Dynamic Prediction of the National Hockey League Draft with Rank-Ordered Logit Models

by

Brendan Kumagai

B.Sc., McMaster University, 2020

Project Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
Department of Statistics and Actuarial Science
Faculty of Science

SIMON FRASER UNIVERSITY
Summer 2023

Declaration of Committee

Name:	**Brendan Kumagai**
Degree:	**Master of Science**
Project title:	**Dynamic Prediction of the National Hockey League Draft with Rank-Ordered Logit Models**
Committee:	**Chair: Haolun Shi** Assistant Professor, Statistics and Actuarial Science
	Tim B. Swartz Supervisor Professor, Statistics and Actuarial Science
	Harsha Perera Committee Member Lecturer, Statistics and Actuarial Science
	Donald Estep Examiner Professor, Statistics and Actuarial Science

Abstract

The National Hockey League (NHL) Entry Draft has been an active area of research in hockey analytics over the past decade. Prior research has explored predictive modelling for draft results using player information and statistics as well as ranking data from draft experts. In this project, we develop a new modelling framework for this problem using a Bayesian rank-ordered logit model based on draft ranking data obtained from industry experts between 2019 and 2022. This model builds upon previous approaches by incorporating team tendencies and needs, addressing within-ranking dependence between players, and solving various other challenges of working with rank-ordered outcomes such as incorporating both unranked players and rankings that only consider a subset of the available pool of players (i.e., North American skaters, European goalies, etc.).

Keywords: Bayesian analysis; sports analytics; rank-ordered logit; multinomial logit; National Hockey League; entry draft

Acknowledgements

First and foremost, I'd like to thank my parents for their love, support and sacrifices throughout my journey from the early mornings in cold hockey rinks when I was a wannabe NHL player to the long trips to McMaster and later SFU when I began new steps in life as a wannabe NHL data scientist. The encouragement from you and extended family as well as the work ethic I have inherited from you has helped me in all aspects of life.

I would like to thank my supervisor, Dr. Tim Swartz, for providing me both the chance to join SFU as an MSc. student and the flexibility to pursue various sports analytics opportunities and competitions during my degree. Thank you for all your feedback and support.

I would also like to give a big thank you to the CANSSI Collaborative Research Teams (CRT) program as well as my TA and RAships for providing the Department of Statistics and Actuarial Sciences at SFU and I with funding to make this research feasible. Being paid to do interesting projects in sports analytics is a dream come true.

This project was initially joint work with Ryker Moreau and Kimberly Kroetch based on an idea that was sparked in conversations with Dr. Tyrel Stokes. They were all instrumental in helping to get this project off the ground in its early stages.

In my time at SFU, I've met some great friends and colleagues that have helped me grow as a person and data scientist. Thank you Kim for helping me survive STAT 830, STAT 851, etc. and teaching me "the song from Big" on the piano; Tyrel for always finding time in your crazy PhD schedule to provide me with feedback and answer my never-ending list of sports analytics questions; Ryker for providing us with occasional therapy dog sessions whenever you brought Bauer to campus; Robyn for piling us all in your minivan for fun excursions and putting up with my OCD in creating data visualizations; Eli for seemingly always having a new Kramer-esque story every day; Mandy for showing me just how much one person can love peanut butter; Gurashish for smothering me with consistently poor food takes; Renny for your enthusiasm surrounding post-work beers; and Lucas, Matt, Dani, and Peter for setting such an amazing culture among the sports analytics grad students at SFU.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Over the past two decades, the National Hockey League (NHL) has imposed a hard salary cap to limit player salaries and control a team's ability to retain and add talented players in an effort to enforce competitive balance throughout the league. This has forced teams to become increasingly savvy in how they allocate resources. The NHL has three main outlets where a team can add, lose or maintain talent: free agency, trades, and the entry draft. Acquiring players through free agency or trades can often be an expensive endeavour costing valuable cap dollars or assets. On the other hand, the entry draft is a low-cost, high-upside way to find and develop NHL-level talent.

The entry draft is an annual event held by the NHL where teams take turns selecting the world's best young, prospective hockey players in an effort to disperse incoming talent throughout the league. The order of the draft is predominantly determined by the reverse order of the standings to ensure that the worst teams from the previous season have the best chance at improving their roster. The draft lasts 7 rounds and each round consists of 32 picks - one for each team. However, there are many wrinkles that may affect the draft order such as trades, playoff results, compensatory picks, and the NHL draft lottery. Knappe (2022) provides a primer for the rules and setup of the NHL draft.

Every NHL team employs a department of scouts to identify and evaluate the top draft-eligible players throughout the season and inform the team's draft selections each year. To strategize and obtain the players they desire, teams make assumptions on how long a player will remain unselected during the draft. They must dynamically consider how the observed selections and upcoming selectors affect the chances that a player will remain available by their next pick.

As an interesting demonstration of the strategy involved in NHL draft decision-making, let's go back to the 2021 draft when the Dallas Stars were positioned to select next with the 15^{th} overall pick. Rather than making a selection, the Stars opted to trade the 15^{th} pick to the Detroit Red Wings in return for the 23^{rd}, 48^{th}, and 138^{th} picks. With the 23^{rd} pick, the Stars chose Wyatt Johnston, a Toronto native who had not played any league hockey in the season leading up to the draft due to COVID-19 restrictions in Ontario. At the time, Johnston was ranked 16^{th} among draft-eligible North American Skaters by NHL Central Scouting with other draft experts providing a wide range of where he ranked overall, fluctuating from 20^{th} to 125^{th}.

Following the draft, the Stars' general manager, Jim Nill, claimed that "[Johnston] was the best player that we had on the board at that time... The other teams that picked before us took players that we had ranked lower than he was" (Nill, 2021). If they believed Johnston had the potential to become a star player in the NHL, then was it worth it to defer from pick 15 down to 23 and acquire the two additional draft picks from Detroit?

The objective of this project is to estimate the probability that any draft-eligible player will be selected at any upcoming draft pick. Additionally, the modelling framework we propose can go beyond this to answer various questions related to the NHL draft. For example, what is the probability that a player will be available at a given pick, such as the case of Wyatt Johnston lasting until pick 23? How does the probability that a player will be selected at each pick change as each new draft pick is made? Which qualities do NHL teams and draft experts value in players and how do these values differ between teams and draft experts? Which draft experts' rankings best align with the tendencies of NHL teams?

Previous research related to the NHL entry draft has predominantly focused on the success of draft selections at the NHL level either through retrospective analysis of past draft selections or building predictive models in an attempt to estimate the level of success prospects will have in their career. Tingling (2017) provides a brief history of drafts in major league sports and a review of research related to the NHL draft. Schuckers (2011) and Tulsky (2013) use historical draft results, player statistics, and draft pick trades to estimate relative pick value. Both find that the value of draft picks have an exponential decaying pattern in the NHL draft (e.g., the decrease in value from the 1^{st} pick to the 2^{nd} pick is much greater than the decrease from the 101^{st} pick to the 102^{nd} pick). Nandakumar (2017) conducts a retrospective analysis of how well teams performed in the draft by comparing their realized draft results to what a 'perfect draft' would look like given the draft capital they had available to them and the future NHL performance of players available in the draft class. Li (2019) uses data from the NHL combine - an annual fitness assessment of draft-eligible players - to look at the relationship between combine results, draft stock, and success at the NHL level. The models found positive relationships between lower-body strength and draft stock as well as aerobic and anaerobic fitness and NHL success.

A subset of the research surrounding major league sports drafts has explored predictive modelling approaches for the outcome of the draft with the goal of answering the questions we have proposed in this project. ESPN Analytics (2022) have developed a model to predict the probability that a player is available at a given pick in the National Football League (NFL) draft; however, their methodology remains private. Burke (2014) and Sprigings (2016) both discuss the foundations of this model at a high-level. It is a Bayesian model based on consensus player rankings and projections from individual experts with weightings for each expert based on historical accuracy. Robinson (2020) uses a Bayesian gamma regression based on mock draft data from draft experts, fans, and media in the NFL to estimate the probability that a player will be available by a given pick.

This project looks to build upon these previous approaches with a Bayesian rank-ordered logit (ROL) model. The ROL model is a sequence of conditional multinomial logit (MNL) models used to model rank-ordered outcomes. The framework we propose addresses within-ranking dependence to allow updated selection probabilities after each new pick. Additionally, it provides natural solutions for incorporating unranked players and ranking sets that contain a subset of the available pool of draft-eligible players such as NHL Central Scouting - which breaks down their rankings into four categories: North American skaters, North American goalies, European skaters, and European goalies. Finally, it accounts for differences between NHL teams and draft experts by leveraging previous years' expert rankings and draft results to estimate team and expert-specific effects. This provides key strategic insights into other teams' draft tendencies and adjusts predictions to align with the tendencies of NHL teams as opposed to draft experts.

Rank-ordered logit models and multinomial logit models, although scarcely applied to hockey, have had many applications within sport and other domains. Tea & Swartz (2022) use a Bayesian MNL model to estimate probabilities of serve placements in tennis. Gerber & Craig (2021) also use a Bayesian MNL model to predict plate appearance outcomes and project batter performance over a season in Major League Baseball (MLB).

The rank-ordered logit model - also referred to as the Plackett-Luce or exploded logit model - was originally developed by Plackett (1975) who leveraged the multinomial choice logit model proposed by Luce (1959) to model rank-ordered outcomes. Many of the applications of ROL models in sport have involved racing sports including horse racing (Bolton & Chapman, 1986; Lo & Bacon-Shone, 1994; Ali, 1998; Johnson, 2010) and automobile racing (Graves et al., 2003; Guiver & Snelson, 2009; Anderson, 2014). More recently, the focus of ROL models in sport have accounted for changes in player abilities over time. Glickman & Hennessy (2015) develop a stochastic ROL model to predict the outcome of multicompetitor competitions while allowing for changes in competitor abilities over time with applications to women's alpine data.

We begin in Chapter 2 by describing the data collected for this project, defining important terminology related to our model, and reviewing the multinomial logit model. In Chapter 3, we define our ROL model specifically designed for the case of the NHL draft. We define a base ROL model and propose enhancements to this model to account for unranked players, subset rankings, and team tendencies. Chapter 4 explores the interesting results and insights obtained from applying our model to the 2019 to 2022 NHL drafts. We demonstrate the effect of incorporating team and expert tendencies into the model, explore differences between teams and experts, and determine how much of a risk the Dallas Stars took by trading down from pick 15 to 23 in acquiring Wyatt Johnston. Additionally, we investigate model evaluation techniques to compare with previous approaches to this problem. We conclude with a short discussion of the implications of our work and future directions that can be taken in Chapter 5.

Chapter 2

Background

2.1 Data

In multicompetitor sports, athletes typically compete in many events each year that can be used to fit the ROL model and predict the outcome of future events. On the other hand, the NHL draft only occurs once per year and each draft involves an entirely different pool of players available to be selected.

Due to the popularity of the draft among NHL fans, there exists a market for draft experts who spend hours producing media content related to the NHL draft including ordered lists which rank the top draft-eligible players. These draft experts are typically employed by sports media networks such as TSN and Sportsnet or websites that produce hockey-related content such as Elite Prospects, Dobber Hockey, and The Hockey News. We leverage these expert rankings along with NHL draft results from previous years to compensate for the fact that we have no true realizations of the draft results until the actual NHL draft occurs.

We will refer to the draft experts from media networks and hockey websites that provide rankings for the NHL draft hereinafter as ‘agencies’. Additionally, we will refer to each set of ranked players provided by an agency hereinafter as a ‘ranking set’ and the observed results from the NHL draft in a given year as the ‘draft results’.

The data used to fit our model includes draft results and ranking sets from the 2019-2022 NHL drafts as well as player information for every draft-eligible player that appears in at least one ranking set or draft result. Each ranking set contains both the rank ordering of players provided by the agency along with the full set of players that were available to be ranked, regardless of whether they were ranked or not by the agency. We obtained this data through a mixture of web scraping with Python Selenium and manual data entry.

