A canonical Stein operator

Yvik Swan, University of Liège

November 7, 2014

Joint work with Christophe Ley (Brussels) and Gesine Reinert (Oxford)

Outline

1 Stein's method

2 A canonical Stein operator

3 Distances between expectations

4 Multivariate extension

Stein's Method for Normal Approximation

Stein (1972, 1986): $Z \sim \mathcal{N}(0,1)$ if and only if for all smooth functions f,

$$\mathbb{E}Zf(Z) = \mathbb{E}f'(Z)$$

Roughly speaking, we deduce that for a random variable W with $\mathbb{E}W = 0, \mathrm{Var}W = 1$, if

$$\mathbb{E}f'(W) - \mathbb{E}Wf(W) \approx 0$$

for many functions f, then $W \approx Z$ in distribution.

Stein's method gives a *systematic* way of quantifying this heuristic, using

$$|\mathbb{E}f'(W) - \mathbb{E}Wf(W)|$$

as a precise measure of non-Gaussianity.

Stein's method for Normal approximation

To do so take h a test function. Then solve for f the *Stein equation*

$$h(x) - \mathbb{E}h(Z) = f'(x) - xf(x).$$

The functions f_h are well understood and have good properties.

Next consider a probability metric of the form

$$d_{\mathcal{H}}(W, Z) = \sup_{h \in \mathcal{H}} |\mathbb{E}h(W) - \mathbb{E}h(Z)|$$

(such as the Kolmogorov metric, the Wasserstein metric, the total variation distance,...).

Deduce

$$d_{\mathcal{H}}(W, Z) = \sup_{h \in \mathcal{H}} \left|\mathbb{E}\left[f_h'(W) - Wf_h(W)\right]\right|. \tag{1}$$

Stein's method for the Gaussian

In other words, Stein's method transforms the problem of bounding distances between W and Z such as

$$\begin{aligned} TV(W,Z) &= \sup_{A\subset\mathbb{R}} |P(W\in A) - P(Z\in B)| \\ \mathcal{W}(W,Z) &= \sup_{h\in Lip(1)} |\mathbb{E}h(W) - \mathbb{E}h(Z)| \\ \mathrm{Kol}(W,Z) &= \sup_{z\in\mathbb{R}} |P(W\le z) - P(Z\le z)| \end{aligned}$$

into that of bounding the quantity

$$\Delta_{\mathcal{H}}(W) = \sup_{h\in\mathcal{H}} \left|\mathbb{E}\left[f_h'(W) - Wf_h(W)\right]\right|.$$

This is a good thing, because that quantity is remarkably amenable to computations.

Version 1 : Comparing score functions

For W sufficiently regular we define its *score* as the random variable $\rho_W(W)$ which satisfies

$$\mathbb{E}\left[\rho_W(W)f(W)\right] = -\mathbb{E}f'(W)$$

for all smooth f (if it exists). Clearly $\rho_Z(Z) = -Z$.

From (1) we get

$$\begin{aligned} d_{\mathcal{H}}(W, Z) &= \sup_{h\in\mathcal{H}} \left|\mathbb{E}\left[f_h'(W) - Wf_h(W)\right]\right| \\ &= \sup_{h\in\mathcal{H}} \left|\mathbb{E}\left[(\rho_W(W) + W)f_h(W)\right]\right| \\ &\leq \kappa_{\mathcal{H}}\mathbb{E}\left|\rho_W(W) + W\right| \end{aligned}$$

with $\kappa_{\mathcal{H}} = \|f_h\|$.

Sharp bounds on $\kappa_{\mathcal{H}}$ are known; properties of $\rho_W(W)$ are often good.

Version 1 : Comparing Score functions

The score function approach was introduced by *Shimizu (1975)* and *Stein (1986, Lesson 6)*. See also *Ley and Swan (2013a, 2013b)*.

The quantity

$$\mathcal{J}_{st}(W) = \mathbb{E}\left[(\rho_W(W) + W)^2\right]$$

is the so-called *Fisher information distance* of W.

If W is a standardized sum then remarkably precise bounds on $\mathcal{J}_{st}(W)$ (involving Poincaré constants) are due to *Johnson and Barron (2004)*.

For instance they show that if $W = \sum_{i=1}^{n} X_i/\sqrt{n}$ with X_i iid with variance 1 then

$$\mathcal{J}_{st}(W) \leq \frac{2R^\star}{n} J(X)$$

for $R^\star$ the restricted Poincaré constant of X. This, combined with Stein's method, provides rates of convergence of the correct order.

Version 1' : Working on the score directly

Stein et al (2004) and Chatterjee and Shao (2012) choose a different route : solving for f the Stein equation

$$h(x) - \mathbb{E}h(W) = f'(x) + \rho_W(x)f(x)$$

they deduce

$$d_{\mathcal{H}}(W, Z) = \sup_{h \in \mathcal{H}} \left| \mathbb{E}\left[f_h'(Z) + \rho_W(Z)f_h(Z)\right]\right|.$$

They work out the properties of f_h in terms of those of ρ_W and apply a technique known as *exchangeable pairs* to bound the rhs directly.

