One approach to the testing the distribution uniformity on the high-dimensional cube

Gennady Martynov

Institute for Information Transmission Problems
of the Russian Academy of Sciences
Moscow, Russia

Asymptotic Statistics of Stochastic Processes and Applications XI
19 July 2017

CLASSICAL CRAMÉR-von MISES MULTIVARIATE UNIFORMITY TEST

We shall use the notation $s = (s_1, ..., s_d)^\top$ and $t = (t_1, ..., t_d)^\top$ for d-vectors.
Let $U = (U_1, ..., U_d)^\top$ be a random vector with the uniform distribution function $F(t)$ on $[0,1]^d$, and let $U_i = (U_{i1}, ..., U_{id}),\ i = 1, ..., n$, are independent observations of U. The empirical distribution function of this sample has the form

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} 1_{U_i \le t}.$$

We can write the multivariate empirical process as

$$b_n(t) = n^{1/2}(F_n(t) - F(t)).$$

The Cramér-von Mises statistic for testing that F is the uniform cdf on $[0,1]^d$ is given by

$$\omega_n^2 = \int_{[0,1]^d} b_n^2(t)dt.$$

Its weighted variant is

$$\omega_n^2 = \int_{[0,1]^d} t^{2B} b_n^2(t)dt,$$

where $t^B = t_1^{\beta_1} \cdot ... \cdot t_d^{\beta_d}$, $B = (\beta_1, ..., \beta_d)$, $B > -1/2$.

$t^B\alpha_n(t)$ converges weakly in $L^2([0,1]^d)$ to $\xi(t) = t^B b(t)$, where $b(t)$ is the multivariate Brownian bridge with the covariance function

$$K_b(s,t) = E(b(s)b(t)) = \prod_{j=1}^{d}\{s_j \wedge t_j\} - \prod_{i=1}^{d}\{s_j t_j\}.$$

Using results from Krivjakova , Martynov, and Tjurin (1977), Deheuvels and Martynov (2003), we can conclude that the limit distribution of the statistic ω_n^2 in general case is the distribunion of the quadratic form

$$\omega^2 = \sum_{i=1}^{\infty} \alpha_i \chi^2_{i,k_i,\delta_i}.$$

It can be noted that the $K_b(s,t)$ tends to zero as d tends to ∞.

CRAMÉR–von MISES TEST FOR THE GAUSSIAN PROCESS IN HILBERT SPACE

One of the problems for the goodness-of-fit tests is the problem to test if an observed random process $S(t)$ on $[0,1]$ is the Gaussian process with zero mean and a covariance function $K_S(t,\tau),\ t,\tau \in [0,1]$,

$$\int_0^1 K_S(t,t)dt < \infty,\ t,\tau \in [0,1]. \tag{1}$$

The decision should be based on n observations $S_1(t), S_2(t), ..., S_n(t)$, $t \in [0,1]$ of the random process $S(t)$.

As a basis for $\mathcal{X}_{[0,1]}$ we choose the orthonormal basis formed by eigenfunctions $g_1(t), g_2(t), ...$ of the integral operator

$$\int_0^1 K_S(t,\tau)g(\tau)d\tau. \tag{2}$$

Realization $S_i(t)$ of $S(t)$ can be represented as $(s_{i1}, s_{i2}, s_{i3}, ...) \in \mathcal{X}_{[0,1]}$, where

$$s_{ij} = \int_0^1 S_i(t)g_j(t)dt. \tag{3}$$

The process $S(t)$ can be represented in the form of expansion in the mentioned basis as $s = (s_1, s_2, s_3, ...)$.

