A generalization of the Lindeberg principle

Sourav Chatterjee

- $\mathbf{X} = (X_1, \ldots, X_n)$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)$ are independent random vectors.

- $f : \mathbb{R}^n \to \mathbb{R}$ is a C^3 function.

- If, say, X_i's and Y_i's are all independent and $\mathbb{E}X_i = \mathbb{E}Y_i$, $\mathbb{E}X_i^2 = \mathbb{E}Y_i^2$, then what are sufficient conditions on f which ensure that $f(\mathbf{X})$ and $f(\mathbf{Y})$ are close in distribution?

- Reason for considering only first two moments: Can be adjusted using linear transformation.

- Conditions based on first two derivatives cannot suffice: Consider $\frac{1}{n}\sum x_i^2$ and $\frac{1}{n}\sum x_i^3$.

- If we let $\mathbf{Z}_i = (X_1, \ldots, X_i, Y_{i+1}, \ldots, Y_n)$, then

$$\mathbb{E}f(\mathbf{X}) - \mathbb{E}f(\mathbf{Y}) = \sum_{i=1}^{n} (\mathbb{E}f(\mathbf{Z}_i) - \mathbb{E}f(\mathbf{Z}_{i-1})).$$

- Let $\mathbf{Z}_i^0 = (X_1, \ldots, X_{i-1}, 0, Y_{i+1}, \ldots, Y_n)$. Taylor expansion gives

$$\begin{aligned} & f(\mathbf{Z}_i) - f(\mathbf{Z}_{i-1}) \\ & = (X_i - Y_i)\partial_i f(\mathbf{Z}_i^0) + \frac{1}{2}(X_i^2 - Y_i^2)\partial_i^2 f(\mathbf{Z}_i^0) \\ & \quad + \frac{1}{6}X_i^3 \partial_i^3 f(\mathbf{Z}_i^*) + \frac{1}{6}Y_i^3 \partial_i^3 f(\mathbf{Z}_i^{**}). \end{aligned}$$

- Under independence, and $\mathbb{E}X_i = \mathbb{E}Y_i$, $\mathbb{E}X_i^2 = \mathbb{E}Y_i^2$, first two terms vanish on taking expectation.

- Thus, if third moments are bounded, then $|\mathbb{E}f(\mathbf{X}) - \mathbb{E}f(\mathbf{Y})| \leq n\psi_3$, where ψ_3 denotes the typical size of the third order derivatives of f.

- Note: Moving from expectations to distributions is easy; just work with $g \circ f$ instead of f, where $g : \mathbb{R} \to \mathbb{R}$ is a smooth function with bounded derivatives. Then error bound is like $n \max\{\psi_1^3, \psi_2\psi_1, \psi_3\}$.

- Note: Suppose only that the Y_i's are independent. Then it suffices that

$$\mathbb{E}(X_i | X_1, \ldots, X_{i-1}) \approx \mathbb{E}(Y_i)$$

and

$$\mathbb{E}(X_i^2 | X_1, \ldots, X_{i-1}) \approx \mathbb{E}(Y_i^2),$$

where the approximations are good enough.

- This means, we only need that the partial sums of the X_i's behave like Brownian motion.

- Donsker Invariance does not suffice for all problems. For example, scan statistics, random matrices, free energy, etc.

- Scan statistics: Let $X_1, \ldots, X_n$ be independent, mean zero, unit variance, bounded third moment.

- Let $\mathcal{A}$ be a collection of subsets of $\{1, \ldots, n\}$.

- Let $M(\mathbf{X}) := n^{-1/2} \max_{A \in \mathcal{A}} \sum_{i \in A} X_i$.

- Question: When can we replace X_i's by standard Gaussians?

- Can prove: If $\mathbf{Y}$ is a vector of independent standard Gaussians, then for any smooth function g,

  $$|\mathbb{E}g(M(\mathbf{X}))-\mathbb{E}g(M(\mathbf{Y}))| \le Cn^{-1/6}(\log|\mathcal{A}|)^{2/3},$$

  where C is constant depending on g and the third absolute moments of the X_i's.

- Method: Uniformly approximate $\max_{A\in\mathcal{A}} S_A$ by a smooth function using

  $$|\max_{A\in\mathcal{A}} S_A - L^{-1}\log\sum_{A\in\mathcal{A}} e^{LS_A}| \le L^{-1}\log|\mathcal{A}|$$

  and optimize the resulting bound over L.

- Free energy of the S-K model in spin glasses:

$$N^{-1} \log \sum_{\sigma \in \{-1,1\}^N} \exp\Big(\sum_{i<j\leq N} g_{ij}\sigma_i\sigma_j\Big)$$

- The limit of this as $N \to \infty$ is known to exist when g_{ij}'s are standard Gaussian.

- Can easily show using our method that same limit holds with g_{ij}'s non-Gaussian. (Already proved by Carmona and Hu.)