Such an approach of course also extends to non-Gaussian approximation via

$$d_{\mathcal{H}}(W, X) = \sup_{h \in \mathcal{H}} \left| \mathbb{E}\left[f_h'(X) + \rho_W(X)f_h(X)\right]\right|,$$

although many conditions on the laws of X and W are necessary in order to get things to work out.

Version 2 : Comparing Stein kernels

For W sufficiently regular we define its *Stein kernel* as the random variable $\tau_W(W)$ which satisfies

$$\mathbb{E}\left[\tau_W(W)f'(W)\right] = \mathbb{E}[Wf(W)]$$

for all smooth f (if it exists). Clearly $\tau_Z(Z) = 1$.

From (1) we get

$$\begin{aligned} d_{\mathcal{H}}(W, Z) &= \sup_{h\in\mathcal{H}} \left|\mathbb{E}\left[f_h'(W) - Wf_h(W)\right]\right| \\ &= \sup_{h\in\mathcal{H}} \left|\mathbb{E}\left[(\tau_W(W) - 1)f_h'(W)\right]\right| \\ &\leq \lambda_{\mathcal{H}}\mathbb{E}\left|\tau_W(W) - 1\right| \end{aligned}$$

with $\lambda_{\mathcal{H}} = \|f_h'\|$.

Sharp bounds on $\kappa_{\mathcal{H}}$ are known; properties of $\tau_W(W)$ are often good.

Version 2 : Comparing Stein kernels

The Stein kernel approach was used e.g. in *Stein (1986, Lesson 6)*, *Cacoullos et al. (1992)* and *Cacoullos, Papathanasiou and Utev (1994)*.

The quantity

$$S(X) = \mathbb{E}\left[(\tau_X(X) - 1)^2\right],$$

is called the *Stein discrepancy* of X.

It can be shown e.g. that if $W = \sum_{i=1}^{n} X_i/\sqrt{n}$ with X_i iid with variance 1 then

$$S(W) \leq \frac{S(X)}{n}.$$

Within this literature a very important breakthrough is due to *Nourdin and Peccati (2009)* who show that if X is "chaotic" with mean 0 and variance 1 then

$$S(X) = \mathrm{Var}(\tau_X(X)) \leq C_1 E\left[X^4 - 3\right],$$

hereby obtaining the famous *fourth moment theorem*.

Version 2' : Working on the Stein factors

Döbler (2014) and *Tudor and Kusuoka (2012, 2014)* and several others introduced the Stein equation

$$h(x) - \mathbb{E}h(W) = \tau_W(x)f'(x) - xf(x).$$

Properties of the solutions f_h are quite good and, using

$$d_{\mathcal{H}}(W, Z) = \sup_{h \in \mathcal{H}} \left| \mathbb{E}\left[\tau_W(Z)f_h'(Z) - Zf_h(Z)\right]\right|,$$

yield good bounds.

Again this is not reserved to Gaussian approximation so that we get

$$d_{\mathcal{H}}(W, X) = \sup_{h \in \mathcal{H}} \left| \mathbb{E}\left[\tau_W(X)f_h'(X) - Xf_h(X)\right]\right|$$

and this yields good bounds in quite some generality.

See also *Ledoux (2012)* and *Azmoodeh, Campese and Poly (2014)* for abstract extensions of the fourth moment theorem.

Version 3 : Working on the Stein operator

In many cases nothing much is known about either $\rho_W(W)$ or $\tau_W(W)$ so that neither of the previous approaches can be used.

Then one can always start from

$$d_{\mathcal{H}}(W, Z) = \sup_{h \in \mathcal{H}} \left| \mathbb{E}\left[f_h'(W) - W f_h(W)\right]\right|$$

and

- apply Taylor expansion;
- use exchangeable pairs;
- use couplings;
- use biasing mechanisms (zero or size biased);
- be smart.

The i.i.d. example

Take $X_1, \ldots, X_n$ i.i.d. copies of X such that $\mathbb{E}X = 0, \mathrm{Var}X = \frac{1}{n}$. Define

$$W = \sum_{i=1}^{n} X_i.$$

Put $W_i = W - X_i = \sum_{j \neq i} X_j$. Then

$$\begin{aligned}
\mathbb{E}Wf(W) &= \sum_i \mathbb{E}X_i f(W) \\
&= \sum_i \mathbb{E}X_i f(W_i) + \sum_i \mathbb{E}X_i^2 f'(W_i) + R \\
&= \frac{1}{n} \sum_i \mathbb{E}f'(W_i) + R
\end{aligned}$$

So

$$\mathbb{E}f'(W) - \mathbb{E}Wf(W) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\{f'(W) - f'(W_i)\} + R.$$

Theorem

It is known that for the sup-norm

$$||f'|| \leq 2||h'||$$

hence for any smooth h

$$|\mathbb{E}h(W) - \mathbb{E}h(Z)| \quad \leq \quad \|h'\| \left(\frac{2}{\sqrt{n}} + \sum_{i=1}^{n} \mathbb{E}|X_i^3| \right).$$

Conclusion :

- If $X_i = O(n^{-\frac{1}{2}})$ then the bound is of order $O(n^{-\frac{1}{2}})$
- Nothing goes to infinity.
- The proof extends to local dependence.
- With couplings we can treat weak global dependence too.

