A new method of normal approximation

Sourav Chatterjee (UCB)

Central limit theorems

- Classical CLT: If $X_1, \ldots, X_n$ are independent random variables with zero mean and finite variance and (...), then

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i$$

  has a gaussian distribution in the limit.
- Often, the X_i's need to be only *approximately* independent.
- Many ways to prove such results, e.g.
  1. Characteristic functions
  2. Martingales & Skorohod embeddings
  3. Little blocks, big blocks
  4. Stein's method (and variants)
  5. Hájek Projections
  6. Specialized techniques for special problems

  ... but they all require luck, and often, very hard work.

An example from statistics

- Let $X_1, \ldots, X_n$ be i.i.d. random d-vectors.
- Let

$$f(X_1, \ldots, X_n) = \frac{1}{\sqrt{n}} \sum_i f_i, \tag{1}$$

  where f_i is a function of X_i and its k nearest neighbors in the set $\{X_1, \ldots, X_n\}$.
- Data might lie on a complicated lower-dimensional manifold.
- Specific example (Levina & Bickel '05): Unbiased estimate of the dimension of the manifold.
- **Routine question:** How to find normal approximation bounds for $f(X_1, \ldots, X_n)$?

An example from statistics

- **Routine answer:** Intuitively plausible; low dependence at distances, etc. But...
- ... technical nightmare, specially for general manifolds.
- **Deeper issue:** Why can't we prove such a plausible result in our sleep?

Another example: Linear statistics of random eigenvalues

- Let $A = (a_{ij})$ be a real symmetric random matrix of order n, with i.i.d. entries on and above the diagonal.
- Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of $n^{-1/2}A$.
- Let f be any function on $\mathbb{R}$ and let $W = \sum_{i=1}^n f(\lambda_i)$. This is called a "linear statistic of the eigenvalues of A".
- Mysterious fact (Sinai & Soshnikov): Under some smoothness assumptions on f, $\mathrm{Var}(W)$ converges to a positive limit σ^2 as $n \to \infty$. Moreover, $W - \mathbb{E}(W)$ converges in law to $N(0, \sigma^2)$.

Linear statistics contd.

- Similar results hold for sample covariance matrices (Bai & Silverstein), random unitary matrices (Diaconis & Evans), and other ensembles (e.g. Rider & Virág).
- All proofs are by method of moments. No insights. No convergence rates.
- These are examples where the usual methods of normal approximation do not work. May be we don't understand everything about normal approximation?

Stein's method

- If $Z \sim N(0,1)$, then $\mathbb{E}(\varphi(Z)Z) = \mathbb{E}(\varphi'(Z))$ for all φ.
- Stein's idea: If $\mathbb{E}(\varphi(W)W) \approx \mathbb{E}(\varphi'(W))$ for many φ's, then W is approximately $N(0,1)$.
- Many variants, e.g.
  - Exchangeable pairs (Stein)
  - Zero bias couplings (Goldstein & Reinert)
  - Size bias couplings (Goldstein & Rinott)
  - Generator approach (Arratia et. al.; Barbour)
- Common complaint: Hard to apply to arbitrary problems.

A conceptual insight

- Define

$$S_p(W) := \sup\{|\mathbb{E}(\varphi(W)W - \varphi'(W))| : \|\varphi'(W)\|_p \le 1\}.$$

- From Stein's lemma: $d_{TV}(W, Z) \le 2S_p(W)$ for every $p > 1$. (Recall: $d_{TV}(X, Y) = \sup_A |\mathbb{P}(X \in A) - \mathbb{P}(Y \in A)|$.)

Theorem

If W has mean zero, unit variance, and density ρ, then

$$S_p(W) = \|h(W) - \mathbb{E}h(W)\|_q,$$

where

$$h(x) = \frac{\int_x^\infty y\rho(y)dy}{\rho(x)},$$

and $\frac{1}{p} + \frac{1}{q} = 1$.

- This shows that approximate normality of W = concentration of $h(W)$.
- Proof is based on L^p-L^q duality from functional analysis.

Idea

- Concentration problems, unlike normal approximation problems, are *transferable* via conditional expectation. That is, if we can write

$$h(W) = \mathbb{E}(T \mid W),$$

where T is a an explicit object arising from the given problem, then

$$\|h(W) - \mathbb{E}h(W)\|_q \leq \|T - \mathbb{E}T\|_q.$$

This might wake you up...

Theorem

Suppose $W = f(X_1, \ldots, X_n)$, *where* X_i*'s are i.i.d.* $N(0,1)$, *and* f *is smooth. Assume* $\mathbb{E}(W) = 0$ *and* $\mathbb{E}(W^2) = 1$. *Let* Z *be a vector of* n *i.i.d. standard gaussians and define* $T : \mathbb{R}^n \to \mathbb{R}$ *as*

$$T(x) := \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x) \int_0^1 \frac{1}{2\sqrt{t}} \mathbb{E}\left(\frac{\partial f}{\partial x_i}(\sqrt{t}x + \sqrt{1-t}Z) \right) dt.$$

Then $h(W) = \mathbb{E}(T(X_1, \ldots, X_n) \mid W)$.

