Underlying Theory and Implementation of the Test

William R. Fairweather

April 2020

Underlying Theory of the Test

Sources

This test of multivariate normality is based on the univariate work of Cs*ö*rg*ö* and Seshadri (1970, 1971) and on my doctoral dissertation (Fairweather 1973). As indicated in the titles of these articles, the test is based on a characterization; that is, it is based on a feature of the normal distribution that is unique to it among all (nondegenerate) multivariate distributions. The MVNtestchar package implements this test.

Transformations and the Characterization

Consider a set of p x 1 random vectors of full rank X_i, i = 1, . . ., 4(p+1). Let $Y_i = X_{2i} - X_{2i-1}$, i = 1, . . ., 2(p+1).

The Y_i have a distribution that is centered at 0. In fact, all of the odd moments of the Y_i are zero regardless of the underlying distribution of the X_i. Now, define

$$W_1 = \sum_{i=1}^{p+1} Y_i {Y'}_i$$

and

$$W_2 = \sum_{i=p+2}^{2(p+1)} Y_i {Y'}_i$$

,

where Y'_i is the transpose of Y_i. The W_i are then independently distributed, symmetric matrices of rank p.

Let $T = W_1 + W_2$ and let $S = T^{-1/2} W_1 T^{-1/2}$. S is a positive definite (symmetric) matrix of rank p regardless of the underlying distribution of the X_i.

It is shown in the Appendix that S is distributed uniformly on its support region **if and only if** the X_i are multivariate normal. It is this characteristic that underlies the test.

The support region for S

The p x p symmetric matrix S is equivalent to a set of p(p+1)/2 random variables V_1, V_2, . . . , V_p, V_{12}, V_{13}, .., V_{1p}, . . . , $V_{p-1,p}$. This is easily seen if we lay out the V_i and the V_{ij} in the matrix format, showing only the upper triangle:

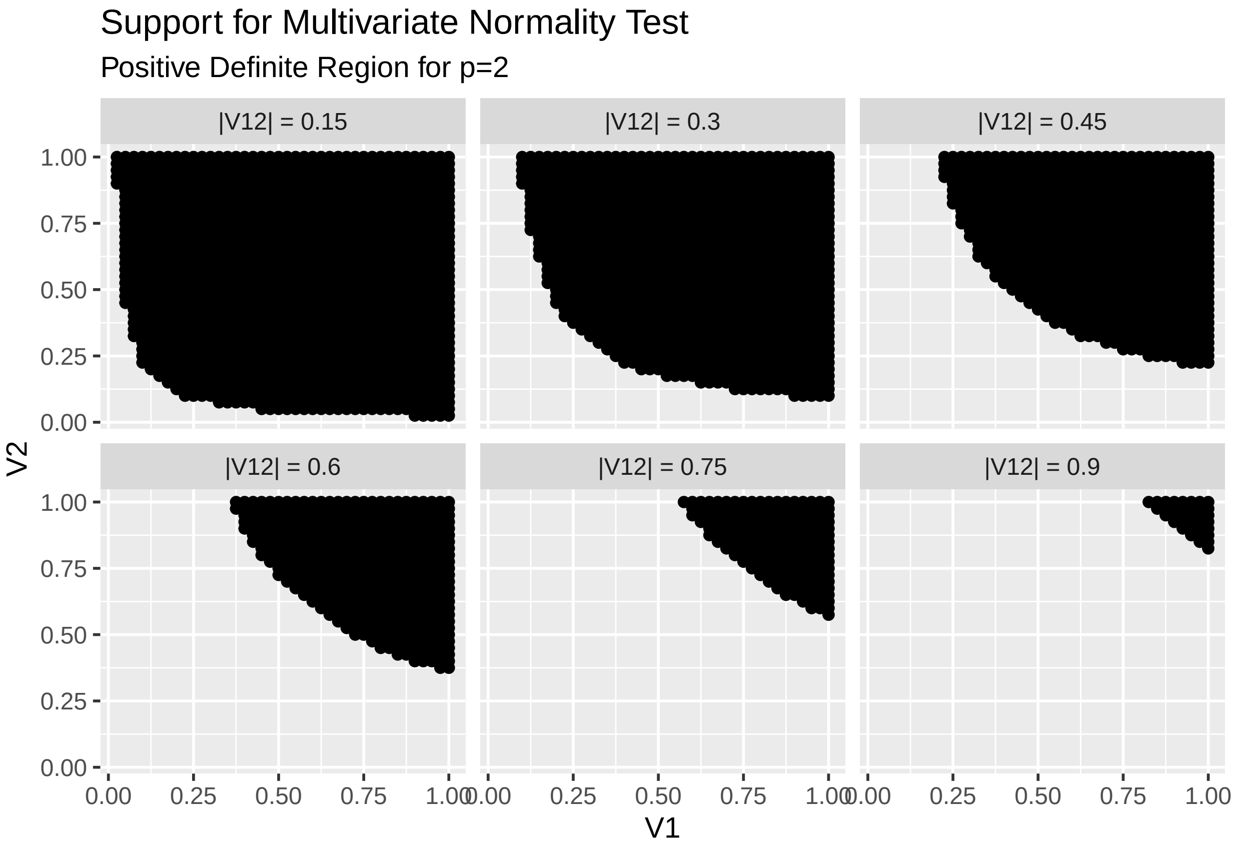
```
V1  V12  V13 ... V1p
    V2   V23 ... V2p
         V3  ... V3p
                .
                 .
                  .
                   Vp
```

