Developing an Analytics Program for Major Junior Hockey

Joshua Pohlkamp-Hartt

Queen's University

August 11, 2015

Motivation

Maple Leafs shake up front office, hire **stats guru** *Kyle Dubas, 28, as assistant GM* *(*Kevin McGran, Toronto Star*)*

There is no statistic to accurately quantify neutral-zone play. *(*Steve Simmons, Toronto Star*)*

Table of Contents

1. Preliminaries
2. The Spericity Tests
3. Simulations and Comparison

Data Collection - Overview

- Tried recording on paper - Impossible for velocity of data!
- Developed JAVA program to meet speed of the game.
- Hired/Trained stats undergrads to collect the data.
- Home games are easy but away games are difficult.
- Including scoring and penalty data through web scraping.

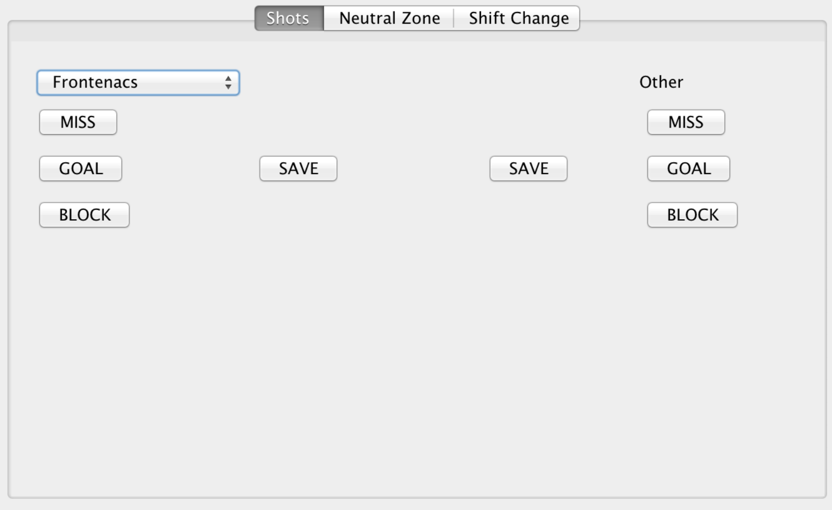
Data Collection - Data Recorded

We had three data collection roles: shots, neutral zone play, shifts. The data recorded was:

- Shots: team, type (blocked, missed net, goal, save), and time.
- Neutral Zone: outcome, team, and time.
- Shifts: player, on/off, time.

Data Collection - Java

Example of Java program shots page.



Data Collection - Home vs Away Games

For home games we were able to watch from the press box. This gave us a full field of view which allowed us to easily monitor all of the play.
For away games we had to watch the scouting film provided by the other team. These films were generally of low quality. An additional issue is the inability to monitor shift changes easily. To avoid collection issues, repeated viewings and strict monitoring of line combinations was needed.

Data Collection - Other Data Sources

In addition to the data we recorded at games we collected information on penalties and scoring.
To do this we scraped the data from the OHL website using the rvest package in R.

Data Analysis - Simple Data

The first things we were able to report were simple statistics that were directly calculated from the available data. Some of these statistics were:

- Shooting: Shot Attempts, Corsi, On Net %, Shooting %, and Even Strength Goals per 60 minutes
- Special Teams: Power Play %, Penalty Kill %, and Power Play Goals per 2 minutes.
- Neutral Zone: Offensive Neutral Zone Success, Defensive Neutral Zone Success

All of these stats were broken down into periods and situation.

Data Analysis - Estimated Neutral Zone Differential Formula

To give an overall assessment of neutral zone play we developed a statistic to take into account both a players offensive and defensive neutral zone skills. Known as the Estimated Neutral Zone Differential (END), we referenced a player's success rate to the average rate for similar players. For a player X and reference group G, we have

$$END(X,G) = (\hat{p}_O(X) - \hat{p}_O(G)) + (\hat{p}_D(X) - \hat{p}_D(G)). \tag{1}$$

The group, G, of players referenced against could be other teammates that play the same position (forward or defense) or league wide. This choice is dependent on what you would like to investigate.

Data Analysis - Estimated Neutral Zone Differential Analysis

To examine if the *END* statistic was a good metric of player quality, we examined the relationship between where in the lineup a player would play and their *END* statistic relative to the team.

We modeled the probability of a player being on the top two lines given the p-value for that player's *END* being positive. The fitted model for this past season's data was $\hat{P}(Top\,2\,lines) = L(.59 - 1.59\times\text{p-value})$ with p-value on the significance for the coefficient being .0947.

This indicates that as a player is more likely to have an *END* value that is not positive, he is more likely to be a bottom end player.

