Schedae Informaticae Vol. 27 (2018): 59–68
doi: 10.4467/20838476SI.18.005.10410

On Some Goodness of Fit Tests for Normality Based on the Optimal Transport Distance

Marcin Mazur and Piotr Kościelniak
Jagiellonian University
Faculty of Mathematics and Computer Science
Lojasiewicza 6, 30-348 Krakow, Poland
e-mail: *{marcin.mazur, piotr.koscielniak}@uj.edu.pl*

Abstract. We apply the optimal transport distance to construct two goodness of fit tests for (univariate) normality. The derived statistics are then compared with those used by the Shapiro-Wilk, the Anderson-Darling and the Cramer-von Mises tests. In particular, we preform Monte Carlo experiments, involving computations of the test power against some selected alternatives and wide range of sample sizes, which show efficiency of the obtained test procedures.

Keywords: Goodness of fit test (for normality), optimal transport distance, Wasserstein distance, autoencoder-based generative model.

1. Introduction

Transportation theory, concerning the problem of optimal mass redistribution, since its formalization in 1781 by G. Monge [6] founded application in many fields of science, including statistics and machine learning.

In [3] del Barrio et al presented the concept of a construction of a goodness of fit test for normality which was based on the ℓ_2-Wasserstein distance (which is a special case of an optimal transport distance) between a sample distribution and the set of normal distributions. They also proved that the obtained procedure (later called the BCMR test) is asymptotically equivalent to the Shapiro-Wilk test. This idea was then investigated in, e.g., [5], where simulations were carried out to show the competitiveness of the BCMR test in relation to the best known normality tests.

The high usefulness of the Wasserstain distance in training autoencoder-based generative models was confirmed by, e.g., the original work of Tolstikhin et al [9], or

the work of Kolouri et al [4], where a "sliced" version of the Wasserstain autoencoder was considered. On the other hand, in [7] the authors proposed a method for training generative autoencoders by explicitly testing the distribution of the code layer output via goodness of fit tests for normality.

In this paper we construct two goodness of fit tests for normality, which are based on the optimal transport distance, but use a slightly different idea than that from [3]. In order to compare them with the Shapiro-Wilk, the Anderson-Darling and the Cramer-von Mises tests, we perform Monte Carlo experiments, involving computations of the test power against some selected alternatives and wide range of sample sizes. The obtained results show some improvement over the existing procedures in many of the cases considered.

2. Optimal transport distance

Let μ_1 and μ_2 be probability measures on $\mathbb{R}^d$. The optimal transport distance between μ_1 and μ_2 is defined as (see, e.g., [10]):

$$W_c(\mu_1, \mu_2) = \inf_{\gamma \in \Gamma(\mu_1, \mu_2)} \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x_1, x_2)\, d\gamma(x_1, x_2),$$

where $\Gamma(\mu_1, \mu_2)$ is the set of joint probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ having μ_1 and μ_2 as marginals, and $c\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$ is a transportation cost function. Then $W_c(\mu_1, \mu_2)$ determines the cheapest way to "pushing forward" μ_1 into μ_2, and can be interpreted as a divergence between μ_1 and μ_2. If $c(\cdot, \cdot) = \rho^p(\cdot, \cdot)$ for some metric ρ in $\mathbb{R}^d$ and $p \geq 1$, then $W_c^{1/p}(\mu_1, \mu_2)$ is called the p-th Wasserstein distance between μ_1 and μ_2.

