multinormality before working with them (e.g., Kelly and Krzysztofowicz, 1997), which has been a strong motivation for development of the methods described in Section 3.4.

The univariate Gaussian PDF (Equation 4.24) describes the individual, or marginal, distribution of probability density for a scalar Gaussian variable. The MVN describes the joint distribution of probability density collectively for the K variables in a vector $\boldsymbol{x}$. The univariate Gaussian PDF is visualized as the bell curve defined on the real line (i.e., in a one-dimensional space). The MVN PDF is defined on the K-dimensional space whose coordinate axes correspond to the elements of $\boldsymbol{x}$, in which multivariate distances were calculated in Sections 11.2 and 11.4.4.

The probability density function for the MVN is

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{K/2}\sqrt{\det[\Sigma]}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T[\Sigma]^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right], \tag{12.1}$$

where $\boldsymbol{\mu}$ is the K-dimensional mean vector, and $[\Sigma]$ is the $(K \times K)$ covariance matrix for the K variables in the vector $\boldsymbol{x}$. In $K = 1$ dimension, Equation 12.1 reduces to Equation 4.24, and for $K = 2$ it reduces to the PDF for the bivariate normal distribution (Equation 4.31). The key part of the MVN PDF is the argument of the exponential function, and regardless of the dimension of $\boldsymbol{x}$ this argument is a squared, standardized distance (i.e., the difference between $\boldsymbol{x}$ and its mean, standardized by the (co-)variance). In the general multivariate form of Equation 12.1 this distance is the Mahalanobis distance, which is a positive-definite quadratic form when $[\Sigma]$ is of full rank, and is not defined otherwise because in that case $[\Sigma]^{-1}$ does not exist. The constants outside of the exponential in Equation 12.1 serve only to ensure that the integral over the entire K-dimensional space is 1,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(\boldsymbol{x})\, dx_1\, dx_2 \cdots dx_K = 1, \tag{12.2}$$

which is the multivariate extension of Equation 4.18.

If each of the K variables in $\boldsymbol{x}$ is separately standardized according to 4.26, the result is the standardized MVN density,

Statistical Methods in the Atmospheric Sciences. https://doi.org/10.1016/B978-0-12-815823-4.00012-2

$$\phi(\boldsymbol{z}) = \frac{1}{(2\pi)^{K/2}\sqrt{\det[R]}} \exp\left[-\frac{\boldsymbol{z}^T[R]^{-1}\boldsymbol{z}}{2}\right], \tag{12.3}$$

where $[R]$ is the (Pearson) correlation matrix (e.g., Figure 3.29) for the K variables. Equation 12.3 is the multivariate generalization of Equation 4.25. The nearly universal notation for indicating that a random vector $\boldsymbol{x}$ follows a K-dimensional MVN with covariance matrix $[\Sigma]$ is

$$\boldsymbol{x} \sim N_K(\boldsymbol{\mu}, [\Sigma]) \tag{12.4a}$$

or, for standardized variables,

$$\boldsymbol{z} \sim N_K(\boldsymbol{0}, [R]), \tag{12.4b}$$

where $\boldsymbol{0}$ is the K-dimensional mean vector whose elements are all zero.

Because the only dependence of Equation 12.1 on the random vector $\boldsymbol{x}$ is through the Mahalanobis distance inside the exponential, contours of equal probability density are ellipsoids of constant Mahalanobis distance D^2 from $\boldsymbol{\mu}$. These ellipsoidal contours centered on the mean enclose the smallest regions in the K-dimensional space containing a given portion of the probability mass, and the link between the size of these ellipsoids and the enclosed probability is the χ^2 distribution:

$$\Pr\left\{D^2 = (\boldsymbol{x}-\boldsymbol{\mu})^T[\Sigma]^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \le \chi^2_K(\alpha)\right\} = \alpha. \tag{12.5}$$

Here $\chi^2_K(\alpha)$ denotes the quantile of the χ^2 distribution with K degrees of freedom, associated with cumulative probability α (Table B.3). That is, the probability of an $\boldsymbol{x}$ being within a given Mahalanobis distance D^2 of the mean is equal to the area to the left of D^2 under the χ^2 distribution with degrees of freedom $\nu = K$. As noted at the end of Section 11.4.4 the orientations of these ellipsoids are defined by the eigenvectors of $[\Sigma]$, which are also the eigenvectors of $[\Sigma]^{-1}$. Furthermore, the elongation of the ellipsoids in the directions of each of these eigenvectors is given by the square root of the product of the respective eigenvalue of $[\Sigma]$ multiplied by the relevant χ^2 quantile. For a given D^2 the (hyper-) volume enclosed by one of these ellipsoids is proportional to the square root of the determinant of $[\Sigma]$,

$$V = \frac{2(\pi D^2)^{K/2}}{K\,\Gamma(K/2)}\sqrt{\det[\Sigma]}, \tag{12.6}$$

where $\Gamma(\cdot)$ denotes the gamma function (Equation 4.7). Here the determinant of $[\Sigma]$ functions as a scalar measure of the magnitude of the matrix, in terms of the volume occupied by the probability dispersion it describes. Accordingly, $\det[\Sigma]$ is sometimes called the *generalized variance*. The determinant, and thus also the volumes enclosed by constant-D^2 ellipsoids, increases as the K variances $\sigma_{k,k}$ increase; but also the determinant and these volumes decrease as the correlations among the K variables increase, because larger correlations result in the ellipsoids being less spherical and more elongated.

Example 12.1. Probability Ellipses for the Bivariate Normal Distribution

It is easiest to visualize multivariate ideas in two dimensions. Consider the MVN distribution fit to the Ithaca and Canandaigua minimum temperature data in Table A.1. Here $K = 2$, so this is a bivariate normal distribution with sample mean vector $[13.0,\ 20.2]^{\mathrm{T}}$ and (2×2) covariance matrix as shown in Equation 11.59. Example 11.3 shows that this covariance matrix has eigenvalues $\lambda_1 = 254.76$ and $\lambda_2 = 8.29$, with corresponding eigenvectors $\boldsymbol{e}_1^{\mathrm{T}} = [0.848,\ 0.530]$ and $\boldsymbol{e}_2^{\mathrm{T}} = [-0.530,\ 0.848]$.

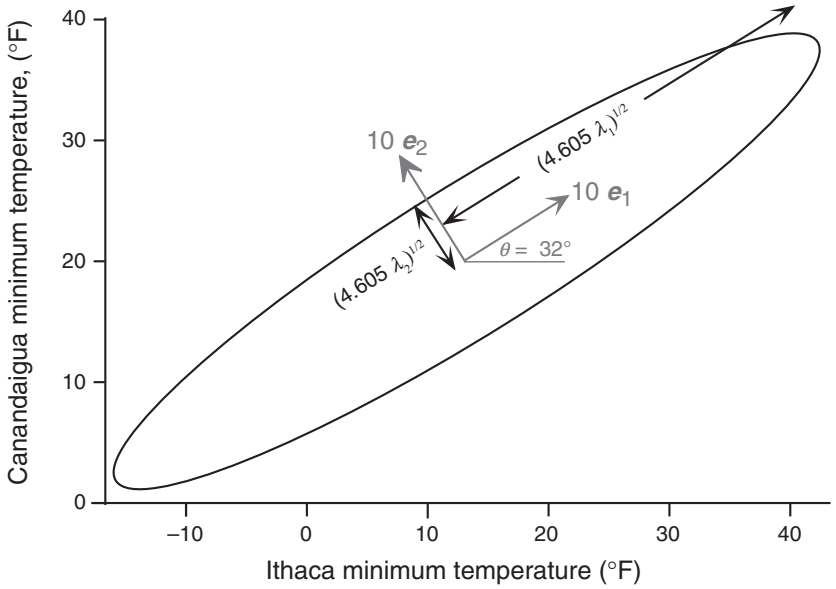

FIGURE 12.1 The 90% probability ellipse for the bivariate normal distribution representing the minimum temperature data in Table A.1, centered at the vector sample mean. Its major and minor axes are oriented in the directions of the eigenvectors *(gray)* of the covariance matrix in Equation 11.59, and stretched in these directions in proportion to the square roots of the respective eigenvalues. The constant of proportionality is the square root of the appropriate χ^2_2 quantile. The eigenvectors are drawn 10 × larger than unit length for clarity.

Figure 12.1 shows the 90% probability ellipse for this distribution. All the probability ellipses for this distribution are oriented 32 degrees from the data axes, as shown in Example 11.6. (This angle between $\boldsymbol{e}_1$ and the horizontal unit vector $[1, 0]^T$ can also be calculated using Equation 11.15.) The extent of this 90% probability ellipse in the directions of its two axes is determined by the 90% quantile of the χ^2 distribution with $\nu = K = 2$ degrees of freedom, which is ${\chi_2}^2(0.90) = 4.605$ from Table B.3. Therefore the ellipse extends to $\left({\chi_2}^2(0.90)\lambda_k\right)^{1/2}$ in the directions of each of the two eigenvectors $\boldsymbol{e}_k$; or the distances $(4.605 \cdot 254.67)^{1/2} = 34.2$ in the $\boldsymbol{e}_1$ direction, and $(4.605 \cdot 8.29)^{1/2} = 6.2$ in the $\boldsymbol{e}_2$ direction.

The volume enclosed by this ellipse is actually an area in two dimensions. From Equation 12.6 this area is $V = 2(\pi\, 4.605)^1 \sqrt{2103.26/(2 \cdot 1)} = 663.5$, since $\det[S] = 2103.26$. ◇

Example 12.2. Approximate Confidence Region for Maximum Likelihood Estimates

Section 4.6.4 noted that when a training data sample is large, the sampling distribution of maximum likelihood parameter estimates is well approximated by a MVN distribution with the same dimensionality K as the number of parameters being estimated simultaneously. Accordingly K-dimensional confidence regions for the parameters can be represented by MVN distributions. Because the large-n bias for the maximum likelihood estimators is small, these confidence regions are centered at the maximum likelihood estimates, and the corresponding covariance matrix (Equation 4.91) is the inverse of the observed Fisher information matrix (Equation 4.92).

Equation 4.87 in Example 4.13 outlines use of Newton-Raphson iteration to estimate the two parameters α and β of a gamma distribution, which involves the observed Fisher information for this problem,

$$[I(\boldsymbol{\theta})] = -\begin{bmatrix} \partial^2 L/\partial\alpha^2 & \partial^2 L/\partial\alpha\partial\beta \\ \partial^2 L/\partial\beta\partial\alpha & \partial^2 L/\partial\beta^2 \end{bmatrix} = \begin{bmatrix} n\Gamma''(\alpha) & n/\beta \\ n/\beta & -\dfrac{n\alpha}{\beta^2} + \dfrac{2\,\Sigma x}{\beta^3} \end{bmatrix}, \tag{12.7}$$

where $\boldsymbol{\theta}$ represents the parameter vector $[\alpha, \beta]^{\mathrm{T}}$. Applying the maximum likelihood algorithm in Example 4.13 to the 1933–1982 Ithaca January precipitation data in Table A.2 yields the parameter estimates $\hat{\alpha} = 3.76$ and $\hat{\beta} = 0.52$ in (Figure 4.16). Substituting these estimates into the second equality of Equation 12.7 yields

$$[I(\boldsymbol{\theta})] = \begin{bmatrix} 15.203 & 13.284 \\ 13.284 & 693.606 \end{bmatrix} \tag{12.8}$$

so that the covariance matrix appropriate to the sampling distribution of these two parameters is

$$\hat{Var}\left(\hat{\boldsymbol{\theta}}\right) = \left[I\left(\hat{\boldsymbol{\theta}}\right)\right]^{-1} = \begin{bmatrix} 0.0669 & -0.00128 \\ -0.00128 & 0.00147 \end{bmatrix}. \tag{12.9}$$

Figure 12.2 shows the resulting 95% joint confidence region for the two parameters, which is bounded by an ellipse in the $K = 2$-dimensional space. The two eigenvalues of the covariance matrix in Equation 12.9 are $\lambda_1 = 0.0669$ and $\lambda_2 = 0.00145$, and the corresponding eigenvectors (defining the directions of the dashed arrows locating the major and minor axes of the confidence ellipse) are $\boldsymbol{e}_1 = [0.9998, -0.0196]^{\mathrm{T}}$ and $\boldsymbol{e}_2 = [0.0196, 0.9998]^{\mathrm{T}}$. These eigenvectors are rotated away from the coordinate axes by only 1.1 degrees, indicating that sampling errors for the two parameters are only weakly correlated (Equation 12.9 implies a correlation of -0.129). Constructing the 95% confidence region requires finding the 95th percentile of the χ_2^2 distribution (=5.991, Table B.3), which together with the eigenvalues defines the extent of the confidence ellipse in the $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ directions. ◇

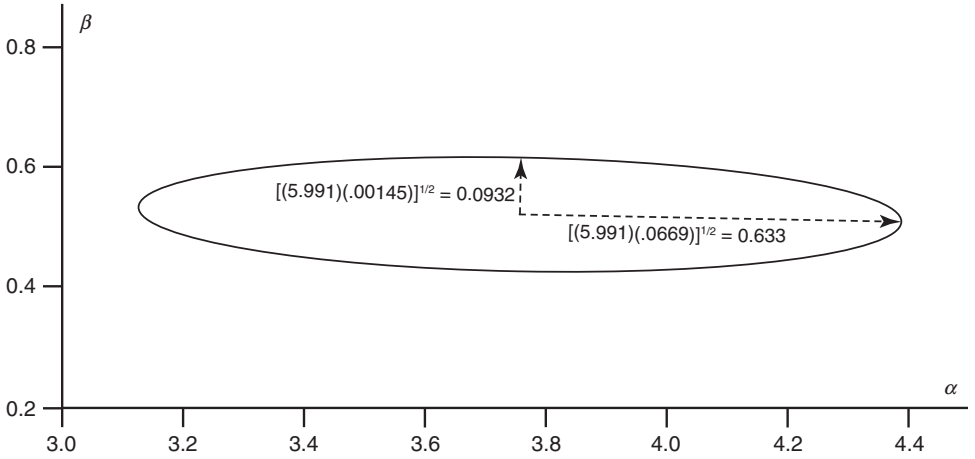

FIGURE 12.2 The 95% confidence ellipse for gamma distribution parameters fit to the 1933–1982 Ithaca January precipitation data in Table A.2.

