On computationally efficient methods for testing multivariate distributions with unknown parameters

Sara Algeri

School of Statistics, University of Minnesota

PHYSTAT Seminar
Dedicated to the memory of Sir David R. Cox,

March 23, 2022.

February 11, 2015

A younger me and Sir D.R. Cox at Nuffield College, Oxford.

His recommendations for a young (astro-)statistician

"Astrostatistics is a very interesting field and aims to address very important problems. What is particularly good for you is that it will allow you to explore many different areas of statistics."

"You need to know the maths. You don't just need the substance, what is more important in statistics is the method."

"Always do and focus on what interests you, not what they make you do."

Sir D.R. Cox.

On computationally efficient methods for testing multivariate distributions with unknown parameters

Goodness-of-fit vs test of hypothesis

- **Goodness-of-fit tests (GOF):** Given a postulated model for the data we test it against all possible alternatives.
  E.g., we expect that $X \sim N(\mu, 1)$, we test

  $$H_0 : X \sim N(\mu, 1) \quad \text{versus} \quad H_1 : X \not\sim N(\mu, 1).$$

  $\Rightarrow$ we have some power against all alternative models.

- **Tests of hypotheses:** Given a postulated model for the data, we test it against an alternative model.
  E.g., we expect that $X \sim N(\mu, 1)$, we test

  $$H_0 : \mu = 0 \quad \text{versus} \quad H_1 : \mu \neq 0.$$

  $\Rightarrow$ we have high power only against the alternative model under H_1.

Which Goodness-of-Fit test should we use? (1)

Discrete data

We typically rely on Pearson's X^2 or its asymptotically equivalent counterparts.

Main advantages

- Simple to implement
- When the expected counts are large we have a good χ^2 approximation (even if there are parameters to estimate).

Which Goodness-of-Fit test should we use? (2)

Continuous data

We have quite a few options:

- Kolmogorov-Smirnov
- Cramer-von Mises
- Anderson-Darling
- etc...

What do they have in common?

They can all be specified as functionals of the *empirical process*.

The empirical distribution function

Given a set of observations $x_1, \ldots, x_n$ from an unknown cumulative distribution function (cdf) $P(x) = P(X \leq x)$. We are interested in testing

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

for some postulated distribution $Q(x)$.

Since $P(x)$ is unknown, we begin by identifying an estimate of $P(x)$. A natural choice is the *empirical cumulative distribution function*

$$P_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{x_i \leq x\}} = \frac{\#\text{ observations} \leq x}{\text{sample size}}.$$

How can we use it to construct our test?

The empirical process

To test

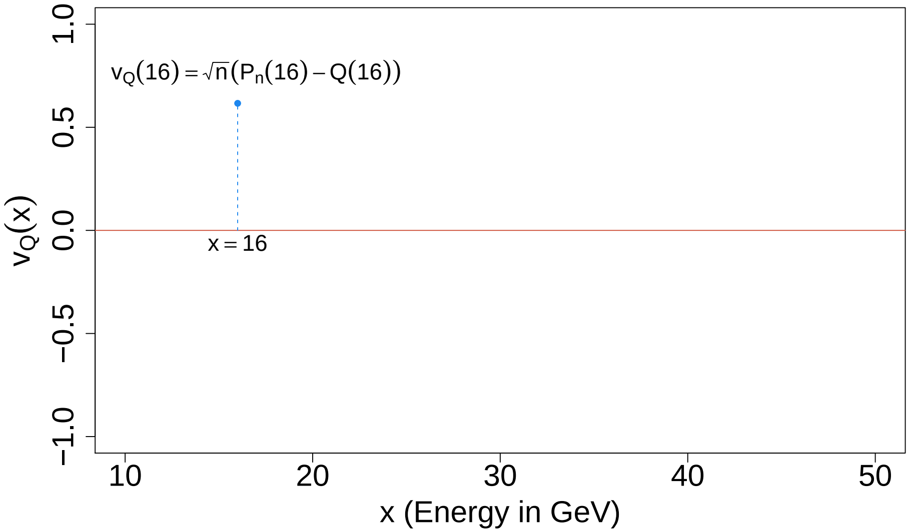
$$H_0 : P = Q \quad \text{versus} \quad H_1 : P \neq Q$$

