Normal approximation by Stein's method

Normal approximation by Stein's method **0**
1 Characterization of distributions 0
2 Facts about the tails of the normal 2
3 Facts about the normal characterizing d.e. in $\mathbb{R}^1$ 2
4 Exchangeable pairs (IMS Lect.1) 5
5 Smoothing 8
6 OLD version of Exchangeable pairs (IMS Lect.1) 11
7 Problems 12

- Stein86 =Stein (1986)
- Stein72 = Stein (1972). This paper appears to be the first published version of the normal approximation method.
- BR76 = Bhattacharya and Ranga Rao (1976)
- RR96 = Rinott and Rotar (1996) and RR97= Rinott and Rotar (1997)
- MT = Meckes (2006) Ph.D. thesis
- Alea = Chatterjee and Meckes (2008)
- Luminy = Meckes (2009), Luminy paper
- PTTM = Pollard (2029) and UGMTP = Pollard (2001). Timothy and Sekhar should have read-only access to the PTTM directory on Dropbox.
- $\gamma = N(0,1)$ with density $\phi(x)$ and $\Phi(x) = \gamma(-\infty, x] = 1 - \bar{\Phi}(x)$.
- $\mathfrak{m}$ = lebesgue measure on $\mathcal{B}(\mathbb{R})$.

1 Characterization of distributions

S:characterize

The Stein method for normal approximation in one dimension is based on a differential equation. This part of the method also works for approximation of any probability measure P_0 on $\mathcal{B}(\mathbb{R})$ with a smooth density $p_0(x) = e^{-g(x)}$ with respect to Lebesgue measure. Note that $\dot{p}_0 = -\dot{g}p_0$.

In this Section I am not being very careful about the necessary regularity assumptions. For example, I seem to need $g(x) \to \infty$ as $x \to -\infty$.

Suppose h is a member of $\mathcal{L}^1(P_0)$ for which there is a smooth function f defined by the relation

\E@ Stein-d.e. <1>

$$\dot{f}(x) - f(x)\dot{g}(x) = H(x) := h(x) - P_0 h.$$

Multiply through by $g(x)$ to get

$$\frac{d}{dx}\left(p_0(x)f(x)\right) = p_0(x)\dot{f}(x) - \dot{g}(x)p_0(x)f(x) = p_0(x)H(x).$$

Integrate with respect to $\mathfrak{m}$ over the interval $(-\infty, w]$:

\E@ f.soln <2>

$$p_0(w)f(w) = \int_{-\infty}^{w} p_0(x)H(x)\,dx$$

This equality shows that $f(x)$ is uniquely determined by h, at least on the set $\{x : p_0(x) > 0\}$. Thus we could think of the left-hand side of <1> as defining a map $\mathcal{K}$ from a set $\mathbb{H}$ of functions on $\mathbb{R}$ back into functions on $\mathbb{R}$:

\E@ kk.def <3>

$$\mathcal{K}(x,h) := (\mathcal{K}h)(x) := \dot{f}(x) - f(x)\dot{g}(x) \qquad \text{for each } h \in \mathbb{H}.$$

Now suppose P is another probability measure on $\mathcal{B}(\mathbb{R})$ for which

$$Ph - P_0 h = PH = P^x\mathcal{K}(x,h) = 0 \qquad \text{for each } h \in \mathbb{H}.$$

Then we have $Ph = P_0 h$ for each h in $\mathbb{H}$. If $\mathbb{H}$ is a large enough subset of $\mathcal{L}^1(P)$ to uniquely determine P_0 then we deduce that $P = P_0$, as measures on $\mathcal{B}(\mathbb{R})$. Here I am thinking about examples like $\mathbb{H}$ equal to $\{h : \|h\|_{BL} \le 1\}$ or to the smaller set $\{h \in \mathcal{C}^\infty(\mathbb{R}) : \|\dot{h}\| \le 1\}$.

For the purposes of showing that $P \approx P_0$ we need some sort of continuity property to translate $P(\mathcal{K}h) \approx 0$ into $PH \approx 0$, together with a way of interpreting an approximation like $Ph \approx P_0 h$ for h in $\mathbb{H}$. Some ϵ's and δ's might turn this idea into a useful inequality.

Remark. These vague ideas suggest to me that the $\mathcal{K}$ need not be defined by a differential equation. A family of operators came to mind for higher dimensions:

$$(\mathcal{K}_r h)(x) = \frac{\mathfrak{m}^t H(t)\{t \in B(x,r)\}}{\mathfrak{m}B(x,r)} \qquad \text{where } H(t) = h(t) - P_0 h.$$

Or maybe it would be better to replace the indicator function of a ball by some smooth function ψ with compact support and define

$$(\mathcal{K}_r h)(x) = C_r(x)^{-1}\mathfrak{m}^t\psi(x+rt)H(t)? \approx ?r(???)$$

where $\{C_r : r > 0\}$ is a family of normalizing functions. I would hope that $C_r(x) \approx c_0(x) + c_1(x)r$ for small r. Maybe some sort of limit as $r \to 0$ could be involved. The invariance of $\mathfrak{m}$ under translations should then play a role. For small r we would have

$$(\mathcal{K}_r h)(x)\,? \approx ?(c_0(x) + c_1(x)r)^{-1}\left(r(???)\right).$$

Something wrong there. I was hoping for something almost linear in r.

2 Facts about the tails of the normal

S:tails

The following facts are proved in Section 7, which contains Problems from PTTM Chapter 3.

(Ni) The function $\rho(x) := \phi(x)/\bar{\Phi}(x)$ is convex, strictly positive and strictly increasing on $\mathbb{R}$, with $r(0) = \rho(0) = \sqrt{2/\pi} \approx 0.798$.