12.2. FOUR HANDY PROPERTIES OF THE MVN

(1) *All subsets of variables from a MVN distribution are themselves distributed MVN*. Consider the partition of a $(K \times 1)$ MVN random vector $\boldsymbol{x}$ into the vectors $\boldsymbol{x}_1 = (x_1, x_2, \ldots, x_L)$, and $\boldsymbol{x}_2 = (x_{L+1}, x_{L+2}, \ldots, x_K)$, as in Equation 11.80a. Then each of these two subvectors themselves follows MVN distributions, with $\boldsymbol{x}_1 \sim N_L(\boldsymbol{\mu}_1, [\Sigma_{1,1}])$ and $\boldsymbol{x}_2 \sim N_{K-L}(\boldsymbol{\mu}_2, [\Sigma_{2,2}])$. Here the two mean vectors compose the corresponding partition of the original mean vector as in Equation 11.80b, and the covariance matrices are the indicated submatrices in Equations 11.81b and 11.81c. Note that the original ordering of the elements of $\boldsymbol{x}$ is immaterial, and that a MVN partition can be constructed from any subset. If a subset of the MVN $\boldsymbol{x}$ contains only one element (e.g., the scalar x_1) its distribution is univariate Gaussian: $x_1 \sim N_1(\mu_1, \sigma_{1,1})$. That is, this first handy property implies that all the marginal distributions for the K elements of a MVN distribution $\boldsymbol{x}$ are univariate Gaussian. The converse may not be true: it is not necessarily the case that the joint distribution of an arbitrarily selected set of K Gaussian variables will follow a MVN distribution.

(2) *Linear combinations of a MVN $\boldsymbol{x}$ are Gaussian*. If $\boldsymbol{x}$ is a MVN random vector, then a single linear combination in the form of Equation 11.82 will be univariate Gaussian with mean and variance given by Equations 11.85a and 11.85b, respectively. This fact is a consequence of the property that sums of Gaussian variables are themselves Gaussian, as noted in connection with the sketch of the Central Limit Theorem in Section 4.4.2. Similarly the result of L simultaneous linear transformations, as in Equation 11.86, will have an L-dimensional MVN distribution, with mean vector and covariance matrix given by Equations 11.87a and 11.87b, respectively, provided the covariance matrix $[\Sigma_y]$ is invertible. This condition will hold if $L \le K$, and if none of the transformed variables y_ℓ can be expressed as an exact linear combination of the others. In addition, the mean of a MVN distribution can be shifted without changing the covariance matrix. If $\boldsymbol{c}$ is a $(K \times 1)$ vector of constants then

$$\boldsymbol{x} \sim N_K(\boldsymbol{\mu}_x, [\Sigma_x]) \Rightarrow \boldsymbol{x} + \boldsymbol{c} \sim N_K(\boldsymbol{\mu}_x + \boldsymbol{c}, [\Sigma_x]). \tag{12.10}$$

(3) *Independence implies zero correlation, and vice versa, for Gaussian distributions*. Again consider the partition of a MVN $\boldsymbol{x}$ as in Equation 11.80a. If $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are independent then the off-diagonal matrices of cross-covariances in Equation 11.81 contain only zeros: $[\Sigma_{1,2}] = [\Sigma_{2,1}]^{\mathrm{T}} = [0]$. Conversely, if $[\Sigma_{1,2}] = [\Sigma_{2,1}]^{\mathrm{T}} = [0]$ then the MVN PDF can be factored as $f(\boldsymbol{x}) = f(\boldsymbol{x}_1) f(\boldsymbol{x}_2)$, implying independence (cf. Equation 2.12), because the argument inside the exponential in Equation 12.1 then breaks cleanly into two factors.

(4) *The conditional distribution of subset of a MVN $\boldsymbol{x}$, given fixed values for other elements, is also MVN*. This is the multivariate generalization of Equations 4.35, which is illustrated in Example 4.7, expressing this idea for the bivariate normal distribution. Consider again the partition $\boldsymbol{x} = [\boldsymbol{x}_1, \boldsymbol{x}_2]^{\mathrm{T}}$ as defined in Equation 11.80b and used to illustrate properties (1) and (3) before. The conditional mean of one subset of the variables $\boldsymbol{x}_1$ given particular values for the remaining variables $\boldsymbol{X}_2 = \boldsymbol{x}_2$ is

$$\boldsymbol{\mu}_1 | \boldsymbol{x}_2 = \boldsymbol{\mu}_1 + [\Sigma_{12}][\Sigma_{22}]^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2), \tag{12.11a}$$

and the conditional covariance matrix is

$$[\Sigma_{11} | \boldsymbol{x}_2] = [\Sigma_{11}] - [\Sigma_{12}][\Sigma_{22}]^{-1}[\Sigma_{21}], \tag{12.11b}$$

where the submatrices of $[\Sigma]$ are again as defined in Equation 11.81. As was the case for the bivariate normal distribution, the conditional mean shift in Equation 12.11a depends on the particular value of the conditioning variable $\boldsymbol{x}_2$, whereas the conditional covariance matrix in Equation 12.11b does not. If $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are independent, then knowledge of one provides no additional information about the other. Mathematically, if $[\Sigma_{1,2}] = [\Sigma_{2,1}]^{\mathrm{T}} = [0]$ then Equation 12.11a reduces to $\boldsymbol{\mu}_1 | \boldsymbol{x}_2 = \boldsymbol{\mu}_1$, and Equation 12.11b reduces to $[\Sigma_1 | \boldsymbol{x}_2] = [\Sigma_1]$.

Example 12.3. Three-Dimensional MVN Distributions as Cucumbers

Imagine a three-dimensional MVN PDF being represented by a cucumber, which is a solid, three-dimensional ovoid. Since the cucumber has a distinct edge, it would be more correct to imagine that it represents that part of a MVN PDF enclosed within a fixed-D^2 ellipsoidal surface. The cucumber would be an even better metaphor if its density increased toward the core and decreased toward the skin.

Figure 12.3a illustrates property (1), which is that all subsets of a MVN distribution are themselves MVN. Here are three hypothetical cucumbers floating above a kitchen cutting board in different orientations, and illuminated from above. Their shadows represent the joint distribution of the two variables whose axes are aligned with the edges of the board. Regardless of the orientation of the cucumber relative to the board (i.e., regardless of the covariance structure of the three-dimensional distribution) each of these two-dimensional joint shadow distributions for x_1 and x_2 is bivariate normal, with probability contours within fixed Mahalanobis distances of the means of the shadow ovals in the plane of the board.

Figure 12.3b illustrates property (4) that conditional distributions of subsets given particular values for the remaining variables in a MVN distribution are themselves MVN. Here portions of two cucumbers are lying on the cutting board, where the long axis of the left cucumber (indicated by the direction of the arrow or the corresponding eigenvector) is oriented parallel to the x_1 axis of the board, and the long axis of the right cucumber has been placed diagonally to the edges of the board. The three variables represented by the

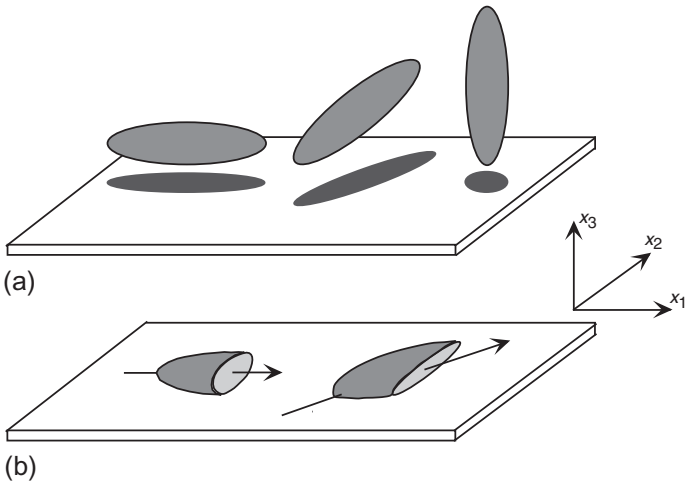

FIGURE 12.3 Three-dimensional MVN distributions as cucumbers on a kitchen cutting board. (a) Three cucumbers floating slightly above the cutting board and illuminated from above, illustrating that their shadows (the bivariate normal distributions representing the two-dimensional subsets of the original three variables in the plane of the cutting board) are ovals, regardless of the orientation (covariance structure) of the cucumber. (b) Two cucumbers resting on the cutting board, with faces exposed by cuts made perpendicularly to the x_1 coordinate axis; illustrating bivariate normality in the other two (x_2, x_3) dimensions, given the left-right location of the cut. Arrows indicate directions of the cucumber long-axis eigenvectors.

left cucumber are thus mutually independent, whereas the two horizontal (x_1 and x_2) variables for the right cucumber are positively correlated. Each cucumber has been sliced perpendicularly to the x_1 axis of the cutting board, and the exposed faces represent the joint conditional distributions of the remaining two (x_2 and x_3) variables. Both faces are ovals, illustrating that both of the resulting conditional distributions are bivariate normal. Because the cucumber on the left is oriented parallel to the cutting board edges (coordinate axes) it represents independent variables.

If parallel cuts had been made elsewhere on these cucumbers, the shapes of the exposed faces would have been the same, illustrating (as in Equation 12.11b) that the conditional covariance (shape of the exposed cucumber face) does not depend on the value of the conditioning variable (location left or right along the x_1 axis at which the cut is made). On the other hand, the conditional means (the centers of the exposed faces projected onto the $x_2 - x_3$ plane, Equation 12.11a) depend on the value of the conditioning variable (x_1), but only if the variables are correlated as in the right-hand cucumber. Making the cut further to the right shifts the location of the center of the exposed face of the right-hand cucumber toward the back of the board (the x_2 component of the conditional bivariate vector mean is greater). On the other hand, because the axes of the left cucumber ellipsoid are aligned with the coordinate axes, the location of the center of the exposed face in the $x_2 - x_3$ plane is the same regardless of where on the x_1 axis the cut has been made. ◇

12.3. TRANSFORMING TO, AND ASSESSING MULTINORMALITY

It was noted in Section 3.4.1 that one strong motivation for transforming data to approximate normality is the ability to use the MVN distribution to describe the joint variations of a multivariate data set. Usually either the Box-Cox power transformations (Equation 3.20), or the Yeo and Johnson (2000) generalization to possibly nonpositive data (Equation 3.23), are used. The Hinkley statistic (Equation 3.21), which reflects the degree of symmetry in a transformed univariate distribution, is the simplest way to decide among power transformations. However, when the goal is specifically to approximate a Gaussian distribution, as is the case when we hope that each of the transformed distributions will form one of the marginal distributions of a MVN distribution, it is probably better to choose transformation exponents that maximize the Gaussian likelihood function (Equation 3.22). It is also possible to choose transformation exponents simultaneously for multiple elements of $\boldsymbol{x}$, by choosing the corresponding vector of exponents $\boldsymbol{\lambda}$ that maximize the MVN likelihood function, although this approach requires substantially more computation than fitting the individual exponents independently, and in most cases is probably not worth the additional effort.

Choices other than the power transformations are also possible and may sometimes be more appropriate. For example, bimodal and/or strictly bounded data, such as might be well described by a beta distribution (see Section 4.4.6) with both parameters less than 1, will not power transform to approximate normality. However, if such data are adequately described by a parametric CDF $F(x)$, they can be transformed to approximate normality by matching cumulative probabilities, that is,

$$z_i = \Phi^{-1}[F(x_i)]. \tag{12.12}$$

Here $\Phi^{-1}[\cdot]$ is the quantile function for the standard Gaussian distribution, so Equation 12.12 transforms a data value x_i to the standard Gaussian z_i having the same cumulative probability as that associated with x_i within its CDF.