we consider the *empirical process* $v_Q(x)$

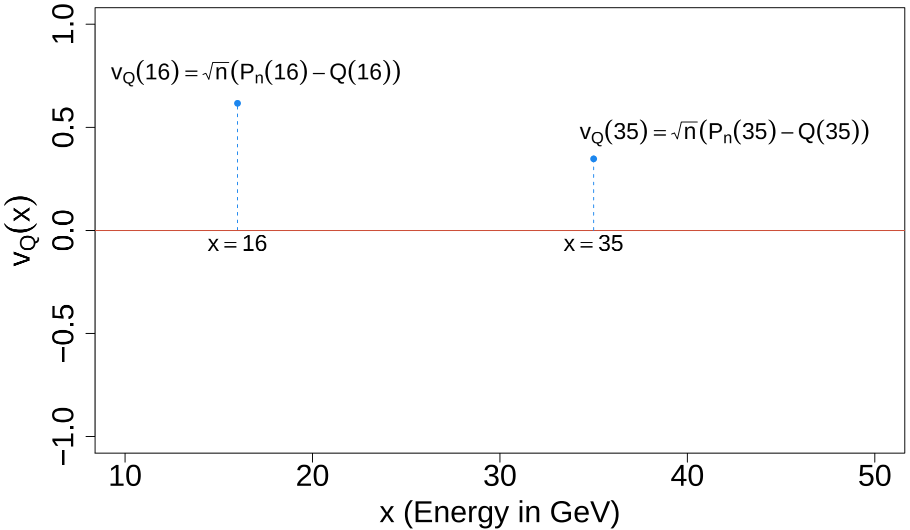
$$v_Q(x) = \sqrt{n}\Big[P_n(x) - Q(x)\Big] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\mathbb{1}_{\{x_i \leq x\}} - Q(x)\right]$$

Let's invest a few seconds to understand this fundamental object for a moment...

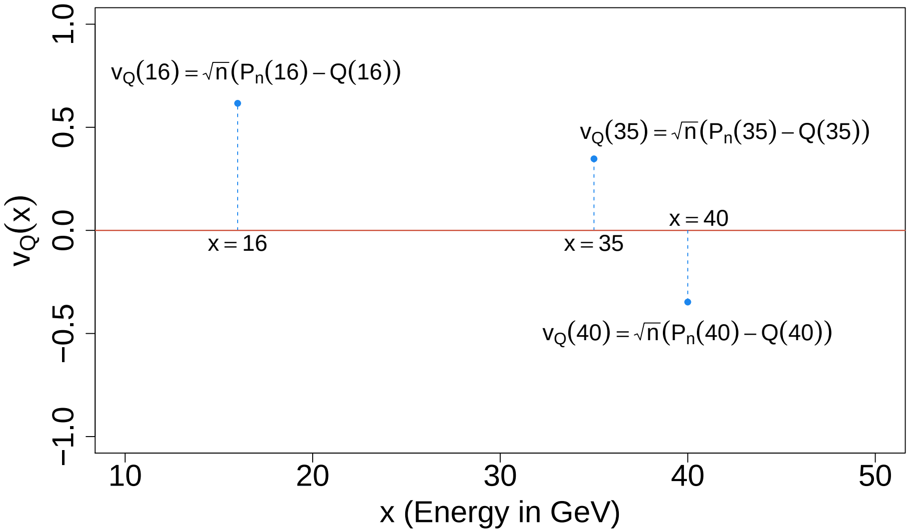
Empirical process: $v_Q(x) = \sqrt{n}\left[P_n(x) - Q(x)\right], \quad x \in [10, 50]$



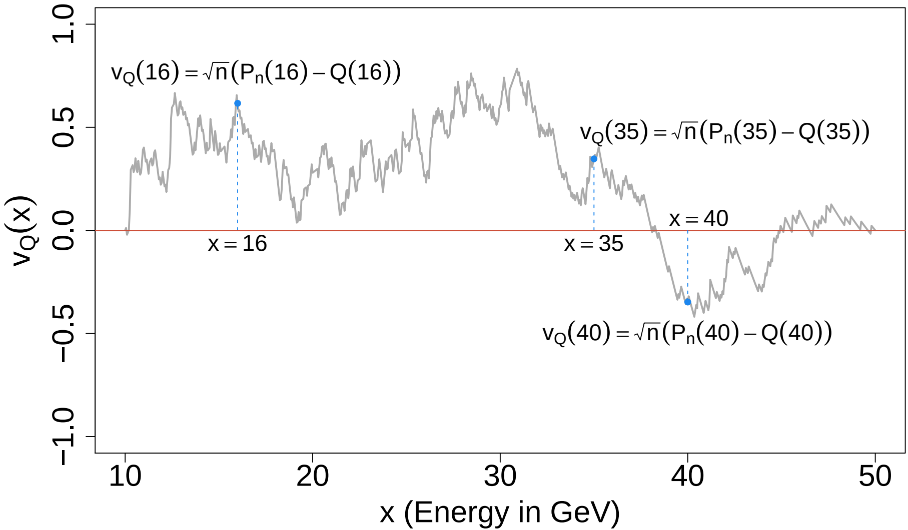
Empirical process: $v_Q(x) = \sqrt{n}\big[P_n(x) - Q(x)\big], \quad x \in [10, 50]$



Empirical process: $v_Q(x) = \sqrt{n}\left[P_n(x) - Q(x)\right], \quad x \in [10, 50]$



