SECTION 12

Random Convex Sets

Donoho (1982) and Donoho and Gasko (1987) studied an operation proposed by Tukey for extending the idea of trimming to multidimensional data. Nolan (1989a) gave a rigorous treatment of the asymptotic theory. Essentially the arguments express the various statistics of interest as differentiable functionals of an empirical measure. The treatment in this section will show how to do this without the formal machinery of compact differentiability for functionals, by working directly with almost sure representations. [Same amount of work, different packaging.]

To keep the discussion simple, let us consider the case of an independent sample $\xi_1, \xi_2, \ldots$ of random vectors from the symmetric bivariate normal distribution P on $\mathbb{R}^2$, and consider only the analogue of 25% trimming.

The notation will be cleanest when expressed (using traditional empirical process terminology) in terms of the *empirical measure* P_n, which puts mass $1/n$ at each of the points $\xi_1(\omega), \ldots, \xi_n(\omega)$.

Let $\mathcal{H}$ denote the class of all closed halfspaces in $\mathbb{R}^2$. Define a random compact, convex set $K_n = K_n(\omega)$ by intersecting all those halfspaces that contain at least $3/4$ of the observations:

$$K_n(\omega) = \bigcap \{H \in \mathcal{H} : P_n H \geq \tfrac{3}{4}\}.$$

It is reasonable to hope that K_n should settle down to the set

$$B(r_0) = \bigcap \{H \in \mathcal{H} : PH \geq \tfrac{3}{4}\},$$

which is a closed ball centered at the origin with radius r_0 equal to the 75% point of the one-dimensional standard normal distribution. That is, if Φ denotes the $N(0,1)$ distribution function, then $r_0 = \Phi^{-1}(3/4) \approx .675$. Indeed, a simple continuity argument based on a uniform strong law of large numbers,

$$\sup_{\mathcal{H}} |P_n H - PH| \to 0 \qquad \text{almost surely,} \tag{12.1}$$

would show that, for each $\epsilon > 0$, there is probability one that

$$B(r_0 - \epsilon) \subseteq K_n(\omega) \subseteq B(r_0 + \epsilon) \qquad \text{eventually.}$$

In a natural sense, K_n is a strongly consistent estimator. Let us not dwell on the details here, because the next argument, which gives the finer asymptotics for K_n, is much more interesting. [The almost sure representation that will appear soon would imply the "in probability" version of (12.1). This would give consistency in probability, which is all that we really need before embarking upon the asymptotic distribution theory for K_n.]

Once K_n contains the origin as an interior point it makes sense to describe its boundary in polar coordinates. Let $R_n(\theta) = R_n(\omega, \theta)$ denote the distance from the origin to the boundary in the direction θ. The consistency result then has the reformulation:

$$\sup_\theta |R_n(\omega, \theta) - r_0| \to 0 \qquad \text{almost surely.}$$

With the help of the functional central limit theorems from Section 10, we can improve this to get convergence in distribution of a random process,

$$\sqrt{n}(R_n(\omega, \theta) - r_0) \qquad \text{for } -\pi \le \theta \le \pi,$$

to a Gaussian process indexed by θ. [It would be more elegant to take the unit circle as the index set, identifying the points $\theta = \pi$ and $\theta = -\pi$.] Such a result would imply central limit theorems for a variety of statistics that could be defined in terms of K_n.

Heuristics. We need to establish a functional central limit theorem for the standardized *empirical process*,

$$\nu_n(\omega, H) = \sqrt{n}(P_n H - PH),$$

as a stochastic process indexed by $\mathcal{H}$. We must show that $\{\nu_n\}$ converges in distribution to a Gaussian process ν indexed by $\mathcal{H}$.