Methods for evaluating normality are necessary, both to assess the need for transformations, and to evaluate the effectiveness of candidate transformations. Mecklin and Mundfrom (2004) provide an extensive review and comparison of methods that have been proposed for assessing multivariate

normality. There is no single best approach to this problem, and in practice we usually look at multiple indicators, which may include both quantitative formal tests and qualitative graphical tools.

Because all marginal distributions of a MVN distribution are univariate Gaussian, goodness-of-fit tests are often calculated for the univariate distributions corresponding to each of the elements of the $\boldsymbol{x}$ whose multinormality is being assessed. The Filliben test for the Gaussian Q-Q plot correlation (Table 5.5) is a good choice for the specific purpose of testing Gaussian distribution. Gaussian marginal distributions are a necessary consequence of joint multinormality, but are not sufficient to guarantee it. In particular, looking only at marginal distributions will not identify the presence of multivariate outliers (e.g., Figure 11.6c), which are points that are not extreme with respect to any of the individual variables, but are unusual in the context of the overall covariance structure.

Two tests for multinormality (i.e., jointly for all K dimensions of $\boldsymbol{x}$) with respect to multivariate skewness and kurtosis are available (Mardia, 1970; Mardia et al., 1979). Both rely on the function of the point pair $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ given by

$$g_{i,j} = (\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T [S]^{-1} (\boldsymbol{x}_j - \overline{\boldsymbol{x}}), \tag{12.13}$$

where $[S]$ is the sample covariance matrix. This function is used to calculate the multivariate skewness measure

$$b_{1,K} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} g_{i,j}^3, \tag{12.14}$$

which reflects high-dimensional symmetry and will be near zero for MVN data. This test statistic can be evaluated using

$$\frac{n\, b_{1,K}}{6} \sim \chi_v^2, \tag{12.15a}$$

where the degrees-of-freedom parameter is

$$v = \frac{K(K+1)(K+2)}{6}, \tag{12.15b}$$

and the null hypothesis of multinormality, with respect to its symmetry, is rejected for sufficiently large values of $b_{1,K}$.

Multivariate kurtosis (appropriately heavy tails for the MVN relative to probability density near the center of the distribution) can be tested using the statistic

$$b_{2,K} = \frac{1}{n} \sum_{i=1}^{n} g_{i,i}^2, \tag{12.16}$$

which is equivalent to the average of $(D^2)^2$ because for this statistic $i = j$ in Equation 12.13. Under the null hypothesis of multinormality,

$$\left[\frac{b_{2,K} - K(K+2)}{8K(K+2)/n} \right]^{1/2} \sim N[0, 1]. \tag{12.17}$$

Scatterplots of variable pairs are valuable qualitative indicators of multinormality, since all subsets of variables from a MVN distribution are jointly normal also, and two-dimensional graphs are easy to plot

and grasp. Thus looking at a scatterplot matrix (see Section 3.6.5) is typically a valuable tool in assessing multinormality. Point clouds that are elliptical or circular are indicative of multinormality. Outliers away from the main scatter in one or more of the plots may be multivariate outliers, as in Figure 11.6c. Similarly, it can be valuable to look at rotating scatterplots (Section 3.6.3) of various three-dimensional subsets of $\boldsymbol{x}$.

Absence of evidence for multivariate outliers in all possible pairwise scatterplots does not guarantee that none exist in higher-dimensional combinations. An approach to exposing the possible existence of high-dimensional multivariate outliers, as well as to detecting other possible problems, is to use Equation 12.5. This equation implies that if the data $\boldsymbol{x}$ are MVN, the (univariate) distribution for $D_i^2, i = 1, \ldots, n$, is ${\chi_K}^2$. That is, the Mahalanobis distance D_i^2 from the sample mean for each $\boldsymbol{x}_i$ can be calculated, and the closeness of this distribution of D_i^2 values to the χ^2 distribution with K degrees of freedom can be evaluated. The easiest and most usual evaluation method is to visually inspect the Q-Q plot. It would also be possible to derive critical values to test the null hypothesis of multinormality according to the correlation coefficient for this kind of plot, using the method sketched in Section 5.2.5.

Because any linear combination of variables that are jointly multinormal will be univariate Gaussian, it can also be informative to look at and formally test linear combinations for Gaussian distribution. Often it is useful to look specifically at the linear combinations given by the eigenvectors of $[S]$,

$$y_i = \boldsymbol{e}_k^T \boldsymbol{x}_i. \tag{12.18}$$

It turns out that the linear combinations defined by the elements of the eigenvectors associated with the smallest eigenvalues can be particularly useful in identifying multivariate outliers, either by inspection of the Q-Q plots, or by formally testing the Q-Q correlations. (The reason behind linear combinations associated with the smallest eigenvalues being especially powerful in exposing outliers relates to principal component analysis, as explained in Section 13.1.5). Inspection of scatterplots of linear combinations in the rotated 2-dimensional spaces defined by pairs of eigenvectors of $[S]$ can also be revealing.

Example 12.4. Assessing Bivariate Normality for the Canandaigua Temperature Data

Are the January 1987 Canandaigua maximum and minimum temperature data in Table A.1 consistent with the proposition that they were drawn from a bivariate normal distribution? Figure 12.4 presents four plots indicating that this assumption is not unreasonable, considering the rather small sample size.

Figures 12.4a and b are Gaussian Q-Q plots for the maximum and minimum temperatures, respectively. The temperatures are plotted as functions of the standard Gaussian variables with the same cumulative probability, which has been estimated using a median plotting position (Table 3.2). Both plots are close to linear, supporting the notion that each of the two data batches was drawn from univariate Gaussian distributions. Somewhat more quantitatively, the correlations of the points in these two panels are 0.984 for the maximum temperatures and 0.978 for the minimum temperatures. If these data were serially independent, we could refer to Table 5.5 and find that both are larger than 0.970, which is the 10% critical value for $n = 30$. Since these data are serially correlated, the Q-Q correlations provide even weaker evidence against the null hypotheses that these two marginal distributions are Gaussian. Figure 12.4c shows the scatterplot for the two variables jointly. The distribution of points appears to be reasonably elliptical, with greater density near the sample mean, $[31.77, 20.23]^{\mathrm{T}}$, and less density at the extremes. This assessment is supported by Figure 12.4d, which is the Q-Q plot for the Mahalanobis distances of each of the points from the sample mean. If the data are bivariate normal, the distribution of these D_i^2 values will be χ^2, with two degrees of freedom, which is an exponential distribution (Equations 4.52 and 4.53), with $\beta = 2$. Values of its quantile function on the horizontal axis of

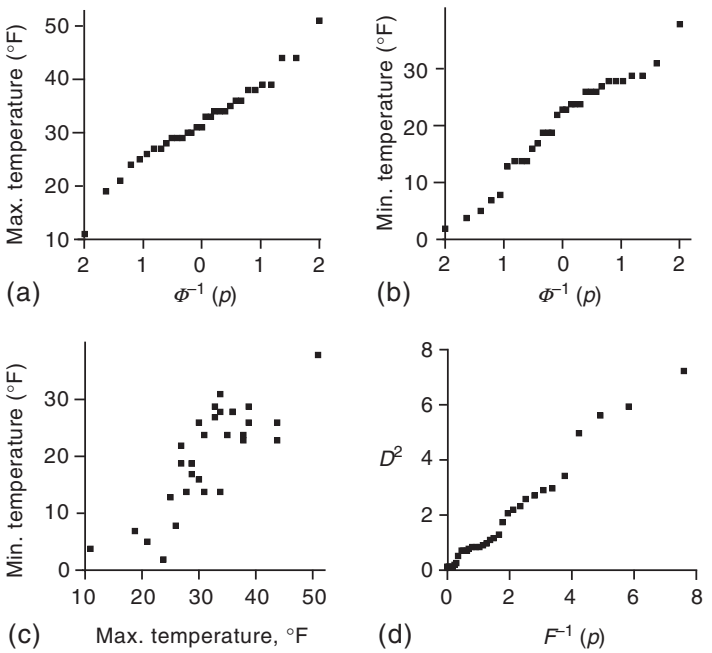

FIGURE 12.4 Graphical assessments of bivariate normality for the Canandaigua maximum and minimum temperature data. (a) Gaussian Q-Q plot for the maximum temperatures, (b) Gaussian Q-Q plot for the minimum temperatures, (c) scatterplot for the bivariate temperature data, and (D) Q-Q plot for Mahalanobis distances relative to the χ^2 distribution.

Figure 12.4d have been calculated using Equation 4.94. The points in this Q-Q plot are also reasonably straight, with the largest bivariate outlier ($D^2 = 7.23$) obtained for 25 January. This is the leftmost point in Figure 12.4c, corresponding to the coldest maximum temperature. The second-largest D^2 of 6.00 results from the data for 15 January, which is the warmest day in both the maximum and minimum temperature data.

The correlation of the points in Figure 12.4d is 0.989, but it would be inappropriate to use Table 5.5 to judge its unusualness relative to a null hypothesis that the data were drawn from a bivariate normal distribution, for two reasons. First, Table 5.5 was derived for Gaussian Q-Q plot correlations, and the null distribution (under the hypothesis of MVN data) for the Mahalanobis distance is χ^2. In addition, these data are not independent. However, it would be possible to derive critical values analogous to those in Table 5.5, by synthetically generating a large number of samples from a bivariate normal distribution with (bivariate) time correlations that simulate those in the Canandaigua temperatures, calculating the D^2 Q-Q plot for each of these samples, and tabulating the distribution of the resulting correlations. Methods appropriate to constructing such simulations are described in the next section. ◇

12.4. SIMULATION FROM THE MULTIVARIATE NORMAL DISTRIBUTION

12.4.1. Simulating Independent MVN Variates

Statistical simulation of MVN variates is accomplished through an extension of the univariate ideas presented in Section 4.7. Generation of synthetic MVN values takes advantage of the second property in Section 12.2 that linear combinations of MVN values are themselves MVN. In particular, realizations

of K-dimensional MVN vectors $\boldsymbol{x} \sim \mathrm{N}_K(\boldsymbol{\mu}, [\Sigma])$ are generated as linear combinations of K-dimensional standard MVN vectors $\boldsymbol{z} \sim \mathrm{N}_K(\mathbf{0}, [I])$, each of the K elements of which are independent standard univariate Gaussian. These standard univariate Gaussian realizations are in turn generated on the basis of uniform variates (see Section 4.7.1) transformed according to an algorithm such as that described in Section 4.7.4.

Specifically, the linear combinations used to generate MVN variates with a given mean vector and covariance matrix are given by the rows of a square-root matrix (see Section 11.3.4) for $[\Sigma]$, with the appropriate element of the mean vector added:

$$\boldsymbol{x}_i = [\Sigma]^{1/2}\boldsymbol{z}_i + \boldsymbol{\mu}. \tag{12.19}$$

As a linear combination of the K standard Gaussian values in the vector $\boldsymbol{z}$, the generated vectors $\boldsymbol{x}$ will have a MVN distribution. It is straightforward to see that they will also have the correct mean vector and covariance matrix:

$$E(\boldsymbol{x}) = E\left([\Sigma]^{1/2}\boldsymbol{z} + \boldsymbol{\mu}\right) = [\Sigma]^{1/2}E(\boldsymbol{z}) + \boldsymbol{\mu} = \boldsymbol{\mu} \tag{12.20a}$$

because $\mathrm{E}(\boldsymbol{z}) = \mathbf{0}$, and

$$\begin{aligned}[\Sigma_x] &= [\Sigma]^{1/2}[\Sigma_z]\left([\Sigma]^{1/2}\right)^T = [\Sigma]^{1/2}[I]\left([\Sigma]^{1/2}\right)^T \\ &= [\Sigma]^{1/2}\left([\Sigma]^{1/2}\right)^T = [\Sigma]\end{aligned}. \tag{12.20b}$$

Different choices for the nonunique matrix $[\Sigma]^{1/2}$ will yield different simulated $\boldsymbol{x}$ vectors for a given input $\boldsymbol{z}$, but Equation 12.20 shows that collectively, the resulting $\boldsymbol{x} \sim N_K(\boldsymbol{\mu}, [\Sigma])$ so long as $[\Sigma]^{1/2}([\Sigma]^{1/2})^{\mathrm{T}} = [\Sigma]$.