The draft is typically held each year between June 20^{th} and 30^{th}. We only consider the final ranking set by each agency prior to the draft in our model. This provides us with 14, 17, 22, and 22 ranking sets in each draft year from 2019 to 2022, respectively. Additionally, all ranking sets published before May 1^{st} are used to set the prior distribution for our model with an empirical Bayes approach as outlined in 3.5.1 and Appendix B.

2.2 Notation

The notation required to express the ROL model proposed in this project involves many nested subscripts. To make the model easier to digest, we use two different types of notation throughout the paper for different purposes. When we wish to illustrate how the model is structured, we use notation in terms of a single ranking set as outlined in Table 2.1. Alternatively, when we wish to expand the model to encompass all ranking sets over all draft years, we use notation that specifies which ranking set is being considered as outlined in Table 2.2. Notice that most of the notation stays consistent between both cases; however, in Table 2.1, we do not specify which ranking set or draft year we are considering.

Variable	Description
n	The number of picks made in the ranking set
N	The number of draft-eligible players considered in the ranking set
s_i	The i^{th} ranked player in the ranking set
S_i'	The set of available players prior to the i^{th} ranking in the ranking set
a_i	The agency that selected the i^{th} ranked player in the ranking set

Table 2.1: A list of variables to be used when describing the ROL model for a single ranking set.

2.3 Multinomial Logit Models

Before laying out the methodology of the ROL model, we first provide a brief overview of the multinomial logit model applied to the NHL draft. MNL models are the main building blocks in defining the rank-ordered logit model in Section 3.1. They are a model used in statistics to classify observations into one of two or more discrete outcome categories.

Let's consider an example where we are building a model to predict the 1^{st} overall pick where there are N total draft-eligible players in the data. This is a special case of the MNL model where each ranking set has one trial being taken from N categories with the trial being the 1^{st} ranked player in the ranking set and the categories, $i = 1, \ldots, N$, representing

Variable	Description
R	The set of draft years considered in the model. Here $R = \{2019, 2020, 2021, 2022\}$
K_r	The number of ranking sets or draft results in year r where $r \in R$
n_r	The number of picks made in the draft results of year r where $r \in R$
N_r	The number of draft-eligible players considered in year r where $r \in R$
n_{rk}	The number of players ranked in the k^{th} ranking set or draft result of year r
N_{rk}	The number of players available to be ranked in the k^{th} ranking set or draft result of year r
s_{irk}	The i^{th} ranked player in the k^{th} ranking set or draft result of year r
S'_{irk}	The set of available players before the i^{th} rank is made in the k^{th} ranking set or draft result of year r
a_{irk}	The agency or team that selected the i^{th} ranked player in the k^{th} ranking set or draft result of year r

Table 2.2: A list of subscript variables to be used when describing the ROL model while considering all ranking sets in all draft years between 2019-2022.

each of the N draft-eligible players. The goal of the MNL model is to estimate the probabilities that each of the available players will be selected 1^{st} overall based on the ranking sets. In other words, we wish to estimate probabilities $\boldsymbol{\pi}_1 = [\pi_{11}, \pi_{21}, \ldots, \pi_{N1}]$ such that π_{i1} is the probability that player i will be selected with the 1^{st} overall pick subject to the constraint that $\sum_{j=1}^{N} \pi_{j1} = 1$.

These probabilities are given by

$$\pi_{i1} = \frac{\exp(\theta_i)}{\sum_{j=1}^{N} \exp(\theta_j)} \tag{2.1}$$

where θ_i represents a random effect for player i that we refer to as player i's 'ability parameter', $\forall i = 1, \ldots, N$. This conversion of real numbered-parameters to probabilities is often referred to as the multinomial choice probability or the softmax function.

In this specification of the MNL model, we wish to estimate the vector of player ability parameters, $\boldsymbol{\theta}$, and, in turn, use (2.1) to estimate the probability that player i will be selected 1^{st} overall $\forall i = 1, \ldots, N$.

2.3.1 Latent Variable Formulation of the MNL Model

The MNL model can also be formulated as a latent variable model by introducing an ability score as shown in Appendix A. This formulation is popular in the literature of discrete choice models in economics. Here, we assume that each player in each ranking set has an unobserved random variable, Y_i, $i = 1, \dots, N$, that can be thought of as the latent rating of player i's ability by the ranking agency.

In this formulation, we assume that $Y_i = \theta_i + \epsilon_i$ where $\epsilon_i \sim \text{Gumbel}(0, 1)$, $\forall i = 1, \dots, N$. That is, Y_i is assumed to be a realization from the player ability parameter for player i with an i.i.d. standard Gumbel error term. We also assume that the agency will select the player that they assign the highest rating to out of all available players. Thus, π_{i1} is considered the probability that Y_i is greater than Y_j, $\forall j \in \{1, \dots, N\} \setminus \{i\}$. In this context, we can extend the expression provided in (2.1) as follows,

$$\pi_{i1} = P(Y_i > Y_j, \forall j \in \{1, \dots, N\} \setminus \{i\}) = \frac{\exp(\theta_i)}{\sum_{j=1}^{N} \exp(\theta_j)}$$

where the equivalency of the right hand side of the equation is derived in Appendix A.

The reasoning behind the Gumbel assumption is made clear by Luce & Suppes (1965), who provide the original derivation of the multinomial choice probability in (2.1) while assuming $\epsilon_i \sim \text{Gumbel}(0, 1)$, $\forall i = 1, \dots, N$. Furthermore, McFadden (1974) shows that if (2.1) is true, then the errors of Y_i must follow the i.i.d. standard Gumbel distribution.

The assumption of i.i.d. standard Gumbel errors is almost identical to an assumption of i.i.d. standard normal errors, with extreme value distributions such as the Gumbel having slightly fatter tails (Train, 2009). However, the latent formulation of the MNL model with Gumbel errors is much more convenient to work with than the equivalent formulation of the multinomial probit (MNP) model with normal errors. Additionally, McFadden & Train (2000) prove that a mixed MNL model can approximate any random utility model with any degree of accuracy by using appropriate variables and mixing distributions. Srinivasan & Mahmassani (2005) provide a special case of this proof to show that the mixed MNL model can approximate the MNP model with appropriate variables and mixing distributions.

Agresti (2019) provides further details on the theoretic framework behind multinomial logit models along with various examples of MNL models applied to real-world scenarios.

Chapter 3

A Rank-Ordered Logit Model for the NHL Draft

3.1 Rank-Ordered Logit Models

The MNL model provides us with a simple framework for describing the probability that a player is selected with the 1^{st} pick in the draft, but there are still many questions that this model cannot address such as: What is the probability of a player being drafted 2^{nd}, 3^{rd} or beyond? How do these probabilities differ depending on which player(s) are previously selected? If a player is consistently ranked top five but is never ranked 1^{st} by an agency, would his probability of being selected 1^{st} be the same as a player rarely ranked in the top 200?

These questions can be addressed using the rank-ordered logit model. The ROL model can be thought of as a product of conditional multinomial logit models where the 1^{st} overall pick is modelled as a MNL model with a single pick taken from the pool of all draft-eligible players, then the 2^{nd} pick is modelled as a MNL model with a single pick taken from all draft-eligible players excluding the player selected 1^{st}, and so on until the N^{th} pick, which is modelled using the MNL model with a single pick taken from all draft-eligible players excluding the $N-1$ players that have already been selected. The ROL model provides a full rank ordering from the 1^{st} pick to the N^{th} pick.

Recall the latent variable formulation of the MNL model introduced in Section 2.3.1. We defined Y_i as the latent rating of player i's ability by the ranking agency and assumed that $Y_i = \theta_i + \epsilon_i$ where $\epsilon_i \sim \text{Gumbel}(0,1)$ are i.i.d. error terms $\forall i = 1,\dots,N$. This can alternatively be written as

$$Y_i \sim \text{Gumbel}(\theta_i, 1) \tag{3.1}$$

where Gumbel$(\theta_i, 1)$ denotes a Gumbel distribution with location parameter θ_i and scale parameter 1.

We can generalize equation (2.1) to obtain multinomial choice probabilities for the m^{th} draft pick $\forall m = 1, \ldots, N$ by conditioning on the previous $m-1$ draft selections. Recall that S'_m is the set of players that remain available at pick m. We express the probability that player i, $\forall i \in S'_m$, will be selected with the m^{th} pick given the previous $m-1$ picks as

$$\pi_{im} = P(Y_i > Y_j, \forall j \in S'_m \setminus \{i\}) = \frac{\exp\{\theta_i\}}{\sum_{j \in S'_m} \exp\{\theta_j\}} \tag{3.2}$$

Appendix A describes the full derivation of π_{im}, $\forall m = 1, \ldots, N$.

The likelihood for a full ranking set is proportional to the probability of obtaining the exact ordering provided by the ranking set. This probability can be expressed using the multinomial choice probabilities from (3.2) as follows

$$L(\boldsymbol{\theta}) = P(Y_{s_1} > Y_{s_2} > \cdots > Y_{s_N} | \boldsymbol{\theta}) = \prod_{i=1}^{N} \pi_{s_i i} = \prod_{i=1}^{N} \frac{\exp\{\theta_{s_i}\}}{\sum_{j=i}^{N} \exp\{\theta_{s_j}\}} \tag{3.3}$$

Notice this is the joint probability that player s_1 is ranked 1^{st}, player s_2 is ranked 2^{nd} given player s_1 is no longer available, and so on down to player s_N being ranked last given players $s_1, \ldots, s_{N-1}$ are no longer available. Recall that s_i is the index of the player ranked i^{th} by the ranking agency.

To provide further intuition on the latent ratings, Y_i, and player abilities, θ_i, consider an example where the player abilities can be thought of as some real-numbered value on a scale of 0-100 in a similar fashion to a player rating system in any modern sports video game. Suppose we know the players' abilities with the top three players in the 2022 draft - Shane Wright, Juraj Slafkovsky, and Logan Cooley - having abilities of $\theta_{\text{Wright}} = 94.2$, $\theta_{\text{Slafkovsky}} = 91.9$, and $\theta_{\text{Cooley}} = 90.1$. Under the model proposed in this chapter, we assume that an agency creating a ranking set will assign each of the players a rating that is the sum of the player's ability parameter and a realization from an i.i.d. standard Gumbel error distribution. For example, let's say $Y_{\text{Wright}} = 93.9$, $Y_{\text{Slafkovsky}} = 90.9$, and $Y_{\text{Cooley}} = 91.6$. Since $Y_{\text{Wright}} > Y_{\text{Cooley}} > Y_{\text{Slafkovsky}}$, we will observe a ranking set of Shane Wright ranked 1^{st}, Logan Cooley ranked 2^{nd}, and Juraj Slafkovsky ranked 3^{rd} in this case.

Figure 3.1 provides an illustration of this hypothetical scenario considering Wright, Slafkovsky, and Cooley.