Empirical process: $v_Q(x) = \sqrt{n}\left[P_n(x) - Q(x)\right], \quad x \in [10, 50]$



An entire family of GOF tests

Recall that

$$v_Q(x) = \sqrt{n}\left[P_n(x) - Q(x)\right] \tag{1}$$

By taking functionals of $v_Q(x)$ we can construct a variety of GOF tests statistics. E.g.,

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

Advantages

If X is 1-dimensional and Q does not depend on unknown parameters, we consider the transformation

$$T = Q(X), \quad \text{and} \quad t_i = Q(x_i),$$

for $i = 1, \ldots, n$. We know that $T \sim \text{Unif}[0,1]$, hence, use the *uniform empirical process*

$$u_n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\mathbb{1}_{\{t_i \leq t\}} - t \right]$$

instead of $v_Q(x)$, and take functionals of $u_n(t)$ as test statistic $\Rightarrow$ we know the distribution of KS, CvM, and AD statistics and we have **distribution-freeness**.

Distribution-freeness

We have *distribution-freeness* whenever the distribution of the test statistic considered does not depend on the model Q being tested.

Limitations

If $\boldsymbol{X}$ is multidimensional and/or Q depends on unknown parameters, $\boldsymbol{\theta}$, estimated by means of some estimator $\widehat{\boldsymbol{\theta}}$, then

$$T = Q(\boldsymbol{X}, \widehat{\boldsymbol{\theta}}) \not\sim \text{Uniform}[0,1]$$

$\Rightarrow$ **we loose distribution-freeness**.

The simplest possible solutions

If X is multi-dimensional and/or Q depends on unknown parameters

- Discretize the data and use Pearson X^2 (or asymptotic equivalent).

Cons: Loss of information/power + in a low counts regime we run into serious problems (e.g., Haberman, 1988).

- Simulate the distribution of our KS, CvM, and AD statistics numerically via Monte Carlo or 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.

The parametric empirical process

Given a set of observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ from an unknown cumulative distribution function (cdf) $P(\boldsymbol{x}) = P(\boldsymbol{X} \leq \boldsymbol{x})$, $\boldsymbol{X} \in \mathcal{X} \subseteq \mathbb{R}^D$. 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* $v_Q(\boldsymbol{x}, \boldsymbol{\theta})$

$$\boxed{v_Q(\boldsymbol{x}, \boldsymbol{\theta}) = \sqrt{n}\Big[P_n(\boldsymbol{x}) - Q(\boldsymbol{x}, \boldsymbol{\theta})\Big]} \tag{2}$$

Estimating the empirical process

Let $\widehat{\boldsymbol{\theta}}$ be the MLE of $\boldsymbol{\theta}$, plug-it in $v_Q(\boldsymbol{x}, \boldsymbol{\theta})$:

$$v_Q(\boldsymbol{x}, \widehat{\boldsymbol{\theta}}) = \sqrt{n}\Big[P_n(\boldsymbol{x}) - Q(\boldsymbol{x}, \widehat{\boldsymbol{\theta}})\Big].$$

Simulating $v_Q(\boldsymbol{x}, \widehat{\theta})$ via the parametric bootstrap

- Let $\widehat{\boldsymbol{\theta}}_{obs}$= MLE of $\boldsymbol{\theta}$ obtained on the data observed.
- For b=1,. . . , B:
  - Simulate a bootstrap sample $\boldsymbol{x}_n^{(b)} = (x_1^{(b)}, \dots, x_n^{(b)})$ from $Q(\boldsymbol{x}, \widehat{\boldsymbol{\theta}}_{obs})$;
  - Estimate $\boldsymbol{\theta}$ on $\boldsymbol{x}_n^{(b)}$ and obtain $\widehat{\boldsymbol{\theta}}^{(b)}$,
  - For each point $\boldsymbol{x}$ considered evaluate

$$v_Q(\boldsymbol{x}, \widehat{\boldsymbol{\theta}}^{(b)}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big[\mathbb{1}_{\{\boldsymbol{x}_i^{(b)} \leq \boldsymbol{x}\}} - Q(\boldsymbol{x}, \widehat{\boldsymbol{\theta}}^{(b)})\Big].$$

Warning: If we evaluate the process at R points $\boldsymbol{x}$ over the search region, we have to evaluate $Q(\boldsymbol{x}, \widehat{\boldsymbol{\theta}}^{(b)})$, a total of RxB times.

Can we make it faster?