Generalizing

Since 1972, Stein's method for the Gaussian has generated hundreds of papers.

Chen (1975) extends to Poisson target; this has also generated hundreds of paper.

Outside these two important settings, the method is also available for

- Exponential
- Geometric
- Gamma and χ^2
- Negative binomial
- binomial
- ...

See e.g.

https://sites.google.com/site/yvikswan/about-stein-s-method

for a list of (70+) papers.

Stein's method in general

For μ a target distribution, with support $\mathcal{I}$:

1. Find a suitable operator $\mathcal{A}$ and a wide class of functions $\mathcal{F}(\mathcal{A})$ such that $X \sim \mu$ if and only if

   $$\mathbb{E}\mathcal{A}f(X) = 0$$

   for all functions $f \in \mathcal{F}(\mathcal{A})$.

2. Let $\mathcal{H}(\mathcal{I})$ be a measure-determining class on $\mathcal{I}$. For each $h \in \mathcal{H}$ find a solution $f = f_h \in \mathcal{F}(\mathcal{A})$ of

   $$h(x) - \mathbb{E}h(X) = \mathcal{A}f(x),$$

   where $X \sim \mu$. If the solution exists and if it is unique in $\mathcal{F}(\mathcal{A})$ then we can write

   $$f(x) = \mathcal{A}^{-1}(h(x) - \mathbb{E}h(X)).$$

   Obtain good bounds on these functions f.

Comparison of distributions

3. Let X and Y have distributions μ_X and μ_Y with Stein operators $\mathcal{A}_X$ and $\mathcal{A}_Y$.

   Suppose that $\mathcal{F}(\mathcal{A}_X) \cap \mathcal{F}(\mathcal{A}_Y) \neq \emptyset$ and choose $\mathcal{H}(\mathcal{I})$ such that all solutions f of the Stein equation belong to this intersection.

   Then

   $$\mathbb{E}h(X) - \mathbb{E}h(Y) = \mathbb{E}\mathcal{A}_Y f(X) = \mathbb{E}\mathcal{A}_Y f(X) - \mathbb{E}\mathcal{A}_X f(X)$$

   and

   $$\sup_{h \in \mathcal{H}(\mathcal{I})} |\mathbb{E}h(X) - \mathbb{E}h(Y)| \leq \sup_{f \in \mathcal{F}(\mathcal{A}_X) \cap \mathcal{F}(\mathcal{A}_Y)} |\mathbb{E}\mathcal{A}_X f(X) - \mathbb{E}\mathcal{A}_Y f(X)|.$$

   A general version of the method then boils down to : *find a way to bound the right hand side*.

Stein's method

The devil is in the detail : everything relies on discovering a good operator $\mathcal{A}$ whose inverse yields good bounds.

How to choose the $\mathcal{A}$? For any given target, there are infinitely many...

We still need a *definition* of these operators.

Outline

1. Stein's method
2. A canonical Stein operator
3. Distances between expectations
4. Multivariate extension

Our set-up

Let $(\mathcal{X}, \mathcal{B}, \mu)$ be a measure space.

Example : $\mathcal{X} = \mathbb{R}$.

Let $\mathcal{X}^\star$ be the set of real-valued functions on $\mathcal{X}$ and take a linear operator $\mathcal{D} : dom(\mathcal{D}) \subset \mathcal{X}^\star \to im(\mathcal{D})$ such that $dom(\mathcal{D}) \setminus \{0\} \neq \emptyset$.

Example : $\mathcal{D}f = f'$.

Let $\mathcal{D}^{-1} : im(\mathcal{D}) \to dom(\mathcal{D})$ be a linear operator which sends any $h = \mathcal{D}f$ onto f. Then

$$\mathcal{D}\left(\mathcal{D}^{-1}h\right) = h$$

for all $h \in im(\mathcal{D})$ and, for $f \in dom(\mathcal{D})$,

$$\mathcal{D}^{-1}\left(\mathcal{D}f\right)$$

is defined up to addition with an element of $ker(\mathcal{D})$.

Example: $\mathcal{D}^{-1}f = \int f(x)dx$.

Our set-up

Assumption

There exists a linear operator $\mathcal{D}^\star : dom(\mathcal{D}^\star) \subset \mathcal{X}^\star \to im(\mathcal{D}^\star)$ and a constant $l := l_{\mathcal{X},\mathcal{D}}$ such that

$$\mathcal{D}(f(x)g(x+l)) = g(x)\mathcal{D}f(x) + f(x)\mathcal{D}^\star g(x)$$

for all $(f,g) \in dom(\mathcal{D}) \times dom(\mathcal{D}^\star)$.