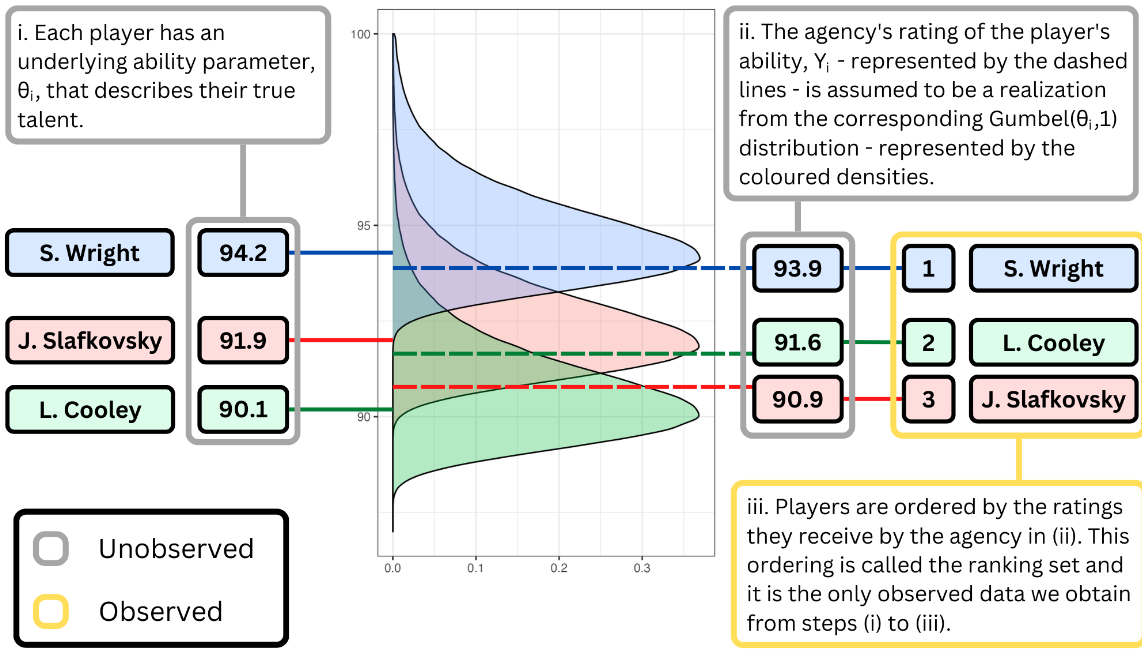


Figure 3.1: A hypothetical illustration of the assumed structure that agencies use in forming a ranking set.

In order for another player, i, to displace Wright as the 1^{st} ranked player, he must be given a rating greater than 93.9 by the ranking agency. If that were the case, his rating Y_i is greater than all other players considered in the ranking set or, equivalently, $Y_i > Y_j, \forall j \in \{1, \ldots, N\} \setminus \{i\}$.

Note that we only observe the *ordering* of the agencies' player ratings through the ranking sets. Agencies typically do not release player ratings and might not use a quantitative rating system in formulating their rankings at all. Thus, the agencies' ratings of each player, Y_i, are unobserved.

The likelihood given by (3.3) is the ROL model for a single ranking set. In Sections 3.2 and 3.3 we alter this model to better address unique issues in modelling the NHL draft. Beyond that, Sections 3.4 and 3.5 describe the full model for all ranking sets and the estimation methods.

3.2 Unranked Players and Subset Rankings

The model likelihood proposed in (3.3) considers a ranking set where all N draft-eligible players are both ranked and available to be ranked. However, this is not always the case in practice.

Typically, an agency provides a ranking set up to some number of picks $n < N$. In this case, (3.3) does not reflect the agency's ordering of the remaining $N - n$ players who were not ranked. To address this, we adjust the likelihood based on the work of Fok et al. (2012) which takes the product of multinomial choice probabilities up until the n^{th} pick as follows

$$P(Y_{s_1} > Y_{s_2} > \cdots > Y_{s_n} > \max\{Y_{s_{n+1}}, \ldots, Y_{s_N}\}|\boldsymbol{\theta}) = \prod_{i=1}^{n} \frac{\exp\{\theta_{s_i}\}}{\sum_{j=i}^{N} \exp\{\theta_{s_j}\}} \tag{3.4}$$

Notice that this updated model likelihood still considers all N players to be available to the agency when selecting each pick, despite that only n players are ranked. This is an important distinction as it incorporates the unranked players into the likelihood and recognizes that the latent rating for each of the unranked players is less than the latent rating of the last *ranked* player.

Additionally, there are also occasional ranking sets that only consider a subset of the total pool of players. A popular example of this is the NHL Central Scouting agency who divides their ranking sets into four categories: North American skaters, North American goalies, European skaters, and European goalies. To address this, we allow N to vary for each ranking set. No change is needed in the likelihood for a single ranking set proposed in (3.4); however, we address this in Section 3.4 when we state the full likelihood for all ranking sets by letting n_{rk} and N_{rk} be the number of players ranked and number of players available for ranking set k in draft year r where $r \in R$ and $k = 1, \ldots, K_r$, respectively.

3.3 Agency and Team Tendencies

The proposed ROL model to this point has assumed a linear predictor of $\eta_{ij} = \theta_i$ for agency or team j's latent rating of player i where θ_i is player i's ability parameter and all other variation in player ratings is due to noise from an i.i.d. standard Gumbel error term. However, the agency or team doing the ranking can have a major influence on the player ratings.

As an easily measurable example, consider the 2020 draft selections made by the Toronto Maple Leafs and Ottawa Senators. That year, the Maple Leafs' picks had an average height of approximately 5'10.5", while the Senators averaged approximately 6'2" among their new draftees. Perhaps this was random chance, but - based on the disparity between the two teams - it is far more likely that the Senators' management team valued tall players more than the Maple Leafs.

To address this, we define a general linear predictor for player i being ranked by agency or team j as

$$\eta_{ij} = \theta_i + \beta_{j1}x_{i1} + \beta_{j2}x_{i2} + \cdots + \beta_{jp}x_{ip} \tag{3.5}$$

where θ_i is the player ability parameter, x_{ik} is a covariate with player-specific information and β_{jk} is an agency or team-specific parameter corresponding to x_{ik}, $\forall k = 1, \ldots, p$.

By including this updated linear predictor that varies with the ranking agency or team, the likelihood is now updated to be

$$\prod_{i=1}^{n} P\left(Y_{s_i a_i} > Y_{j a_i}, \forall j \in S_i' \setminus \{s_i\} \mid \boldsymbol{\theta}, \boldsymbol{\beta}\right) = \prod_{i=1}^{n} \frac{\exp\{\eta_{s_i a_i}\}}{\sum_{j=i}^{N} \exp\{\eta_{s_j a_i}\}} \tag{3.6}$$

where $\boldsymbol{\beta}$ is an $A \times p$ matrix with rows representing each of the A total agencies and teams in the data and columns representing each of the p covariates in the data.

Since we are now assuming that the linear predictor is dependent on the agency or team, the latent rating, $Y_{ij} = \eta_{ij} + \epsilon_{ij}$, where ϵ_{ij} remains as an i.i.d. standard Gumbel error term, will also be dependent on the agency or team. As a consequence, we can no longer use the $Y_1 > Y_2 > \cdots > Y_n$ notation as expressed in 3.3 and 3.4 since each subsequent pick may be made by a different team. Thus, we only care about the ordering of the next *pick* by each team.

3.4 Final Model

We have defined a ROL model for the NHL draft with respect to a single ranking set in Section 3.1 and built upon it to address draft-related issues in Sections 3.2 and 3.3. Now we extend this model to all ranking sets over all draft years available in our data. Recall Table 2.2 provides a guide of the notation used for the ROL model with all ranking sets.

We now let $Y_{ijrk} = \eta_{ij} + \epsilon_{ijrk}$ be the latent rating by agency j for player i in the k^{th} ranking set of year r where ϵ_{ijrk} remains an i.i.d. standard Gumbel error term $\forall r, k, i, j$. The full model likelihood can now be stated as the product of (3.6) taken over all ranking sets in all draft years as follows

$$L(\boldsymbol{\theta}, \boldsymbol{\beta}) = \prod_{r \in R} \prod_{k=1}^{K_r} \prod_{i=1}^{n_{rk}} \frac{\exp\{\eta_{s_{irk} a_{irk}}\}}{\sum_{j=i}^{N_{rk}} \exp\{\eta_{s_{jrk} a_{irk}}\}} \tag{3.7}$$

where s_{irk} and a_{irk} are the i^{th} ranked player and the agency or team that ranked the i^{th} player in the k^{th} ranking set or draft result in year r, respectively. Note that the denominator in 3.7 has subscript a_{irk} rather than a_{jrk} to ensure the probabilities at the i^{th} pick reflect the tendencies of the agency making that pick.

The model defined in (3.5) and (3.7) is our full ROL model proposed for the NHL draft. This triple product represents the probability of observing all selections over all ranking sets in all draft years considered by the model.

3.5 Parameter Estimation

The ROL model we have proposed for the NHL draft typically involves the estimation of hundreds - or even thousands - of parameters. To estimate this high-dimensional set of parameters, we turn to Bayesian inference. In the Bayesian setting, we can express the posterior distribution of parameters $\boldsymbol{\theta}, \boldsymbol{\beta}$ as proportional to the model likelihood stated in (3.7) multiplied by the prior distribution, $\pi(\boldsymbol{\theta}, \boldsymbol{\beta})$, as follows

$$\pi(\boldsymbol{\theta}, \boldsymbol{\beta}|\boldsymbol{Z}) \propto \prod_{r\in R}\prod_{k=1}^{K_r}\prod_{i=1}^{n_{rk}} \frac{\exp\{\eta_{s_{irk}a_{irk}}\}}{\sum_{j=i}^{N_{rk}} \exp\{\eta_{s_{jrk}a_{irk}}\}}\pi(\boldsymbol{\theta}, \boldsymbol{\beta}) \tag{3.8}$$

where $\boldsymbol{Z}$ is the ranking set data for all players considered in the ranking set.

Here, we are interested in estimating unknown parameters $\boldsymbol{\theta}, \boldsymbol{\beta}$ which quantify the ability of each player and the level at which each agency and team values particular player traits, respectively.

3.5.1 Prior Distribution

We assume a $\sum_{r\in R} N_r$ dimensional multivariate normal prior on the ability parameters with covariance matrix $\Sigma = \sigma_\theta^2 \boldsymbol{I}$ and mean vector $\boldsymbol{\mu}_\theta$ which is determined through an empirical Bayes procedure based on ranking sets that were released prior to May 1^{st} of the draft year. Thus, our prior distribution is $\boldsymbol{\theta} \sim \text{MVN}(\boldsymbol{\mu}_\theta, \sigma_\theta^2 \boldsymbol{I})$ where we place a hyperprior of Inv-Gamma$(1,1)$ on σ_θ^2. A full description of the empirical Bayes procedure to obtain $\boldsymbol{\mu}_\theta$ is found in Appendix B.

We assume a hierarchical structure on each $\boldsymbol{\beta}_k$ in the parameter matrix $\boldsymbol{\beta} = [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p]$. That is, $\beta_{jk} \sim N(\mu_{\beta k}, \sigma_{\beta k}^2)$ where we set standard hyperpriors $\mu_{\beta k} \sim N(0,1)$ and $\sigma_{\beta k}^2 \sim$ Inv-Gamma$(1,1)$, $\forall k = 1, \ldots, p$ where p is the total number of player information covariates in the model.

From (3.5) it is evident that $\boldsymbol{\theta}$ and $\boldsymbol{\beta}_k$ are each only identifiable up to an additive constant. To address this, we impose a constraint on the model that all player ability parameters in a given draft year must sum to zero. We also impose the same constraint for each $\boldsymbol{\beta}_k$, $\forall k = 1, \ldots, p$. As a result, zero represents the average value of a parameter and parameter values can be interpreted as relative to each other or to the average (zero).

3.5.2 Computation

The model is fit with Stan (Stan Development Team, 2022a), an open-source software that uses Hamiltonian Monte Carlo (HMC) methods to obtain draws from the posterior distribution of model parameters. We access Stan through RStudio and the R package 'rstan' (Stan Development Team, 2022b), which provides an interface for integrating Stan code into the R programming language.

We run the model on 4 chains with 4500 iterations for each chain, 2000 of which are used as the 'burn-in' stage. Thus, we obtain 10000 total posterior draws for our model parameters. We denote $\boldsymbol{\theta}^{(t)}, \boldsymbol{\beta}^{(t)}$ as a posterior draw provided from Stan's HMC simulations for our model where $t = 1, \ldots, 10000$. Thus, posterior means for model parameters can be expressed as $\hat{\boldsymbol{\theta}} = \frac{1}{10000} \sum_{t=1}^{10000} \boldsymbol{\theta}^{(t)}$ and $\hat{\boldsymbol{\beta}} = \frac{1}{10000} \sum_{t=1}^{10000} \boldsymbol{\beta}^{(t)}$.