Recall that

$$v_Q(\boldsymbol{x}, \widehat{\boldsymbol{\theta}}) = \sqrt{n}\Big[P_n(\boldsymbol{x}) - Q(\boldsymbol{x}, \widehat{\boldsymbol{\theta}})\Big].$$

A Taylor expansion of $v_Q(\boldsymbol{x}, \widehat{\boldsymbol{\theta}})$ around $\boldsymbol{\theta}$ leads to

$$v_Q(\boldsymbol{x}, \widehat{\boldsymbol{\theta}}) \approx v_Q(\boldsymbol{x}, \boldsymbol{\theta}) - \sqrt{n}\,(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T \frac{\partial}{\partial \boldsymbol{\theta}} Q(\boldsymbol{x}, \boldsymbol{\theta}).$$

Moreover, let $q(\boldsymbol{x}, \boldsymbol{\theta})$ be the density of Q, a know theoretical result is

$$\sqrt{n}\,(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \approx \frac{1}{\sqrt{n}} \underbrace{\Gamma_{\boldsymbol{\theta}}^{-1}}_{\substack{\text{Inverse of}\\ \text{the Fisher}\\ \text{information}}} \underbrace{\sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \log q(\boldsymbol{x}_i, \boldsymbol{\theta})}_{\substack{\text{Score}\\ \text{function}}}$$

The projected empirical process

Putting everything together

$$\underbrace{v_Q(\boldsymbol{x},\widehat{\boldsymbol{\theta}})}_{\substack{\text{Empirical}\\\text{process}\\\text{at } \widehat{\boldsymbol{\theta}}}} \approx \underbrace{v_Q(\boldsymbol{x},\boldsymbol{\theta})}_{\substack{\text{Empirical}\\\text{process}\\\text{at } \boldsymbol{\theta}}} - \frac{1}{\sqrt{n}}\sum_{j=1}^{p} \frac{\partial}{\partial\theta_j} Q(\boldsymbol{x},\boldsymbol{\theta}) \underbrace{\Gamma_{\boldsymbol{\theta}}^{-1}}_{\substack{\text{Inverse of}\\\text{the Fisher}\\\text{information}}} \underbrace{\sum_{i=1}^{n}\frac{\partial}{\partial\theta_j}\log q(\boldsymbol{x}_i,\boldsymbol{\theta})}_{\substack{\text{Score}\\\text{functions}}}$$

- The error of the approximation is $o_p(1)$, that is, it quickly converges to zero in probability as $n \to \infty$.
- We call the right-hand-side of the approximation above *projected empirical process* (Khmaladze, 1980) and we denote it by $\widetilde{v}_Q(\boldsymbol{x},\boldsymbol{\theta})$.
- The projected empirical process **does not depend on $\widehat{\boldsymbol{\theta}}$!**
- Why "projected"? (I will tell you in a few slides).

Simulating $\widetilde{v}_Q(x, \theta)$ via the parametric bootstrap

- Let $\widehat{\boldsymbol{\theta}}_{obs}$= MLE of $\boldsymbol{\theta}$ obtained on the data observed.
- Evaluate $Q(\boldsymbol{x}, \widehat{\boldsymbol{\theta}}_{obs})$ and $\frac{\partial}{\partial\theta_j}Q(\boldsymbol{x}, \widehat{\boldsymbol{\theta}}_{obs})$ at each point $\boldsymbol{x}$ considered.
- For b=1,..., B:
  - Simulate a bootstrap sample $\boldsymbol{x}_n^{(b)} = (x_1^{(b)}, \ldots, x_n^{(b)})$ from $Q(\boldsymbol{x}, \widehat{\boldsymbol{\theta}}_{obs})$;
  - For each point $\boldsymbol{x}$ considered evaluate

$$\widetilde{v}_Q(\boldsymbol{x}, \widehat{\boldsymbol{\theta}}_{obs}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\mathbb{1}_{\{\boldsymbol{x}_i^{(b)} \leq \boldsymbol{x}\}} - Q(\boldsymbol{x}, \widehat{\boldsymbol{\theta}}_{obs})\right] - \frac{1}{\sqrt{n}}\sum_{j=1}^{p}\frac{\partial}{\partial\theta_j}Q(\boldsymbol{x}, \widehat{\boldsymbol{\theta}}_{obs})\Gamma^{-1}_{\widehat{\boldsymbol{\theta}}_{obs}}\sum_{i=1}^{n}\frac{\partial}{\partial\theta_j}\log q(\boldsymbol{x}_i^{(b)}, \widehat{\boldsymbol{\theta}}_{obs})$$

Note: If we evaluate the process at R points $\boldsymbol{x}$ over the search region, we have to evaluate $Q(\boldsymbol{x}, \widehat{\boldsymbol{\theta}}_{obs})$ and $\frac{\partial}{\partial\theta_j}Q(\boldsymbol{x}, \widehat{\boldsymbol{\theta}}_{obs})$, a total of R times (instead of $R \times B$ times!)