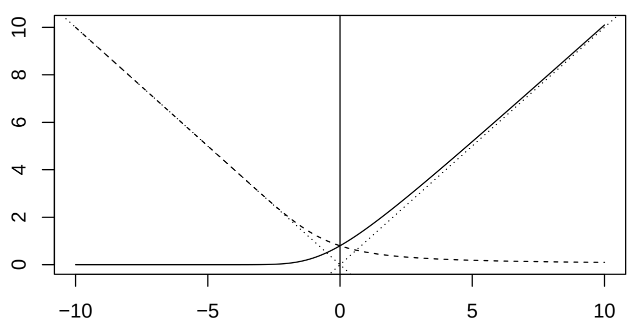
(Nii) The function $r(x) := \rho(x) - x$ is strictly positive, convex and decreasing, with and $\rho(x)r(x) < 1$ for all x.

(Niii) From the literature:

$$(3x + \sqrt{x^2+8})/4 < \rho(x) \le \left(x + \sqrt{x^2+4}\right)/2 \qquad \text{for all } x \in \mathbb{R}.$$

I don't think we need such exquisite detail.

The solid line in the picture shows $\rho(\cdot)$ and the dashed line shows $r(\cdot)$.



For my purposes, the representations

$$\Phi(t) = \phi(t)/\rho(t) \quad \text{AND} \quad \Phi(t) = \bar{\Phi}(-t) = \phi(t)/\rho(-t)$$

will be useful.

3 Facts about the normal characterizing d.e. in $\mathbb{R}^1$

S:de-facts

The method of exchangeable pairs seems to come down to a Taylor expansion with good bounds on remainder terms.

Compare with Stein86 pages 22–28. In particular, his Lemma 3 on page 25 gave a way to bound various derivatives of the f defined by <1>.

Stein.p25 <4> **Lemma.** *Suppose f is a smooth function for which*

\E@ Stein.normal <5>
$$\dot{f}(x) - xf(x) = H(x) := h(x) - \gamma h \qquad \text{for each } x.$$

where h is absolutely continuous with almost sure derivative $\dot{h}$ for which $\|\dot{h}\|_\infty$ is finite. Then:

(i) $\|f\|_\infty \le \sqrt{\pi/2}\|H\|_\infty$.

(ii) $\|\dot f\|_\infty \le 2\|H\|_\infty$.

(iii) $\|\ddot f\|_\infty \le 2\|\dot h\|_\infty$.

Remarks.

(a) Absolute continuity allows for things like h being of the form $\max(h_1, h_2)$ with $h_i \in \mathcal{C}^1(\mathbb{R})$. We need $\dot h$ to be locally Lebesgue integrable with

$$h(b) - h(a) = \int_a^b \dot h(r)\,dr \qquad \text{for all } -\infty < a < b < \infty.$$

(b) If h is differentiable at x then the function f is twice differentiable:

\E@ fdotdot <6>

$$\begin{aligned}\ddot f(x) &= \frac{d}{dx}\left(xf(x) + H(x)\right) = x\dot f(x) + f(x) + \dot H(x)\\ &= (1+x^2)f(x) + xH(x) + \dot h(x).\end{aligned}$$

(c) For $h \in \mathcal{L}^1(\gamma)$,

\E@ solution <7>

$$\phi(x)f(x) = \gamma^r H(r)\{r \le x\} = -\gamma^r H(r)\{r \ge x\}.$$

The $\ge$ and $\le$ could be replaced by strict inequalities, because γ has no atoms.

Proof of (i)

Write C_0 for $\|H\|_\infty$. From (c) we have

$$\begin{aligned}\phi(x)|f(x)| &\le C_0\gamma(-\infty, x] = C_0\Phi(x) = C_0\phi(x)/\rho(-x),\\ \phi(x)|f(x)| &\le C_0\gamma[x,\infty) = C_0\phi(x)/\rho(x).\end{aligned}$$

Thus $|f(x)| \le C_0/\max[\rho(x), \rho(-x)] \le C_0/\sqrt{2/\pi}$.

Proof of (ii)

$$|\dot f(x)| \le |xf(x)| + |H(x)| \le \frac{C_0|x|}{\max[\rho(x), \rho(-x)]} + C_0.$$

A sloppy proof of (iii)

Here is a sloppy transcription of Stein's argument that gives (iii) with 2 replaced by some universal constant c.

Homogeneity of the defining equation <5> lets me assume, without loss of generality, that $\|\dot h\|_\infty = 1$.

The argument is mostly a matter of expressing $\ddot f$ using integrals of $\dot h$ then appealing to classical tail bounds (see Section 2) for the normal.

I think it suffices to consider $\ddot f(x)$ for $x \ge 0$. The bounds for negative x should come from the analogous argument with H replaced by $-H(-x)$.

First get an integral representation for H:

\E@ H.rep <8>

$$
\begin{aligned}
H(r) &= \gamma^s\left(h(r)-h(s)\right)\\
&= \gamma^s \mathfrak{m}^t \dot{h}(t)\left(\{s<t<r\}-\{r<t<s\}\right)\\
&= \mathfrak{m}^t \dot{h}(t)\left[\{t<r\}\Phi(t)-\{r<t\}\bar{\Phi}(t)\right] \qquad \text{by Fubini}\\
&= \int_{-\infty}^{r} \dot{h}(t)\Phi(t)\,dt - \int_r^{\infty}\dot{h}(t)\bar{\Phi}(t)\,dt.
\end{aligned}
$$

With $r=x$, this agrees with Stein86 page 27, equation (56).