Under this specification, it took approximately 36 hours to run on a 16-core MacBook Pro with 64 GB of RAM.

3.5.3 Predictive Distributions

The desired output of the ROL model for the NHL draft is an estimate of the joint probability distribution over all remaining players and draft picks. To do this, we wish to obtain the predictive distribution, which averages the density of future data over the posterior densities of the unknown model parameters to provide a forecast of future outcomes. The predictive distribution allows us to gain a better understanding of the uncertainty in our model by not only accounting for the uncertainty of the likelihood but also - by integrating over $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ - we average over the uncertainty in the parameter values.

Recall that the posterior distribution, $\pi(\boldsymbol{\theta}, \boldsymbol{\beta}|\boldsymbol{Z})$, corresponds to the ability parameters and team and agency preferences based on past ranking sets and draft results. The data $\boldsymbol{Z}$ corresponds to all previous ranking sets and draft results as well as the results of the current draft up to the time of interest. Fortunately, the Bayesian framework provides a convenient approach for prediction of future draft results, $\tilde{\boldsymbol{Z}}$. The predictive distribution of $\tilde{\boldsymbol{Z}}$ is given by

$$p(\tilde{\boldsymbol{Z}}|\boldsymbol{Z}) = \int\int p(\tilde{\boldsymbol{Z}}|\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{\beta})\pi(\boldsymbol{\theta},\boldsymbol{\beta}|\boldsymbol{Z})d\boldsymbol{\theta}d\boldsymbol{\beta} \tag{3.9}$$

where $p(\tilde{\boldsymbol{Z}}|\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{\beta})$ is the joint probability mass function across all remaining draft picks, given $\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{\beta}$.

As a by-product of MCMC, we generate realizations of the predictive distribution by first generating $(\boldsymbol{\theta}^{(t)},\boldsymbol{\beta}^{(t)})$, $\forall t = 1,\ldots,10000$ from the posterior distribution, then generating $\tilde{\boldsymbol{Z}}$ from $p(\tilde{\boldsymbol{Z}}|\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{\beta})$ using these posterior draws. This is done by repeatedly simulating from the softmax function with a diminishing choice set that removes players as they are 'selected' in the simulation. Thus, we do not directly solve this integral but rather approximate it using posterior samples.

Let $\pi_{ijm}^{(t)} = \frac{\exp\{\eta_{ij}^{(t)}\}}{\sum_{\ell=i}^{N_r}\exp\{\eta_{\ell j}^{(t)}\}}$ be the probability that player i is selected by team j with the m^{th} pick in the draft, given the previous $m-1$ picks are known where $\eta_{ij}^{(t)}$ is the linear predictor as defined in (3.5) under the t^{th} posterior draw.

Suppose we wish to begin drafting from pick m, where $1 \leq m \leq n_r, m \in \mathbb{Z}$ and the ordered set of size $m-1$ of players that have already been selected is contained within $\boldsymbol{Z}$, while S'_m is the set of size $N_r - (m-1)$ of players that have not been selected. We obtain a simulation of the draft by repeatedly (i) simulating a selection from the set of multinomial choice probabilities, $\left(\pi_{ijm}^{(t)}\right)_{i\in S'_m}$, (ii) removing the selected player from S'_m and adding him to an ordered set of players selected by simulation, $\tilde{\boldsymbol{Z}}^{(t)}$, (iii) repeating (i) and (ii) until n_r draft selections are made.

By repeating this process $\forall t = 1,\ldots,10000$, we can estimate the probability that a player will be selected with each upcoming pick given $\tilde{\boldsymbol{Z}}$ as expressed in (3.9).

Chapter 4

Results

We now explore the results of the ROL model applied to ranking set and draft result data from the 2019 to 2022 NHL drafts as described in Section 2.1. All models were fit by modifying code published by Stokes (2022).

Appendix C provides tables with all teams and agencies considered in the model along with corresponding abbreviations used in this chapter.

4.1 Without Agency and Team Tendencies

We begin by exploring the ROL model *without* considering agency and team tendencies. Here, we use the likelihood posed in (3.7) with linear predictor $\eta_{ij} = \theta_i$ rather than (3.5). Thus, we are estimating the posterior distribution of the player ability parameters with no consideration of agency or team tendencies.

Figure 4.1 displays posterior summaries from this model for the ability parameters of the top 32 players in the 2022 NHL draft. The posterior draws from the model are used to form the credible intervals shown in this figure. By ordering players by the posterior means of their ability parameter, we obtain a consolidated draft ranking based on the ranking sets provided.

Notice the unusually large gaps between the posterior means of Shane Wright and Logan Cooley as well as David Jiricek and Joakim Kemell. These are due to the natural tiers that form in the draft when ranking sets consistently rank a group of players above those below them. For example, agencies consistently ranked Wright, Cooley, Slafkovsky, Nemec and Jiricek as the top five in the draft, while Kemell and Savoie rarely achieved this feat. Thus, despite being ranked 6^{th} and 7^{th}, respectively, the model estimates a relatively large gap between them and the top five.

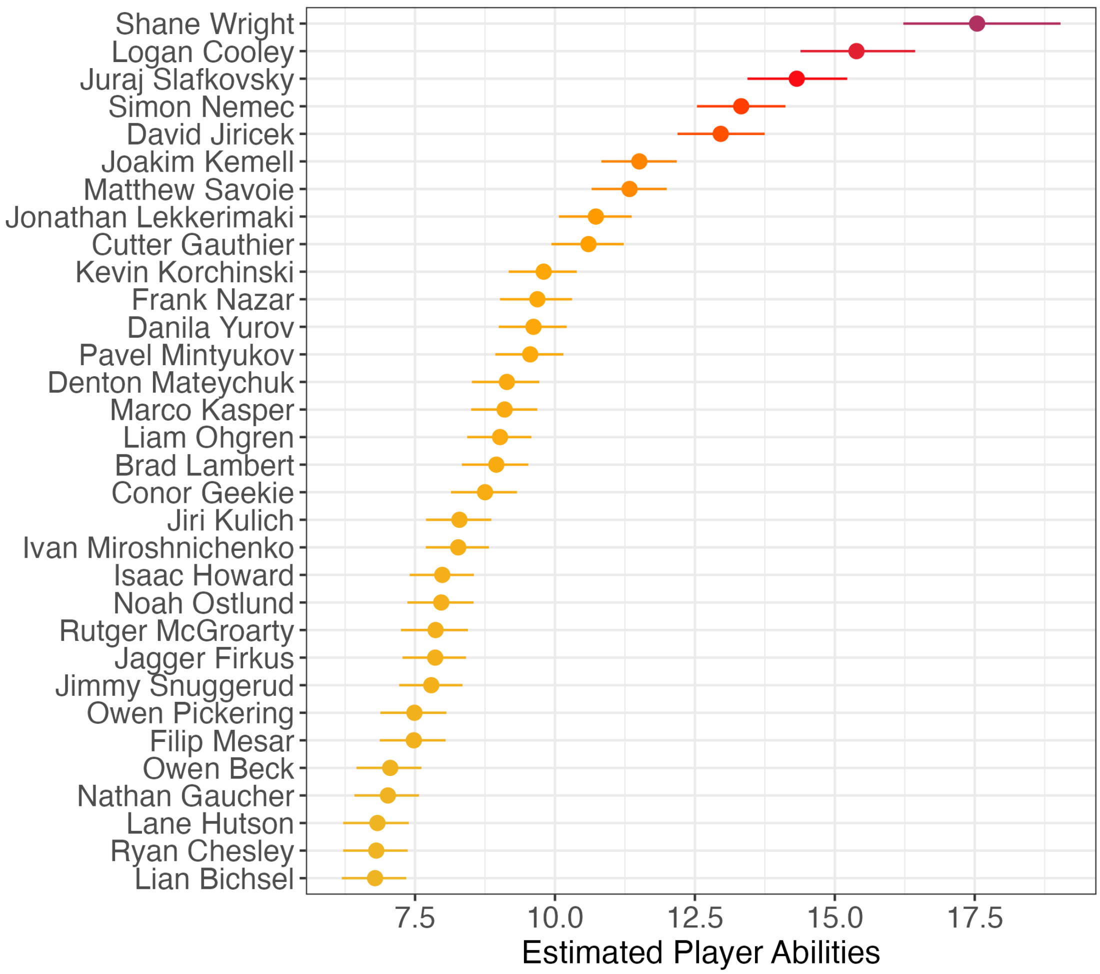


Figure 4.1: Posterior distribution summaries of the player ability parameters, $\boldsymbol{\theta}$, for the top 32 of 601 draft-eligible players in the 2022 NHL draft immediately prior to the draft. Players are listed in descending order based on posterior mean estimates of $\boldsymbol{\theta}$. Points represent the posterior means of θ_i for player i; lines represent the corresponding 95% credible intervals of the parameter.

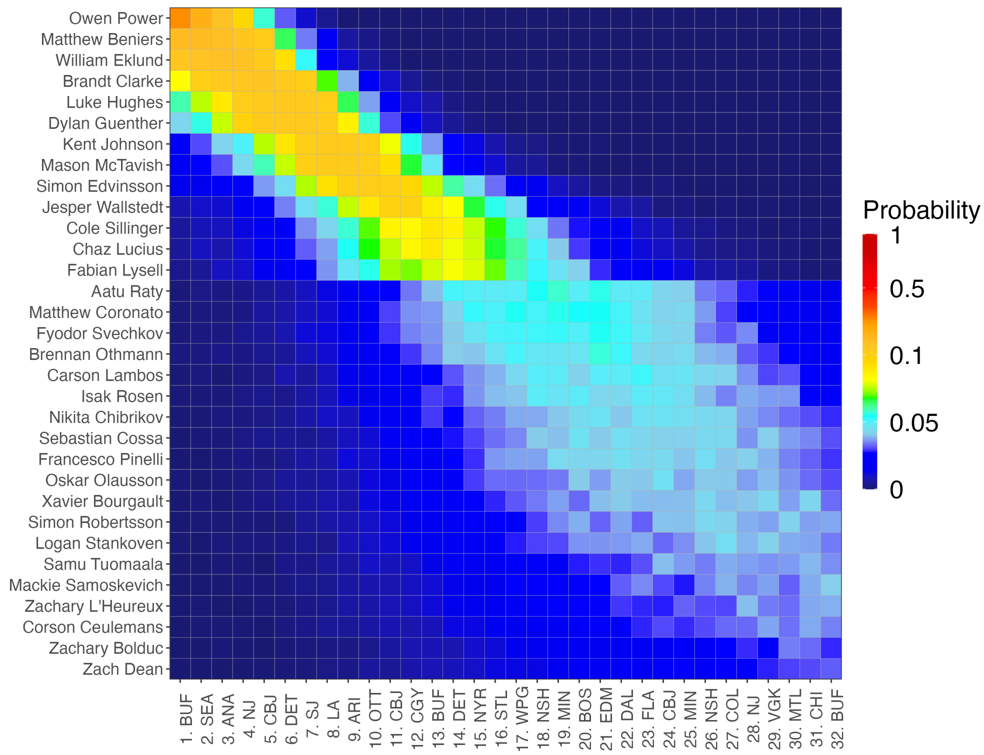
Figures 4.2a and 4.2b illustrate the estimated predictive probabilities that a player is selected at any pick as described in Section 3.5.3. Here, we look at the top 32 players and picks in the 2021 and 2022 drafts. We refer to this visual as the player-pick probability mass function (pmf) since it represents the joint pmf that a player will be selected at a corresponding draft pick. The i^{th} row can be referred to as the pick pmf for player i up until pick 32 as it represents the marginal pmf for player i being selected at any pick in the top 32.

