A Novel Scale-Space Approach for Multinormality Testing and the k-Sample Problem

Kristian Hindberg[a,*], Jan Hannig[b], Fred Godtliebsen[a]

[a] *Department of Mathematics and Statistics, University of Tromsø, N-9037 Tromsø, Norway*
[b] *Department of Statistics, University of North Carolina, Chapel Hill, NC 27599-3260, US*

Abstract

Two classical statistical problems, multivariate testing of normality and the k-sample problem, are explored by a novel analysis on several levels of resolutions/scales simultaneously. The presented methods do not invert any estimated covariance matrix. Thereby, the methods also work in the High Dimension Low Sample Size situation, i.e. $n \leq p$. The output, a significance map, is produced by doing a one-dimensional test for all possible scale/location pairs. The significance map shows for which scale/location pairs the null hypothesis is rejected. For the testing of multinormality, the Anderson-Darling test is utilized to detect potential departures from multinormality at different combinations of scales and locations. In the k-sample case, it is tested whether or not the k data sets can be said to originate from the same unspecified discrete or continuous multivariate distribution. This is done by testing the k vectors corresponding to the same scale/location pair of the different data sets through the k-sample Anderson-Darling test. Successful demonstrations of the new methodology on artificial and real data sets are presented, and a feature selection scheme is demonstrated.

Keywords: Multi-scale, significance map, Anderson-Darling testing, high dimension low sample size, Cramér-Wold theorem

*Corresponding address: Department of Mathematics and Statistics, Faculty of Science and Technology, University of Tromsø, N-9037 Tromsø, Norway. Tel.: +47 77644001; fax: +47 77644765

Email addresses: kristian.hindberg@uit.no (Kristian Hindberg), jan.hannig@unc.edu (Jan Hannig), fred.godtliebsen@uit.no (Fred Godtliebsen)

1. Introduction

In practice, it is frequently assumed that a data set can be described by a multivariate normal distribution. Many common statistical procedures rely on the data being multinormal, something which is often not adequately checked before use of the procedures (Cox and Wermuth, 1994; Farrell et al., 2007; Looney, 1995). Often, this assumption is false for either the whole data set or parts of it. Another classical problem is the testing of whether k multivariate data sets originate from the same distribution. For each of the two problems, a scale-space inspired methodology, testing all scales and locations simultaneously, is presented. The two presented algorithms are very similar, the main difference being which one-dimensional test is used. For both algorithms, a weighted summation is performed across the dimensions/locations. The notion of scale is connected to the number of dimensions being summed across, while the different dimensions typically are temporal or spatial samples.

The presented algorithms have two aspects that make them useful in many situations. As will be shown, the algorithms avoid the need to estimate the covariance matrix, leading to algorithms that can handle the High Dimension Low Sample Size (HDLSS) situation. Furthermore, the algorithms allow an evaluation of the data set for all scales and all locations simultaneously. By this approach, it may, for the multinormality testing, be detected if only some parts of the data set are originating from a multinormal distribution. For the k-sample case, the scale-space approach can detect if one or more of the k samples differ on different scales and/or locations. By not estimating the covariance matrix, the tests potentially loose some power compared to tests that incorporate the information from the estimated covariance matrix. This loss of power is acceptable on the grounds of being able to handle the HDLSS situation. As a result of the summation, the algorithms will include a large number of one-dimensional tests. For the results presented, the Anderson-Darling (AD) test is used for both the multinormality testing and the k-sample problem, see Section 2.5. The choice of this test is a result of its excellent power against all alternatives and existence of very good approximations for the asymptotic distribution and formulas adjusting for the finite sample sizes (Marsaglia and Marsaglia, 2004; Pettitt, 1976; Sinclair and Spurr, 1988).

A simple artificial example is presented to illustrate the main ideas of the paper. The artificial data set has a distribution that is multivariate normal for some of the dimensions and a mixture of two different normal distributions for the rest of the dimensions. Figure 1 shows 40 signals of length 50, i.e. the data matrix has size $[n, p] = [40, 50]$. There are two different underlying

true signals. In the first population, 20 signals are sampled from a zero mean multinormal distribution with covariance element (i, j) equal to $0.5 \cdot \phi^{|i-j|}$ with $\phi = 0.5$ (i.e. the signal is an autoregressive process of order 1). The remaining 20 signals have the same covariance structure, while the expected value equals -2.15 for index 6 to 12 and -3.5 for index 20. For indices $26, \ldots, 40$, the expected value increases linearly from 0.1 to 2.5, while the rest of the dimensions have expectation equal to zero.

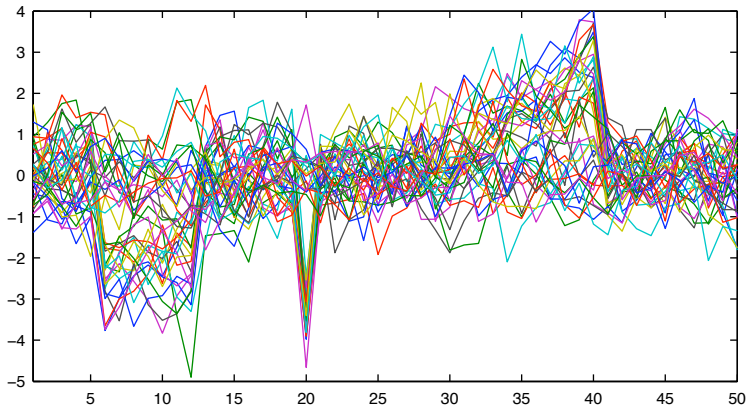

Figure 1: All 40 artificial signals of length 50

The output (called a significance map) from the proposed multinormality test of the data in Figure 1, is shown in Figure 2. On the horizontal axis are the 50 dimensions, while different window widths are given on the vertical axis. These different window widths represent the scale part of the presented algorithms. To be specific, a window width of 7 means that to produce the pixel in the significance map corresponding to window width/scale 7 and location 9, a weighted summation is performed across columns 6 to 12. The summation compresses this $[40, 7]$ part of the data matrix into a one-dimensional vector of length 40, which is then tested for normality. By going through all scale/location pairs, the significance map is produced. Red pixels mark rejections of the null hypothesis, i.e. indicating that the part of the data matrix which has been summed across cannot be considered as a sample from a multinormal distribution. The output is presented with the well-known Bonferroni approach for handling multiple testing (Hochberg and Tamhane, 1987). Later outputs will also present the result from the not so conservative False Discovery Rate (FDR) approach (Benjamini and Hochberg, 1995), see Section 2.3. For a distribution to be multinormal, all marginals must be normally distributed. This is a necessary, but not sufficient condition for multinormality. Scale 1 corresponds to testing the marginal distribution of each dimension. Note that the negative peak at dimension

20 is found on small scales, while the mixture density from dimension 6 to 12 and the increasing trend from dimensions 25 to 40 are found on larger scales. Therefore, this example shows that both short and long scales may be of importance in the same data set.

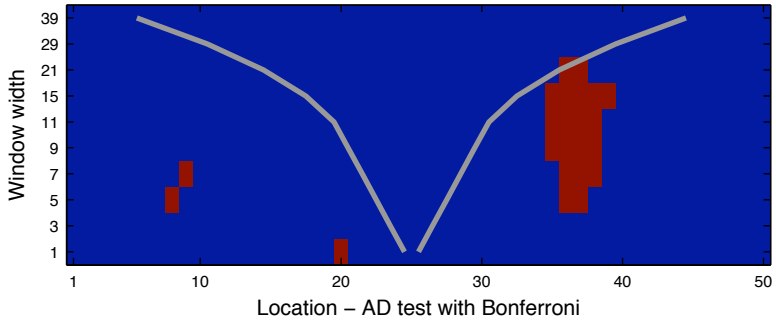

Figure 2: Significance map of the test for multinormality of a artificial data set. Red indicates rejection of the null hypothesis (multinormality) for that window width/location. For a given scale, the horizontal distance between the two gray lines equals the width of the summation window of that scale.