Then get a representation for f:

$$
\begin{aligned}
\phi(x)f(x) &= \gamma^r H(r)\{r<x\}\\
&= \mathfrak{m}^t \dot{h}(t)\gamma^r\{r<x\}\left[\{t<r\}\Phi(t)-\{r<t\}\bar{\Phi}(t)\right]\\
&= \mathfrak{m}^t \dot{h}(t)\left[\{t<x\}\left(\Phi(x)-\Phi(t)\right)\Phi(t)-\Phi(t\wedge x)\bar{\Phi}(t)\right]
\end{aligned}
$$

Rewrite $\Phi(x)-\Phi(t)$ as $\bar{\Phi}(t)-\bar{\Phi}(x)$ and $\Phi(t\wedge x)$ as $\{t<x\}\Phi(t)+\{t\ge x\}\Phi(x)$, then cancel out as $\{t<x\}\bar{\Phi}(t)\Phi(t)$ to conclude that

\E@ f.rep <9>

$$
\phi(x)f(x) = -\mathfrak{m}^t \dot{h}(t)\left[\{t<x\}\bar{\Phi}(x)\Phi(t)+\{t>x\}\Phi(x)\bar{\Phi}(t)\right]
$$

From <6> and the representations of H and f we get

\E@ fdotdot.rep <10>

$$
\begin{aligned}
\ddot{f}(x)-\dot{h}(x) &= (1+x^2)f(x)+xH(x)\\
&= g_1(x)\int_{-\infty}^{x}\dot{h}(t)\Phi(t)\,dt + g_2(x)\int_x^{\infty}\dot{h}(t)\bar{\Phi}(t)\,dt\\
\text{where}\qquad & g_1(x) := x-(1+x^2)\bar{\Phi}(x)/\phi(x)\\
& g_2(x) := -x-(1+x^2)\Phi(x)/\phi(x).
\end{aligned}
$$

Stein86 page 28:

$$
\begin{aligned}
\int_{-\infty}^{x}\Phi(t)\,dt &= x\Phi(x)+\phi(x)\\
\int_x^{\infty}\bar{\Phi}(t)\,dt &= -x\bar{\Phi}(x)+\phi(x)
\end{aligned}
$$

[Proof: In both cases, check equality of derivatives then note equality of limits as $x\to-\infty$ or $x\to\infty$.]

Here is a way to bound the g_1 contribution.

$$
\begin{aligned}
|g_1(x)\int_{-\infty}^{x}\dot{h}(t)\Phi(t)\,dt| &\le |x-(1+x^2)/\rho(x)|\int_x^{\infty}\bar{\Phi}(t)\,dt|\\
&= \frac{|x(x+r(x))-(1+x^2)|}{\rho(x)}\phi(x)|1-x\bar{\Phi}(x)/\phi(x)|\\
&\le \frac{1}{\rho(x)}\frac{e^{-x^2/2}}{\sqrt{2\pi}}\frac{r(x)}{\rho(x)} \le c.
\end{aligned}
$$

I hope the g_2 contribution can be handled similarly.

4 Exchangeable pairs (IMS Lect.1)

S:exch

Suppose (X,Y) is an exchangeable pair of real random variables with joint distribution $\mathbb{Q}^{x,y} = P^x K_x^y$. (That is, K_x is the conditional distribution of Y given $X = x$.)

Assume the setting of Lemma <4>: h is an absolutely continuous function with almost sure derivative $\dot{h}$ for which $\|\dot{h}\|_\infty$ is finite and f is a smooth function (actually twice differentiable, Lebesgue almost everywhere) for which

$$\dot{f}(x) - xf(x) = H(x) := h(x) - \gamma h \qquad \text{for each } x.$$

That is,

$$\phi(x)f(x) = \gamma^r H(r)\{r \le x\} = -\gamma^r H(r)\{r \ge x\}.$$

By Taylor, for each real δ,

\E@ Taylor.f <11>

$$\begin{aligned} f(x+\delta) - f(x) - \delta\dot{f}(x) &\\ &= \delta \int_0^1 \dot{f}(x+t\delta) - \dot{f}(x)\,dt \\ &= \delta^2 \int_0^1 \int_0^1 \{0 < s < t < 1\}\ddot{f}(x+s\delta)\,ds\,dt \\ &= \delta^2 \int_0^1 (1-s)\ddot{f}(x+s\delta)\,ds. \end{aligned}$$

Remark. I should check carefully that this version of Taylor holds if $\dot{f}$ is only absolutely continuous with almost sure derivative $\ddot{f}$.

If h is only piecewise continuous then $\dot{f}$ is only piecewise continuous. If h is smoother then f inherits higher derivatives. If we want to bound Wasserstein distances then we might get away with something like h differentiable with $\|\dot{h}\|_\infty \le 1$, as a way to approximate functions with $\|h\|_{\text{lip}} \le 1$. For the case where $h(x) = \{x \le w\} - \Phi(w)$, we only get one-times differentiabilty with a piecewise continuous derivative for f. That makes the approximation argument more delicate (Ho and Chen 1978, IMS Lect II).

I am suspicious of MT Lemma 1.3, page 4, where the analog of Lemma <4> was cited for g only bounded and continuous, even though the third assertion involved $\|\dot{g}\|_\infty$.

Stein.hdot *<12>* **Lemma.** *(cf. Stein86 pages 13–15 and 33–36)*
Suppose (X,Y) is an exchangeable pair with joint distribution $\mathbb{Q}^{x,y} = P^x K_x^y$ and $\mathbb{P}X = \mathbb{P}Y = 0$ and $\mathbb{P}X^2 = \mathbb{P}Y^2 = 1$. (That is, $Px = 0$ and $Px^2 = 1$.) Define $D(x,y) := y - x$.

cf.$\mathbb{P}(Y - X \mid X = x)$

(i) $K_x D = m(x) = -\lambda x + r_1(x)$ *and* $K_x D^2 = \tau^2(x) = 2\lambda + r_2(x)$.

