Comparing dominance of tennis' big three

via multiple-output Bayesian quantile regression models

Bruno Santos
University of Kent

Introduction

Big Three

- Roger Federer
- Rafael Nadal
- Novak Djokovic

- Won63 out of 77 Grand Slam tournaments, between Wimbledon in 2003 until 2022.

Dominance

- List of Grand Slam Winners:

January-2022

Player	Titles
1. Roger Federer	20
1. Rafael Nadal	20
1. Novak Djokovic	20
4. Pete Sampras	14
5. Roy Emerson	12

Currently, August-2022

Player	Titles
1. Rafael Nadal	22
2. Novak Djokovic	21
3. Roger Federer	20
4. Pete Sampras	14
5. Roy Emerson	12

Question: Who is more dominant between the Big Three?

How to measure dominance in a tennis match

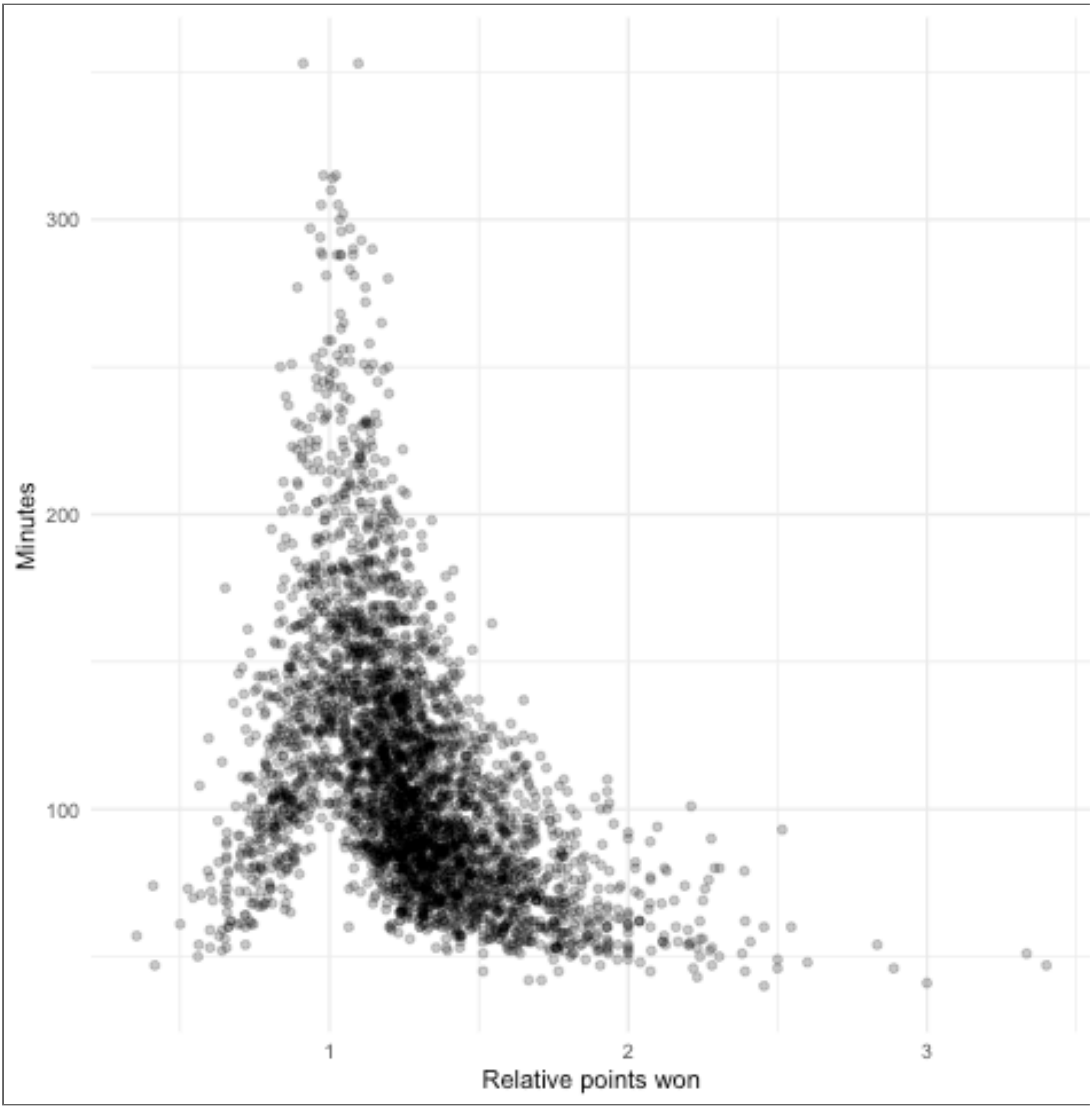
Important notes:

- A tennis match is divided into sets and games.
- A player with most sets wins the match.
- A player can win more games, but still lose the match.
  - Example: 7-6, 0-6, 7-6.
- Solution:
  - **Relative points:** ratio points won/lost in a match.
  - **Duration** of the match.

