Kernelized goodness-of-fit tests for discrete variables

Marie Ernst* and Yvik Swan

University of Liège, Belgium

Liège, June 5th, 2018

Target distribution $\mathcal{L}^0$

General context

- $\mathcal{L}^0$ admits a Stein operator $\mathcal{T}^0$ acting on a collection $\mathcal{F}^0$ of functions $f : \mathbb{R} \to \mathbb{R}$, i.e.,

$$\mathbb{E}[\mathcal{T}^0 f(X)] = 0 \ \forall f \in \mathcal{F}^0 \text{ iff } X \sim \mathcal{L}^0$$

- If $X \sim \mathcal{L}^0$, for some linear operator $\mathcal{D}$, we suppose

$$\mathbb{E}[\mathcal{T}^0 f(X) g(X)] = -\mathbb{E}[f(X) \mathcal{D} g(X)]$$

for all $f \in \mathcal{F}^0$ and all appropriate g.

Target distribution $\mathcal{L}^0$

Gaussian case

- $\mathcal{N}(0,1)$ admits the Stein operator $\mathcal{T}^0 f(x) = f'(x) - xf(x)$ acting on $\mathcal{F}^0 = \{\text{abs. cont. function } f : \mathbb{R} \to \mathbb{R} \text{ st } f' \in L^1\}$, i.e.,

$$X \sim \mathcal{N}(0,1) \text{ iff}$$
$$\mathbb{E}[Xf(X)] = \mathbb{E}[f'(X)] \ \forall f \in \mathcal{F}^0$$

- if $X \sim \mathcal{N}(0,1)$, for all $f \in \mathcal{F}^0$, we have

$$\mathbb{E}[(f'(X) - Xf(X))g(X)] = -\mathbb{E}[f(X)g'(X)]$$

for all g for which these expectations exist.

Stein kernel and Fisher information

If $\mathcal{L}^0$ is continuous and $X \sim \mathcal{L}^0$, for all $f \in \mathcal{F}^0$, we have

$$\mathbb{E}[\mathcal{T}^0 f(X) g(X)] = -\mathbb{E}[f(X) g'(X)]$$

for all g for which these expectations exist.

Score and Fisher information

For $f(x) = 1$,

- $\mathcal{T}^0 1(x) = \frac{p'(x)}{p(x)} =$ score
- $\mathbb{E}[(\mathcal{T}^0 1(X))^2] = \mathcal{I}(X)$

Stein kernel

For $\mathcal{T}^0 f(x) = x - \mu$,

- $\tau(x) = f(x) =$ Stein kernel
- $\mathbb{E}[(\tau(X))^2] = \mathbb{S}(X)$

Kernelizing the Stein identities

2-Stein kernel

A function $k : \mathbb{R}^2 \to \mathbb{R}$ for which both marginals $x \mapsto k(x, x')$ and $x \mapsto k(x', x)$ are in $\mathcal{F}^0$ at all x'.

Considering the function $K(x, x') = \mathcal{T}_1^0 \mathcal{T}_2^0 k(x, x')$, we get a relation between k and K:

$$\mathbb{E}[k(X, X')\mathcal{D}g_1(X)\mathcal{D}g_2(X')] = \mathbb{E}[K(X, X')g_1(X)g_2(X')]$$

for appropriate g_1, g_2.

Discrepancy between $\mathcal{L}^0$ and $\mathcal{L}^1$

k-based kernelized Stein discrepancy

Considering $Y, Y' \overset{i.i.d.}{\sim} \mathcal{L}^1$ and following Chwialkowski et al (2016) and Liu et al (2016), we define

$$\mathbb{S}_K(\mathcal{L}^1, \mathcal{L}^0) := \mathbb{E}\left[K(Y, Y')\right] = \mathbb{E}\left[\mathcal{T}_1^0 \mathcal{T}_2^0 k(Y, Y')\right]$$

Gaussian case

$$\begin{aligned} K(x, y) = \partial_x \partial_y k(x, y) - y \partial_x k(x, y) \\ - x \partial_y k(x, y) + xy k(x, y) \end{aligned}$$

and $\mathbb{S}_K(\mathcal{L}^1, \mathcal{L}^0) = \mathbb{E}\left[K(Y, Y')\right]$ for any $Y, Y' \overset{i.i.d.}{\sim} \mathcal{L}^1$.

Continuous case

If the distribution $\mathcal{L}^0$ (resp. $\mathcal{L}^1$) admits the Stein operator

$\mathcal{T}^0 f(x) = f'(x) + \rho^0(x) f(x)$ (resp. $\mathcal{T}^1 f(x) = f'(x) + \rho^1(x) f(x)$),

we have, for $Y \sim \mathcal{L}^1$,

$$\mathbb{E}[\mathcal{T}^0 f(Y)] = \mathbb{E}[f(Y) + \rho_0(Y) f(Y)] = \mathbb{E}[(\rho_0 - \rho_1)(Y) f(Y)]$$

Continuous case

If the distribution $\mathcal{L}^0$ (resp. $\mathcal{L}^1$) admits the Stein operator

$\mathcal{T}^0 f(x) = f'(x) + \rho^0(x) f(x)$ (resp. $\mathcal{T}^1 f(x) = f'(x) + \rho^1(x) f(x)$),

we have, for $Y \sim \mathcal{L}^1$,

$$\mathbb{E}[\mathcal{T}^0 f(Y)] = \mathbb{E}[f(Y) + \rho_0(Y) f(Y)] = \mathbb{E}[(\rho_0 - \rho_1)(Y) f(Y)]$$