It is clear that (for the purposes of this work) we can limit ourselves to the case of $d = 1$, in which the optimal transport distance between one dimensional probability measures μ_1 and μ_2 with cumulative probability functions P_{μ_1} and P_{μ_2}, respectively, is given by the following closed formula (see, e.g., [8]):

$$W_c(\mu_1, \mu_2) = \int_0^1 c(P_{\mu_1}^{-1}(t), P_{\mu_2}^{-1}(t))\, dt, \tag{1}$$

where $P_{\mu_1}^{-1}(t) = \inf\{x \in \mathbb{R} : P_{\mu_1}(x) \geq t\}$ and $P_{\mu_2}^{-1}(t) = \inf\{x \in \mathbb{R} : P_{\mu_2}(x) \geq t\}$ for $t \in (0, 1)$. Since in our work we are only interested in the distance between an empirical distribution $\mu_{\underline{x}}$ of a sample $\underline{x} = (x_1, \ldots, x_n)$, and a given reference distribution μ_0, from now on we will be using the notation $W_c(\underline{x}, \mu_0)$ instead of $W_c(\mu_{\underline{x}}, \mu_0)$.

As one can easily guess, the usefulness of the optimal transport distance (for various purposes) depends on the specific cost function applied. In the following few paragraphs we make appropriate choices of c, in order to obtain respective closed forms of $W_c(\underline{x}, \mu_0)$, which will allow us to construct appropriate goodness of fit tests.

The ℓ_2 case. We define a cost function as a square of the ℓ_2 distance, i.e., $c_{\ell_2}(x, y) = (x - y)^2$. Then (1) gives, in fact, the square of the 2-nd Wasserstein

distance, and we can calculate:

$$W_{c_{\ell_2}}(\underline{x},\mu_0)=\int_0^1 (P_{\underline{x}}^{-1}(t)-P_0^{-1}(t))^2\,dt=\int_0^1\left(\sum_{i=1}^n x_{(i)}\mathbf{1}_{\frac{i-1}{n}<t\le\frac{i}{n}}-P_0^{-1}(t)\right)^2 dt$$
$$=\frac{1}{n}\sum_{i=1}^n x_{(i)}^2-2\sum_{i=1}^n x_{(i)}\int_{\frac{i-1}{n}}^{\frac{i}{n}} P_0^{-1}(t)\,dt+\int_{-\infty}^{\infty} y^2\cdot p_0(y)\,dy$$
$$=\frac{1}{n}\sum_{i=1}^n x_{(i)}^2-2\sum_{i=1}^n x_{(i)}\int_{Q_{\frac{i-1}{n}}}^{Q_{\frac{i}{n}}} y\cdot p_0(y)\,dy+m^2+\sigma^2. \tag{2}$$

Here and henceforth: (i) $(x_{(1)},\ldots,x_{(n)})$ means an ordered sample $\underline{x}$, (ii) $P_{\underline{x}}=P_{\mu_{\underline{x}}}$ and $P_0=P_{\mu_0}$, (iii) p_0, m, σ and Q_r denote the density function, the mean, the variance and the r-th quantile of μ_0, respectively.

The ℓ_2 case with a factor. Let $\mathrm{supp}(\mu_0)$ denote the interior of the set of all $x\in\mathbb{R}$ such that $\mu_0((x-\varepsilon,x+\varepsilon))>0$ for every $\varepsilon>0$. We multiply the cost function c_{ℓ_2} by a function factor $f\colon \mathrm{supp}(\mu_0)\to\mathbb{R}^+$, i.e., we take $c_{\ell_2,f}(x,y)=(x-y)^2\cdot f(y)$, and then, making similar calculations as in (2), we obtain:

$$W_{c_{\ell_2,f}}(\underline{x},\mu_0)=\int_0^1 (P_{\underline{x}}^{-1}(t)-P_0^{-1}(t))^2\cdot f(P_0^{-1}(t))\,dt=\sum_{i=1}^n x_{(i)}^2\int_{Q_{\frac{i-1}{n}}}^{Q_{\frac{i}{n}}} f(y)\cdot p_0(y)\,dy$$
$$-2\sum_{i=1}^n x_{(i)}\int_{Q_{\frac{i-1}{n}}}^{Q_{\frac{i}{n}}} y\cdot f(y)\cdot p_0(y)\,dy+\int_{-\infty}^{\infty} y^2\cdot f(y)\cdot p_0(y)\,dy, \tag{3}$$

provided all the above integrals exist.