- Random matrices: The *Empirical Spectral Distribution* (ESD) of a matrix is the probability distribution which puts equal mass on each of its eigenvalues.

- Let $X = (x_{ij})_{1\le i\le p, 1\le j\le n}$ be a data matrix of i.i.d. N(0,1) variables, and let S be the corresponding sample covariance matrix.

- If $p/n \to \lambda \in (0,\infty)$, then the ESD of S converges to a nonrandom limiting law depending only on λ (the Marčenko-Pastur family of distributions).

- Known (also, provable by our method) that the x_{ij}'s need not be Normal.

- Question: Can we have a multidimensional version of the bivariate permutation test for correlation?

- More precisely, if we permute each row of a (nonrandom) data matrix independently, does the resulting sample covariance matrix have the same asymptotic properties as in the independent Gaussian case?

- Answer: Yes, at least as far as spectral distributions are concerned.

- Suppose we have exchangeable random variables $V_1, \ldots, V_n$. What is the version of the previous result in this situation?

- Answer:
  - Let $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^n V_i$.
  - Let $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (V_i - \widehat{\mu})^2$.
  - Let $Y_1, \ldots, Y_n$ be i.i.d. $N(0,1)$, independent of the V_i's.
  - Let $Z_i = \widehat{\mu} + \widehat{\sigma}(Y_i - \bar{Y})$, $i = 1, \ldots, n$.

  Then, the vector $(V_1, \ldots, V_n)$ "behaves like" $(Z_1, \ldots, Z_n)$, in the same sense as before.

- For a smooth function f, $|\mathbb{E}f(\mathbf{V}) - \mathbb{E}f(\mathbf{Z})|$ is bounded by

$$\sqrt{n}M^2\psi_2 + nM^3\psi_3,$$

where, as before, ψ_r is the typical size of the r^{th} order derivatives, while M is the typical size of $\max|V_i|$.

- Note that using Y_i instead of $Y_i - \bar{Y}$ won't work. Example: Sampling without replacement.

- Possible applications: May be used to simplify situations which involve complicated but exchangeable random variables, e.g. occupancy problems, nearest neighbors, permutation statistics, and so on.

Theorem 1 *Suppose $V_1, \ldots, V_n$ are exchangeable random variables with finite third moment. Define*

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} V_i \text{ and } \widehat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(V_i - \widehat{\mu})^2}.$$

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ be a C^3 function, and ψ_2, ψ_3 are monotone functions such that for $r = 2, 3$, all r^{th} order partial derivatives of f are dominated by the function $\psi_r(\max_{1\le i\le n}|x_i|)$. Let $Y_1, \ldots, Y_n$ be i.i.d. standard Gaussian random variables, independent of $(V_1, \ldots, V_n)$. Let $Z_i = \widehat{\mu} + \widehat{\sigma}(Y_i - \bar{Y})$, $i = 1, \ldots, n$.

$$\begin{aligned} &|\mathbb{E}f(V_1, \ldots, V_n) - \mathbb{E}f(Z_1, \ldots, Z_n)| \\ &\le 10\sqrt{n}\mathbb{E}(A^4)^{1/2}\mathbb{E}(\psi_2(R)^2)^{1/2} \\ &\quad + 7n\mathbb{E}(A^6)^{1/2}\mathbb{E}(\psi_3(R)^2)^{1/2}, \end{aligned} \tag{1}$$

where $A = 2\max_{1\le i\le n}|V_i|$ and $R = \max\{2A, \max_{1\le i\le n}|Z_i|\}$.

- Steps in the proof: First, note that we can assume that $\sum V_i = 0$ and $\sum V_i^2 = n$, since we can standardize the V_i's and work conditionally given $\widehat{\mu}$ and $\widehat{\sigma}$.

- Next, let $\mathcal{F}_i$ be the sigma-algebra generated by $V_1, \ldots, V_i$, and define

$$X_i = V_i + \frac{1}{n-i+1} \sum_{j=1}^{i-1} V_j.$$

- Then $\mathbb{E}(X_i | \mathcal{F}_{i-1}) = 0$ and

$$\mathbb{E}(X_i^2 | \mathcal{F}_{i-1}) = 1 + O_P((n-i+1)^{-1/2}).$$

  We use our previous result to replace the X_i's by i.i.d. N(0,1) variables $Y_1, \ldots, Y_n$.

- Easy to check:

$$V_i = X_i - \sum_{j=1}^{i-1} \frac{X_j}{n-j}.$$

- If we let

$$Y_i' = Y_i - \sum_{j=1}^{i-1} \frac{Y_j}{n-j},$$

then for $i > j$,

$$\mathrm{Cov}(Y_i', Y_j') = -\frac{1}{n-j} + \sum_{k=1}^{j-1} \frac{1}{(n-k)^2}.$$

- Can manipulate to show that this is approximately

$$\mathrm{Cov}(Y_i - \bar{Y}, Y_j - \bar{Y}) + O((n - i \wedge j + 1)^{-2}).$$

Similar approximation holds for $i = j$, too.

