Approximate Bayesian binary and ordinal regression for prediction with structured uncertainty in the inputs

Aleksandar Dimitriev Erik Štrumbelj

Abstract

We present a novel approach to binary and ordinal prediction with uncertainty in the input variables. We were initially motivated by sports analytics data and the important task of predicting future outcomes from past performances. Unlike typical prediction settings, where inputs are assumed to be constants, the uncertainty in team characteristics is substantial, especially in the beginning of a season, and changes over time. Failing to account for this leads to overly confident predictions and prediction bias. We deal with this problem in two stages. First, we treat the inputs as random and apply a simple Bayesian structured measurement error model to estimate their distribution. And second, we use a Bayesian approach to prediction that can handle structured uncertainty in the inputs. The second-stage model could be fit using sampling-based approximation, but the computational cost is too high for practical use. Instead, we efficiently approximate the prediction model conditional on the inputs and then marginalize the conditional model over the input space using Monte Carlo approximation. For binary predictions, Bayesian logistic regression can be efficiently approximated using the well-known Laplace approximation and we extend the same approach to Bayesian ordinal logistic regression. We empirically evaluated the proposed model's prediction error on sports data and compared it to models that are typically used in this setting. Our approach substantially outperformed all other models, highlighting the severity of the problem of uncertain inputs. This makes the proposed approach a valuable contribution to the sports forecasting toolbox, but it can also be applied to other similar domains.

1 Introduction.

Machine learning, statistical analysis, and other quantitative approaches play an increasingly important role in sports analytics and professional sports. Applications include decision-support systems for coaching, ranking teams and players, sports forecasting, and many others (see [11, 12, 14] for a review).

Sport differs from most other application domains in that it features competition and recurring entities (teams, players) that potentially evolve over time, with both gradual (within season) and abrupt changes (injuries, player trades, between-season breaks). As such, sports gives rise to several unique and interesting data analysis challenges. In this paper we focus on one of these challenges that motivated our research and is particularly important in sports forecasting, but also relevant to prediction tasks in general.

To illustrate the problem, we use the case of predicting the outcome of a football match. The outcome of a football match depends on many things, but most importantly the characteristics of the two competing teams, which are typically derived from past performances. One characteristic that is known to be a very good predictor of football game outcomes is a team's scoring rate − the number goals scored per match. Using a historical data set of teams' scoring rates (and usually many other important input variables that can be derived from statistical summaries) and outcomes, we can then fit a model that connects the two and can be used to predict future outcomes.

One aspect of the above approach that has so far not been taken into account is that sports match summaries (typically counts, such as rebounds made in basketball, goals scored in football, etc...) are noisy, and that noise is transferred to any input variables we derive from them. When fitting a model, failing to account for this uncertainty will result in being overconfident in the parameters of the model. This is partially mitigated by the fact that uncertainty in the inputs decreases with the number of games available and we typically have many seasons worth of data available for fitting the model. However, the problem becomes much more pronounced when predicting future outcomes. Team characteristics change over time, in particular between the end of one season and the beginning of another, often to the extent that only matches from the new season are relevant to this season. This effectively resets the number of games available to estimate team characteristics and substantially increases the noise in the derived input variables. Therefore, we hypothesized that the accuracy of prediction models in sports can be substantially improved by first modelling the uncertainty in the inputs and second taking this uncertainty into account when fitting the model and making predictions. In this paper we focus on the second, while briefly addressing

the first in a limited case.

In its essence, the problem is similar to modelling data with measurement error, where input variables X (like scoring rates) are measured with error, with W being the actual measurements (goals scored). This area has received a lot of attention [3, 5] and measurement error models can roughly be split into two groups: functional or structural modelling [3]. Functional models are models that make no or very few assumptions about the distribution of X, while in strucured models, a model is placed on X, typically a parametric one.

In sports, most measurements are counts or derived from counts. Another difference compared to typical measurement error settings is that teams recur, so training samples are not independent or identically distributed, but we have multiple measurements of the same input variable. This allows us to use a structural approach to modelling measurement error. However, to fit a model and make predictions, we also need a model that can handle uncertain inputs.

