Multivariate model assessment without χ^2

Sara Algeri
salgeri@umn.edu

School of Statistics, University of Minnesota

PHYSTAT 2-Samples,
Virtual meeting.

June 2, 2023.

To truly understand what are the main objects involved in our analysis, we begin with the one-dimensional setting...

Goodness-of-fit vs 2-Samples tests

Goodness-of-fit (GOF)

- Inputs: A sample $x_1, \ldots, x_n$ from P and a postulated distribution Q
- Test:

$$H_0 : P = Q \quad \text{vs} \quad H_1 : P \neq Q$$

- Test statistics: Functionals of the empirical process

$$v_n(x) = \sqrt{n}\left[\widehat{P}(x) - Q(x)\right]$$

with $\widehat{P}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{x_i \leq x\}}$.

Note: we can incorporate unknown parameters through $Q(x)$.

2-Samples

- Inputs: Samples $x_{11}, \ldots, x_{1n_1}$ from P_1 and $x_{21}, \ldots, x_{2n_2}$ from P_2
- Test:

$$H_0 : P_1 = P_2 = P \quad \text{vs} \quad H_1 : P_1 \neq P_2$$

- Test statistics: Functionals of the empirical process

$$v_n(x) = \sqrt{\frac{n_1 n}{n_2}}\left[\widehat{P}_1(x) - \widehat{P}(x)\right]$$

with $\widehat{P}(x) = \frac{n_1}{n}\widehat{P}_1(x) + \frac{n_2}{n}\widehat{P}_2(x)$

$\widehat{P}_j(x) = \frac{1}{n_j}\sum_{i=1}^{n_j} \mathbb{1}_{\{x_{ji} \leq x\}}$, $j = 1, 2$,

and $n = n_1 + n_2$.

A very famous example

Recall that our empirical process specifies as

- $v_n(x) = \sqrt{n}\left[\widehat{P}(x) - Q(x)\right]$, for GOF problems.
- $v_n(x) = \sqrt{\frac{n_1 n}{n_2}}\left[\widehat{P}_1(x) - \widehat{P}(x)\right]$, for 2-Samples problems.

Then the Kolmogorov-Smirnov statistic specifies as

$$\boxed{KS = \sup_x |v_n(x)|.} \tag{1}$$

In the one-dimensional setting and if, in the context of GOF, Q does not depend on unknown parameters, KS is **asymptotically distribution-free**.

Distribution-freeness in GOF

We have *distribution-freeness* whenever the null distribution of the test statistic considered does depend on the distribution Q being tested.

Another desirable property of GOF

- Let's keep in mind that distribution-freeness is not all we need.
- We also want sensitivity (power) against "all" alternatives.

Note

The latter is essentially what differentiates GOF tests from tests of hypotheses (e.g., Neyman-Pearson) where the power is concentrated towards the specific alternative model specified under H_1.

But what does "all" actually mean?

...there exist alternatives that cannot be detected even by Neyman-Pearson, so there is no hope we can detect those via GOF.

Power against (converging) contiguous alternatives

A more sensible criterion...

We require our GOF test to have some power against all (converging) **contiguous** alternatives.

What is a (converging) contiguous alternative?

Heuristically...

- They are alternatives that get progressively closer to the null as the sample size increases.
- They are detectable via Neyman-Pearson*.
- In the limit, we can identify the direction from which they approach the null.

*See "Oosterhoff J., van Zwet W.R. A note on contiguity and Hellinger distance. *Springer New York*, 2012."

...a little more formally

Converging contiguous alternatives

Let Q be the null distribution postulated for the underlying continuous data generating process, and let q be its density. We say that the distribution $\widetilde{P}_n$ is a contiguous alternative to Q if its density can be specified as

$$\widetilde{p}_n(x) = q(x)\left[1 + \frac{h_n(x)}{\sqrt{n}}\right],$$

with $\|\, h_n(x) \,\|_Q^2 < \infty$ and $\|\, h_n(x) - h \,\|_Q^2 \rightarrow 0$. This last condition is what makes them "converging" and the function h which corresponds to the direction from which $\widetilde{p}_n$ approaches q.

