# Improving Upon Existing RAPM Models for NHL Player Evaluation by Using Informative Priors in a Bayesian Hierarchical Model

RAPM (Regularized Adjusted Plus-Minus) is a metric that is commonly used to evaluate the offensive and defensive skill of hockey players. Traditionally, a ridge regression is run on shift-level event data to quantify how much each player contributes to various on-ice outcomes. The shift-level event data can be compiled from publicly available sources. The outcome chosen (e.g. goals/expected goals/shots for or against) is used as the dependent variable and dummy variables indicating the players on the ice during a shift are the predictors in the model. Often times other predictors are also included (e.g. number of skaters on the ice for each team, score of the game at the time of the shift, etc). Then, the estimated coefficient for a player is used to represent that player's offensive or defensive ability.

A ridge regression is usually used instead of OLS, since the shrinkage of the regression coefficients can help to prevent overfitting. However, this regularization is still unable to address the substantial multicollinearity present in the shift-level data. This multicollinearity exists because forwards often play on the same lines together, and similarly defense usually play most of their minutes with a single defense partner. Some hockey analytics professionals calculate new measures (e.g. Box Plus-Minus or Statistical Plus-Minus) as a function of RAPM to try and get away from this multicollinearity. They then use these new measures for player evaluation, or even feed these into additional calculations to calculate GAR/xGAR/WAR (Goals/Expected Goals/Wins above Replacement). This approach still produces player assessments that are sometimes quite divorced from reality, since they are still rooted in RAPM metrics that are plagued by multicollinearity. Also these calculations get unnecessarily complicated.

I am currently working on a Bayesian RAPM model which is more immune to issues of multicollinearity and can therefore provide more accurate estimates of a player's offensive and defensive ability. Others have implemented Bayesian RAPM models before, but my unique contribution is two-fold:

1) How I define my normally-distributed informative priors
    a. I combine information about a player's ice time and box score statistics to yield reasonable prior distributions.
    b. Using a player's ice time helps to level the playing field across different teams (elite players on bad teams will still be viewed favorably due to high TOI).
    c. Using box score statistics helps the model to tease out which players on a team are driving outcomes on the ice versus which players have less of an impact.
2) The hierarchical structure of my Bayesian model
    a. Player TOI and box score statistics only inform the means of players' prior distributions relative to each other.
    b. Defining hyperparameters in a hierarchical model helps to turn this relative knowledge about players into concrete values to use for prior distribution means.
    c. The variance of the prior distributions for each player can also be modeled as a random variable whose distribution is defined by hyperparameters.