We adopt a two-stage Bayesian approach. In the first stage we use simple conjugate analysis to model uncertainty in the inputs. This is described in more detail in Section 3.1. In the second stage, which is described in Section 2 and is the focus of this paper, we again use a Bayesian approach to model the relationship between the noisy inputs and the outputs and incorporate uncertainty associated with the inputs in the form of prior information. The proposed model can be fit using sampling-based approximation methods (Metropolis-Hastings [4], Hamiltonian Monte Carlo [2]), however, running times make it infeasible for practical use. Instead, we propose approximate prediction conditional on the inputs and marginalizing the inputs using a Monte Carlo approach, which we describe in more detail in Section 2. To achieve practical running times, this procedure requires an efficient fitting of the conditional model, so we utilize the well-known Laplace approximation to Bayesian logistic regression (Section 2.1) and extend the approach to Bayesian ordinal logistic regression (Section 2.2), the two most commonly used models in sports forecasting.

We empirically evaluate our approach and compare it to a baseline of typically used prediction models on two prediction tasks: predicting binary outcomes in basketball and predicting ordinal outcomes in football. The empirical evaluation and data used are described in Section 3. In Section 4 we present the results of the empirical comparison, which confirm our initial hypothesis that the proposed model would substantially improve predictive accuracy, in particular in the early stages of the season. With Section 5 we conclude the paper and offer directions for further work.

2 Methods.

The standard prediction setting is the following: we are given a dataset $X \in \mathbb{R}^{n \times m}$, which consists of n observations and m input variables, and a target variable $y \in \mathbb{R}^n$. We then train a model that can produce predictions y^* for a given new x^* by using the training data X. We adopt the Bayesian approach:

$$p(\beta|X, y) = \frac{p(X, y|\beta)p(\beta)}{p(X, y)} \propto p(X, y|\beta)p(\beta),$$

where β are the parameters of the model. For the prediction setting, we obtain the probability by marginalizing over the model parameters:

$$p(y^*|x^*, X, y) = \int p(y^*|\beta, x^*)p(\beta|X, y)\, d\beta.$$

In the above, the data are treated as constant. A more general problem arises in our setting, where we instead treat the inputs X and x^* as random variables, with densities $p(X|w)$ and $p(x^*|w^*)$, where w and w^* are known constants.

In this case, to train a model and obtain its posterior distribution $p(\beta|w, y)$ we must marginalize over the input space of possible training data sets X:

$$\begin{aligned} p(\beta|w, y) &= \int p(\beta, X|w, y)\, dX \\ &= \int p(\beta|X, y)p(X|w)\, dX. \end{aligned}$$

Similarly, to obtain a prediction for y^* we must not only marginalize over the parameters β, but also over the distribution of the test sample x^* as follows:

$$p(y^*|w,w^*,y) = \int_{x^*} \int_{\beta} p(y^*|x^*, \beta)p(x^*|w^*)p(\beta|w, y)\, dx^*\, d\beta \tag{2.1}$$

This approach is general, since it can be applied with any model that produces a posterior probability distribution over its parameters $p(\beta|X)$ and a distribution over its predictions $p(y^*|x^*, X)$ for a given test sample x^*. What remains for the model to be fully specified is defining the distributions that generate the training and test data sets.

The integral in Eq. (2.1) will generally be intractable even for the simplest of Bayesian models and is typically approximated using Monte Carlo approximation

$$E[y^*|w^*, w, y] \approx \frac{1}{N}\sum_{i=1}^{N} y^*_{(i)},$$

where $y^*_{(i)}$ is a random sample from the posterior predictive distribution in Eq. (2.1) and can be obtained by sequentially sampling $X_{(i)}$ from $p(X|w)$, $\beta_{(i)}$ from $p(\beta|X = X_{(i)}, y)$, $x^*_{(i)}$ from $p(x^*|w^*)$, and finally, $y^*_{(i)}$ from $p(y^*|x^* = x^*_{(i)}, \beta = \beta_{(i)})$.

