On Hypothesis Testing for Comparing Image Quality Assessment Metrics

Rui Zhu, Fei Zhou, Wenming Yang and Jing-Hao Xue

In developing novel image quality assessment (IQA) metrics, researchers should compare their proposed metrics with state-of-the-art metrics. A commonly adopted approach is by comparing two residuals between the nonlinearly mapped scores of two IQA metrics and the difference mean opinion score, which are assumed from Gaussian distributions with zero means. An F-test is then used to test the equality of variances of the two sets of residuals. If the variances are significantly different, then we conclude that the residuals are from different Gaussian distributions and that the two IQA metrics are significantly different. The F-test assumes that the two sets of residuals are independent. However, given that the IQA metrics are calculated on the same database, the two sets of residuals are paired and may be correlated. We note this improper usage of the F-test by practitioners, which can result in misleading comparison results of two IQA metrics. To solve this practical problem, we introduce the Pitman test to investigate the equality of variances for two sets of correlated residuals. Experiments on the LIVE database show that the two tests can provide different conclusions.

Introduction

Image quality assessment (IQA) is a popular research topic in image processing. Several widely used IQA metrics, such as noise quality measure (NQM) [1], structural similarity index (SSIM) [2], multiscale structural similarity (MS-SSIM) [3], visual information fidelity [4], feature similarity index (FSIM) [5] and gradient similarity (GSM) [6], have been proposed in the last several decades.

Researchers should compare their proposed IQA metrics with state-of-the-art metric to validate the superiority of their metrics. Such comparisons are typically performed following the procedures proposed in [7].

To test whether two IQA metrics are significantly different, Sheikh et al. [7] used the hypothesis test on two sets of residuals between the nonlinearly mapped scores calculated from each of the two IQA metrics and the difference mean opinion score (DMOS). In [7], one assumption is that the two sets of residuals are samples from Gaussian distributions with zero means. Therefore, to test whether the two sets of residuals are from the same distribution, we only need to test whether the two sets of residuals present the same variance. Sheikh et al. [7] adopted a simple F-test to investigate the equality of variances of two sets of residuals.

The F-test assumes that the two samples are independent [8]. However, we note that the two sets of residuals in IQA can be correlated, thereby invalidating the independence assumption in the F-test. In particular, when comparing two IQA metrics, we apply the metrics to the same database, resulting in paired scores calculated from the two IQA metrics, with one residual in the first IQA metric uniquely matched with one residual in the second IQA metric on the same image. The paired scores of the two IQA metrics are correlated; for example, as degradation on an image increases, the scores from the two IQA metrics can both decrease. Thus, the two residuals between the DMOS and the two nonlinearly mapped scores may also be correlated.

When the two samples are correlated, the F-test cannot provide reliable results on the equality of variances. Therefore, the conclusion whether the two IQA metrics are statistically different based on the F-test is not reliable.

Pitman and Morgan [9, 10] developed a test to examine the equality of variances for two correlated samples. In the Pitman test statistic, the Pearson correlation coefficient is involved to consider the effect of the correlation between samples. Instead of using the F-test, we introduce the Pitman test to examine the equality of variances for comparing two IQA metrics.

Using the F-test to compare IQA metrics

Supposing that the aim is to compare the scores $\mathbf{x}$ and $\mathbf{y}$ calculated from two IQA algorithms (after the nonlinear mapping) on the same database with $\mathbf{x} = [x_1, x_2, \ldots, x_N]^T \in \mathbb{R}^{N\times 1}$ and $\mathbf{y} = [y_1, y_2, \ldots, y_N]^T \in \mathbb{R}^{N\times 1}$, where N is the number of images in the database and supposing that the DMOS for the images in the database is denoted as $\mathbf{z} = [z_1, z_2, \ldots, z_N]^T \in \mathbb{R}^{N\times 1}$, then the two residuals between the DMOS and the two nonlinearly mapped scores are $\mathbf{d}_x = \mathbf{z} - \mathbf{x}$ and $\mathbf{d}_y = \mathbf{z} - \mathbf{y}$.