It is interesting to note that the transformation in Equation 12.19 can be inverted to produce standard MVN vectors $\boldsymbol{z} \sim N_K(\mathbf{0}, [I])$ corresponding to MVN vectors $\boldsymbol{x}$ of known distributions. Usually this manipulation is done to transform a sample of vectors $\boldsymbol{x}$ to the standard MVN according to their estimated mean vector and covariance matrix, analogously to the standardized anomaly (Equation 3.27),

$$\boldsymbol{z}_i = [S]^{-1/2}(\boldsymbol{x}_i - \overline{\boldsymbol{x}}) = [S]^{-1/2}\boldsymbol{x}_i'. \tag{12.21}$$

This relationship is called the *Mahalanobis transformation*. It is distinct from the scaling transformation (Equation 11.34), which produces a vector of standard Gaussian variates having unchanged correlation structure. It is straightforward to show that Equation 12.21 produces uncorrelated z_k values, each with unit variance:

$$\begin{aligned}[S_z] &= [S_x]^{-1/2}[S_x]\left([S_x]^{-1/2}\right)^T \\ &= [S_x]^{-1/2}[S_x]^{1/2}\left([S_x]^{1/2}\right)^T\left([S_x]^{-1/2}\right)^T = [I][I] = [I]\end{aligned}. \tag{12.22}$$

12.4.2. Simulating Multivariate Time Series

The autoregressive processes for scalar time series described in Sections 10.3.1 and 10.3.2 can be generalized to stationary multivariate, or vector, time series. In this case the variable $\boldsymbol{x}$ is a vector quantity

observed at discrete and regularly spaced time intervals. The multivariate generalization of the AR(p) process in Equation 10.23 is

$$\boldsymbol{x}_{t+1} - \boldsymbol{\mu} = \sum_{i=1}^{p} [\Phi_i](\boldsymbol{x}_{t-i+1} - \boldsymbol{\mu}) + [B]\boldsymbol{\varepsilon}_{t+1}. \tag{12.23}$$

Here the elements of the vector $\boldsymbol{x}$ consist of a set of K correlated time series, $\boldsymbol{\mu}$ contains the corresponding mean vector, and the elements of the vector $\boldsymbol{\varepsilon}$ are mutually independent (and usually Gaussian) random variables with zero mean and unit variance. The matrices of autoregressive parameters $[\Phi_i]$ correspond to the scalar autoregressive parameters ϕ_k in Equation 10.23. The matrix $[B]$, operating on the vector $\boldsymbol{\varepsilon}_{t+1}$, allows the random components in Equation 12.23 to have different variances, and to be mutually correlated at each time step (although they are uncorrelated in time). Note that the order, p, of the autoregression was denoted as K in Chapter 10 and does not indicate the dimension of a vector there. Multivariate autoregressive-moving average models, extending the scalar models in Section 10.3.6 to vector data, can also be defined.

The multivariate AR(1) process is the most common special case of Equation 12.20,

$$\boldsymbol{x}_{t+1} - \boldsymbol{\mu} = [\Phi](\boldsymbol{x}_t - \boldsymbol{\mu}) + [B]\,\boldsymbol{\varepsilon}_{t+1}, \tag{12.24}$$

which is obtained from Equation 12.23 for the autoregressive order $p = 1$. It is the multivariate generalization of Equation 10.16 and will be stationary process if all the eigenvalues of $[\Phi]$ are between -1 and 1. Matalas (1967) and Bras and Rodríguez-Iturbe (1985) describe use of Equation 12.24 in hydrology, where the elements of $\boldsymbol{x}$ are typically simultaneously measured (possibly transformed) streamflows at different locations. This equation is also often used as part of a common synthetic *weather generator* formulation (Richardson, 1981). In this second application $\boldsymbol{x}$ usually has three elements, corresponding to daily maximum temperature, minimum temperature, and solar radiation at a given location.

The two parameter matrices in Equation 12.24 are most easily estimated using the simultaneous and lagged covariances among the elements of $\boldsymbol{x}$. The simultaneous covariances are contained in the usual covariance matrix $[S]$, and the lagged covariances are contained in the matrix

$$[S_1] = \frac{1}{n-1}\sum_{t=1}^{n-1} \boldsymbol{x}'_{t+1}\boldsymbol{x}'^{T}_{t} \tag{12.25a}$$

$$= \begin{bmatrix} s_1(1\rightarrow 1) & s_1(2\rightarrow 1) & \cdots & s_1(K\rightarrow 1) \\ s_1(1\rightarrow 2) & s_1(2\rightarrow 2) & \cdots & s_1(K\rightarrow 2) \\ \vdots & \vdots & \ddots & \vdots \\ s_1(1\rightarrow K) & s_1(2\rightarrow K) & \cdots & s_1(K\rightarrow K) \end{bmatrix}. \tag{12.25b}$$

This equation is similar to Equation 11.35 for $[S]$, except that the pairs of vectors whose outer products are summed are data (anomalies) at pairs of successive time points. The diagonal elements of $[S_1]$ are the lag-1 autocovariances (the lagged autocorrelations in Equation 3.36 multiplied by the respective variances, as in Equation 3.39) for each of the K elements of $\boldsymbol{x}$. The off-diagonal elements of $[S_1]$ are the lagged covariances among unlike elements of $\boldsymbol{x}$. The arrow notation in this equation indicates the time sequence of the lagging of the variables. For example, $s_1(1\rightarrow 2)$ denotes the correlation between x_1 at time t, and x_2 at time $t+1$, and $s_1\,(2\rightarrow 1)$ denotes the correlation between x_2 at time t, and x_1 at time t+1. Notice that the matrix $[S]$ is symmetric, but that in general $[S_1]$ is not.

The matrix of autoregressive parameters $[\Phi]$ in Equation 12.24 is obtained from the lagged and unlagged covariance matrices using

$$[\Phi] = [S_1]\,[S]^{-1}. \tag{12.26}$$

Obtaining the matrix $[B]$ requires finding a matrix square root (Section 11.3.4) of

$$[B]\,[B]^T = [S] - [\Phi]\,[S_1]^T. \tag{12.27}$$

Chapman et al. (2015) provide a generalization of the Yule-Walker equations (Equation 10.24) for estimating the parameter matrices for the general vector autoregression (Equation 12.23) with order $p \geq 2$.

Having defined a multivariate autoregressive model, it is straightforward to simulate from it using the defining equation (e.g., Equation 11.24) together with an appropriate random number generator to provide time series of realizations for the random-forcing vector $\boldsymbol{\varepsilon}$. Usually these are taken to be standard Gaussian, in which case they can be generated using the algorithm described in Section 4.7.4. In any case the K elements of $\boldsymbol{\varepsilon}$ will have zero mean and unit variance, will be uncorrelated with each other at any one time t, and will be uncorrelated with other forcing vectors at different times $t+\tau$:

$$E[\boldsymbol{\varepsilon}_t] = \mathbf{0} \tag{12.28a}$$

$$E\left[\boldsymbol{\varepsilon}_t\,\boldsymbol{\varepsilon}_t^T\right] = [I] \tag{12.28b}$$

$$E\left[\boldsymbol{\varepsilon}_t\,\boldsymbol{\varepsilon}_{t+\tau}^T\right] = [0], \tau \neq 0. \tag{12.28c}$$

If the $\boldsymbol{\varepsilon}$ vectors contain realizations of independent Gaussian variates, then the resulting $\boldsymbol{x}$ vectors will have a MVN distribution, because they are linear combinations of (standard) MVN vectors $\boldsymbol{\varepsilon}$. If the original data that the simulated series are meant to emulate are clearly non-Gaussian, they may be transformed before fitting the time series model.

Example 12.5. Fitting and Simulating from a Bivariate Autoregression

Example 12.4 examined the Canandaigua maximum and minimum temperature data in Table A.1, and concluded that the MVN distribution is a reasonable model for their joint variations. The first-order autoregression (Equation 12.24) is a reasonable model for their time dependence, and fitting the parameter matrices $[\Phi]$ and $[B]$ will allow statistical simulation of synthetic bivariate series that statistically resemble these data. This process can be regarded as an extension of Example 10.3, which illustrated the univariate AR(1) model for the time series of Canandaigua minimum temperatures alone.

The sample statistics necessary to fit Equation 12.24 are easily computed from the Canandaigua temperature data in Table A.1 as

$$\overline{\boldsymbol{x}} = [31.77\ 20.23]^T \tag{12.29a}$$

$$[S] = \begin{bmatrix} 61.85 & 56.12 \\ 56.12 & 77.58 \end{bmatrix} \tag{12.29b}$$

and

$$[S_1] = \begin{bmatrix} s_{\max\rightarrow\max} & s_{\min\rightarrow\max} \\ s_{\max\rightarrow\min} & s_{\min\rightarrow\min} \end{bmatrix} = \begin{bmatrix} 37.32 & 44.51 \\ 42.11 & 51.33 \end{bmatrix}. \tag{12.29c}$$

The matrix of simultaneous covariances is the ordinary covariance matrix $[S]$, which is of course symmetric. The matrix of lagged covariances (Equation 12.29c) is not symmetric. Using Equation 12.26, the estimated matrix of autoregressive parameters is

$$[\Phi] = [S_1]\,[S]^{-1} = \begin{bmatrix} 37.32 & 44.51 \\ 42.11 & 51.33 \end{bmatrix} \begin{bmatrix} .04705 & -.03404 \\ -.03404 & .03751 \end{bmatrix} = \begin{bmatrix} .241 & .399 \\ .234 & .492 \end{bmatrix}. \tag{12.30}$$

The matrix $[B]$ can be anything satisfying (c.f. Equation 12.27)

$$[B][B]^T = \begin{bmatrix} 61.85 & 56.12 \\ 56.12 & 77.58 \end{bmatrix} - \begin{bmatrix} .241 & .399 \\ .234 & .492 \end{bmatrix} \begin{bmatrix} 37.32 & 42.11 \\ 44.51 & 51.33 \end{bmatrix} = \begin{bmatrix} 35.10 & 25.49 \\ 25.49 & 42.47 \end{bmatrix}, \tag{12.31}$$

with one solution given by the Cholesky factorization (Equations 11.65 and 11.66),

$$[B] = \begin{bmatrix} 5.92 & 0 \\ 4.31 & 4.89 \end{bmatrix}. \tag{12.32}$$

Using the estimated values in Equations 12.30 and 12.32, and substituting the sample mean from Equation 12.29a for the mean vector, Equation 12.24 becomes an algorithm for simulating bivariate $\boldsymbol{x}_t$ series with the same (sample) first- and second-moment statistics as the Canandaigua temperatures in Table A.1. The Box-Muller algorithm (see Section 4.7.4) is especially convenient for generating the vectors $\boldsymbol{\varepsilon}_t$ in this case because it produces them in pairs. Figure 12.5a shows a 100-point realization of a bivariate time series generated in this way. Here the vertical lines connect the simulated maximum and minimum temperatures for a given day, and the light horizontal lines locate the two mean values (Equation 12.29a). These two time series statistically resemble the January 1987 Canandaigua temperature data to the extent that Equation 12.24 is capable of doing so. They are unrealistic in the sense that the generating-process statistics do not change through the 100 simulated days, since the underlying generating model is covariance stationary. That is, the means, variances, and covariances are constant throughout the 100 time points, whereas in nature these statistics change over the course of a winter. Also, the time series is potentially unrealistic in the sense that it is possible (although rare) to statistically simulate maximum temperatures that are colder than the simulated minimum temperature for the day. Recalculating the simulation, but starting from a different random number seed, would yield a different series, but with the same statistical characteristics.

Figure 12.5b shows a scatterplot for the 100 point pairs, corresponding to the scatterplot of the actual data in the lower-right panel of Figure 3.31. Since the points were generated by forcing Equation 12.24

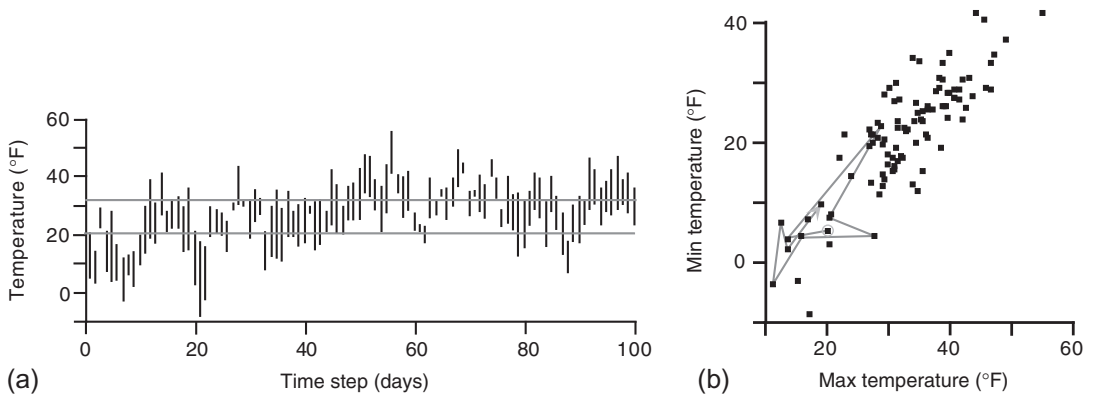

FIGURE 12.5 (a) A 100-point realization from the bivariate AR(1) process fit to the January 1987 Canandaigua daily maximum and minimum temperatures. Vertical lines connect the simulated maximum and minimum for each day, and light horizontal lines locate the two means. (b) Scatterplot of the 100 bivariate points. Light gray line segments connect the first 10 pairs of values.

with synthetic Gaussian variates for the elements of $\boldsymbol{\varepsilon}$, the resulting distribution for $\boldsymbol{x}$ is bivariate normal by construction. However, the points are not independent and exhibit time correlation mimicking that found in the original data series. The result is that successive points do not appear at random within the scatterplot, but rather tend to cluster. The light gray line illustrates this time dependence by tracing a path from the first point (circled) to the tenth point (indicated by the arrow tip). ◇

Since the statistics underlying Figure 12.5a remained constant throughout the simulation, it is a realization of a stationary time series—in this case a perpetual January. Simulations of this kind can be made to be more realistic by allowing the parameters, based on the statistics in Equations 12.29, to vary periodically through an annual cycle. The result would be a *cyclostationary* autoregression whose statistics are different for different dates, but the same on the same date in different years. Cyclostationary autoregressions are described in Richardson (1981), Von Storch and Zwiers (1999), and Wilks and Wilby (1999), among others.

