Ranking Elite Swimmers using Extreme Value Theory

Harry Spearing
Supervisors: Jon Tawn*, Tim Paulden†, David Irons†, Grace Bennett†
*Lancaster University, †ATASS sports

June 29, 2019
Motivation

FINA rankings are currently used to compare swimmers across all 34 Olympic events. Based on a points system, $p_{i,j} \propto (b_j/t_i)^3$, the points $p_{i,j}$ given to swimmer $i$ in event $j$ is where $b_j$ is the base time for event $j$. Very sensitive to changes in world record. Bias between events.
Motivation

- FINA rankings are currently used to compare swimmers across all 34 Olympic events.
Motivation

- FINA rankings are currently used to compare swimmers across all 34 Olympic events.
- Based on points system,
  \[ p_{i,j} \propto (b_j/t_{i,j})^3 \]
  
  \( p_{i,j} \) given to swimmer \( i \) in event \( j \) is where \( b_j \) is the base time for event \( j \).
Motivation

- FINA rankings are currently used to compare swimmers across all 34 Olympic events.
- Based on points system,
  \[ p_{i,j} \propto (b_j/t_{i,j})^3 \]
  where \( p_{i,j} \) given to swimmer \( i \) in event \( j \) is where \( b_j \) is the base time for event \( j \).
- Very sensitive to changes in world record.
Motivation

- FINA rankings are currently used to compare swimmers across all 34 Olympic events.
- Based on points system,
  \[ p_{i,j} \propto \left( \frac{b_j}{t_{i,j}} \right)^3 \]
  \( p_{i,j} \) given to swimmer \( i \) in event \( j \) is where \( b_j \) is the base time for event \( j \).
- Very sensitive to changes in world record.
- Bias between events.
A good ranking system should be **robust** to changes in the data, and quantify **uncertainty** in the ranks.

Other features of interest:

- What is the ultimate possible swim-time for a given event?
- What will be the swim-time of the next world record?
- When will current world records next be broken?
- Which event will next have a new record set?
- Can we adjust for technological advances?
A good ranking system should be robust to changes in the data, and quantify uncertainty in the ranks. Other features of interest:

▶ What is the ultimate possible swim-time for a given event?
A good ranking system should be robust to changes in the data, and quantify uncertainty in the ranks. Other features of interest:

▶ What is the ultimate possible swim-time for a given event?
▶ What will be the swim-time of the next world record?
A good ranking system should be **robust** to changes in the data, and quantify **uncertainty** in the ranks.

Other features of interest:

- What is the ultimate possible swim-time for a given event?
- What will be the swim-time of the next world record?
- When will current world records next be broken?
A good ranking system should be **robust** to changes in the data, and quantify **uncertainty** in the ranks.

Other features of interest:

▶ What is the ultimate possible swim-time for a given event?
▶ What will be the swim-time of the next world record?
▶ When will current world records next be broken?
▶ Which event will next have a new record set?
A good ranking system should be **robust** to changes in the data, and quantify **uncertainty** in the ranks.

Other features of interest:

- What is the ultimate possible swim-time for a given event?
- What will be the swim-time of the next world record?
- When will current world records next be broken?
- Which event will next have a new record set?
- Can we adjust for technological advances?
Figure: Data for the men’s 100m butterfly. The fastest 200 times over the period 2001 to late 2018.
Figure: Rate of observations
Figure: Distribution of swim-times.
Want to model both rate and distribution of these extreme events. By definition, there are very few observations. Data of the swimmer population not available. Even if it was, this would not be useful! A Separate framework for analysing extreme events is required.
Theory

▶ Want to model both rate and distribution of these extreme events.
Theory

- Want to model both *rate* and *distribution* of these extreme events.
- By definition, there are very few observations.
Theory

- Want to model both rate and distribution of these extreme events.
- By definition, there are very few observations.
- Data of the swimmer population not available.

A separate framework for analyzing extreme events is required.
Theory

- Want to model both rate and distribution of these extreme events.
- By definition, there are very few observations.
- Data of the swimmer population not available.
- Even if it was, this would not be useful!
Theory

- Want to model both rate and distribution of these extreme events.
- By definition, there are very few observations.
- Data of the swimmer population not available.
- Even if it was, this would not be useful!
Theory

▶ Want to model both rate and distribution of these extreme events.
▶ By definition, there are very few observations.
▶ Data of the swimmer population not available.
▶ Even if it was, this would not be useful!

▶ A Separate framework for analysing extreme events is required.
Extreme Value Theory

**Figure:** Block maxima

**Generalised extreme value (GEV) distribution function:**
\[
G(x) = \exp\left(-\left[1 + \frac{\xi (x - \mu)}{\sigma}\right]^{-\frac{1}{\xi}}\right),
\]
\(\mu, \xi \in \mathbb{R}, \sigma \in \mathbb{R}^+\).