V_1 through V_p are the diagonal elements of the matrix and V_{12} … $V_{p\text{-}1,p}$ are the off-diagonal elements.

By construction, the diagonal elements of S all lie in the interval [0,1]. Because S is positive definite, the off-diagonal elements all lie in the interval [-1,1]. The support region of S is within the hyperrectangle

$R_p = [0,1]^p$ x $[-1,1]^m$,where m = p(p-1)/2.

The support region is difficult to envision in higher dimensions. For p = 2, the matrix S has two diagonal elements and one off-diagonal element: V_1, V_2, and V_{12}. The support region of S is then that part of the 3-dimensional rectangle bounded by $R\ _2 = [0,1]$ x [0,1] x [-1,1] that satisfies $V_1\ V_2 - V^2{}_{12} > 0$. By stepping $|V_{12}|$ up from 0 to 1.0, the following graph shows the support region (shaded) as a function of V_1 and V_2.



The package contains four functions that produce rotatable graphs depicting this hyperrectangle and the positive definite region within it. The function *support.p2( )* shows the entire positive definite region. The functions *slice.v1( )* and *slice.v12( )* show slices through the region for fixed values of V_1 and V_{12}, respectively. Finally, *maxv12( )* shows the maximum value of V_{12} as a function of V_1 and V_2.

The latest values of the rotational parameters are output in list format upon exit from the graphics functions to facilitate return to that rotation, if desired. To capture these, assign the function to a new variable.

Implementing the Test

The function *testunknown(x, pvector, k)* implements the test of multivariate normality of the n x p matrix, x. x can be either a matrix or an array. The input parameters will be discussed more fully below.

The matrix x is assumed to be a sample of n observations on the unknown p-variate distribution. Here, n = 4r(p+1) for r a positive integer. The transformations involve exact numbers of random variables, and *testunknown( )* will discard observations at random to ensure that this condition holds.

testunknown( ) divides the sample into r groups of 4(p+1) observations and performs the transformations described above on each of the r sets independently. This results in positive definite matrices S_1, . . . , S_r distributed independently as described above regardless of the distribution of the X_i.

S_1, . . . , S_r are distributed uniformly on the positive definite subspace of the hyperrectangle

R_p = $[0,1]^p$ x $[-1,1]^m$ if and only if x is a sample from a multivariate normal distribution.

testunknown( ) performs a chisquare goodness of fit test by filling this hyperrectangle with minicubes and counting the number of S_i in each minicube. As implied, the minicubes are hypercubical (equal length in every dimension).

The input variable pvector is essentially a check to ensure that the matrix is oriented properly; it should equal the value of p taken from x directly. If this test fails, the function aborts. The parameter k defines the number of cuts to be made on each edge of the hyperrectangle. The minicubes will then be of size

1/k x 1/k x . . . x 1/k. There are p(p+1)/2 terms in this set. It is possible to undertake various research projects with this test function, and an array with mobs layers is allowed in order to facilitate this possibility. We have in mind simulations with mobs repetitions. If x is an array with 1 layer, x should have dimension

n x p x 1.

Relating Sample Size to the Size of Minicubes

Any goodness of fit test, univariate or multivariate, must consider the relationship of the sample size to the number of "bins" into which the support is subdivided. The sample size is always finite and the samples are continuous, so that creating too many bins will always result in exactly 1 observation per occupied bin. With too many bins, there can be no distinction here between null and alternative distributions.

In our case, the number of minicubes ("bins") into which the hypercube is divided is

N(k,p) = k^p $(2k)^m$, where m = p(p-1)/2. For p=2, m = 1 and N(k,p) = $2k^3$. Similarly, N(k,3) = $8k^6$ and N(k,4) = $64k^{10}$.