Data Analysis - Line Optimization Formulation

The next thing that we thought would be useful for the team was a way to identify strong line combinations. To do this we examined modeled players and their lines (2 or 3 term interactions) with the probability of a goal being scored in the next minute. We performed this for goals for and against separately then evaluated the difference in cumulative pessimistic confidence bounds on the coefficients of our model. That is for a line combination we get the metric,

$$\Delta_\alpha = (\sum_{i=1}^{18}(\beta_{i,gf} + \Phi(\alpha)S(\beta_{i,gf})) + \sum_{j=1}^{3}(\gamma_{j,gf} + \Phi(\alpha)S(\gamma_{j,gf})) + \sum_{k=1}^{4}(\zeta_{k,gf} +$$

$$- (\sum_{i=1}^{18}(\beta_{i,ga} + \Phi(1-\alpha)S(\beta_{i,ga}) + \sum_{j=1}^{3}(\gamma_{j,ga} + \Phi(1-\alpha)S(\gamma_{j,ga})) + \sum_{k=1}^{4}(\zeta \tag{2}$$

Where α is how pessimistic we are.

Data Analysis - Line Optimization Application

To apply this metric and we start by identifying the line combinations we are interested in using. Then to avoid issues with unequal use of line combinations in the data, we employed a bagging algorithm. For each sampling from our data we selected the line combination with the largest Δ_α. Then we reported the optimal line combination to the combination selected most often across the samples.

Data Analysis - Line Optimization Example

Example of Histogram for bagged line combinations. We wanted to determine the optimal line combination for use in the second round of the playoffs from the data collected in the last 10 games of the season and first round of the playoffs.

The choice of α will alter the selection. Smaller values will promote line combinations that have more evidence of being successful, larger values will promote more unproven combinations. We often used $\alpha = .1$.

Data Analysis - Player Monitoring Formulation

To provide the coaches with a way to monitor and evaluate player performance within games we used exponentially weighted moving average quality control charts.

$$\begin{aligned}\Delta_\alpha = (&\sum_{i=1}^{18}(\beta_{i,gf} + \Phi(\alpha)S(\beta_{i,gf})) + \sum_{j=1}^{3}(\gamma_{j,gf} + \Phi(\alpha)S(\gamma_{j,gf})) + \sum_{k=1}^{4}(\zeta_{k,gf} + \\ - (&\sum_{i=1}^{18}(\beta_{i,ga} + \Phi(1-\alpha)S(\beta_{i,ga}) + \sum_{j=1}^{3}(\gamma_{j,ga} + \Phi(1-\alpha)S(\gamma_{j,ga})) + \sum_{k=1}^{4}(\zeta\end{aligned} \tag{3}$$

Where α is how pessimistic we are.

Simulations

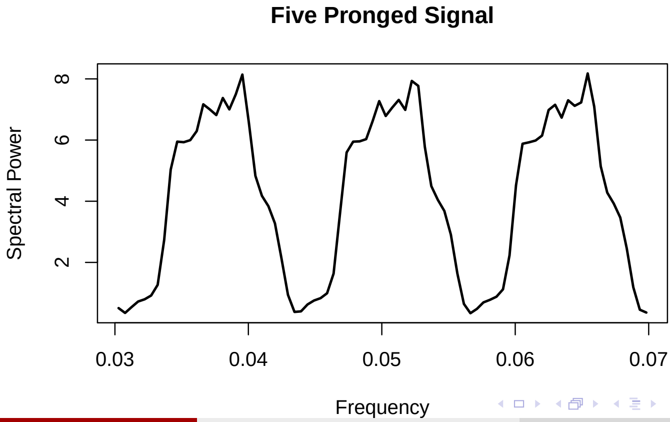
- We wanted to check if the Sphericity tests would give reasonable choices of parameters on a synthetic data set.
- We used a data set of 1000 points that was made up of 38 5-pronged sinusoidal signals of width .08Hz that are evenly spaced across the frequency band at .13Hz apart.

$$X(t) = \sum_{l=1}^{38} \alpha_l \sum_{j=-2}^{2} (.3 - .1 \mid j \mid) \sin(2\pi(.013l + .002j)t) + z_t, \quad (4)$$

  where α_l is a random amplitude for each signal that is taken from $U(.5, 1)$ and $z_t \sim N(0, 2)$.
- For the best MTM estimate, we wanted the frequency band to cover one signal, so ideally $NW \in [4, 6.5]$.