EXAMPLE: WIENER PROCESS

We will test the hypothesis H_0 that the observed random process on [0,1] is Gaussian with the zero mean and the covariance function

$$K_0(t,\tau) = \min(t,\tau).$$

This process (and all its observations) can be multiplied on

$$1/\sqrt{t}.$$

Resulting process has the unit variance. Its covariance function is

$$K(t,\tau) = \frac{\min(t,\tau)}{\sqrt{t\tau}}.$$

Corresponding covariance operator has the eigenvalues $\lambda_k = (z_{0,k}/2)^2$ and eigenfunctions

$$\varphi_k(t) = J_0(z_{0,k}t)/\sqrt{t}.$$

TEST FOR GAUSSIAN MEASURE IN HILBERT SPACE

The easiest way is to look at the problem in the general case. We will consider the probability space $(\mathcal{X},\ \mathcal{B},\ \nu)$ where $\mathcal{X}$ is a separable Hilbert space, $\mathcal{B}$ is the σ-algebra of Borel set on $\mathcal{X}$ and ν is a probability measure. Let we have n observations $X^1, X^2, ..., X^n$ of the random element X of $(\mathcal{X},\ \mathcal{B},\ \nu)$. We will test hypothesis

$$H_0 :\ \ \nu = \mu,$$

where μ is a Gaussian measure on $(\mathcal{X}, \mathcal{B})$ with a mean a and a covariance operator $K(z, w),\ z, w \in \mathcal{X}$. As a also $K(z, w)$ supposed be known. We can take $a = 0$.

Let $e = (e_1, e_2, ...)$ be the orthonormal basis of the eigenvectors of K and $\sigma_1^2, \sigma_2^2, ...$ be the eigenvalues of K. Let $x = (x_1, x_2, \ldots)$ be the representation of x in the basis e. Random element $X = (X_1, X_2, ...)$ has the independent components with the distributions $N(0, \sigma_i^2),\ i = 1, 2, ...$ In result, we can transform the probability space $(\mathcal{X},\ \mathcal{B},\ \nu)$ to a probability space

$$([0,1]^\infty,\ C^\infty,\ \Gamma).$$

Here C is the Borel set on $[0, 1]$ and Γ is the measure corresponding to ν.
Let Υ be the "uniform" measure on $([0, 1]^\infty,\ C^\infty)$. Now we will test the hypothesis

$$H0 :\ \Gamma = \Upsilon. \tag{4}$$

For application of the Cramér-von Mises-type test for testing the hypothesis (4), it need to introduce the function F on $[0,1]^\infty$ such, that it defines the measure Υ. In the finite dimensional case, as a variant of F can be chosen the obvious distribution function. This function must be nonzero for all points $t = (t_1, t_2, ...)$ in $[0,1]^\infty$, with of exception of a set of measure zero. The trasformed random variable X is defined now as $T = (T_1, T_2, ...)$.

The convenient example is

$$F(t) = P\{T_1 \le t_1^{r_1}, T_2 \le t_2^{r_2}, T_3 \le t_3^{r_3} ...\} \\ = t_1^{r_1} t_2^{r_2} t_3^{r_3} ..., \tag{5}$$

when r_i tends to zero sufficiently rapidly. Let $T^{(i)} = (T_{i1}, T_{i2}, ...)$ be the observations of T. The empirical function is

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{\infty} 1_{T_{i1} \le t_1^{r_1},\ T_{i2} \le t_2^{r_2}, \cdots}. \tag{6}$$

With this distribution function $F(t),\ t \in [0,1]^{\infty}$, the measure μ can be restored. The Cramér-von-Mises statistics is

$$\omega_n^2 = n \int_{[0,1]^{\infty}} \left(F_n(t) - \prod_{i=1}^{\infty} t_i^{r_i} \right)^2 dt_1 dt_2 ....,$$

The "empirical process"

$$\xi_n(t) = \sqrt{n}\left(F_n(t) - \prod_{i=1}^{\infty} t_i^{r_i}\right), \quad t \in [0,1]^{\infty},$$

converges weakly in $L_2(\mathcal{X})$ to the Gaussian process with the covariance function

$$K(s,t) = \prod_{i=1}^{\infty} \min(s_i^{r_i}, t_i^{r_i}) - \prod_{i=1}^{\infty} s_i^{r_i} t_i^{r_i}, \; s,t \in [0,1]^{\infty}.$$