Fisher information distance

$$\mathcal{I}(\mathcal{L}^0, \mathcal{L}^1) = \mathbb{E}[(\rho_0(Y) - \rho_1(Y))^2]$$

Continuous case

If the distribution $\mathcal{L}^0$ (resp. $\mathcal{L}^1$) admits the Stein operator

$$\mathcal{T}^0 f(x) = f'(x) + \rho^0(x) f(x) \quad (\text{resp. } \mathcal{T}^1 f(x) = f'(x) + \rho^1(x) f(x)),$$

we have, for $Y \sim \mathcal{L}^1$,

$$\mathbb{E}[\mathcal{T}^0 f(Y)] = \mathbb{E}[f(Y) + \rho_0(Y) f(Y)] = \mathbb{E}[(\rho_0 - \rho_1)(Y) f(Y)]$$

Fisher information distance

$$\mathcal{I}(\mathcal{L}^0, \mathcal{L}^1) = \mathbb{E}[(\rho_0(Y) - \rho_1(Y))^2]$$

Generalized Fisher information

If $Y, Y' \overset{i.i.d.}{\sim} \mathcal{L}^1$,

$$\begin{aligned}
\mathbb{S}_K(\mathcal{L}^1, \mathcal{L}^0) &= \mathbb{E}[\mathcal{T}_1^0 \mathcal{T}_2^0 k(Y, Y')] \\
&= \mathbb{E}[(\rho_0(Y) - \rho_1(Y)) \mathcal{T}_2^0 k(Y, Y')] \\
&= \mathbb{E}[(\rho_0(Y) - \rho_1(Y)) k(Y, Y') (\rho_0(Y') - \rho_1(Y'))]
\end{aligned}$$

Discrepancy between $\mathcal{L}^0$ and $\mathcal{L}^1$

Proposition

If k admits a representation of the form

$$k(x, x') = \sum_{j=0}^{\infty} \alpha_j e_j(x) e_j(x')$$

where $\alpha_j \geq 0$ and $(e_j)_{j \geq 0}$ a basis of $\mathcal{F}^0$, then

- $\mathbb{S}_K(\mathcal{L}^1, \mathcal{L}^0) \geq 0$
- $\mathbb{S}_K(\mathcal{L}^1, \mathcal{L}^0) = 0$ iff $\mathcal{L}^0 = \mathcal{L}^1$.

Estimation of kernelized Stein discrepancy

Given a sample $Y_1, \ldots, Y_n \sim \mathcal{L}^1$,

$$\hat{\mathbb{S}}_K^{(n)}(Y_1, \ldots, Y_n) = \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} \sum K(Y_i, Y_j).$$

Asymptotic behavior

Theorem

Using asymptotic properties of U-statistics (Serfling (2009))

1. If $\mathcal{L}_1 \neq \mathcal{L}_0$, then

$$\sqrt{n}\left(\hat{\mathbb{S}}_K^{(n)}(Y_1, \ldots, Y_n) - \mathbb{S}_K(\mathcal{L}^1, \mathcal{L}^0)\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

   where $\sigma^2 = \mathrm{Var}\left[\mathbb{E}[K(Y, Y') \mid Y]\right] \neq 0$.

2. If $\mathcal{L}_1 = \mathcal{L}_0$, then

$$n\hat{\mathbb{S}}_K^{(n)}(Y_1, \ldots, Y_n) \xrightarrow{d} \sum_{j=0}^{\infty} \lambda_j (Z_j^2 - 1)$$

   where $Z_j \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$ and $\{\lambda_j\}$ are the eigenvalues of the operator $\mathcal{A}$ where $\mathcal{A}g(x) = \mathbb{E}\left[g(X')K(X, X') \mid X = x\right]$.

Goodness of fit test

kernelized Stein based test statistic

For $\mathcal{H}_0 : \mathcal{L}_0 = \mathcal{L}_1$ vs $\mathcal{H}_1 : \mathcal{L}_0 \neq \mathcal{L}_1$, $R\mathcal{H}_0$ if $\hat{\mathbb{S}}_K^{(n)}(y_1, \ldots, y_n)$ is larger than the quantile of the limit distribution under the null.

Drawback

In general, the asymptotic null distribution is intractable.

$\Rightarrow$ We use bootstrap

Choosing a 2-Stein kernel

Mehler kernel

$$k(x,y) = \frac{1}{\sqrt{1-\rho^2}} \exp\left(-\frac{\rho^2(x^2+y^2)-2\rho xy}{2(1-\rho^2)}\right)$$
$$= \sum_{n=0}^{\infty} \frac{\rho^n}{n!} H_n(x) H_n(y)$$

for $\rho \in (-1,1)$.

Radial basis function kernel

$$k(x,y) = \exp\left(-(x-y)^2/2\right)$$
$$= \sum_{j=0}^{\infty} \frac{x^j}{\sqrt{j!}} e^{-x^2/2} \frac{y^j}{\sqrt{j!}} e^{-y^2/2}$$

Choosing a kernel

Kernel associated to $\mathcal{T}^0$

If $(e_j)_{j\geq 0}$ is an ONB of $\mathcal{F}^0$, the kernel could be

$$k(x, x') = \sum_{j=0}^{\infty} \alpha_j e_j(x) e_j(x')$$

Examples

- If $\mathcal{L}^0$ is Gaussian: Hermite polynomials
- If $\mathcal{L}^0$ is Beta: Jacobi polynomials
- If $\mathcal{L}^0$ is Gamma: Laguerre polynomials

Notations

- Probability mass function p
- Support of p : $\mathcal{S} = \{0, \ldots, n\}$ for some $n \in \mathbb{N} \cup \{\infty\}$
- Finite difference operators:

$$\Delta^{\ell} f(x) = \frac{f(x+\ell) - f(x)}{\ell}$$

- $L^1(p)$ denotes the collection of $g : \mathbb{Z} \to \mathbb{R}$ such that $\mathbb{E}[|g(X)|] = \sum_{x=0}^{n} |g(x)| p(x) < \infty$.