3. Goodness of fit test statistics

As it has been already mentioned, the distance $W_c(\underline{x},\mu_0)$, with all the variants of a transportation cost function c, can be a candidate for a test statistic to verify the hypothesis that the sample $\underline{x}$ comes from the distribution μ_0 (we reject such hypothesis for large values of $W_c(\underline{x},\mu_0)$). Moving on this way, if we are interested in the construction of a goodness of fit test for normality, we can simply test the hypothesis that the standardized sample $\underline{y}=(y_1,\ldots,y_n)$ with $y_i=\frac{x_i-\overline{x}}{s}$, where $\overline{x}$ and s are the mean and the standard deviation of $\underline{x}$, i.e, $\overline{x}=\frac{1}{n}\sum_{i=1}^n x_i$ and $s^2=\frac{1}{n-1}\sum_{i=1}^n (x_i-\overline{x})^2$, comes from the standard normal distribution $N(0,1)$. This idea looks compatible[1] with that used in [3] to construct the BCMR test. However, in this paper we propose a slightly different concept of construction of a test statistics, basing on fact that if Y is a random variable then the variable $P_Y(Y)$ has the continuous uniform distribution $U(0,1)$. Therefore, as a "measure of nonnormality" we can use

[1] It also results from our simulations, which we do not include here due to the page limit.

$W_c(\underline{z}, U(0,1))$, where $\underline{z} = (P_{N(0,1)}(y_1), \ldots, P_{N(0,1)}(y_n))$, instead of $W_c(\underline{y}, N(0,1))$. In the following paragraphs we follow this approach (in reference to the choices of a cost function made in the previous section) to obtain a few goodness of fit test for normality, while in the next section we compare them (by examining the test power against several selected alternatives and wide range of sample sizes) with the Shapiro-Wilk, the Anderson-Darling and the Cramer-von Mises tests.

The ℓ_2 case. Using (2) we calculate:

$$W_{c_{\ell_2}}(\underline{z}, U(0,1)) = \frac{1}{n}\sum_{i=1}^{n} z_{(i)}^2 - 2\sum_{i=1}^{n} z_{(i)} \int_{\frac{i-1}{n}}^{\frac{i}{n}} y\,dy + \frac{1}{4} + \frac{1}{12} = \frac{1}{n}\sum_{i=1}^{n} z_{(i)}^2 - \frac{1}{n^2}\sum_{i=1}^{n} z_{(i)} \cdot (i^2 - (i-1)^2) + \frac{1}{3} = \frac{1}{n}\sum_{i=1}^{n} z_{(i)}^2 + \frac{1}{n^2}\sum_{i=1}^{n} z_{(i)}(2i-1) + \frac{1}{3}. \tag{4}$$

Then applying (4) to construct a goodness of fit test for normality, we obtain a well-known Cramér-von Mises procedure. Indeed, it is easy to verify that $W_{c_{\ell_2}}(\underline{z}, U(0,1))$ coincides with the Cramér-von Mises distance between $P_{\underline{y}}$ and $P_{N(0,1)}$ (see, e.g., [1]), hence we have $\omega^2(\underline{y}) = n \cdot W_{c_{\ell_2}}(\underline{z}, U(0,1))$, where ω^2 denotes the Cramér-von Mises statistics.

The ℓ_2 case with a factor that improves the sensitivity in tails ("divergent" version). We use (3) with a factor function defined by the following formula: $f_1(y) = [P_0(y) \cdot (1 - P_0(y))]^{-1}$. Note that f tends to ∞ as $P_0(y)$ goes to 0 or 1, which causes that tails of μ_0 influence the cost function much more than in the previous case. Then we obtain:

