Geophysical Statistics Project
National Center for Atmospheric Research

Presentation for 7400, April 14th 2003

U-Statistics and PWM in Modeling Extremes

Reinhard Furrer, GSP-CGD, NCAR

- Motivation for the project
- Description of the project
- Outline of the talk

Motivation for the Project

To model extremes, we often use the distributions

$$\text{GEV}(\mu, \sigma, \xi) \qquad \text{GP}(\sigma, \xi)$$

The parameters can be estimated with:

Maximum likelihood (ML):
can be generalized, seasonality in parameters, dependence structure, well established theory, non-regular setting, ...

Probability weighted moments (PWM):
used in hydrology, simple and works well for small n, not well elaborated theory, ...

What if the data are dependent?

Description of the Project

We consider PWM estimators as U-statistics.

Then we proceed:

- familiarize with U-statistics
- use existing literature to derive
  - asymptotic properties in iid case
  - asymptotic properties in stationary case
- develop theory to 'extreme' context
- ...

Outline of the Talk

U-statistics

- Motivation with questions and answers
- Definition of an U-statistic
- Variance of an U-statistic
- H-Decomposition
- Asymptotics for iid and stationary case
- Conclusion and outlook

Fundamental Questions

Consider a functional θ defined on a set $\mathcal{F}$ of distributions:

$$\theta = \theta(F), \qquad F \in \mathcal{F}.$$

Suppose we wish to estimate θ based on a random sample $X_1, \ldots, X_n$ iid F, with F unknown, but $\mathcal{F}$ known.

1. Does there exist an unbiased estimator of θ, independent of the distribution F?
   Can we characterize the sets $\mathcal{F}$ and the functionals θ for which the answer is yes?

2. If such an estimator exists, what is it?
   If several exist, which is the best?

Surprising Answers

Theorem

A functional θ admits an unbiased estimator if and only if there exists a function ψ of k variables such that

$$\theta(F) = \int \cdots \int \psi(x_1, \ldots, x_k)\, \mathrm{d}F(x_1) \cdots \mathrm{d}F(x_k)$$

Theorem

There exists an unique symmetric unbiased estimator $\psi^{[n]}$.

Theorem

For any other unbiased estimator ϕ, $\mathrm{Var}(\phi) \geq \mathrm{Var}(\psi^{[n]})$.

Definition of a U-Statistic

Suppose there exists a function $\psi(x_1, \ldots, x_k)$, called kernel of degree k.
The estimate

$$\psi^{[n]}(x_1, \ldots, x_n) = \frac{(n-k)!}{n!} \sum_{\mathcal{S}} \psi(x_{i_1}, \ldots, x_{i_k}),$$

is unbiased and essentially unique.
Symmetrize the kernel $\psi^{[k]}(x_1, \ldots, x_k) = \sum \psi(x_{i_1}, \ldots, x_{i_k})$, then the unique unbiased symmetric estimators of θ are

$$\hat{\theta} = U_n = \binom{n}{k}^{-1} \sum_{(n,k)} \psi^{[k]}(X_{i_1}, \ldots, X_{i_k})$$

$\rightsquigarrow$ U-statistics

Examples 123: Mean

The functional

$$\theta(F) = \int x f(x)\, \mathrm{d}x$$

suggests the kernel $\psi(x) = x$.

We note that

$$\hat{\theta} = \binom{n}{k}^{-1} \sum_i \psi(X_i) = \frac{1}{n} \sum_i X_i = \bar{X}$$

Examples 123: Variance

We can define the variance by

$$\int \int \frac{1}{2}(x_1 - x_2)^2\, \mathrm{d}F(x_1)\, \mathrm{d}F(x_2)$$

which is estimated by

$$U_n = \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} \frac{1}{2}(X_1 - X_2)^2$$

No surprises: $U_n = s_n^2$.

Examples 123: PWM

The density of $\max(X_1, \ldots, X_{r+1})$ is $(r+1)F(x)^r f(x)$, thus

$$\beta_r = \mathrm{E}\Big(XF(X)^r\Big) = \int x F(x)^r f(x)\, \mathrm{d}x$$
$$= \frac{1}{r+1} \mathrm{E}\Big(\max(X_1, \ldots, X_{r+1})\Big)$$

And we have the U-statistic

$$U_n = \hat{\beta}_r = \binom{n}{r+1}^{-1} \sum_{(n,r+1)} \frac{1}{r+1} \max(X_{i_1}, \ldots, X_{i_r})$$

Variance of an U-Statistic

Define

$$\psi_c(x_1, \ldots, x_c) = \mathrm{E}\Big(\psi(x_1, \ldots, x_c, X_{c+1}, \ldots, X_k)\Big) \quad c = 1, \ldots, k$$
$$\sigma_c^2 = \mathrm{Var}\Big(\psi_c(X_1, \ldots, X_c)\Big)$$

Theorem

$$\mathrm{Var}(U_n) = \binom{n}{k}^{-1} \sum_{c=1}^{k} \binom{k}{c}\binom{n-k}{k-c} \sigma_c^2.$$

- Proof is based on enumeration techniques
- Same techniques to calculate covariances
- Upper bounds for higher moments

Examples 12: Variance

Recall that $\psi(x_1, x_2) = \frac{1}{2}(x_1 - x_2)^2$.

$$\begin{aligned}\psi_1(x_1) &= \frac{1}{2}\Big(\sigma^2 + (x_1 - \mu)^2\Big)\\ \sigma_1^2 &= \frac{1}{4}(\mu_4 - \sigma^4)\\ \sigma_2^2 &= \frac{1}{2}(\mu_4 + \sigma^4)\end{aligned}$$

which leads to

$$\mathrm{Var}(s_n^2) = \binom{n}{2}^{-1}\Big(2(n-2)\sigma_1^2 + \sigma_2^2\Big) = \frac{\mu_4}{n} - \frac{(n-3)\sigma^4}{n(n-1)}$$