Notice that each draft class displays unique trends. In 2021, we observe a fair amount of uncertainty towards the top of the draft as there was a lot of variation among agencies in ranking the top 6 players. While in 2022, the vast majority of agencies ranked Shane Wright 1^{st}; consequently, Wright had a very high probability of being selected 1^{st} overall by this model. Additionally, we now observe the same tiers in the draft as Figure 4.1 just on the probability scale as opposed to the latent ability scale. This feature allows teams to gain a sense of where large perceived drops in pick value occur and strategize accordingly.

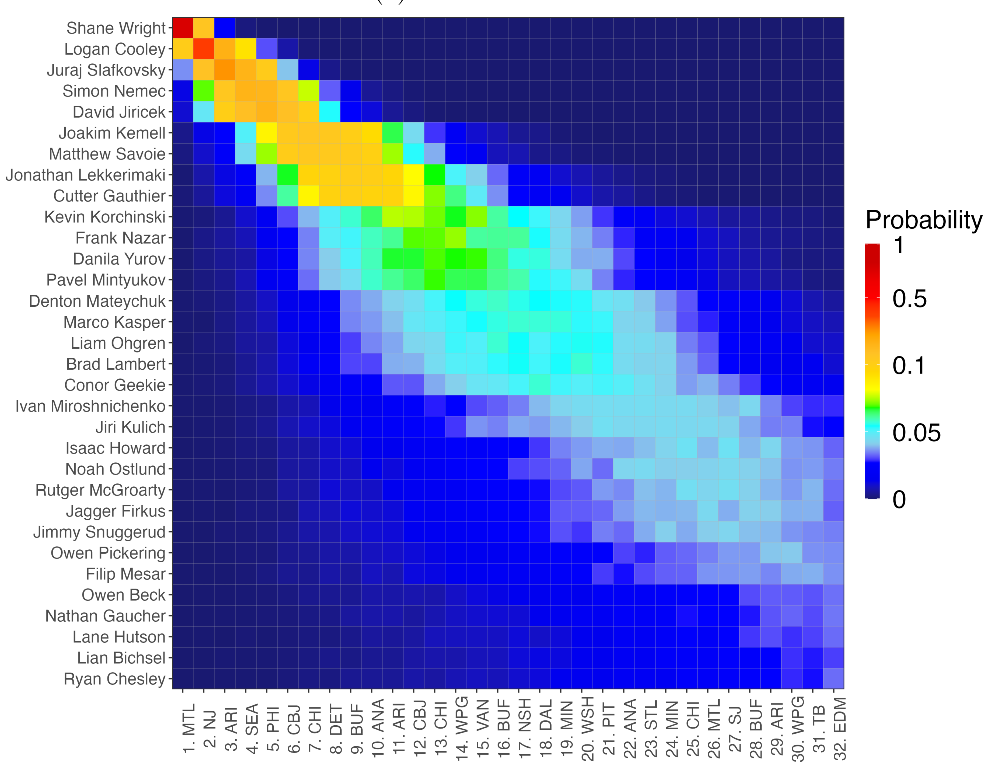
Once the first $m-1$ picks are known, we can also use the same process to simulate forward from the m^{th} pick and obtain the updated probabilities conditional on the known picks. Figures 4.3a and 4.3b display the player-pick pmfs of the 2022 draft after the first three picks in two different scenarios: (a) the actual observed NHL draft results of Slafkovsky, Nemec, and Cooley and (b) the most likely draft result by this model after three hypothesized picks of Wright, Cooley, and Slafkovsky.

In the 5^{th} row of Figures 4.3a and 4.3b we see the pick pmf of David Jiricek. Notice the change in Jiricek's pick pmfs in the two different scenarios. Under the observed draft order with Shane Wright still available, Jiricek has a $<$1% chance of being selected 4^{th} overall; however, if Wright were selected and the best remaining player were Nemec, then Jiricek has approximately a 22% chance of being selected 4^{th}. This example shows how the ROL model can dynamically update selection probabilities to account for changes in the pool of available players.

These player-pick pmfs can be used to provide teams with data-driven information when there are time-sensitive decisions that must be made. Despite slow computational time in fitting the model, it takes approximately 12 seconds to run 10000 simulations of the next 32 picks in the draft, while each team has three minutes to made their pick when they are up next.

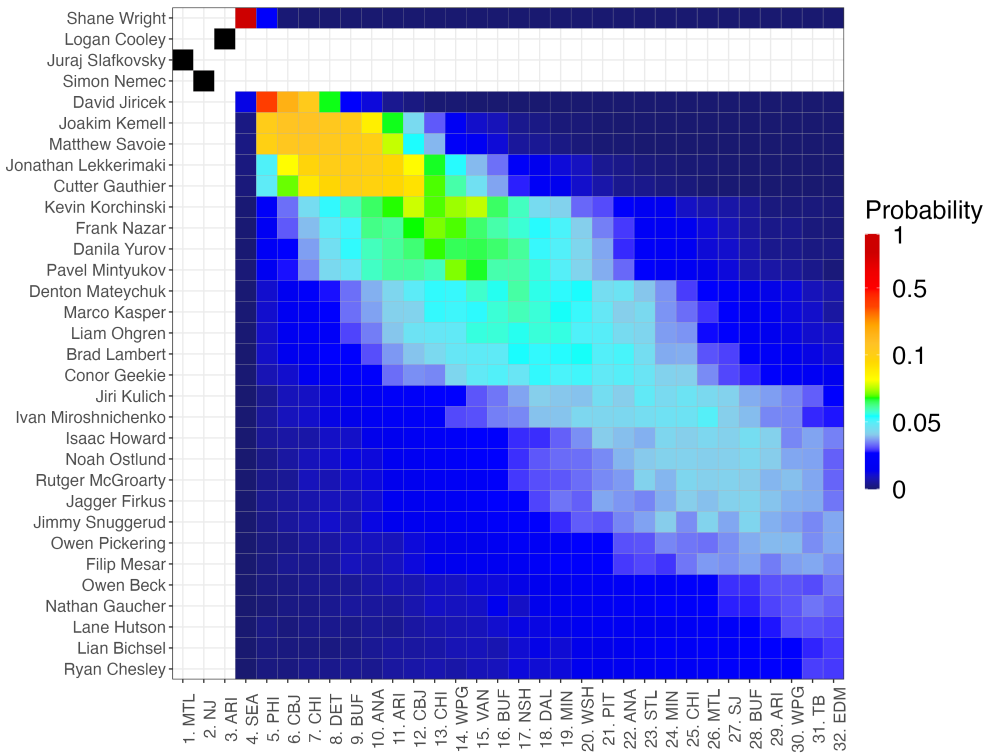


(a) 2021 NHL Draft

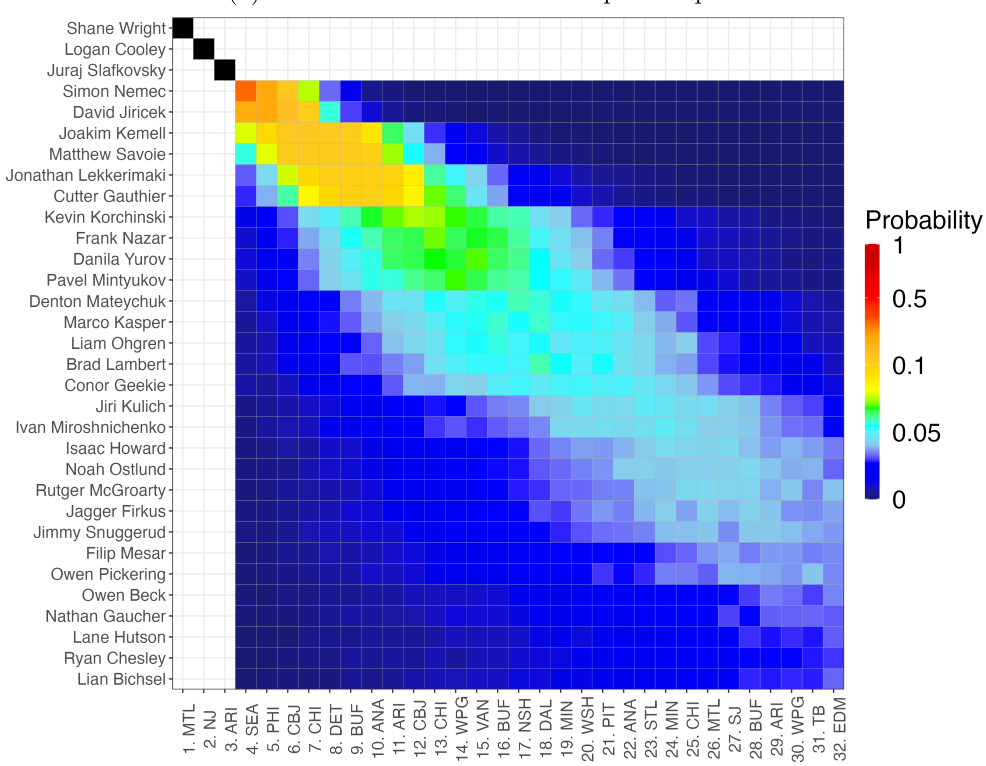


(b) 2022 NHL Draft

Figure 4.2: Player-pick pmfs based on draft simulations immediately prior to the draft from the ROL draft model without agency and team tendencies. Players are listed in descending order based on ability parameters, $\boldsymbol{\theta}$.



(a) Observed outcome for the top three picks



(b) Most likely outcome for the top three picks

Figure 4.3: Player-pick pmfs based on draft simulations from the ROL draft model without agency and team tendencies after the first three picks of the draft.

4.2 With Agency and Team Tendencies

We now generate draft simulations based on posterior draws from the full model with agency and team tendencies as described in (3.5), (**??**), and (3.7) where each parameter is constrained to zero across player or agency and team. Thus, we are able to leverage all of the same insights and features provided in the ROL model without agency and team tendencies that we detailed in Section 4.1. Beyond that, we can also gain insights into draft tendencies from other teams and more accurate estimates of pick probabilities that better align with teams than agencies.

We estimate agency and team tendency parameters for three covariates that we introduce into our model. These covariates are

- x_{i1}: An indicator for whether or not player i is an overager in the current draft year. An overager in the NHL draft is a player that was eligible, but not selected in a prior NHL draft and is thus re-entering the draft as an 'overaged' player.
- x_{i2}: Proportion of games played at professional men's level in draft year.
- x_{i3}: Player height, z-scored by position and draft year.

In Figure 4.4, we display the posterior means of each tendency parameter as shown by the labels for each agency (blue) and team (red) along with horizontal lines to represent the average of the posterior means for both types. The purpose of these plots is to identify differences in draft tendencies between agencies and teams as well as highlight notable teams with respect to each covariate. From these plots, we identify that teams tend to value height more than agencies do, agencies tend to value professional experience slightly more than teams, and overaged players are considered less by agencies than teams.

Figures 4.5a and 4.5b show the updated player-pick pmf after incorporating agency and team tendencies for the 2022 draft both at the beginning of the draft and after the first five picks. These probabilities are not as smooth as Figure 4.2 or Figure 4.3; however, they now adjust for draft tendencies that other teams have shown in the past. For example, notice that the probability Lian Bichsel is selected spikes at Chicago's 13^{th} and 25^{th} overall picks. In Figure 4.4 we see that Chicago has high height and pro experience tendency parameters, while Lian Bichsel is a 6'4" defenceman who played 72.5% of his games in the Swedish Hockey League (SHL), Sweden's top professional league.

4.3 Wyatt Johnston: A Star in the Making?

Recall the example in the introduction of this project about the Dallas Stars' selection of Wyatt Johnston in the 2021 draft. In that example, we posed a question: If the Dallas Stars believed Johnston had the potential to become a star player in the NHL, then would it have been worth it to defer from pick 15 down to 23 and acquire additional picks 48 and 138 from the Detroit Red Wings?