12.5. INFERENCES ABOUT A MULTINORMAL MEAN VECTOR

This section describes parametric multivariate hypothesis tests concerning mean vectors, based on the MVN distribution. There are many instances where multivariate nonparametric approaches are more appropriate. Some of these multivariate nonparametric tests have been described, as extensions to their univariate counterparts, in Sections 5.3 and 5.4. The parametric tests described in this section require the invertibility of the sample covariance matrix of $\boldsymbol{x}$, $[S_x]$, and so will be infeasible if $n \leq K$. In that case nonparametric tests would be indicated. Even if $[S_x]$ is invertible, the resulting parametric test may have disappointing power unless $n \gg K$, and this limitation can be another reason to choose a nonparametric alternative.

12.5.1. Multivariate Central Limit Theorem

The Central Limit Theorem for univariate data was described briefly in Section 4.4.2, and again more quantitatively in Section 5.2.1. It states that the sampling distribution of the average of a sufficiently large number, n, of random variables will be Gaussian, and that if the variables being averaged are mutually independent the variance of that sampling distribution will be smaller than the variance of the original variables by the factor $1/n$. The multivariate generalization of the Central Limit Theorem states that the sampling distribution of the mean of n independent random $(K \times 1)$ vectors $\boldsymbol{x}$ with mean $\boldsymbol{\mu}_x$ and covariance matrix $[\Sigma_x]$ will be MVN with the same covariance matrix, again scaled by the factor $1/n$. That is,

$$\overline{\boldsymbol{x}} \sim N_K\left(\boldsymbol{\mu}_x, \frac{1}{n}[\Sigma_x]\right) \tag{12.33a}$$

or, equivalently

$$\sqrt{n}(\overline{\boldsymbol{x}} - \boldsymbol{\mu}_x) \sim N_K(0, [\Sigma_x]). \tag{12.33b}$$

If the random vectors $\boldsymbol{x}$ being averaged are themselves MVN, then the distributions indicated in Equations 12.33 are exact, because then the sample mean vector is a linear combination of the MVN vectors $\boldsymbol{x}$. Otherwise, the multinormality for the sample mean is approximate, and that approximation improves as the sample size n increases.

Multinormality for the sampling distribution of the sample mean vector implies that the sampling distribution for the Mahalanobis distance between the sample and population means will be χ^2. That is, assuming that $[\Sigma_x]$ is known, Equation 12.5 implies that

$$(\overline{\boldsymbol{x}} - \boldsymbol{\mu})^T \left(\frac{1}{n}[\Sigma_x]\right)^{-1} (\overline{\boldsymbol{x}} - \boldsymbol{\mu}) \sim \chi^2_K, \tag{12.34a}$$

or

$$n\,(\overline{\boldsymbol{x}} - \boldsymbol{\mu})^T [\Sigma_x]^{-1} (\overline{\boldsymbol{x}} - \boldsymbol{\mu}) \sim \chi^2_K. \tag{12.34b}$$

12.5.2. Hotelling's T^2

Usually inferences about means must be made without knowing the population variance, and this is true in both univariate and multivariate settings. Substituting the estimated covariance matrix into Equation 12.34 yields the one-sample *Hotelling* T^2 *statistic*,

$$T^2 = (\overline{\boldsymbol{x}} - \boldsymbol{\mu}_0)^T \left(\frac{1}{n}[S_x]\right)^{-1} (\overline{\boldsymbol{x}} - \boldsymbol{\mu}_0) = n\,(\overline{\boldsymbol{x}} - \boldsymbol{\mu}_0)^T [S_x]^{-1} (\overline{\boldsymbol{x}} - \boldsymbol{\mu}_0). \tag{12.35}$$

Here $\boldsymbol{\mu}_0$ indicates the unknown population mean (under the null hypothesis) about which inferences will be made. Equation 12.35 is the multivariate generalization of (the square of) the univariate one-sample t statistic that is obtained by combining Equations 5.3 and 5.4. The univariate t is recovered from the square root of Equation 12.35 for scalar (i.e., $K = 1$) data. Both t and T^2 express differences between the sample mean being tested and its hypothesized true value under H_0, "divided by" an appropriate characterization of the dispersion of the null distribution. T^2 is a quadratic (and thus nonnegative) quantity, because the unambiguous ordering of univariate magnitudes on the real line that is expressed by the univariate t statistic does not generalize to higher dimensions. That is, the ordering of scalar magnitude is unambiguous (e.g., it is clear that $5 > 3$), whereas the ordering of vectors is not (e.g., is $(3, 5)^T$ larger or smaller than $(5, 3)^T$?).

The one-sample T^2 is simply the Mahalanobis distance between the vectors $\boldsymbol{x}$ and $\boldsymbol{\mu}_0$, within the context established by the estimated covariance matrix for the sampling distribution of the mean vector, $(1/n)[S_x]$. Since $\overline{\boldsymbol{x}}$ is subject to sampling variations, a continuum of T^2 values is possible, and the probabilities for these outcomes are described by a PDF. Under the null hypothesis H_0: $E(\boldsymbol{x}) = \boldsymbol{\mu}_0$, an appropriately scaled version of T^2 follows what is known as the F *distribution*,

$$\frac{(n-K)}{(n-1)\,K} T^2 \sim F_{K,n-K}. \tag{12.36}$$

The F distribution is a two-parameter distribution whose quantiles are tabulated in most introductory statistics textbooks. Both parameters are referred to as degrees-of-freedom parameters, and in the context of Equation 12.36 they are $\nu_1 = K$ and $\nu_2 = n - K$, as indicated by the subscripts in Equation 12.36. Accordingly, a null hypothesis that $E(\boldsymbol{x}) = \boldsymbol{\mu}_0$ would be rejected at the α level if

$$T^2 > \frac{(n-1)\,K}{(n-K)} F_{K,n-K}(1-\alpha), \tag{12.37}$$

where $F_{K,n-K}(1-\alpha)$ is the $1-\alpha$ quantile of the F distribution with K and $n-K$ degrees of freedom.

One way of looking at the F distribution is as the multivariate generalization of the t distribution, which is the null distribution for the t statistic in Equation 5.3. The sampling distribution of Equation 5.3 is t rather than standard univariate Gaussian, and the distribution of T^2 is F rather than χ^2 (as might have been expected from Equation 12.34), because the corresponding dispersion measures (s^2 and $[S]$, respectively) are sample estimates rather than known population values. Just as the univariate t distribution converges to the univariate standard Gaussian as its degrees-of-freedom parameter increases (and the variance s^2 is estimated increasingly more precisely), the F distribution approaches proportionality to the χ^2 with $\nu_1 = K$ degrees of freedom as the sample size (and thus also ν_2) becomes large, because $[S]$ is estimated more precisely:

$$\chi^2_K(1-\alpha) = K F_{K,\infty}(1-\alpha). \tag{12.38}$$

That is, the $(1-\alpha)$ quantile of the χ^2 distribution with K degrees of freedom is exactly a factor of K larger than the $(1-\alpha)$ quantile of the F distribution with $\nu_1 = K$ and $\nu_2 = \infty$ degrees of freedom. Since $(n-1) \approx (n-K)$ for sufficiently large n, the large-sample counterparts of Equations 12.36 and 12.37 are, to good approximation,

$$T^2 \sim \chi^2_K \tag{12.39a}$$

if the null hypothesis is true, leading to rejection at the α level if

$$T^2 > \chi^2_K(1-\alpha). \tag{12.39b}$$

Differences between χ^2 and (scaled) F quantiles are about 5% for $n-K = 100$, so that this is a reasonable rule of thumb for appropriateness of Equations 12.39 as large-sample approximations to Equations 12.36 and 12.37.

The two-sample t-test statistic (Equation 5.5) is also extended in a straightforward way to inferences regarding the difference of two independent sample mean vectors:

$$T^2 = [(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2) - \boldsymbol{\delta}_0]^T [S_{\Delta\overline{\boldsymbol{x}}}]^{-1} [(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2) - \boldsymbol{\delta}_0], \tag{12.40}$$

where

$$\boldsymbol{\delta}_0 = E[\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2] \tag{12.41}$$

is the difference between the two population mean vectors under H_0, corresponding to the second term in the numerator of Equation 5.5. If, is as often the case, the null hypothesis is that the two underlying means are equal, then $\boldsymbol{\delta}_0 = \mathbf{0}$ (corresponding to Equation 5.6). The two-sample Hotelling T^2 in Equation 12.40 is a Mahalanobis distance between the difference of the two sample mean vectors being tested and the corresponding difference of their expected values under the null hypothesis. If the null hypothesis is $\boldsymbol{\delta}_0 = \mathbf{0}$, Equation 12.40 reduces to a Mahalanobis distance between the two sample mean vectors.

The covariance matrix for the (MVN) sampling distribution of the difference of the two mean vectors is estimated differently, depending on whether the covariance matrices for the two samples, $[\Sigma_1]$ and $[\Sigma_2]$, can plausibly be assumed equal. If so, this matrix is estimated using a pooled estimate of that common covariance,

$$[S_{\Delta\overline{\boldsymbol{x}}}] = \left(\frac{1}{n_1} + \frac{1}{n_2}\right) [S_{pool}], \tag{12.42a}$$

where

$$[S_{pool}] = \frac{n_1 - 1}{n_1 + n_2 - 2}[S_1] + \frac{n_2 - 1}{n_1 + n_2 - 2}[S_2] \tag{12.42b}$$

is a weighted average of the two sample covariance matrices for the underlying data. If these two matrices cannot plausibly be assumed equal, and if in addition the sample sizes are relatively large, then the dispersion matrix for the sampling distribution of the difference of the sample mean vectors may be estimated as

$$[S_{\Delta\overline{x}}] = \frac{1}{n_1}[S_1] + \frac{1}{n_2}[S_2], \tag{12.43}$$

which is numerically equal to Equation 12.42 for $n_1 = n_2$.

If the sample sizes are not large, the two-sample null hypothesis is rejected at the α level if

$$T^2 > \frac{(n_1 + n_2 - 2)K}{(n_1 + n_2 - K - 1)} F_{K, n_1 + n_2 - K - 1}(1 - \alpha). \tag{12.44}$$

That is, critical values are proportional to quantiles of the F distribution with $\nu_1 = K$ and $\nu_2 = n_1 + n_2 - K - 1$ degrees of freedom. For ν_2 sufficiently large (>100, perhaps), Equation 12.39b can be used, as before.

Finally, if $n_1 = n_2$ and corresponding observations of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are linked physically—and correlated as a consequence—it is appropriate to account for the correlations between the pairs of observations by computing a one-sample test involving their differences. Defining $\boldsymbol{\Delta}_i$ as the difference between the ith observations of the vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, analogously to Equation 5.10, the one-sample Hotelling T^2 test statistic, corresponding to Equation 5.11, and of exactly the same form as Equation 12.35, is

$$T^2 = (\overline{\boldsymbol{\Delta}} - \boldsymbol{\mu}_\Delta)^T \left(\frac{1}{n}[S_\Delta]\right)^{-1} (\overline{\boldsymbol{\Delta}} - \boldsymbol{\mu}_\Delta) = n\,(\overline{\boldsymbol{\Delta}} - \boldsymbol{\mu}_\Delta)^T [S_\Delta]^{-1} (\overline{\boldsymbol{\Delta}} - \boldsymbol{\mu}_\Delta). \tag{12.45}$$

Here $n = n_1 = n_2$ is the common sample size, and $[S_\Delta]$ is the sample covariance matrix for the n vectors of differences $\boldsymbol{\Delta}_i$. The unusualness of Equation 12.45 in the context of the null hypothesis that the true difference of means is $\boldsymbol{\mu}_\Delta$ is evaluated using the F distribution (Equation 12.37) for relatively small samples, and the χ^2 distribution (Equation 12.39b) for large samples.

Example 12.6. Two-Sample, and One-Sample Paired T^2 Tests

Table 12.1 presents January averages of daily maximum and minimum temperatures at New York City and Boston, for the 30 years 1971 through 2000. Because these are annual values, their serial correlations are quite small. As averages of 31 daily values each, the univariate distributions of these monthly values are expected to closely approximate the Gaussian. Figure 12.6 shows scatterplots for the values at each location. The ellipsoidal dispersions of the two point clouds suggest bivariate normality for both pairs of maximum and minimum temperatures. The two scatterplots overlap somewhat, but the visual separation is sufficiently distinct to suspect strongly that their generating distributions are different.