Example: $\mathcal{D}^* f = \mathcal{D}f = f'$, $l = 0$ because $(fg)' = f'g + fg'$

Under this assumption, $\mathcal{D}$ and $\mathcal{D}^\star$ are skew-adjoint in the sense that

$$\int_{\mathcal{X}} g\mathcal{D}f d\mu = -\int_{\mathcal{X}} f\mathcal{D}^\star g d\mu$$

for all $(f,g) \in dom(\mathcal{D}) \times dom(\mathcal{D}^\star)$ such that

- $g\mathcal{D}f \in L^1(\mu)$ or $f\mathcal{D}^\star g \in L^1(\mu)$
- $\int_{\mathcal{X}} \mathcal{D}(f(\cdot)g(\cdot + l))d\mu = 0$.

Example: $\int fg' = -\int f'g$ for all f, g such that $\int (fg)' = 0$.

Example 2

Let μ be the counting measure on $\mathcal{X} = \mathbb{Z}$ and take $\mathcal{D} = \Delta^+$, the forward difference operator. Then

$$\mathcal{D}^{-1} f(x) = \sum_{k=\bullet}^{x-1} f(k).$$

Also we have the discrete product rule

$$\Delta^+(f(x)g(x-1)) = g(x)\Delta^+ f(x) + f(x)\Delta^- g(x)$$

for all $f, g \in \mathbb{Z}^\star$ and all $x \in \mathbb{Z}$.

Hence our assumption is satisfied with $\mathcal{D}^\star = \Delta^-$, the backward difference operator and $l = -1$.

Example 3

Let $\mu(x)$ be the $\mathcal{N}(0,1)$ measure on $\mathbb{R}$, with density φ, and take

$$\mathcal{D}_\varphi f(x) = f'(x) - xf(x) = \frac{(f(x)\varphi(x))'}{\varphi(x)}.$$

Then

$$\mathcal{D}_\varphi^{-1} f(x) = \frac{1}{\varphi(x)} \int_\bullet^x f(y)\varphi(y)dy.$$

Also we have the product rule

$$\mathcal{D}_\varphi(gf)(x) = (gf)'(x) - xg(x)f(x) = g(x)\mathcal{D}_\varphi f(x) + f(x)g'(x).$$

Hence our assumption is satisfied with $\mathcal{D}^\star g = g'$ and $l = 0$.

Example 4

Let $\mu(x)$ be the Poisson(λ)measure on $\mathbb{Z}^+$ with pmf γ_λ and

$$\Delta_\lambda^+ f(x) = \lambda f(x+1) - xf(x) = \frac{\Delta^+(f(x)x\gamma_\lambda(x))}{\gamma_\lambda(x)}.$$

Then

$$(\Delta_\lambda^+)^{-1} f(x) = \frac{1}{x\gamma_\lambda(x)} \sum_{k=\bullet}^{x-1} f(k)\gamma_\lambda(k)$$

(which is ill-defined at $x = 0$) and

$$\Delta_\lambda^+(g(x-1)f(x)) = g(x)\Delta_\lambda^+ f(x) + f(x)x\Delta^- g(x).$$

Hence our assumption is satisfied with $\mathcal{D}^\star g(x) = x\Delta^- g(x)$ and $l = -1$.

Remark

In all examples the choice of $\mathcal{D}$ is, in a sense, arbitrary and other options are available.

Less conventional choices of $\mathcal{D}$ can be envisaged (even forward differences in the continuous setting, non standard derivatives etc.).

In principle no restriction on dimensions is necessary (more on this at the end of the talk).

From now on for the sake of presentation we mainly concentrate on the Lebesgue measure and $\mathcal{D}$ the usual derivative, i.e.

$$\mathcal{D}^* f = \mathcal{D} f = f'; \, \mathcal{D}^{-1} f = \int f(x) dx.$$

A canonical Stein operator

Let X be a continuous random variable having pdf p with interval support $\mathcal{I} \subset \mathbb{R}$.

Definition

The *Stein class* of X is the class $\mathcal{F}(X)$ of functions $f : \mathbb{R} \to \mathbb{R}$ such that

- fp is differentiable on $\mathbb{R}$
- $(fp)'$ is integrable and $\int (fp)' = 0$.

Definition

To X we associate the *Stein operator* $\mathcal{T}_X$ of X such that

$$\mathcal{T}_X f = \frac{(fp)'}{p}$$

with the convention that $\mathcal{T}_X f = 0$ outside of $\mathcal{I}$.

Example: for $p = \phi$ the standard normal pdf,

$$\mathcal{T}_X f(x) = \frac{(f(x)\phi(x))'}{\phi(x)} = f'(x) + \frac{\phi'(x)}{\phi(x)} f(x) = f'(x) - xf(x)$$

A useful relationship

If X and Y have the same support then

$$Y \overset{\mathcal{D}}{=} X \text{ if and only if } (\mathcal{T}_Y, \mathcal{F}(Y)) = (\mathcal{T}_X, \mathcal{F}(X));$$

see *Ley and Swan (2013)* for more details.