Sheikh et al. [7] assumed that $\mathbf{d}_x = [d_{x1}, d_{x2}, \ldots, d_{xN}]^T \in \mathbb{R}^{N\times 1}$ and $\mathbf{d}_y = [d_{y1}, d_{y2}, \ldots, d_{yN}]^T \in \mathbb{R}^{N\times 1}$ are the samples drawn from Gaussian distributions with zero means. Therefore, testing whether $\mathbf{d}_x$ and $\mathbf{d}_y$ are from the same Gaussian distribution becomes testing whether their population variances σ_x^2 and σ_y^2 are the same.

The F-test is adopted in [7] to test the equality of variances. The null hypothesis H_0 is $\sigma_x^2/\sigma_y^2 = 1$ and the alternative hypothesis H_1 is $\sigma_x^2/\sigma_y^2 \neq 1$. Given the two sets of

samples $\mathbf{d}_x$ and $\mathbf{d}_y$, the F-test statistic is calculated as

$$F = \frac{s_x^2}{s_y^2}, \tag{1}$$

where $s_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(d_{xi}-\bar{d}_x)^2$ and $s_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(d_{yi}-\bar{d}_y)^2$ are the sample variances, and $\bar{d}_x = \frac{1}{N}\sum_{i=1}^{N} d_{xi}$ and $\bar{d}_y = \frac{1}{N}\sum_{i=1}^{N} d_{yi}$ are the sample means. The F-test statistic has an F distribution with $N-1$ and $N-1$ degrees of freedom.

Therefore, the test conclusion is drawn by comparing the value of F in (1) with the critical values of an F distribution with $N-1$ and $N-1$ degrees of freedom. Given the significance level α (which usually takes values of 1%, 5% or 10%), the null hypothesis is rejected $F > F_{1-\alpha/2,N-1,N-1}$ or $F < F_{\alpha/2,N-1,N-1}$, where $F_{1-\alpha/2,N-1,N-1}$ and $F_{\alpha/2,N-1,N-1}$ are the critical values. Then, we conclude that $\mathbf{d}_x$ and $\mathbf{d}_y$ are from different Gaussian distributions. Otherwise, if $F_{\alpha/2,N-1,N-1} \leq F \leq F_{1-\alpha/2,N-1,N-1}$, we do not reject the null hypothesis, and the conclusion is that $\mathbf{d}_x$ and $\mathbf{d}_y$ are from the same Gaussian distribution.

Why the F-test is unsuitable for comparing IQA metrics?

The F-test assumes that $\mathbf{d}_x$ and $\mathbf{d}_y$ are two independent samples from Gaussian populations. However, $\mathbf{d}_x$ and $\mathbf{d}_y$ may be correlated because the samples are paired; one sample d_{xi} in $\mathbf{d}_x$ is uniquely paired with one sample d_{yi} in $\mathbf{d}_y$ because they are calculated on the same ith image. Such samples are called paired samples in statistics. If the two IQA metrics are both well designed, their scores both decrease as the degree of degradation increases in the same image. Such correlations between scores may also render the residuals $\mathbf{d}_x$ and $\mathbf{d}_y$ as correlated. Empirical evidence of the correlation between residuals is provided later in experimental results. Therefore, the conclusion drawn from the F-test of whether $\mathbf{d}_x$ and $\mathbf{d}_y$ are from the same distribution can be unreliable.

The Pitman test as a solution

In statistics, a hypothesis test for paired samples is usually different from that for independent samples. For example, the t-test is used to test the equality of means for independent samples, whereas the paired t-test is used for paired samples. For evaluating the equality of variances, the Pitman test is designed for correlated samples [9, 11, 12].

Here, we introduce the Pitman test to examine the equality of variances for the residuals of two IQA metrics. The null hypothesis H_0 is $\sigma_x^2 = \sigma_y^2$, and the alternative hypothesis H_1 is $\sigma_x^2 \neq \sigma_y^2$. The Pitman test statistic is calculated as

$$t = \frac{(1 - s_x^2/s_y^2)\sqrt{N-2}}{\sqrt{4(1-r^2)(s_x^2/s_y^2)}}, \tag{2}$$

where

$$r = \frac{\sum_{i=1}^{N}(d_{xi}-\bar{d}_x)(d_{yi}-\bar{d}_y)}{\sqrt{\sum_{i=1}^{N}(d_{xi}-\bar{d}_x)^2 \sum_{i=1}^{N}(d_{yi}-\bar{d}_y)^2}} \tag{3}$$

is the Pearson correlation coefficient between the two sets of samples $\mathbf{d}_x$ and $\mathbf{d}_y$. It is clear that in (2) the correlation r is considered in the test statistic. The Pitman test statistic exhibits a Student's t distribution with $N-2$ degrees of freedom.