The approach described in Section 3.5.3 allows us to estimate the probability that Wyatt Johnston is selected at any specified pick given all of the information we had available to us at the time of the 15^{th} pick. By simulating these picks, the estimated probabilities are unconditional of all picks that occur from the start of the simulation up to the pick of interest. Thus, by summing the probabilities that he is selected in all prior picks, we can obtain an estimate of the probability that Johnston is still *available* by the specified pick as well.

Figure 4.6 provides the pick cumulative distribution function (cdf) for Johnston over the first three rounds of the NHL draft given the first 14 picks are known. If the Stars kept their pick, there would have been a 100% that Johnston was available to them at 15. If they opted to select someone else with the 15^{th} overall pick, then there would have been an 81.7% chance that Johnston would last until the Stars' next pick at 47^{th} overall. By making the trade down from 15 to 23, their probability of having Johnston available to them only dropped to 97.9%. Assuming Johnston was the best player on their draft board at pick 15 - as Stars' general manager Jim Nill suggested (Nill, 2021) - the Stars acquired picks 48 and 138 for a 2.1% drop in probability that they would be able to draft Wyatt Johnston.

Whether this risk was worth it will unfold over the next decade as Johnston and the other players involved in the trade continue to develop. Tingling (2017) suggests that there is little evidence that teams are able to differentiate talent from the consensus opinion. With this in mind, despite the Johnston's success since the Stars drafted him, a 2.1% decrease in probability that they get their desired player will generally have far less value in return for two additional draft picks.

4.4 Model Evaluation

To evaluate our proposed ROL models, we compare the accuracy of their predictions in the 2022 NHL draft to a pair of notable ranking sets:

Ranking Set	Spearman	Kendall
ROL-WT	**0.910**	**0.727**
ROL-NT	0.854	0.656
TSN	0.884	0.709
SBN	0.850	0.654

Table 4.1: The Spearman and Kendall rank correlation coefficients between the first 50 picks in the 2022 NHL draft and the ROL model with agency and team tendencies (ROL-WT), the ROL model without agency and team tendencies (ROL-NT), TSN Bob McKenzie's final rankings (TSN), and SB Nation Jared Book's consolidated rankings (SBN).

- TSN Bob McKenzie's final ranking set (TSN) for the 2022 draft; McKenzie's list is highly regarded in the hockey community for its consistent accuracy in predicting the results of the draft. This ranking set contains 90 players.

- SB Nation Jared Book's consolidated ranking set (SBN) for the 2022 draft[1]; Book obtains average rankings for draft-eligible players from a collection of 15 different sources. This ranking set contains 154 players.

We obtain a rank ordering from our ROL models with agency and team tendencies (ROL-WT) and without agency and team tendencies (ROL-NT) by calculating the expected value of each player's pick pmf and ranking all players in ascending order.

We compute the Spearman (Spearman, 1904) and Kendall (Kendall, 1955) rank correlation coefficients with the observed 2022 NHL draft results for each of the four rank orderings (TSN, SBN, ROL-WT, ROL-NT). Since TSN and SBN did not provide extensive ranking sets, players that were not ranked on their lists were selected in the draft as early as pick 52. Thus, we only used the first 50 picks in the NHL draft in correlation calculations to avoid dealing with unranked players biasing the results.

Table 4.1 provides the correlation results for these four ranking sets. Among the four rank orderings, our ROL model with agency and team tendencies appears to provide an edge in predictive performance. On the other hand, the ROL model without agency and team tendencies performs approximately the same as the basic consolidated ranking (SBN) in terms of correlation.

[1]We do not use Book's consolidated rankings in the model fitting process due to its clear dependence on other ranking sets.

The increase in correlation obtained by incorporating agency and team tendencies into the model with three simple covariates provides us with a boost beyond Bob McKenzie's highly-regarded ranking set and general consolidated rankings. This model will help teams gain more accurate and transparent estimates of how long a player will remain available in the NHL draft and, in turn, can help them gain a competitive edge during the draft.

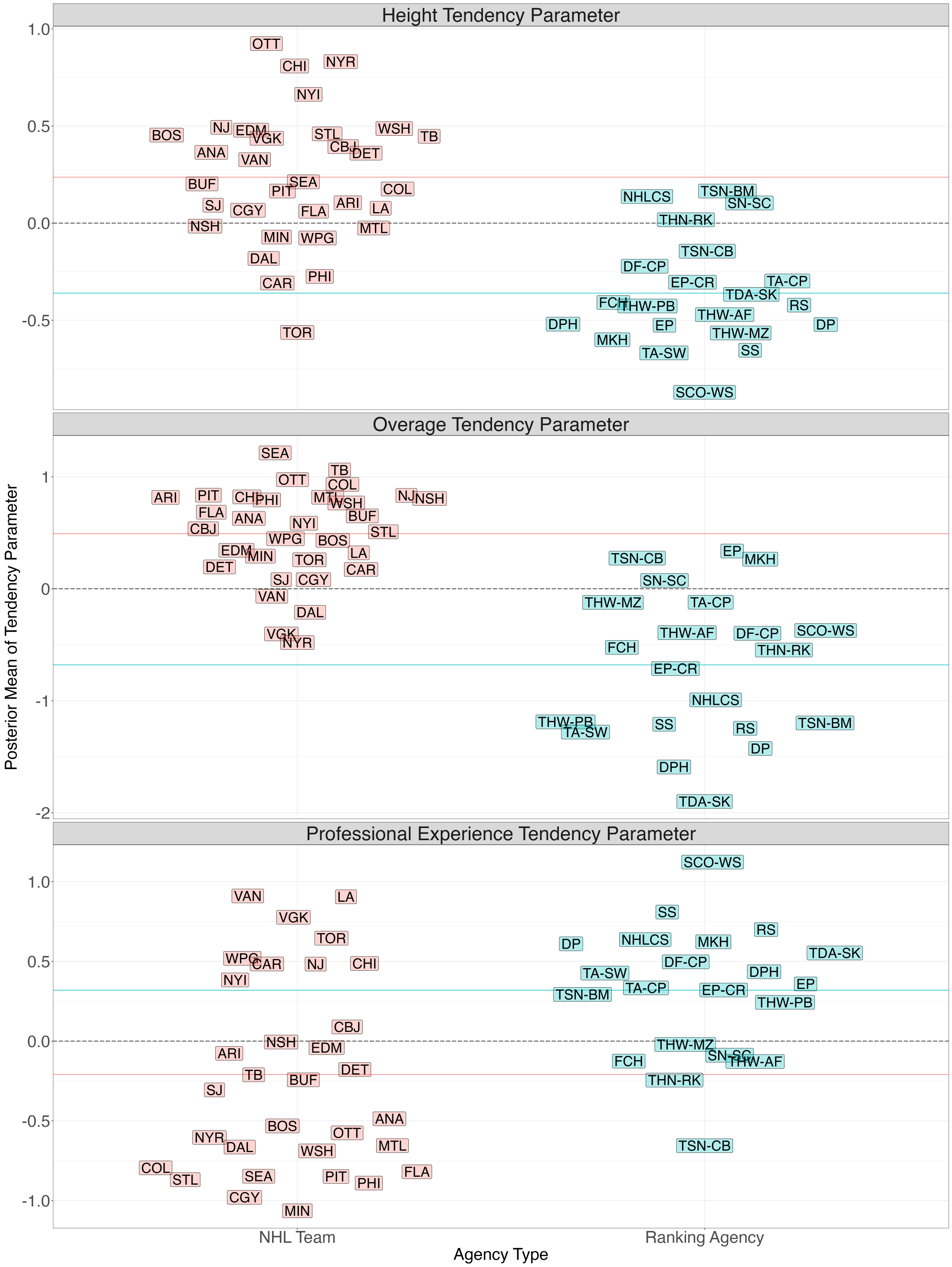
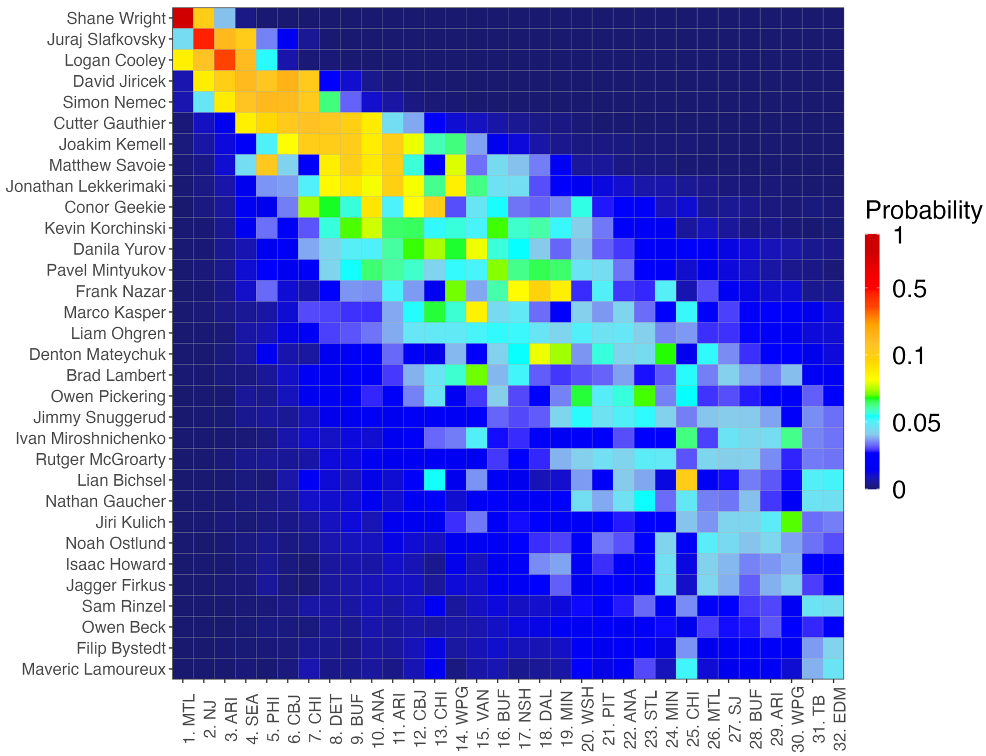
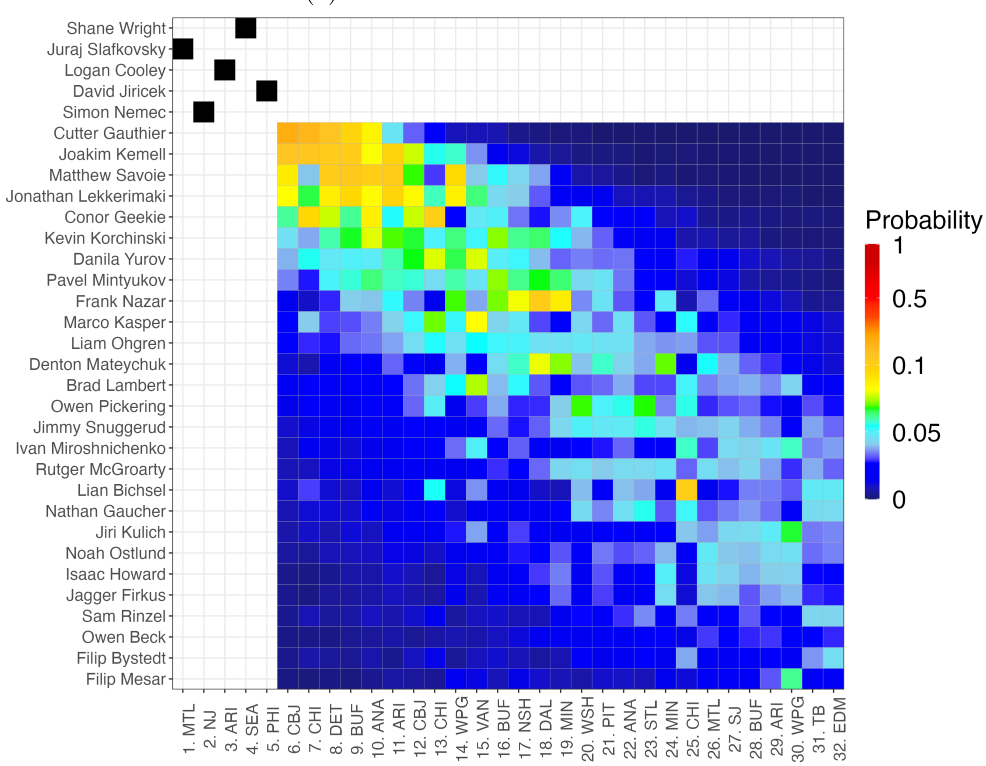


Figure 4.4: A visualization of the posterior means for agency and team tendency parameters; top, middle and bottom represent the height, overager, and professional experience tendency parameters, respectively. The y-axis represents the posterior mean of the parameter and the x-axis differentiates teams and agencies with meaningless jitter to separate individual teams and agencies. The red lines represents the average of the posterior means for teams while the blue lines represents the average of the posterior means for agencies.



