The Annals of Applied Statistics
2013, Vol. 7, No. 3, 1497–1524
DOI: 10.1214/13-AOAS646

COMPETING PROCESS HAZARD FUNCTION MODELS FOR PLAYER RATINGS IN ICE HOCKEY

BY A. C. THOMAS*, SAMUEL L. VENTURA*,
SHANE T. JENSEN† AND STEPHEN MA*

Carnegie Mellon University and University of Pennsylvania†*

Evaluating the overall ability of players in the National Hockey League (NHL) is a difficult task. Existing methods such as the famous "plus/minus" statistic have many shortcomings. Standard linear regression methods work well when player substitutions are relatively uncommon and scoring events are relatively common, such as in basketball, but as neither of these conditions exists for hockey, we use an approach that embraces the unique characteristics of the sport. We model the scoring rate for each team as its own semi-Markov process, with hazard functions for each process that depend on the players on the ice. This method yields offensive and defensive player ability ratings which take into account quality of teammates and opponents, the game situation, and other desired factors, that themselves have a meaningful interpretation in terms of game outcomes. Additionally, since the number of parameters in this model can be quite large, we make use of two different shrinkage methods depending on the question of interest: full Bayesian hierarchical models that partially pool parameters according to player position, and penalized maximum likelihood estimation to select a smaller number of parameters that stand out as being substantially different from average. We apply the model to all five-on-five (full-strength) situations for games in five NHL seasons.

1. Introduction.

In many situations where a desired outcome depends on the performance of a group, it can be difficult to evaluate the individual contributions of its members. The study of sports provides a number of examples; the easier decomposition of baseball into what are essentially head-to-head match-ups makes it comparatively easy to tell whether one batter is superior to another, given enough observations.

The study of goal-based team sports—ice hockey, field hockey, basketball, soccer and lacrosse, among others—is considerably more difficult, as the separation of roles is much more difficult to measure with modern game statistics, especially when player efforts do not directly lead to goals. In ice hockey, player abilities are historically quantified by citing offensive statistics, such as goals and assists, defensive statistics such as blocked shots and a goaltender's saves, and combinations such as the plus/minus statistic ($+/-$), or the net goals scored for a player's team

Received October 2012; revised March 2013.

Key words and phrases. Competing process models, Cox model, ice hockey, Bayesian inference, hierarchical models, player ratings.

when that player was on the ice. However, these are measured across many different combinations of players on the ice who contribute to the play, so an overall assessment of individual ability is not as obvious. Even if we assume that goaltenders have no role in team offense, there is surely a defensive assessment that can be made for other players, which is not as easily captured by these count-based statistical measures.

The nature of ice hockey means that scoring events are often quite rare. Even if we divide an hour-long game into segments of a minute each, the fact that most hockey games have fewer than ten total goals means that the majority of these will be empty of scoring events, requiring a treatment that is considerate of this imbalance; segments of unequal length must also be handled appropriately. This rarity may also mean that the data could be insufficient to adequately separate players from each other in their ratings, or have minimal predictive value, either for a player's own future performance or for in-game outcomes. Properly assessing uncertainty in all parameters, and in the stochastic process of the game itself, is essential to the process.

To manage these factors and generate meaningful player ratings, we propose to measure the abilities of players in ice hockey according to goal-scoring rates when they are on the ice, much as in the plus/minus approach. However, we have two particular features of our approach that improve upon plus/minus. First, we consider goal-scoring to be the combination of at least two semi-Markov processes, modulated by the players on the ice for each team, so that each player on the ice contributes to both their team offense and team defense. Second, we regularize these estimates to ensure better predictive performance, which may also have the benefit of selecting a subset of players to have nonzero (i.e., nonaverage) ratings.

Ideally, our method for obtaining meaningful player ratings will have several important properties. We want ratings that can be interpreted in terms of game outcomes—namely, goals scored or prevented. In that spirit, we want to distinguish the offensive and defensive capabilities of each player separately, allowing for a superior assessment of ability, as well as the quality of a player's team, teammates and opposition by factoring their abilities into each observed event. In some cases, we would also like to distinguish a subset of players as "exceptional" at offense or defense (in either direction).

We continue by describing previous methods for rating the offensive and defensive skill for players in hockey and other sports in Section 2, as well as describing the data available for this work. In Section 3 we describe our methodological approach to the problem, demonstrating many of its applications in Sections 4 and 5. We conclude in Section 6 by discussing potential extensions to our approach.

2. Previous approaches for player ratings.

2.1. *Count-based measures*: *Simple plus/minus*. The notion of tracking the number of goals scored, both for and against, for each player on the ice is decades old, but its full application took years to reach its current state. In the National Hockey League (NHL), the world's premier professional ice hockey organization, its initial use was said to be pioneered by the Montreal Canadiens hockey club in the 1950s, though only for their own purposes and in secret [Klein and Reif (2001), Chapter 1]. The system was popularized by NHL coach Emile Francis during the 1960s, though the existing weaknesses of this approach were obvious even then: it does not take into account each player's quality of teammates, quality of opponents and position. Good players on bad teams often have similar plus/minus statistics as bad players on good teams [Klein and Reif (2001), Chapter 3].

Without changing the basic structure of the statistic, the most obvious weakness one can address is the effective rarity of goals, an average of roughly three per team per game. By adding other events that can lead to goals, more information can be attributed to the efforts of players on the ice. These typically include shots on goal, either unweighted or adjusted for the distance from the net, possibly including those that are blocked by the opposing team's skaters or miss the net entirely; these include the Fenwick- and Corsi-weighted plus/minus; Macdonald (2012a) lists these and others that have been adapted to the general approach.

Lock and Schuckers (2009) and Schuckers et al. (2011) extend this idea by accounting for all events that are recorded in a modern NHL game, including face-offs, turnovers and hits, all of which are thought to change the likelihood of the scoring of goals, either due to changes in puck possession or location on the ice. Each of these has an effective "weight" in terms of the expected number of goals scored or prevented because that event did or did not occur; for example, a team that wins a face-off near their opponent's goal is more likely to score in the following seconds than they are to be scored upon, and have a higher probability of scoring than if their opponent had won the face-off instead. For a player in a game, the sum of the weights of events in which they are involved can then inform us about that player's overall contribution to the game.

2.2. *Regression-adjusted measures.* The other most notable weakness of the standard plus/minus measure, or any of its derivatives, is coincident play: if two or more players are on the ice together for much of their shared time, it can be difficult to distinguish the abilities of each player from each other when so many of the outcomes to which they contribute are common to both. This problem is common to all goal-based team sports.

To handle this issue in basketball, Rosenbaum (2004) proposed to divide a National Basketball Association (NBA) game into intervals marked by the substitution of players onto the court. From this, he derived a number of independent events, each containing a number of scoring opportunities for each team. The outcome of each event is the difference in points scored between the two teams divided by the time elapsed during the interval; the predictors are indicators of the players on the court for each team—positive for the home team, negative for the away team. Using a linear regression model of these player-predictors on the scoring outcome, each player's associated coefficient represents their contribution to the change in score in favor of their team; this is their "adjusted plus/minus" rating. Ideally, this measure will isolate a player's contribution to their own rating and remove it from others, as the quality of their teammates and their opponents is accounted for.

Ilardi and Barzilai (2008) modify this approach by taking every interval as not one but two events—home scoring and away scoring—and treating them as independent, conditional on the length of the event. Each player on the court appears in each of these two events, as an offensive and defense player respectively, and therefore has a distinct rating for each of these "skills;" the combination of the two can then be taken as the total adjusted player rating.

Each of these procedures was conducted by Macdonald (2011) on NHL data by noting player substitutions from official game logs and using these to construct a table of events. Gramacy, Jensen and Taddy (2013) considers a logistic regression model that focuses only on those events where one team scores a goal, which has the benefit of considering a much smaller set of events. Neither of these models allows for a user to simulate an entire game; the outcomes do not correspond to goal scoring processes, but to scoring rates in the former case and relative ability in the latter.

These methods have also been used with regularization methods to improve parameter estimates; ridge regression [Hoerl and Kennard (1970)] is used by Sill (2010) for the NBA, and Macdonald (2012b) for the NHL, to account for these difficulties. Other Gaussian-derived models such as James-Stein estimation [James and Stein (1961)] are compared for the case of batting averages in Brown (2008).

2.3. *Process models.* The nature of substitution and scoring data from the NBA is vastly different from that of the NHL. In the NBA, there are typically several scoring events for either team per rotation (the equivalent of a "shift" in hockey), and there are relatively few substitutions per game. In the NHL, scoring events are much rarer, on the order of 10 minutes between goals, while players typically only spend about 30–60 seconds on the ice before returning to the bench for a substitution. As we show in Section 2.4, roughly 98% of these intervals have a total of zero goals scored. Using this linear regression approach, the event durations will not factor in, and significant information will be lost. Additionally, since the data are clearly non-Gaussian, methods based on Gaussian convergence properties may not be reliable, as the error terms and the prediction terms must be highly dependent to produce the majority-zero data.

The rarity of scoring events relative to the number of observable intervals suggests the use of a Poisson-type process model. Each event represents an observation of the same players on the ice, and any event that does not end in a goal is essentially censored by the change in players. This directly incorporates the observed duration of the event as well as accounting for the relatively sparse number of goals. Simple Poisson models have been used for making strategic decisions in hockey [Morrison (1976), Beaudoin and Swartz (2010)]; these methods can be improved to account for heterogeneity in the scoring rate over time [Thomas (2007)].