Section 2 presents the concept of scale-space, the statistical models being investigated and the details of the two presented algorithms. Some investigations into the power of the tests are also presented. In Section 3 the algorithms are applied to some real data sets, comparisons with other algorithms are done, and a feature selection scheme is presented and tested on real data. Section 4 contains a discussion of the results.

2. Methodology

Recall that an important motivation for applying a scale-space approach is the fact that different phenomena can be visible/detectable on different scales and/or locations of the data set. In classical nonparametric smoothing schemes, some sort of bandwidth parameter has to be chosen (Wand and Jones, 1995). By selecting one bandwidth only, features detectable on other bandwidths will not be found. However, using a scale-space approach, one can look at all bandwidths simultaneously. Scale-space ideas have proven useful in many areas and have been applied to feature detection in curves and images (Chaudhuri and Marron, 1999; Godtliebsen et al., 2004), density estimation (Godtliebsen et al., 2002), curve fitting (Chaudhuri and Marron, 2000), Bayesian time series analysis (Øigård et al., 2006) and spectral feature detection (Sørbye et al., 2009).

2.1. Model Assumptions

For the multinormality testing case, let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be a set of p-dimensional vectors. The null hypothesis assumes that these vectors originate from a non-specified p-dimensional multinormal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, i.e.

$$H_0 \ : \ \mathbf{X}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad \forall \ i,$$

where the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ are unknown. For the presented algorithm, the parameters of this assumed multinormal distribution do not have to be calculated. Note that by avoiding the need for an estimate of the covariance matrix, the algorithm could be applied to data sets with any combination of sample size and sample dimension, as long as the number of samples is high enough for the one-dimensional normality test to be applicable.

The algorithm works with any covariance structure and there are no requirements for smoothness of expected values of neighboring dimensions. As will be presented later, the algorithm performs a weighted summation across neighboring dimensions. A motivation behind this summation is that neighboring dimensions frequently have some sort of logical connection to each other, as for example in a time series. When the data set consists of a multidimensional time series, the different dimensions are equivalent to the different sampling times. If the dimensions are shifted around, the algorithm could produce different results. Therefore, interpretations of the results are easier when the different dimensions have a natural ordering, as for example with spatial or temporal data.

For the k-sample case, each of the k samples consist of a given number (which can be different for each k) of p-dimensional vectors with unknown cumulative distribution functions (CDF), given by $F_1, F_2, \cdots, F_k$, respectively. The null hypothesis is then stated as

$$H_0 \ : \ F_1(\mathbf{x}) = F_2(\mathbf{x}) = \cdots = F_k(\mathbf{x}), \quad \forall \ \mathbf{x} \in \mathbb{R}^p. \tag{1}$$

Since this methodology only tests whether or not the CDFs all are the same, the CDFs can take any form or belong to any class of distributions. Again, the interpretations of the results are easiest when working with data having a natural ordering.

2.2. Concept of Scale and Summation Across Dimensions

One of the main ideas of this manuscript is testing simultaneously for many different scales and locations. The scale values are the number of different dimensions being summed across. At the finest resolution, the scale is 1, corresponding to a test of the marginal distributions. At scale 3, the

result of the summation for location/dimension d is a weighted summation of the sample values with index $d-1$, d and $d+1$. For other scales, completely analogous summations are performed. Note that by this summation, small differences within the data can be detected, even though this difference might not be detected for lower scales. The set of default scales is chosen to be $\{1\ 3\ 5\ 7\ 9\ 11\ 15\ 21\ 29\ 39\ 51\ 65\ 81\ 99\ \ldots\ s_{\max}\}$, where the difference in neighboring scales increases by two for each scale up to a maximum scale $s_{\max} \leq p$. Alternatively, the user can choose to use scales up to some upper scale only. The number of scales used is designated n_s.

For each of the different scales s, a weighted summation across different dimensions/locations is performed, producing $\mathbf{L}_{s,d}$, where d is the location index ranging from 1 to p and $\mathbf{L}_{s,d}$ is a vector of length n. The resulting $\mathbf{L}_{s,d}$'s form a matrix $\underline{\mathbf{L}}$ with size $[n_s, p, n]$. For the summation weights a discrete Epanechikov (Wand and Jones, 1995) window function is used. For a given pair of s and d, the Epanechikov summation window is a column vector given by

$$\mathbf{w}_{s,d}(i) = K \cdot \left[1 - \left(\frac{i-d}{\lceil s/2 \rceil}\right)^2\right]_+, \qquad i = 1, \ldots, p,$$

where K is some normalizing constant, $\lceil \cdot \rceil$ is the ceiling function, and the plus-function is defined as $[f(x)]_+ = \max[0, f(x)]$ for some functional value $f(x)$. The $\mathbf{L}_{s,d}$ vector is generated through

$$\mathbf{L}_{s,d} = \underline{\mathbf{X}} \cdot \mathbf{w}_{s,d},$$

where the data matrix $\underline{\mathbf{X}}$ has size $[n, p]$, with the n samples of length p along each row, and $\cdot$ indicates normal matrix multiplication. The resulting vector $\mathbf{L}_{s,d}$ is thereby a weighted summation across the s dimensions centered on the d'th dimension. Figure 3 shows how the algorithm generates the $\underline{\mathbf{L}}$ matrix and how it is used to generate the output matrix, that is the significance map $\underline{\mathbf{R}}$, with different scales along the vertical axis and location along the horizontal axis.

As an example one can calculate the vector elements of the $\underline{\mathbf{L}}$ matrix corresponding to the scale/location pairs $(1,1)$, $(2,2)$ and $(3,4)$ of the data matrix

$$\underline{\mathbf{X}} = \begin{bmatrix} 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 3 & 2 \\ 1 & 1 & 0 & 1 & 1 \\ 2 & 1 & 1 & 0 & 0 \end{bmatrix}.$$

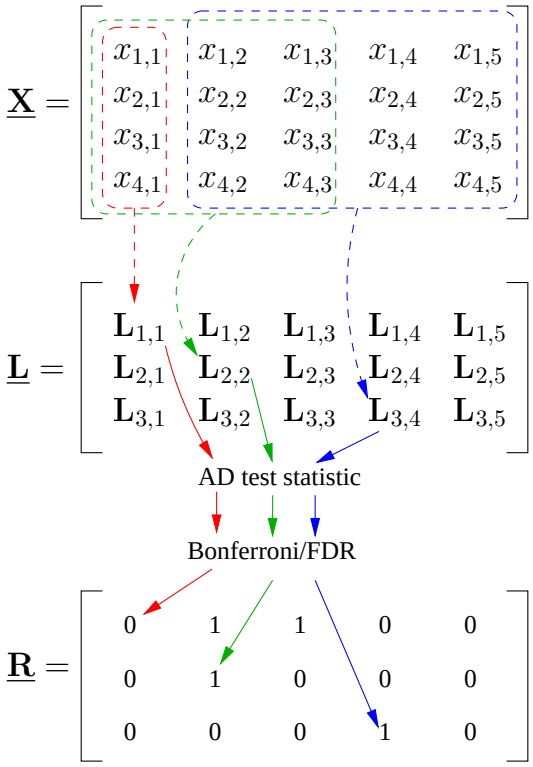

Figure 3: Workflow chart. The data matrix $\underline{\mathbf{X}}$ has dimensions $[n, p] = [4, 5]$. The summation matrix $\underline{\mathbf{L}}$ has dimensions $[n_s, p, n]$ and each $\mathbf{L}_{s,d}$ is a vector of length n. The significance matrix $\underline{\mathbf{R}}$ has dimensions $[n_s, p]$. The red box, which only spans one dimension, indicates that for the lowest scale, no summation is performed across the dimensions. For the green and blue boxes, summation is performed across dimensions 1–3 and 2–5, respectively. Note that two significance maps are produced, one each for the Bonferroni/FDR approaches, with ones in $\underline{\mathbf{R}}$ marking rejections of the null hypothesis for the corresponding scales and locations. When plotting the significance maps, the vertical axis is inverted.