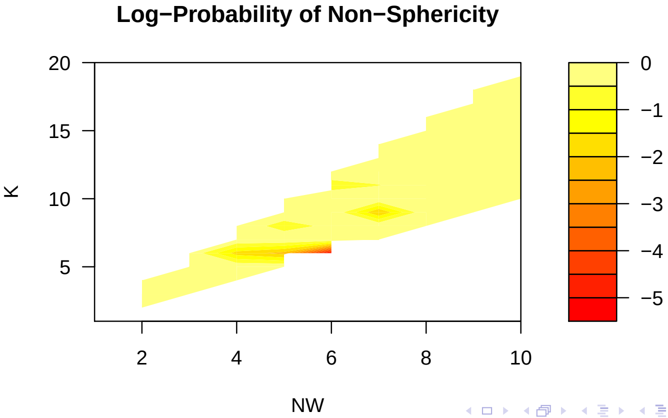
Synthetic data example

- This is an example of 3 of the 5-pronged sinusoidal signals from the data set. For this MTM estimate we used $NW = 4$, $K = 7$, $N = 1000$.



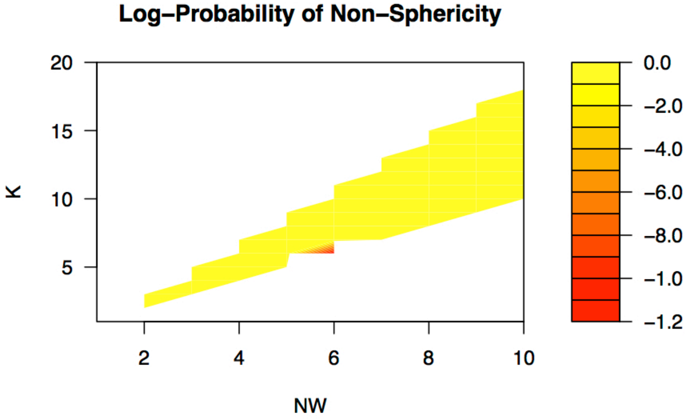
Naive Sphericity Test simulation results

- We tested the naive sphericity of the residuals for parameters $NW \in [2, 10]$ and $K \in [2, 20]$. The test found $NW = 6$ and $K = 6$ had the lowest probability ($p = .006$).



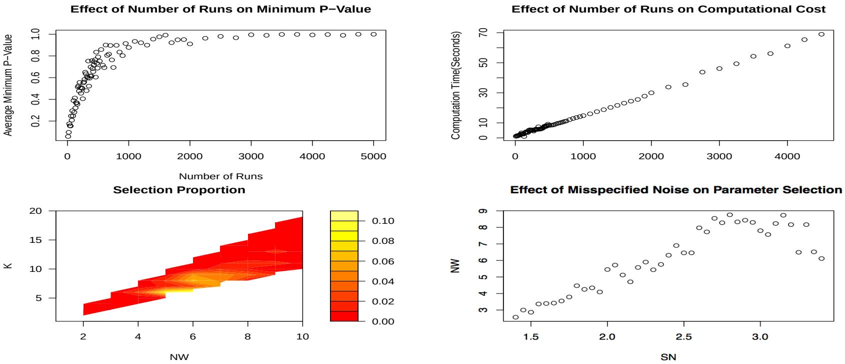
Bagged Sphericity Test simulation results

- For the bagged sphericity test with $M = 4000$ on the parameters $NW \in [2, 10]$ and $K \in [2, 20]$ the results were the same with $NW = 6$ and $K = 6$ having the lowest probability($p = .997$).



Bagging Test Drawbacks

- The bagging test does run into several drawbacks; p-values are dependent on M, computationally expensive for consistent results, variability of minimum probability parameters from test to test, and sensitivity to noise power estimate.



Comparison

- Both tests can provide a reasonable choice of parameters when no theoretical background is possible.
- The bagged sphericity test can be helpful in situations where the noise level is known.
- The bagged sphericity test does suffer from several drawbacks.
- For quick checks for parameter selection we recommend the naive approach but if the noise level is well estimated and computer power is a non-issue then the bagged test is ideal.

Acknowledgments

This research is supported by Queen's University and my supervisors Dr. Takahara and Dr. Thomson.

References

- T.W. Anderson, *An introduction to multivariate statistical analysis*, vol. 2, Wiley, 1958.
- Basil P Korin, *On the distribution of a statistic used for testing a covariance matrix*, Biometrika **55** (1968), no. 1, 171–178.
- S. Lahiri, *Resampling methods for dependent data*, vol. 14, Springer, 2003.
- D. Slepian and H.O. Pollack, *Prolate spheroidal wave functions, fourier analysis and uncertainty - I*, Bell System Technical Journal **40** (1961), no. 1, 43–64.
- R. Tibshirani T. Hastie and J. Friedman, *The elements of statistical learning*, vol. 1, Springer, 2001.
- D.J. Thomson, *Spectrum estimation and harmonic analysis*, Proceedings of the IEEE **70** (1982), no. 09, 1055–1096.