The densities $p(X|w)$, $p(x^*|w^*)$ represent our structural measurement error model and are in most practical cases easy to sample from efficiently. The densities $p(\beta|X = X_{(i)}, y)$ and $p(y^*|x^* = x^*_{(i)}, \beta = \beta_{(i)})$ are the posterior and posterior predictive for the selected prediction model, conditional on the inputs being fixed, and we have to be able to efficiently sample from them. This implies that we need an efficient prediction model or, in the case of Bayesian models, which are typically computationally intensive, an efficient structural approximation to $p(\beta|X = X_{(i)}, y)$. In the remainder of this section we discuss two of the most commonly used prediction models in sports forecasting − logistic regression for binary outcomes, and ordinal logistic regression for ordinal outcomes.

2.1 Approximation to Bayesian logistic regression. When predicting a binary outcome, we will use Bayesian logistic regression in place of the general model that can be trained on a given dataset $X \in \mathbb{R}^{n\times m}$, $y \in \{0, 1\}^n$. It is a discriminative model that for a given example (x, y) yields a probabilistic prediction:

$$P(y = 1|x) = \sigma(\beta^T x),$$

where $\sigma(x) = 1/(1 + e^{-x})$ denotes the logistic sigmoid function and β is the parameter vector of the model. Its likelihood is thus:

$$P(X, Y|\beta) = \prod_{i=1}^{n} P(y_i = 1|x_i)^{y_i} P(y = 0|x_i)^{1-y_i}.$$

Assuming a normal prior on the parameters $\beta \sim \mathcal{N}(\mu_0, \Sigma_0)$, their posterior distribution is given by Bayes' theorem (Eq. 2):

$$p(\beta|X, y) \propto \prod_{i=1}^{n} \sigma(\beta^T x_i)^{y_i}(1 - \sigma(\beta^T x_i))^{1-y_i} \\ e^{-\frac{1}{2}(\beta-\mu_0)^T\Sigma_0(\beta-\mu_0)}.$$

Unfortunately, a closed-form expression does not exist for neither its posterior nor its predictive distribution. We resort to a structural approximation to its posterior distribution using the well-known Laplace approximation (see [1], pages 213-215). Although it is not normally distributed, its logarithm can be approximated well by a Gaussian distribution, because the posterior density has a single maximum. Let q_{map} denote this maximum, which can be easily obtained with any gradient method. This defines the mean of the Gaussian approximation. To obtain the covariance, we take the negative logarithm of the posterior:

$$-\ln p(\beta|X, y) \propto -\sum_{i=1}^{n} y_i\sigma(\beta^T x_i) + (1 - y_i)(1 - \sigma(\beta^T x_i)) \\ +\frac{1}{2}(\beta - \mu_0)^T\Sigma_0(\beta - \mu_0).$$

The inverse of the Hessian of this function, evaluated at the posterior mode, corresponds to the covariance $\Sigma_N = (-\nabla\nabla \ln\, p(\beta|X, y))^{-1}$. Thus, the approximation to the posterior is the following Gaussian:

$$p(\beta|X, y) \sim \mathcal{N}(q_{map}, (\Sigma_0^{-1} + X^T RX)^{-1}),$$

where R is a diagonal matrix with entries $r_{ii} = \sigma(q^T_{map}x_i)(1 - \sigma(q^T_{map}x_i))$.

2.2 Approximation to Bayesian proportional-odds model. In the previous section we discussed the approximation to Bayesian logistic regression and in this section which we now extend to the ordinal case. In particular, this includes the most common type of outcome in sports: win, draw, or lose. Note that the binary (logistic) regression model is a special case (for 2 categories) of the proportional-odds model covered in this section. However, due to all the extra complexity that arises when the number of categories is greater than 2, it makes sense to treat them separately.

Let n and m again be the be the number of samples, and input variables, respectively and k the number of (ordered) categories. The proportional-odds model (also known as ordinal logistic regression or proportional-hazards model) is the most commonly used model for the ordinal setting [9]. The model is based on the assumption that the odds of all binary decisions between categories are proportional to each other or, equivalently, that the $k-1$ logit surfaces are parallel:

$$\text{logit}(P(Y \le j|x)) = \log\left(\frac{P(Y \le j|x)}{P(Y > j|x)}\right) \\ = \beta x + \alpha_j, \text{for } j \in \{1, .., k-1\},$$

where β and α_j are parameters.