(ii) $\dot{f}(x) - xf(x) = H(x) := h(x) - \gamma h$ *with* h *having a bounded, piecewise continuous derivative* $\dot{h}$ *(or* h *absolutely continuous?).*

Then

$$2\lambda|Ph - \gamma h| \le 2\|f\|_\infty P|r_1(x)| + \|\dot{f}\|_\infty P|r_2(x)| + \|\ddot{f}\|_\infty \mathbb{Q}|D|^3.$$

Proof. The function $F(x,y) := (y-x)\,(f(x)+f(y))$ is antisymmetric and

$$\begin{aligned} F(x,y) &:= (y-x)\,(f(x)+f(y)) \\ &= 2Df(x) + D^2\dot{f}(x) + D\left(f(x+D) - f(x) - D\dot{f}(x)\right) \\ &= 2Df(x) + D^2\dot{f}(x) + D^3\int_0^1 (1-s)\ddot{f}(x+sD)\,ds \end{aligned}$$

Antisymmetry gives

$$\begin{aligned} 0 &= \mathbb{Q}F(x,y) \\ &= P^x f(x) K_x^y D + P^x \dot{f}(x) K_x^y D^2 + R_3 \qquad \text{with } |R_3| \le \mathbb{Q}|D|^3\|\ddot{f}\|_\infty \\ &= 2Pm(x)f(x) + P\tau^2(x)\dot{f}(x) + R_3 \\ &= P\left(-2\lambda x f(x) + 2\lambda\dot{f}(x)\right) + P\left(2r_1(x)f(x) + r_2(x)\dot{f}(x)\right) + R_3 \\ &= 2\lambda(Ph - \gamma h) + \text{remainder}, \end{aligned}$$

where

$$\text{remainder} = P\left(2r_1(x)f(x) + r_2(x)\dot{f}(x)\right) + R_3$$

so that

$$|\text{remainder}| \le 2\|f\|_\infty P|r_1(x)| + \|\dot{f}\|_\infty P|r_2(x)| + \|\ddot{f}\|_\infty \mathbb{Q}|\Delta|^3.$$

Lemma <4> bounds $\|f\|_\infty$ and $\|\dot{f}\|_\infty$ by multiples of $\|H\|_\infty$ and and $\|\ddot{f}\|_\infty$ by a multiple of $\|\dot{h}\|_\infty$. □

MT-Thm2.1 <13> **Corollary.** *(MT page 19) For each* $\epsilon > 0$ *let* (X, Y_ϵ) *be an exchangeable pair with* $\mathbb{P}X = 0$ *and* $\mathbb{P}X^2 = 1$ *for which there are functions* $\alpha(X)$ *and* $\beta(X)$ *and* $\mathcal{E}(X)$ *in* $\mathcal{L}^1(\mathbb{P})$ *such that*

(i) $\mathbb{P}_X(Y_\epsilon - X) = -\lambda\epsilon^2 X + \alpha(X)o(\epsilon^2)$

Why not replace $\mathcal{E}$ by $\lambda\mathcal{E}$?

(ii) $\mathbb{P}_X\,(Y_\epsilon - X)^2 = \epsilon^2\,(2\lambda + \mathcal{E}(X)) + \beta(X)o(\epsilon^2)$

(iii) $\mathbb{P}|Y_\epsilon - X|^3 = o(\epsilon^2)$

Then $d_{TV}(P,\gamma) \le P|\mathcal{E}(x)|/\lambda$ *(?) where* P *denotes the distribution of* X*.*

Remark. I have doubts about the total variation distance in the assertion. It is true that $d_{TV}(P,\gamma) = \sup\{|PK - \gamma K| : K \text{ compact }\}$. Each compact K can be approximated by a sequence of continuous functions h_n with compact support: $1 \geq h_n \geq K$ and $h_n \downarrow K$. We have $\|h_n\|_\infty \leq 1$ but, even if h_n is smooth, we don't have control over $\|\dot{h}_n\|_\infty$. However,

$$\|P - \gamma\|_{BL} := \sup\{|Ph - \gamma h| : \|h\|_{\text{BL}} \leq 1\}$$

where, by the definition in UGMTP page 170, $\|h\|_{\text{BL}} = \|h\|_{\text{lip}} + 2\,\|h\|_\infty$. (Why did I bother with the 2?) It is true that an h for which $\|h\|_{\text{BL}} \leq 1$ can be well approximated by smooth functions h_n with $\|\dot{h}_n\|_2 \leq \|h_n\|_{\text{lip}}$ (See PTTM Problem 1 in Chapter 6. In $\mathbb{R}^1$ the ℓ^2 and ℓ^∞ norms of a function are the same, I think. Check.))

Proof. I'll interpret the assertion of the Corollary as an assertion about functions h with both $\|h\|_\infty \leq 1$ and $\|\dot{h}\|_\infty \leq 1$. Also I could assume h is infinitely differentiable and has compact support, if it helps.

I think Lemma <12>, with λ replaced by $\lambda\epsilon^2$ and $r_1(x) = \alpha(x)o(\epsilon^2)$ and $r_2(x) = \epsilon^2\mathcal{E}(x) + \beta(x)o(\epsilon^2)$ gives

$$0 = 2\lambda\epsilon^2\,(Ph - \gamma h) + \text{remainder} + R_3,$$

where $|R_3| \leq \mathbb{P}|Y_\epsilon - X|^3\|\ddot{f}\|_\infty$ and

$$\begin{aligned}\text{remainder} &= P\Big(f(x)2r_1(x) + r_2(x)\dot{f}(x)\Big)\\ &= P\Big(2f(x)\alpha(x)o(\epsilon^2) + \epsilon^2\mathcal{E}(x) + \beta(x)o(\epsilon^2)\dot{f}(x)\Big)\end{aligned}$$

Divide through by ϵ^2. so that

$$\begin{aligned}&|2\lambda(Ph - \gamma h) + P\mathcal{E}(x)|\\ &\quad\leq 2\|f\|_\infty o(1)P|\alpha(x)| + \|\dot{f}\|_\infty o(1)P|\beta(x)| + \|\ddot{f}\|_\infty o(1) = o(1).\end{aligned}$$

□

That is close to what MT asserted, but with d_{TV} replaced by d_{BL}.

Remark. (Probably wrong) For such a result to be useful we would need $P|\mathcal{E}|/\lambda$ to be small. Also, it appears to me that MT wanted the X and Y_ϵ to be random vectors, even though condition (ii) on page 19 was written with $(W_\epsilon - W)^2$. That suggests we should start from the vector analog of the expansion <11>.