**Generalised pareto distribution (GPd) function:**
\[
H_u(x) = 1 - \left[1 + \xi \left(\frac{x - u}{\tilde{\sigma}_u}\right)\right]^{-\frac{1}{\xi}},
\]
\(\xi \in \mathbb{R}, \tilde{\sigma}_u \in \mathbb{R}^+\).
Extreme Value Theory

Figure: Block maxima

Generalised extreme value (GEV) distribution function:

\[ G(x) = \exp \left( -\left[ 1 + \xi \frac{x - \mu}{\sigma} \right]^{1/\xi}_+ \right), \]

\( \mu, \xi \in \mathbb{R}, \sigma \in \mathbb{R}^+ \).
Generalised extreme value (GEV) distribution function:

\[ G(x) = \exp \left( -\left[ 1 + \frac{\xi (x - \mu)}{\sigma} \right]^{-1/\xi} \right), \]

\( \mu, \xi \in \mathbb{R}, \sigma \in \mathbb{R}^+ \).
Extreme Value Theory

**Figure:** Block maxima

**Figure:** Peaks over threshold, \( u \)

Generalised extreme value (GEV) distribution function:

\[
G(x) = \exp \left( -\left[ 1 + \xi (x - \mu) / \sigma \right]_{+}^{-1/\xi} \right),
\]

\( \mu, \xi \in \mathbb{R}, \sigma \in \mathbb{R}^+ \).

Generalised pareto distribution (GPd) function:

\[
H_u(x) = 1 - \left[ 1 + \xi \left( \frac{x - u}{\tilde{\sigma}_u} \right) \right]_{+}^{-\frac{1}{\xi}},
\]

\( \xi \in \mathbb{R}, \tilde{\sigma} \in \mathbb{R}^+ \).
Extreme Value Theory

**Figure:** Block maxima

**Figure:** Peaks over threshold, $u$

**Figure:** GEV density function

**Figure:** GP density function
Extreme Value Theory

Figure: Change in location, $\mu$

Figure: Change in scale, $\sigma$

Figure: Change in shape, $\xi$

Figure: Change in threshold, $u$
Point process

Point process framework combines annual maxima (GEV), and peaks over threshold (GPd), to allow for non-homogeneous rates of occurrence.

Connects GEV parameters ($\mu, \sigma, \xi$) to GPd parameters ($\tilde{\sigma}, \xi$) via:

$$\tilde{\sigma} = \sigma + \xi (u - \mu).$$

Model both rate and distribution of extreme data, for any event.

Observations occur via a Poisson random variable with rate $R(\mu, \sigma, \xi | x, t)$ for a swim-time $x$ at time $t$. 
Point process

- Point process framework combines annual maxima (GEV), and peaks over threshold (GPd), to allow for non-homogeneous rates of occurrence.

\[
\tilde{\sigma} = \sigma + \xi (u - \mu)
\]

- Model both rate and distribution of extreme data, for any event.

- Observations occur via a Poisson random variable with rate \( R(\mu, \sigma, \xi | x, t) \) for a swim-time \( x \) at time \( t \).
Point process

- Point process framework combines annual maxima (GEV), and peaks over threshold (GPd), to allow for non-homogeneous rates of occurrence.

- Connects GEV parameters \((\mu, \sigma, \xi)\) to GPd parameters \((\tilde{\sigma}, \xi)\) via:
  \[
  \tilde{\sigma} = \sigma + \xi(u - \mu).
  \]
Point process

- Point process framework combines annual maxima (GEV), and peaks over threshold (GPd), to allow for non-homogeneous rates of occurrence.

- Connects GEV parameters \((\mu, \sigma, \xi)\) to GPd parameters \((\tilde{\sigma}, \xi)\) via:

  \[
  \tilde{\sigma} = \sigma + \xi(u - \mu).
  \]

- Model both rate and distribution of extreme data, for any event.
Point process

- Point process framework combines annual maxima (GEV), and peaks over threshold (GPd), to allow for non-homogeneous rates of occurrence.

- Connects GEV parameters \((\mu, \sigma, \xi)\) to GPd parameters \((\tilde{\sigma}, \xi)\) via:
  \[
  \tilde{\sigma} = \sigma + \xi(u - \mu).
  \]

- Model both \textbf{rate} and \textbf{distribution} of extreme data, for any event.

- Observations occur via a Poisson random variable with rate \(R(\mu, \sigma, \xi|x, t)\) for a swim-time \(x\) at time \(t\).
Model: independent events

Figure: Data for the men’s 100m butterfly.
Model: independent events

Figure: Data for the men’s 100m butterfly.

- A trend in the rates of occurrences.
Model: independent events

Figure: Data for the men’s 100m butterfly.

- A trend in the rates of occurrences.
- Observations above threshold have a time independent distribution.
Model: independent events

Figure: Data for the men’s 100m butterfly.