Examples 12: PWM

Recall that $\psi(x_1, x_2) = \frac{1}{2}\max(x_1, x_2)$.

$$\psi_1(x_1) = \frac{1}{2}\Big(x_1 F(x_1) + \int_{x_1}^{\infty} x_2 \,\mathrm{d}F(x_2)\Big) = \frac{1}{2}\Big(x_1 F(x_1) + g(x_1)\Big)$$

Then

$$\sigma_1^2 = \frac{1}{4}\mathrm{Var}\Big(XF(X) + g(X)\Big)$$

Many distributions do not have an analytic form for $g(\cdot)$.
For the GPD we have

$$\int_x^{\infty} y \,\mathrm{d}F(y) = \frac{\sigma + x}{1 - \xi}\Big(1 - F(x)\Big)$$

H-Decomposition 12: Definition

Define recursively the kernels

$$\begin{aligned}h^{(1)}(x_1) &= \psi_1(x_1) - \theta\\ h^{(c)}(x_1, \ldots, x_c) &= \psi_c(x_1, \ldots, x_c) - \sum_{j=1}^{c-1}\sum_{(c,j)} h^{(j)}(x_{i_1}, \ldots, x_{i_j}) - \theta\end{aligned}$$

To each kernel $h^{(c)}$ we let $H_n^{(c)}$ be the associated U-statistic.

Theorem

$$U_n = \theta + \sum_{j=1}^{k}\binom{k}{j} H_n^{(j)}.$$

H-Decomposition 12: Properties

- Is a representation of a U-statistic of degree k in terms of uncorrelated U-statistics of degree $1, \ldots, k$
- The terms have variances of decreasing order in n.
- Defining $\delta_j^2 = \mathrm{Var}\Big(h^{(j)}(X_1, \ldots, X_j)\Big) = 0$, we have

$$\mathrm{Var}(U_n) = \sum_{j=1}^{k}\binom{k}{j}^2\binom{n}{j}^{-1}\delta_j^2$$

- σ_c^2 are a linear combination of δ_j^2.

Example: Variance

Recall that $\psi(x_1, x_2) = \frac{1}{2}(x_1 - x_2)^2$.

$$\begin{aligned}h^{(1)}(x_1) &= \psi_1(x_1) - \theta = \frac{1}{2}\Big(\sigma^2 + (x_1 - \mu)^2\Big) - \sigma^2\\ &= \frac{1}{2}\Big((x_1 - \mu)^2 - \sigma^2\Big)\\ h^{(2)}(x_1, x_2) &= \psi(x_1, x_2) - \psi_1(x_1) - \psi_1(x_2) + \theta\\ &= -(x_1 - \mu)(x_2 - \mu)\end{aligned}$$

Thus

$$s_n^2 = \sigma^2 + \Big(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - \sigma^2\Big) + \binom{n}{2}^{-1}\sum_{i<j}(X_i - \mu)(X_j - \mu)$$

Asymptotics 12: iid Case

Theorem
Let $\sigma_1^2 > 0$. Then $\sqrt{n}(U_n - \theta) \xrightarrow{\mathcal{D}, n\to\infty} \mathcal{N}(0, k^2\sigma_1^2)$.

- The proof is based on the H-decomposition:

$$\begin{aligned}\sqrt{n}(U_n - \theta) &= \sqrt{n}(kH_n^{(1)} + R_n)\\ &= \frac{k}{\sqrt{n}}\sum_{i=1}^{n} kh^{(1)}(X_i) + \sqrt{n}R_n\end{aligned}$$

- Multivariate version exists.

Asymptotics 12ab: Stationary Case

We need to characterize the type of stationarity.

For example absolutely regular processes:

The stationary process is said to be absolutely regular if

$$\beta(n) = \mathsf{E} \sup_{A \in \mathcal{M}(t+n,\infty)} \Big| \mathsf{P}\{A|\mathcal{M}(-\infty,t)\} - \mathsf{P}\{A\} \Big| \xrightarrow{n\to\infty} 0$$

with $\mathcal{M}(a,b)$ the σ-field generated by events of the form $\Big\{\omega : \big(X_{t_1}(\omega), \ldots, X_{t_k}(\omega)\big) \in \mathcal{B}\Big\}$, where $a \le t_1 < \cdots < t_k = b$ for $k = 1, \ldots, 1 + b - a$ and $\mathcal{B}$ is a k-dimensional Borel set.

Asymptotics 12ab: Stationary Case

Theorem

Let X_t be a stationary process satisfying $\beta(n) = \mathcal{O}(n^{-(2+\epsilon)/\epsilon})$, $\epsilon > 0$, and let U_n be a U-statistic based on a kernel of degree 2 satisfying

$$\sup_{i,j} \mathsf{E}\Big|\psi(X_i, X_j)\Big|^{2+\delta} < \infty \qquad \delta < \epsilon$$

$$\int\int \Big|\psi(x_1, x_2)\Big|^{2+\delta} \,\mathrm{d}F(x_1)\,\mathrm{d}F(x_2) < \infty$$

Then $\sqrt{n}(U_n - \theta)/(2\sigma)$ tends asymptotically to a standard normal distribution.

- The proof is based on the H-decomposition.
- Can be extended to strong mixing processes.

Further Research

- Apply asymptotics to PWM
- Rewrite theorem for $D(u)$ and $D'(u)$
- Lets get started . . .