A toy example

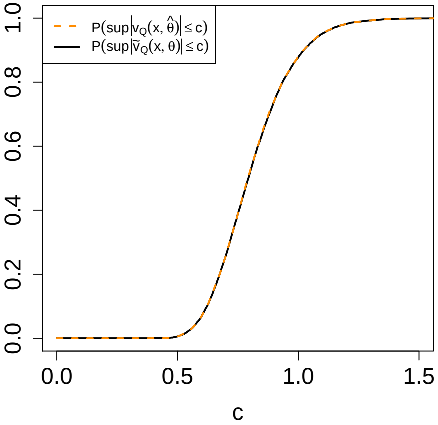
We draw a sample of $n = 100$ observations from

$$q(\boldsymbol{x}, \boldsymbol{\theta}) \propto e^{-\frac{1}{2\theta_3}\left[(x_1-\theta_1)^2+(x_2-\theta_2)^2\right]} \quad \boldsymbol{x} \in \mathcal{X} = [1, 20] \times [1, 25], \tag{3}$$

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

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

Simulated distributions of the KS statistic



The two simulated distributions are basically overlapping.

Which simulation procedure should we use?

- In theory, we would expect that bootstrapping the projected empirical process will be faster. But how much faster?

In our toy example...

Overall (system+user) CPU time needed to simulate the distributions of the Kolmogorov statistic $\sup_{\boldsymbol{x}} |v_Q(\boldsymbol{x}, \widehat{\boldsymbol{\theta}})|$ and $\sup_{\boldsymbol{x}} |\widetilde{v}_Q(\boldsymbol{x}, \boldsymbol{\theta})|$ via the parametric bootstrap over $10,000$ replicates and $n = 100$ observations.

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

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?)

Why "projected"?

Consider the normalized score vector defined as

$$b(\boldsymbol{x},\boldsymbol{\theta}) = \Gamma_{\boldsymbol{\theta}}^{-1/2} \frac{\partial}{\partial \boldsymbol{\theta}} \log q(\boldsymbol{x}_i,\boldsymbol{\theta}). \quad (4)$$

That is, conversely from $\frac{\partial}{\partial \theta_j} \log Q(\boldsymbol{x},\boldsymbol{\theta})$, each component $b_j(\boldsymbol{x},\boldsymbol{\theta})$ of (4) has mean zero, unit variance and is uncorrelated with each $b_k(\boldsymbol{x},\boldsymbol{\theta})$, $k \neq j$.

Our projected empirical process $\widetilde{v}_Q(\boldsymbol{x},\boldsymbol{\theta})$ is a projection of $v_Q(\boldsymbol{x},\boldsymbol{\theta})$ orthogonal to the normalized scored functions $b_j(\boldsymbol{x},\boldsymbol{\theta})$.

A useful (re-)formulation

Specifically

$$\widetilde{v}_Q(\boldsymbol{x},\boldsymbol{\theta}) = \overbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big\{\big[\mathbb{1}_{\{\boldsymbol{x}_i\leq\boldsymbol{x}\}} - Q(\boldsymbol{x},\boldsymbol{\theta})\big]}^{v_Q(\boldsymbol{x},\boldsymbol{\theta})} - \sum_{j=1}^{p} b_j(\boldsymbol{x}_i,\boldsymbol{\theta})\int_{-\infty}^{\boldsymbol{x}} b_j(\boldsymbol{x},\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{x}\Big\}$$

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

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

We will see very soon that the functions $\psi_{\boldsymbol{x}}(\boldsymbol{x}_i,\boldsymbol{\theta})$ play a fundamental role here.

A projected Brownian motion

The limiting process of $\widetilde{v}_Q(\boldsymbol{x}, \boldsymbol{\theta})$ can be shown to be a projected Brownian motion orthogonal to the normalized score functions $b_j(\cdot, \boldsymbol{\theta})$ (Khmaladze, 1980).

$\Rightarrow$ the limit of $\widetilde{v}_Q(\boldsymbol{x}, \boldsymbol{\theta})$ is Gaussian!
$\Rightarrow$ it is characterized by its mean and covariance functions, i.e.,

$$E_Q[\widetilde{v}_Q(\boldsymbol{x}, \boldsymbol{\theta})] = \int \psi_{\boldsymbol{x}}(\boldsymbol{t}, \boldsymbol{\theta}) \mathrm{d}Q(\boldsymbol{t}, \boldsymbol{\theta}) = E_Q\left[\psi_{\boldsymbol{x}}\right] = 0$$

$$E_Q[\widetilde{v}_Q(\boldsymbol{x}, \boldsymbol{\theta})\widetilde{v}_Q(\boldsymbol{x}', \boldsymbol{\theta})] = \int \psi_{\boldsymbol{x}}(\boldsymbol{t}, \boldsymbol{\theta})\psi_{\boldsymbol{x}'}(\boldsymbol{t}, \boldsymbol{\theta}) \mathrm{d}Q(\boldsymbol{t}, \boldsymbol{\theta}) = E_Q\left[\psi_{\boldsymbol{x}}\psi_{\boldsymbol{x}'}\right]$$

$\Rightarrow$ what really characterizes the limit are our $\psi_{\boldsymbol{x}}$.