Moreover, the game can often be divided into a number of discrete states that give additional information about the game. Hirotsu and Wright (2002) examine soccer as a continuous-time Markov process with 6 states: 2 teams can possess the ball on either half of the field, plus the state of having a goal scored in either net. Thomas (2006) considers a larger state space for hockey with a semi-Markov process instead. Only when a team has possession of the ball/puck in their opponent's territory can they score a goal, so that this underlying state will then directly influence the scoring rate for each team. This method can be applied if data on location and possession is available, but this is not currently available to the public.

We expect that players in the game will similarly affect the scoring rates for each team. The Cox process model [Cox (1972)] decomposes the rate of this process, described by the hazard function $h(t, X) = \lambda(t, X)$, into a time-varying component $\lambda_0(t)$ and a time-independent term for the inclusion of covariates $\lambda_x(X)$. Just as in the linear model case, these models can also be regularized, such as with the Lasso [Tibshirani (1997)].

2.4. *Source of data.* Records of many National Hockey League (NHL) games are available to varying levels of detail. For the sake of dividing the game into discrete intervals, we use the interpretation of Rosenbaum (2004) and Macdonald (2011) that an interval should end either when a player substitution is made by either team or when an event occurs (e.g., when a goal is scored). This level of detail is available with ease in game records from the 2007–2008 season until the 2011–2012 season. We select those shifts in which both teams are at full strength—each team has five skaters and one goaltender on the ice—and note the duration of the event in seconds. The outcome is one of three possibilities: the home team scores, the away team scores, or neither team scores and at least one player substitution occurs. As Table 1 shows, over 98% of the observations are nongoal outcomes, which is highly disproportionate compared to basketball.

For this analysis, we consider a process whose only events are goals scored by each team. We have additional information on shots on goal that did not result in goals, on penalties called that result in man-advantage situations, and on timeouts called (extremely rarely) by coaches. We do not include these at this stage

TABLE 1
A count of the events of each type in the database. A home team advantage is apparent

Seasons: 2007–2012	**Away goal**	**No goal (Changes)**	**Home goal**
Total events	10,935	1,301,799	11,981
Percent of total events	0.83	98.27	0.90

to keep the analysis on events that directly influence the final result of winning or losing the game, since shots on goal only lead to goals a fraction of the time, and the relationship between shots on goal and goals is not as simple as a fixed fraction of events. Any processes that lead to shots must also lead to goals, and to add additional competing processes to the model would add an additional level of complexity that is beyond the scope of this investigation. [See Macdonald (2012a) for how shots can be used in a standard regression scheme.]

For each season, we divide the games into two groups, uniformly at random—one for in-sample training (all observations from 80% of the games) and one for out-of-sample validation (the remaining 20% of games). When we perform any tuning parameter selection, we further subdivide the in-sample training set for cross-validation.

3. Model specification.

We model the stochastic nature of the game as a model of two competing processes for the scoring of a goal, censored by player substitutions. Each process has parameters for offensive and defensive characteristics, and these parameters are regularized by partial pooling. We use penalized maximum likelihood and full hierarchical Bayesian models to infer parameters of interest.

3.1. *Events obey a competing processes model.*

There are, at a minimum, two opposing processes in a hockey game: the home team tries to score on the away team, and vice versa. Both of these events are relatively rare compared to the number of observed event intervals, so that it is natural to model these as competing stochastic processes. Predictors that modulate these processes can be the teams in the game, the score of the game, the players on the ice or some other combination, and the same predictors appear in each process.

We choose a Cox proportional hazards model for each process, so that the hazard function has separate components for time dependence and predictors, as $h(X, t) = h_0(t)h_1(X)$, where X can represent various factors such as the players and/or team on the ice. For this investigation we begin with $h_0(t) = 1$, so that the time until the next event is exponentially distributed with rate $h_1(X)$; the baseline rate is established with a corresponding intercept term in X. More information on the location of the puck at each $t = 0$ may allow us to refine the time-based component in future investigations.

From this, each team's scoring rate is modeled as a log-linear Poisson process. The intercept terms, labeled r^h and r^a, represent the baseline scoring rates for the home and away teams, since, as we see in Table 1, the overall scoring rate for the home team is greater than for the away team; in this way, we explicitly detect a home-ice advantage. For each predictor indexed by p, let (ω_p, δ_p) be a measure of the offensive and defensive contribution for that predictor, so that a rating of zero corresponds to an "average" contribution; the corresponding indicators are X_p^h and X_p^a.

The scoring rates for each process are

$$\lambda^h = \exp\biggl(r^h + \sum_p (X_p^h \omega_p + X_p^a \delta_p)\biggr);$$

$$\lambda^a = \exp\biggl(r^a + \sum_p (X_p^a \omega_p + X_p^h \delta_p)\biggr)$$

for this combination. For each instance of this process, T^h and T^a are the times to each event for these processes, and let t be the first time at which any players on the ice are substituted, thereby censoring the scoring process. We assume that the (unmodeled) censoring time is independent of these event times and that, conditional on the predictors, these events are independent of each other. The outcome can then be registered as

$$Y = \begin{cases} 1, & \text{if } T^h < T^a, T^h < t, \\ -1, & \text{if } T^a < T^h, T^a < t, \\ 0, & \text{otherwise} \end{cases}$$

so that $(1, 0, -1)$ represents a home goal, no goal and away goal, respectively. Let $T = \min\{t, T_h, T_a\}$ be the observed time of the event.

Because of the independence condition, the likelihood for this event is then the product of the individual likelihoods, noting if either or each of the events was censored. With the survival function form $S(x) = P(T > x)$, we have

$$\begin{aligned} f(Y|\lambda_h, \lambda_a, T) = {} & f_h(T|\lambda_h)^{\mathbb{I}(Y=1)} S_h(T|\lambda_h)^{\mathbb{I}(Y\neq 1)} \\ & \times f_a(T|\lambda_a)^{\mathbb{I}(Y=-1)} S_a(T|\lambda_a)^{\mathbb{I}(Y\neq -1)}. \end{aligned}$$

Using this approach, each predictor's offensive parameter coefficient represents the change in the team goal-scoring rate with respect to a baseline rate (in particular, if they are replaced by another player of typical ability), and likewise for their defensive parameter and the opposing goal rate.

This method has several advantages for this class of data. Rather than trying to model a single outcome, such as goal differential, we can simultaneously calculate both the offensive and the defensive player ability parameters for each player, which are known to be distinct. The parameters we calculate have a meaningful interpretation in terms of game outcomes, since it reflects an increase or decrease in scoring rate. We can assess a player's marginal goal fraction over data in question by comparing the expected number of goals scored and allowed by their team given their ratings against the same data with ratings set to zero.

In addition to the offensive and defensive abilities of each player, we can account for several other possible influences. We can fit parameters to a whole team to capture their average ability, rather than simply including all the players independently. If we include both teams and players as predictors, this would change the interpretation of a "player effect" to be relative to the performance of one's team. We can also model an effect for the in-game score differential, since many teams may change their offensive and defensive strategies depending on how far ahead or behind they are in the game. This may best be accomplished by selecting a different intercept term depending on the score.

3.2. *Regularization of parameter estimates.* Even though we observe hundreds of thousands of discrete shift intervals in a season, the potential number of parameters in this model is also very large, and many of the player ability measures will be made with only a small number of observations, such as players who appear in only one game. Worse yet are those players who are not on the ice for any goal by one team and therefore have a maximum likelihood estimate of minus infinity for each of their parameters. To account for this, we use a hierarchical model to shrink parameter estimates toward a common mean (namely, zero), with the possibility that different positions (center, goaltender, winger and defenseman) have different shrinkage behavior. We have a number of choices for how to carry out this *regularization*: the choice of prior distribution or penalty term, the degree of hierarchical structure we impose, and whether we choose to minimize a function or integrate over a distribution.

The two standard choices for a prior/penalty distribution are the Gaussian and the Laplace, which penalize the mean squared error and absolute error, respectively; the latter case corresponds to the Lasso method [Tibshirani (1996, 1997)]. We can also consider a third class that joins the two, in the spirit of the Elastic Net method [Zou and Hastie (2005)], the Laplace–Gaussian distribution:

Prior type	**PDF**
Lasso/L1	$f(x \mid \lambda) = \frac{\lambda}{2} \exp(-\lambda \lvert x \rvert)$
Ridge/L2	$f(x \mid \sigma^2) = \exp(-x^2/(2\sigma^2))/\sqrt{2\pi\sigma^2}$
Elastic Net/L1 + L2	$f(x \mid \lambda, \sigma^2) = \frac{\exp(-\sigma^2\lambda^2/2 - \lambda \lvert x \rvert - x^2/(2*\sigma^2))}{\sqrt{8\pi\sigma^2}\Phi(-\sigma\lambda)}$

While each of these regularization options act to stabilize parameter estimates, both in cases with few observations and in those pairs or multiples with high collinearity, each family gives a different interpretation for the shrinkage behavior of the covariates. There are fully Bayesian interpretations for the Lasso which have been explored for standard linear models [Park and Casella (2008)], as well as those for the Elastic Net [Li and Lin (2010), Hans (2011)].