Discrete operator

Stein operator

(Ley, Reinert and Swan 2017)

- Canonical discrete $\ell-$Stein operator:

$$\mathcal{T}_p^{\ell} f(x) = \frac{\Delta^{\ell}(f(x)p(x))}{p(x)}$$

  NB: convention: $\mathcal{T}_p^{\ell} f(x) = 0$ if $x \notin \mathcal{S}$.

- Canonical discrete $\ell-$Stein class: $\mathcal{F}^{\ell}(p)$ is the collection of $f : \mathbb{Z} \to \mathbb{R}$ st $f \in L^1(p)$ and $\mathcal{T}_p^{\ell} f \in L^1(p)$ has mean 0.

Particular cases

- Forward Stein operator: $\mathcal{T}_p^{+} f(x)$ if $\ell = 1$
- Backward Stein operator: $\mathcal{T}_p^{-} f(x)$ if $\ell = -1$

Properties of canonical operators

Δ^ℓ respects the following product formula:

$$\Delta^\ell\big(f(x)g(x)\big) = \big(\Delta^\ell f(x)\big)g(x+\ell) + f(x)\big(\Delta^\ell g(x)\big)$$

Then $\mathcal{T}_p^\ell$ respects the following relations:

$$\begin{aligned}\mathcal{T}_p^\ell\big(f(x)g(x)\big) &= \frac{1}{\ell}\left(f(x+\ell)g(x+\ell)\frac{p(x+\ell)}{p(x)} - f(x)g(x)\right)\\ &= \big(\mathcal{T}_p^\ell f(x)\big)g(x+\ell) + f(x)\big(\Delta^\ell g(x)\big)\end{aligned}$$

By definition of Stein operator, we deduce the "Stein identities":

- $\mathbb{E}\big[f(X)g(X)\big] = \mathbb{E}\left[f(X+\ell)g(X+\ell)\frac{p(X+\ell)}{p(X)}\right]$
- $\mathbb{E}\left[\big(\mathcal{T}^\ell f(X)\big)g(X+\ell)\right] = -\mathbb{E}\left[f(X)\big(\Delta^\ell g(X)\big)\right]$

Properties of canonical operators

Δ^ℓ respects the following product formula:

$$\Delta^\ell\big(f(x)g(x)\big) = \big(\Delta^\ell f(x)\big)g(x+\ell) + f(x)\big(\Delta^\ell g(x)\big)$$

Then $\mathcal{T}_p^\ell$ respects the following relations:

$$\begin{aligned}\mathcal{T}_p^\ell\big(f(x)g(x)\big) &= \frac{1}{\ell}\left(f(x+\ell)g(x+\ell)\frac{p(x+\ell)}{p(x)} - f(x)g(x)\right)\\ &= \big(\mathcal{T}_p^\ell f(x)\big)g(x+\ell) + f(x)\big(\Delta^\ell g(x)\big)\end{aligned}$$

By definition of Stein operator, we deduce the "Stein identities":

- $\mathbb{E}\big[f(X)g(X)\big] = \mathbb{E}\left[f(X+\ell)g(X+\ell)\frac{p(X+\ell)}{p(X)}\right]$
- $\mathbb{E}\left[\big(\mathcal{T}^\ell f(X)\big)g(X+\ell)\right] = -\mathbb{E}\left[f(X)\big(\Delta^\ell g(X)\big)\right]$

A class of discrete distributions

Class of discrete distributions

Focus on distributions p which satisfy a recurrence relation

$$p(x+\ell) = \frac{a(x)}{b(x)} p(x)$$

or, equivalently

$$\Delta^{\ell} p(x) = \frac{1}{\ell} \left(\frac{a(x) - b(x)}{b(x)} \right) p(x)$$

with $a, b : \mathbb{Z} \to \mathbb{R}$ two "well-behaved" functions.

Stein operators

$$\mathcal{T}_p^{\ell}\big(f(x)g(x)\big) = \frac{1}{\ell} \left(f(x+\ell)g(x+\ell) \frac{p(x+\ell)}{p(x)} - f(x)g(x) \right)$$

A class of discrete distributions

Class of discrete distributions

Focus on distributions p which satisfy a recurrence relation

$$p(x+\ell) = \frac{a(x)}{b(x)} p(x)$$

or, equivalently

$$\Delta^{\ell} p(x) = \frac{1}{\ell} \left(\frac{a(x) - b(x)}{b(x)} \right) p(x)$$

with $a, b : \mathbb{Z} \to \mathbb{R}$ two "well-behaved" functions.

Stein operators

$$\mathcal{T}_p^{\ell}\big(f(x)g(x)\big) = \frac{1}{\ell} \left(f(x+\ell)g(x+\ell)\frac{a(x)}{b(x)} - f(x)g(x) \right)$$

A class of discrete distributions

Class of discrete distributions

Focus on distributions p which satisfy a recurrence relation

$$p(x+\ell) = \frac{a(x)}{b(x)} p(x)$$

or, equivalently

$$\Delta^{\ell} p(x) = \frac{1}{\ell} \left(\frac{a(x) - b(x)}{b(x)} \right) p(x)$$

with $a, b : \mathbb{Z} \to \mathbb{R}$ two "well-behaved" functions.