Let $H(r, \theta)$ denote the closed halfspace containing the origin with boundary line perpendicular to the θ direction at a distance r from the origin. That is, $H(r, \theta)$ consists of all points whose projections onto a unit vector in the θ direction are $\le r$. For a given point with polar coordinates (r, θ), the halfspace $H(r, \theta)$ maximizes PH over all H that have (r, θ) as a boundary point. The boundary point of $B(r_0)$ in the direction θ is determined by solving the equation $PH(r, \theta) = 3/4$ for r, giving $r = r_0$. Similarly, the boundary point of K_n in the direction θ is almost determined by solving the equation $P_n H(r, \theta) = 3/4$, as we will soon see. (Discreteness of P_n might prevent us from getting exact equality; and the halfspace that determines the boundary point will be rotated slightly from the $H(r, \theta)$ position.) That is, $R_n(\theta)$ is approximately determined by solving the following equation for r:

$$\tfrac{3}{4} \approx P_n H(r, \theta) = PH(r, \theta) + \frac{1}{\sqrt{n}} \nu_n H(r, \theta).$$

Asymptotically the right-hand side is distributed as

$$\Phi(r) + \frac{1}{\sqrt{n}} \nu H(r, \theta) \approx \Phi(r_0) + (r - r_0)\Phi'(r_0) + \frac{1}{\sqrt{n}} \nu H(r_0, \theta)..$$

Thus $\sqrt{n}(R_n(\theta) - r_0)$ should behave asymptotically like $-\nu H(r_0, \theta)/\Phi'(r_0)$, which is a Gaussian process indexed by θ.

The functional limit theorem for ν_n. Define a triangular array of processes,

$$f_{ni}(\omega, H) = \frac{1}{\sqrt{n}}\{\xi_i(\omega) \in H\} \qquad \text{for } H \in \mathcal{H} \text{ and } i \le n.$$

They have constant envelopes $F_{ni} = 1/\sqrt{n}$. We will apply the Functional Central Limit Theorem of Section 10 to the processes

$$\nu_n H = \sum_{i \le n} \Big(f_{ni}(\omega, H) - \mathbb{P} f_{ni}(\cdot, H) \Big).$$

It is easy to show, by an appeal to Lemma 4.4, that the processes define random subsets of $\mathbb{R}^n$ with pseudodimension 3. Every closed halfspace has the form

$$H = \{x \in \mathbb{R}^2 : \alpha \cdot x + \beta \ge 0\}$$

for some unit vector α in $\mathbb{R}^2$ and some real number β. Notice that $f_{ni}(\omega, H) = 1/\sqrt{n}$ if and only if $\alpha \cdot \xi_i + \beta \ge 0$. The points in $\mathbb{R}^n$ with coordinates $\alpha \cdot \xi_i + \beta$ trace out a subset of a 3-dimensional subspace as α and β vary.

The other conditions of the Theorem are just as easy to check. For every pair of halfspaces H_1 and H_2, and every n,

$$\mathbb{P}(\nu_n H_1 \nu_n H_2) = P H_1 H_2 - P H_1 P H_2,$$

and

$$\rho(H_1, H_2)^2 = \rho_n(H_1, H_2)^2 = P|H_1 - H_2|.$$

[Typically, manageability is the only condition that requires any work when the Functional Central Limit Theorem is applied to the standardized sums of independent, identically distributed processes.]

The Theorem asserts that ν_n converges in distribution, as a random element of the function space $B(\mathcal{H})$, to a Gaussian process concentrated on $U(\mathcal{H})$, the set of all bounded, ρ-uniformly continuous functions. The Representation Theorem from Section 9 provides perfect maps ϕ_n and a Gaussian process $\widetilde{\nu}$ with sample paths in $U(\mathcal{H})$ such that the random processes $\widetilde{\nu}_n = \nu_n \circ \phi_n$ satisfy

$$\sup_{\mathcal{H}} |\widetilde{\nu}_n(H) - \widetilde{\nu}(H)| \to 0 \qquad \text{almost surely.}$$

We need not worry about measurability difficulties here, because the supremum over $\mathcal{H}$ is equal to the supremum over an appropriate countable subclass of $\mathcal{H}$. The representation also gives a new version of the empirical measure,

$$\widetilde{P}_n H = P H + \frac{1}{\sqrt{n}} \widetilde{\nu}_n H = P H + \frac{1}{\sqrt{n}} \big(\widetilde{\nu} H + o(1) \big), \tag{12.2}$$

where the $o(1)$ represents a function of H that converges to zero uniformly over $\mathcal{H}$.