For convenience, we introduce $\alpha_0 = -\infty$ and $\alpha_k = +\infty$. The outcome probabilities can then be written as

$$\begin{aligned} P(Y = j|x) &= P(Y \leq j|x) - P(Y \leq j-1|x) \\ &= \sigma(\beta x+\alpha_j)-\sigma(\beta x+\alpha_{j-1}), \text{for } j \in \{1,..,k\}, \end{aligned}$$

where σ is again the inverse logit function.

The proportional odds model and its generalizations (see [10]), the non-proportional odds model (each surface has its own β, allowing for non-parallel logit surfaces) and partial-proportional odds (some of the β are surface-dependent) have received very little attention in the Bayesian setting. Here we focus on the proportional-odds model, but the approach could be extended to the non-proportional odds model as well (see Discussion).

First, we place prior distributions on the parameters. To ensure in-order intercepts, we introduce parameters d_j, $j \in \{1,..,k-1\}$ and a stick-breaking reparametrization of the $k-1$ parameters a_j with $a_j = \sum_{i=1}^{j} = d_i$. We place flat priors on the parameters $p(d_i) \propto 1$ and all d_i are restricted to be positive, except for d_1. As in the binary case, we place normal priors on the coefficients $\beta_1, ..., \beta_m \sim N(0, \sigma_\beta)$.

The model's likelihood is

$$p(\beta, d|\mathcal{D}) \propto \prod_{i=1}^{n}\prod_{j=1}^{k}(R_{i,j} - R_{i,j-1})^{(y_i=j)} \prod_{i=1}^{m} e^{\frac{(\beta_i)^2}{2\sigma_\beta^2}},$$

where $R_{i,j} = \sigma(\beta x_i + \alpha_j)$, and the log-likelihood

$$L = C - \frac{1}{2\sigma_\beta^2}(\beta \circ \beta) + \sum_{i=1}^{n}\sum_{j=1}^{k} I(y_i = j)\log(R_{i,j} - R_{i,j-1}),$$

where I is the indicator function and C is a constant.

The gradient of the log-likelihood cannot be written as succinctly as is the case with logistic regression. Instead, we start with the derivative of the log-likelihood of an arbitrary parameter θ:

$$\frac{\partial}{\partial\theta}L = -\frac{1}{2\sigma_\beta^2}\frac{\partial}{\partial\theta}(\beta \circ \beta) + \sum_{i=1}^{n}\sum_{j=1}^{k}\frac{\partial}{\partial\theta}L_{i,j},$$

where $\frac{\partial}{\partial\theta}L_{i,j} = 0$ if $y_i \neq j$ and

$$\begin{aligned} \frac{\partial}{\partial\theta}L_{i,j} &= \frac{1}{R_{i,j} - R_{i,j-1}}(\frac{\partial}{\partial\theta}R_{i,j} - \frac{\partial}{\partial\theta}R_{i,j-1}) \\ &= \frac{1}{R_{i,j} - R_{i,j-1}}\Big[R_{i,j}(1 - R_{i,j})\frac{\partial}{\partial\theta}(\beta x_i + \alpha_j) \\ &\qquad - R_{i,j-1}(1 - R_{i,j-1})\frac{\partial}{\partial\theta}(\beta x_i + \alpha_{j-1})\Big], \end{aligned}$$

otherwise.

Note that we do not derive the Hessian, because it had little effect on the accuracy and running times. All results reported in the empirical evaluation were obtained using a numerical approximation to the Hessian.

3 Empirical evaluation.

We will empirically evaluate our approach on two distinct sets of data. Binary regression will be evaluated on basketball data and ordinal regression on football (in football, draw has a very high probability and needs to be taken into account).

In both cases, the count data are first preprocessed to obtain uncertainty in the input variables, which is described in Section 3.1. Also in both cases, we include a typically used model in the comparison, to serve as a baseline for comparison (binary logistic regression and ordered logistic regression, for basketball and football, respectively). For the baseline models, only the mean counts are used, ignoring any uncertainty. We also include the baseline with noisy test cases. This is obtained by treating the test input variables as random, using the same structural model for the inputs as for the proposed model, and approximating the expected prediction of the baseline models with Monte Carlo sampling.