Stein operators

$$\mathcal{T}_p^{\ell}\big(f(x)g(x)\big) = \frac{1}{\ell} \left(f(x+\ell)g(x+\ell)\frac{a(x)}{b(x)} - f(x)g(x) \right)$$

$\rightsquigarrow$ Natural choice for f: $f(x+\ell) = b(x)$

A class of discrete distributions

Class of discrete distributions

Focus on distributions p which satisfy a recurrence relation

$$p(x+\ell) = \frac{a(x)}{b(x)} p(x)$$

or, equivalently

$$\Delta^{\ell} p(x) = \frac{1}{\ell} \left(\frac{a(x) - b(x)}{b(x)} \right) p(x)$$

with $a, b : \mathbb{Z} \to \mathbb{R}$ two "well-behaved" functions.

Stein operators

$$\mathcal{A}_1^{\ell} g(x) := g(x+\ell)a(x) - g(x)b(x-\ell)$$
$$\mathcal{A}_2^{\ell} g(x) := a(x) \left(\Delta^{\ell} g(x) \right) + r(x)g(x)$$

where $r(x) = \mathcal{T}_p^{\ell} f(x) = (a(x) - b(x-\ell))/\ell$

Examples

Binomial (n, p)

- $p(x) = \binom{n}{x} p^x (1-p)^{n-x}$ for $x \in \{0, \ldots, n\}$

Examples

Binomial (n, p)

- $p(x) = \binom{n}{x} p^x (1-p)^{n-x}$ for $x \in \{0, \ldots, n\}$
- $\dfrac{p(x+1)}{p(x)} = \dfrac{p(n-x)}{(x+1)(1-p)}$

Examples

Binomial (n, p)

- $p(x) = \binom{n}{x} p^x (1-p)^{n-x}$ for $x \in \{0, \ldots, n\}$
- $\dfrac{p(x+1)}{p(x)} = \dfrac{p(n-x)}{(x+1)(1-p)}$
- $\mathcal{A}_1^+ f(x) := p(n-x) f(x+1) - x(1-p) f(x)$

Examples

Binomial (n, p)

- $p(x) = \binom{n}{x} p^x (1-p)^{n-x}$ for $x \in \{0, \ldots, n\}$
- $\dfrac{p(x+1)}{p(x)} = \dfrac{p(n-x)}{(x+1)(1-p)}$
- $\mathcal{A}_1^+ f(x) := p(n-x) f(x+1) - x(1-p) f(x)$
- $\mathcal{A}_2^+ f(x) := p(n-x) \Delta^+ f(x) + (x - np) f(x)$

Examples

Binomial(n, p)	Poisson(λ)
$p(x) = \binom{n}{x} p^x (1-p)^{n-x}$ $x \in \{0, \ldots, n\}$ $a(x) = p(n-x)$ $b(x) = (x+1)(1-p)$ $r(x) = x - np$	$p(x) = e^{-\lambda} \lambda^x (x!)^{-1}$ $x \in \mathbb{N}$ $a(x) = \lambda$ $b(x) = x+1$ $r(x) = x - \lambda$

Beta binomial(α, β, n)	Hypergeometric(n, N, R)
$p(x) = \binom{n}{x} B(\alpha+x, n+\beta-x)(B(\alpha,\beta))^{-1}$ $x \in \{0, \ldots, n\}$ $a(x) = (n-x)(\alpha+x)$ $b(x) = (x+1)(n+\beta-(x+1))$ $r(x) = (\alpha+\beta)x - n\alpha$	$p(x) = \binom{R}{x}\binom{N-R}{n-x}\left(\binom{N}{n}\right)^{-1}$ $x \in \{0, \ldots, n\}$ $a(x) = \frac{1}{N}(n-x)(R-x)$ $b(x) = \frac{1}{N}(x+1)(N-R-(n-(x+1))$ $r(x) = x - n\frac{R}{N}$

Panjer(α, β, p_0)	Ord family (1968)
$p(x) = \left(\alpha + \frac{\beta}{x}\right) p(x-1);\ p_0 = p(0)$ $x \in \mathbb{N}$ $a(x) = \alpha(x+1) + \beta$ $b(x) = x+1$ $r(x) = x(1-\alpha) + \beta - \alpha$	$\Delta p(x-1) = \frac{(a-x)p(x-1)}{b_0 + b_1 x + b_2 x(x-1)}$ $x \in \mathbb{N}$ $a(x) = c_0 + c_1 x + c_2 x^2$ $b(x) = d_0 + d_1 x + d_2 x^2$ $r(x) = r_0 + r_1 x$

Fulman-Goldstein(q, m, n) (finite case)	Fulman-Goldstein(q, m) (limit case)
$p(x) = q^{-x(m+x)} \frac{\prod_{i=1}^{m+n}(1-q^{-i}) \prod_{i=x+1}^{\infty}(1-q^{-i})}{\prod_{i=1}^{n-x}(1-q^{-i}) \prod_{i=1}^{m+x}(1-q^{-i})}$ $x \in \{0, \ldots, n\}$ $a(x) = q(1-q^{-n+x-1})$ $b(x) = (q^{x+1}-1)(q^{m+x+1}-1)$ $r(x) = (q^x-1)(q^{m+x}-1) - q(1-q^{-n+x-1})$	$p(x) = q^{-x(m+x)} \prod_{i=x+1}^{\infty}(1-q^{-i}) \left(\prod_{i=1}^{m+x}(1-q^{-i})\right)^{-1}$ $x \in \mathbb{N}$ $a(x) = q$ $b(x) = (q^{x+1}-1)(q^{m+x+1}-1)$ $r(x) = (q^x-1)(q^{m+x}-1) - q$