Perhaps I should abandon MT and look for analogous results in Alea or Luminy.

5 Smoothing

S:smoothing

BR §11 derived a bunch of smoothing inequalities, which were used by RR96 Lemma 4.1 RR97 Lemma 4.1 with little in the way of proof.

BR stated results for $\mu - \nu$, for a bounded measure μ and a bounded signed measure ν on $\mathcal{B}(\mathbb{R}^k)$. The signed measure ν has a unique representation (the Jordan decomposition) as $\nu_\oplus - \nu_\ominus$, with $\nu_\oplus$ and $\nu_\ominus$ nonnegative measures with disjoint supports. The inequalities involved integrals with respect to $\nu_\oplus$.

I never did discover what μ and ν would be in specific cases. Instead I write λ for the signed measure $\mu - \nu = \mu + \nu_\ominus - \nu_\oplus$. By general results about the Jordan decomposition, we have $\lambda_\oplus \le \mu + \nu_\ominus$ and $\lambda_\ominus \le \nu_\oplus$. I think the BR bounds involving $\nu_\oplus$ are larger than my bounds involving $\lambda_\ominus$.

Even better, why not just assume that μ and ν are finite measures with disjoint supports right from the start? If we are interested in $\lambda = P - Q$ for probability measures P and Q then we could take $\mu = (P - Q)^+$ and $\nu = (P - Q)^- = (Q - P)^+$. The special case where $Q = N(0, I_k)$, or maybe $Q = N(0, V)$ for some covariance matrix, seems relevant for the Stein's theory.

My notation, based on BR

As usual, I write $B(x, r)$ for $\{y \in \mathbb{R}^K : |x - y|_2 < r\}$ and $B[x, r]$ for the corresponding closed ball.

For a given locally bounded, $\mathcal{B}(\mathbb{R}^k)$-measurable function h on $\mathbb{R}^k$ and $\delta > 0$, define

$$M_\delta(x, h) = \sup\{h(y) : |x - y|_2 < \delta\}.$$

The map $x \mapsto M_\delta(x, h)$ is lower semi-continuous and thus is borel-measurable.

Remark. Notice that $-\inf\{h(y) : |x - y|_2 < \delta\} = M_\delta(x, -h)$.

The BR results involve smoothing by convolution with a probability measure ρ on $\mathcal{B}(\mathbb{R}^k)$:

$$\rho \star \lambda(h) = \rho^x \lambda^y h(x + y) \qquad \text{for 'reasonable' } h.$$

Equivalently, by courtesy of Fubini,

$$\rho \star \lambda(h) = \lambda^x h_\rho(x) \qquad \text{where } h_\rho(x) := \rho^y h(x + y).$$

For example, under suitable integrability assumptions on $M_\epsilon(\cdot, h)$ we have $\rho \star \lambda(M_\epsilon) = \rho^x \lambda^y M_\epsilon(x + y, h)$. The main trick in the proof of the first Lemma consists of two inequalities involving $M_\epsilon(x + y, h) := \sup\{h(x + y + w) : |w|_2 < \epsilon\}$, which hold whenever $|x|_2 < \epsilon$:

\E@ Meps.lower <14> $$M_\epsilon(x + y, h) \ge h(x + y - x) = h(y);$$

\E@ Meps.upper <15> $$M_\epsilon(x + y, h) \le \sup\{h(y + w) : |w|_2 < 2\epsilon\} = M_{2\epsilon}(y, h).$$

For each probability measure ρ (K_ϵ in BR notation), define

$$g_\delta(f,\rho) := (\rho\star\lambda)^w M_\delta(w,f); \qquad \text{AND} \qquad \tau_\delta(f) := \nu^y\left(M_\delta(y,f) - f(y)\right).$$

BR11.1 <16> **Lemma.** *(≈ BR Lemma 11.1) Let ρ be a probability measure that concentrates on $B(0,\epsilon)$, for some $\epsilon > 0$, and $\lambda = \mu - \nu$, a difference of two finite measures. Assume some sort of integrability regarding h and M and m. Then*

$$\lambda h \le \rho\star\lambda(M_\epsilon) + \nu^y\big(M_{2\epsilon}(y,h) - h(y)\big) = g_\epsilon(h,\rho) + \tau_{2\epsilon}(h).$$

Proof. As $\rho\{x : |x|_2 < \epsilon\} = 1$, inequality <15> implies $\rho^x M_\epsilon(x+y) \le M_{2\epsilon}(y)$. For the asserted inequality, start with the $\rho\star\lambda(M_\epsilon)$ term:

$$\begin{aligned}\rho^x\lambda^y M_\epsilon(x+y) &\ge \rho^x\left(\mu^y h(y) - \nu M_\epsilon(x+y)\right) \qquad \text{by <14>}\\ &= \mu h - \nu^y\rho^x M_\epsilon(x+y)\\ &\ge \mu h - \nu h - \nu^y\big(M_{2\epsilon}(y) - h(y)\big).\end{aligned}$$

□ as asserted.

two-sided <17> **Corollary.**

□ $$|\lambda h| = \max\left(\lambda h, \lambda(-h)\right) \le \max\left(g_\epsilon(h,\rho), g_\epsilon(-h,\rho)\right) + \max\left(\tau_{2\epsilon}(h), \tau_{2\epsilon}(-h)\right)$$

BR11.4 <18> **Lemma.** *(≈ BR Lemma 11.4 and RR97 Lemma 4.1) Suppose P is a probability measure on $\mathcal{B}(\mathbb{R}^k)$ for which $\alpha := PB(0,\epsilon) > 1/2$ for some $\epsilon > 0$. Let λ be as in Lemma <16> and $\mathcal{H}$ be a uniformly bounded set of measurable functions on $\mathbb{R}^k$ for which*

(a) if $h \in \mathcal{H}$ then $h_\theta \in \mathcal{H}$, where $h_\theta(y) := h(\theta + y)$,

(b) if $h \in \mathcal{H}$ then $-h \in \mathcal{H}$.