For the binary case we also include the proposed model without marginalization. That is, jointly modelling the β coefficients and X, including the uncertainty in the form of an informative prior on X. We implement this model in the probabilistic programming language and tool for Bayesian inference Stan [13]. Stan uses a variant of Hamiltonian Monte Carlo [2], the no-U-turn sampler [7].

Our empirical evaluation framework is a straightforward measurement of out-of-sample predictive accuracy, while respecting the time line. That is, all forecasts are made ex-ante, using only data available prior to a sports match. Sports data are naturally grouped by season and we use train-test season pairs, training on one season and predicting on the next, but for each test case prediction only data available prior to that match are used in the structural estimation of uncertainty (or computing the means, for the baseline models).

The structure of this sports data is such that we can group the features that correspond to a team e.g. by taking the mean of the shots missed for each game where that particular team played. We use this in training, so that when two teams play a game, instead of using the counts for that particular game, we use a preprocessed version of the average of all the games in the season where the team played. The preprocessing will be explained in the following section.

We measure predictive accuracy with mean squared error (MSE) in the binary case and the rank probability score (RPS) in the ordinal case. Note that both are proper scoring rules [6] and that for two categories, RPS is equivalent to MSE.

3.1 Preprocessing count data. In sports, the vast majority of sports statistics recorded during matches are count variables. Subsequently, almost all input variables that are used as predictors are either count variables or ratios of count variables, in particular ratios of the form $\frac{A}{A+B}$ where A and B are sums of count variables. For example, shot percentage (the fraction of shots that are successful) is a good predictor of victory in basketball, and it is of the form $\frac{n_s}{n_s+n_f}$, where n_s and n_f denote the number of successful and missed shots, respectively.

Our treatment of modelling the measurement error in the inputs will be straightforward and we will assume that the count variables follow a Poisson distribution. A natural choice for the prior distribution of the rate parameter λ is the Gamma distribution $\lambda \sim \text{Gamma}(a_0, b_0)$, the conjugate prior of the Poisson distribution. Therefore, for each count variable with mean $\bar{\lambda}_i$ over n_i observations games, the posterior is again Gamma with shape parameter $a_0 + \lambda_i n_i$ and scale parameter $b_0 + n_i$

$$\lambda_i | \bar{\lambda_i}, n_i \sim \text{Gamma}(a_0 + \lambda_i n_i, b + n_i),$$

where we select weakly informative priors $a_0 = b_0 = 0.001$.

A sum of Gamma distributed random variables with the same scale is again Gamma distributed with same scale. Furthermore, if A and B are Gamma distributed random variables with the same scale $A \sim \text{Gamma}(\alpha, \theta)$ and $B \sim \text{Gamma}(\beta, \theta)$, then the random variable $X = \frac{A}{A+B}$ is distributed $X \sim Beta(\alpha, \beta)$.. Therefore, for a ratio variable derived from Poisson posterior rates of the form $R = \frac{\sum_i \lambda_{A,i}}{\sum \lambda_{A,i} + \sum \lambda_{B,i}}$ is Beta distributed:

$$R | \bar{\lambda}_{A,i}, \bar{\lambda}_{B,i} \sim \text{Beta}(\sum \bar{\lambda}_{A,i}, \sum \bar{\lambda}_{B,i}).$$

The scale parameters for the θ are always the same, since the second parameter in the Gamma distribution corresponds to the number of games played, and it is the same for all counts at a particular point in time.

Two assumptions are made for the above to be true. First, the Poisson count variable rates are constant within a season and second, the are Poisson variables are independent. In practice, these assumptions are violated, because team characteristics do change during a season and are correlated. While a more sophisticated model for measurement error would increase the predictive accuracy, our main goal is to illustrate the benefits of taking into account the uncertainty in the inputs.