Similar to that in the F-test, the test conclusion is drawn by comparing the value of t in (2) with the critical values of a t distribution with $N-2$ degrees of freedom.

The F-test versus the Pitman test

In Fig. 1, we illustrate the use of the F-test and the Pitman test in comparing IQA metrics. Two IQA metrics $\mathbf{M}_x$ and $\mathbf{M}_y$ are applied to the same IQA database, providing two residuals $\mathbf{d}_x$ and $\mathbf{d}_y$, respectively. In the comparison of $\mathbf{d}_x$ and $\mathbf{d}_y$ by using the F-test, two assumptions are applied: 1) independence between $\mathbf{d}_x$ and $\mathbf{d}_y$ and 2) normality of $\mathbf{d}_x$ and $\mathbf{d}_y$, as shown in Fig. 1(a). By contrast, when the Pitman test is used, the only assumption is the normality of $\mathbf{d}_x$ and $\mathbf{d}_y$, as shown in Fig. 1(b). Given that $\mathbf{d}_x$ and $\mathbf{d}_y$ are paired and correlated, the Pitman test is more appropriate to test the equality of variances than the F-test.

Experimental results

In the following experiments, we aim to test whether $\mathbf{d}_x$ and $\mathbf{d}_y$ are from Gaussian distributions with the same variances on the LIVE database. We show that different conclusions can be drawn from the F-test and the Pitman test.

Following the experiments in [7], all experiments are performed on five types of degradations (JPEG2000, JPEG, Gaussian noise, Gaussian blur and fast-fading wireless) separately and then on the overall database.

We compare the scores of the following seven IQA metrics: FSIM [5], GSM [6], most apparent distortion (MAD) [13], MS-SSIM [3], NQM [1], peak signal to noise ratio (PSNR) and SSIM [2]. All scores and their nonlinearly mapped scores are obtained from `http://sse.tongji.edu.cn/linzhang/IQA/IQA.htm`.

The significance levels of the F-test and the Pitman test are both set to 5%.

Are two conclusions different?

The results show that, for all types of degradations and the overall database, the F-test does not always produce the same conclusion as that by the Pitman test regarding whether two IQA metrics are statistically significantly different. Here, we