The Epanechikov weights for the given scale/location pairs are $\mathbf{w}_{1,1} = [1,0,0,0,0]^T$, $\mathbf{w}_{2,2} = 1/10 \cdot [3,4,3,0,0]^T$, and $\mathbf{w}_{3,4} = 1/30 \cdot [0,5,8,9,8]^T$, where ^{T} indicates the transpose. The resulting vector elements are

$$\mathbf{L}_{1,1} = [0,0,1,2]^T \quad , \quad \mathbf{L}_{2,2} = \frac{1}{10} \cdot [0,7,7,13]^T \quad , \quad \mathbf{L}_{3,4} = \frac{1}{30} \cdot [17,56,22,13]^T.$$

2.3. Normality Testing

From the matrix $\underline{\mathbf{L}}$, the actual one-dimensional normality test statistics are calculated. For each of the (s,d) pairs, the p-value of the AD test statistics of the vector $\mathbf{L}_{s,d}$ is stored. To address the problem of multiple testing, the algorithm outputs two significance maps, one based on the Bonferroni approach and one based on FDR. The p-dimensional vector of p-values of each scale is fed into FDR, generating the FDR-based significance map scale by scale. For the Bonferroni approach, the critical value is obtained from the nominal significance level α divided by the number of dimensions p, producing on average at least one false alarm every $1/\alpha$ scale. This follows the usual SiZer recommendation of adjusting the significance for each scale separately. The alternative, adjusting the output map for all the time scales simultaneously, is known from the SiZer literature to be overly conservative (Chaudhuri and Marron, 1999). The nominal significance level can be chosen by the user, with a default value of $\alpha = 0.05$.

When using the multiple testing corrections, it is assumed that the tests are independent. This is of course not the case, both since the data might contain dependencies and since for scales larger than 1, the summations across dimensions will generate dependencies in the weighted sums. When the null hypothesis is true, these dependent tests will typically give a somewhat lower rejection ratio than what is expected from the nominal significance level for scales larger than 1.

2.4. The k-Sample Problem

For the k-sample problem, the k data matrices $\underline{\mathbf{X}}_i$, $i = 1, \cdots, k$ are all put through the summation procedure of Figure 3, producing $\underline{\mathbf{L}}_i$, $i = 1, \cdots, k$. For each scale/location pair (s,d), the k corresponding vectors (of size n_i, $i = 1, \cdots, k$) from the $\underline{\mathbf{L}}_i$ matrices are fed into the k-sample AD test (Pettitt, 1976; Scholz and Stephens, 1987). The distributions of the sums along the dimensions will in general be different from the marginal distributions. Nevertheless, if the k data sets do have the same multivariate distribution, for a given scale/location pair (s,d), the distributions of the k different summation vectors will be the same. The p-values of the tests are stored and used in the generation of the FDR-based significance map, while the Bonferroni approach finds the critical value as for the multinormality

testing. If the null hypothesis is rejected, the (s, d)-element of the output matrix is marked as a significant element, indicating that at least one of the empirical distributions are significantly different from the others.

2.5. Anderson-Darling Testing

The two algorithms presented use three different AD tests. The AD goodness-of-fit test is used in the case of checking for multinormality (Anderson and Darling, 1952, 1954; Lewis, 1961). For the two-sample/k-sample case, the versions of the AD test suggested by Pettitt (1976) and Scholz and Stephens (1987), respectively, are used.

The AD goodness-of-fit test checks the simple null hypothesis that a sample is from a distribution with a known continuous CDF, $F(x)$. Let $x_1 \leq x_2 \leq \cdots \leq x_n$ be the ordered sample of size n, and let $u_i = F(x_i), i = 1, \ldots, n$. The AD test statistic is defined as

$$A_n^2 \equiv -n - \frac{1}{n} \sum_{i=1}^{n} (2i - 1) \ln \left[u_i (1 - u_{n-i+1}) \right]. \tag{2}$$

This clearly shows that the AD test is distribution free, as long as the null distribution is fully known. Approximate expressions for the asymptotic distribution of the AD test are given by Marsaglia and Marsaglia (2004); Sinclair and Spurr (1988).

When testing for multinormality with unknown distributional parameters, i.e. testing a composite hypothesis, $F(x)$ is some unknown normal CDF, something which changes the distribution of the AD test statistic. In this case, the sorted data are normalized, producing $z_i, i = 1, \ldots, n$. Then, $u_i' = F_0(z_i)$ is produced, where $F_0(\cdot)$ is the standard normal CDF. These u_i' values are fed into Equation (2), and the final test statistic is obtained by applying the correction factor for finite sample sizes given on page 123 of D'Agostino and Stephens (1986). The p-values and critical values are calculated from the approximations given on page 127 of D'Agostino and Stephens (1986). Following page 373 of D'Agostino and Stephens (1986), the presented algorithm requires $n \geq 8$. The presence of ties in the data is a good indicator of non-normality, something which the AD test will reflect too. For instance, if normally distributed data is in some way rounded off, the rejection rate will be higher than the rate expected from the prescribed significance level.

For the k-sample case, there is no need to estimate any parameters, and the test statistic reduces to a rank statistic. Hence, the distribution of the test statistic is independent of the distribution of the k samples. The two-sample case and the k-sample case are treated separately, even though the k-sample

reduces to the two-sample case in Pettitt (1976) when $k = 2$. The correction factor in Pettitt (1976) is used to produce the final two-sample test statistic. Pettitt (1976) shows that the distribution of the sample-size adjusted two-sample AD test statistic can be approximated well by the asymptotic distribution of the AD goodness-of-fit test for a fully known null distribution. The presented algorithm uses Equation (3.6) in Sinclair and Spurr (1988) to produce the approximate p-value of the test statistic when $k = 2$.

The general k-sample AD test statistic in Scholz and Stephens (1987) is given as

$$A_{kN} \equiv \frac{1}{N} \sum_{i=1}^{k} \frac{1}{n_i} \sum_{j=1}^{N-1} \frac{(NM_{ij} - jn_i)^2}{j(N-j)},$$

where $N = n_1 + n_2 + \cdots + n_k$, and M_{ij} is the number of observations in the i'th sample that are not greater than the j'th observation of the pooled sample of all k samples. Equation (6) in Scholz and Stephens (1987) modifies the expression for A_{kN}, to be able to handle ties in the data. The presented algorithm uses the expression adjusted for ties, both for the two-sample and k-sample cases. Thereby, $F_i(x)$ in Equation (1) can be connected to a continuous or discrete random vector. The interpolation scheme of Scholz and Stephens (1987) is used to determine the p-value of A_{kN} when $k > 2$. Inspired by Pettitt (1976), it is required that all $n_i \geq 8,\ i = 1, \ldots, k$.