$$\begin{aligned} W_{c_{\ell_2,f_1}}(\underline{z}, U(0,1)) &= \sum_{i=1}^{n} z_{(i)}^2 \int_{\frac{i-1}{n}}^{\frac{i}{n}} \frac{dy}{y(1-y)} - 2\sum_{i=1}^{n} z_{(i)} \int_{\frac{i-1}{n}}^{\frac{i}{n}} \frac{dy}{1-y} + \int_0^1 \frac{y}{1-y}\,dy \\ &= \sum_{i=1}^{n} z_{(i)}^2 \left(\ln\frac{n-i+1}{n-i} + \ln\frac{i}{i-1}\right) + \sum_{i=1}^{n}(1 - 2z_{(i)}) \ln\frac{n-i+1}{n-i} \\ &\quad - 1 = \sum_{i=1}^{n} z_{(i)}^2 \ln\frac{i}{i-1} + \sum_{i=1}^{n}(1 - z_{(i)})^2 \ln\frac{n-i+1}{n-i} - 1. \end{aligned} \tag{5}$$

However, as it can be evidently noticed, there is a problem with the above formula consisting in the fact that its two components are infinite (since the respective integrals are, in fact, divergent). In order to deal with this matter, we simply omit the infinite components in (5), obtaining the following formula:

$$\overline{W}_{c_{\ell_2,f_1}}(\underline{z}, U(0,1)) = \sum_{i=2}^{n} z_{(i)}^2 \ln\frac{i}{i-1} + \sum_{i=1}^{n-1}(1 - z_{(i)})^2 \ln\frac{n-i+1}{n-i} - 1. \tag{6}$$

It turns out (as evidenced by our experiments) that applying $\overline{W}_{c_{\ell_2,f_1}}(\underline{z}, U(0,1))$ as a statistics for a goodness of fit test for normality seems to give quite good results, excluding cases when we were dealing with samples from distributions with short tails, such as, e.g., a continuous uniform distribution (see Figure 1). Hence, in order

to increase the test power, we consider an additional simple "distance" $C(\underline{z}, U(0,1))$, defined as:

$$C(\underline{z}, U(0,1)) = z_{(1)} + (1 - z_{(n)}), \tag{7}$$

which works well as a test statistics for samples from distributions with short tails (see Figure 1 for the results of our experiments). Then, to finish a construction of a required test procedure for verifying normality, we combine (6) and (7) with suitable adjustments required for multiple testing problems, i.e., Bonferroni correction, and call this test W^B. More precisely, if pv_1 and pv_2 are p-values of the tests using respectively (6) and (7) as test statistics, then p-value of the W^B test is established as $\min\{2pv_1, 2pv_2, 1\}$ (see, e.g., [2] for further details).