If we choose the L1 method and set each λ to a constant, then we have a (relatively standard) Lasso implementation, in which the penalized MLE or MAP estimates for the parameter may be exactly zero with nonzero probability, which yields a smaller subset of predictors for which the scoring rate change is distinguishable from zero. The L2 method with constant σ^2 terms yields a result in which the penalized MLE or MAP estimates for each parameter are brought closer but not exactly to zero. Compromising with the L1 + L2 method allows for some of the benefits of both properties, but may sacrifice the ease of implementation that can be found in the simpler cases. In the case of simple optimization, the L1 and L2 cases are suited to using cross-validation to choose the penalty weights λ and σ^2. If we are considering multiple partially pooled groups, cross-validation may no longer be computationally feasible, since searching the space of possible parameters becomes more difficult the more dimensions we add.

3.3. *Implementation.* We have two types of problems that we consider: those in which the total distribution of predictors and their group-level variance terms is of direct interest, and those in which we are only interested in selecting a subset of relevant predictors. The former case requires simultaneous estimation of a number of shrinkage parameters, and this dimensionality makes a search of the space difficult to accomplish with cross-validated methods, so we use the full hierarchical Bayesian approach. In the latter case, there is typically only one dimension of interest, as we wish to select from only one relevant subset of predictors, and so here we can use penalized maximum likelihood estimation much more easily.

3.3.1. *Optimization of penalized likelihood.* We use maximization of a penalized likelihood to get rough parameter estimates, with modest levels of L1 and/or L2 shrinkage to handle parameters with minimal information in the data, such as players who played in only one game. We can use this as a starting point for Markov Chain Monte Carlo to obtain estimates for the pooled variance/shrinkage parameters. For each MCMC routine, we discard a sufficient number of initial samples as burn-in and thin the chain sufficiently so that the thinned chain has negligible autocorrelation for all parameters and a sufficient number of uncorrelated samples (in each of our cases, a minimum of 500) for use in inference.

We can also simply scan through a series of values for each shrinkage parameter, selecting the optimal value through out-of-sample validation. This is easiest when there is only one shrinkage parameter to estimate.

3.3.2. *Full posterior estimation with MCMC.* The full hierarchical model has three levels: from the data, to the predictor coefficients, and finally to their partial pooling prior distributions. We use a Gibbs sampler blocked on pairs of variables to estimate model parameters.

- *Level* 1: Each outcome $(Y|X^h, X^a, \omega, \delta, t)_i$ is distributed as the competing process model. Each predictor block (X_i^h, X_i^a) is stored as a sparse vector, given that there are typically no more than 16 total nonzero terms in each row.

- *Level* 2: Each coefficient pair $(\omega, \delta)_p$ is distributed according to its prior distribution. In the Laplace–Gaussian case, this has four terms corresponding to the group $g(p)$ that has predictor p as a member: the Laplace terms $(\lambda_{\omega,g}, \lambda_{\delta,g})$ and the Gaussian terms $(\sigma_{\omega,g}^2, \sigma_{\delta,g}^2)$.

As the intercept terms r^h and r^a effectively correspond to their own (ω, δ) pair and belong to their own group, each acts as their own group mean; weak hyperpriors on their own prior terms act marginally as weak prior distributions.

Each pair (ω_p, δ_p) is updated using a Metropolis sampler with a bivariate Gaussian proposal distribution. Indexing each observed shift with i, the target distribution

$$f(\omega_p, \delta_p | Y, X, \sigma_{\omega,g(p)}, \sigma_{\delta,g(p)}, \lambda_{\omega,g(p)}, \lambda_{\delta,g(p)})$$

equals the product

$$f(\omega_p, \delta_p | \sigma_{\omega,g(p)}, \sigma_{\delta,g(p)}, \lambda_{\omega,g(p)}, \lambda_{\delta,g(p)}) \prod_{i:p\in(X_i^h, X_i^a)} f(Y|X^h, X^a, \omega, \delta, t)_i.$$

We initialize all (ω, δ) terms with a penalized maximum likelihood estimate using relatively loose shrinkage parameters.

- *Level* 3: Each Laplace λ term has a weak Gamma conjugate prior; each Gaussian σ^2 term has a weak Inverse Gamma conjugate prior. If the Laplace–Gaussian is used, these priors are no longer conjugate to their respective parameter forms.

  Each pair $(\lambda_{\omega,g}, \sigma^2_{\omega,g})$ is updated through a pair of univariate grid approximation samplers. The first samples according to the density along the sum of approximate total shrinkage, $1/\sigma_{\omega,g} + \lambda_{\omega,g}/\sqrt{2}$, while keeping the relative fraction of shrinkage $\frac{\lambda_{\omega,g}/\sqrt{2}}{\lambda_{\omega,g}/\sqrt{2}+1/\sigma_{\omega,g}}$ constant;[1] after updating these values, the second samples the relative fraction while keeping the approximate total constant. This is repeated for each pair $(\lambda_{\delta,g}, \sigma^2_{\delta,g})$. (One can always sample directly from the bivariate grid approximation as well, though this is less computationally efficient.)

We constructed the sampler using the `R` programming language with supporting back-end code in `C++`. Execution time varies with the total number of covariates, with the simplest cases (200,000 outcomes and 60 covariates) taking 30 processor-minutes, to the more complicated runs (200,000 outcomes and 2600 covariates) requiring roughly 60 processor-hours. We used multiple parallel chains with sufficient burn-in periods to collect a sufficient number of uncorrelated samples. We validated the sampler using the method of posterior quantiles [Cook, Gelman and Rubin (2006)].

In each of these cases, we can judge the performance of each selected model initially using in-sample measures, then confirming goodness of fit by checking against our held-out data. For MCMC, we use the Deviance Information Criterion, calculated using the individual samples and the average overall samples, applied to the likelihood of the original (fitted) data for in-sample fit, as well as to our withheld data for out-of-sample validation.

[1] The $\sqrt{2}$ factor is added to reflect the fact that a Laplace distribution with scale 1 has a variance of 2.

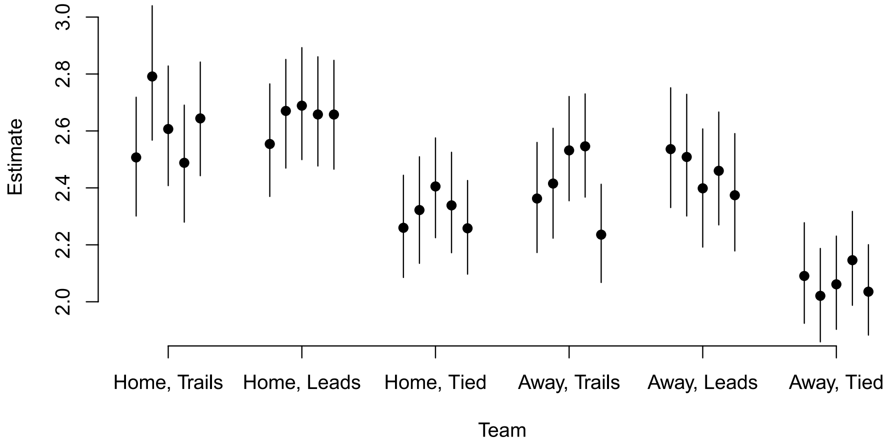

FIG. 1. *The scoring rates per* 60 *minutes for generic home and away teams in each individual season, divided by game score. Points are posterior means, lines are central* 95% *credible intervals. The home team consistently outscores the away team in all five seasons and overall within each game situation; scoring rates for both sides are elevated when the game is not tied, whichever team is winning.*

4. Analysis of full posterior distribution. Since all analyses in this investigation are conducted on events where both teams are at full strength, we refer to any particular coefficient pair (ω_p, δ_p) as the Mean Even Strength Hazard (MESH) rating for the corresponding predictor, such as the team (as in Section 4.2), a particular player (Section 4.3) or the extra contribution of a pair of players (Section 5.2). We estimate the net MESH rating as offensive ability minus defensive liability, $\omega_p - \delta_p$.

4.1. *Home-ice advantage and game score.* The simplest version of this process model has only two coefficients, the intercepts for the home team and away team processes:

$$\lambda^h = \exp(r^h), \qquad \lambda^a = \exp(r^a).$$

We can extend this by specifying different intercepts for different game score situations. For this analysis we choose three: when the home team is winning, tied or trailing.[2] Figure 1 shows the estimates of these intercepts in each of the five seasons under consideration, for home and away teams, by taking each $\exp(r^h)$ and $\exp(r^a)$, the per-second rates, and multiplying up to a full (hypothetical) 60-minute game.

It is clear that the home team has a consistent advantage. Whether or not the effective home scoring rate is actually identical in each of the five seasons, they

[2] While we can extend this more generally to all combinations of game score, the results of this division are quite robust.

are so close as to be indistinguishable from each other; this is similar for the away scoring rate. The year-to-year variability in home and away mean rates is consistent with a common goal-scoring rate across all five seasons; simulations verify that the change in estimated means is consistent with the spread in estimation based on the generation of a season's worth (1230 games) of goals for each team from the Poisson model.

It is also clear that there is a change in scoring rates by game score. Interestingly, the scoring rates for each team are raised by the same amount when a team is leading or trailing, compared to when the score is tied. This suggests that teams are more cautious during tie scores, and that efforts by the trailing team to increase their own scoring rate, or by the leading team to increase their margin, result in a corresponding and roughly equal increase in their opponents' rate.