An important note on data binning

Can't we just bin the data and rely Pearson?

In multidimensional settings and/or when Q depends on unknown parameters, it is common practice to bin the data and use Pearson X^2 (or asymptotic equivalent) to perform GOF.

Warning

It can be shown* that Pearson (and many other similar statistics) have no power against infinitely many converging contiguous alternatives

*See "Algeri S. and Khmaladze E.V. When Pearson χ^2 and other divisible statistics are not goodness-of-fit tests. *In preparation*."

First, let's extend what we known for 1D...

1D

- Sample: $x_1, \dots, x_n$. Each observation x_i is a scalar.
- Distribution function under H_0 evaluated at x:

$$Q(x) = P(X \leq x | H_0)$$

- Empirical distribution function under H_0 evaluated at x:

$$\widehat{P}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{x_i \leq x\}}$$

- Empirical process:

$$v_n(x) = \sqrt{n}\left[\widehat{P}(x) - Q(x)\right]$$

multi-D

- Sample: $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$. Each observation $\boldsymbol{x}_i = (x_{1i}, \dots, x_{Di})$ is a vector.
- Distribution function under H_0 evaluated at $\boldsymbol{x} = (x_1, \dots, x_D)$:

$$Q(\boldsymbol{x}) = P(X_1 \leq x_1, \dots, X_D \leq x_D | H_0)$$

- Empirical distribution function under H_0 evaluated at $\boldsymbol{x}$:

$$\widehat{P}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{x_{1i} \leq x_1, \dots, x_{Di} \leq x_D\}}$$

- Empirical process:

$$v_n(\boldsymbol{x}) = \sqrt{n}\left[\widehat{P}(\boldsymbol{x}) - Q(\boldsymbol{x})\right]$$

...and then let's get parametric...

Given a set of n observations from an unknown cumulative distribution function (cdf) $P(\boldsymbol{x}) = P(X_1 \leq x_1, \dots, X_d \leq x_d)$, we are interested in testing

$$H_0 : P(\boldsymbol{x}) = Q(\boldsymbol{x}, \boldsymbol{\theta}) \quad \text{versus} \quad H_1 : P(\boldsymbol{x}) \neq Q(\boldsymbol{x}, \boldsymbol{\theta})$$

for some postulated distribution $Q(\boldsymbol{x}, \boldsymbol{\theta})$. To perform the test above, we consider the *parametric empirical process* $v_Q(\boldsymbol{x}, \boldsymbol{\theta})$

$$\boxed{v_Q(\boldsymbol{x}, \boldsymbol{\theta}) = \sqrt{n}\left[\widehat{P}(\boldsymbol{x}) - Q(\boldsymbol{x}, \boldsymbol{\theta})\right]} \tag{2}$$

...we now have access to an entire family of tests!

Recall that our empirical process is $v_n(\boldsymbol{x},\boldsymbol{\theta}) = \sqrt{n}\big[P_n(\boldsymbol{x}) - Q(\boldsymbol{x},\boldsymbol{\theta})\big]$ and let $\widehat{\boldsymbol{\theta}}$ be the Maximum Likelihood Estimate (MLE) of $\boldsymbol{\theta} \Rightarrow$ we can construct tests statistics as functionals of $v_n(\boldsymbol{x},\widehat{\boldsymbol{\theta}})$. E.g.,

- Kolmogorov-Smirnov statistic: $KS = \sup_{\boldsymbol{x}} v_n(\boldsymbol{x},\widehat{\boldsymbol{\theta}})$.
- Cramer-von Mises statistic: $CvM = \int |v_n(\boldsymbol{x},\widehat{\boldsymbol{\theta}}))|^2 dQ(\boldsymbol{x},\widehat{\boldsymbol{\theta}}))$.
- Anderson-Darling statistic: $AD = \int \left| \frac{v_n(\boldsymbol{x},\widehat{\boldsymbol{\theta}})}{\sqrt{Q(\boldsymbol{x},\widehat{\boldsymbol{\theta}}))(1-Q(\boldsymbol{x},\widehat{\boldsymbol{\theta}}))}} \right|^2 dQ(\boldsymbol{x},\widehat{\boldsymbol{\theta}}))$.