Then

$$(2\alpha - 1)\sup\nolimits_{h\in\mathcal{H}} |\lambda h| \le \sup\nolimits_{\mathcal{H}} g_\epsilon(h,P) + \alpha\tau_{2\epsilon}(\mathcal{H}) + (1-\alpha)\tau_\epsilon(\mathcal{H}),$$

□ *where $\tau_\delta(\mathcal{H}) := \sup_{h\in\mathcal{H}} \tau_\delta(h)$.*

Proof. Define $\Delta := \sup_{\mathcal{H}} |\lambda h|$. By assumption (b),

$$\sup\nolimits_{\mathcal{H}} \lambda h = \Delta \qquad \text{AND} \qquad \inf\nolimits_{h\in\mathcal{H}} \lambda h = -\Delta.$$

The main idea is to decompose P as $\alpha\rho + \overline{\alpha}\,\overline{\rho}$ where $\rho := P(\cdot \mid B(0,\epsilon))$ and $\overline{\rho} := P(\cdot \mid B(0,\epsilon)^c)$ and $\overline{\alpha} := 1 - \alpha$. The argument for the ρ contribution will be essentially the same as for Lemma <16>: for each h in $\mathcal{H}$,

$$\rho^x\lambda^y M_\epsilon(x+y,h) \ge \lambda h - \tau_{2\epsilon}(h) \ge \lambda h - \tau_{2\epsilon}(\mathcal{H}).$$

For $\overline{\rho}$ use

$$M_\epsilon(x+y,h) \geq h(x+y) = h_x(y),$$
$$M_\epsilon(x+y,h) = \sup_{|w|<\epsilon} h(x+y+w) = M_\epsilon(y,h_x),$$

and the trivial fact that $\inf_x F(x) \leq \overline{\rho}^x F(x) \leq \sup_x F(x)$ to get

$$\begin{aligned}
\overline{\rho}^x \lambda^y M_\epsilon(x+y,h) &\geq \overline{\rho}^x \left(\mu^y h_x(y) - \nu^y M_\epsilon(x+y,h)\right) \\
&= \overline{\rho}^x \left[\lambda^y h_x(y) - \nu^y \left(M_\epsilon(y,h_x) - h_x(y)\right)\right] \\
&\geq \inf_x \lambda^y h_x(y) - \sup_x \nu^y \left(M_\epsilon(y,h_x) - h_x(y)\right) \\
&\geq -\Delta - \tau_\epsilon(\mathcal{H}).
\end{aligned}$$

Take a weighted average:

$$\begin{aligned}
(P \star \lambda)^w M_\epsilon(w,h) &= P^x \lambda^y M_\epsilon(x+y,h) \\
&= \alpha\lambda h - \alpha\tau_{2\epsilon}(\mathcal{H}) - \overline{\alpha}\left(\Delta + \tau_\epsilon(\mathcal{H})\right).
\end{aligned}$$

Then take a supremum over h in $\mathcal{H}$:

$$\sup_{h\in\mathcal{H}} g_\epsilon(h,P) \geq \alpha\Delta - \alpha\tau_{2\epsilon}(h) - \overline{\alpha}\left(\Delta + \tau_\epsilon(\mathcal{H})\right),$$

□ which rearrange to give the asserted upper bound for $(2\alpha-1)\Delta$.

6 OLD version of Exchangeable pairs (IMS Lect.1)

S:OLDep

I had a hard time deciphering Stein's explanations related to the diagram labeled (28) on Stein86 page 12.

Remark. I was very confused. Maybe better to skip the mess.

I think the idea is that we have a probability measure P on $\mathcal{B}(\mathbb{R})$ that we want to show is close, in some sense, to some other P_0. In the cases that seem of most interest $P_0 = \gamma := N(0,1)$.

The construction involves an exchangeable pair of random variables W, W', each with distribution P. If $\mathbb{Q}$ denotes the joint distribution of the pair then we can take the variables as the coordinate maps on $\mathbb{R}^2$: $W(x,y) = x$ and $W'(x,y) = y$. (Maybe it would be better to use (w, w') instead of (x,y), but then I would have a lot of pesky superscript w''s.) Both the marginals of $\mathbb{Q}$ are equal to P. Exchangeability means that

\E@ exch <19>

$$\mathbb{Q}F(x,y) = \mathbb{Q}F(y,x) \qquad \text{for all } F \text{ in } \mathcal{L}^1(\mathbb{Q}).$$

In particular, if $F(y,x) = -F(x,y)$ (antisymmetry) then $\mathbb{Q}F = 0$. Stein's $\mathcal{F}$ corresponds to the subspace of all antisymmetric members of $\mathcal{L}^1(\mathbb{Q})$.

The conditional distribution of W' given W corresponds to a markov kernel $\{K_w : w \in \mathbb{R}\}$ for which

$$\mathbb{Q}F(x,y) = P^w K_w^{w'} F(w, w').$$

The kernel also defines a linear map $F \mapsto K_w^{w'} F(w,w')$ from $\mathcal{L}^1(\mathbb{Q})$ into $\mathcal{L}^1(P)$. Stein wrote T for the restriction of this linear map to the set

\E@ antisymm <20>

$$\mathcal{F} := \{F \in \mathcal{L}^1(\mathbb{P}) : F(x,y) = -F(y,x)\}$$

of antisymmetric, $\mathbb{Q}$-integrable functions on $\mathbb{R}^2$. Consequently, $PTF = \mathbb{Q}F = 0$ for F in $\mathcal{F}$. That is, if $F \in \mathcal{F}$ then TF is an element of $\mathcal{L}^1(P)$ with zero P-expectation: $PTF = 0$ for $F \in \mathcal{F}$.

The hope appears to be (IMS p 10, Lemma 2) that for each f in (some subspace of) $\mathcal{L}^1(P)$ with $Pf = 0$ there should exist an antisymmetric F in $\mathcal{L}^1(\mathbb{Q})$ for which $f(x) = TF$.