(a) Start of the 2022 NHL draft



(b) After the first five picks of the 2022 NHL draft

Figure 4.5: Player-pick pmfs based on draft simulations from the ROL draft model with agency and team tendencies at the start and after five picks into the 2022 NHL draft.

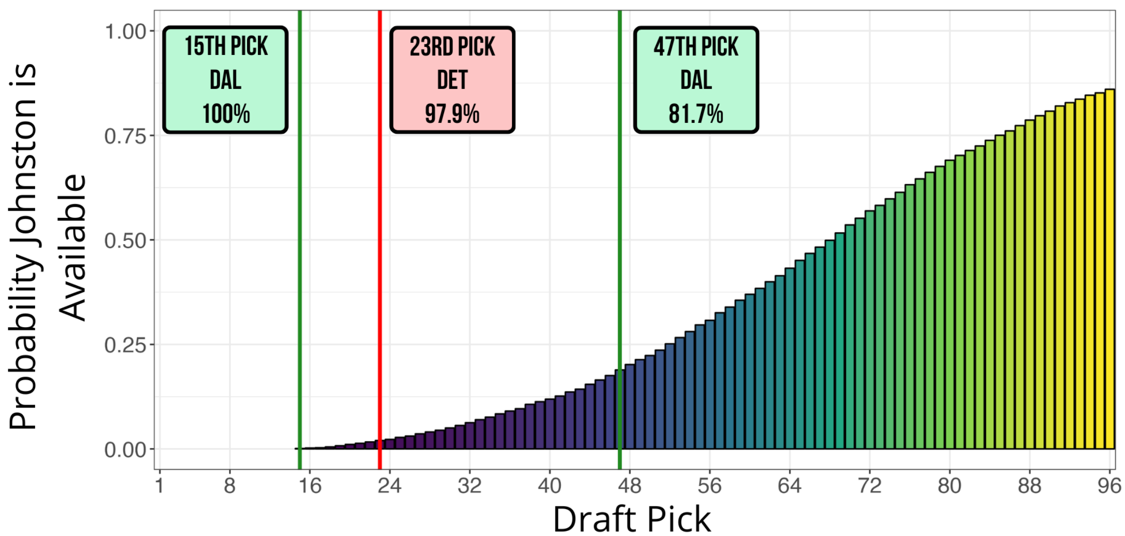


Figure 4.6: The pick cdf for Wyatt Johnston from the 15^{th} overall pick in the 2021 draft until the end of the 3^{rd} round of the draft. Markers represent the probability that Johnston is available at the current pick (15), the Red Wings' next pick (23) and the Stars' next pick (48) from left to right, respectively.

Chapter 5

Discussion

In this project we introduce a novel framework for modelling the outcome of the NHL draft that overcomes many obstacles not tackled by previous approaches. We leverage a Bayesian rank-ordered logit model, which can be thought of a sequence of multinomial logit models, to model the rank-ordered structure of the draft and estimate the probability that each player will be selected at each pick through simulations. In doing so we find ways to update probabilities within the draft conditional on previous picks that have been observed, address unranked players and subset rankings into the model likelihood, and incorporate team and agency tendencies to uncover insights into team strategies and adjust probabilities to better align with team tendencies as opposed to agencies.

Despite building upon current methods, there are still areas of improvement that can enhance model performance. Some areas of exploration include further investigation into both the prior distribution for model parameters and the covariates included in the model, incorporation of team needs, adjusting agency and team tendencies to account for changes in personnel, addressing the dependence between ranking sets, exploring further into model evaluation approaches, and application of the ROL model framework to other sports drafts such as the NFL, NBA or MLB. However, the two main limitations with the proposed framework are computational feasibility and the dependence of team tendencies on previous draft picks. We currently require over 24 hours of computation time to run the model with a sufficient number of iterations, which greatly hinders our ability for investigation into model improvements. Additionally, the model currently assumes that team tendencies will remain the same throughout the draft. This is not necessarily the case as a team will likely adapt their preferences as they satisfy team needs. For example, if a team drafts a goalie with their 1^{st} round pick, we suspect that they will be less likely to select a goalie with any of their future picks in the draft as they will focus on filling other areas of concern on their roster.

We see the ROL model as a state-of-the-art approach to modelling the outcome of major league sports drafts. Teams that implement this framework can gain insights into their opponents draft strategies and have a stronger, data-driven approach to estimate how long a desired player will last before he is selected by another team in the draft.

Bibliography

Agresti, A. (2019). *An Introduction to Categorical Data Analysis.* 3rd ed. John Wiley and Sons, New York. Chap. 6.

Agresti, A., Caffo, B. (2000). Simple and Effective Confidence Intervals for Proportions and Differences of Proportions Result from Adding Two Successes and Two Failures. *The American Statistician* Vol. 54, No. 4, 280-288.

Ali, M.M. (1998). Probability Models on Horse-Race Outcomes. *Journal of Applied Statistics*, Vol. 25, 221–229.

Anderson, A. (2014). Maximum Likelihood Ranking in Racing Sports. *Applied Economics*, Vol. 46, No. 15, 1778-1787.

Bolton, R., Chapman, R. (1986). Searching for Positive Returns at the Track. *Management Science*, Vol. 32, 1040-1060.

Burke, B. (2014). Bayesian Draft Analysis Tool. Accessed online October 15, 2022 at https://www.advancedfootballanalytics.com/2014/05/bayesian-draft-analysis-tool.html

ESPN Analytics (2022). *Draft Day Predictor Tool.* Accessed online October 15, 2022 at http://espnsportsanalytics.com/about.html.

Fok, D., Paap, R., Van Dijk, B. (2010). A Rank-Ordered Logit Model with Unobserved Heterogeneity in Ranking Capabilities. *Journal of Applied Econometrics*, Vol. 27, No. 5, 831–846. https://doi.org/10.1002/jae.1223.

Gerber, E.A.E., Craig, B.A. (2021). A Mixed Effects Multinomial Logistic-Normal Model for Forecasting Baseball Performance. *Journal of Quantitative Analysis in Sports*, Vol. 27, No. 3, 221-239. https://doi.org/10.1515/jqas-2020-0007.

Glickman, M., Hennessy, J. (2015). A Stochastic Rank Ordered Logit Model for Rating Multi-Competitor Games and Sports. *Journal of Quantitative Analysis in Sports*, Vol. 11, No. 3, 131-144. https://doi.org/10.1515/jqas-2015-0012.

Graves, T., Reese, C.S., Fitzgerald, M. (2003). Hierarchical Models for Permutations: Analysis of Auto Racing Results. *Journal of the American Statistical Association*, Vol. 98, 282–291.

Guiver, J., Snelson, E. (2009). Bayesian Inference for Plackett-Luce Ranking Models. In Proceedings of the *26th Annual International Conference on Machine Learning*, ACM, 377–384.

Gumbel, E.J. (1958). *Statistics of Extremes*. Columbia University Press, New York.

Johnson, J., Bruce, A, Yu, J. (2010). The Ordinal Efficiency of Betting Markets: An Exploded Logit Approach. *Applied Economics*, Vol. 42, No. 29, 3703-3709

Kendall, M.G., (1955). *Rank Correlation Methods*. New York: Hafner Publishing Co.

Knappe, K. (2022). How Does the NHL Entry Draft Really Work?. *SB Nation: Pension Plan Puppets*. Available at Accessed online October 18, 2022 at https://www.pensionplanpuppets.com/2022/7/4/23145545/how-does-the-nhl-entry-draft-really-work

Li, M.C. (2019). A Statistical Investigation of Data from the NHL Combine. [Master's thesis, Simon Fraser University]. Simon Fraser University Statistical and Actuarial Sciences Graduate Theses. Accessed online October 18, 2022 at https://www.sfu.ca/stat-actsci/research/research-resources/graduate-theses/2015-2019-thesis.html

Lo, V.S., Bacon-Shone, J. (1994). A Comparison Between Two Models for Predicting Ordering Probabilities in Multiple-Entry Competitions. *The Statistician*, Vol. 43, 317–327.

Luce, R.D. (1959). *Individual Choice Behavior a Theoretical Analysis*. John Wiley and Sons, New York.

Luce, R.D., Suppes, P. (1965). Preferences, Utility and Subjective Probability. In *Handbook of Mathematical Psychology*, Luce, R. Bush, R., Galanter, E. editors, John Wiley and Sons, New York, 249–410.

McFadden, D. (1974). Conditional Logit Analysis of Qualitative Choice Behavior. In *Frontiers in Econometrics*, Zarembka, P. editor, Academic Press, New York. 105–142.

McFadden, D., Train, K. (2000). Mixed MNL Models for Discrete Response. *J. Appl. Econ.*, Vol. 15, 447-470. https://doi.org/10.1002/1099-1255(200009/10)15:5<447::AID-JAE570>3.0.CO;2-1

Nandakumar, N. (2017). What Does It Mean To Draft Perfectly? An Evaluation Of Draft Strategy In The National Hockey League. *Joseph Wharton Scholars*. Available at https://repository.upenn.edu/joseph_wharton_scholars/26

Nill, J. (2021, Jul 24). Wyatt Johnston & Dallas Stars GM Jim Nill React to Dallas Selecting Wyatt in 1st Round | NHL Draft. *Hockey on Fanatics View*. Accessed online October 15, 2022 at https://www.youtube.com/watch?v=Lwv-XLIhUn0

Plackett, R.L. (1975). The Analysis of Permutations. *Applied Statistics*, Vol. 24, 193–202.

Robinson, B. (2020). A Hierarchical Modeling Approach to Predicting the NFL Draft. In proceedings of the *2020 Carnegie Mellon Sports Analytics Conference*. https://www.stat.cmu.edu/cmsac/conference/2020/assets/pdf/Robinson.pdf.

Schuckers, M.E. (2011). What's An NHL Draft Pick Worth? A Value Pick Chart for the National Hockey League. Accessed online October 15, 2022 at https://myslu.stlawu.edu/ msch/sports/Schuckers_NHL_Draft.pdf

Spearman, C. (1904). The Proof and Measurement for Association Between Two Things. *Amer. Journal of Psychology*, Vol. 15, 72–101.

Sprigings, D. (2016). NHL Draft Probability Tool. Accessed online October 15, 2022 at https://hockey-graphs.com/2016/06/08/nhl-draft-probability-tool/

Srinivasan, K.K., Mahmassani, H.S. (2005). A Dynamic Kernel Logit Model for the Analysis of Longitudinal Discrete Choice Data: Properties and Computational Assessment. *Transportation Science*, Vol. 39, No. 2, 160-181.