- Now use the following result about normal random vectors:

Lemma 1 *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be independent vectors of centered Gaussian random variables. Suppose* $f : \mathbb{R}^n \to \mathbb{R}$ *is a twice differentiable function with bounded derivatives. Then*

$$\mathbb{E}f(\mathbf{Y}) - \mathbb{E}f(\mathbf{X})$$
$$= \frac{1}{2}\int_0^1 \sum_{1\leq i,j\leq n} \mathbb{E}\left[\frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{Z}_t)\right](\mathbb{E}Y_iY_j - \mathbb{E}X_iX_j)\, dt$$

where $\mathbf{Z}_t = \sqrt{1-t}\mathbf{X} + \sqrt{t}\mathbf{Y}$*, provided the expectations on the right side exist.*

- Back to matrices: The Stieltjes transform of a probability distribution F is defined on $\mathbb{C}\backslash\mathbb{R}$ by

$$m_F(z) := \int_{-\infty}^{\infty} \frac{1}{x-z} dF(x)$$

- The Stieltjes transform of the ESD of an $n \times n$ matrix A is given by

$$m_A(z) = \frac{1}{n}\mathrm{Tr}((A - zI)^{-1}).$$

- Stieltjes transforms have a continuous characterizing relationship with distribution functions.

- Stieltjes transforms are amenable to differentiation: Suppose $A = A(u)$ is a matrix-valued function of some scalar parameter u. Let

  $$G(u) = (A(u) - zI)^{-1}.$$

  Then

  $$\frac{dG}{du} = -G\frac{dA}{du}G$$

  Continuing, we can arrive at a cumbersome but explicit expression for third derivatives.

- Bounds on the derivatives can be obtained using the properties of the Hilbert-Schmidt norm; in particular, the following crucial property:

  If A and B are square matrices, and A is normal, with spectral radius ρ, then $\|AB\| \leq \rho\|B\|$.

  This is useful because of the fact that $\|G\| \leq |\mathrm{Im}(z)|^{-1}$.

- For instance, we have

$$\begin{aligned}&|\mathrm{Tr}((\partial_{ij}^2 S)G(\partial_{ij}S)G^2)|\\ &\le \|\partial_{ij}^2 S\|\|G(\partial_{ij}S)G^2\|\\ &\le \|\partial_{ij}^2 S\|\|\partial_{ij}S\||\mathrm{Im}(z)|^{-3}.\end{aligned}$$

- Returning to the sample covariance matrix, let f denote its Stieltjes transform at a fixed $z \in \mathbb{C}\backslash\mathbb{R}$. When $p/n \to \lambda \in (0, \infty)$, we can show that

$$\psi_2(f) \le Cn^{-2}, \ \psi_3(f) \le Cn^{-5/2}.$$

- Theorem 1 can now be invoked to complete the argument.

- Stein's method of Normal approximation: If (W, W') is an exchangeable pair of random variables, and

$$\mathbb{E}(W' - W|W) \approx -\lambda W,$$
$$\mathbb{E}((W' - W)^2|W) \approx 2\lambda + o(\lambda),$$
$$\mathbb{E}|W' - W|^3 \ll \lambda^{3/2},$$

where λ is a very small number, then W is approximately standard Gaussian.

- Idea: If we generate a reversible Markov chain $W_0, W_1, \ldots$, with $W_0 = W$ and $W_1 = W'$, then it behaves like a discrete approximation of a stationary Ornstein-Uhlenbeck process.

- Let

$$X_i = W_i - (1-\lambda)W_{i-1}.$$

- Then $\mathbb{E}(X_i|\mathcal{F}_{i-1}) \approx 0$ and $\mathbb{E}(X_i^2|\mathcal{F}_{i-1}) \approx 2\lambda$.

- Reconstruct W_n from $X_1, \ldots, X_n$ as

$$W_n = (1-\lambda)^n W_0 + \sum_{i=1}^{n} (1-\lambda)^{n-i} X_i.$$

- Use Lindeberg approach to get a bound on

  $$\mathbb{E}f(W_n) - \mathbb{E}f((1-\lambda)W_0 + \sum_{i=1}^{n}(1-\lambda)^{n-i} Y_i),$$

  where Y_i's are i.i.d. $N(0, 2\lambda)$.

- Finally, note that $\mathbb{E}f(W_n) = \mathbb{E}f(W)$, and take $n \to \infty$.

- This gives an approach to getting general diffusion approximation bounds: Recover the "pretend Brownian motion increments" and write the diffusion as a function of those; then use Lindeberg method on the "reconstruction function" to get error bounds.