Stein also assumed existence of another vector space $\mathcal{F}_0$ and a linear map $T_0 : \mathcal{F}_0 \to \mathcal{X}_0$, with $\mathcal{X}_0$ a subspace of $\mathcal{L}^1(P)$ and a linear(?) map $A : \mathcal{F}_0 \to \mathcal{F}$. (My A is Stein's α.) For normal approximation I think $\mathcal{F}_0$ can be taken as a subspace of $\mathcal{L}^1(P)$ consisting of suitably smooth (piecewise continuous first derivatives and ...?) functions and

$$(T_0 f)(x) = \dot{f}(x) - x f(x).$$

Compare with the definition of Stein's T_N on IMS page 19. The map A in this case is (I think) given by

\E@ A.def <21>

$$(Af)(x,y) = (y-x)\left(f(x) + f(y)\right).$$

He mentioned that $T(Af)$ should be thought of as an approximation to TF if $F = Af$.

The range of T_0 is a subspace $\mathcal{X}_0$ of $\mathcal{L}^1(P) \cap \mathcal{L}^1(P_0)$. He assumed existence of a sort of inverse $U_0 : \mathcal{X}_0 \to \mathcal{F}_0$, a linear map such that

\E@ U0 <22>
$$T_0 f = h(w) - P_0 h \qquad \text{if } f = U_0 h \text{ with } h \in \mathcal{X}_0.$$

In the normal case, $f = U_0 h$ is the solution to the differential equation $\dot{f}(x) - x f(x) = h - \gamma h$:

\E@ U_0.normal <23>
$$\begin{aligned}(U_0 h)(w) &= \exp(w^2/2) \int_{-\infty}^{w} (h(r) - \gamma h) \exp(-y^2/2)\, dy \\ &= -\exp(w^2/2) \int_{w}^{\infty} (h(r) - \gamma h) \exp(-y^2/2)\, dy\end{aligned}$$

The second form comes from the fact that $\gamma(h - \gamma h) = 0$. Stein (IMS page 14, Lemma 4) needed to assume that h is piecewise continuous with $h(w) = O(w^2)$ as $|w| \to \infty$.

7 Problems

S:Problems

The function $\phi(x)$ denotes the $N(0,1)$ density and $\bar{\Phi}(x) = \int_x^\infty \phi(t)\, dt$; the functions $\mathcal{R}(x)(\cdot)$ and $\rho(\cdot)$ and $r(\cdot)$ are defined on $\mathbb{R}$ by $1/\rho(x) = \mathcal{R}(x) = \bar{\Phi}(x)/\phi(x)$ and $r(x) = \rho(x) - x$.

P:Laplace [1] The following bounds are apparently due to Laplace. Each bound is of the form $p(1/x)$ with p a polynomial.

(i) Show that $p(1/x) > \mathcal{R}(x)$ for all $x > 0$ if

\E@ Mill.upper <24>
$$-\frac{d}{dt}\left(p(1/t)\phi(t)\right) > \phi(t) \qquad \text{for all } t > 0$$

and $p(1/x) < \mathcal{R}$ for all $x > 0$ if

\E@ Mill.lower <25>
$$-\frac{d}{dt}\left(p(1/t)\phi(t)\right) < \phi(t) \qquad \text{for all } t > 0$$

Hint: $\int_x^\infty$.

(ii) Show that<24> holds if and only if $p(t) + t^3 p'(t) > t$ for all $t > 0$. Characterize <25> by the reverse inequality.

(iii) Define a sequence of monomials by $\Delta_0(t) = t$ and $\Delta_k(t) = -t^3 \Delta'_{k-1}(t)$ for $k \geq 1$. Show that

$$\Delta_k(t) = (-1)^k a_k t^{2k+1} \qquad \text{where } a_k = 1 \times 3 \times \cdots \times (2k-1).$$

(iv) Define $p_k(t) = \sum_{i=0}^{k} \Delta_i(t)$. Show that $p_k(t) + t^3 p_k'(t) = t + \Delta_{k+1}(t)$.

(v) Conclude that $p_k(1/x) > \mathcal{R}(x) > p_{k+1}(1/x)$ for each even k. For example, for $k = 1$ and $k = 2$ we have, for all $x > 0$,

$$x^{-1} > \mathcal{R}(x) > x^{-1} - x^{-3}$$
$$x^{-1} - x^{-3} + 3x^{-5} > \mathcal{R}(x) > x^{-1} - x^{-3} + 3x^{-5} - 15x^{-7}$$

(vi) From the inequality for $k = 2$ deduce that $xr(x) \to 1$ as $x \to \infty$.

P:rho.facts [2] Here are the basic facts about $\rho(\cdot)$ and $r(\cdot)$.

(i) Suppose $Z \sim N(0, 1)$. Show that

$$\mathcal{R}(x) = \int_0^\infty \phi(x+t)/\phi(x)\,dt = \int_0^\infty e^{-xt-t^2/2}dt = \sqrt{\pi/2}\,\mathbb{P}e^{-x|Z|}.$$

(ii) Show that $L(x) := \log \mathcal{R}(x) = -\log \rho(x)$ is a strictly decreasing, convex function. Deduce that $\rho(x) = e^{-L(x)}$ is strictly increasing with $\rho(x) \to 0$ as $x \to -\infty$ and $\rho(x) \to \infty$ as $x \to \infty$. Also $\log \rho(x) = -x^2/2 - \log \sqrt{2\pi} - \log \bar{\Phi}(x)$ is concave.

(iii) Using the bounds from Problem [1], show that $\rho(x) > x$ and $r(x) > 0$ for all x. (Note that $\rho(x) > x$ is trivially true for $x \le 0$.) Also show that $xr(x) \to 1$ as $x \to \infty$.

(iv) Show that $\log(\bar{\Phi})$ has derivative $-\rho$ so that $\rho'(x)/\rho(x) = d\log\rho(x)/dx = \rho(x) - x$. That is, $\rho'(x) = \rho(x)r(x)$ for all x. From the concavity of $\log\rho$ deduce that r is a decreasing function.