- A trend in the rates of occurrences.
- Observations above threshold have a time independent distribution.
- Swim-suit effect.
Model: independent events

This is achieved via two additional parameters:

\((\xi^{(e)}, \mu^{(e)}_0, \sigma^{(e)}_0, \beta^{(e)}, \gamma^{(e)})\),
Model: independent events

This is achieved via two additional parameters:

\[(\xi^{(e)}, \mu_0^{(e)}, \sigma_0^{(e)}, \beta^{(e)}, \gamma^{(e)})\],

such that:

\[
\xi^{(e)}(t) = \xi^{(e)}, \\
\mu^{(e)}(t) = \mu_0^{(e)} + \beta^{(e)} t + \gamma^{(e)} 1_{\{t \in S_t\}}, \\
\sigma^{(e)}(t) = \sigma_0^{(e)} + \xi^{(e)} \beta^{(e)} t + \xi^{(e)} \gamma^{(e)} 1_{\{t \in S_t\}},
\]
...independent event model

- Have model with five parameters: \( \mu_0, \sigma_0, \xi, \beta, \gamma \).

- For all 34 Olympic swimming events, this means 170 parameters in total.

- Can improve model robustness and predictive power by pooling information across events.

- Try introducing distance as a covariate.

- A log-log relationship with distance works well for athletics, but...
…independent event model

▶ Have model with five parameters: $\mu_0$, $\sigma_0$, $\xi$, $\beta$, $\gamma$.
▶ For all 34 Olympic swimming events, this means 170 parameters in total.
...independent event model

- Have model with five parameters: $\mu_0$, $\sigma_0$, $\xi$, $\beta$, $\gamma$.
- For all 34 Olympic swimming events, this means 170 parameters in total.
- Can improve model **robustness** and **predictive power** by pooling information across events.
...independent event model

- Have model with five parameters: $\mu_0, \sigma_0, \xi, \beta, \gamma$.
- For all 34 Olympic swimming events, this means 170 parameters in total.
- Can improve model **robustness** and **predictive power** by pooling information across events.
- Try introducing **distance** as a covariate.
...independent event model

- Have model with five parameters: $\mu_0$, $\sigma_0$, $\xi$, $\beta$, $\gamma$.
- For all 34 Olympic swimming events, this means 170 parameters in total.
- Can improve model robustness and predictive power by pooling information across events.
- Try introducing distance as a covariate.
- A log-log relationship with distance works well for athletics, but..
Across event model

- The different strokes mean this isn’t much of an improvement.

**Figure**: Transformed parameter estimates against log threshold swim-time $\log(u)$: $\log(\tilde{\sigma}), \xi, \log(\beta), \log(-\mu_0), \sqrt{\gamma}$. 

[Graph showing transformed parameter estimates against log distance]
Across event model

Therefore, use the threshold time as a covariate:

Figure: Transformed parameter estimates against log threshold swim-time, $\log(u)$: $\log(\tilde{\sigma})$, $\xi$, $\log(\beta)$, $\log(-\mu_0)$, $\sqrt{\gamma}$.
Across event model

These pooled estimates would result in model:

\[
\xi^{(e)} = \xi, \\
\log \left( \mu^{(e)} \right) = \alpha_1 + \theta_1 \tilde{u}_e, \\
\log \left( \beta^{(e)} \right) = \alpha_2 + \theta_2 \tilde{u}_e, \\
\sqrt{\gamma^{(e)}} = \alpha_3 + \theta_3 \tilde{u}_e, \\
\log \left( \tilde{\sigma}_u^{(e)} \right) = \alpha_4 + \theta_4 \tilde{u}_e,
\]

\[\tilde{u}_e = \log (u_e).\]
Across event model

These pooled estimates would result in model:

\[ \xi^{(e)} = \xi, \]
\[ \log \left( \mu^{(e)} \right) = \alpha_1 + \theta_1 \tilde{u}_e, \]
\[ \log \left( \beta^{(e)} \right) = \alpha_2 + \theta_2 \tilde{u}_e, \]
\[ \sqrt{\gamma^{(e)}} = \alpha_3 + \theta_3 \tilde{u}_e, \]
\[ \log \left( \tilde{\sigma}_u^{(e)} \right) = \alpha_4 + \theta_4 \tilde{u}_e, \]

\[ \tilde{u}_e = \log (u_e). \]
Splines

Figure: different degree B-spline.
Figure: different degree B-spline.
Splines

Figure: different degree B-spline.

\[
\log (\tilde{\sigma}(\tilde{u})) = \sum_{k=1}^{q} a_k B_k(\tilde{u})
\]

where \(a_k\) is the \(k^{th}\) element of the spline coefficient vector \(a\).
Final model

**Figure:** parameters for the final model: $\log(\tilde{\sigma})$, $\xi$, $\log(\beta)$, $\log(-\mu_0)$, $\sqrt{\gamma}$.