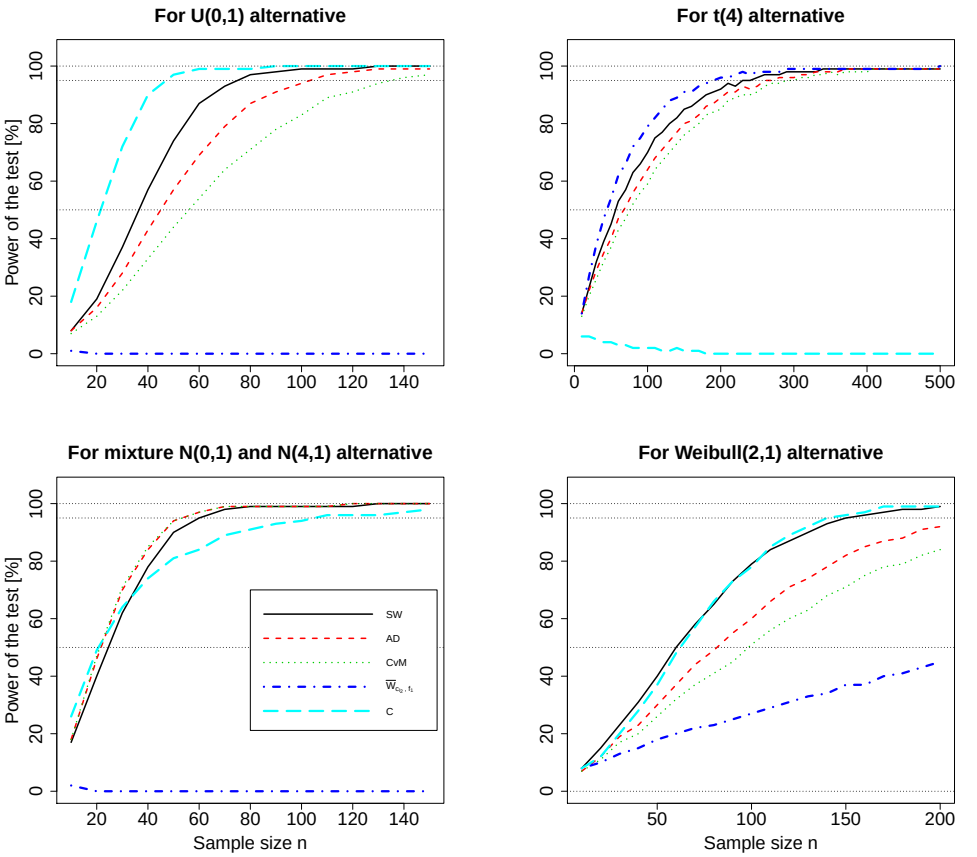

Figure 1. Simulation of the power of the $\overline{W}_{c_{\ell_2,f_1}}$ and the C tests against 4 different alternatives.

The ℓ_2 case with a factor that improves the sensitivity in tails ("convergent" version). In order to avoid problems with "infinities", as it was in the previous case, using the following formula we define slightly different factor function for which all

the integrals in (3) exist: $f_2(y) = [P_0(y) \cdot (1 - P_0(y))]^{-\frac{1}{2}}$. Then we can calculate:

$$W_{c_{\ell_2,f_2}}(\underline{z}, U(0,1)) = \sum_{i=1}^{n} z_{(i)}^2 \int_{\frac{i-1}{n}}^{\frac{i}{n}} \frac{dy}{\sqrt{y(1-y)}} - 2\sum_{i=1}^{n} z_{(i)} \int_{\frac{i-1}{n}}^{\frac{i}{n}} \frac{y}{1-y}\, dy + \int_0^1 \sqrt{\frac{y^3}{1-y}}\, dy$$
$$= \sum_{i=1}^{n} z_{(i)}(1 - z_{(i)}) \left(\arcsin \frac{n-2i}{n} - \arcsin \frac{n-2(i-1)}{n} \right)$$
$$+ \sum_{i=1}^{n} 2z_{(i)} \left(\frac{\sqrt{i(n-i)}}{n} - \frac{\sqrt{(i-1)(n-(i-1))}}{n} \right) + \frac{3}{8}\pi, \qquad (8)$$

and use (8) to construct a goodness of fit test for normality.

4. Power comparisons

In this section we present the results of our Monte Carlo experiments that compare the power of the constructed tests with the other known normality tests.

All the simulations were performed by applying suitable procedures (some of them implemented by the authors) in R programming language (version 3.5.1). Distributions of test statistics (for the tests presented in this work) were obtained (for each sample size n) by simulation using 100000 repetitions, while the power (and probability of Type I error) of each test was calculated from 5000 repetitions. In all cases the level of significance was set at $\alpha = 0.05$.

n	10	20	30	40	50	60	70	80	90	100
SW	0.0488	0.052	0.05	0.0528	0.0598	0.051	0.047	0.0518	0.0512	0.048
AD	0.051	0.0534	0.0464	0.0556	0.0554	0.0482	0.0464	0.0516	0.048	0.0462
CvM	0.049	0.0516	0.0498	0.0534	0.0582	0.0456	0.048	0.0514	0.0472	0.0478
W^B	0.0434	0.0492	0.046	0.0508	0.0496	0.0482	0.0502	0.05	0.0466	0.0524
$W_{c_{\ell_2,f_2}}$	0.0474	0.0522	0.045	0.0546	0.0564	0.044	0.0456	0.0516	0.0482	0.0452

Table 1. The Type I error simulation.