Fisher information and Stein kernel

If $\mathcal{L}^0$ is discrete and $X \sim \mathcal{L}^0$, for all $f \in \mathcal{F}^0$, we have

$$\mathbb{E}[\mathcal{T}^0 f(X) g(X)] = -\mathbb{E}[f(X) \mathcal{D} g(X)]$$

for all g for which these expectations exist. For $Y \sim q$,

p-score and Fisher information

- p-score for Y: $\rho(Y)$ st $\mathbb{E}[\rho(Y) g(Y)] = \mathbb{E}[b(Y-1)\Delta^- g(Y)]$
- *a* Fisher information: $\mathbb{E}[(\mathcal{T}^0 f(Y))^2] = \mathbb{E}[(\rho(Y))^2] = \mathcal{I}_b(q)$

p-kernel

- p-kernel for Y: $\tau(Y)$ st $\mathbb{E}[r(Y) g(Y)] = \mathbb{E}[\tau(Y)\Delta^- g(Y)]$
- *a* Stein information: $\mathbb{E}\left[(\tau(Y))^2\right]$

p-score and p-kernel: example

Binomial(n, p)

- Binomial score for distributions on $\mathbb{N}$: $\rho(Y)$ st

$$\mathbb{E}[\rho(Y)g(Y)] = \mathbb{E}[(1-p)Y\Delta^{-}g(Y)]$$

- Binomial kernel for distributions on $\mathbb{N}$: $\tau(Y)$ st

$$\mathbb{E}[\tau(Y)\Delta^{-}g(Y)] = \mathbb{E}[(Y-np)g(Y)]$$

If $X \sim \text{Bin}(n, p)$, $\rho(X) = X - np$ and $\tau(X) = (1-p)X$.

Goodness-of-fit tests for discrete distributions

- Pearson Chi-squared test

Goodness-of-fit tests for discrete distributions

- Pearson Chi-squared test
- Larger family: power-divergence statistics

$$2n\mathcal{I}^{\lambda} := \frac{2}{\lambda(\lambda+1)} \sum_{i=1}^{k} O_i \left[\left(\frac{O_i}{E_i} \right)^{\lambda} - 1 \right]$$

Goodness-of-fit tests for discrete distributions

- Pearson Chi-squared test
- Larger family: power-divergence statistics

$$2n\mathcal{I}^{\lambda} := \frac{2}{\lambda(\lambda+1)} \sum_{i=1}^{k} O_i \left[\left(\frac{O_i}{E_i} \right)^{\lambda} - 1 \right]$$

- Even larger family: ϕ-divergence defined by Csiszar (1967):

$$D_{\phi} = \sum_{i=1}^{k} E_i \phi(O_i/E_i)$$

Examples of GOF statistics

Particular power-divergence statistics

- $\lambda = 1$: Pearson chi-square
- $\lambda = 0$: log-likelihood ratio G^2
- $\lambda = -1/2$: Freeman-Tukey or Hellinger distance statistics H^2
- $\lambda = -1$: modified log-likelihood ratio statistics GM
- $\lambda = -2$: Neyman's modified chi-square statistic
- $\lambda = 2/3$: Cressie and Read statistic

Particular ϕ-divergence statistics

The choice of

$$\phi(x) = \frac{1}{\lambda(\lambda+1)}(x^{\lambda+1} - x)$$

leads to the power divergence statistic.

Comparison of GOF tests

1. Power divergence tests:
   - Pearson ($\lambda = 1$)
   - Cressie-Read ($\lambda = 2/3$)
   - loglikelihood ratio ($\lambda = 0$)
   - Freeman Tukey ($\lambda = -1/2$)
   - modified loglikelihood ratio ($\lambda = -1$)
   - Neyman modified chi-squared ($\lambda = -2$)
2. Kernelized discrepancy tests:
   - Using a RBF kernel
   - Using a kernel defined by Afendras et al's polynomials

Afendras polynomials

Afendras et al (2011) polynomials

Discrete orthogonal polynomials $P_k(x)$ which satisfy the identify

$$\mathbb{E}[P_k(X)g(X)] = \mathbb{E}[\tau^k(X)g^{(k)}(X)]$$

where τ is the Stein kernel of X.

Adhoc discrete 2-Stein kernel

Considering $N+1$ polynomials ($N = 0, \ldots, n$),

$$k_N(x, x') := \sum_{j=0}^{N} \alpha_j P_j(x) P_j(x')$$

where $\alpha_j^{-1} = \mathbb{E}_p[P_j^2(X)]$.