Data

- Data organised by Jeff Sackmann in the repository:
  - **https://github.com/JeffSackmann/tennis_atp**
- All matches from the Big Three, between 1998 and the US Open in 2021.
  - Excluding Davis Cup and Olympic Games matches.
  - Also matches played on carpet.
- We should condition on some variables:
  - type of tournament (Grand Slam, Masters 1000, …);
  - surface (clay, grass and hard courts);
  - wins and losses;
  - rank of opponent.

Data distribution



Bayesian quantile regression for multiple output response variables

Directional quantile regression model

- Response variable is defined as $\boldsymbol{Y} \in \mathbb{R}^k$.
- Directional index can be defined by $\boldsymbol{\tau} \in \mathcal{B}^k := \{\boldsymbol{v} \in \mathbb{R}^k : 0 < ||\boldsymbol{v}|| < 1.\}$.
  - $\boldsymbol{\tau} = \tau \boldsymbol{u}, \tau \in (0, 1)$.
  - Direction: $\boldsymbol{u} \in \mathcal{S}^{k-1} := \{z \in \mathbb{R}^k : ||z|| = 1\}$;
- Define $\boldsymbol{\Gamma}_u$, an arbitrary $k \times (k-1)$ matrix of unit vectors.
  - $(u \vdots \Gamma_u)$ is an orthonormal basis of $\mathbb{R}^k$.

DEFINITION:

The $\boldsymbol{\tau}$th quantile of $\boldsymbol{Y}$ is the τth quantile hyperplane obtained from the regression:

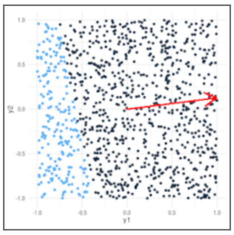
- $\boldsymbol{Y}_u := \boldsymbol{u}' \boldsymbol{Y}$ on the marginals of $\boldsymbol{Y}^\perp := \boldsymbol{\Gamma}_u{}' \boldsymbol{Y}$ with an intercept term.

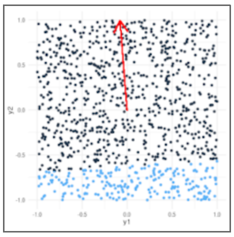
Estimation setup

The $\boldsymbol{\tau}$th quantile of $\boldsymbol{Y}$ is any element of the collection Λ_τ of hyperplanes
$\lambda_\tau := \{\boldsymbol{y} \in \mathbb{R}^k : \boldsymbol{u}'\boldsymbol{y} = \hat{\boldsymbol{b}_\tau}\boldsymbol{\Gamma}_u'\boldsymbol{y} + \hat{a_\tau}\}$,
such that $(\hat{a_\tau}, \hat{\boldsymbol{b}_\tau})$ are the solutions of the minimization problem
$\min_{(a_\tau, \boldsymbol{b}_\tau) \in \mathbb{R}^k} E[\rho_\tau(\boldsymbol{u}'\boldsymbol{y} - \boldsymbol{b}_\tau\boldsymbol{\Gamma}_u'\boldsymbol{y} - a_\tau)]$.
where $\rho_\tau(u)$ is a known loss function in the quantile regression literature defined as
$\rho_\tau(u) = u(\tau - \mathbb{I}(u < 0)), \quad 0 < \tau < 1.$