Asymptotics. With (12.2) we have enough to establish an almost sure limit result for $\widetilde{R}_n(\widetilde{\omega}, \theta) = R_n(\phi_n(\widetilde{\omega}), \theta)$, which will imply the corresponding distributional result for $R_n(\omega, \theta)$. Let $\{\delta_n\}$ be a sequence of random variables on $\widetilde{\Omega}$ that

converges almost surely to zero at a rate to be specified soon. Define

$$Z(\theta) = \tilde{\nu}\big(H(r_0,\theta)\big)/\Phi'(r_0),$$
$$\ell_n(\theta) = r_0 - \frac{1}{\sqrt{n}}\big(Z(\theta) + \delta_n\big),$$
$$u_n(\theta) = r_0 - \frac{1}{\sqrt{n}}\big(Z(\theta) - \delta_n\big).$$

If we can find δ_n uniformly of order $o(1)$ such that, eventually,

$$\ell_n(\theta) \le \widetilde{R}_n(\theta) \le u_n(\theta) \qquad \text{for all } \theta,$$

then it will follow that

$$\sqrt{n}(\widetilde{R}_n(\theta) - r_0) \to -Z(\theta) \qquad \text{uniformly in } \theta,$$

as desired.

Consider first the upper bound on $\widetilde{R}_n(\theta)$. Temporarily write $H_n(\theta)$ for the half-space $H(u_n(\theta),\theta)$. Then

$$\widetilde{P}_n H_n(\theta) = PH_n(\theta) + \frac{1}{\sqrt{n}}\big(\tilde{\nu}H_n(\theta) + o(1)\big) \qquad \text{uniformly in } \theta.$$

Apply the Mean Value Theorem to approximate the contribution from P:

$$\begin{aligned} PH_n(\theta) &= \Phi(u_n(\theta)) \\ &= \Phi(r_0) + (u_n(\theta) - r_0)\big(\Phi'(r_0) + o(1)\big) \\ &= \tfrac{3}{4} - \frac{1}{\sqrt{n}}\Big(\tilde{\nu}H(r_0,\theta) - o(1) - \big(\Phi'(r_0) + o(1)\big)\delta_n\Big), \end{aligned}$$

where the $o(1)$ represent functions of θ that converge to zero uniformly in θ. For the contribution from $\tilde{\nu}$ consider first the difference $|H_n(\theta) - H(r_0,\theta)|$. It is the indicator function of a strip of width $|Z(\theta) - \delta_n|/\sqrt{n}$; its P measure converges to zero uniformly in θ. Thus

$$\rho\big(H_n(\theta), H(r_0,\theta)\big) \to 0 \qquad \text{uniformly in } \theta.$$

By the uniform continuity of the $\tilde{\nu}$ sample paths it follows that

$$\tilde{\nu}(H_n(\theta)) = \tilde{\nu}H(r_0,\theta) + o(1) \qquad \text{uniformly in } \theta.$$

Adding the two contributions to $\widetilde{P}_n H_n(\theta)$ we get

$$\widetilde{P}_n H_n(\theta) = \tfrac{3}{4} + \frac{1}{\sqrt{n}}\Big(\big(\Phi'(r_0) + o(1)\big)\delta_n - o(1)\Big).$$

We can choose δ_n converging to zero while ensuring that the coefficient of $1/\sqrt{n}$ is always positive. With that choice, the set $H_n(\theta)$ becomes one of the half spaces whose intersection defines $\widetilde{K}_n$; the boundary point in the θ direction must lie on the ray from the origin to the boundary of $H_n(\theta)$; the distance $\widetilde{R}_n(\theta)$ must be less than $u_n(\theta)$.