Test for binomiality

Target distribution

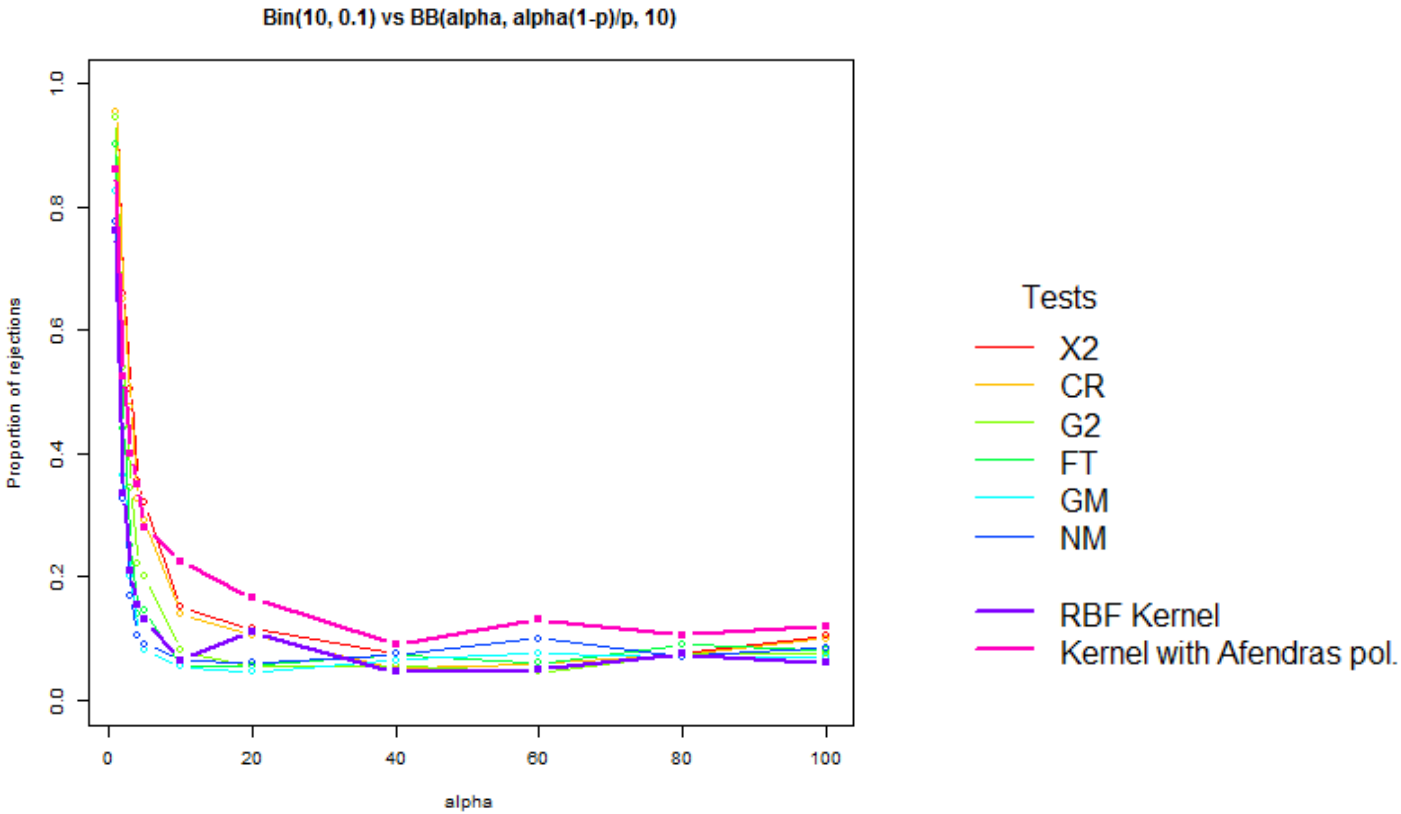
Binomial (n, p)

Alternative distribution

1. Binomial (n_2, p_2)
2. Beta Binomial (α, β, n) with $p = \frac{\alpha}{\alpha+\beta}$
3. Hypergeometric (n, N, R) with $p = R/N$
4. Poisson binomial, i.e., sum of indep. Bern(p_i)
5. A sum of indicators with some totally correlated, i.e., $\text{Bin}(n-j, p) \oplus j\text{Bern}(p)$ for $j = 2, \ldots, n-2$.