(Recall: This implies that $d_{TV}(W, N(0,1)) \le 2\sqrt{\mathrm{Var}(h(W))}$.)

Simple examples

$f(X_1, \ldots, X_n)$	$T(X_1, \ldots, X_n)$
$\frac{\sum_{i=1}^n X_i}{\sqrt{n}}$	1
$\frac{\sum_{i=1}^n X_i^2 - n}{\sqrt{2n}}$	$\frac{\sum_{i=1}^n X_i^2}{n}$
$\frac{\sum_{i=1}^n X_i X_{i+1}}{\sqrt{n}}$	$\frac{1}{2n}\sum_{i=2}^{n}(X_{i-1} + X_{i+1})^2 + \frac{X_1^2 + X_{n+1}^2}{2n}$

Sketch of proof

- h is characterized by $\mathbb{E}(\varphi(W)W) = \mathbb{E}(\varphi'(W)h(W))$.
- Suffices to show $\mathbb{E}(\varphi(W)W) = \mathbb{E}(\varphi'(W)T(X))$.
- Let $X^t = \sqrt{t}X + \sqrt{1-t}Z$. Then

$$\begin{aligned}\mathbb{E}(\varphi(W)W) &= \mathbb{E}(\varphi(W)(f(X^1) - f(X^0))) \\ &= \mathbb{E}\left(\varphi(W)\int_0^1 \frac{d}{dt} f(X^t)dt\right).\end{aligned}$$

- Note that

$$\frac{d}{dt} f(X^t) = \sum_{i=1}^n \left(\frac{X_i}{2\sqrt{t}} - \frac{Z_i}{2\sqrt{1-t}}\right)\frac{\partial}{\partial x_i} f(X^t).$$

- Proof is completed by a sequence of tricky integration-by-parts steps.

How does one bound $\mathrm{Var}(T)$ in general?

Answer: By the gaussian Poincaré inequality. If $X_1, \ldots, X_n$ are i.i.d. $N(0,1)$, and $T : \mathbb{R}^n \to \mathbb{R}$ is absolutely continuous, then

$$\mathrm{Var}(T(X_1, \ldots, X_n)) \le \mathbb{E}\|\nabla T(X_1, \ldots, X_n)\|^2.$$

(Recall: $\nabla T = (\partial T/\partial x_1, \ldots, \partial T/\partial x_n)$ is the gradient of T.)

A general result

- Let $\mathcal{L}(c_1, c_2)$ be the class of probability measures on $\mathbb{R}$ that arise as laws random variables like $h(Z)$, where $Z \sim N(0,1)$ and $h \in C^2(\mathbb{R})$ satisfies
$$|h'(x)| \leq c_1 \text{ and } |h''(x)| \leq c_2.$$

- Let $\mathcal{L}$ be the class of all distributions that belong to $\mathcal{L}(c_1, c_2)$ for some finite c_1, c_2.
- In the next slide, we have a general normal approximation theorem for smooth functions of independent r.v.'s with laws in $\mathcal{L}$.

Only for the professionals...

Theorem

Let $X = (X_1, \ldots, X_n)$ be a vector of independent random variables in $\mathcal{L}(c_1, c_2)$ for some finite c_1, c_2. Take any $g \in C^2(\mathbb{R}^n)$ and let ∇g and $\nabla^2 g$ denote the gradient and Hessian of g. Let

$$\kappa_0 = \left(\mathbb{E}\sum_{i=1}^n \left|\frac{\partial g}{\partial x_i}(X)\right|^4\right)^{1/2},$$

$$\kappa_1 = (\mathbb{E}\|\nabla g(X)\|^4)^{1/4}, \text{ and}$$

$$\kappa_2 = (\mathbb{E}\|\nabla^2 g(X)\|^4)^{1/4}.$$

Suppose $W = g(X)$ has a finite fourth moment and let $\sigma^2 = \mathrm{Var}(W)$. Let Z be a normal random variable having the same mean and variance as W. Then

$$d_{TV}(W, Z) \le \frac{2\sqrt{5}(c_1 c_2 \kappa_0 + c_1^3 \kappa_1 \kappa_2)}{\sigma^2}.$$

Applications to linear statistics of eigenvalues

- If $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of a Wigner matrix with entries in $\mathcal{L}$ and $W = \sum_{i=1}^n f(\lambda_i)$, and let Z be a gaussian random variable with the same mean and variance as W. Then

$$d_{TV}(W, Z) \le \frac{C(f)}{\sigma^2 \sqrt{n}},$$

  where $C(f)$ is an explicit constant depending on f.
- For the sample covariance matrix given by a $p \times N$ data matrix of i.i.d. entries, the corresponding bound is

$$\frac{C(f)(p \wedge N)}{\sigma^2 N^{3/2}}.$$

  (Bai and Silverstein have shown that when $p/N \to \alpha \in (0, 1)$, σ^2 converges to a positive limit.)
- Works for Double Wishart matrices and Gaussian Toeplitz matrices (new results).