We started our simulations by checking whether the probability of Type I error is controlled at the level $\alpha = 0.05$. The results obtained indicate that Type I error is controlled correctly for each of the tests considered (see Table 1). In order to study the power of the tests we considered 8 typical alternative distributions with different properties: symmetric distributions with short tails (U(0,1) and Beta(2,2)) and long tails (Student's $t(4)$), as well as asymmetric distributions with different tails (lognormal, Gamma(2,1), Weibull(2,1) and $\chi^2(5)$). Finally, we examined also one two-modal symmetric distribution (the mixture of $N(0,1)$ and $N(4,1)$). We compared the

W^B and the $W_{c_{\ell_2,f_2}}$ tests with the Shapiro-Wilk (SW), the Anderson-Darling (AD) and the Cramer-von Mises (CvM) tests. The results of our simulations are presented in Tables 2-9 and Figure 2.

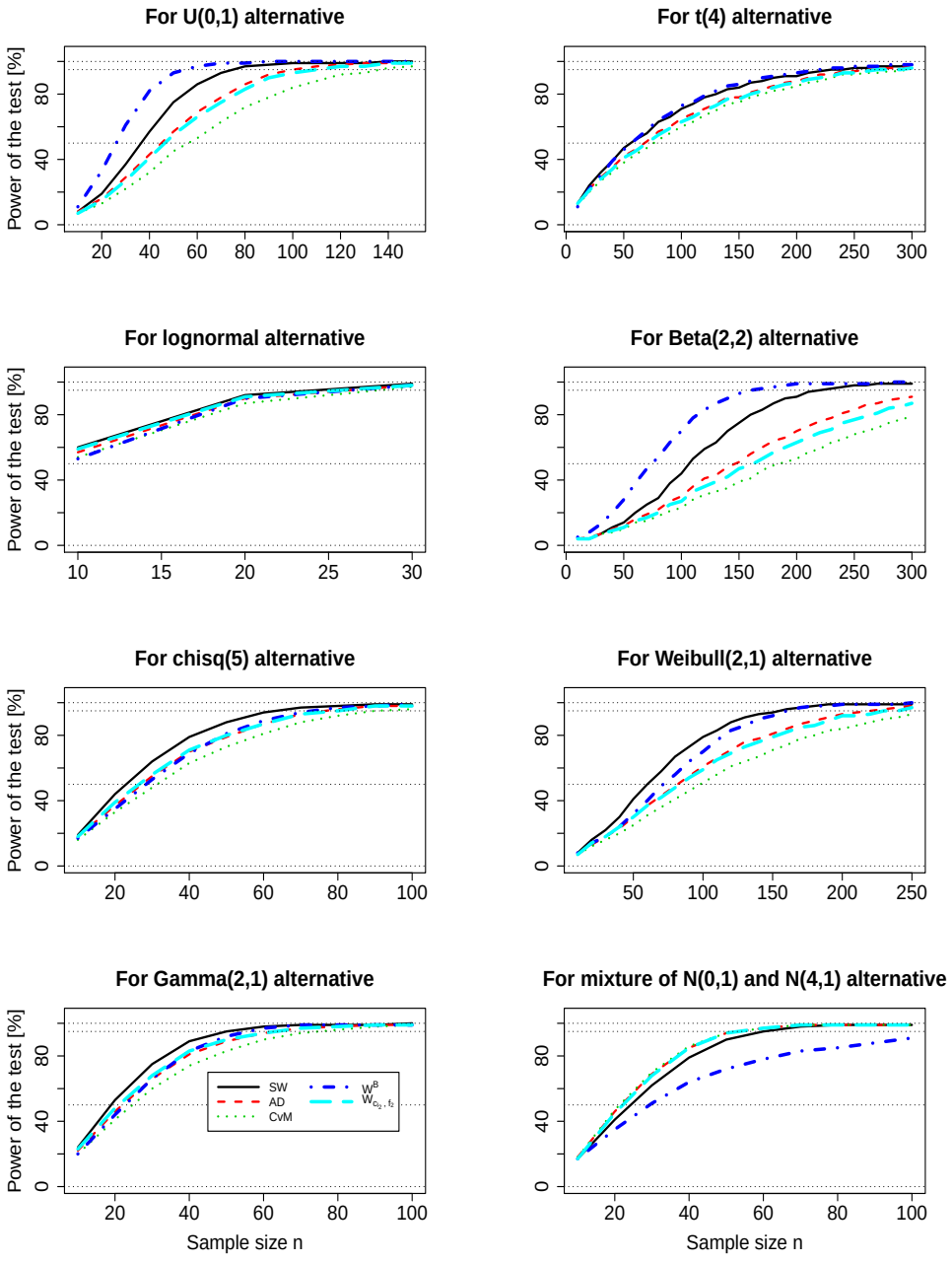

Figure 2. Power simulation against 8 different alternatives.

n	10	20	30	40	50	60	70	80	90	100	150
SW	8	19	37	57	75	86	93	97	98	99	**100**
AD	8	16	29	43	57	69	78	86	92	95	99
CvM	7	13	22	32	45	53	63	72	78	84	97
W^B	**11**	**33**	**61**	**82**	93	**97**	**99**	**99**	**100**	**100**	**100**
$W_{c\ell_2,f_2}$	7	15	27	41	55	66	75	83	90	93	99

Table 2. Power simulation against $U(0,1)$ alternative.

n	10	20	30	40	50	100	150
SW	**19**	**44**	**64**	**79**	**88**	**99**	**100**