Now consider the lower bound on $\widetilde{R}_n(\theta)$. Let $\mathbf{t}_n(\theta)$ denote the point a distance $\ell_n(\theta)$ from the origin in the θ direction. It is enough if we show that $\widetilde{K}_n$ contains every $\mathbf{t}_n(\theta)$.

If, for a particular θ, the point $\mathbf{t}_n(\theta)$ were outside $\widetilde{K}_n$, there would exist a halfspace H with $\widetilde{P}_n H \geq 3/4$ and $\mathbf{t}_n(\theta) \notin H$. By sliding H towards $\mathbf{t}_n(\theta)$ we would get an H' with $\widetilde{P}_n H' \geq 3/4$ and $\mathbf{t}_n(\theta)$ on the boundary of H'. The right choice for δ_n will ensure that such an H' cannot exist.

For each θ let $H_n(\theta)$ denote the halfspace with $\mathbf{t}_n(\theta)$ on its boundary and the largest $\widetilde{P}_n$ measure. (Of course this is not the same $H_n(\theta)$ as before.) The maximum of PH over all halfspaces with $\mathbf{t}_n(\theta)$ on the boundary is achieved at $H(\ell_n(\theta), \theta)$. So, uniformly in θ,

$$\tfrac{3}{4} \leq \widetilde{P}_n H_n(\theta) = PH_n(\theta) + O(1/\sqrt{n}) \leq PH(\ell_n(\theta), \theta) + O(1/\sqrt{n}) \to \tfrac{3}{4}.$$

It follows that $PH_n(\theta)$ also converges uniformly to $3/4$. This forces the boundary of $H_n(\theta)$ to orient itself more and more nearly perpendicular to the θ direction. Consequently,

$$\rho\big(H_n(\theta), H(r_0, \theta)\big) \to 0 \qquad \text{uniformly in } \theta.$$

Uniform continuity of the $\widetilde{\nu}$ sample paths now lets us assert

$$\widetilde{P}_n H_n(\theta) = PH_n(\theta) + \frac{1}{\sqrt{n}}\Big(\widetilde{\nu} H(r_0, \theta) + o(1)\Big) \qquad \text{uniformly in } \theta.$$

Again using the fact that the maximum of PH over all halfspaces with $\mathbf{t}_n(\theta)$ on the boundary is achieved at $H(\ell_n(\theta), \theta)$, we deduce that, uniformly in θ,

$$\begin{aligned} PH_n(\theta) &\leq PH(\ell_n(\theta), \theta) \\ &= \Phi(\ell_n(\theta)) \\ &= \Phi(r_0) + (\ell_n(\theta) - r_0)\Big(\Phi'(r_0) + o(1)\Big) \\ &= \tfrac{3}{4} - \frac{1}{\sqrt{n}}\Big(\widetilde{\nu} H(r_0, \theta) - o(1) + (\Phi'(r_0) + o(1))\delta_n\Big). \end{aligned}$$

With δ_n converging to zero slowly enough to cancel out all the $o(1)$ terms, plus a little bit more, we get a contradiction, $\widetilde{P}_n H_n(\theta) < 3/4$ for all θ. There can therefore be no halfspace with $\widetilde{P}_n H' \geq 3/4$ and $\mathbf{t}_n(\theta)$ on its boundary. The point $\mathbf{t}_n(\theta)$ must lie inside K_n. The argument for the lower bound on $\widetilde{R}_n(\theta)$ is complete.

REMARKS. Nolan (1989b) has studied an estimator related to K_n, following Donoho (1982). Its analysis is similar to the arguments given in this section, but more delicate.