(where all the integrals are multivariate).
They are not distribution-free but we can simulate their distribution via the parametric bootstrap.

Cons: Computational complexity may be high $+$ simulations must be repeated on a case-by-case basis.

$$\Downarrow$$

In the remaining of the talk we will see two approaches which will help us to overcome these two limitations.

Estimated and projected empirical process

We can approximate our estimated empirical process via

$$\underbrace{v_Q(\boldsymbol{x}, \widehat{\boldsymbol{\theta}})}_{\substack{\text{Empirical}\\\text{process}\\\text{at } \widehat{\boldsymbol{\theta}}}} = \underbrace{v_Q(\boldsymbol{x}, \boldsymbol{\theta})}_{\substack{\text{Empirical}\\\text{process}\\\text{at } \boldsymbol{\theta}}} - \sum_{j=1}^{p} \int_{[-\infty, \boldsymbol{x}]} \underbrace{b_j(\boldsymbol{t}, \boldsymbol{\theta})}_{\substack{j\text{-th normalized}\\\text{score}\\\text{function}}} \mathrm{d}\boldsymbol{t} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} b_j(\boldsymbol{x}_i, \boldsymbol{\theta}) + o_p(1)$$

where $[-\infty, \boldsymbol{x}] = [-\infty, x_1] \times \cdots \times [-\infty, x_D]$

- We denote the right hand side with $\widetilde{v}_Q(\boldsymbol{x}, \boldsymbol{\theta})$. It is a projection of $v_Q(\boldsymbol{x}, \boldsymbol{\theta})$ orthogonal to the normalized score functions $b_j(\boldsymbol{x}, \boldsymbol{\theta})$, i.e., the components of

$$b(x, \boldsymbol{\theta}) = \underbrace{\Gamma_{\boldsymbol{\theta}}^{-1/2}}_{\substack{\text{Inverse sqrt of}\\\text{the Fisher info}}} \underbrace{\frac{\partial}{\partial \boldsymbol{\theta}} \log q(\boldsymbol{x}, \boldsymbol{\theta})}_{\substack{\text{Score}\\\text{vector}}}. \quad (3)$$

- The *projected empirical process** **does not depend on $\widehat{\theta}$!**

* See "Khmaladze, E.V. (1980). The use of ω^2 tests for testing parametric hypotheses. *Theory of Probability & Its Applications*."

A toy example to assess the computational gain

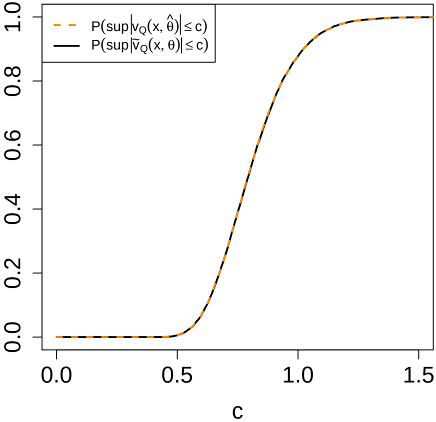
We draw a sample of $n = 100$ observations from

$$q(\boldsymbol{x}, \boldsymbol{\theta}) \propto e^{-\frac{1}{2\boldsymbol{\theta}_3}\left[(x_1-\boldsymbol{\theta}_1)^2+(x_2-\boldsymbol{\theta}_2)^2\right]} \quad \boldsymbol{x} \in [1,20] \times [1,25], \tag{4}$$