This assertion follows from the representation $\xi_n(t)$ by the sum

$$\xi_n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\prod_{i=1}^{\infty} t_i^{r_i} - I_{T_i < t_i^{r_i}}\right), \quad t \in [0,1]^{\infty}$$

of i.i.d. random functions and, for example, from Vaart and Wellner (1996)

If, for example,

$$r_i = i^{-v}, \quad v > 1,$$

then the condition of the weak convergence

$$\int_{[0,1]^\infty} K(s,s)ds$$

$$= \prod_{i=1}^{\infty} \frac{1}{r_i + 1} - \prod_{i=1}^{\infty} \frac{1}{2r_i + 1} < \infty$$

is fulfilled. As $i \to \infty$ $r_i \downarrow 0$. It can be written

$$K(s,t) = K_0(s,t) - w(s)w(t),$$

where

$$K_0(s,t) = \prod_{i=1}^{\infty} \min(s_i^{r_i}, t_i^{r_i}) = \prod_{i=1}^{\infty} K_{0i}(s_i^{r_i}, t_i^{r_i}),$$

$$w(s) = \prod_{i=1}^{\infty} s_i^{r_i}, \quad s, t \in [0,1]^\infty.$$

EIGENVALUES AND EIGENFUNCTIONS FOR $R_{0i}(s_i^{r_i}, t_i^{r_i})$

Now, we can consider the elementary kernel $r_i(t_i, s_i) = \min(t_i^{r_i}, s_i^{r_i})$, $i = 1, \ldots$. Its eigenfunctions and eigenvalues can be found from the integral equation

$$\varphi(t) = \lambda \int_0^1 \min(t^{r_i}, \tau^{r_i})\varphi(\tau)d\tau, \tau \in [0, 1].$$

Let

$$z(t) = \int_t^1 \varphi(\tau)d\tau.$$

The corresponding differential equation is

$$z''(t) + \lambda r_i t^{r_i - 1} z(t) = 0,\ z(1) = z'(0) = 0.$$

Let $\mu_i = r_i/(1+r_i)$. Then the eigenvalues of R_{0i} are

$$\lambda_{r_i,k} = r_i \left(\frac{z_{\mu_i - 1,k}}{2\mu_i} \right)^2, \quad k = 1, 2, \ldots . \tag{7}$$

In our study, r_i varies from 1 to zero while μ_i varies from 1/2 to zero.

Corresponding the non-normalized eigenfunctions are

$$\widetilde{\varphi}_{r_i,k}(t) = t^{r_i/2} J_{\mu_i}\left(z_{\mu_i-1,k} t^{(1+r_i/2)}\right), \quad k = 1, 2, ...,$$

The squared normalizing divisor for $\widetilde{\varphi}_{r_i,k}(t)$ is

$$\begin{aligned} D^2_{r_i,k} &= \frac{(\mu_i - 1) z^{2\mu_i}_{\mu_i-1,k} \Gamma(2+\mu_i)\Gamma(1+2\mu_i)}{\sqrt{\pi}} \\ &\quad \times {}_1F_2\left(\frac{1}{2} + \mu_i; 2+\mu_i, 1+2\mu_i; -z^2_{\mu_i-1,k}\right). \end{aligned}$$

EIGENVALUES AND EIGENFUNCTIONS OF K_0

Further, we obtain the eigenvalues and eigenfunctions of the kernel $K_0^*(t,\tau)$, $t,\tau \in [0,1]^\infty$. In this case, the Fredholm equation becomes

$$\varphi_r(t) = \lambda \int_{[0,1]^\infty} \prod_{j=1}^{\infty} \min(t_j^{r_j}, \tau_j^{r_j}) \varphi_r(\tau) d\tau, \quad t,\tau \in [0,1]^\infty. \tag{8}$$

Eigenvalues and eigenfunctions of K_0 can be represented as the product of all possible combinations of eigenvalues and eigenfunctions K_{0i} taken one by one for each $i = 1, 2, \ldots$ Denote
$\alpha_{j,k} = 1/\lambda_{r_j,k}$, $k = 1, 2, 3, \ldots$.