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, namely $\phi_{\boldsymbol{x}}(\boldsymbol{x}_i, \boldsymbol{\theta}, \boldsymbol{\beta})$, we have that the processes

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

will have the same limit, under F and Q, respectively.

This can be done by means of the Khmaladze-2 (K-2) transform (Khmaladze, 2016).

The K-2 transform in a nutshell

The K-2 transform applied to the functions $\psi_{\boldsymbol{x}}(\boldsymbol{x}_i, \boldsymbol{\theta})$ is

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

- The isometry $l_{\boldsymbol{\theta},\beta}(\boldsymbol{x}) = \sqrt{\frac{q(\boldsymbol{x},\boldsymbol{\theta})}{f(\boldsymbol{x},\beta)}}$ ensures $E_F\left[(l_{\boldsymbol{\theta},\beta}\psi_{\boldsymbol{x}})(l_{\boldsymbol{\theta},\beta}\psi_{\boldsymbol{x}'})\right] = E_Q\left[\psi_{\boldsymbol{x}}\psi_{\boldsymbol{x}'}\right]$.
- The unitary operator K ensures that $E_F\left[K\left[(l_{\boldsymbol{\theta},\beta}\psi_{\boldsymbol{x}})\right]\right] = E_Q\left[\psi_{\boldsymbol{x}}\right] = 0$.
- The unitary operator $\boldsymbol{U}$ ensures orthogonality w.r.t. the normalized score functions under F, namely $a_j(\boldsymbol{x}, \boldsymbol{\theta})$, $j = 1, \ldots, p$.

See Algeri (2022) for the explicit expressions of K and $\boldsymbol{U}$.

A new family of test statistics

Recall that

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

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

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

which have the same limiting distribution as

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

Where is the computational advantage?

- The test statistics $\text{KS}_{F|Q}$, $\text{CvM}_{F|Q}$, and $\text{AD}_{F|Q}$ need to be computed only once on the data observed.
- We can then compare their observed values with the simulated distribution of KS_Q, CvM_Q, and AD_Q.

Requirements on F and Q

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

- 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

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

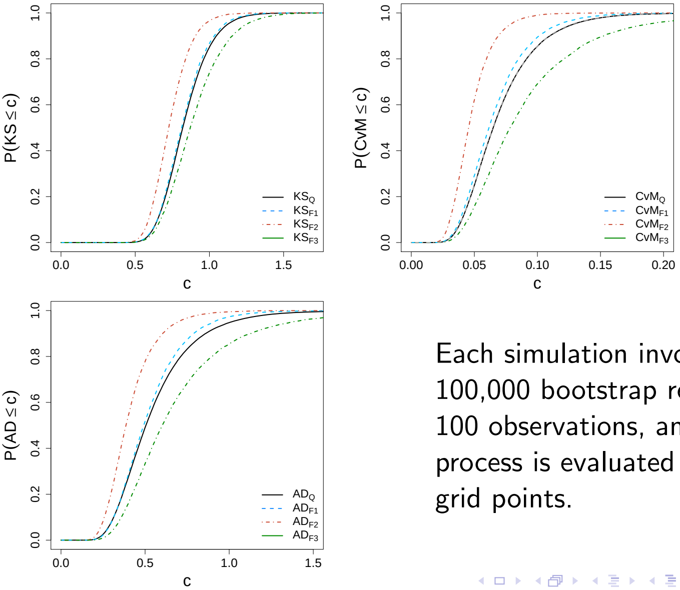
where $\mu = (0, 3)^T$, $\boldsymbol{\Sigma} = \begin{bmatrix} 20 & 10 \\ 10 & 20 \end{bmatrix}$, $\boldsymbol{x} \in \mathcal{X} = [1, 20] \times [1, 25]$.

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

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

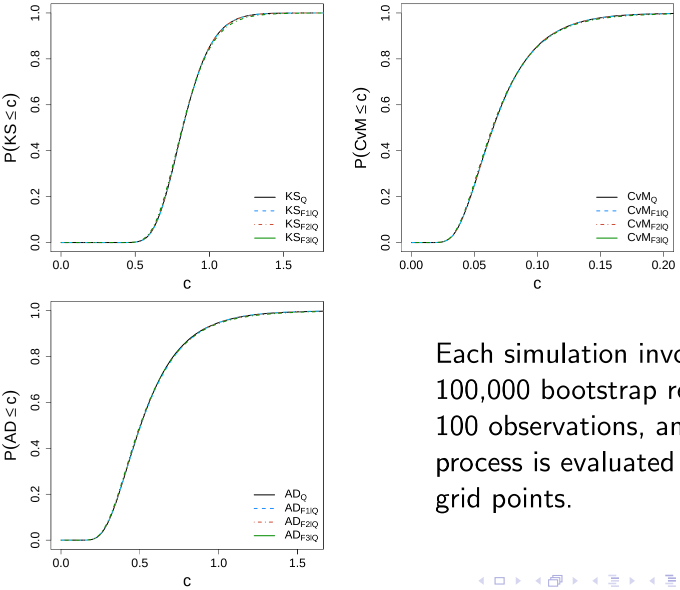
- **Reference distribution:** $q(\boldsymbol{x}, \boldsymbol{\theta}) \propto e^{-\frac{1}{2\theta_3}\left[(x_1 - \theta_1)^2 + (x_2 - \theta_2)^2\right]}$.