Upper and lower halfspaces

With predictor variables, we have
$\lambda_\tau(\boldsymbol{X}) = \{\boldsymbol{u}'\boldsymbol{y} = \hat{\boldsymbol{b}}_\tau \boldsymbol{\Gamma}_u' \boldsymbol{y} + \boldsymbol{x}'\hat{\boldsymbol{\beta}}_\tau + \hat{a}_\tau\}$,
We can say that each element $(\hat{a}_\tau, \hat{\boldsymbol{b}}_\tau, \hat{\boldsymbol{\beta}}_\tau)$ define an upper closed quantile halfspace
$H_{\tau\boldsymbol{u}}^+ = H_{\tau\boldsymbol{u}}^+(\hat{a}_\tau, \hat{\boldsymbol{b}}_\tau, \hat{\boldsymbol{\beta}}_\tau) = \{\boldsymbol{y} \in \mathbb{R}^k : \boldsymbol{u}'\boldsymbol{y} \geq \hat{\boldsymbol{b}}_\tau \boldsymbol{\Gamma}_u' \boldsymbol{y} + \boldsymbol{x}'\hat{\boldsymbol{\beta}}_\tau + \hat{a}_\tau\}$
and an analogous lower open quantile halfspace switching $\geq$ for $<$.

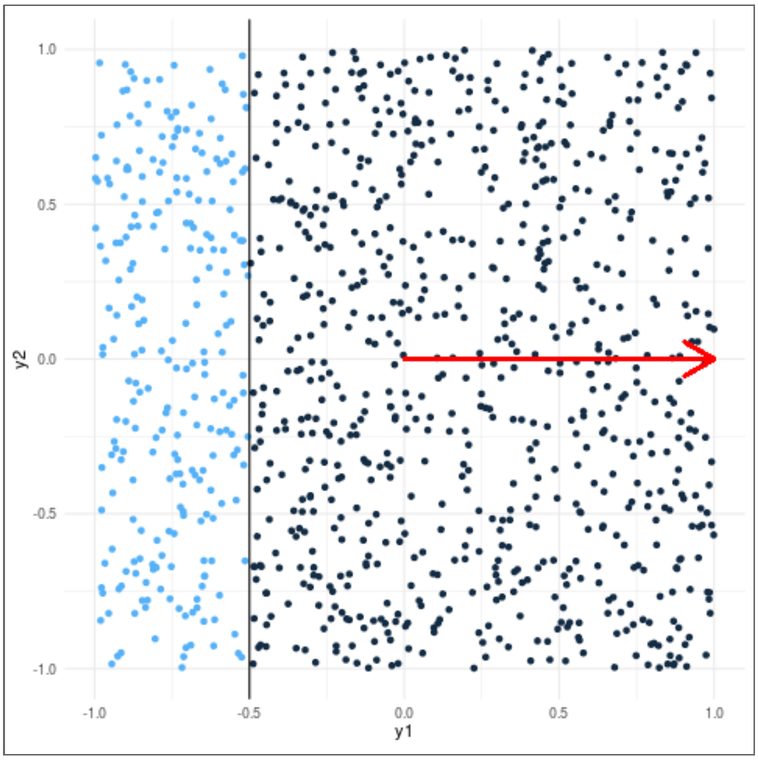




Properties

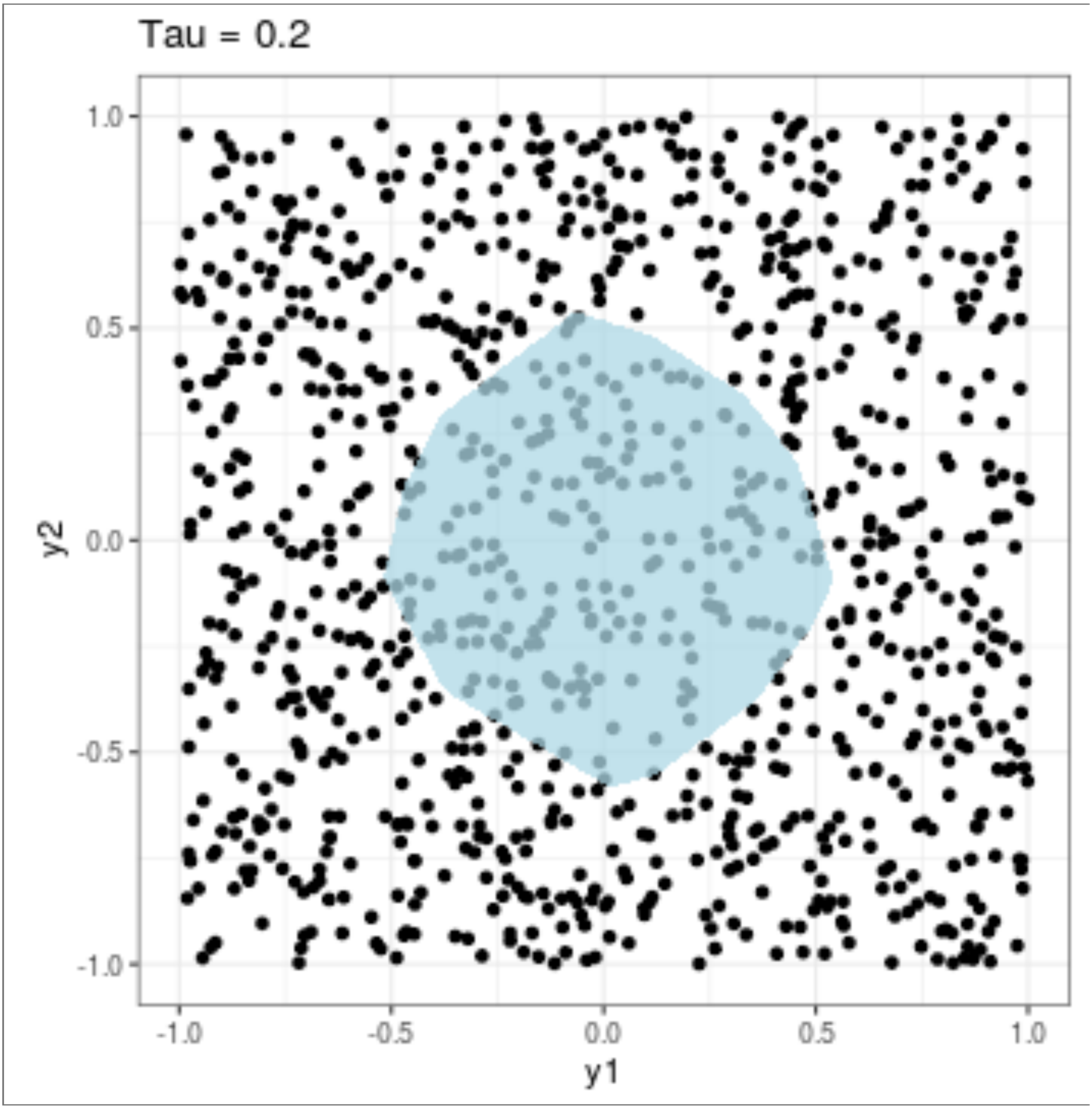
- Probabilistic nature of quantiles:

$$P(\boldsymbol{Y} \in H^{-}_{\tau\boldsymbol{u}}) = \tau,$$



Quantile region

Moreover, fixing τ we are able to define the τ quantile region $R(\tau)$ as
$R(\tau) = \bigcap_{\boldsymbol{u} \in \mathcal{S}^{k-1}} H^{+}_{\tau \boldsymbol{u}}$.



Bayesian directional quantile regression model

Consider the mixture representation of the asymmetric Laplace distribution