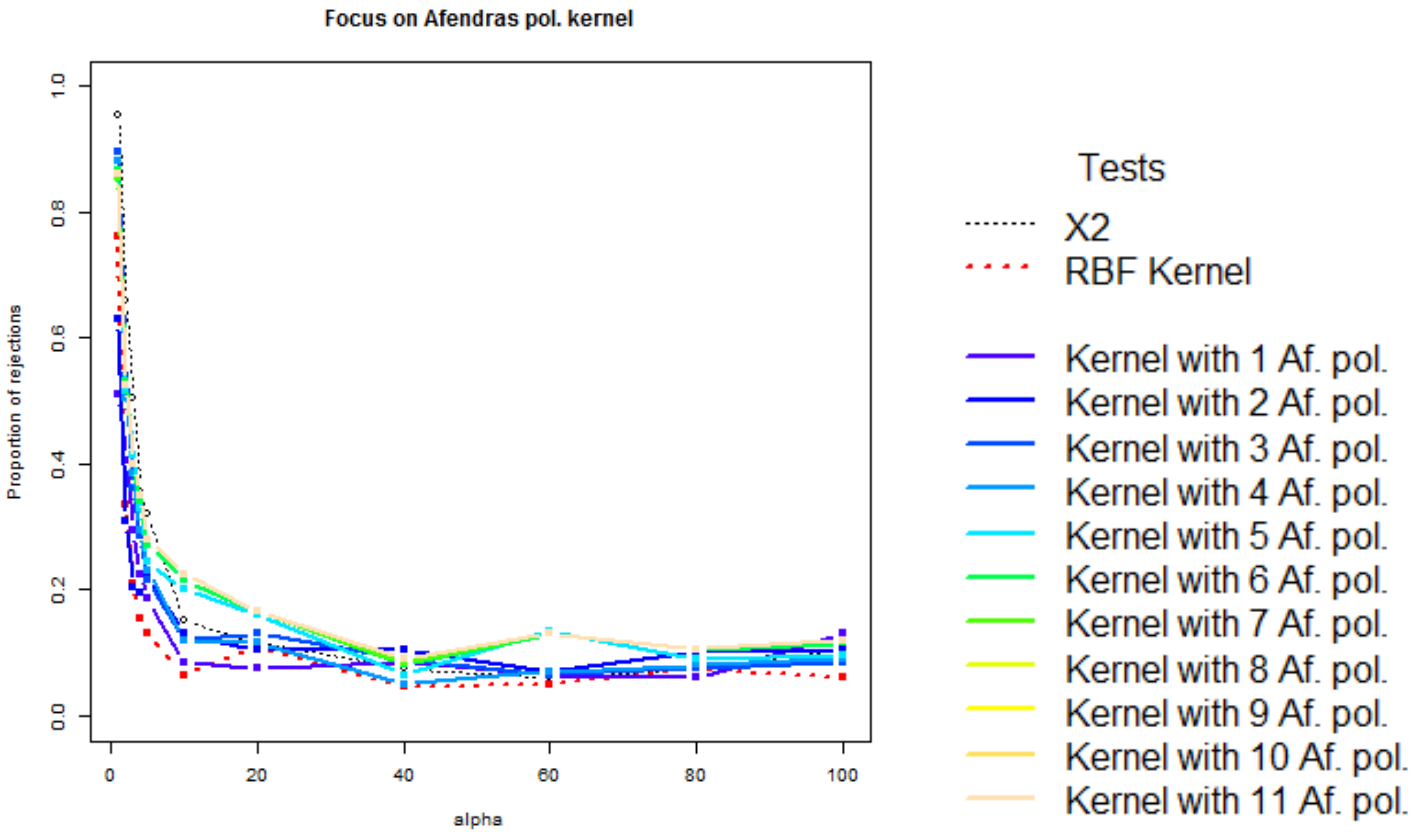
Some results

Beta Binomial (α, β, n) vs Bin(10,0.1)



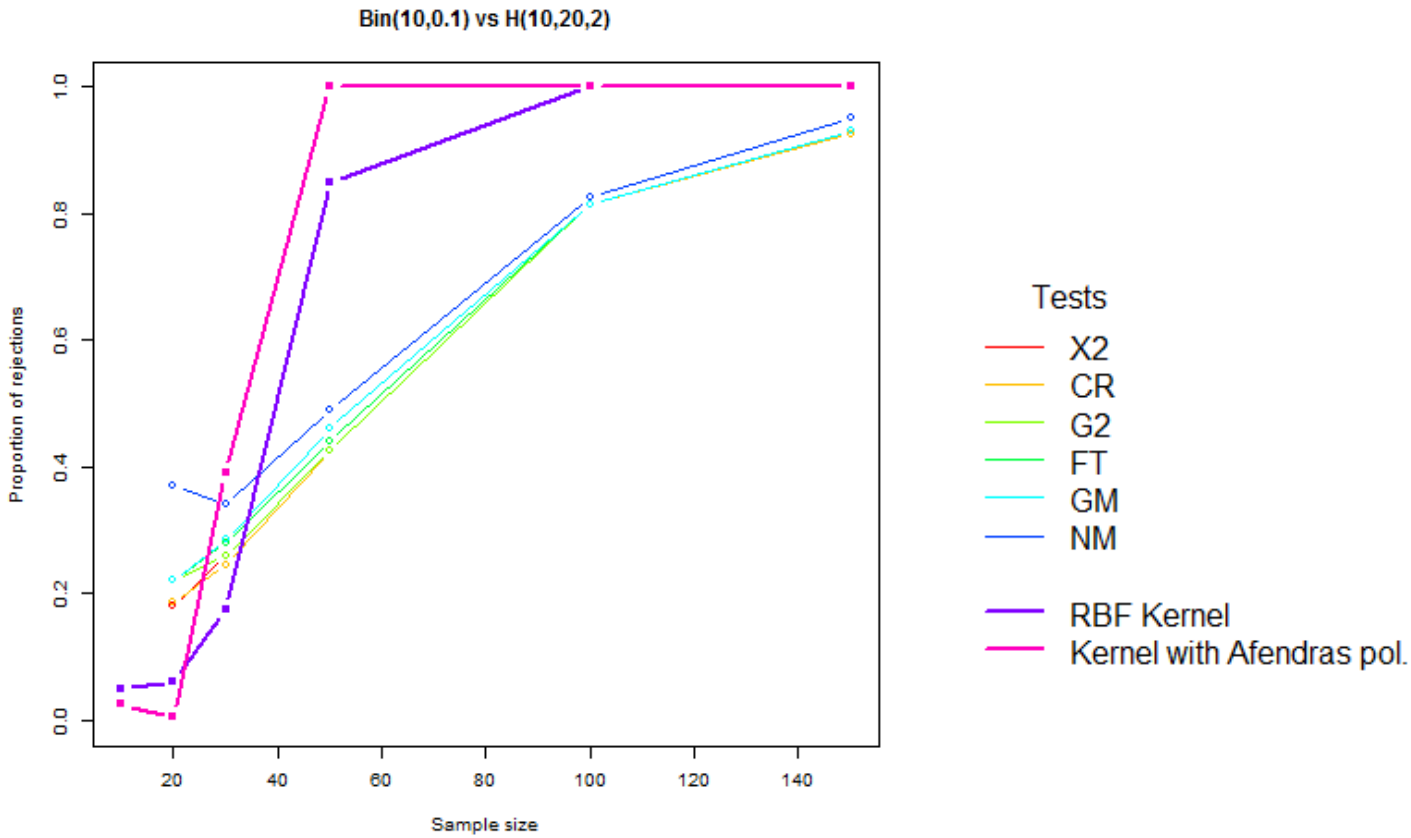
Some results

Beta Binomial (α, β, n) vs Bin(10,0.1)