4.2. *Overall team performance, per season.* Because each of the 30 teams in the data is present in roughly one-fifteenth of the total events, we do not expect the degree of sparsity as when we model the impact of individual players. This does not mean, however, that the model cannot benefit from partial pooling on team parameters, both to reduce the effective dimensionality of the model and to improve predictive accuracy. This model is then specified as

$$\lambda^h = \exp(r^h + \omega_{\text{home}} + \delta_{\text{away}}); \qquad \lambda^a = \exp(r^a + \omega_{\text{away}} + \delta_{\text{home}}),$$

with partial pooling under one of our chosen schemes; in general, this is of the form

$$\omega_{\text{team}} \sim \text{Laplace–Gaussian}(\lambda_{\text{team}}, \sigma^2_{\text{team}}),$$

where the shrinkage behavior depends on the prior specification for the parameters $(\lambda_{\text{team}}, \sigma^2_{\text{team}})$. We include three sets of intercepts for game score situation, though additional analysis shows that our parameter estimates are insensitive to this model choice.

We estimate these parameters within each season using MCMC for each of the three submodels for pooling. For each shrinkage mode, two variance components are estimated, for total offensive and defensive ability, respectively. For the Laplace–Gaussian prior form, there are four total parameters, rather than two, and this mode has the lowest Deviance Information Criterion for all five seasons, both in- and out-sample, as shown in Table 2. From this point on, we focus on results using only the full Laplace–Gaussian prior.

Figure 2 shows the posterior distributions for each team's net MESH rating within each season, using the Laplace–Gaussian prior. As expected, these track well with the number of goals scored and allowed by each team during these seasons, since the correlation of parameters across teams is minimal; teams play each other no more than eight times per season out of a total of 82 games. There are also several significant deviations for some teams for one season compared to the

TABLE 2

DIC for the Laplace, Gaussian and Laplace–Gaussian pooling priors for the model with teams as explanatory variables. The Laplace–Gaussian performs the best in each season and overall in both the in-sample and out-of-sample cases

	Insample			Outsample		
Season	**L1**	**L2**	**L1 + L2**	**L1**	**L2**	**L1 + L2**
2007–2008	57701	57703	**57691**	14088	14088	**14085**
2008–2009	59996	59966	**59962**	15071	15064	**15064**
2009–2010	61415	61359	**61348**	15710	15704	**15702**
2010–2011	62552	62521	**62515**	15541	15538	**15537**
2011–2012	62398	62398	**62377**	15983	15983	**15982**

rest, such as St. Louis in 2012 (very positive) and Minnesota in 2012 (very negative), that are not statistically distinguishable from their other performances but still illuminating nonetheless.

It is worth noting that very few of these parameters have 95% credible intervals that do not contain zero, suggesting that the amount of information to distinguish a team from being truly "average" is quite small; however, it is clear that some teams are highly probable to be better (or worse) than other teams in the league, so that distinguishing a statistically significant ordering is within the reach of this model. In particular, the overall rating for the Boston Bruins over these five seasons is

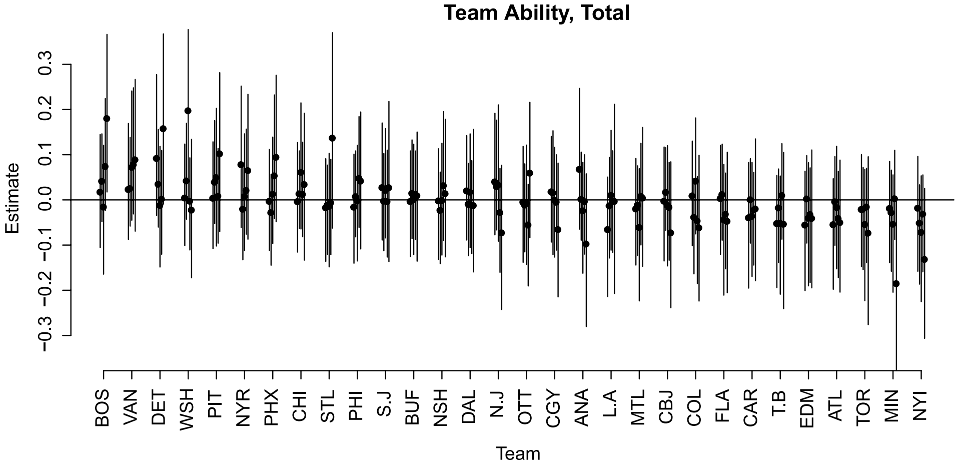

FIG. 2. *Total team ability estimates for each team in the NHL, grouped by team for each season; order is by overall team rating. Points are posterior means, lines are central* 95% *credible intervals. A rating of* 0.1 *corresponds to a differential of roughly* 0.3 *goals per game scored or prevented. Note that only two team-years, the* 2012 *Boston Bruins and* 2010 *Washington Capitals, have effects that are significantly different than average.*

greater than 12 of their 29 opponents with at least 95% probability; the rating for the Detroit Red Wings is greater than 11 of theirs with 95% probability or greater. Similarly, the New York Islanders and Minnesota Wild are worse than 12 and 10 of their opponents, respectively.

Given the limited amount of information, the question arises as to whether we can reliably distinguish player abilities from average when the information on a single player is much less than that for a single team. We address this question in the following section.

4.3. *Distribution of player abilities, across all seasons*. The estimation procedure for team effects is relatively straightforward, given the relative balance of the design matrix. Once we consider individual players, more questions arise since the design matrix can be far more unbalanced; for example, a player's defensive rating may be trickier to estimate because they share the majority of their shifts with a single goaltender. Arguably, it gets worse if both players are *great* players, since they may both be retained by a single team for much of their careers.

This is made easier when dealing with data from multiple seasons, as the more players change teams, the more the players in the league will mix. We therefore model player abilities as constant over all five seasons, which we refer to as the "grand model," specified with the following terms:

- Overall home and away effects with score differential effects.
- Offensive and defensive parameters for all skaters (centers, wingers and defensemen).
- Defensive parameters only for goaltenders.
- Laplace–Gaussian pooling for each type of ability and each position parameter (center, left wing, right wing, defenseman, goaltender).

We do not include team effects at this stage specifically because we are trying to compare players across teams, and their collinearity with goaltenders is needlessly complicating. We are still resigned to the degree of confounding in defensive estimates, since the goaltender not only plays a large role, but also is not typically replaced throughout the game, most often only relieved during a poor outing. We use the standard MCMC implementation to estimate parameters.

There is only a small subset of players whose ratings can be considered statistically significant. Of 1592 total players over five seasons, 37 have player ratings whose effective total (offensive skill minus defensive liability) have 95% central credible intervals that do not contain zero. Of these, 36 are positive; only one player, Stephane Veilleux, had a negative total rating with statistical significance, suggesting that he is a good enough player to log regular ice time with a major league team, but not so good that his contributions in even strength are less than the league average. (This is not necessarily the same as a "replacement"-level player.) The top five players at each position group are given in Table 3.

TABLE 3

Top five players at each position, by overall rating, over five NHL seasons (2007–2012). *Listed are mean ratings*, 95% *credible intervals and posterior probabilities that the player is the best at his position*

Player	**Total MESH rating**	**95% credible interval**	**% probability best player**
Center			
Pavel Datsyuk	0.463	(0.262, 0.668)	39.5
Sidney Crosby	0.388	(0.155, 0.598)	18.1
Henrik Sedin	0.355	(0.096, 0.606)	13.3
Patrice Bergeron	0.280	(0.075, 0.535)	8.7
Evgeni Malkin	0.266	(0.048, 0.429)	4.5
Winger			
Alexander Semin	0.321	(0.167, 0.459)	3.9
Alex Ovechkin	0.318	(0.160, 0.478)	6.6
Marian Gaborik	0.308	(0.128, 0.478)	7.6
Loui Eriksson	0.258	(0.097, 0.407)	6.0
Alexander Radulov	0.249	(0.003, 0.490)	5.5
Defense			
Zdeno Chara	0.077	(−0.015, 0.244)	11.7
Mark Streit	0.0427	(−0.038, 0.207)	6.1
Jaroslav Spacek	0.0373	(−0.033, 0.163)	4.5
Mike Green	0.036	(−0.031, 0.185)	2.6
Matt Carle	0.034	(−0.026, 0.161)	3.2
Goaltender			
Henrik Lundqvist	0.186	(0.076, 0.292)	36.0
Tim Thomas	0.120	(0.005, 0.233)	20.6
Jonathan Quick	0.102	(−0.012, 0.221)	14.2
Martin Brodeur	0.101	(−0.009, 0.209)	7.0
Roberto Luongo	0.100	(−0.010, 0.211)	5.3

4.3.1. *Overall variability of rating by position.* Figure 3 shows the variability of player abilities at each position according to their respective Laplace–Gaussian distributions. The first graph shows us an approximate proportion of the fraction of variability best explained by the Laplace term, as an indicator of the degree to which a distribution of players has heavier tails; the higher this is, the higher the number of "extreme" players. The second graph shows the total variability of player abilities as the standard deviation of player estimates at each iteration of the MCMC.

Several matters are apparent. There is considerable variability in offensive ability for forwards (centers and wingers) but far less for defensemen. This is consistent with the notion that defensemen have less impact on offensive output during even-strength situations.