The two vector means, and their difference vector, are

$$\overline{\boldsymbol{x}}_N = \begin{bmatrix} 38.68 \\ 26.15 \end{bmatrix}, \tag{12.46a}$$

$$\overline{\boldsymbol{x}}_B = \begin{bmatrix} 36.50 \\ 22.13 \end{bmatrix}, \tag{12.46b}$$

TABLE 12.1 Average January Maximum and Minimum Temperatures (°F) for New York City and Boston, 1971–2000, and the Corresponding Year-by-Year Differences

	New York		Boston		Differences	
Year	T_{max}	T_{min}	T_{max}	T_{min}	Δ_{max}	Δ_{min}
1971	33.1	20.8	30.9	16.6	2.2	4.2
1972	42.1	28.0	40.9	25.0	1.2	3.0
1973	42.1	28.8	39.1	23.7	3.0	5.1
1974	41.4	29.1	38.8	24.6	2.6	4.5
1975	43.3	31.3	41.4	28.4	1.9	2.9
1976	34.2	20.5	34.1	18.1	0.1	2.4
1977	27.7	16.4	29.8	16.7	−2.1	−0.3
1978	33.9	22.0	35.6	21.3	−1.7	0.7
1979	40.2	26.9	39.1	25.8	1.1	1.1
1980	39.4	28.0	35.6	23.2	3.8	4.8
1981	32.3	20.2	28.5	14.3	3.8	5.9
1982	32.5	19.6	30.5	15.2	2.0	4.4
1983	39.6	29.4	37.6	24.8	2.0	4.6
1984	35.1	24.6	32.4	20.9	2.7	3.7
1985	34.6	23.0	31.2	17.5	3.4	5.5
1986	40.8	27.4	39.6	23.1	1.2	4.3
1987	37.5	27.1	35.6	22.2	1.9	4.9
1988	35.8	23.2	35.1	20.5	0.7	2.7
1989	44.0	30.7	42.6	26.4	1.4	4.3
1990	47.5	35.2	43.3	29.5	4.2	5.7
1991	41.2	28.5	36.6	22.2	4.6	6.3
1992	42.5	28.9	38.2	23.8	4.3	5.1
1993	42.5	30.1	39.4	25.4	3.1	4.7
1994	33.2	17.9	31.0	13.4	2.2	4.5
1995	43.1	31.9	41.0	28.1	2.1	3.8
1996	37.0	24.0	37.5	22.7	−0.5	1.3
1997	39.2	25.1	36.7	21.7	2.5	3.4
1998	45.8	34.2	39.7	28.1	6.1	6.1
1999	40.8	27.0	37.5	21.5	3.3	5.5
2000	37.9	24.7	35.7	19.3	2.2	5.4

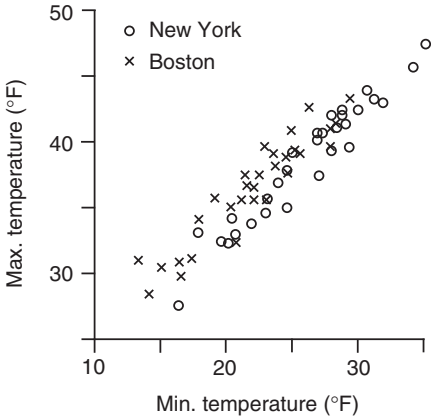

FIGURE 12.6 January average maximum and minimum temperatures, 1971–2000, for New York City (circles) and Boston (x's).

and

$$\overline{\boldsymbol{\Delta}} = \overline{\boldsymbol{x}}_N - \overline{\boldsymbol{x}}_B = \begin{bmatrix} 2.18 \\ 4.02 \end{bmatrix}. \tag{12.46c}$$

As might have been expected from its lower latitude, the average temperatures at New York are warmer. The sample covariance matrix for all four variables jointly is

$$[S] = \left[\begin{array}{c|c} [S_N] & [S_{N-B}] \\ \hline [S_{B-N}] & [S_B] \end{array}\right] = \left[\begin{array}{cc|cc} 21.485 & 21.072 & 17.150 & 17.866 \\ 21.072 & 22.090 & 16.652 & 18.854 \\ \hline 17.150 & 16.652 & 15.948 & 16.070 \\ 17.866 & 18.854 & 16.070 & 18.386 \end{array}\right]. \tag{12.47}$$

Because the two locations are relatively close to each other and the data were taken in the same years, it is appropriate to treat them as paired values. This assertion is supported by the large cross-covariances in the submatrices $[S_{B-N}] = [S_{N-B}]^T$, corresponding to correlations ranging from 0.89 to 0.94: the data at the two locations are clearly not independent of each other. Nevertheless, it is instructive to first carry through T^2 calculations for differences of mean vectors as a two-sample test, ignoring these large cross-covariances for the moment.

Regarding the Boston and New York temperatures as mutually independent, the appropriate test statistic would be Equation 12.40. If the null hypothesis is that the underlying vector means of the two distributions from which these data were drawn are equal, $\boldsymbol{\delta}_0 = \mathbf{0}$. Both the visual impressions of the two data scatters in Figure 12.6, and the similarity of the covariance matrices $[S_N]$ and $[S_B]$ in Equation 12.47, suggest that assuming equality of covariance matrices would be reasonable. The appropriate covariance for the sampling distribution of the mean difference would then be calculated using Equation 12.42, although because the sample sizes are equal the same numerical result is obtained with Equation 12.43:

$$[S_{\Delta\overline{\boldsymbol{x}}}] = \left(\frac{1}{30} + \frac{1}{30}\right)\left(\frac{29}{58}[S_N] + \frac{29}{58}[S_B]\right) = \frac{1}{30}[S_N] + \frac{1}{30}[S_B] = \begin{bmatrix} 1.248 & 1.238 \\ 1.238 & 1.349 \end{bmatrix}. \tag{12.48}$$

The test statistic (Equation 12.40) can now be calculated as

$$T^2 = [2.18 \;\; 4.02] \begin{bmatrix} 1.248 & 1.238 \\ 1.238 & 1.349 \end{bmatrix}^{-1} \begin{bmatrix} 2.18 \\ 4.02 \end{bmatrix} = 32.34. \tag{12.49}$$

The $1-\alpha = 0.9999$ quantile of the F distribution with $\nu_1 = 2$ and $\nu_2 = 57$ degrees of freedom is 10.9, so the null hypothesis is rejected at the $\alpha = 0.0001$ level because $[(30+30-2)(2)/(30+30-2-1)]\,10.9 = 22.2 \ll T^2 = 32.34$ (cf. Equation 12.44). The actual p-value is smaller than 0.0001, but more extreme F-distribution quantiles are not commonly tabulated. Using the χ^2 distribution will provide only a moderately close approximation (Equation 12.38) because $\nu_2 = 57$, but the cumulative probability corresponding to ${\chi_2}^2 = 32.34$ can be calculated using Equation 4.53 (because ${\chi_2}^2$is the exponential distribution with $\beta = 2$) to be 0.99999991, corresponding to a p value of 0.00000001 (Equation 12.39b).

Even though the two-sample T^2 test provides a definitive rejection of the null hypothesis, it underestimates the statistical significance, because it does not account for the positive covariances between the New York and Boston temperatures that are evident in the submatrices $[S_{N-B}]$ and $[S_{B-N}]$ in Equation 12.47. In effect, the estimate in Equation 12.48 has assumed $[S_{N-\ B}] = [S_{B-N}] = [0]$. One way to account for these correlations is to compute the differences between the maximum temperatures as the linear combination ${\boldsymbol{b}_1}^T = [1, 0, -1, 0]$; compute the differences between the minimum temperatures as the linear combination ${\boldsymbol{b}_2}^T = [0, 1, 0, -1]$; and then use these two vectors as the rows of the transformation matrix $[B]^T$ in Equation 11.87b to compute the covariance $[S_\Delta]$ of the $n = 30$ vector differences, from the full covariance matrix $[S]$ in Equation 12.47. Equivalently, we could compute this covariance matrix from the 30 data pairs in the last two columns of Table 12.1. In either case the result is

$$[S_\Delta] = \begin{bmatrix} 3.133 & 2.623 \\ 2.623 & 2.768 \end{bmatrix}. \tag{12.50}$$

The null hypothesis of equal mean vectors for New York and Boston implies $\boldsymbol{\mu}_\Delta = \boldsymbol{0}$ in Equation 12.45, yielding the test statistic

$$T^2 = 30\,[2.18 \;\; 4.02] \begin{bmatrix} 3.133 & 2.623 \\ 2.623 & 2.768 \end{bmatrix}^{-1} \begin{bmatrix} 2.18 \\ 4.02 \end{bmatrix} = 298. \tag{12.51}$$

Because these temperature data are spatially correlated, much of the variability that was ascribed to sampling uncertainty for the mean vectors separately in the two-sample test in Equation 12.49 is actually shared and does not contribute to sampling uncertainty about the temperature differences. The numerical consequence is that the variances in the matrix $(1/30)[S_\Delta]$ are much smaller than their counterparts in Equation 12.48 for the two-sample test. Accordingly, T^2 for the paired test in Equation 12.51 is much larger than for the two-sample test in Equation 12.49. In fact it is huge, leading to the rough (because the sample sizes are only moderate) estimate, through Equation 4.53, for the p value of 2×10^{-65}.

Both the (incorrect) two-sample test and the (appropriate) paired test yield strong rejections of the null hypothesis that the New York and Boston mean vectors are equal. But what can be concluded about the way(s) in which they are different? This question will be taken up in Example 12.8. ◇

The T^2 tests described so far are based on the assumption that the data vectors are mutually uncorrelated. That is, although the K elements of $\boldsymbol{x}$ may have nonzero correlations, the vector observations $\boldsymbol{x}_i$, $i = 1, \ldots, n$, have been assumed to be mutually independent. As noted in Section 5.2.4, ignoring serial correlation may lead to large errors in statistical inference, typically because the sampling distributions of the test statistics have greater dispersion (the test statistics are more variable from batch to batch of data) than would be the case if the underlying data were independent.

A simple adjustment (Equation 5.13) is available for scalar t tests if the serial correlation in the data is consistent with a first-order autoregression (Equation 10.16). The situation is more complicated for the multivariate T^2 test because, even if the time dependence for each of K elements of $\boldsymbol{x}$ is reasonably represented by an AR(1) process, their autoregressive parameters ϕ may not be the same, and the lagged correlations among the different elements of $\boldsymbol{x}$ must also be accounted for. However, if the multivariate AR (1) process (Equation 12.24) can be assumed as reasonably representing the serial dependence of the data, and if the sample size is large enough to produce multinormality as a consequence of the Central Limit Theorem, the sampling distribution of the sample mean vector is

$$\overline{\boldsymbol{x}} \sim N_K\left(\boldsymbol{\mu}_x, \frac{1}{n}\left[\Sigma_\phi\right]\right), \tag{12.52a}$$

where

$$\left[\Sigma_\phi\right] = \left([I] - [\Phi]\right)^{-1}[\Sigma_x] + [\Sigma_x]\left([I] - [\Phi]^T\right)^{-1} - [\Sigma_x]. \tag{12.52b}$$

Equation 12.52a corresponds to Equation 12.33a for independent data, and $[\Sigma_\Phi]$ reduces to $[\Sigma_x]$ if $[\Phi] = [0]$ (i.e., if the $\boldsymbol{x}$'s are serially independent). For large n, sample counterparts of the quantities in Equation 12.52 can be substituted, and the matrix $[S_\Phi]$ used in place of $[S_x]$ in the computation of T^2 test statistics.

12.5.3. Simultaneous Confidence Statements

As noted in Section 5.1.7, a confidence interval is a region around a sample statistic, containing values that would not be rejected by a test whose null hypothesis is that the observed sample value is the true value. In effect, confidence intervals are constructed by working hypothesis tests in reverse. The difference in multivariate settings is that a confidence interval defines a region in the K-dimensional space of the data vector $\boldsymbol{x}$ rather than an interval in the one-dimensional space (the real line) of the scalar x. That is, multivariate confidence intervals are K-dimensional hypervolumes, rather than one-dimensional line segments.