$\boldsymbol{\theta} = (-2, 5, 25)$ and its MLE is $\widehat{\boldsymbol{\theta}}_{obs} = (-0.77, 6.32, 22.02)$.
We proceed by simulating the distribution of the KS statistic, i.e.,

1. We simulate $\sup_{\boldsymbol{x}} |v_Q(\boldsymbol{x}, \widehat{\boldsymbol{\theta}})|$ by sampling from $Q(\boldsymbol{x}, \widehat{\boldsymbol{\theta}}_{obs})$ via the parametric bootstrap.
2. We simulate $\sup_{\boldsymbol{x}} |\widetilde{v}_Q(\boldsymbol{x}, \boldsymbol{\theta})|$ by sampling from $Q(\boldsymbol{x}, \widehat{\boldsymbol{\theta}}_{obs})$ via the parametric bootstrap.

Simulated distributions of the KS statistic



	$\sup_x \lvert\widetilde{v}_Q(\boldsymbol{x},\boldsymbol{\theta})\rvert$	$\sup_x \lvert v_Q(\boldsymbol{x},\widehat{\boldsymbol{\theta}})\rvert$
CPU time	9.429 mins	12.198 hrs

B=10,000 n=100 R=2000

But what if we want to test another model, $F(\boldsymbol{x}, \beta)$ for which all of this is not at all feasible?

(Can we somehow retrieve distribution-freeness?)

A useful (re-)formulation

We rewrite our empirical process as,

$$\widetilde{v}_Q(\boldsymbol{x},\boldsymbol{\theta}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{ \left[\mathbb{1}_{\{x_{1i}\leq x_1,\ldots,x_{Di}\leq x_D\}} - Q(\boldsymbol{x},\boldsymbol{\theta})\right] - b^T(x_i,\boldsymbol{\theta})\int_{[-\infty,\boldsymbol{x}]} b(\boldsymbol{t},\boldsymbol{\theta})\mathrm{d}\boldsymbol{t} \right\}$$

Setting everything in the curly brackets equal to $\psi_{\boldsymbol{x}}(\boldsymbol{x}_i,\boldsymbol{\theta})$, we have

$$\widetilde{v}_Q(\boldsymbol{x},\boldsymbol{\theta}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{\boldsymbol{x}}(\boldsymbol{x}_i,\boldsymbol{\theta}). \tag{5}$$

- One can show that the limit of $\widetilde{v}_Q(\boldsymbol{x},\boldsymbol{\theta})$ is Gaussian.
- its mean and covariance are $E_Q\left[\psi_{\boldsymbol{x}}\right]=0$ and $E_Q\left[\psi_{\boldsymbol{x}}\psi_{\boldsymbol{x}'}\right]$

$\Rightarrow$ **the $\psi_{\boldsymbol{x}}$ fully characterize the limiting distribution of $\widetilde{v}_Q(\boldsymbol{x},\boldsymbol{\theta})$.**

Towards (asymptotic) distribution-freeness

Can we construct another process whose limit, under $F(x,\beta)$, will be the same as that of $\widetilde{v}_Q(x,\theta)$ under Q?

The key here is to "play" with our $\psi_{\boldsymbol{x}}(\boldsymbol{x}_i,\boldsymbol{\theta})$ functions so that, by taking a suitable transformation of them, we can construct a new process that, under F, will have the same limiting distribution as $\widetilde{v}_Q(\boldsymbol{x},\boldsymbol{\theta})$, under Q.

This can be done by means of the Khmaladze-2 (K-2) transform*.

*See "Khmaladze, E.V. (2016). Unitary transformations, empirical processes and distribution free testing. *Bernoulli*, 2016."