In theory, any omnibus test which achieves a specified significance level could be used in the presented framework. Relying on power studies by Shapiro and Wilk (1965); Stephens (1974); Thode Jr (2002), the well-known Shapiro-Wilk test (Shapiro and Wilk, 1965; Rahman and Govindarajulu, 1997; Royston, 1992) is seen as the best alternative to the AD test for the multinormality testing. Other tests which were considered include Watson's U^2 test (Watson, 1961), Kuiper's test (Stephens, 1965), Lilliefors' test (Lilliefors, 1967), the Cramér-von-Mises test (Stephens and Maag, 1968), the Shapiro-Francia test (Shapiro and Francia, 1972), D'Agostino-Pearson's K^2 test (D'Agostino et al., 1990; Pearson et al., 1977), the Jarque-Bera test (Jarque and Bera, 1980), and Doornik's test (Doornik and Hansen, 2008). Other tests considered for the k-sample case include the Kolmogorov-Smirnov test (Kiefer, 1959), the Cramér-von-Mises test (Kiefer, 1959), and Watson's U_k^2 test (Maag, 1966).

2.6. Cramér-Wold

The Cramér-Wold theorem states that two random column vectors $\mathbf{X}$ and $\mathbf{Y}$ have the same distribution if and only if for all row vectors $\mathbf{a}$, the random variables $\mathbf{a} \cdot \mathbf{X}$ and $\mathbf{a} \cdot \mathbf{Y}$ have the same distribution (Lehmann, 1998). In the presented algorithms, the different summation weights of the

Epanechikov window take the role of $\mathbf{a}$. Thereby, when doing the summation and testing for normality/difference between samples for many scales, a set of $\mathbf{a}$ vectors are applied to the single or many data sets. The Cramér-Wold theorem requires that the distribution of $\mathbf{a} \cdot \mathbf{X}$ and $\mathbf{a} \cdot \mathbf{Y}$ are equal for all possible $\mathbf{a}$ vectors. In the presented setting, only a finite number of vectors are tested. Since the presented algorithms are most suitable for data with some sort of neighboring structure (e.g. time series or spatial data), the important $\mathbf{a}$ vectors should be those that look at dimensions close to each other to a varying degree. Hence, following the Cramér-Wold theorem, a lack of rejection for (almost) all scales/locations should be seen as a good indication of the null hypothesis actually being true for the whole data set.

2.7. Significance of Rejections

For all the scale/location pairs, the p-value is available. The lower the p-value of a "rejection pair", the more significant the rejection of the null hypothesis is on that scale/location. In the graphical user interface (GUI) for the presented algorithms, the user can constantly change the desired significance level. By moving this significance level up and down, one can determine on which scales/locations the null hypothesis is most significantly rejected. In Figure 4 the example of the Introduction is revisited, where significance levels of 0.001 and 0.005 have been used, compared to the 0.05 in the Introduction. By comparing Figure 4 to Figure 2, it is clear that for this realization, the most significant region is the single non-normal dimension of index 20, and the region from index 26 to 40 is the second most non-normal.

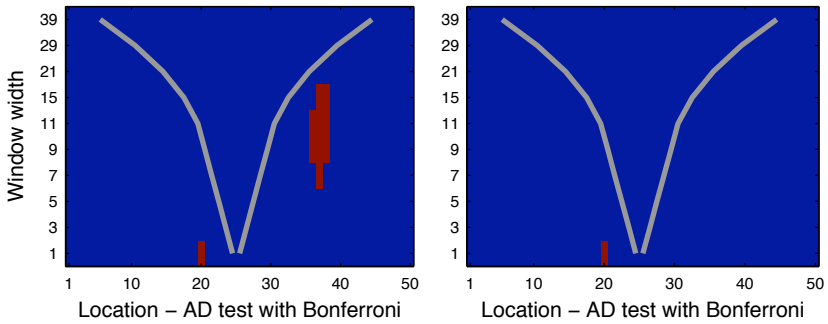

Figure 4: Significance maps of the scale-space multinormality test for the data of the Introduction with a significance level of 0.005/0.001 for left/right panel

2.8. Power of the Scale-Space Tests

There are no clear templates for power studies of the proposed scale-space tests. After the summations are done, the tests use the well-documented

AD tests. Thereby, the power of the scale-space tests is connected to the power of the AD tests. Instead, it can be informative to illustrate how the power varies over the different scale/location pairs of the output matrix for a given example. Assume that the data set has the same structure as in the motivational example of the Introduction. Figure 5 shows the rejection ratio (from 1 000 data sets) of the scale-space test for multinormality. As can be seen, one finds the highest powers for the scale/location pairs that best fit the non-normal dimensions. Similar results would be obtained for the test for comparing k data sets.

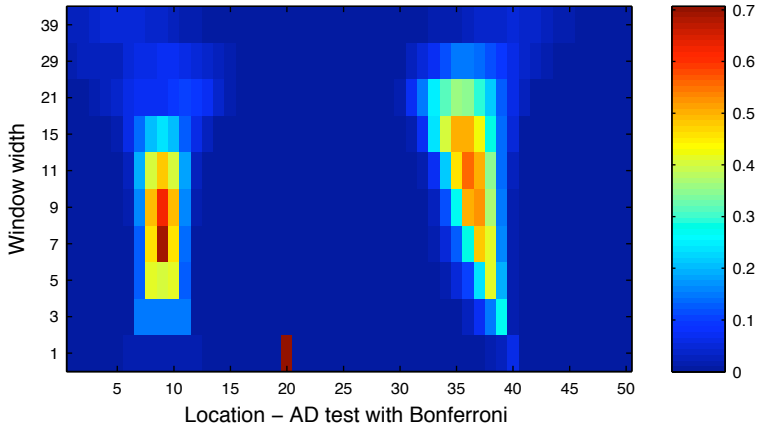

Figure 5: Rejection ratios of all scale/location pairs for 1 000 replications of the motivational example

To investigate the effect of increased number of dimensions, a number of normally distributed dimensions have been added to the right side of the signal of the Introduction. Table 1 shows the power of the multinormality test for different number of dimensions and for the FDR/Bonferroni correction. The case of 50 dimensions in total corresponds to the power of the pairs of Figure 5. From this it is clear that the power decreases as the number of dimensions grows.

3. Results

The suggested algorithms have been tested on a number of different data sets. A five percent significance level has been used for all the figures, unless otherwise stated. First, the initial example of the Introduction is investigated in more detail.

	Window width/location pair					
	1/20		7/9		9/37	
Dimensions in total	FDR	Bonf.	FDR	Bonf.	FDR	Bonf.
50	0.735	0.687	0.863	0.684	0.863	0.563
100	0.619	0.583	0.773	0.596	0.773	0.440
250	0.457	0.443	0.578	0.433	0.578	0.272
500	0.331	0.312	0.411	0.295	0.411	0.197
1 000	0.264	0.248	0.298	0.219	0.227	0.116

Table 1: Power of test for multinormality when the signal of the Introduction is augmented with a number of normally distributed dimensions

3.1. Introductory Example Revisited

For larger scales, the scale-space test for multinormality can be shown to increase the mode separation if the distribution has more than one mode. This is demonstrated through some simple examples. Assume that all the dimensions of some data set are unimodal normal (that is, all the n samples of a given dimension have the same normal distribution) with different means and/or variances for different dimensions. The result of the summation will then be some other normal distribution.