Consider the one-sample T^2 test, Equation 12.35. After the data $\boldsymbol{x}_i$, $i = 1, \ldots, n$, have been observed and their sample covariance matrix $[S_x]$ has been computed, a $(1-\alpha)\cdot 100\%$ confidence region for the true vector mean consists of the set of points satisfying

$$n(\boldsymbol{x} - \overline{\boldsymbol{x}})^T [S_x]^{-1} (\boldsymbol{x} - \overline{\boldsymbol{x}}) \le \frac{K(n-1)}{n-K} F_{K,n-K}(1-\alpha), \tag{12.53}$$

because these are the $\boldsymbol{x}'$s that would not trigger a rejection of the null hypothesis that the true mean is the observed sample mean. For sufficiently large $n-K$, the right-hand side of Equation 12.53 would be well approximated by ${\chi_K}^2(1-\alpha)$. Similarly, for the two-sample T^2 test (Equation 12.40) a $(1-\alpha)\cdot 100\%$ confidence region for the difference of the two means consists of the points $\boldsymbol{\delta}$ satisfying

$$[\boldsymbol{\delta} - (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)]^T [S_{\Delta\overline{\boldsymbol{x}}}]^{-1} [\boldsymbol{\delta} - (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)] \le \frac{K(n_1+n_2-2)}{n_1+n_2-K-1} F_{K,n_1+n_2-K-1}(1-\alpha), \tag{12.54}$$

where again the right-hand side is approximately equal to ${\chi_K}^2(1-\alpha)$for large samples.

The points $\boldsymbol{x}$ satisfying Equation 12.53 are those whose Mahalanobis distance from $\overline{\boldsymbol{x}}$ is no larger than the scaled $(1-\alpha)$ quantile of the F (or χ^2, as appropriate) distribution on the right-hand side, and similarly for the points $\boldsymbol{\delta}$ satisfying Equation 12.54. Therefore the confidence regions defined by these equations

are bounded by (hyper-) ellipsoids whose characteristics are defined by the covariance matrix for the sampling distribution of the respective test statistic; for example, by $(1/n)[S_x]$ for Equation 12.53. Because the sampling distribution of $\bar{\boldsymbol{x}}$ approximates the MVN distribution on the strength of the Central Limit Theorem, the frontiers of the confidence regions defined by Equation 12.53 are probability ellipsoids for the MVN distribution with mean $\bar{\boldsymbol{x}}$ and covariance $(1/n)[S_x]$ (cf. Equation 12.5). Similarly, the confidence regions defined by Equation 12.54 are bounded by hyper-ellipsoids centered on the vector mean difference between the two sample means.

As illustrated in Example 12.1, the properties of these confidence ellipses, other than their center, are defined by the eigenvalues and eigenvectors of the covariance matrix for the sampling distribution in question. In particular, each axis of one of these ellipsoids will be aligned in the direction of one of the eigenvectors, and each will be elongated in proportion to the square root of the corresponding eigenvalue. In the case of the one-sample confidence region, for example, the limits of $\boldsymbol{x}$ satisfying Equation 12.53 in the directions of each of the axes of the ellipse are

$$\boldsymbol{x} = \bar{\boldsymbol{x}} \pm \boldsymbol{e}_k \sqrt{\lambda_k \frac{K\,(n-1)}{n-K} F_{K,n-K}(1-\alpha)}, \; k = 1, \ldots, K, \tag{12.55}$$

where λ_k and $\boldsymbol{e}_k$ are the kth eigenvalue-eigenvector pair of the matrix $(1/n)[S_x]$. Again, for sufficiently large n, the quantity under the radical would be well approximated by $\lambda_k {\chi_K}^2(1-\alpha)$. Equation 12.55 indicates that the confidence ellipses are centered at the observed sample mean $\bar{\boldsymbol{x}}$, and extend further in the directions associated with the largest eigenvalues. They also extend further for smaller α because these produce larger cumulative probabilities for the distribution quantiles $F(1-\alpha)$ and ${\chi_K}^2(1-\alpha)$.

It would be possible and computationally simpler to conduct K univariate t tests, and to then compute K univariate confidence intervals separately for the means of each of the elements of $\boldsymbol{x}$ rather than the T^2 test examining the vector mean $\bar{\boldsymbol{x}}$. What is the relationship between an ellipsoidal multivariate confidence region of the kind just described, and a collection of K univariate confidence intervals? Jointly, these univariate confidence intervals would define a hyper-rectangular region in the K-dimensional space of $\boldsymbol{x}$; but the probability (or confidence) associated with outcomes enclosed by it will be substantially less than $1-\alpha$, if the lengths of each of its K sides are the corresponding $(1-\alpha)$•100% scalar confidence intervals. The problem is one of test multiplicity: if the K tests on which the confidence intervals are based are independent, the joint probability of all the elements of the vector $\boldsymbol{x}$ being simultaneously within their scalar confidence bounds will be $(1-\alpha)^K$. To the extent that the scalar confidence interval calculations are not independent, the joint probability will be different, but difficult to calculate.

An expedient workaround for this multiplicity problem is to calculate the K one-dimensional *Bonferroni confidence intervals* and use these as the basis of a joint confidence statement:

$$\Pr\left\{ \bigcap_{k=1}^{K} \left[\bar{x}_k + z\left(\frac{\alpha/K}{2}\right)\sqrt{\frac{s_{k,k}}{n}} \le \mu_k \le \bar{x}_k + z\left(1-\frac{\alpha/K}{2}\right)\sqrt{\frac{s_{k,k}}{n}} \right] \right\} \ge 1-\alpha. \tag{12.56}$$

The expression inside the square bracket defines a univariate, $(1-\alpha/K)$•100% confidence interval for the kth variable in $\boldsymbol{x}$. Each of these confidence intervals has been expanded relative to the nominal $(1-\alpha)$•100% confidence interval, to compensate for the multiplicity in K dimensions simultaneously. For convenience, it has been assumed in Equation 12.56 that the sample size is adequate for standard Gaussian quantiles to be appropriate, although quantiles of the t distribution with $n-1$ degrees of freedom usually would be used for n smaller than about 30.

There are two problems with using Bonferroni confidence regions in this context. First, Equation 12.56 is an inequality rather than an exact specification. That is, the probability that all the

K elements of the hypothetical true mean vector $\boldsymbol{\mu}$ are contained simultaneously in their respective one-dimensional confidence intervals is at least $1-\alpha$, not exactly $1-\alpha$. That is, in general the K-dimensional Bonferroni confidence region is too large, but exactly how much more probability than $1-\alpha$ may be enclosed by it is not known.

The second problem is more serious. As a collection of univariate confidence intervals, the resulting K-dimensional hyper-rectangular confidence region ignores the covariance structure of the data. Bonferroni confidence statements can be reasonable if the correlation structure is weak, for example, in the setting described in Section 10.5.6. But Bonferroni confidence regions are inefficient when the correlations among elements of $\boldsymbol{x}$ are strong, in the sense that they will include large regions having very low plausibility. As a consequence they are too large in a multivariate sense and can lead to silly inferences.

Example 12.7. Comparison of Unadjusted Univariate, Bonferroni, and MVN Confidence Regions

Assume that the covariance matrix in Equation 11.59, for the Ithaca and Canandaigua minimum temperatures, had been calculated from $n=100$ independent temperature pairs. This many observations would justify large-sample approximations for the sampling distributions (standard Gaussian z and χ^2, rather than t and F quantiles), and assuming independence obviates the need for the nonindependence adjustments in Equation 12.52.

What is the best two-dimensional confidence region for the true climatological mean vector, given the sample mean $[13.00, 20.23]^{\mathrm{T}}$, and assuming the sample covariance matrix for the data in Equation 11.59? Relying on the multivariate normality for the sampling distribution of the sample mean implied by the Central Limit Theorem, Equation 12.53 defines an elliptical 95% confidence region when the right-hand side is the χ^2 quantile ${\chi_2}^2(0.95) = 5.991$. The result is shown in Figure 12.7, centered on the sample mean (+). Compare this ellipse to Figure 12.1, which is centered on the same mean and based on the same covariance matrix (although drawn to enclose slightly less probability). Figure 12.7 has exactly the same shape and orientation, but it is much more compact, even though it encloses somewhat more probability. Both ellipses have the same eigenvectors, ${\boldsymbol{e}_1}^T = [0.848, 0.530]$ and ${\boldsymbol{e}_2}^T = [-0.530, 0.848]$, but the eigenvalues for Figure 12.7 are 100-fold smaller, that is, $\lambda_1 = 2.5476$ and $\lambda_2 = 0.0829$. The difference is that Figure 12.1 represents one contour of the MVN distribution for the data, with covariance $[S_x]$ given by Equation 11.59, but Figure 12.7 shows one contour of the MVN distribution with covariance $(1/n)[S_x]$, appropriate to Equation 12.53 and relevant to the sampling distribution of the mean rather than the distribution for the data. This ellipse is the smallest region enclosing 95% of the probability of this distribution for the sampling variations of the sample mean. Its elongation reflects the strong correlation between the minimum temperatures at the two locations, so that differences between the sample and true means due to sampling variations are much more likely to involve differences of the same sign for both the Ithaca and Canandaigua means.

The gray rectangle in Figure 12.7 outlines the 95% Bonferroni confidence region. It has been calculated using $\alpha = 0.05$ in Equation 12.56, and so is based on the 0.0125 and 0.9875 quantiles of the standard Gaussian distribution, or $z = \pm 2.24$. The resulting rectangular region encloses at least $(1-\alpha)\cdot 100\% = 95\%$ of the probability of the joint sampling distribution. It occupies much more area in the plane than does the confidence ellipse, because the rectangle includes large regions in the upper left and lower right that contain very little probability. However, from the standpoint of univariate inference—that is, confidence intervals for one location without regard to the other—the Bonferroni limits are narrower.

The dashed rectangular region results jointly from the two standard 95% confidence intervals. The length of each side has been computed using the 0.025 and 0.975 quantiles of the standard Gaussian

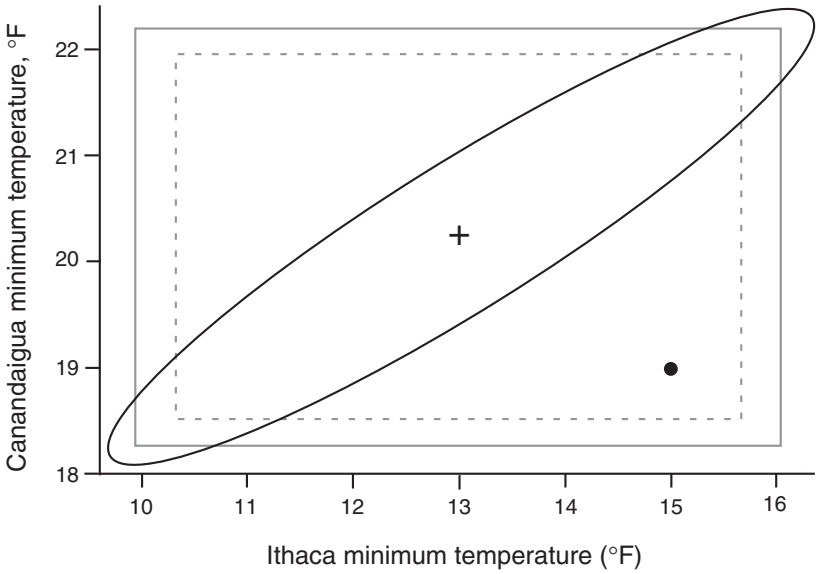

FIGURE 12.7 Hypothetical 95% joint confidence regions for the mean Ithaca and Canandaigua minimum temperatures, assuming that $n = 100$ independent bivariate observations had been used to calculate the covariance matrix in Equation 11.59. The ellipse encloses points within a Mahalanobis distance of $\chi^2 = 5.991$ of the sample mean (indicated by +) $[13.00, 20.23]^{\mathrm{T}}$. Horizontal and vertical limits of the dashed rectangle are defined by two independent confidence intervals for the two variables, with $\pm z(0.025) = \pm 1.96$. Gray rectangle indicates corresponding Bonferroni confidence region, calculated with $\pm z(0.0125) = \pm 2.24$. The point $[15, 19]^{\mathrm{T}}$ (large dot) is comfortably within both rectangular confidence regions, but is at Mahalanobis distance $\chi^2 = 1006$ from the mean relative to the joint covariance structure of the two variables, and is thus highly implausible.

distribution, which are $z = \pm 1.96$. They are, of course, narrower than the corresponding Bonferroni intervals, and according to Equation 12.56 the resulting rectangle includes at least 90% of the probability of this sampling distribution. Like the Bonferroni confidence region, it portrays large areas with very low probabilities as containing plausible values for the true mean.

The main difficulty with Bonferroni confidence regions is illustrated by the point $(15, 19)^{\mathrm{T}}$, located by the large dot in Figure 12.7. It is comfortably within the gray rectangle delineating the Bonferroni confidence region, which carries the implication that this is a plausible value for the true mean vector. However, a Bonferroni confidence region is defined without regard to the multivariate covariance structure of the distribution that it purports to represent. In the case of Figure 12.7 the Bonferroni confidence region ignores the fact that sampling variations for these two positively correlated variables are much more likely to yield differences between the two sample and true means that are of the same sign. The Mahalanobis distance between the points $(15, 19)^{\mathrm{T}}$ and $(13.00, 20.23)^{\mathrm{T}}$, according to the covariance matrix $(1/n)[S_x]$, is 1006, implying an astronomically small probability for a separation this large and of this orientation for these two vectors (cf. Equation 12.34a). The vector $[15, 19]^{\mathrm{T}}$ is an extremely implausible candidate for the true mean $\boldsymbol{\mu}_x$. ◇

12.5.4. Interpretation of Multivariate Statistical Significance

What can be said about multivariate mean differences if the null hypothesis for a T^2 test is rejected, that is, if Equation 12.37 or 12.44 (or their large-sample counterpart, Equation 12.39b) is satisfied? This

question is complicated by the fact that there are many ways for multivariate means to differ from one another, including but not limited to one or more of the pairwise differences between the elements that would be detected by the corresponding univariate tests.