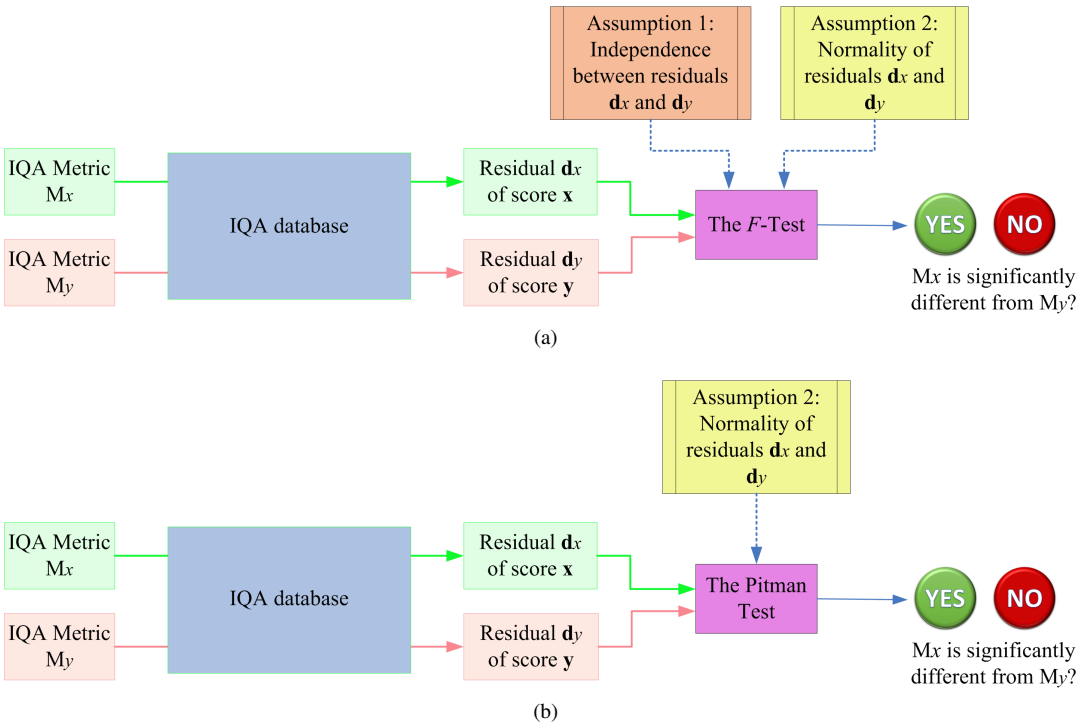

Fig. 1. Illustrations of the use of (a) the F-test and (b) the Pitman test in the comparison of IQA metrics.

Table 1. Differences between the conclusions drawn from the F-test and the Pitman test for the overall database and Gaussian noises.

	Same conclusions	Different conclusions		Total
		Pitman: $\mathbf{M}_x$ and $\mathbf{M}_y$ are different F: $\mathbf{M}_x$ and $\mathbf{M}_y$ are the same	Pitman: $\mathbf{M}_x$ and $\mathbf{M}_y$ are the same F: $\mathbf{M}_x$ and $\mathbf{M}_y$ are different	
Overall database	19	2	0	21
Gaussian noises	18	3	0	21

show two examples on the overall database and the Gaussian noises in Table 1.

A total of 21 pairs of IQA metrics are compared in the experiments. For the overall database, we obtain 19 same conclusions and 2 different conclusions from the two tests. The two pairs of IQA metrics with different conclusions are (PSNR, GSM) and (PSNR, MS-SSIM). Similar results are obtained for the Gaussian noises, that is, 18 same conclusions and 3 different conclusions on (PSNR, FSIM), (PSNR, MS-SSIM) and (NQM, MS-SSIM). In addition, all different conclusions present the same pattern: the Pitman test concludes that $\mathbf{M}_x$ and $\mathbf{M}_y$ are different, whereas the F-test concludes that $\mathbf{M}_x$ and $\mathbf{M}_y$ are the same. For example, the well-known MS-SSIM is empirically superior to PSNR. However, the F-test cannot tell their difference, whereas the Pitman test can statistically distinguish between the pair.

We can formulate two observations from the above results. First, the F-test and the Pitman test can provide the same conclusions for most comparisons of IQA metrics. However, different conclusions exist for certain cases. Second, the Pitman test can detect more statistically significantly unequal IQA metrics than the F-test for correlated samples. This finding is reasonable because a high correlation r results in increased absolute value of t in (2), given the fixed s_x^2 and s_y^2. Thus, with a larger absolute value of t, the Pitman test is more likely to reject the null hypothesis, compared with the F-test.

Are two residuals correlated?

The Pearson correlation coefficients between the two sets of residuals for all tests with different conclusions and the same conclusions are box-plotted in Fig. 2. From these two boxplots, we can observe the following patterns.

First, almost all correlations are nonzeros, with the exception of several outliers. This finding empirically demonstrates our argument that the two sets of residuals from the two IQA metrics may be correlated.

Second, the median of the correlations in the left boxplot

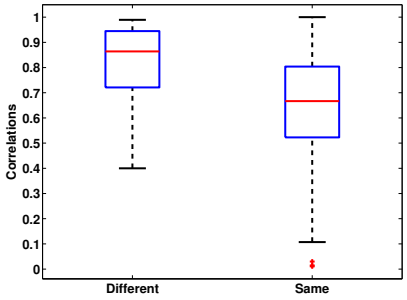

Fig. 2. Correlations between the pairs of residuals for IQA metrics with different conclusions (the left boxplot) and the same conclusions (the right boxplot) between the F-test and the Pitman test.

is close to 0.9 and is markedly higher than that in the right boxplot. This finding suggests that with a high correlation between the two sets of residuals, the two statistical hypothesis tests tend to provide different conclusions.