We have $\lambda_{r_j,1} \to 1$, but $\lambda_{r_j,k} \to \infty$ $(\alpha_{j,k} \to 0)$, $k = 2, 3, ...,$ as $j \to \infty$. The behavior of the characteristic numbers when r_j tends to zero is shown in Table. Here and below, we take $r_i = i^{-a(1-i^{-b})}$, $a = 3$, $b = 0.5$.

Table: The behavior of $\alpha_{r_j,k}$ when r_i varies from one to zero.

j	r_j	$\alpha_{j,1}$	$\alpha_{j,2}$	$\alpha_{j,3}$	$\alpha_{j,4}$	$\alpha_{j,5}$	...
1	1	0.4053	0.0450	0.0162	0.0083	0.0050	...
2	0.60197	0.5298	0.0463	0.0160	0.0081	0.0048	...
3	0.31323	0.6829	0.0400	0.0132	0.0065	0.0039	...
4	0.17678	0.7917	0.0303	0.0097	0.0047	0.0028	...
5	0.10816	0.8610	0.0219	0.0068	0.0033	0.0019	...
10	0.01952	0.9716	0.0050	0.0015	0.0007	0.0004	...
25	0.00160	0.9976	0.0004	0.0001	0.0001	0.0000	...
50	0.00023	0.9997	0.0001	0.0000	0.0000	0.0000	...
99	0.00003	1.0000	0.0000	0.0000	0.0000	0.0000	...
∞	0	1	0	0	0	0	...

We can compute the largest characteristic number. It is equal to

$$\alpha_1 = \prod_{j=1}^{\infty} \alpha_{j,1} \approx 0.0642. \tag{9}$$

It is also obvious that the second largest characteristic number is

$$\alpha_2 = \alpha_1 \alpha_{1,2} / \alpha_{1,1} \approx 0.007137. \tag{10}$$

The corresponding eigenfunctions are

$$\varphi_1(t) = \prod_{j=1}^{\infty} \varphi_{r_j,1}(t) \quad \text{and} \quad \varphi_2(t) = \varphi_1(t)\varphi_{r_1,2}(t)/\varphi_{r_1,1}(t).$$

EIGENVALUES OF K^*

The characteristic numbers α_j^* of $K^*(s,t)$ are solutions of the equation

$$\sum_{k=1}^{\infty} \frac{C_k^2}{\alpha_k - \alpha} = 1, \; C_k = \prod_{j=1}^{\infty} C_{j,k}, \; k = 1, 2, ...$$

and

$$C_{j,k} = \int_0^1 w_j(t)\varphi_{j,k}(t)dt = \int_0^1 t^{\mu_j/1-\mu_j}\varphi_{j,k}(t)dt, \; j = 1, 2, \dots,$$

or

$$C_{r,k}^2 = \left(2^{-\mu}(\mu-1)\Gamma(2+\mu)z_{\mu-1,k}^{\mu}\, {}_0F_1\left(;2+\mu;-\frac{1}{4}z_{\mu-1,k}^2\right)\right)^2 / D_{r,k}^2,$$

where ${}_0F_1$ is the *generalized hypergeometric function*

$${}_0F_1(;a;z) = \sum_{k=0}^{\infty} \frac{z^k}{(a)_k k!}.$$

LIMIT DISTRIBUTION OF ω^2

In the previous section, we obtained the eigenvalues $\lambda_k^* = 1/\alpha_k^*$ for the covariance function $K^*(s,t)$. For calculation of the limiting distribution function of Ω^2 we can use the *Smirnov formula*, namely, for $t > 0$,

$$P\left(\omega^2 > t\right) = \frac{1}{\pi}\sum_{k=1}^{\infty}(-1)^{k+1}\int_{\lambda_{2k-1}^*}^{\lambda_{2k}^*}\frac{e^{-tu/2}du}{u\sqrt{\left|\prod_{k=1}^{\infty}\left\{1-\frac{u}{\lambda_k^*}\right\}\right|}}.$$

The Smirnov formula is designed for distinct eigenvalues.