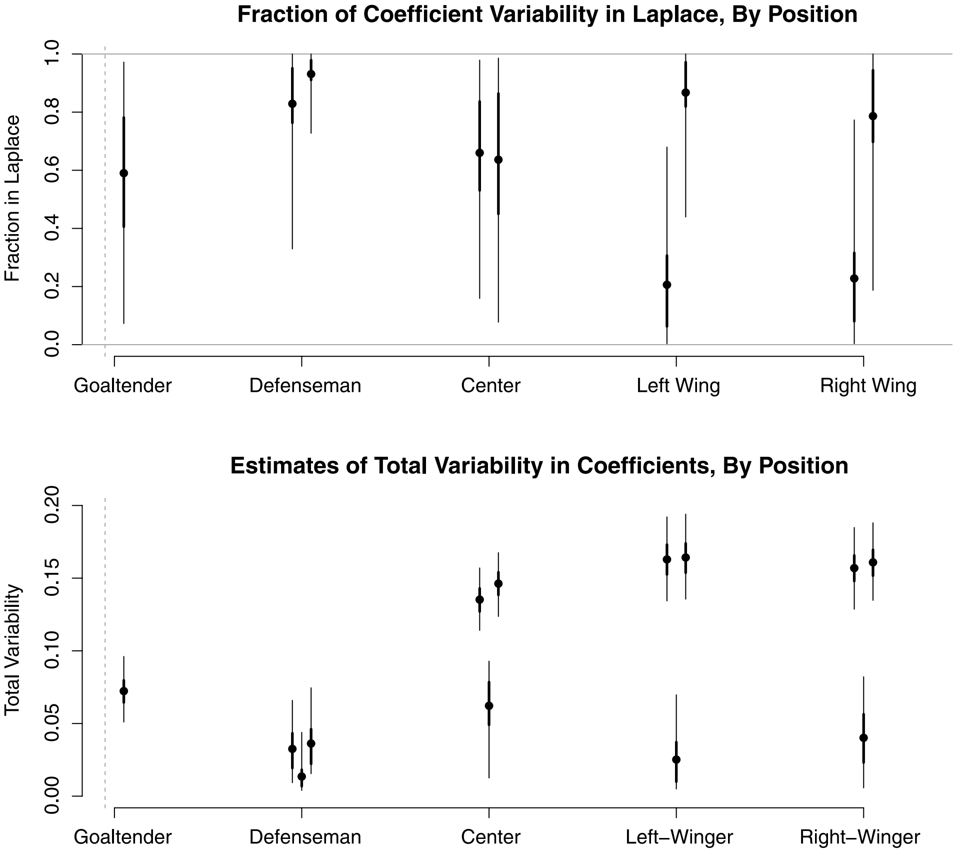

FIG. 3. *Variability properties of coefficient estimates by position. Thick and thin lines represent 50% and 95% credible intervals*; *points represent estimated means*. Top, *the approximate fraction of the variability that can be attributed to the Laplace component of the Laplace–Gaussian error distribution, for each position and for offense and defense. For most positions there is a strong tendency toward the Laplacian distribution*, *with heavier tails and more outliers*; *this is less pronounced for winger offense*. Bottom, *the offensive*, *defensive and total variability by position. Goaltenders have only defensive variability, which is considerably more variable than defense for any skating position. Offensive players* (*centers and wingers*) *have more variability in offense*, *and every skating position has minimal variability in defence*.

For all positions other than goaltender, defensive variability is far smaller than it is for offense. Two explanations are immediate. First, it may be that the collinearity between skaters and goaltenders is causing our estimates of goaltender ability to be more variable than they are in reality, and less variable for the skaters. Second, since the total defensive burden is shared by six players (five skaters plus one goaltender) rather than the five for offense, and the bulk of defensive skill is taken up by the goaltender, the total amount of "defensive skill" available to be shared by skaters is considerably smaller, and therefore there is less total variability between players.

How valuable is an individual position to a team? A typical starting goaltender plays about 60 full games a season for their team, while first-line offensive and defensive players will have the equivalent of roughly 30 and 35 full games, respectively. On average, a good goaltender is worth roughly what a good offensive player is to a team's total output with respect to "average" players, while a good defensive player appears to be worth considerably less.

The center position has, on the whole, more effect on defensive performance than a defenseman does, and wingers seem to have roughly equal defensive variability as the defense position has total variability. This would seem to confirm the case that when forwards have control of the puck, particularly in their offensive zone, they deny the likelihood of their opponents being able to score. As we show soon, this does not mean that a player with a high ω rating must therefore have a high δ rating.

From these overall results, we move on to describe the individual performances of players over the five-season period, as organized by position. Table 3 lists the top five players in each position group under the grand model; we provide a more complete list of players at each position in the supporting material, including several of the worst players at each position.

4.3.2. *Assessing model fit.* We compare the models with basic intercepts, team parameters and player parameters by calculating the likelihood of points withheld from the original model fit, in each of the five seasons and altogether, with respect to the posterior mean for each parameter. As we show in Table 4, the likelihood is highest in all five seasons for the player parameter model, even with a much higher number of parameters.

The adequacy of the fit of the model to data is harder to assess. The process is inherently noisy—the number of goals in a game for any team varies wildly—

TABLE 4

Comparing the out-of-sample doubled negative log likelihood for the models for game score only, team parameters and player parameters, respectively. Even with a large number of extra parameters, the model with player effects yields considerably better fit for the withheld data than the alternatives

Group	**Score**	**Team**	**Player**
2007–2008	14096.1	14088.7	**14002.5**
2008–2009	15077.2	15072.1	**15012.1**
2009–2010	15704.7	15702.1	**15643.1**
2010–2011	15543.7	15542.1	**15488.5**
2011–2012	15994.8	15992.3	**15940.1**
Total	76416	76397	**76087**

and so our ability to predict the behavior of any one game is minimally improved when adding player parameters. To check the adequacy of our estimates for player parameters, we simulate data for each game in the withheld set using the posterior mean and check the sum of the goals scored by the home and away teams in each simulated season against the truth; we find that the true data lies within the 95% simulated confidence interval each time, and with every model (score alone, teams and players, resp.).

4.3.3. *Players that make the greatest total difference.* Since the ratings represent multipliers to the default scoring rate, we can quickly estimate the total contribution of a player over the observation period as the difference in expected goals, scored and allowed by any average team, relative to an average player,

$$G_{\text{net}} = \bigl[\bigl(\exp(r_{\text{base}} + \omega_p) - \exp(r_{\text{base}})\bigr) - \bigl(\exp(r_{\text{base}} - \delta_p) - \exp(r_{\text{base}})\bigr)\bigr] \times T_{\text{total},p}.$$

A mean intercept parameter $r_{\text{base}} = -7.3$ corresponds to roughly 2.4 goals per 60 minutes. Table 5 lists the top 25 total goal producers and preventers over the five season period. Five goaltenders make the top 25 list; despite the fact that defensemen typically log more ice time than forwards, only one defencemen make the top 25. If necessary, we can adjust these ratings to reflect teammates and opponents by using the expected goals in each shift given all other player ratings, to handle nonlinearity in the rate relationship.

G_{net} compares directly to plus/minus in terms of a player's overall impact. We plot all player's total G_{net} against Plus-Minus in Figure 4.

5. Applications with variable selection. Many problems of interest have to do with selecting a relevant subset of predictors from a much larger set. There are several such examples we can carry out with our method that we present here. These methods tend to be considerably faster than operations with Markov Chain Monte Carlo, since we're more concerned with the selection of a subset than in the evaluation of its stochastic properties. A negative consequence of this is that this estimation approach is nonregular, making assessment of uncertainty difficult [Dawid (1994)]. Our primary purpose here is identification rather than quantification (which is handled well by the full hierarchical Bayesian treatment) and our numerical estimates are presented so that we can compare their magnitudes with effects from the full model.

5.1. *"Most valuable player" awards, per team, per season.* The term Most Valuable Player has many interpretations throughout the sports world. One that appeals to us is the notion that a player is most valuable to their team if their team's performance suffers the most compared to a "replacement" player in their stead. In the context of this model, we propose that each player should be judged

TABLE 5

The top 25 *even-strength players in the NHL over* 5 *seasons* (2007–2012) *according to the net number of goals scored or prevented* G_{net}, *assuming a baseline scoring rate of roughly* 2.4 *goals per team per* 60 *minutes*. *At position* 20, *Zdeno Chara is the highest-ranked defenseman in this time period*

Rank	**Player**	**Pos**	**Time (s)**	**+Scored**	**+Scored**	G_{net}	**% Pr(Best)**
1	Henrik Lundqvist	G	928100	0	139.5	139.5	39
2	Pavel Datsyuk	C	320200	109.7	18.74	138	19
3	Henrik Sedin	C	350300	114.5	2.471	118.2	18
4	Joe Thornton	C	360000	62.63	8.347	73.13	2
5	Sidney Crosby	C	240400	98.85	−17.74	70.3	2
6	Tim Thomas	G	732300	0	70.18	70.18	4
7	Alexander Semin	L	271100	68.57	−1.287	66.8	<1
8	Loui Eriksson	L	330900	66.54	−0.1943	66.29	3
9	Alex Ovechkin	L	373500	66.96	−1.125	65.53	<1
10	Evgeni Malkin	C	309400	86.38	−15.62	64.31	<1
11	Marian Gaborik	R	276700	64.16	−0.05811	64.08	<1
12	Martin Brodeur	G	814100	0	60.95	60.95	<1
13	Patrice Bergeron	C	239400	51.14	4.423	56.97	<1
14	Jonathan Toews	C	306800	56.69	0.1341	56.86	<1
15	Ilya Kovalchuk	L	376500	74.55	−14.78	55.44	<1
16	Jarome Iginla	R	393100	71.55	−13.66	54.21	<1
17	Jason Spezza	C	318400	73.31	−14.84	53.4	<1
18	Stephen Weiss	C	327900	58.86	−4.605	53.03	<1
19	Roberto Luongo	G	799600	0	52.55	52.55	<1
20	Zdeno Chara	D	436700	40.35	8.899	50.47	<1
21	Patrick Sharp	R	300200	51.78	−1.818	49.5	<1
22	Thomas Vanek	L	290400	51.92	−3.624	47.33	<1
23	Henrik Zetterberg	L	339700	47.79	−1.238	46.29	<1
24	Alex Tanguay	L	278000	46.85	−0.6684	46.02	<1
25	Jonathan Quick	G	661600	0	45.5	45.5	<1

with respect to the rest of their team. Since selecting an exceptional player can be treated as a special case of variable selection, we propose a scheme to pick exceptional players on each team.