AD	18	37	55	70	79	98	99
CvM	16	33	48	63	73	96	99
W^B	17	35	53	69	81	**99**	**100**
$W_{c\ell_2,f_2}$	18	39	56	71	80	98	99

Table 3. Power simulation against $\chi^2(5)$ alternative.

n	10	20	30	40	50	60	70	80	90	100	150	200	300	400
SW	**13**	**24**	**32**	**39**	**47**	**52**	56	63	66	71	84	91	**98**	**99**
AD	**13**	22	29	35	42	46	51	57	60	65	78	88	97	**99**
CvM	12	20	27	32	38	43	47	52	56	60	75	85	96	98
W^B	11	22	30	**39**	46	**52**	**58**	**64**	**68**	**73**	**86**	**93**	**98**	**99**
$W_{c\ell_2,f_2}$	13	21	28	34	41	45	50	55	59	63	77	87	96	**99**

Table 4. Power simulation against $t(4)$ alternative.

n	10	20	30	40
SW	**60**	**92**	**99**	**99**
AD	57	90	98	**99**
CvM	54	87	97	**99**
W^B	53	90	98	**99**
$W_{c\ell_2,f_2}$	59	91	98	**99**

Table 5. Power simulation against lognormal alternative.

n	10	20	30	40	50	60	70	100
SW	17	41	62	79	90	95	98	**99**
AD	**18**	46	69	85	**94**	**97**	**99**	**99**
CvM	**18**	**47**	**70**	**86**	**94**	**97**	**99**	**99**
W^B	17	35	51	64	72	78	83	91
$W_{c\ell_2,f_2}$	17	45	68	85	**94**	**97**	**99**	**99**

Table 6. Power simulation against the mixture of $N(0,1)$ and $N(4,1)$ alternative.

n	10	20	30	40	50	100	150	200	280
SW	4	4	7	11	14	44	75	91	99
AD	4	5	7	10	12	30	51	70	89
CvM	4	4	7	8	10	23	39	53	75
W^B	**5**	**8**	**13**	**20**	**28**	**70**	**93**	**99**	**100**
$W_{c\ell_2,f_2}$	4	4	7	9	11	27	47	63	84

Table 7. Power simulation against Beta(2,2) alternative.