Minicubes are entirely, partially, or not at all within the positive definite region of the hyperrectangle. Whenever k=1, the region of the hyperrectangle representing the diagonal elements of the matrix is not subdivided and the region representing the off-diagonal elements is only subdivided into the positive and negative values of each element. In all cases with k=1 each minicube is partially within the positive definite region. We can calculate analytically the fraction of the hyperrectangle that is within the positive definite region only for p=2. This fraction is 4/9. For p>2, the calculation appears to be intractable.

The number of minicubes clearly grows rapidly with the dimensionality of the sample. However, the support of the S_i is only a subset of the hyperrectangle, namely the positive definite region. The following table shows the number of minicubes in the hyperrectangle, N(k,p), the ratio of the positive definite region to the overall volume of the hyperrectangle, and the approximate number of minicubes in the positive definite region, as a function of k and $p \leq 4$.

Table 1. The number of minicubes, N(k,p) in the positive definite region of the hyperrectangle, as a function of the number of cuts, k and the dimensionality, p of the sample.

	p=2				p=3				p=4			
	k	N(k,p)	ratio (%)	pos def	k	N(k,p)	ratio (%)	pos def	k	N(k,p)	ratio (%)	pos def
1	2	16	44.4	7	2	512	14.8	76	2	65536	0.6	344
2	5	250	44.4	111	5	125000	7.2	8948	3	3779136	0.5	20410
3	10	2000	44.4	889	6	373248	7.8	29213	4	67108864	0.5	335544
4	15	6750	44.4	3000	7	941192	7.8	73688	5	6.3e+08	0.5	3e+06
5	20	16000	44.4	7110	9	4251528	7.8	331619	6	3.9e+09	0.5	2e+07

N(k,p) is easily calculated in each case. For p=2, the calculated ratio was multiplied by the number of minicubes to get the approximate number of positive definite minicubes. For $p > 2$, rows 1 through 4 of Table 1 were calculated as follows: The hyperrectangle was filled with minicubes as described above. A minicube was defined to be within the positive definite region if a point very near the center of the minicube represented a positive definite matrix. The last row of the table is an extrapolation obtained by applying the asymptotic ratio to the calculated value of N(k,p).

For each value of p the ratio of minicubes in the positive definite region to the overall number in the hyperrectangle is fairly constant. For $p > 2$, this ratio is a very small part of the overall volume of the hyperrectangle. Nevertheless, Table 1 shows that a very large number of "bins" in the support region will result if k is set too large.

In performing the characterization transformations, the number of vector samples is substantially reduced to form the positive definite matrices that are tested for uniformity of distribution. Table 1 refers to the number of bins into which the matrices S_i will fall. 4(p+1) vectors X_i will result in a single matrix S_i. This multiplier is 12 for p=2, is 16 for p=3, and is 20 for p=4. Assuming that the expected number of S_i in each bin should be 3 or 5, Table 2 gives the number of X_i that should be in the sample for each value of k.

Table 2. Relationship of sample size n to number of cuts k, as a function of the expected number E of S_i per minicube.

	p = 2			p = 3			p = 4		
	k	E=3	E=5	k	E=3	E=5	k	E=3	E=5
1	2	256	427	2	3648	6080	2	20640	34400
2	5	4000	6666	5	429504	715840	3	1224600	2041000
3	10	31997	53328	6	1402224	2337040	4	20132640	33554400
4	15	107989	179982	7	3537024	5395040	5	1.87e+08	3.12e+08
5	20	255974	426624	9	15917712	26529520	6	1.19e+09	1.98e+09

From Table 2, we conclude that if we wish to test a trivariate sample for normality, and we have about 715,000 vector observations, we should set k to be no more than 5 to ensure that there are about E=5 observations per minicube.

Example Databases Included in MVNtestchar Package

In order to gain experience with the test function, the package contains four sample databases:

- unknown.Np2
- unknown.Np4
- unknown.Bp2
- unknown.Bp4

In these names N symbolizes a sample of normal random vectors and B symbolizes a sample of modified Bernoulli random vectors. The number in the name indicates the vector rank. True Bernoulli random variables cause the test program to crash because of colinearity, so a normal variable with extremely small variance was added to each one to make the Bernoulli vectors continuous random variables. Finally, unknown.Bp2 is a matrix; the others are arrays with a single layer.

References

Csörgö, M and Seshadri, V (1970). On the problem of replacing composite hypotheses by equivalent simple ones, *Rev. Int. Statist. Instit.*, **38**, 351-368

Csörgö, M and Seshadri,V (1971). Characterizing the Gaussian and exponential laws by mappings onto the unit interval, *Z. Wahrscheinlickhkeitstheorie verw. Geb.*, **18**, 333-339.

Fairweather, WR (1973). A test for multivariate normality based on a characterization. Dissertation submitted in partial fulfillment of the requirements for the Doctor of Philosophy, University of Washington, Seattle WA.

APPENDIX

In preparation