We use a model with teams and individual players as predictors. (We omit goaltenders for this ranking due to the confounding with team ratings.) We fix the estimates for team ability and the grand means to be those obtained in Section 4.2. This is to ensure that all subsequent player ratings obtained will roughly sum to zero, since all ratings are relative to their team rating for each of offense and defense.

We use a single shrinkage penalty for player ratings. Here we choose a single Lasso penalty of $\lambda = 8$, as it produces the highest likelihood for the out-of-sample data in three of five seasons; in the other two, the optimal penalty was such that

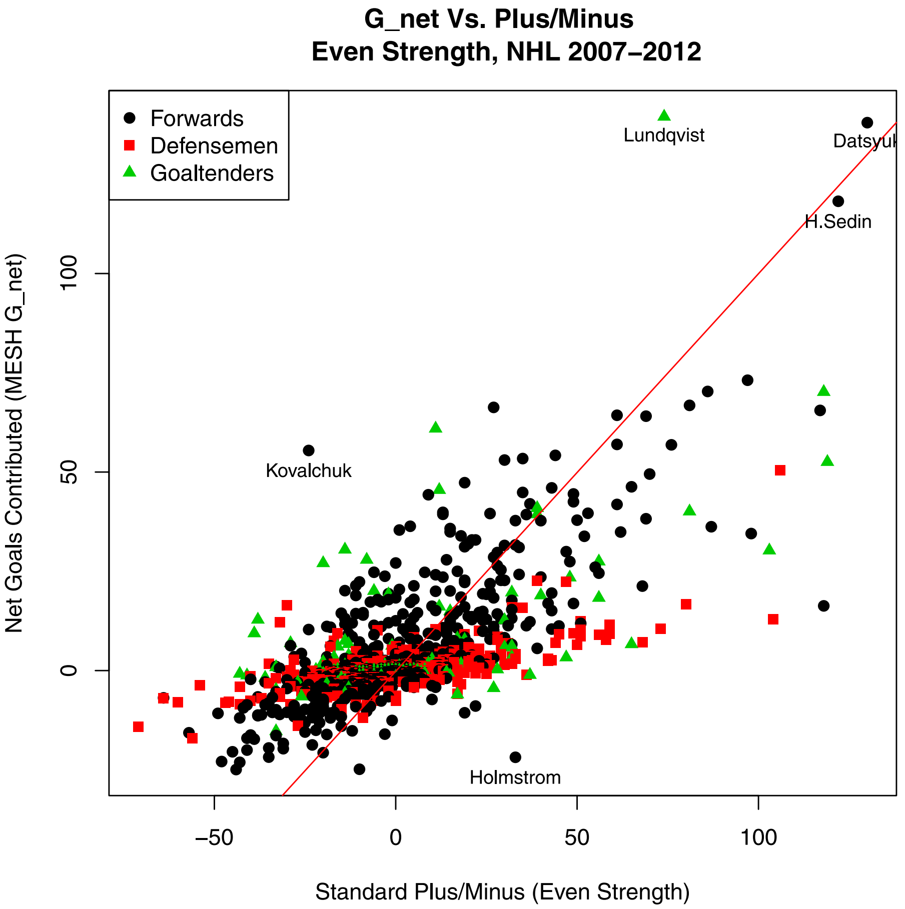

FIG. 4. *Comparing estimated mean G_{net} against measured plus/minus for all players. Many player ratings are shrunk toward zero, particularly players with lower ice time, but also defensemen (red squares); goaltender ratings tend to be unaffected (green triangles). In the upper-right corner, Pavel Datsyuk and Henrik Sedin (players ranked 2 and 3 in G_{net}) are comparatively unshrunk; goaltender Henrik Lundqvist has a considerable boost compared to his observed plus/minus when accounting for his teammates and his opposition. Also highlighted are forward Ilya Kovalchuk, whose ratings difference is the largest of all players, and Tomas Holmstrom, whose negative G_{net} contrasts with his modestly positive plus/minus.*

no player had a nonzero relative rating. In each case, the fit to out-of-sample data was virtually identical for penalties greater than 5. For each team, we select players with the highest and lowest offensive, defensive and overall ratings, and place them in the appropriate MVP and LVP tables. When there are empty cells in the table, we steadily decrease the penalty, filling in empty cells in the MVP and LVP table

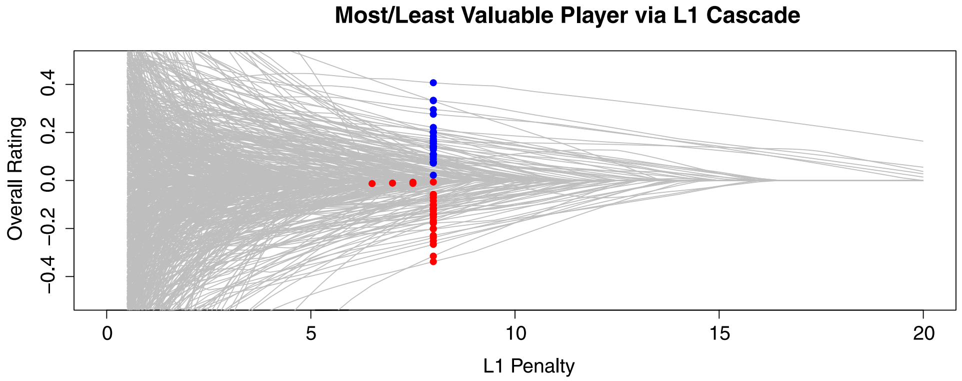

FIG. 5. *The Lasso Cascade method for picking team Most/Least Valuable Players for the* 2011–2012 *season. Team-level effects are fixed, and player effects are subjected to a steadily decreasing penalty beginning with* $\lambda = 8$ *as chosen by out-of-sample validation. Points indicate where MVPs* (*in blue*) *and LVPs* (*in red*) *are first declared for overall ability.*

as new players emerge, and stopping when all cells in the table are filled. (This occurs in the case of 2 to 5 teams.)[3]

Figure 5 shows a demonstration of the method for the 2011–2012 season, and Table 6 lists the top 10 MVPs and bottom 10 LVPs for that year; a full list of named MVPs and LVPs, for offense, defense and overall, is provided in the supplementary material. Most of the results are consistent with expectations, though we can spot

TABLE 6
The top 10 *MVPs and bottom* 10 *LVPs for the* 2011–2012 *season, calculated as the rating of a player relative to their team's average and selected by the Lasso method*

Team	**MVP**	**Rel. rating**	**Team**	**LVP**	**Rel. rating**
EDM	Jordan Eberle	0.407	N.J	Ryan Carter	−0.338
T.B	Steven Stamkos	0.334	NYI	Nino Niederreiter	−0.315
PIT	Sidney Crosby	0.332	DET	Tomas Holmstrom	−0.266
NYI	John Tavares	0.295	BOS	Shawn Thornton	−0.252
FLA	Stephen Weiss	0.276	CHI	Michael Frolik	−0.238
PHX	Adrian Aucoin	0.221	MTL	Alexei Emelin	−0.229
OTT	Marcus Foligno	0.203	T.B	Dominic Moore	−0.202
WSH	Alexander Semin	0.200	WSH	Michael Knuble	−0.200
STL	David Perron	0.200	BUF	Robyn Regehr	−0.178
DAL	Jamie Benn	0.184	CGY	Tim Jackman	−0.173

[3] We can perform a full Bayesian analysis with this model and these selected shrinkage parameters to determine uncertainties on each player selected. We demonstrate the Lasso method here due to its relative speed in identifying the correct shrinkage factor for this variable selection.