A short example of this is given. Assume that the data matrix $\underline{\mathbf{X}}$ has dimensions $[10, 3]$ and that all ten samples of column 1 is $\mathcal{N}(0, 1)$ distributed, while column 2 is $\mathcal{N}(4, 1)$ distributed, and column 3 is $\mathcal{N}(8, 1)$ distributed. The summation (for simplicity, assuming even weights of $1/3$) over these three columns would produce a 10-element long vector with distribution $\mathcal{N}(4, 1/3)$, which the AD test would detect as normal, i.e. the test would not reject it.

Now assume that the ten samples of a given dimension do not have the same distribution. Assume that the five first samples of the three columns are distributed as $\mathcal{N}(1, 1)$, while the last five are distributed as $\mathcal{N}(0, 1)$. When checking the columns separately, the 10-element vector might not "look" enough different from a unimodal normal distribution to be rejected by the AD test. When summing (again, assuming even weights of $1/3$) over the three columns, the distribution of the sum of the first five samples is given by $\mathcal{N}(1, 1/3)$, while the last five have a $\mathcal{N}(0, 1/3)$ distribution. This shows that the peaks have larger separation (both variances have decreased) as a result of the summation.

3.2. Multinormality of Temperature Data

A data set obtained from the Norwegian Meteorological Institute is analyzed. The data show daily mean temperature for the 92 days of June–August

for the period 1937 to 2008 at Blindern, Oslo. This gives a data matrix of dimensions $[n, p] = [72, 92]$, making algorithms that rely on inversion of the estimated covariance matrix impossible to use. A plot of all the 72 years is given in Figure 6.

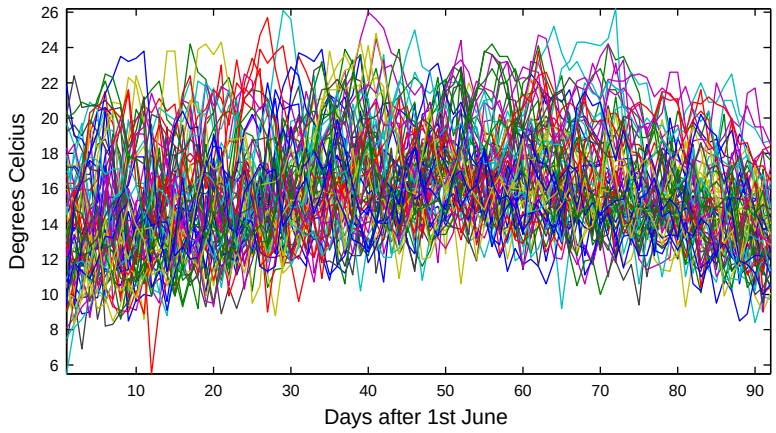

Figure 6: Daily mean temperatures at Oslo, Blindern, for the period 1937-2008

Figure 7 gives the results of checking for multinormality. Note that significant features are found both for the FDR and Bonferroni correction. To see what is going on, the period around time point 75 (i.e. in the middle of August) is shown in Figure 8. From this figure it seems that the mean temperature is around 15°C, but the temperature distribution around this time is skewed upwards. This means that Oslo at this time of the year experiences larger positive than negative deviations from the mean, something that is not a surprising result if you have knowledge about the temperature in that area.

3.3. Comparison of Temperature Records

For the comparison of data sets, temperature data sets are once more used. Temperature data sets from two different meteorological stations in the Oslo area are compared. The first one is located at Ferder lighthouse at the start of the 100 km long Oslo fjord, while the second one is located at Fornebu, which is at the very inner part of the Oslo fjord. The two data sets consist of more or less overlapping yearly records, respectively 64 and 45 complete years (years with missing data in the months of interest have been removed). Figure 9 shows the two data sets, and Figure 10 shows the resulting significance maps. From this, it is clear that the temperature distribution at the two stations differ early and late in the summer. From

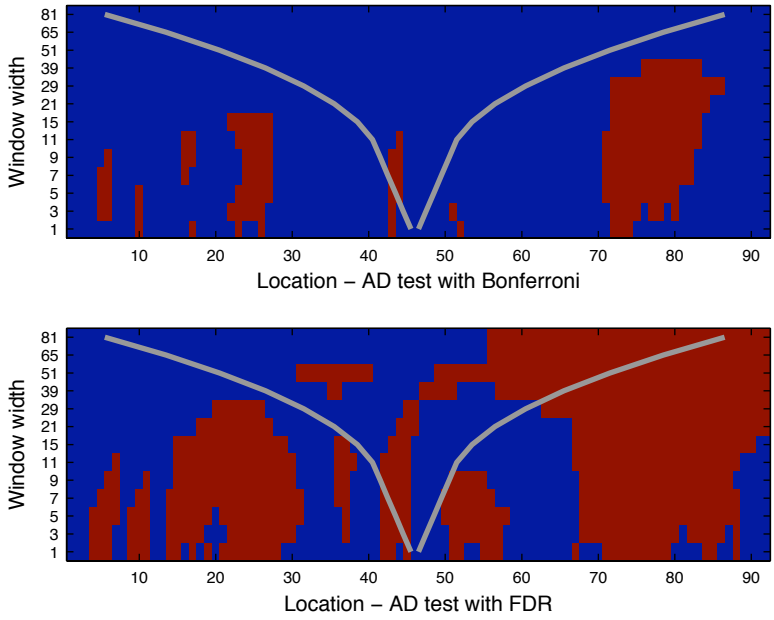

Figure 7: Significance maps for summer temperatures in Oslo

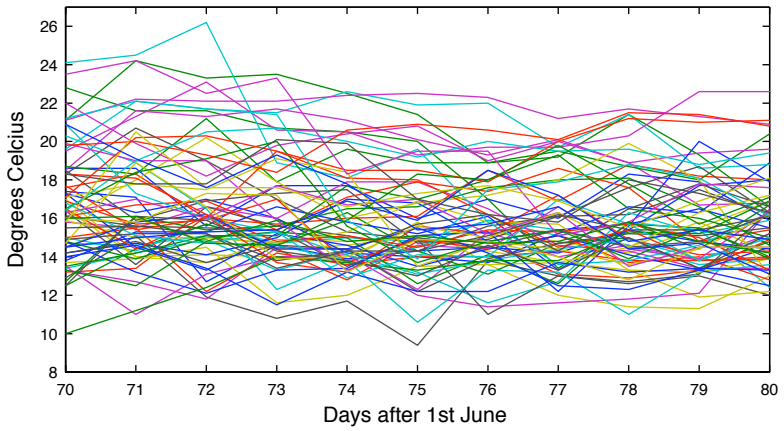

Figure 8: Mid-August temperatures in Oslo for the years 1937–2008

a closer inspection, it is clear that Fornebu is warmer during early summer, while it is the opposite during late summer.