The K-2 transform in a nutshell

The K-2 transform applied to the functions $\psi_x(x_i, \boldsymbol{\theta})$ is

$$\phi_{\mathbf{x}}(\boldsymbol{x}_i, \boldsymbol{\theta}, \boldsymbol{\beta}) = \underbrace{\boldsymbol{U}\left[K\left[l_{\boldsymbol{\theta},\boldsymbol{\beta}}(\boldsymbol{x}_i)\right.\right.}_{\text{K-2 transform}} \left.\left.\psi_{\mathbf{x}}(\boldsymbol{x}_i, \boldsymbol{\theta})\right]\right]$$

- The isometry $l_{\boldsymbol{\theta},\boldsymbol{\beta}}(\boldsymbol{x}) = \sqrt{\frac{q(\boldsymbol{x},\boldsymbol{\theta})}{f(\boldsymbol{x},\boldsymbol{\beta})}}$ ensures $E_F\left[(l_{\boldsymbol{\theta},\boldsymbol{\beta}}\psi_{\mathbf{x}})(l_{\boldsymbol{\theta},\boldsymbol{\beta}}\psi_{\mathbf{x}'})\right] = E_Q\left[\psi_{\mathbf{x}}\psi_{\mathbf{x}'}\right]$.
- The unitary operator K ensures that $E_F\left[K l_{\boldsymbol{\theta},\boldsymbol{\beta}}\psi_{\mathbf{x}}\right] = E_Q\left[\psi_{\mathbf{x}}\right] = 0$.
- The unitary operator $\boldsymbol{U}$ ensures orthogonality w.r.t. to the normalized score functions under $F(\boldsymbol{x}, \boldsymbol{\beta})$.

A new family of test statistics

Recall that

$$\boxed{\widetilde{v}_F(\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{\beta})=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_{\boldsymbol{x}}(\boldsymbol{x}_i,\boldsymbol{\theta},\boldsymbol{\beta})} \quad \text{and} \quad \boxed{\widetilde{v}_Q(\boldsymbol{x},\boldsymbol{\theta})=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{\boldsymbol{x}}(\boldsymbol{x}_i,\boldsymbol{\theta})}$$

We can now construct our *K-2 transformed* test statistics as

$$\begin{aligned}\mathrm{KS}_{F|Q}&=\sup_{\boldsymbol{x}}\left|\,\widetilde{v}_F(\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{\beta})\,\right|,\quad \mathrm{CvM}_{F|Q}=\int \widetilde{v}_F^2(\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{\beta})\,\mathrm{d}Q(\boldsymbol{x},\boldsymbol{\theta}),\\ \text{and}\quad \mathrm{AD}_{F|Q}&=\int\frac{\widetilde{v}_F^2(\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{\beta})}{Q(\boldsymbol{x},\boldsymbol{\theta})[1-Q(\boldsymbol{x},\boldsymbol{\theta})]}\mathrm{d}Q(\boldsymbol{x},\boldsymbol{\theta}),\end{aligned} \tag{6}$$

which have the same limiting distribution as

$$\begin{aligned}\mathrm{KS}_{Q}&=\sup_{\boldsymbol{x}}\left|\,\widetilde{v}_Q(\boldsymbol{x},\boldsymbol{\theta})\,\right|,\quad \mathrm{CvM}_{Q}=\int \widetilde{v}_Q^2(\boldsymbol{x},\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{x},\boldsymbol{\theta}),\\ \text{and}\quad \mathrm{AD}_{Q}&=\int\frac{\widetilde{v}_Q^2(\boldsymbol{x},\boldsymbol{\theta})}{Q(\boldsymbol{x},\boldsymbol{\theta})[1-Q(\boldsymbol{x},\boldsymbol{\theta})]}\mathrm{d}Q(\boldsymbol{x},\boldsymbol{\theta}),\end{aligned} \tag{7}$$

Requirements on F and Q

Can we use any $F(\boldsymbol{x}, \beta)$ and any $Q(\boldsymbol{x}, \theta)$?