- $\xi^{(e)} = \xi$,
- $\log(-\mu^{(e)}) = \alpha_1 + \theta_1 \tilde{u}_e$,
- $\log(\beta^{(e)}) = \alpha_2 + \theta_2 \tilde{u}_e$,
- $\sqrt{\gamma^{(e)}} = \alpha_3 + \theta_3 \tilde{u}_e$,
- $\log(\tilde{\sigma}_u^{(e)}) = \sum_{k=1}^{q} a_k B_k(\tilde{u}_e)$,
Model fits: rate of observations

Figure: Data for the men’s 100m butterfly.

Figure: Fitted rate of occurrences.
Model fits: distribution of observations.

Figure: Fitted distribution above threshold, men’s 100m butterfly: Threshold, World record, Ultimate possible swim-time.
Results: Rankings

Can define ranking based on $\Pr\{X_e > x\}$, which gives:
Results: Rankings

Can define ranking based on $Pr\{X_e > x\}$, which gives:
Results: Expected times

Figure: Expected next world record swim-time. Ultimate possible swim-time
Results: Time of next record
Results: Probability of next record in a given event
If $r_e := \max(x_e)$, then

$$\Pr\{ T^{(e)} < t \} = 1 - \Pr\{ T^{(e)} > t \} \quad (1)$$

$$= 1 - \sum_{m=0}^{\infty} \Pr\{ \max(X_{1:N_t}^{(e)}) < r_e | N_t = m \} \Pr\{ N_t = m \}$$

$$= 1 - \sum_{m=0}^{\infty} \left[ H_{ue}^{(e)}(r_e) \right]^m \left[ \Lambda^{(e)}(A_{(1,t),ue}) \right]^m \exp\left[ -\Lambda^{(e)}(A_{(1,t),ue}) \right] / m!$$

$$= 1 - \exp \left\{ -\Lambda^{(e)}(A_{(1,t),ue}) H_{ue}^{(e)}(r_e) \right\}$$

$$= F_{T^{(e)}}(t),$$
...derivations

Let

\[ T^{(-e)} := \min_k \{ T^{(k)}, k \in E \setminus \{e\} \}. \]

Then the probability that the next world record is set in event \( e \) is

\[
\Pr\{ T^{(-e)} > T^{(e)} \}
= \int_1^\infty \Pr\{ T^{(-e)} > T^{(e)} \mid T^{(e)} = t \} \Pr\{ T^{(e)} = t \} \, dt
= \int_1^\infty \prod_{k \in E \setminus \{e\}} \left\{ \exp \left[ -\Lambda^{(k)}(A_{(1,t)}, u_k) \, \tilde{H}^{(k)}_{u_k}(r_k) \right] \right\} \left[ 1 + \xi \left( \frac{u_e - \mu^{(e)}(t)}{\sigma^{(e)}(t)} \right) \right]^{-\frac{1}{\xi}} \tilde{H}^{(e)}_{u_e}(r_e) \exp \left[ -\Lambda^{(e)}(A_{(1,t), u_e}) \, \tilde{H}^{(e)}_{u_e}(r_e) \right] \, dt
= \int_1^\infty \left\{ \exp \left[ -\sum_{k \in E} \Lambda^{(k)}(A_{(1,t), u_k}) \, \tilde{H}^{(k)}_{u_k}(r_e) \right] \right\} \left[ 1 + \xi \left( \frac{u_e - \mu^{(e)}(t)}{\sigma^{(e)}(t)} \right) \right]^{-\frac{1}{\xi}} \tilde{H}^{(e)}_{u_e}(r_e) \right\}
\]

where the second equality follows because

\[
\Pr\{ T^{(-e)} > T^{(e)} \mid T^{(e)} = t \} = \prod_{k \in E \setminus \{e\}} \left\{ \exp \left[ -(\Lambda^{(k)}(A_{(1,t), u_k}) \, \tilde{H}^{(k)}_{u_k}(r_k) \right] \right\}
\]
Figure: Pooled pp plot over all events, with 95 % tolerance interval.
Figure: The expected number of observations for Women’s 100m freestyle and 95% confidence intervals, against the number of observations in the data (red crosses).
\[ L(\theta; \mathbf{x}, \mathbf{t}) = \prod_{e \in E} \left\{ \exp \left[ -\Lambda^{(e)}(\mathcal{A}_1, u_e) \right] \prod_{i=1}^{200} \int_{x_i^{(e)} - s_i/2}^{x_i^{(e)} + s_i/2} \lambda^{(e)}(t, x) \, dx \right\} \exp \left[ -(\phi_1 p_r + \phi_2 p_m) \right] \]
Figure: Expected next world record swim-time. Ultimate possible swim-time