It is worth notifying that the $W_{c\ell_2,f_2}$ test has almost the same power as the Anderson-Darling test (against all considered alternatives). Additionally, looking at Tables 2 and 7 we can conclude that the W^B test is much better than other tests against symmetric unimodal short-tailed distributions (in such cases it is apparently the most powerful for all sample sizes). In turn, for symmetric long-tailed distributions (see Table 4) the W^B test is comparable with the Shapiro-Wilk test and they and they are both the best in this case (W^B is a little bit worse for sample sizes smaller than 60 and a little bit better for larger ones).

Tables 3, 5, 8 and 9 show that the W^B test is slightly worse than the Shapiro-Wilk

test against asymmetric alternatives, but it is comparable to the other tests. We can notice here that differences are the smallest (practically 0) for long-tailed distribution (see Table 5), and the biggest for short-tailed ones (see Table 8). The W^B test is relatively the worst for two-modal symmetric alternative (see Table 6), however, in this case even the Shapiro-Wilk test is not the best one (the winners here are the Anderson-Darling, the Cramer-von Mises and the $W_{c_{\ell_2,f_2}}$ tests).

n	10	20	30	40	50	100	150	200	250
SW	**8**	**16**	**22**	**30**	**41**	**79**	**94**	**99**	99
AD	7	13	18	23	30	61	81	93	98
CvM	7	12	16	20	25	51	71	84	93
W^B	8	14	18	24	32	70	92	**99**	**100**
$W_{c_{\ell_2,f_2}}$	7	14	18	24	30	59	79	92	97

Table 8. Power simulation against Weibull(2,1) alternative.

n	10	20	30	40	50	100
SW	**24**	**53**	**75**	**89**	**95**	**100**
AD	22	46	66	81	89	99
CvM	20	41	60	74	83	99
W^B	20	44	66	83	92	99
$W_{c_{\ell_2,f_2}}$	23	48	68	83	90	99

Table 9. Power simulation against Gamma(2,1) alternative.

Summing up, the W^B test can be considered as comparable with the Shapiro-Wilk test, whereas the $W_{c_{\ell_2,f_2}}$ test as compatible with the Anderson-Darling test.

5. Conclusions and future work

In the paper we presented the construction of two goodness of fit tests for normality that are based on the optimal transport distance. We showed experimentally that they bring some advantages over the Shapiro-Wilk, the Anderson-Darling and the Cramer-von Mises tests. Specifically, the W^B test turned out to be apparently more powerful against symmetric short-tailed alternatives, while it was comparable or slightly worse in the other cases. As our future work, inspired by this paper and [4, 9], we consider application of the obtained test statistics (measures of nonnormality) in loss functions used to study autoencoder-based generative models.

6. References

[1] L. Baringhaus and N. Henze. Cramér-von mises distance: probabilistic interpretation, confidence intervals, and neighbourhood-of-model validation. *J. Nonparametr. Stat.*, 29:167–188, 2017.

[2] F. Bretz, T. Hothorn, and P. Westfall. *Multiple Comparisons Using R*. CRC Press, Boca Raton, 2010.

[3] E. del Barrio, J.A Cuesta-Albertos, C. Matrán, and Rodríguez-Rodríguez J.M. Tests of goodness of fit based on the l_2-wasserstein distance. *The Annals of Statistics*, 27:1230–1239, 1999.

[4] S. Kolouri, P.E. Pope, C.E. Martin, and G.K. Rohde. Sliced wasserstein auto-encoders. In *ICLR*, 2019.

[5] E. Krauczí. A study of the quantile correlation test for normality. *Test*, 18:156–165, 2009.

[6] G. Monge. *Mémoire sur la théorie des déblais et des remblais*. De lÍmprimerie Royale, Paris, 1781.

[7] A. Palmer, D. Dey, and J. Bi. Reforming generative autoencoders via goodness-of-fit hypothesis testing. In *UAI*, 2018.

[8] S.T. Rachev and L. Rüschendorf. *Mass Transportation Problems, Volume I: Theory*. Springer, New York, 1998.

[9] I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schoelkopf. Wasserstein auto-encoders. In *ICLR*, 2018.

[10] C. Villani. *Optimal Transport: Old and New*. Springer, Berlin, 2008.