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.

Power

	$\alpha = 0.001$					
H_0	KS	CvM	AD	KS	CvM (K-2 rotated)	AD
Q	.4773	.7785	.4633	-	-	-
F_1	.3872	.6762	.4815	.1578	1	1
F_2	.0036	.0025	.0053	.0058	.0226	.0156
F_3	.6452	.7947	.0295	.5062	.7975	.6036

	$\alpha = 0.05$					
H_0	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.

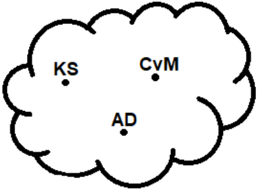
A few practical considerations

and possible points of discussion

- We should NOT expect the K-2 rotated statistics to always dominate their classical counterparts or vice-versa!
- The "closer" our reference distribution, Q, is to the F model we want to test, the "quicker" we will achieve distribution-freeness.
- The K-2 transform involves the operators K and U, these are linear operators $\Rightarrow$ while their implementation may be tedious when dealing with many parameters, it is not very difficult.
- In situations where the likelihood is not tractable in closed-form, a possible solution is that of constructing templates for the score, starting from the likelihood templates and applying the definition of derivative.
  - Recall that their evaluation does not need to be repeated on multiple runs, and it is only needed to evaluate the K-2 rotated test statistics on the data observed.

Conclusions

"You need to know the maths. You don't just need the substance, what is more important in statistics is the method." - Sir D.R. Cox.



If we focus on the substance we stop here.

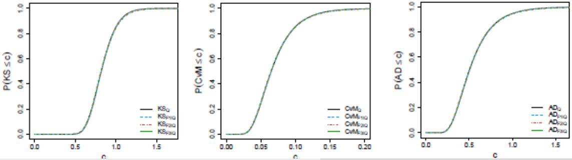
If we focus on the method we can unify them ...

$$v_Q(x) = \sqrt{n}\left[P_n(x) - Q(x)\right]$$

... and extend them to addess our needs !

$$\widetilde{v}_F(\mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\beta}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_{\mathbf{x}}(\mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\beta})$$

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



References

- <u>Main reference:</u> **Algeri S. (2022). K-2 rotated goodness-of-fit for multivariate data. Physical Review D**.
- Haberman, S. (1988). A warning on the use of chi-squared statistics with frequency tables with small expected cell counts. *Journal of the American Statistical Association*.
- Khmaladze, E. (1980). The use of ω^2 tests for testing parametric hypotheses. *Theory of Probability & Its Applications*.
- Khmaladze, E. (2016). Unitary transformations, empirical processes and distribution free testing. *Bernoulli*.

Thank you all for your time.

Extra slides

Material from: Algeri S. (2022+). Model assessment in counting experiments: a look beyond χ^2. *In preparation.*

Binned data: a toy example

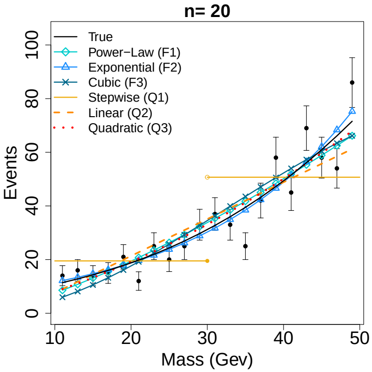
We aim to test three plausible representations of the background intensity functions typically used in the the context of the CMS Higgs-to-two photon analysis:

$$\lambda_{F_1}(x,\boldsymbol{\beta}) = \beta_0 x^{\beta_1}, \quad \lambda_{F_2}(x,\boldsymbol{\beta}) = \beta_0 e^{\beta_1 x}, \quad \text{and} \quad \lambda_{F_3}(x,\boldsymbol{\beta}) = \beta_0 x^2 + \beta_1 x^3, \tag{10}$$