(v) (Sampford, 1953) Show that $\gamma(x) := \rho''(x)/\rho(x) = 2r(x)^2 + xr(x) - 1$. Show that

$$\gamma'(x) = (4r(x) + x)\,r'(x) + r(x) = 2r(x)r'(x) + \rho(x)\gamma(x) < \rho(x)\gamma(x).$$

Argue as follows to show that $\gamma(x) > 0$ for all $x \in \mathbb{R}$, which implies that ρ is strictly convex. Suppose there were an x_0 for which $\gamma(x_0) \le 0$. By the preceding argument, $\gamma'(x_0)$ would be < 0. There would therefore be some $\delta > 0$ and $x_1 > x_0$ at which $\gamma(x_1) < -\delta$. By part (iii), $\gamma(x) \to 0$ as $x \to \infty$. For some finite K there would exist some $K > x_1$ for which $|\gamma(x)| < \delta$ for $x > K$. The difefrentiable function γ would achieve its minimum value on $[x_0, \infty)$ at some point x_2 in $[x_0, K]$ at which $0 > -\delta \ge \gamma(x_2)$ and $\gamma'(x_2) = 0$, a contradiction.

P:half [3] Show that $\bar{\Phi}(x) \le \frac{1}{2}e^{-x^2/2}$ for $x \ge 0$. Hint: From Problem [2](i) we have $\mathcal{R}(0) > \mathcal{R}(x)$.

P:Px [4] For each x show (by means of integration-by-parts) that $\int_x^\infty t\phi(t)\,dt = \phi(x)$ and $\int_x^\infty t^2\phi(t)\,dt = x\phi(x) + \bar{\Phi}(x)$. Let P_x be the probability measure with density $p_x(t) = \phi(t)\{t \ge x\}/\bar{\Phi}(x)$ with respect to Lebesgure measure on the real line. Show that the variance of P_x is $1 - \rho(x)r(x)$. Deduce that $\rho(x)r(x) < 1$ for all x.

P:Birnbaum [5] (Birnbaum, 1942). Use Cauchy-Schwarz and facts from Problem [4] to show that

$$\phi(x)^2 = \left(\int_x^\infty t\sqrt{\phi(t)}\sqrt{\phi(t)}\,dt\right)^2 \leq \left(x\phi(x) + \bar{\Phi}(x)\right)\bar{\Phi}(x).$$

Deduce that $1 \leq (x + \mathcal{R}(x))\,\mathcal{R}(x) = \left(\mathcal{R}(x) + \frac{1}{2}x\right)^2 - x^2/4$, which implies $\rho(x) \leq \left(x + \sqrt{x^2+4}\right)/2$.

P:Sampford [6] (Sampford, 1953) Let $\gamma(x) = 2r(x)^2 + xr(x) - 1$, as in Problem [2](v). Remember that $\gamma(x) > 0$ for all $x \in \mathbb{R}$. Argue that $r(x)$ cannot belong to the closed interval $I_x := \{t \in \mathbb{R} : 2t^2 + xt - 1 \leq 0\}$, which has endpoints $\left(-x \pm \sqrt{x^2+8}\right)/4$. Deduce that $r(x) > (-x + \sqrt{x^2+8})/4 = 2/(x + \sqrt{x^2+8}\,)$ and $\rho(x) > (3x + \sqrt{x^2+8})/4$. Note: $r(x) > 0$.

References

BhattacharyaRao76 Bhattacharya, R. N. and R. Ranga Rao (1976). *Normal Approximation and Asymptotic Expansions*. New York: Wiley.

Birnbaum1942AMS Birnbaum, Z. W. (1942). An inequality for Mills' ratio. *Ann. Math. Statist. 13*, 245–246.

ChatterjeeMeckes2008Alea Chatterjee, S. and E. Meckes (2008). Multivariate normal approximation using exchangeable pairs. *Alea 4*, 257–283.

HoChen1978AnnProb Ho, S.-T. and L. H. Y. Chen (1978). An L_p bound for the remainder in a combinatorial central limit theorem. *Annals of Probability 6*(2), 231 – 249.

Meckes-thesis Meckes, E. (2006). *An infinitesimal version of Stein's method of exchangeable pairs*. Ph. D. thesis, Stanford University (Mathematics). Documents available at https://case.edu/artsci/math/mwmeckes/elizabeth/.

Meckes2009Luminy Meckes, E. (2009). On Stein's method for multivariate normal approximation. In *High Dimensional Probability V: The Luminy Volume*, Volume 5 of *Institute of Mathematical Statistics COLLECTIONS*, pp. 153–178.

PollardUGMTP Pollard, D. (2001). *A User's Guide to Measure Theoretic Probability*. Cambridge University Press.

Pttm Pollard, D. (2029). *Probability Tools, Tricks, and Miracles*. Cambridge University Press. (Under construction).

RinottRotar1996JMultivA Rinott, Y. and V. Rotar (1996). A multivariate CLT for local dependence with $n^{-1/2}\log n$ rate and applications to multivariate graph related statistics. *Journal of Multivariate Analysis 56*(2), 333–350.

inottRotar1997AnnApplProb Rinott, Y. and V. Rotar (1997). On coupling constructions and rates in the CLT for dependent summands with applications to the antivoter model and weighted U-statistics. *The Annals of Applied Probability 7*(4), 1080–1105.

Sampford1953AMS Sampford, M. R. (1953). Some inequalities on Mill's ratio and related functions. *The Annals of Mathematical Statistics 24*(1), 130–132.

Stein1972Berk6 Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proc. Sixth Berkeley Symp. Math. Stat. Prob.*, Volume 2, pp. 583–602.

Stein86ims Stein, C. (1986). *Approximate Computation of Expectations*, Volume 7 of *Lecture Notes–Monograph series*. Institute of Mathematical Statistics.