Stan Development Team (2022a). Stan Modeling Language Users Guide and Reference Manual, Version 2.21.0. https://mc-stan.org

Stan Development Team (2022b). RStan: The R interface to Stan. R package version 2.21.5. https://mc-stan.org/

Stokes, T. (2022). Monaco_ranking. *GitHub*. Accessed online October 15, 2022 at https://github.com/tyrelstokes/Monaco_ranking.

Tea, P., Swartz, T.B. (2022). The Analysis of Serve Decisions in Tennis using Bayesian Hierarchical Models. *Annals of Operational Research*. https://doi.org/10.1007/s10479-021-04481-7

Tingling, P.M. (2017). Educated Guesswork: Drafting in the National Hockey League. In *Handbook of Statistical Methods and Analyses in Sports*, J. Albert, M.E. Glickman, T.B. Swartz and R.H. Koning, editors, CRC Press: Boca Raton, 327-340.

Tulsky, E. (2013). NHL Draft: What Does it Cost to Trade Up?. Accessed online October 15, 2022 at https://www.broadstreethockey.com/2013/4/25/4262594/nhl-draft-pick-value-trading-up

Train, K. (2009). *Discrete Choice Methods with Simulation*. 2nd ed. Cambridge University Press (2009). Chap. 3.

Appendix A

Derivation of Multinomial Choice Probability for Next Pick

Train (2009) provides a full derivation of the multinomial choice probability in the context of the ROL model. Here, we provide a modified excerpt of this derivation in the context of the proposed NHL draft model.

Suppose the agency, a, has ranked the best $m-1$ players out of a pool of N available players and they are moving on to select the next player. Let η_{ia} be the linear predictor for the i^{th} player remaining, $\forall i = m, \ldots, N$. Also assume that the agency assigns each player a latent rating, $Y_{ia} = \eta_{ia} + \epsilon_{ia}$ where $\epsilon_{ia} \sim \text{Gumbel}(0,1)$ is i.i.d. $\forall i = m, \ldots, N$. Since we are only considering one agency, a, in this derivation, we will omit a from the subscripts of Y_i, η_i and ϵ_i for simplicity.

Note that we describe the linear predictor as $\eta_{ia} = \theta_i$ in Chapters 2.3 and 3.1 where θ_i is player i's latent ability parameter, while we generalize it as (3.5) in the final model when incorporating team and agency tendencies. The derivation provided is agnostic for either notation.

Let π_{im} be the probability that player i is ranked m^{th} by the agency, $\forall i = m, \ldots, N$. This can be expressed as

$$\begin{aligned}
\pi_{im} &= P(Y_i > Y_j, \forall j = m, \ldots, N, j \neq i) \\
&= P(\eta_i + \epsilon_i > \eta_j + \epsilon_j, \forall j = m, \ldots, N, j \neq i) \\
&= P(\epsilon_j < \epsilon_i + \eta_i - \eta_j, \forall j = m, \ldots, N, j \neq i)
\end{aligned} \tag{A.1}$$

Recall that the cumulative distribution function and probability density function for the standard Gumbel distribution are given as follows, respectively.

$$F_{\epsilon_j}(y) = \exp\{-\exp\{-y\}\} \tag{A.2}$$

$$f_{\epsilon_j}(y) = \exp\{-y\}\exp\{-\exp\{-y\}\}$$

If we assume ϵ_i is given, then (A.2) becomes the cumulative distribution $\forall \epsilon_j$ evaluated at $\epsilon_i + \eta_i - \eta_j$. Since ϵ_k is i.i.d. $\forall k = m, \ldots, N$, the cumulative distribution over all $j = m, \ldots, N, j \neq i$ is the product of the individual cdfs.

$$\pi_{im}|\epsilon_i = \prod_{j \in \{m,\ldots,N\} \setminus \{i\}} F_{\epsilon_j}(\epsilon_i + \eta_i - \eta_j) = \prod_{j \in \{m,\ldots,N\} \setminus \{i\}} \exp\{-\exp\{-(\epsilon_i + \eta_i - \eta_j)\}\}$$

However, ϵ_i is random, not given. Thus, the choice probability becomes the integral of $\pi_{im}|\epsilon_i$ over all values of ϵ_i, weighted by its density. We express this below where x is a dummy variable used to integrate across ϵ_i.

$$\begin{aligned}
\pi_{im} &= \int f_{\epsilon_i}(x) \left(\prod_{j \in \{m,\ldots,N\} \setminus \{i\}} F_{\epsilon_j}(x + \eta_i - \eta_j) \right) dx \\
&= \int \exp\{-x\}\exp\{-\exp\{-x\}\} \left(\prod_{j \in \{m,\ldots,N\} \setminus \{i\}} \exp\{-\exp\{-(x + \eta_i - \eta_j)\}\} \right) dx \\
&= \int_{-\infty}^{\infty} \exp\{-x\} \left(\prod_{j \in \{m,\ldots,N\}} \exp\{-\exp\{-(x + \eta_i - \eta_j)\}\} \right) dx \\
&= \int_{-\infty}^{\infty} \exp\{-x\} \exp\left\{ - \sum_{j \in \{m,\ldots,N\}} \exp\{-(x + \eta_i - \eta_j)\} \right\} dx \\
&= \int_{-\infty}^{\infty} \exp\{-x\} \exp\left\{ - \exp\{-x\} \sum_{j \in \{m,\ldots,N\}} \exp\{\eta_j - \eta_i\} \right\} dx
\end{aligned}$$

Define $t = \exp\{-x\}$ such that $dt = -\exp\{-x\}dx$.

Note that, as $x \to \infty$, $t \to 0$, while as $x \to -\infty$, $t \to \infty$. Thus, we can express π_{im} as

$$
\begin{aligned}
\pi_{im} &= \int_{\infty}^{0} \exp\left\{-t\sum_{j=m}^{N}\exp\{\eta_j-\eta_i\}\right\}(-dt) \\
&= \int_{0}^{\infty} \exp\left\{-t\sum_{j=m}^{N}\exp\{\eta_j-\eta_i\}\right\}dt \\
&= \left.\frac{\exp\left\{-t\sum_{j=m}^{N}\exp\{\eta_j-\eta_i\}\right\}}{-\sum_{j=m}^{N}\exp\{\eta_j-\eta_i\}}\right|_{0}^{\infty} \\
&= \frac{1}{\sum_{j=m}^{N}\exp\{\eta_j-\eta_i\}} \\
\pi_{im} &= \frac{\exp\{\eta_i\}}{\sum_{j=m}^{N}\exp\{\eta_j\}}
\end{aligned}
$$

Therefore, the probability that player i will be selected with the next pick can be expressed as a multinomial choice probability with categories representing each of the draft-eligible players that remain unranked. This can be applied $\forall m = 1, \ldots, N$.

Appendix B

Composite Score for Ability Parameter Prior Distribution

Each draft-eligible player receives a composite score which is used as the mean in the prior distribution for his ability parameter. The vector of all composite scores is denoted $\boldsymbol{\mu}_\theta = [\mu_{\theta 1}, \ldots, \mu_{\theta Q}]$ where $Q = \sum_{r \in R} N_r$ is the total number of draft-eligible players considered in the model over all draft years.

The goal of this composite score is to adjust our prior distribution to better match popular opinion prior to the ranking agencies releasing their final ranking sets at the end of the hockey season.

The composite score is determined through an empirical Bayes procedure based on ranking sets that were published before May 1^{st} leading up to the draft. We define this composite score for player i as

$$\mu_{\theta i} = \frac{c_1(w_{ri1} + z) + c_5(w_{ri5} + z) + c_{32}(w_{ri32} + z) + c_{\mathrm{u}}(w_{\mathrm{riu}} + z)}{w_r + 4z}$$

where

- r is the draft year for player i.
- w_r is the total number of ranking sets in draft year r that were released before May 1^{st}.
- w_{ri1} is the total number of ranking sets in draft year r that were released before May 1^{st} where player i was ranked 1^{st} overall.
- w_{ri5} is the total number of ranking sets in draft year r that were released before May 1^{st} where player i was ranked between 2^{nd} and 5^{th} overall.

- w_{ri32} is the total number of ranking sets in draft year r that were released before May 1^{st} where player i was ranked between 6^{th} and 32^{nd} overall.

- w_{riu} is the total number of ranking sets in draft year r that were released before May 1^{st} where player i was ranked outside of the top 32 players.

- c_* are constants corresponding to each w_{ri*} that must be tuned in assigning the prior score.

- z is a regularizing constant that is used to add $4z$ pseudo-observations with z of each assigned to each of the four categories $(w_{ri1}, w_{ri5}, w_{ri32}, w_{riu})$. This is based on the 'add 2 successes and 2 failures' approach by Agresti & Caffo (2000) to improve the performance of confidence intervals for proportions.

After manual tuning based on our domain knowledge of the NHL draft, we set $c_1 = 12$, $c_5 = 6$, $c_{32} = 2$, $c_u = 0$ and $z = 2$, to best align with our prior beliefs. Since the prior means of the ability parameters produced from this score, $\mu_{\theta i}$, are not naturally interpretable, we use the softmax function to convert $\mu_{\theta i}$ into an estimate of the probability that each player will be selected first and performed manual tuning in these terms.

We also explored a prior of $\boldsymbol{\mu}_\theta = \mathbf{0}$; however, we found the composite score system described in this section converged quicker and provided an increase in computational speed as compared to the $\mathbf{0}$-mean case.

Appendix C

Agency and Team Abbreviations

Agency	Abbreviation
The Athletic - Corey Pronman	TA-CP
The Athletic - Scott Wheeler	TA-SW
Daily Faceoff - Chris Peters	DF-CP
Dobber Prospects - Full Staff	DP
The Draft Analyst - Steve Kournianos	TDA-SK
Draft Prospects Hockey - Full Staff	DPH
Elite Prospects - Full Staff	EP
Elite Prospects - Cam Robinson	EP-CR
Future Considerations Hockey - Full Staff	FCH
The Hockey News - Ryan Kennedy	THN-RK
The Hockey Writers - Peter Baracchini	THW-PB
The Hockey Writers - Andrew Forbes	THW-AF
The Hockey Writers - Matthew Zator	THW-MZ
McKeen's Hockey - Full Staff	MKH
NHL Central Scouting - Full Staff	NHLCS
Recruit Scouting - Full Staff	RS
Scouching - Will Scouch	SCO-WS
Smaht Scouting - Full Staff	SS
Sportsnet - Sam Cosentino	SN-SC
TSN - Craig Button	TSN-CB
TSN - Bob McKenzie	TSN-BM

Table C.1: A list of all agencies considered in the model accompanied by corresponding abbreviations.

Team	Abbreviation
Anaheim Ducks	ANA
Arizona Coyotes	ARI
Boston Bruins	BOS
Buffalo Sabres	BUF
Calgary Flames	CGY
Carolina Hurricanes	CAR
Chicago Blackhawks	CHI
Colorado Avalanche	COL
Columbus Blue Jackets	CBJ
Dallas Stars	DAL
Detroit Red Wings	DET
Edmonton Oilers	EDM
Florida Panthers	FLA
Los Angeles Kings	LA
Minnesota Wild	MIN
Montreal Canadiens	MTL
Nashville Predators	NSH
New Jersey Devils	NJ
New York Islanders	NYI
New York Rangers	NYR
Ottawa Senators	OTT
Philadelphia Flyeers	PHI
Pittsburgh Penguins	PIT
San Jose Sharks	SJ
Seattle Kraken	SEA
St. Louis Blues	STL
Tampa Bay Lightning	TB
Toronto Maple Leafs	TOR
Vancouver Canucks	VAN
Vegas Golden Knights	VGK
Washington Capitals	WSH
Winnipeg Jets	WPG

Table C.2: A list of all teams considered in the model accompanied by corresponding abbreviations.