Moreover, for all $f \in \mathcal{F}(X)$,

$$\begin{aligned} \mathbb{E}\left[g'(X)f(X)\right] &= \int g'(x)f(x)p(x)dx \\ &= -\int g(x)\frac{(fp)'(x)}{p(x)}p(x)dx \\ &= -\mathbb{E}\left[g(X)\mathcal{T}_X f(X)\right] \end{aligned}$$

for all differentiable functions g such that

- $\int (gfp)' dx = 0$, and
- $\int |g'fp| dx < \infty$.

We collect all such g in a class $dom((\cdot)', X, f)$.

Stein operators as skew-adjoints

In all generality we get the following :

$$\mathbb{E}\left[\mathcal{D}g(X)f(x)\right] = -\mathbb{E}\left[g(X)\mathcal{T}_X f(X)\right]$$

for all $f \in \mathcal{F}(X)$ and all $g \in dom(\mathcal{D}, f, X)$.

The canonical Stein operator for X is thus, in some sense, skew adjoint to $\mathcal{D}$ with respect to integration in X.

With this definition, up to the choice of $\mathcal{D}$, there is therefore only one *canonical* Stein operator.

Theorem

Characterization :

$$Y \overset{\mathcal{D}}{=} X \text{ if and only if } \mathbb{E}\left[f(Y)\mathcal{D}g(Y)\right] = -\mathbb{E}\left[g(Y)\mathcal{T}_X f(Y)\right]$$

for all $f \in \mathcal{F}(X)$ *and for all* $g \in dom(\mathcal{D}, X, f)$ *.*

Stein characterisations

Let Y be continuous with pdf q, and same support as X with pdf p.

- Suppose that $\frac{q}{p}$ is differentiable and fix $g \in \cap_{f \in \mathcal{F}(X)} dom((\cdot)', X, f)$ such that g is X-a.s. never 0 and $g\frac{q}{p}$ is differentiable.

  Then

  $$Y \overset{\mathcal{D}}{=} X \text{ if and only if } \mathbb{E}\left[f(Y)g'(Y)\right] = -\mathbb{E}\left[g(Y)\mathcal{T}_X f(Y)\right]$$

  for all $f \in \mathcal{F}(X)$.
- Let $f \in \mathcal{F}(X)$ be X-a.s. never zero and assume that $dom((\cdot)', X, f)$ is dense in $L^1(X)$.

  Then

  $$Y \overset{\mathcal{D}}{=} X \text{ if and only if } \mathbb{E}\left[f(Y)g'(Y)\right] = -\mathbb{E}\left[g(Y)\mathcal{T}_X f(Y)\right]$$

  for all $g \in dom((\cdot)', X, f)$.

Some special cases

Take $g \equiv 1$ (this is always permitted) to obtain the Stein characterization

$$Y \overset{\mathcal{D}}{=} X \text{ if and only if } \mathbb{E}\left[\mathcal{T}_X f(Y)\right] = 0 \text{ for all } f \in \mathcal{F}(X).$$

Example: for p the standard normal pdf,

$$\mathbb{E}\mathcal{T}_X f(Y) = \mathbb{E}\left[f'(Y) - Yf(Y)\right] = 0.$$

If $f \equiv 1$ is in $\mathcal{F}(X)$ then we obtain the Stein characterization

$$Y \overset{\mathcal{D}}{=} X \Longleftrightarrow \mathbb{E}[g'(Y)] = -E\left[\frac{p'(Y)}{p(Y)} g(Y)\right] \text{ for all } g \in dom((\cdot)', X, 1).$$

Example: for p the standard normal pdf,

$$\mathbb{E}g'(Y) = \mathbb{E}Yg(Y).$$

The inverse Stein operator

For $h \in \mathcal{F}^{(0)}(X) = \left\{ h : \mathbb{R} \to \mathbb{R} \text{ such that } E\left[h(X)\right] = 0 \right\}$ we define the *inverse Stein operator* $\mathcal{T}_X^{-1} : \mathcal{F}^{(0)}(X) \to \mathcal{F}(X)$ as

$$\mathcal{T}_X^{-1} h(x) = \frac{1}{p(x)} \int_a^x p(y)h(y)dy = -\frac{1}{p(x)} \int_x^b p(y)h(y)dy.$$

Example: for p the standard normal pdf, and h with standard normal mean 0,

$$\mathcal{T}_X^{-1} h(x) = -e^{\frac{x^2}{2}} \int_x^\infty h(y) e^{-\frac{y^2}{2}} dy.$$

In all generality, this operator is such that

$$\mathbb{E}\left[g(X)h(X)\right] = -\mathbb{E}\left[\mathcal{D}g(X)\mathcal{T}_X^{-1}h(X)\right]$$

for all $h \in \mathcal{F}^{(0)}(X)$ and all $g \in dom(\mathcal{D})$.