- Let $f(\boldsymbol{x}, \boldsymbol{\beta})$ and $q(\boldsymbol{x}, \boldsymbol{\theta})$ be the densities of $F(\boldsymbol{x}, \boldsymbol{\beta})$ and $Q(\boldsymbol{x}, \boldsymbol{\theta})$. We require that:
  - $f(\boldsymbol{x}, \boldsymbol{\beta}) = 0$ iff $q(\boldsymbol{x}, \boldsymbol{\theta}) = 0$ (they have the same support).
  - $\boldsymbol{\theta}$, $\boldsymbol{\beta}$ are both of size p (the have the same size).
- **These are rather general criteria!** $\Rightarrow$ $Q(\boldsymbol{x}, \boldsymbol{\theta})$ can be chosen to be arbitrarily simple to ease the computations.
- We call $Q(\boldsymbol{x}, \boldsymbol{\theta})$ "*reference distribution*" because, for any $F_1, \dots, F_M$ satisfying these criteria, we can construct a process $\widetilde{v}_{F_m}$, $m = 1, \dots, M$ with the same distribution as $\widetilde{v}_Q$.

An illustrative example

- **Data:** a sample of $n = 100$ observations generated from

$$p(\boldsymbol{x}) \propto (2\pi)^{-1}|\boldsymbol{\Sigma}|^{-1/2}\left[1 + (\boldsymbol{x} - \mu)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \mu)\right]^{-3/2}, \qquad (8)$$

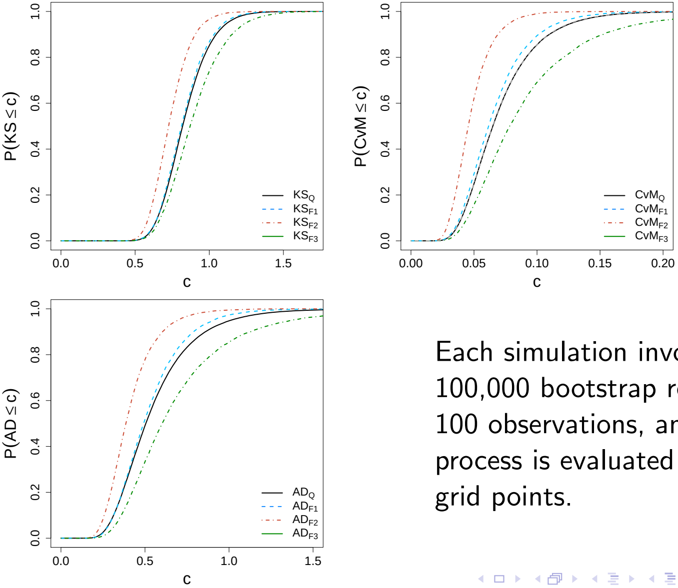
where $\mu = (0,3)^T$, $\boldsymbol{\Sigma} = \begin{bmatrix} 20 & 10 \\ 10 & 20 \end{bmatrix}$, $\boldsymbol{x} \in [1,20] \times [1,25]$.

- **Null models** we aim to test:

$$\begin{aligned} f_1(x;\boldsymbol{\beta}) &\propto x_1^{(\beta_1-1)}x_2^{(\beta_2-1)}\exp\left\{-\beta_3(x_1+x_2)\right\}, \\ f_2(x;\boldsymbol{\beta}) &\propto \frac{\beta_3}{2\pi}[(x_1-\beta_1)^2 + (x_2-\beta_2)^2 + \beta_3]^{-3/2}, \\ f_3(x;\boldsymbol{\beta}) &\propto e^{-\frac{1}{200}\left[\left(\frac{x_1}{\beta_1}-1\right)^2+\left(\frac{x_2}{\beta_2}-1\right)^2-\beta_3\left(\frac{x_1}{\beta_1}-1\right)\left(\frac{x_2}{\beta_2}-1\right)\right]}, \end{aligned} \qquad (9)$$