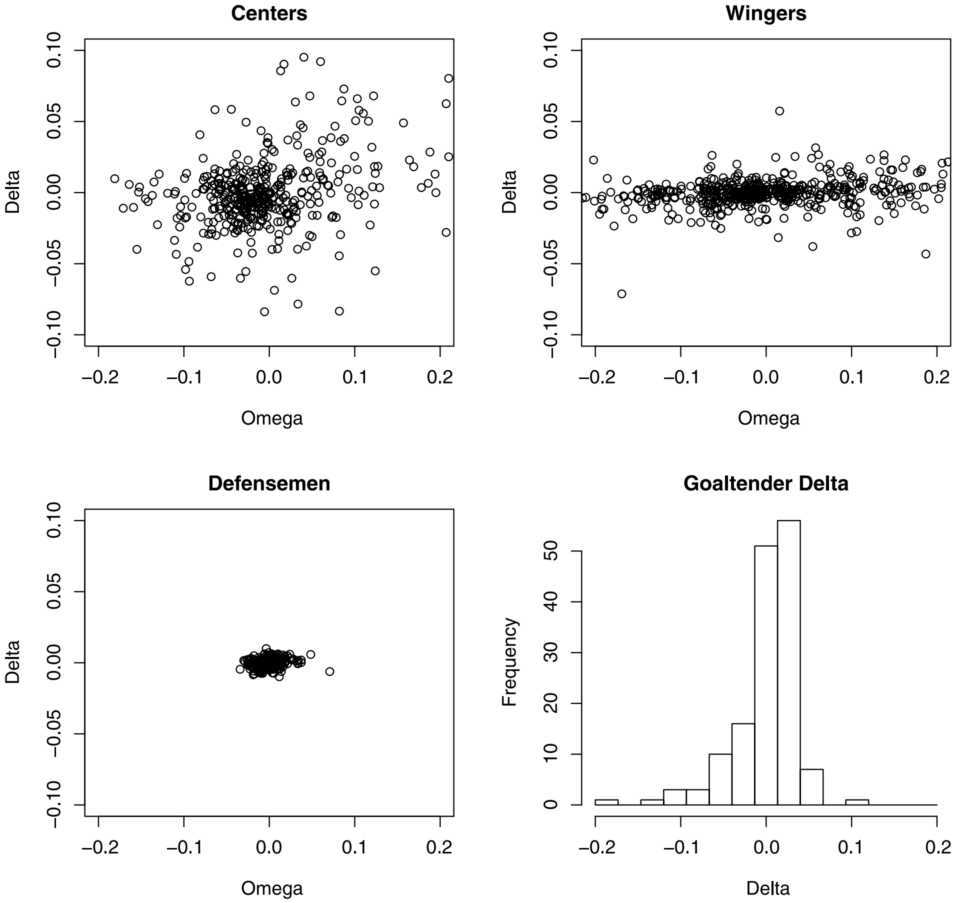

FIG. 6. *Scatterplots of player ability estimates by position. There is little if any correlation between a player's estimates of offensive and defensive ability.*

some interesting trends. First, quite often, the most valuable player for offense will be the least valuable player for defense, such as Joffrey Lupul with the 2011–2012 Maple Leafs, or vice versa. In many ways this is not surprising; since the best players have the most ice time, they would be more likely to have ratings that are not shrunk completely to zero on that basis alone, and because these ratings tend to not be correlated (see Figure 6 for ratings in the five-season grand model), it is not surprising that this rating will sometimes be negative.

Second, some of the more surprising Least Valuable Players are centers who specialize in taking face-offs, often at critical times, such as David Steckel of the Washington Capitals in 2009–2010 and again with the Toronto Maple Leafs in 2011–2012. These players are often brought into the game specifically to take face-offs, often in their team's defensive zone, before switching off for another player at their next opportunity. Because they are given fewer opportunities to score goals, merely to help prevent them, their offensive ratings will suffer accordingly; their defensive ratings can be insignificant by comparison. Taking puck location into account has been the subject of previous research [Thomas (2006)] and its role in this model will be the subject of a future investigation.

5.2. *Identifying exceptional player pair interactions.* If we can select a smaller subset of predictors from a much larger collection, we allow for the possibility of including a substantially large number of extra predictors to any of our models. One compelling inclusion is player interactions; in this context, this would allow us to see whether two players have an additional, detectable "chemistry" that yields a higher or lower total in their offensive or defensive abilities. If this is the case, we must see whether there are any corresponding changes to the individual player abilities as well.

Since the MCMC procedure gets considerably slower with the addition of a large number of predictors and coefficients, we use the Lasso method of penalized maximum likelihood to detect a number of nonzero coefficients for the new group.

We begin by specifying the grand model in Section 4.3, and we use the mean value of each σ_g and λ_g as Laplace–Gaussian penalty terms that we will keep fixed for the individual player effects, to allow for and moderate adjustments due to the pair terms.

We then select a subset of player pairs from the database. For this analysis, we took the top 1000 pairs of players in terms of the number of shifts they played together over the five-year period. We use the condition that both players played forward positions or both players played defense, since these groups tend to coordinate their play among themselves. We add these pairs as predictors to the model. We then estimate the model parameters for a series of Lasso penalty values, labeled λ_{pair}, on the player-pair terms, in order from strictest to loosest for computational ease.[4] The choice of penalty term depends on the goal in question; if the goal is to increase predictive accuracy, a penalty term that minimizes out-of-sample error is appropriate.[5]

In this case, we find that the penalty $\lambda_{\text{pair}} = 8.5$ minimizes the test-set likelihood under cross-validation for these events. Of the 2000 possible parameters to select from (1000 each for ω and δ), this routine selects 247 nonzero parameters for player pairs for 221 unique player pairs.

Table 7 shows the top and bottom five player pair ratings from the analysis; a more complete list is available in the supporting material. Of particular note is the most extreme case, the pairing of Ilya Kovalchuk and Todd White, whose mutual rating is so low that they effectively wiped out their positive total individual

[4] We maintain the previously obtained penalty values for player effects.

[5] If the goal is to select a fixed number of significant partnerships, we would choose the penalty term that yields that count.

TABLE 7

The top and bottom five player-pair interactions over 5 NHL seasons. These effects represent the additional total rate beyond the abilities of the players themselves

Rank	**Player 1**		**Player 2**		**Team**	**Time (s)**	**Rating**
1	Brad Boyes	R	Jay McClement	C	STL	35466	0.393
2	Matt Carle	D	Andrej Meszaros	D	PHI	41011	0.314
3	Patrice Bergeron	C	Brad Marchand	C	BOS	85678	0.31
4	Jussi Jokinen	L	Jeff Skinner	C	CAR	46196	0.287
5	Kris Letang	D	Paul Martin	D	PIT	40034	0.275
217	Zach Bogosian	D	John Oduya	D	WPG	57215	−0.235
218	David Booth	L	Michael Santorelli	C	FLA	34158	−0.241
219	Alex Frolov	L	Anze Kopitar	C	LA	45982	−0.269
220	Sidney Crosby	C	Evgeni Malkin	C	PIT	69217	−0.283
221	Ilya Kovalchuk	L	Todd White	C	ATL	70421	−0.545

ratings during their time together. Both recorded very high-scoring seasons when they played together, but this accolade effectively masks their mutual liability on defense. The next-lowest pair of Sidney Crosby and Evgeni Malkin is similar; their presence together does not increase their (considerable) offensive prowess beyond their individual levels, but does lead to a substantial increase in the rate of goals scored against their team while they are both on the ice.

Interestingly, the pair of Henrik and Daniel Sedin, twin brothers who play most of their even-strength shifts together, does not appear in the selected group. Indeed, the most total ice time in the top/bottom five is the 135th-most coincident pair of Patrice Bergeron and Brad Marchand from Boston. This suggests that the levels of shrinkage are appropriate for obtaining a reasonable subset of player pairs that have reasonable deviations.

As a final check, the positions of players in the grand rating table are mostly unchanged, so that the original player ratings are reasonably robust to these new additions. Worth noting is that the top two positions in the grand ratings reverse; Sidney Crosby now has the highest player rating over Pavel Datsyuk, due to the removal of the poorer outcomes when he plays with Evgeni Malkin, as opposed to other potential linemates.

6. Discussion and extensions

We have presented a model-based method for assessing player ability in ice hockey by treating the game as a competing stochastic process. Given the sheer number of predictors, and the relatively weak explanatory power of each, we use shrinkage methods to improve our estimation of model parameters. We also allow for the possibility of expanding the model specification from a simple flat hazard model to a more general Cox proportional hazards semi-Markov process, to account for other phenomena. In terms of comparisons between players, our method produces similar results for player effects as other approaches [Macdonald (2011), Gramacy, Jensen and Taddy (2013)], suggesting that there is sufficient information in the data to distinguish player ability at a grand level, despite different models. Our method has a key advantage in that it has a specific mechanism for generating hypothetical games, as long as a mechanism for player substitutions is known, and that the physical units of our coefficient estimates correspond directly to a change in the scoring rate.

Here we address potential ways to better extend the model as a useful interpretation of the game. One obvious issue is that the methods for estimating parameters in this model are considerably slower than simple regression, whether we use Monte Carlo methods or functional maximization, especially when more parameters or data points are added. If this method is to ever see conventional and public use, the computation must either be considerably faster or a new method of estimation must be used. Because this is a highly nonstandard likelihood function, it is a complicated matter to improve parameter estimates in a general way. Sequential updating may prove to be the easiest method to improve both methods, particularly with regard to particle filtering for hierarchical Bayesian methods.

We have also assumed that player ability is constant over the period considered, whether this is one season or five. There is considerable reason to expect that player abilities change from year to year in a meaningful way, such as a "career curve" [Berry, Reese and Larkey (1999)] or as simple deviations from a career mean. In this analysis, we chose to use the constant approach for several reasons, mainly that it would grossly magnify the number of parameters in a model where the data is already information-poor. We leave the introduction of single-player variability into this model as a subject of future research.

As a practical matter, there are several factors that can be explored immediately. Many have to do with the use of the time-dependent component of the Cox model, which we have kept as constant and unit-valued to this point.

Knowing location affects the short-term scoring rate. A game of hockey begins with a face-off at center ice, immediately after which neither team is very likely to score in the next few seconds. A distribution for the goal hazard after face-off was proposed by Thomas (2007), which begins at 0 for both teams and rises to a plateau with an exponential decay. If a team has the puck in their offensive zone, they are more likely to score a goal in the immediate future than the mean rate, and their opponents far less likely.

One approach is to include known puck possession and location terms as covariates in a general model; Macdonald (2012a), in particular, uses the zone in which the play starts as a mean-altering covariate. In our case, the natural point to include this is in the time-varying component to the Cox model, by choosing a relative hazard that starts at a rate given the state of play and returns to the overall mean. We expect that this will alleviate the issues highlighted in Section 5.1, wherein some players are frequently substituted in for defensive zone face-offs, a choice that unfairly penalizes their offensive ratings.