Product rule

The Stein operator satisfies the product rule

$$\mathcal{T}_X(fg(\cdot + l)) = f\mathcal{D}g + g\mathcal{T}_X f$$

with $f \in \mathcal{F}(X)$ and $g \in dom(\mathcal{D}, X, f)$.

We introduce the class

$$dom(\mathcal{D}, X) = \bigcap_{f \in \mathcal{F}(X)} dom(\mathcal{D}, X, f).$$

The classes $dom(\mathcal{D}, X)$ and $\mathcal{F}(X)$ are defined with minimal conditions, and often quite complicated to write out explicitly.

Stein equations

Let $h \in L^1(X)$. We introduce the *Stein equation for the target X*

$$h(x) - \mathbb{E}h(X) = f(x)g'(x) + g(x)\mathcal{T}_X f(x),$$

whose solutions are *pairs* of functions (f, g) such that $f \in \mathcal{F}(X)$, $g \in dom((\cdot)', X, f)$ and

$$fg = \mathcal{T}_X^{-1}(h - \mathbb{E}_p h).$$

Although fg is unique, the individual f and g are not (just consider multiplication by constants).

Example: for X standard normal,

$$h(x) - \mathbb{E}h(X) = f(x)g'(x) + g(x)(f'(x) - xf(x)).$$

Special Stein operators

Our general Stein operator acts on pairs of functions (f, g);

$$\mathcal{A}(f, g)(x) = \mathcal{T}_X(fg)(x) = f(x)g'(x) + g(x)\mathcal{T}_X f(x).$$

Several operators are immediate to obtain from here :

- Fix a differentiable g and use

  $$\mathcal{A}_X f = \mathcal{T}_X\,(fg) = fg' + g\mathcal{T}_X f$$

  with $f \in \mathcal{F}_{\mathcal{A}}(X) \subset \mathcal{F}(X)$.
- Fix $f = c \in \mathcal{F}(X)$ and use

  $$\mathcal{A}_X g(x) = c(x)g'(x) + g(x)\mathcal{T}_X c(x)$$

  with $g \in dom((\cdot)', X, c)$. Sometimes we call this the c-operator (see *Goldstein and Reinert (2013)*).

Infinitely many other options are available.

The score function

Suppose that X is such that the constant function $1 \in \mathcal{F}(X)$ (this is no small assumption).

Then taking $c = 1$ in

$$\mathcal{A}_X g(x) = c(x)g'(x) + g(x)\mathcal{T}_X c(x).$$

we get

$$\mathcal{A}_X g(x) = g'(x) + g(x)\rho(x)$$

with

$$\rho(x) = \mathcal{T}_X 1(x) = \frac{p'(x)}{p(x)}$$

the score function of X; indeed

$$\mathbb{E}\left[g(X)\mathcal{T}_X 1(X)\right] = -\mathbb{E}\left[g'(X)\right]$$

for all $g \in dom((\cdot)', X, 1)$.

The Stein kernel

Suppose that X has finite mean ν.

Take $c = \mathcal{T}_X^{-1}(\nu - Id)$ with Id the identity function in

$$\mathcal{A}_X g(x) = c(x)g'(x) + g(x)\mathcal{T}_X c(x)$$

to get

$$\mathcal{A}_X g(x) = \tau(x)g'(x) + (\nu - x)g(x).$$

with

$$\tau = \mathcal{T}_X^{-1}(\nu - Id)$$

the Stein kernel of X; indeed

$$\mathbb{E}\left[g'(x)\mathcal{T}_X^{-1}(\nu - Id)\right] = \mathbb{E}\left[g(X)(X - \nu)\right]$$

for all $g \in dom((\cdot)', X, \tau)$.

In general

There are now Stein operators in the literature for dozens of types of distributions; some operators are first order,

$$\mathcal{T}_{\text{centeredgamma}} f(x) = 2(x+\nu) f'(x) - x f(x)$$

some are second order

$$\mathcal{T}_{\text{variancegamma}} f(x) = x f''(x) + (2\nu + 1) f'(x) - x f(x),$$

some are even more complicated.

All these can be written in the form

$$\mathcal{A}_X f = \mathcal{T}_X(A(f))$$

for A a suitable transformation of the test function, and we call the process leading to specific expressions of the operators *standardizations of the test functions*.

Example: Normal

In the example of a $\mathcal{N}(0, \sigma^2)$ random variable, our operator translates to

$$\mathcal{T}_N f(x) = f'(x) - \frac{1}{\sigma^2} x f(x)$$

which contrasts with

$$\sigma^2 f'(x) - x f(x),$$

the standard Stein operator for this case. The score function is $-\frac{x}{\sigma^2}$. We compute the Stein kernel

$$\tau(x) = \sigma^2.$$

The c-Stein operator is the standard operator.