Recommendations for practitioners

On the basis of experimental results, we offer the following suggestions for the comparison of IQA metrics.

When the correlation between two scores (or particularly residuals) is low, the Pitman test and the F-test can provide the same comparison result. However, with a high correlation, the Pitman test and the F-test tend to provide different answers. In this case, we trust the results of the Pitman test, which is specifically designed for correlated samples. Therefore, to obtain reliable results for all cases, we suggest for practitioners to use the Pitman test to evaluate the equality of variances for comparing two IQA metrics. With the Pitman test, several methods that were reported to be statistically indistinguishable in the literature can be determined to be statistically significantly different.

Summary

In this article, we introduce the Pitman test to address the problem of using the F-test in comparing IQA metrics when the independence assumption is invalidated. However, if the normality assumption is also violated, then the power of the Pitman test also decreases. In this case, nonparametric tests without the normality assumption may provide superior solutions.

Authors

Rui Zhu (r.zhu@kent.ac.uk) received her Ph.D. degree in statistics from University College London in 2017. She is a lecturer in the School of Mathematics, Statistics & Actuarial Science, University of Kent. Her research interests include spectral data analysis, hyperspectral image analysis, subspace-based classification methods and image quality assessment.

Fei Zhou (flying.zhou@163.com, corresponding author) received his Ph.D. degree from the Department of Electronic Engineering, Tsinghua University in 2013. He is currently a visiting scholar of the Department of Statistical Science, University College London. His research interests include applications of image processing and pattern recognition techniques in video surveillance, image super-resolution, image interpolation, image quality assessment and object tracking.

Wenming Yang (yangelwm@163.com) received his Ph.D. degree in electronic engineering from Zhejiang University in 2006. He is an associate professor in the Department of Electronic Engineering, Graduate School at Shenzhen, Tsinghua University. His research interests include image processing, pattern recognition, computer vision, biometrics, video surveillance and image super-resolution.

Jing-Hao Xue (jinghao.xue@ucl.ac.uk) received his Dr.Eng. degree in signal and information processing from Tsinghua University in 1998 and Ph.D. degree in statistics from the University of Glasgow in 2008. He is a senior lecturer in the Department of Statistical Science, University College London. His research interests include statistical machine learning, high-dimensional data analysis, pattern recognition and image analysis.

References

[1] Niranjan Damera-Venkata, Thomas D Kite, Wilson S Geisler, Brian L Evans, and Alan C Bovik, "Image quality assessment based on a degradation model," *IEEE Transactions on Image Processing*, vol. 9, no. 4, pp. 636–650, 2000.

[2] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[3] Zhou Wang, Eero P Simoncelli, and Alan C Bovik, "Multiscale structural similarity for image quality assessment," in *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on*. IEEE, 2003, vol. 2, pp. 1398–1402.

[4] Hamid R Sheikh and Alan C Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, 2006.

[5] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.

[6] Anmin Liu, Weisi Lin, and Manish Narwaria, "Image quality assessment based on gradient similarity," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1500–1512, 2012.

[7] Hamid R Sheikh, Muhammad F Sabir, and Alan C Bovik, "A statistical evaluation of recent full reference image quality as-

sessment algorithms," *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3440–3451, 2006.

[8] Robert L Mason, Richard F Gunst, and James L Hess, *Statistical design and analysis of experiments: with applications to engineering and science*, vol. 474, John Wiley & Sons, 2003.

[9] EJG Pitman, "A note on normal correlation," *Biometrika*, vol. 31, no. 1/2, pp. 9–12, 1939.

[10] WA Morgan, "A test for the significance of the difference between the two variances in a sample from a normal bivariate population," *Biometrika*, vol. 31, no. 1/2, pp. 13–19, 1939.

[11] William G Cochran, "Testing two correlated variances," *Technometrics*, vol. 7, no. 3, pp. 447–449, 1965.

[12] James Lee, "Comparison of variance between correlated samples," *Bioinformatics*, vol. 8, no. 4, pp. 405–406, 1992.

[13] Eric C Larson and Damon M Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *Journal of Electronic Imaging*, vol. 19, no. 1, pp. 011006–011006, 2010.