We also consider three different reference distributions Q_1, Q_2, and Q_3, with associated intensity functions

$$\lambda_{Q_1}(x,\boldsymbol{\theta}) = \theta_0 \mathbb{1}_{\{x\leq 30\}} + \theta_1 \mathbb{1}_{\{x>30\}}, \quad \lambda_{Q_2}(x,\boldsymbol{\theta}) = \theta_0 x + \theta_1 x^2, \\ \text{and} \quad \lambda_{Q_3}(x,\boldsymbol{\theta}) = \theta_0 x^2 + \theta_1 x^3. \tag{11}$$

The data

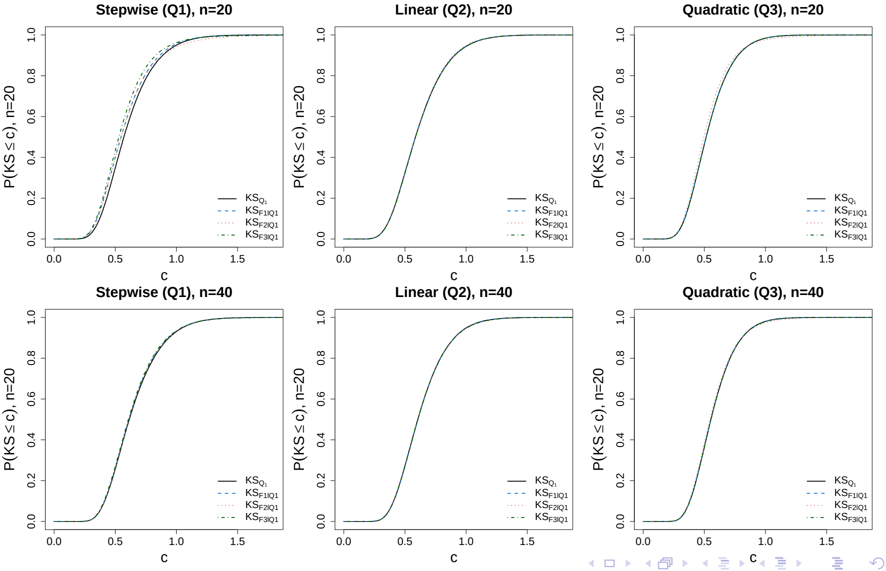


We consider a sample from a Poisson process with intensity function is

$$\lambda(x) \propto 700 \exp\left\{-\frac{1}{2}\left(\frac{x}{45.5} - 130\right)\right\}$$

For now we consider $n = 20$ bins.

Null distribution of K-2 rotated KS statistic



Power comparison

H_0	N	X^2	G^2	KS	CvM	AD	$KS_{F\|Q_1}$	$CvM_{F\|Q_1}$	$AD_{F\|Q_1}$	$KS_{F\|Q_2}$	$CvM_{F\|Q_2}$	$AD_{F\|Q_2}$	$KS_{F\|Q_3}$	$CvM_{F\|Q_3}$	$AD_{F\|Q_3}$
Q_1		.998	.999	**1**	**1**	**1**	-	-	-	-	-	-	-	-	-
Q_2		.307	.306	.671	**.747**	.739	-	-	-	-	-	-	-	-	-
Q_3		.107	.098	.176	.211	**.220**	-	-	-	-	-	-	-	-	-
F_1	20	.152	.137	.286	.348	.356	.196	.197	.255	.377	**.440**	.436	.271	.333	.348
F_2		.059	.061	.084	.096	.094	.103	.111	**.114**	.069	.073	.071	.064	.066	.067
F_3		.548	.456	.615	.664	.795	.580	.594	.639	.824	.875	**.883**	.719	.815	.835
Q_1		.987	.987	**1**	**1**	**1**	-	-	-	-	-	-	-	-	-
Q_2		.216	.215	.661	**.736**	.729	-	-	-	-	-	-	-	-	-
Q_3		.091	.081	.181	.215	**.228**	-	-	-	-	-	-	-	-	-
F_1	40	.124	.105	.277	.337	.340	.171	.200	.257	.369	**.435**	.432	.279	.337	.355
F_2		.055	.059	.076	.089	.085	.079	.085	**.091**	.071	.071	.071	.072	.075	.075
F_3		.447	.309	.580	.660	.788	.513	.576	.648	.835	.883	**.886**	.737	.835	.853
Q_1		.930	.926	**1**	**1**	**1**	-	-	-	-	-	-	-	-	-
Q_2		.154	.151	.680	**.760**	.752	-	-	-	-	-	-	-	-	-
Q_3		.084	.070	.182	.219	**.231**	-	-	-	-	-	-	-	-	-
F_1	80	.105	.083	.281	.343	.348	.135	.164	.238	.379	**.442**	.438	.278	.340	.358
F_2		.052	.057	.081	**.091**	.089	.083	.086	.089	.078	.083	.079	.080	.086	.086
F_3		.382	.205	.588	.667	.792	.523	.608	.701	.853	.900	**.901**	.752	.843	.858