Example: Beta

Consider beta distributions with density

$$p(x;\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}\mathbf{1}_{\{x\in[0,1]\}}.$$

Here

$$\mathcal{T}_B f(x) = -f'(x) - \frac{1}{x(1-x)} f(x)((\alpha-1)x - (\beta-1)(1-x)).$$

The standard Stein operator for this case is

$$\mathcal{A}f(x) = x(1-x)f'(x) + (\alpha(1-x) - \beta x)f(x),$$

see *Doebler (2012)*. The score function, defined when $\alpha > 1$ and $\beta > 1$, is

$$\rho(x) = \frac{\alpha}{x} - \frac{\beta-1}{x-1}.$$

The beta Stein kernel is

$$\tau(x) = \frac{x(1-x)}{\alpha+\beta}.$$

Outline

1 Stein's method

2 A canonical Stein operator

3 Distances between expectations

4 Multivariate extension

Approximate computation of expectations

Theorem

Let h be a function such that $\mathbb{E}_i|h| < \infty$ for $i = 1, 2$.

- *Let (f, g) solve the X_1-Stein equation for h. Then*

  $$\mathbb{E}_2 h - \mathbb{E}_1 h = \mathbb{E}\left[f(X_2)\mathcal{D}_1^\star g(X_2) - g(X_2)\mathcal{T}_1 f(X_2)\right].$$

- *Fix $f_1 \in \mathcal{F}_1$ and define $g_h := \frac{1}{f_1}\mathcal{T}_1^{-1}(h - \mathbb{E}_1 h)$. Then*

  $$\mathbb{E}_2 h - \mathbb{E}_1 h = \mathbb{E}_2\left[f_1\mathcal{D}_1^\star g_h - f_2\mathcal{D}_2^\star g_h + g_h\mathcal{T}_1 f_1 - g_h\mathcal{T}_2 f_2\right]$$

  for all $f_2 \in \mathcal{F}_2$ such that $g_h \in dom(\mathcal{D}_2, X_2, f_2)$.

- *Fix $g_1 \in dom(\mathcal{D}_1, X_1)$ and define $f_h := \frac{1}{g_1}\mathcal{T}_1^{-1}(h - \mathbb{E}_1 h)$. Then*

  $$\mathbb{E}_2 h - \mathbb{E}_1 h = \mathbb{E}_2\left[f_h\mathcal{D}_1^\star g_1 - f_h\mathcal{D}_2^\star g_2 + g_1\mathcal{T}_1 f_h - g_2\mathcal{T}_2 f_h\right].$$

  if $f_h \in \mathcal{F}_1 \cap \mathcal{F}_2$, for all $g_2 \in dom(\mathcal{D}_2, X_2)$.

A Corollary

Corollary

Let $\mathcal{H}$ be any class of functions $h : \mathbb{R} \to \mathbb{R}$ such that $E_i|h| < \infty$ for $i = 1, 2$.

1. *For all c such that [mild assumptions]*

$$\sup_{h \in \mathcal{H}} |E_1 h - E_2 h| \leq \kappa_{\mathcal{H},1} E_2 |\mathcal{T}_1 c - \mathcal{T}_2 c| \tag{2}$$

   with $\kappa_{\mathcal{H},1} = \sup_{h \in \mathcal{H}} \|(1/c)\, \mathcal{T}_1^{-1}(h - E_1 h)\|_\infty$.

2. *For all g such that [mild assumptions]*

$$\sup_{h \in \mathcal{H}} |E_1 h - E_2 h| \leq \kappa_{\mathcal{H},2} E_2 |\mathcal{T}_1^{-1}(g) - \mathcal{T}_2^{-1}(g)| \tag{3}$$

   with $\kappa_{\mathcal{H},2} = \sup_{h \in \mathcal{H}} \|((1/\mathcal{T}_1^{-1} g)\, \mathcal{T}_1^{-1}(h - E_1 h))'\|_\infty$.

Binomial approximation

Let $X \sim \mathrm{Bin}(n,p)$ and $W = \sum_{i=1}^n X_i$ with $X_i \sim \mathrm{Bin}(1,p_i)$, $i = 1,\ldots,n$, and $np = \sum_{i=1}^n p_i$. Use $\mathcal{D} = \Delta^+$, the forward difference.

It is easy to show

$$\tau_X(x) = (1-p)x \text{ and } \tau_{X_i}(x) = (1-p_i)x.$$

Then

$$|\mathbb{E}h(X) - \mathbb{E}h(W)| \leq ||\mathcal{D}g_h||_\infty \sum_{i=1}^n |p_i - p|p_i.$$

Similarly comparing “scores” we get

$$|\mathbb{E}h(X) - \mathbb{E}h(W)| \leq \frac{2\|g_h\|_\infty}{1-p} \sum_{i=1}^n |p - p_i|p_i.$$

Example: Distance between Gaussians

For $X_i \sim \mathcal{N}(0, \sigma_i^2), i = 1, 2.$

Using $\tau_i(x) = \sigma_i^2$ we get

$$d_{\mathrm{TV}}(X_1, X_2) \leq 2\frac{|\sigma_1^2 - \sigma_2^2|}{\sigma_2^2}.$$

Using $\rho_i(x) = -x/\sigma_i^2$ we get

$$d_{\mathrm{TV}}(X_1, X_2) \leq \sigma_1\sqrt{\frac{\pi}{2}}\mathbb{E}|X_2|\left|\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}\right| = \frac{\left|\sigma_1^2 - \sigma_2^2\right|}{\sigma_1\sigma_2}.$$

If $\sigma_1 < \frac{\sigma_2}{2}$ then this bound beats the first bound.