Part II: Arbitrary functions of independent random variables

Notation

- Suppose $X = (X_1, \ldots, X_n)$ is a vector of independent random variables.
- Let $W = f(X)$ be an arbitrary function of X .
- Let X' be an independent copy of X.
- For each $A \subseteq \{1, \ldots, n\}$, define the vector X^A as

$$X_i^A = \begin{cases} X_i' & \text{if } i \in A, \\ X_i & \text{if } i \notin A. \end{cases}$$

Notation

- Let
$$\Delta_j f(X) = f(X) - f(X^j).$$
(Recall: $X^j = (X_1, \ldots, X_{j-1}, X_j', X_{j+1}, \ldots, X_n)$.)
- Define
$$T(X, X') = \frac{1}{2} \sum_A \frac{1}{\binom{n}{|A|}(n - |A|)} \sum_{j \notin A} \Delta_j f(X) \Delta_j f(X^A).$$

Our main result

Theorem

Suppose $\mathbb{E}(W) = 0$ *and let* $\sigma^2 = \mathbb{E}(W^2)$. *Then*

$$\begin{aligned}&\mathcal{W}(\sigma^{-1}W, N(0,1))\\ &\le \frac{\sqrt{\mathrm{Var}(\mathbb{E}(T|W))}}{\sigma^2} + \frac{1}{2\sigma^3}\sum_{j=1}^{n}\mathbb{E}|\Delta_j f(X)|^3,\end{aligned}$$

where $\mathcal{W}$ *is the Wasserstein distance and* T *is defined as in the previous slide.*

(Recall: $\mathcal{W}(X, Y) = \sup\{|\mathbb{E}f(X) - \mathbb{E}f(Y)| : \|f\|_{\mathrm{Lip}} \le 1\}$.)

Simplest example

- Suppose $f(X) = n^{-1/2}\sum_{i=1}^n X_i$. Then

$$T(X, X') = \frac{1}{2n}\sum_{j=1}^n (X_j - X_j')^2.$$

  Thus, $\mathrm{Var}(T(X, X')) = O(n^{-1})$.

- Also,

$$\sum_{j=1}^n \mathbb{E}|\Delta_j f(X)|^3 = n^{-3/2}\sum_{j=1}^n \mathbb{E}|X_j - X_j'|^3 = O(n^{-1/2}).$$

- Combining, we get an $O(n^{-1/2})$ error bound.

Quadratic forms

- Suppose $X_1, \ldots, X_n$ are i.i.d. symmetric ± 1-valued random variables.
- Let $\mathbf{A} = (a_{ij})$ be a real symmetric matrix of order n.
- Let $W = \sum_{i<j} a_{ij} X_i X_j$ and $\sigma^2 = \mathrm{Var}(W) = \frac{1}{2}\,\mathrm{Tr}(\mathbf{A}^2)$.
- Can compute: $\mathbb{E}(T|X) = \frac{1}{2} X^t \mathbf{A}^2 X$.

Putting $W' = (W - \mathbb{E}(W))/\sigma$, we have

$$\mathcal{W}(W', N(0,1)) \le \left(\frac{\mathrm{Tr}(\mathbf{A}^4)}{2\sigma^4}\right)^{1/2} + \frac{5}{2\sigma^3} \sum_{i=1}^{n} \left(\sum_{j=1}^{n} a_{ij}^2\right)^{3/2}.$$

This is slightly stronger than the best known results (Rotar '73, Hall '84, Götze & Tikhomirov '99, '02).

Proof of main theorem (brief sketch)

- For any g, f, A, and $j \notin A$,

$$\begin{aligned}
&\mathbb{E}(\Delta_j g(X)\Delta_j f(X^A)) \\
&= \mathbb{E}(g(X)\Delta_j f(X^A)) - \mathbb{E}(g(X^j)\Delta_j f(X^A)) \\
&= 2\mathbb{E}(g(X)\Delta_j f(X^A)) \text{ by exchangeability of } (X_j, X_j').
\end{aligned}$$