One benign side effect of this is that “garbage goals”—those scored after a longer scrum in an offensive zone, taking advantage of continued pressure rather than pure skill—would be down-weighted, since we would expect a goal to be much more likely in that scenario.

Including more events as outcomes. Since a goal is preceded by a shot on goal in the vast majority of cases, one method to improve the modeling framework is to consider shots to be a noncensored terminating state of a model instead of a goal. Since this would lead to a roughly tenfold count in the number of uncensored events, it would represent a great increase in the precision of estimates, especially if there was no individual variability on what fraction of shots on goal became goals. But this is certainly not the case, since there is significant variety on the fraction of shots that become goals (let alone shots on net) depending on the player; a defenseman’s slap shot is considerably less likely than a forward’s wrist shot. How we can include this feature in this model framework is an open problem, but may include information on the success rate of shots based on location and type as a post-processing step.

Censoring may be slightly informative. Shift lengths are either obtained directly or censored by player changes. One assumption we make is that the censoring mechanism is roughly exogenous and does not depend on or influence the state of the game in progress. While this assumption is clearly incorrect, the distributions of shift time are quite similar, as shown in Figure 7. Two immediate reasons

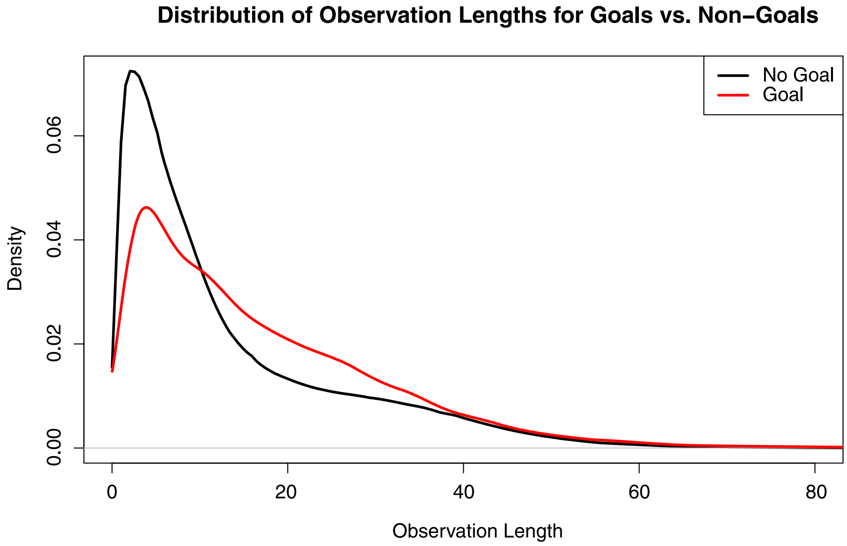

FIG. 7. *The lengths of shifts, conditioned on whether or not a goal was scored to terminate the observation. Shifts that end in goals are slightly longer.*

for this are clear. First, a goal is often scored following a longer scrum in the offensive zone, during which players have no opportunity to change off. Second, the changing process can be sequential; three players change, then shortly after, the other two change off, leading to a bias in short shifts. We expect that this factor can be accounted for, either through modeling or stratification, once we take puck possession and location into account.

Does the power play look like the process model? When a team has a man-advantage over their opponents, the game tends to look very different than a smooth stochastic process: the team on the power play sets up shop in their offensive zone, plays keep-away from their opponents and maneuvers to make a shot on goal. The short-handed team's prime goal in this period is not to score, but to remove the danger by clearing the puck from their own zone. (Scoring a short-handed goal is often seen as a bonus rather than the main objective while killing a penalty.)

To extend this model to the power-play situation, we would need to account for this in a principled manner. It may be sufficient to simply change the baseline scoring rates or to replace the penalized player with an indicator for the power play state, but this is subject to a future investigation and not at all obvious given the apparent differences in game play.

Acknowledgments We thank Brian Macdonald for introducing us to the shift-based method of dividing the game into components and for fruitful discussions. We also thank an Associate Editor and two anonymous referees for their helpful comments.

REFERENCES

BEAUDOIN, D. and SWARTZ, T. B. (2010). Strategies for pulling the goalie in hockey. *Amer. Statist.* **64** 197–204. MR2757163

BERRY, S. M., REESE, C. S. and LARKEY, P. D. (1999). Bridging different eras in sports. *J. Amer. Statist. Assoc.* **94** 661–676.

BROWN, L. D. (2008). In-season prediction of batting averages: A field test of empirical Bayes and Bayes methodologies. *Ann. Appl. Stat.* **2** 113–152. MR2415597

COOK, S. R., GELMAN, A. and RUBIN, D. B. (2006). Validation of software for Bayesian models using posterior quantiles. *J. Comput. Graph. Statist.* **15** 675–692. MR2291268

COX, D. R. (1972). Regression models and life-tables. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **34** 187–220. MR0341758

DAWID, A. P. (1994). Selection paradoxes of Bayesian inference. In *Multivariate Analysis and Its Applications* (*Hong Kong*, 1992). *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **24** 211–220. IMS, Hayward, CA. MR1479468

GRAMACY, R. B., JENSEN, S. T. and TADDY, M. (2013). Estimating player contribution in hockey with regularized logistic regression. Available at arXiv:1209.5026.

HANS, C. (2011). Elastic net regression modeling with the orthant normal prior. *J. Amer. Statist. Assoc.* **106** 1383–1393. MR2896843

HIROTSU, N. and WRIGHT, M. (2002). Using a Markov process model of an association football match to determine the optimal timing of substitution and tactical decisions. *Journal of the Operational Research Society* **53** 88–96.

HOERL, A. E. and KENNARD, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12** 55–67.

ILARDI, S. and BARZILAI, A. (2008). Adjusted plus-minus ratings: New and improved for 2007–2008. Available at http://www.82games.com/ilardi2.htm.

JAMES, W. and STEIN, C. (1961). Estimation with quadratic loss. In *Proc*. 4*th Berkeley Symp*. *Probab*. *Statist*. *Sympos*. *Math*. *Statist*. *and Prob*. **1** 361–379. Univ. California Press, Berkeley, CA. MR0133191

KLEIN, J. Z. and REIF, K.-E. (2001). *The Hockey Compendium*: *NHL Facts*, *Stats and Stories*. McClelland & Stewar, Toronto, Ontario, Canada.

LI, Q. and LIN, N. (2010). The Bayesian elastic net. *Bayesian Anal*. **5** 151–170. MR2596439

LOCK, D. and SCHUCKERS, M. (2009). Beyond +/-: A rating system to compare NHL players. Presentation at joint statistical meetings.

MACDONALD, B. (2011). A regression-based adjusted plus-minus statistic for NHL players. *J*. *Quant*. *Anal*. *Sports* **7**.

MACDONALD, B. (2012a). An expected goals model for evaluating NHL teams and players. In *MIT Sloan Sports Analytics Conference 2012*. MIT, Cambridge, MA.

MACDONALD, B. (2012b). Adjusted plus-minus for NHL players using ridge regression. Available at http://arxiv.org/abs/1201.0317v1.

MORRISON, D. G. (1976). On the Optimal Time to Pull the Goalie: A Poisson Model Applied to a Common Strategy Used in Ice Hockey. In *TIMS Studies in Management Science* **4**. North-Holland, Amsterdam.

PARK, T. and CASELLA, G. (2008). The Bayesian lasso. *J*. *Amer*. *Statist*. *Assoc*. **103** 681–686. MR2524001

ROSENBAUM, D. T. (2004). Measuring how NBA players help their teams win. Available at http://www.82games.com/comm30.htm.

SCHUCKERS, M. E., LOCK, D. F., WELLS, C., KNICKERBOCKER, C. J. and LOCK, R. H. (2011). National hockey league skater ratings based upon all on-ice events: An adjusted minus/plus probability (AMPP) approach. Unpublished manuscript. Available at http://myslu.stlawu.edu/~msch/sports/LockSchuckersProbPlusMinus113010.pdf.

SILL, J. (2010). Improved NBA adjusted +/- using regularization and out-of-sample testing. In *MIT Sloan Sports Analytics Conference*. MIT.

THOMAS, A. C. (2006). The impact of puck possession and location on ice hockey strategy. *J*. *Quant*. *Anal*. *Sports* **2** 18 pp. (electronic). MR2224831

THOMAS, A. C. (2007). Inter-arrival times of goals in ice hockey. *J*. *Quant*. *Anal*. *Sports* **3** 17. MR2326667

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J*. *R*. *Stat*. *Soc*. *Ser*. *B Stat*. *Methodol*. **58** 267–288. MR1379242

TIBSHIRANI, R. (1997). The lasso method for variable selection in the Cox model. *Stat*. *Med*. **16** 385–395.

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J*. *R*. *Stat*. *Soc*. *Ser*. *B Stat*. *Methodol*. **67** 301–320. MR2137327

A. C. THOMAS
S. L. VENTURA
S. MA
DEPARTMENT OF STATISTICS
CARNEGIE MELLON UNIVERSITY
5000 FORBES AVENUE
BAKER HALL 132
PITTSBURGH, PENNSYLVANIA 15213
USA
E-MAIL: act@acthomas.ca

S. T. JENSEN
DEPARTMENT OF STATISTICS
THE WHARTON SCHOOL
UNIVERSITY OF PENNSYLVANIA
463 HUNTSMAN HALL
3730 WALNUT STREET
PHILADELPHIA, PENNSYLVANIA 19104
USA