From Student to Gauss

Set $X_1 = Z$ standard Gaussian and $X_2 = W_\nu$ a Student t random variable with $\nu > 2$ degrees of freedom.

We have $\tau_1 = 1$ and $\tau_2(x) = \frac{x^2+\nu}{\nu-1}$ which yields

$$d_{\mathrm{TV}}(Z, W_\nu) \leq 2\mathbb{E}\left|\frac{W_\nu^2+\nu}{\nu-1} - 1\right| \leq \frac{4}{\nu-2}.$$

We can also use score functions to get

$$d_{\mathrm{TV}}(Z, W_\nu) \leq \sqrt{\frac{\pi}{2}} \frac{-2 + 8\left(\frac{\nu}{1+\nu}\right)^{(1+\nu)/2}}{(\nu-1)\sqrt{\nu}B(\nu/2, 1/2)},$$

which is of the same order, with a better constant.

Outline

1. Stein's method

2. A canonical Stein operator

3. Distances between expectations

4. Multivariate extension

A canonical Stein operator

Let $X \sim p : \mathbb{R}^d \to \mathbb{R}$ a differentiable distribution on $\mathbb{R}^d$.

Given a d-dimensional vector field F we write div the usual divergence operator $\mathrm{div} F(x) = \sum_{i=1}^d \partial_i F(x)$.

The canonical Stein operator of p is

$$\mathcal{T}_X F = \frac{\mathrm{div}(Fp)}{p}$$

acting on differentiable vector fields $F : \mathbb{R}^d \to \mathbb{R}^d$ or $d \times d$ matrices F.

We define $\mathcal{F}(X)$ as the class of all vector fields for which $E\left[\mathcal{T}_X F(X)\right] = 0$.

Literature review

Many authors have considered (unwittingly) the operator $\mathcal{T}_X F$:

- Landsman and Neslehova and coauthors (2010:2014) in the context of elliptical distributions;
- Chatterjee and Meckes (2008) and Reinert and Röllin (2009) for multivariate normal
- Brown et al. (2006) in the context of the heat equation
- Nourdin, Peccati and Swan (2014:2014) specifically via Stein matrices
- Artstein et al (2004:2014) with variational considerations in mind.

We can use these results to propose a general version of Stein's method also in arbitrary (finite) dimension.

Stein operator for multivariate Gaussian

For instance, taking $F = G\nabla f$ with G a symmetric $d \times d$ matrix then (1) becomes

$$\begin{aligned}\mathcal{T}_X f &= \sum_{i,j=1}^{d} \partial_i (G_{ij} \partial_j f) + \sum_{i,j=1}^{d} G_{ij} \partial_j f \frac{\partial_i p}{p} \\ &= \nabla^t \cdot (G \nabla f) + \nabla f^t G \nabla \log p\end{aligned}$$

In particular, if $p = \phi$ is the density of a $\mathcal{N}_d(0, \Sigma)$ random vector then

$$\nabla \log p(x) = -\Sigma^{-1} x$$

so that, taking $G = \Sigma$, the operator becomes

$$\mathcal{T}_X f(x) = \sum_{i,j=1}^{d} \sigma_{ij} \partial_{ij} f(x) + (\nabla f(x))' x,$$

which one recognizes as the standard operator for the Gaussian,

Stein operators

In all generality, given three functions $G \in \mathbb{R}^{d \times d}$, $g \in \mathbb{R}^d$ and $f \in \mathbb{R}$ we introduce the family of scalar operators

$$\mathcal{A}_X(G, f, g) = \mathcal{T}_X(Ggf) = \mathrm{div}(Gg)f + g^t G^t \frac{\nabla(fp)}{p},$$

and vector operators

$$\mathcal{A}_X(G, f) = \mathcal{T}_X(Gf) = \mathrm{div}(G)f + G^t \mathcal{T}_X(f),$$

different choices of G, g and f will lead to different operators.

These generalize in particular the Stein kernels (now matrices) and score functions (now vectors) to arbitrary dimensions.

We even have explicit inverses under certain circumstances.

Conclusion

Further reading :

- Approximate computation of expectations : a canonical Stein operator (with C. Ley and G. Reinert).
- A handbook of Stein operators (with C. Döbler, R. Gaunt, C. Ley and G. Reinert).
- Integration by parts and representation of information functionals (with G. Peccati and I. Nourdin). Proceedings of the 2014 IEEE International Symposium on Information Theory (2014)
- Entropy and the fourth moment phenomenon (with G. Peccati and I. Nourdin). The Journal of Functional Analysis 266 (5), 3170-3207 (2014)