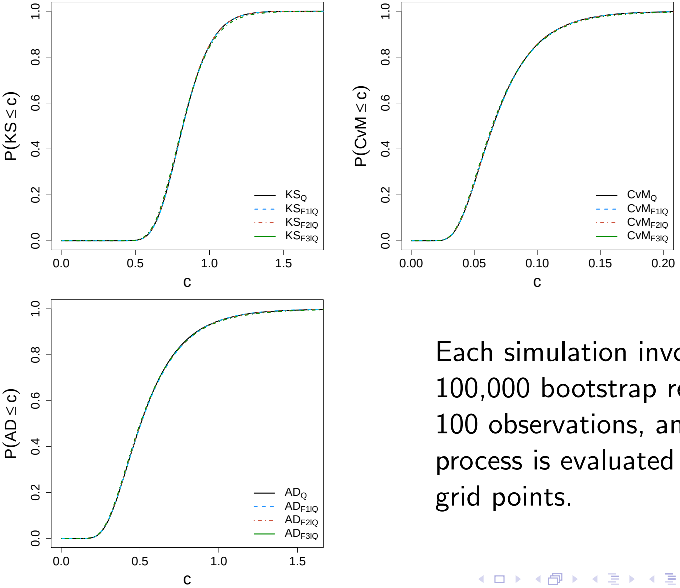
- **Reference distribution:** $q(x,\boldsymbol{\theta}) \propto e^{-\frac{1}{2\boldsymbol{\theta}_3}\left[(x_1-\boldsymbol{\theta}_1)^2+(x_2-\boldsymbol{\theta}_2)^2\right]}$.

Classical KS, CvM and AD: null distribution



Each simulation involves 100,000 bootstrap replicates, 100 observations, and the process is evaluated at 2000 grid points.

Rotated KS, CvM and AD: null distribution



Each simulation involves 100,000 bootstrap replicates, 100 observations, and the process is evaluated at 2000 grid points.

What about the power?

H_0	$\alpha = 0.05$					
	KS	CvM	AD	KS	CvM (K-2 rotated)	AD
Q	.9331	.9817	.9382	-	-	-
F_1	.8623	.9529	.9092	.6971	1	1
F_2	.1078	.1019	.1237	.1336	.2422	.2541
F_3	.9528	.9820	.6356	.9153	.9746	.9470

Each simulation involves 100,000 bootstrap replicates, 100 observations, and the process is evaluated at 2000 grid points.

Note: We should NOT expect the K-2 rotated statistics to always dominate their classical counterparts or vice-versa!

Key take-home ideas

When testing one continuous distribution which is multidimensional and/or depends on unknown parameters

- and we can simulate from it/evaluate it reasonably fast
  - $\Rightarrow$ we can reduce substantially the computational effort by simulating for the *projected empirical process*.
- but we cannot simulate from it/evaluate it reasonably fast
  - $\Rightarrow$ we can construct asymptotically distribution-free tests by means of the *K-2 transform*.

When testing $M > 1$ continuous distributions, $F_1, \ldots, F_M$, which are multidimensional and/or depend on unknown parameters

- $\Rightarrow$ we can avoid M different simulations and run just one by performing goodness-of-fit via *K-2 transform*.

References

- **Main reference:** Algeri S. (2022). K-2 rotated goodness-of-fit for multivariate data. *Physical Review D*.
- Oosterhoff J., van Zwet W.R. A note on contiguity and Hellinger distance. *Springer New York*, 2012.
- Algeri S. and Khmaladze E.V. When Pearson χ^2 and other divisible statistics are not goodness-of-fit tests. *In preparation*.
- Khmaladze, E.V. (1980). The use of ω^2 tests for testing parametric hypotheses. *Theory of Probability & Its Applications*.
- Khmaladze, E.V. (2016). Unitary transformations, empirical processes and distribution free testing. *Bernoulli*, 2016.

Thank you all for your time.