$$
\begin{aligned}
Y_i|w_i &\sim N(\mu + \theta w_i, \psi^2 \sigma w_i) \\
w_i &\sim \mathrm{Exp}(\sigma) \\
&\Updownarrow \\
Y &\sim AL(\mu, \sigma, \tau)
\end{aligned}
$$

Then one can consider that, for each direction u,

$Y_u|\boldsymbol{b}_\tau, \boldsymbol{\beta}_\tau, \sigma, w \sim N(Y^\perp b_\tau + \boldsymbol{x}'\boldsymbol{\beta}_\tau + \theta w_i, \psi^2 \sigma w_i)$,

Application results

Model choices

- Y_1 : Relative points won.
- Y_2 : Minutes played.
- Covariates:
  - Player (Federer, Nadal, Djokovic);
  - Surface;
  - Win or loss;
  - Type of tournament;
  - Top 20 player opponent or not;
- For the model, we fix $\tau = 0.25$ and consider 180 directions in the unit circle.
- We consider interaction effects between player and the other covariates.

Effect of win and losses

Effect of tournament

Effect of Top 20

Effect of surface

Final discussion

Conclusions

- This model does not need to make any probability assumptions in order to reach its conclusions.
- Nadal's dominance in clay courts is unmatched.
- Federer dominance in grass courts is also visible.
- The same way as Djokovic dominance in hard courts.
- In the time dimension, Federer shows an edge during wins.
- For most comparisons, Djokovic seems the most dominant player.

Thank you!

- b.santos@kent.ac.uk