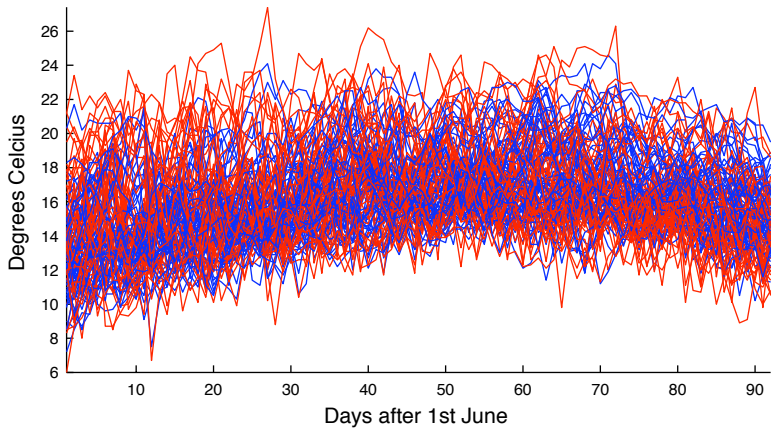

Figure 9: Temperature data from Ferder (blue) and Fornebu (red)

3.4. Comparison to Other Methods

Just about all methods for testing for multinormality rely in some way on inverting the estimated covariance matrix. When the number of samples is less or equal to the number of dimensions, i.e. $n \leq p$, the estimated covariance matrix is non-invertible. To the authors' knowledge, only the projection methods in Liang et al. (2000, 2009) work when $n \leq p$. The preferred method of Liang et al. (2000) first transforms the data matrix, and then projects it onto some lower-dimensional space of dimension $d \leq \min(n-2, p)$. The transformed data will under the null hypothesis be distributed as a d-dimensional standard multinormal distribution, something which is checked using the skewness and kurtosis test of Mardia (1970). Asymptotic distributions are given, but in the setting of interest (n is not large compared to p), the use of the Liang test relies on a permutation procedure for generating p-values.

It is not straightforward to compare the presented scale-space method to the Liang procedure since the presented scale-space method does not produce one single answer to the hypothesis testing problem. To illustrate that the presented method outperforms the Liang test in some settings, a simple example has been tested. Assume the same data set structure as in the example of the Introduction, except that the only non-normal part is the mixture of dimensions 6 to 12, the other dimensions are zero mean normally distributed. This setup results in the optimal scale/location pair being $(4, 9)$, i.e. summing over dimensions 6 to 12. When the non-zero mean value in this area is

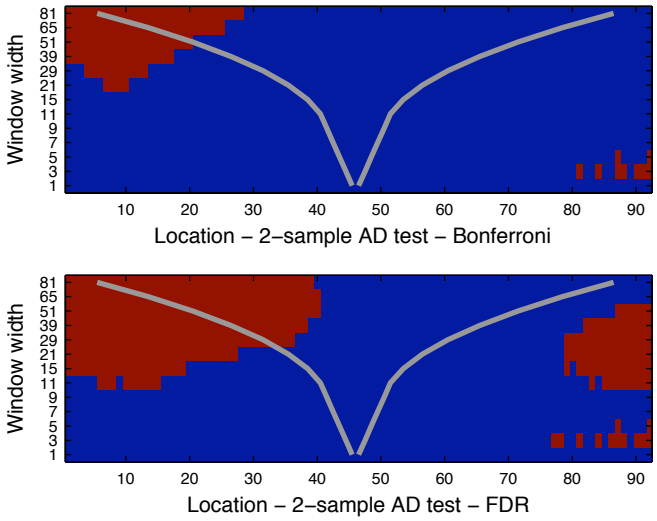

Figure 10: Significance maps from comparing the temperature data of Ferder and Fornebu with the scale-space method

2.35, the presented scale-space method has a detection ratio of 0.884/0.918 (Bonferroni/FDR) for the pair $(4, 9)$ (based on 1 000 Monte Carlo repetitions). The Liang test has for the same data sets a rejection ratio of 0.659. For the Liang test only the kurtosis test and only the optimal projection dimension ($d = 1$) are used. In a real setting, the optimal projection dimension would not be known and both the skewness and kurtosis test would be used, leading to a significantly lower power when the correction for multiple testing is done. In the same way, when the non-zero mean value is 2.05, the presented scale-space method has a rejection ratio of 0.569/0.628 for the pair $(4, 9)$, while the Liang test has for the same data sets a rejection ratio of 0.480.

For the comparison of two or more data sets, there are several methods which handle the $n \leq p$ situation: Friedman and Rafsky (1979); Hall and Tajvidi (2002); Henze (1988); Székely and Rizzo (2004). The test by Székely and Rizzo is a k-sample extension of the two-sample test suggested by Baringhaus and Franz (2004). A similar two-sample test was suggested by Aslan and Zech (2005). This test performed very similar to, but not better than, the Székely-Rizzo/Baringhaus test in the two-sample test case of Table 2. All these methods use some kind of distance measure between the data vectors, and from these distances the test statistics are generated, without estimating any covariance matrices. For the case of interest ($n \leq p$), the tests all rely on permutation procedures to determine the p-value of the test statistic. The

case of two data sets X and Y is first investigated. The expected value of X is zero for all dimensions, while Y has one region of a number of neighboring dimensions with a non-zero expected value. Both X and Y have the same covariance structure as the example of the Introduction. The number of dimensions of Y that have a non-zero mean value is varied, along with this non-zero value. The upper part of Table 2 shows the results, where as before the result of the scale-space algorithm refers to the scale/location pair with the highest rejection ratio.

	Dim: 1	**Dim: 3**	**Dim: 5**	**Dim: 7**
Two-sample	$\boldsymbol{\delta = 0.85}$	$\boldsymbol{\delta = 0.75}$	$\boldsymbol{\delta = 0.65}$	$\boldsymbol{\delta = 0.55}$
Scale-space	0.549/0.560	0.713/0.739	0.694/0.760	0.622/0.740
Székely-Rizzo	0.412	0.780	0.845	0.843
Hall-Tajvidi	0.158	0.374	0.488	0.519
Nearest Neighbor	0.246	0.460	0.542	0.539
Friedman-Rafsky	0.265	0.498	0.576	0.571
Three-sample	$\boldsymbol{\delta = 0.75}$	$\boldsymbol{\delta = 0.55}$	$\boldsymbol{\delta = 0.50}$	$\boldsymbol{\delta = 0.45}$
Scale-space	0.706/0.721	0.623/0.648	0.675/0.737	0.676/0.763
Székely-Rizzo	0.373	0.559	0.740	0.771
Seven-sample	$\boldsymbol{\delta = 0.15}$	$\boldsymbol{\delta = 0.11}$	$\boldsymbol{\delta = 0.10}$	$\boldsymbol{\delta = 0.09}$
Scale-space	0.739/0.755	0.617/0.642	0.694/0.744	0.697/0.771
Székely-Rizzo	0.299	0.477	0.632	0.688

Table 2: Power of comparing a number of different data sets with a varying number of dimensions ("Dim") for which there is a distributional difference $\boldsymbol{\delta}$ in the tested data sets. For the Hall test, the T and S tests gave very similar results. Three nearest neighbors were used in the Nearest Neighbor test. The results of the Friedman-Rafsky test are for three trees, which consistently performed better than one and two threes in this setting. The scale-space results are for the Bonferroni/FDR correction, respectively, and a 0.10 significance level is used. 2 000 Monte Carlo samples are used.

Of the alternative tests, the method of Székely and Rizzo (2004) consistently shows the greatest power in the settings tested. When the difference between X and Y is across many dimensions, the power of the Székely test is higher than the power of the scale-space approach. If there instead is only one dimension with a different distribution of X and Y, the power of the scale-space approach is greater than for the Székely test. This means that the Székely is a good alternative approach, but by using the scale-space approach one can determine where in the data set the difference is located.