- With $g = \varphi \circ f$, we have $\Delta_j g(X) \approx \varphi'(f(X))\Delta_j f(X)$, and hence

$$\begin{aligned}
&\frac{1}{2}\mathbb{E}(\varphi'(f(X))\Delta_j f(X)\Delta_j f(X^A)) \\
&\approx \frac{1}{2}\mathbb{E}(\Delta_j g(X)\Delta_j f(X^A)) = \mathbb{E}(\varphi(f(X))\Delta_j f(X^A)).
\end{aligned}$$

- Thus,

$$\begin{aligned}
&\mathbb{E}(\varphi'(f(X))T(X, X')) \\
&\approx \mathbb{E}\left(\varphi(f(X))\sum_A \frac{1}{\binom{n}{|A|}(n-|A|)}\sum_{j\notin A}\Delta_j f(X^A)\right).
\end{aligned}$$

Proof of main theorem (brief sketch)

- Now note that

$$\sum_A \frac{1}{\binom{n}{|A|}(n-|A|)} \sum_{j \notin A} \Delta_j f(X^A) = f(X) - f(X'),$$

  which is just an algebraic identity.
- Thus,

$$\begin{aligned}\mathbb{E}(\varphi'(f(X))T(X,X')) &\approx \mathbb{E}[\varphi(f(X))(f(X)-f(X'))] \\ &= \mathbb{E}(\varphi(f(X))f(X)).\end{aligned}$$

- Exact equality holds for $\varphi(u) = u$, which gives

$$\mathbb{E}(T(X,X')) = \mathrm{Var}(f(X)) = \sigma^2.$$

- Thus, if $\mathrm{Var}(T(X,X'))$ is tiny, then

$$\mathbb{E}(\varphi(f(X))f(X)) \approx \sigma^2 \mathbb{E}(\varphi'(f(X))),$$

  which shows that $f(X) \stackrel{\cdot}{\sim} N(0,\sigma^2)$.

Again, how to bound $\mathrm{Var}(T)$ in general?

Very useful tool: Analog of Poincaré inequality, known as the *Efron-Stein* inequality in the statistical literature.

Theorem (Efron-Stein inequality)

Let $Z = f(Y_1, \ldots, Y_m)$ be a function of independent random objects $Y_1, \ldots, Y_m$. Let Y_i' be an independent copy of Y_i, $i = 1, \ldots, m$. Then

$$\mathrm{Var}(Z) \leq \frac{1}{2}\sum_{i=1}^{m} \mathbb{E}\big[(f(Y_1, \ldots, Y_{i-1}, Y_i', Y_{i+1}, \ldots, Y_m) - f(Y_1, \ldots, Y_m))^2\big].$$

A nearest neighbor problem

- Let $X_1, \ldots, X_n$ be i.i.d. random variables lying on a nice manifold of unknown dimension m.
- For a fixed $k \geq 2$, the Levina-Bickel estimate of m with tuning parameter k is given by

$$\hat{m}_k = \frac{1}{n} \sum_{\ell=1}^{n} \left(\frac{1}{k-1} \sum_{j=1}^{k-1} \log \frac{D_{\ell k}}{D_{\ell j}} \right)^{-1}, \tag{2}$$

  where $D_{\ell j}$ is the distance between X_ℓ and its j^{th} nearest neighbor.
- Under assumptions, $\hat{m}_k$ is consistent. How to prove a CLT?
- Existing results and techniques (Bickel-Breiman, Avram-Bertsimas, Penrose-Yukich, etc.) provide no immediate help.

A general nearest neighbor CLT

Theorem

Suppose $X_1, \ldots, X_n$ are i.i.d. $\mathbb{R}^d$-valued random vectors such that $\|X_1 - X_2\|$ is a continuous r.v. Fix $k \geq 1$, and suppose that for each i, f_i is a function of only X_i and its k nearest neighbors, and $W = n^{-1/2} \sum_i f_i$. Suppose for some $p \geq 8$, $\gamma_p := \max_i \mathbb{E}|f_i|^p$ is finite. Let $\sigma^2 = \mathrm{Var}(W)$ and $W' = (W - \mathbb{E}(W))/\sigma$. Then

$$\mathcal{W}(W', N(0,1)) \leq C \frac{\alpha(d)^3 k^4 \gamma_p^{2/p}}{\sigma^2 n^{(p-8)/2p}} + C \frac{\alpha(d)^3 k^3 \gamma_p^{3/p}}{\sigma^3 n^{(p-6)/2p}},$$

where $\alpha(d)$ is the minimum number of 60° cones at the origin required to cover $\mathbb{R}^d$, and C is a universal constant.

Summary

- Aim of the work: To reduce normal approximation problems to technically manageable variance bounding exercises. Gives explicit bounds.
- Variance bounds can be handled effectively by Poincaré or martingale inequalities.
- Hardest applications till now: the eigenvalue CLT and the nearest neighbor CLT.
- Future plan: (1) Work on other examples that I have in mind. (2) Find more applications. (3) Convince others to use the method.