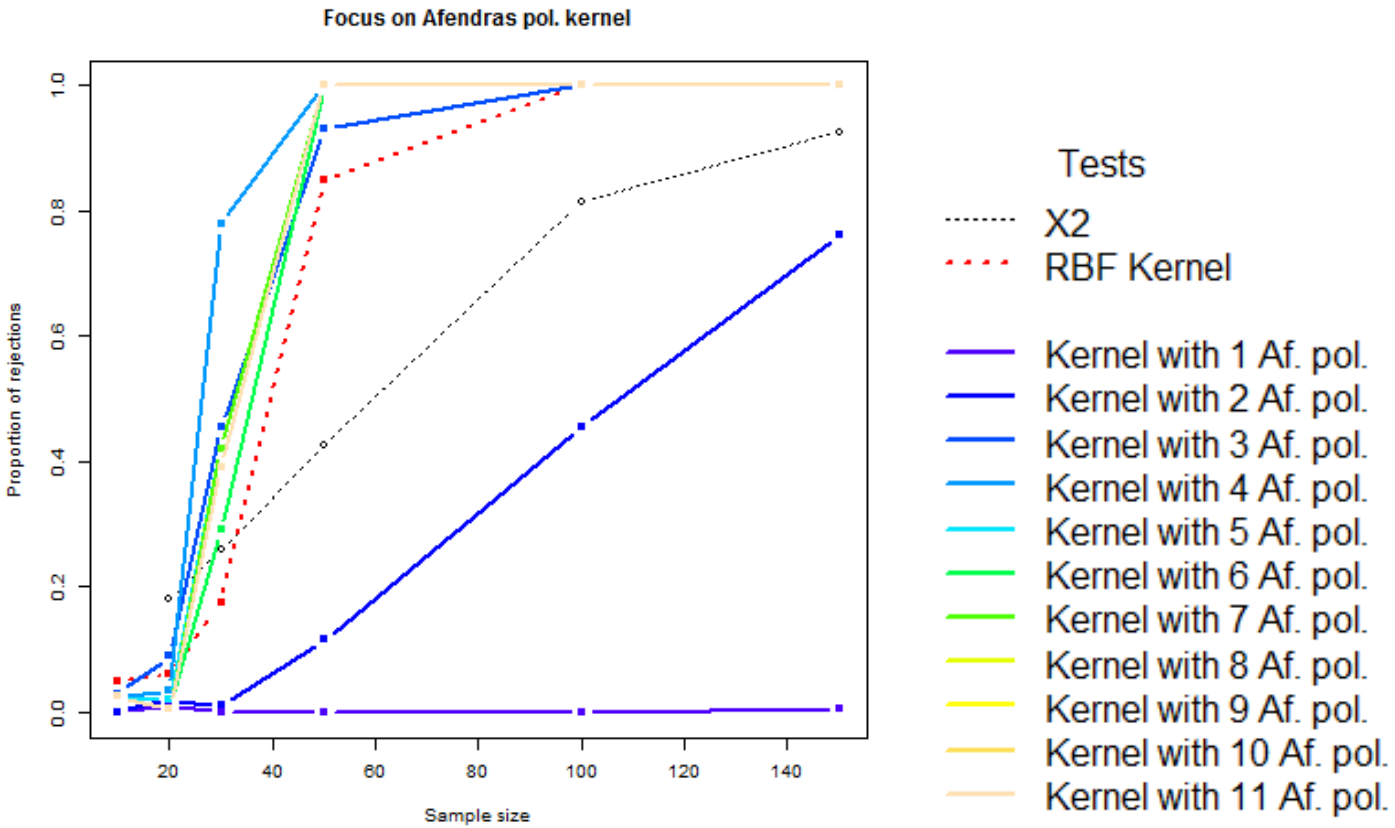
Some results

Hypergeometric $(n, N, 2)$ vs Bin(10,0.1)



Some results

Hypergeometric $(n, N, 2)$ vs Bin(10,0.1)



Preliminary conclusions

When the samples are drawn from another family of distributions than the target one,

- the discrepancy tests have similar/better power than power divergence tests;
- a kernel associated to the target distribution is more adequate than an arbitrary kernel;
- the power increases with the number of polynomials used in the kernel.

When the samples are drawn from the same family of distributions, tests based on the first polynomial seems to have the best power.

Perspectives

- Optimization of the choice of kernel for each distribution?
- Null and non-null distribution?
- Extension to the multivariate case?

References

- G. Afendras et al. An extended stein-type covariance identity for the pearson family with applications to lower variance bounds. *Bernoulli*, **17**(2), 2011.
- K. Chwialkowski et al. A kernel test of goodness of fit. *International Conference on Machine Learning*, 2016.
- J. Fulman, L. Goldstein. Stein's method and the rank distribution of random matrices over finite fields. *The Annals of Probability*, **43**(3), 2015.
- C. Ley and Y. Swan. Stein's density approach and information inequalities. *Electronic Communications in Probability*, **18**, 2013.
- C. Ley, G. Reinert, Y. Swan. Stein's method for comparison of univariate distributions. *Probability Surveys*, **14**, 2017.
- Q. Liu, J. Lee, and M. Jordan. A kernelized stein discrepancy for goodness-of-fit tests. *International Conference on Machine Learning*, 2016.
- H. H. Panjer. Recursive evaluation of a family of compound distributions. *ASTIN Bulletin: The Journal of the IAA*, **12**(1), 1981.
- R. J. Serfling. *Approximation theorems of mathematical statistics*, **162**. John Wiley & Sons, 2009.