Except for the Hall method, the methods can all be extended to the case of $k > 2$ data sets. For the case of $k = 3$, the presented scale-space method

has only been compared to the method of Székely and Rizzo. The same covariance structure as for the two-sample case has been used for the three data sets X, Y and Z. X is zero mean, while Y has for some neighboring dimensions a non-zero expected value of $\boldsymbol{\delta}$, and Z has for the same dimensions a non-zero expected value of $-\boldsymbol{\delta}$, see the middle part of Table 2. The case of $k = 7$ is finally investigated in the lower part of Table 2. Here, the different data sets have the same structure as for the case of $k = 3$, but the different data sets $X_i,\ i = 1, 2, \ldots, 7$ have mean value equal to $i \cdot \boldsymbol{\delta}$ for the non-zero dimensions. From these results, the scale-space method seems to improve compared to the Székely-Rizzo method when the number of data sets increase, and the methods are giving comparable results in the tested settings.

3.5. Feature Selection

In a classification setting, the p-values of the different scale/location pairs could be used to find useful scale-space features. The pairs with the smallest p-values should be good candidate features for classification algorithms. The p-values of neighboring pairs will be correlated (for all scales larger than 1). An ad hoc strategy to avoid the selection of neighboring pairs has been used. That is, say that the most significant pair is at window width 7 (i.e. scale number 4) and location 5. Then, all pairs for two scales down (scale number 2 and 3) and two scales up (scale number 5 and 6) which sum over the data of dimension/location 5, are excluded from being selected as a feature as a result of pair $(4, 5)$ being selected as a feature. The next feature to be selected corresponds to the scale/location pair, which has not been excluded in the steps before, with the lowest p-value of the pairs not already selected. This is repeated until a wanted number of features are found or there are no good features left to pick from, where a potential feature's "goodness" would be connected to its p-value.

The suggested feature selection algorithm has been tested on a setting similar to the example of the Introduction. Here, instead of having one data set with two parts, there are two data sets X and Y. X is distributed as the 20 first samples of the motivational example, while Y is distributed as the 20 remaining samples, except that the expected value equals -0.65 for index 6 to 12 and -1 for index 20. For indices $26, \ldots, 40$, the expected value increases linearly from 0.05 to 0.75.

The suggested feature selection scheme has been compared to using all dimensions as inputs to classification algorithms. This is meant as a proof of concept, more than a thorough comparison to other methods. The tested sample sizes of both X and Y were 20, 30 and 60. For the classification,

k Nearest Neighbor classification (with k=1 and k = 3), Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) were used, when applicable (Hastie et al., 2009). For the scale-space feature selection, the number of features selected ranged from 1 to 15. One pair of X and Y data sets was used to find the training features. These features were then used to classify 500 X and 500 Y data sets. This was repeated 100 times, making up in total 100 000 tests, and the ratio of correct classification was averaged across these 100 000 tests, as shown in Figure 11. The splitting up was done to average out the fact that different features will be selected depending on the training data set. With three well-selected features, one could capture the main differences in the two data sets, but as the figure shows, one needs on average more than three features to have the maximum ratio of correct classification. The figure shows that using the suggested scale-space features is better than using the raw data in this example.

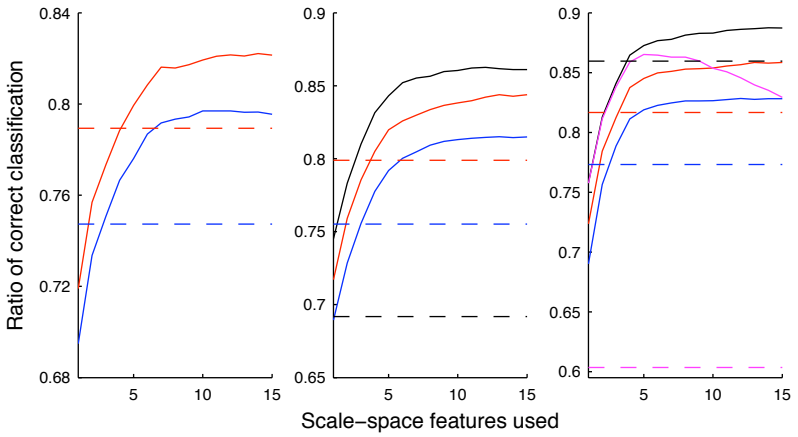

Figure 11: Results of using a varying number of scale-space features (solid lines) compared to using all dimensions (dashed lines) for classification through 1NN (blue), 3NN (red), LDA (black), QDA (magenta). The vertical axis shows the ratio of correct classifications based on 100 000 simulations. Sample sizes from left to right are: 20, 30 and 60.

4. Discussion

The scale-space approach is applied to the testing for multivariate normality and to the k-sample problem. The summation across scales reduces the multivariate problem to a large number of one-dimensional tests. A significance map, showing where and for which scales the null hypothesis is rejected, is generated by going through all combinations of the location and

scale parameters. The summation throws away all information of the dependency structure of the data. When there are more samples than dimensions, i.e. $n > p$, the discharging of covariance information will lower the power of the scale-space tests compared to tests which use this information gained through estimation of the covariance matrix. What is gained on the other hand, is the ability to check for multinormality and compare data sets in the High Dimension Low Sample Size setting, something which almost all other methods fail to handle.

The presented algorithms have been tested on artificial data and real temperature data sets, showing how both the check for multinormality and the comparison of data sets could be done through the scale-space approach.

Within the scale-space framework, to the authors' best knowledge, there is no other algorithm to compare the presented work with, even though a large number of tests for assessing the multinormality of a given data set exist (Alva and Estrada, 2009; Mecklin and Mundfrom, 2004; Romeu and Ozturk, 1993; Thode Jr, 2002). To the knowledge of the authors, the only multivariate methods for testing multinormality that handle the case when $n \leq p$, are the methods by Liang et al. (2000, 2009), where the method from 2000 is the preferred method, as stated in Liang et al. (2009). As have been shown, this method is inferior to the presented method in some relevant aspects and cases.

In the case of comparing k data sets, there exist some methods which handle the case where at least one of the sample sizes are less than the number of dimensions. In general, these methods are based on some distance measure between the data vectors, and do not estimate the covariance matrix. The suggested scale-space method has been compared to these methods. In the tested settings, the power of the method of Székely and Rizzo (2004) is comparable to the power of the scale-space approach. The Szekely test does not on the other hand provide any info about where the data sets differ, information that is essential for doing feature selection. Selection of relevant features based on the presented scale-space k-sample problem algorithm is demonstrated in Section 3.

Acknowledgement

Fred Godtliebsen was financially supported by project 176872/V30 from the eVita program in the Norwegian Research Council. Jan Hannig's research was supported in part by the National Science Foundation under Grant No. 0707037 and 1007543. Our MATLAB implementation of the k-sample AD test is strongly influenced by the implementation of Trujillo-Ortiz et al. (2007).

References

Alva, J. A. V., Estrada, E. G., 2009. A generalization of Shapiro-Wilk's test for multivariate normality. Communications in Statistics - Theory and Methods 38 (11), 1870–1883.

Anderson, T. W., Darling, D. A., June 1952. Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. Annals of Mathematical Statistics 23 (2), 193–212.

Anderson, T. W., Darling, D. A., December 1954. A test of goodness of fit. Journal of the American Statistical Association 49 (268), 765–769.

Aslan, B., Zech, G., February 2005. Statistical energy as a tool for binning-free, multivariate goodness-of-fit tests, two-sample comparison and unfolding. Nuclear Instruments and Methods in Physics Research A 537 (3), 626–636.