If a T^2 test results in the rejection of its multivariate null hypothesis, the implication is that at least one scalar test for a linear combination $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}$ or $\boldsymbol{a}^{\mathrm{T}}(\boldsymbol{x}_1 - \boldsymbol{x}_2)$, for one- and two-sample tests, respectively, will be statistically significant. In any case, the scalar linear combination providing the most convincing evidence against the null hypothesis (regardless of whether or not it is sufficiently convincing to reject at a given test level) will satisfy

$$\boldsymbol{a} \propto [S]^{-1}(\overline{\boldsymbol{x}} - \boldsymbol{\mu}_0) \tag{12.57a}$$

for one-sample tests, or

$$\boldsymbol{a} \propto [S]^{-1}\left[(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2) - \boldsymbol{\delta}_0\right] \tag{12.57b}$$

for two-sample tests. At minimum, then, if a multivariate T^2 calculation results in a null hypothesis rejection, then linear combinations corresponding to the K-dimensional direction defined by the vector $\boldsymbol{a}$ in Equation 12.57 will lead to significant results also. It can be very worthwhile to interpret the meaning, in the context of the data, of the direction $\boldsymbol{a}$ defined by Equation 12.57. Of course, depending on the strength of the overall multivariate result, other linear combinations may also lead to scalar test rejections, and it is possible that all linear combinations will be significant. The direction $\boldsymbol{a}$ also indicates the direction that best discriminates between the populations from which $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ were drawn (see Section 15.2.2).

The reason that any linear combination $\boldsymbol{a}$ satisfying Equation 12.57 yields the same test result can be seen most easily in terms of the corresponding confidence interval. Consider for simplicity the confidence interval for a one-sample T^2 test, Equation 12.53. Using the results in Equation 11.85, this scalar confidence interval is defined by

$$\boldsymbol{a}^T\overline{\boldsymbol{x}} - c\sqrt{\frac{\boldsymbol{a}^T[S_x]\,\boldsymbol{a}}{n}} \le \boldsymbol{a}^T\boldsymbol{\mu} \le \boldsymbol{a}^T\overline{\boldsymbol{x}} + c\sqrt{\frac{\boldsymbol{a}^T[S_x]\,\boldsymbol{a}}{n}}, \tag{12.58}$$

where c^2 equals $[K(n-1)/(n-K)]\,F_{K,\,n-K}(1-\alpha)$, or ${\chi_K}^2$, as appropriate. Even though the length of the vector $\boldsymbol{a}$ is arbitrary, so that the magnitude of the linear combination $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}$ is also arbitrary, the quantity $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{\mu}$ is scaled identically.

Another remarkable property of the T^2 test is that valid inferences about any and all linear combinations can be made, even though they may not have been specified *a priori*. The price that is paid for this flexibility is that such inferences will be less precise than those made using conventional scalar tests for linear combinations that were specified in advance. This point can be appreciated in the context of the confidence regions shown in Figure 12.7. If a test regarding the Ithaca minimum temperature alone had been of interest, corresponding to the linear combination $\boldsymbol{a} = [1, 0]^{\mathrm{T}}$, the appropriate confidence interval would be defined by the horizontal extent of the dashed rectangle. The corresponding interval for this linear combination from the full T^2 test is substantially wider, being defined by the projection, or shadow, of the ellipse onto the horizontal axis. But what is gained from the multivariate test is the ability to make valid simultaneous probability statements regarding as many linear combinations as may be of interest.

Example 12.8. Interpreting the New York and Boston Mean January Temperature Differences
Return now to the comparisons made in Example 12.6, between the vectors of average January maximum and minimum temperatures for New York City and Boston. The difference between the

sample means was $[2.18, 4.02]^{\mathrm{T}}$, and the null hypothesis was that the true means were equal, so the corresponding difference $\boldsymbol{\delta}_0 = \mathbf{0}$. Even assuming, erroneously, that there is no spatial correlation between the two locations (or, equivalently for the purpose of the test, that the data for the two locations were taken in different years), T^2 in Equation 12.49 indicates that the null hypothesis should be strongly rejected.

Both means are warmer at New York, but Equation 12.49 does not necessarily imply significant differences between the average maxima or the average minima. Figure 12.6 shows substantial overlap between the data scatters for both maximum and minimum temperatures, with each scalar mean near the center of the corresponding univariate data distribution for the other city. Computing the separate univariate tests (Equation 5.8) yields $z = 2.18/\sqrt{1.248} = 1.95$ for the maxima and $z = 4.02/\sqrt{1.349} = 3.46$ for the minima. Even leaving aside the problem that two simultaneous comparisons are being made, the result for the difference of the average maximum temperatures is not quite significant at the 5% level, although the difference for the minima is stronger.

The significant result in Equation 12.49 ensures that there is at least one linear combination $\boldsymbol{a}^{\mathrm{T}}(\boldsymbol{x}_1 - \boldsymbol{x}_2)$ (and possibly others, although not necessarily the linear combinations resulting from $\boldsymbol{a}^{\mathrm{T}} = [1, 0]$ or $[0, 1]$) for which there is a significant difference. According to Equation 12.57b, the vectors producing the most significant linear combinations are proportional to

$$\boldsymbol{a} \propto [S_{\Delta\overline{x}}]^{-1}\overline{\boldsymbol{\Delta}} = \begin{bmatrix} 1.248 & 1.238 \\ 1.238 & 1.349 \end{bmatrix}^{-1} \begin{bmatrix} 2.18 \\ 4.02 \end{bmatrix} = \begin{bmatrix} -13.5 \\ 15.4 \end{bmatrix}. \tag{12.59}$$

This linear combination of the mean differences, and the estimated variance of its sampling distribution, are

$$\boldsymbol{a}^T\overline{\boldsymbol{\Delta}} = [-13.5 \;\; 15.4] \begin{bmatrix} 2.18 \\ 4.02 \end{bmatrix} = 32.5, \tag{12.60a}$$

and

$$\boldsymbol{a}^T[S_{\Delta\overline{x}}]\boldsymbol{a} = [-13.5 \;\; 15.4] \begin{bmatrix} 1.248 & 1.238 \\ 1.238 & 1.349 \end{bmatrix} \begin{bmatrix} -13.5 \\ 15.4 \end{bmatrix} = 32.6, \tag{12.60b}$$

yielding the univariate test statistic for this linear combination of the differences $z = 32.5/\sqrt{32.6} = 5.69$. This is, not coincidentally, the square root of Equation 12.49. The appropriate benchmark against which to compare the unusualness of this result in the context of the null hypothesis is not the standard Gaussian or t distributions (because this linear combination was derived from the test data, not *a priori*), but rather the square roots of either ${\chi_2}^2$ quantiles or of appropriately scaled $F_{2,30}$ quantiles. The result is still very highly significant, with $p \approx 10^{-7}$.

Equation 12.59 indicates that the most significant aspect of the difference between the New York and Boston mean vectors is not the warmer temperatures at New York relative to Boston (which would correspond to $\boldsymbol{a} \propto [1, 1]^{\mathrm{T}}$). Rather, the elements of $\boldsymbol{a}$ are of opposite sign and of nearly equal magnitude, and so describe a *contrast*. Since $-\boldsymbol{a} \propto \boldsymbol{a}$, one way of interpreting this contrast is as the difference between the average maxima and minima, corresponding to the choice $\boldsymbol{a} \approx [1, -1]^{\mathrm{T}}$. That is, the most significant aspect of the difference between the two mean vectors is closely approximated by the difference in the average diurnal range, with the range for Boston being larger. The null hypothesis that that the two diurnal ranges are equal can be tested specifically, using the contrast vector $\boldsymbol{a} = [1, -1]^{\mathrm{T}}$ in Equation 12.60, rather than the linear combination defined by Equation 12.59. The result is $z = -1.84/\sqrt{0.121} = -5.29$. This test statistic is negative because the diurnal range at New York is

smaller than the diurnal range at Boston. It is slightly smaller (in absolute value) than the result obtained when using $\boldsymbol{a} = [-13.5, 15.4]$, because that is the most significant linear combination, although the result is almost the same because the two vectors are aligned in nearly the same direction. Comparing the result to the ${\chi_2}^2$ distribution yields the very highly significant result $p \approx 10^{-6}$. Visually, the separation between the two point clouds in Figure 12.6 is consistent with this difference in diurnal range: The points for Boston tend to be closer to the upper left, and those for New York are closer to the lower right. On the other hand, the relative orientation of the two means is almost exactly opposite, with the New York mean closer to the upper right corner, and the Boston mean closer to the lower left. ◇

12.6. EXERCISES

12.1. Assume that the Ithaca and Canandaigua maximum temperatures in Table A.1 constitute a sample from a MVN distribution, and that their covariance matrix [*S*] has eigenvalues and eigenvectors as given in Exercise 11.8. Sketch the 50% and 95% probability ellipses of this distribution.

12.2. Assume that the four temperature variables in Table A.1 are MVN distributed, with the ordering of the variables in $\boldsymbol{x}$ being $[\text{Max}_{\text{Ith}}, \text{Min}_{\text{Ith}}, \text{Max}_{\text{Can}}, \text{Min}_{\text{Can}}]^{\text{T}}$. The respective means are also given in Table A.1, and the covariance matrix [*S*] is given in the answer to Exercise 11.9a. Assuming the true mean and covariance are the same as the sample values,

a. Specify the conditional distribution of $[\text{Max}_{\text{Ith}}, \text{Min}_{\text{Ith}}]^{\text{T}}$, given that $[\text{Max}_{\text{Can}}, \text{Min}_{\text{Can}}]^{\text{T}} = [31.77, 20.23]^{\text{T}}$ (i.e., the average values for Canandaigua).

b. Consider the linear combinations $\boldsymbol{b}_1 = [1, 0, -1, 0]$, expressing the difference between the maximum temperatures, and $\boldsymbol{b}_2 = [1, -1 -1, 1]$, expressing the difference between the diurnal ranges, as rows of a transformation matrix $[B]^{\text{T}}$. Specify the distribution of the transformed variables $[B]^{\text{T}}\boldsymbol{x}$.

12.3. Consider the bivariate normal distribution for the random vector $\boldsymbol{x}$, defined by:

$$\boldsymbol{\mu}^{\text{T}} = [5.2 \;\; 16.1] \text{ and } [\sum] = \begin{bmatrix} 174.7 & 285.2 \\ 285.2 & 525.0 \end{bmatrix}.$$

If x_1 and x_2 are Box-Cox ("power-") transformed versions of the variables y_1 and y_2, respectively, defined by
$x_1 = (y_1^{1/2} - 1)/0.5$ and $x_2 = (y_2^{1/2} - 1)/0.5$, evaluate $\Pr\{y_1 \le 10 \,|\, y_2 = 100\}$.

12.4. Use the mean vector for the New York January maximum and minimum temperatures in Equation 12.46a, and their covariance matrix in a portion of Equation 12.47, and assume that these define a bivariate normal distribution.

a. Fully specify the distribution of the "mean" January temperature in a given year, which is computed as the average of the maximum and minimum temperatures.

b. January 2012 was unusually warm in New York city, with $[x_{\text{max}}, x_{\text{min}}]^{\text{T}} = [44.2, 30.4]^{\text{T}}$. Evaluate the probability of seeing a January temperature vector at least as far removed from the long-term mean as that observed in 2012, assuming that the Chi-square distribution is an adequate approximation to the sampling distribution of the Mahalanobis distance.

12.5. The eigenvector associated with the smallest eigenvalue of the covariance matrix [*S*] for the January 1987 temperature data referred to in Exercise 11.2 is ${\boldsymbol{e}_4}^T = [-0.665, 0.014, 0.738, \;\; -0.115]$. Assess the normality of the linear combination ${\boldsymbol{e}_4}^T\boldsymbol{x}$,

a. Graphically, with a Q-Q plot. For computational convenience, evaluate $\Phi(z)$ using Equation 4.30.

b. Formally, with the Filliben test (see Table 5.5), assuming no autocorrelation.

12.6. a. Compute the 1-sample T^2 testing the linear combinations $[B]^T\overline{\boldsymbol{x}}$ with respect to H_0: $\boldsymbol{\mu}_0 = \mathbf{0}$, where $\boldsymbol{x}$ and $[B]^T$ are defined as in Exercise 12.2. Ignoring the serial correlation, evaluate the plausibility of H_0, assuming that the χ^2 distribution is an adequate approximation to the sampling distribution of the test statistic.
b. Compute the most significant linear combination for this test.

12.7. Repeat Exercise 12.4, assuming spatial independence (i.e., setting all cross-covariances between Ithaca and Canandaigua variables to zero).