QUADRATIC FORM DISTRIBUTION WITH MULTIPLE EIGENVALUES

$$F(x) = 1 - \frac{1}{\pi} \arctan \frac{1}{a}$$

$$- \frac{1}{\pi} \int_0^{\infty} \frac{\exp(-atx/2) \sin(\zeta(t) - tx/2)}{t\beta(t)} dt,$$

where

$$\zeta(t) = \sum_{k=1}^{\infty} \left\{ \frac{s_k}{2} w_k(t) + \frac{\mu_k \delta_k^2 t}{2[(\mu_k - at)^2 + t^2]} \right\},$$

$$\beta(t) = \prod_{k=1}^{\infty} \left\{ \left[\left(1 - \frac{at}{\mu_k} \right)^2 + \frac{t^2}{\mu_k^2} \right]^{s_k/4} \right\}$$

$$\exp \left\{ - \frac{\delta_k^2 t[a\mu_k - (a^2+1)t]}{2[(\mu_k - at)^2 + t^2]} \right\},$$

$$w_k(t) = \operatorname{arcctg} \frac{\mu_k - at}{t}, \quad 0 \le w_k(t) < \pi.$$

NUMERICAL RESULTS

The Cramér-von Mises statistic can be represented as

$$\omega_n^2 = n \int_{[0,1]^\infty} \left(\frac{1}{n} \sum_{i=1}^{n} \prod_{j=1}^{\infty} I_{T_{i,j} < t_j^{r_i}} - \prod_{i=1}^{\infty} t_i^{r_i} \right)^2 dt.$$

The limit distribution of the Cramér-von Mises statistic can be calculated by the methods described above for multidimensional case. The statistic ω_n^2 can be calculated by the Monte-Carlo method. In turn, the distribution of the statistic was calculated also using the Monte Carlo method.

QUANTILES of ω^2

We will present the estimated quantiles of the distribution ω_n^2 with $r_i = i^{-a(1-i^{-b})}$, $a = 3$, $b = 0.5$.

Dimension	quantile 0.9	quantile 0.95		
1	0.3473	0.4614	0.7435	0.8694
5	0.214	0.270		
10	0.180	0.222		
15	0.172	0.212		
25	0.169	0.209		
30	0.168	0.209		
40	0.166	0.206		
75	0.167	0.205		
∞	0.16450	0.20371	0.30039	0.34350

BIBLIOGRAPHY

1. Deheuvels, P., Martynov, G. (2003) Karhunen-Loève expansions for weighted Wiener processes and Brownian bridges via Bessel functions., *Progress in Probability, Birkhäuser, Basel/Switzerland*, **55**, 57–93.
2. Darling, D. A. (1955). The Cramér-Smirnov test in the parametric case. *Ann. Math. Statist.* **26** 1–20.
3. Kac, M., Kiefer, J., Wolfowitz, J. (1955) On tests of normality and other tests of goodness-of-fit based on distance methods. *Ann. Math. Statist.*, **30**, 420–447.
4. Martynov, G. V. (1975). Computation of distribution function of quadratic forms of normally distributed random variables. *Theor. Probab. Appl.* **20** 782–793.
5. Krivjakova E. N., Martynov G. V., and Tjurin J. N. (1977) The distribution of the omega square statistics in the multivariate case. *"Theory of Probability and its Applications."*, **V. 22**, N 2, 415–420.
6. Martynov, G. V., (1979) The omega square tests, *Nauka, Moscow*, 80pp., (in Russian).
7. Martynov, G. V., (1992) Statistical tests based on empirical processes and related questions, *J. Soviet. Math*, **61**, 2195–2271.
8. Van der Vaart, A. W., Wellner, J. A. (1996) Weak Converge and Empirical Processes, Springer-Verlag, 508 pp.