Baringhaus, L., Franz, C., 2004. On a new multivariate two-sample test. Journal of Multivariate Analysis 88 (1), 190–206.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological) 57 (1), 289–300.

Chaudhuri, P., Marron, J., September 1999. SiZer for exploration of structures in curves. Journal of the American Statistical Association 94 (447), 807–823.

Chaudhuri, P., Marron, J. S., April 2000. Scale space view of curve estimation. The Annals of Statistics 28 (2), 408—428.

Cox, D. R., Wermuth, N., 1994. Tests of linearity, multivariate normality and the adequacy of linear scores. Journal of the Royal Statistical Society. Series C (Applied Statistics) 43 (2), 347–355.

D'Agostino, R. B., Belanger, A., Ralph B. D'Agostino, J., November 1990. A suggestion for using powerful and informative tests of normality. The American Statistician 44 (4), 316–321.

D'Agostino, R. B., Stephens, M. A. (Eds.), 1986. Goodness-of-fit Techniques. Vol. 68 of Statistics: Textbooks and Monographs. Marcel Dekker, New York, NY.

Doornik, J. A., Hansen, H., December 2008. An omnibus test for univariate and multivariate normality. Oxford Bulletin of Economics and Statistics 70 (s1), 927–939.

Farrell, P. J., Salibian-Barrera, M., Naczk, K., December 2007. On tests for multivariate normality and associated simulation studies. Journal of Statistical Computation and Simulation 77 (12), 1065–1080.

Friedman, J. H., Rafsky, L. C., July 1979. Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. The Annals of Statistics 7 (4), 697–717.

Godtliebsen, F., Marron, J. S. ., Chaudhuri, P., March 2002. Significance in scale space for bivariate density estimation. Journal of Computational and Graphical Statistics 11 (1), 1–21.

Godtliebsen, F., Marronb, J. S., Chaudhuri, P., November 2004. Statistical significance of features in digital images. Image and Vision Computing 22 (13), 1093–1104.

Hall, P., Tajvidi, N., June 2002. Permutation tests for equality of distributions in high-dimensional settings. Biometrika 89 (2), 359–374.

Hastie, T., Tibshirani, R., Friedman, J. H., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition. Springer Series in Statistics. Springer, New York, NY.

Henze, N., June 1988. A multivariate two-sample test based on the number of nearest neighbor type coincidences. The Annals of Statistics 16 (2), 772–783.

Hochberg, Y., Tamhane, A. C., 1987. Multiple Comparison Procedures. Wiley series in probability and mathematical statistics. Applied probability and statistics. John Wiley & Sons, New York, NY.

Jarque, C. M., Bera, A. K., 1980. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. Economics Letters 6 (3), 255–259.

Kiefer, J., June 1959. K-sample analogues of the Kolmogorov-Smirnov and Cramér-V. Mises tests. Annals of Mathematical Statistics 39 (2), 420–447.

Lehmann, E. L., 1998. Elements of Large-Sample Theory. Springer, New York, NY.

Lewis, P. A. W., December 1961. Distribution of the Anderson-Darling statistic. Annals of Mathematical Statistics 32 (4), 1118–1124.

Liang, J., Li, R., Fang, H., Fang, K.-T., April 2000. Testing multinormality based on low-dimensional projection. Journal of Statistical Planning and Inference 86 (1), 129–141.

Liang, J., Tang, M.-L., Chan, P. S., September 2009. A generalized Shapiro-Wilk W statistic for testing high-dimensional normality. Computational Statistics and Data Analysis 53 (11), 3883–3891.

Lilliefors, H., June 1967. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. Journal of the American Statistical Association 62 (318), 399–402.

Looney, S. W., February 1995. How to use tests for univariate normality to assess multivariate normality. The American Statistician 49 (1), 64–70.

Maag, U. R., December 1966. A k-sample analogue of Watson's U^2 statistic. Biometrika 53 (3/4), 579–583.

Mardia, K. V., December 1970. Measures of multivariate skewness and kurtosis with applications. Biometrika 57 (3), 519–530.

Marsaglia, G., Marsaglia, J., February 2004. Evaluating the Anderson-Darling distribution. Journal of Statistical Software 9 (2), 1–5.

Mecklin, C. J., Mundfrom, D. J., April 2004. An appraisal and bibliography of tests for multivariate normality. International Statistical Review 72 (1), 123–138.

Øigård, T. A., Rue, H., Godtliebsen, F., December 2006. Bayesian multiscale analysis for time series data. Computational Statistics and Data Analysis 51 (3), 1719—1730.

Pearson, E. S., D'Agostino, R. B., Bowman, K. O., August 1977. Tests for departure from normality: Comparison of powers. Biometrika 64 (2), 231–246.

Pettitt, A. N., April 1976. A two-sample Anderson-Darling rank statistic. Biometrika 63 (1), 161–168.

Rahman, M. M., Govindarajulu, Z., 1997. A modification of the test of Shapiro and Wilk for normality. Journal of Applied Statistics 24 (2), 219–235.

Romeu, J. L., Ozturk, A., August 1993. A comparative study of goodness-of-fit tests for multivariate normality. Journal of Multivariate Analysis 46 (2), 309–334.

Royston, P., September 1992. Approximating the Shapiro-Wilk W-test for non-normality. Statistics and Computing 2 (3), 117–119.

Scholz, F. W., Stephens, M. A., September 1987. K-sample Anderson-Darling tests. Journal of the American Statistical Association 82 (399), 918–924.

Shapiro, S. S., Francia, R. S., March 1972. An approximate analysis of variance test for normality. Journal of the American Statistical Association 67 (337), 215–216.

Shapiro, S. S., Wilk, M. B., December 1965. An analysis of variance test for normality (complete samples). Biometrika 52 (3/4), 591–611.

Sinclair, C. D., Spurr, B. D., December 1988. Approximations to the distribution function of the Anderson-Darling test statistic. Journal of the American Statistical Association 83 (404), 1190–1191.

Sørbye, S. H., Hindberg, K., Olsen, L. R., Rue, H., September 2009. Bayesian multiscale feature detection of log-spectral densities. Computational Statistics and Data Analysis 53 (11), 3746–3754.

Stephens, M. A., December 1965. The goodness-of-fit statistic V_n: Distribution and significance points. Biometrika 52 (3/4), 309–321.

Stephens, M. A., September 1974. EDF statistics for goodness of fit and some comparisons. Journal of the American Statistical Association 69 (347), 730–737.

Stephens, M. A., Maag, U. R., July 1968. Further percentage points for W_N^2. Biometrika 55 (2), 428–430.

Székely, G. J., Rizzo, M. L., November 2004. Testing for equal distributions in high dimensions. InterStat (5), 1–16.

Thode Jr, H. C., 2002. Testing for Normality. Marcel Dekker, New York, NY.

Trujillo-Ortiz, A., Hernandez-Walls, R., Barba-Rojo, K., Cupul-Magana, L., Zavala-Garcia, R. C., 2007. AnDarksamtest: Anderson-Darling k-sample procedure to test the hypothesis that the populations of the drawned groups are identical. A MATLAB file.
URL `http://www.mathworks.com/matlabcentral/fileexchange/17451`

Wand, M. P., Jones, M. C., 1995. Kernel Smoothing. Vol. 60 of Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, Boca Raton, FL.

Watson, G. S., June 1961. Goodness-of-fit tests on a circle. Biometrika 48 (1/2), 109–114.