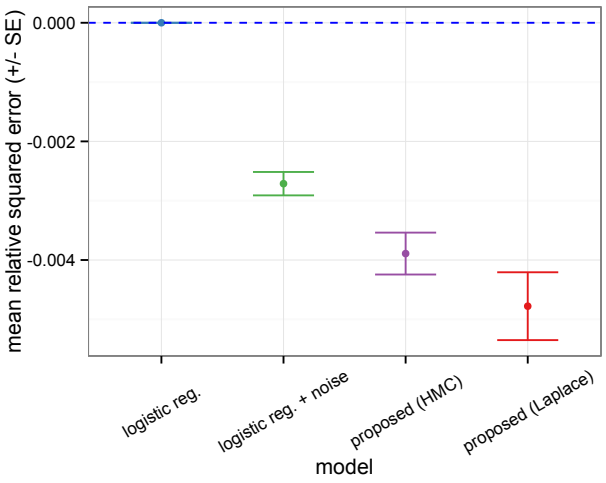

Figure 1: Estimated prediction errors across all train-test season pairs in the NBA basketball data set. All errors are relative to the baseline for comparison, the logistic regression model. That is, for each game, the baseline model's squared error was subtracted from all squared errors for that match to simplify interpretation of differences in the presence of strongly correlated errors.

3.2 Basketball data. In basketball, the outcome can be either a win or a loss. While there is also the chance for a draw after regular time (the chance of a draw occurring in the NBA is relatively low, approximately 6%), games go into overtime until a winner emerges. Therefore, prediction outcomes in basketball is a binary regression problem.

The basketball data used in our experiments consist of all the regular season and play-off games in the past 13 seasons (12 train-test season pairs) of the National Basketball Association (NBA) from 2001/02 to 2013/14. The data were obtained from `http://www.basketball-reference.com/`. Each NBA season contains approximately 1200 games. The exception is season 2011/12 which featured only 66 regular season games instead of the typical 82 regular season games per team, due to a lockout.

The count variables included in the data are counts of two-point shots made and missed, three-point shots made and missed, turnovers, offensive rebounds and defensive rebounds. We use these counts indirectly by transforming them into 8 ratios, described in [15], which

are known to be good predictors of basketball match outcomes. Note that these ratios are based on the well-known four factors effective field goal percentage, rebounding percentage, turnover percentage, and free throw percentage [8].

3.3 Football data. In a football match, there are three basic outcomes: win, draw, lose. Unlike most other sports, draws are common in football and have to be taken into account when modelling the outcome (approximately one third of all outcomes are draws). Furthermore, the nature of the three outcomes is ordinal - a win is more similar to a draw that it is similar to a loss - and it should be modelled as such. Therefore, forecasting football outcomes is an ideal application domain for illustrating the advantages of taking uncertainty into account in an ordinal setting.

Our football data set consists of the last 5 complete seasons (2010/11 - 2014/15) of each of the top 5 European national club competitions: English Premier League, French Ligue 1, German Bundesliga, Italian Serie A, and Spanish La Primera. That is, 4 train-test season pairs for each of the 5 leagues for a total of 20. In addition to the outcome, we include, for each match and each of the two teams that participated in the match, the number of goals scored, shots and shots on target, corners, fouls committed, and yellow and red cards received. The data were obtained from `http://football-data.co.uk/data.php`.

We assume the latent utility in the ordinal model is a linear combination of the of the rates of the recorded input variables: goals scored per match, shots per match, corners per match, etc... As described in the beginning of the section, we transform the count data inputs using Poisson-Gamma models with weakly-informative priors.

Similar to taking into account the within-season changes and correlations between input variables, better predictive accuracy could also be achieved by using a more informative set input variables. However, the main purpose is to illustrate the relative improvement achieved by taking uncertainty in the inputs into account. The same principles apply even if another (better) set of input variables is used.

4 Results.

Before presenting the results of the empirical evaluation for binary (basketball) and ordinal predictions (football), we first discuss the computation times of the experiments.

On an off-the-shelf laptop computer and for a single train-test season pair the baseline models (logistic and ordinal logistic regression) have negligible running times, the structural approximation variants of proposed model run in the order of minutes, and the HMC variant of the binary logistic regression model takes in the order of hours to achieve similar accuracy. Note that the HMC variant of the proposed ordinal model was not included in the comparison on football data sets, because the computation times make it infeasible for practical use.

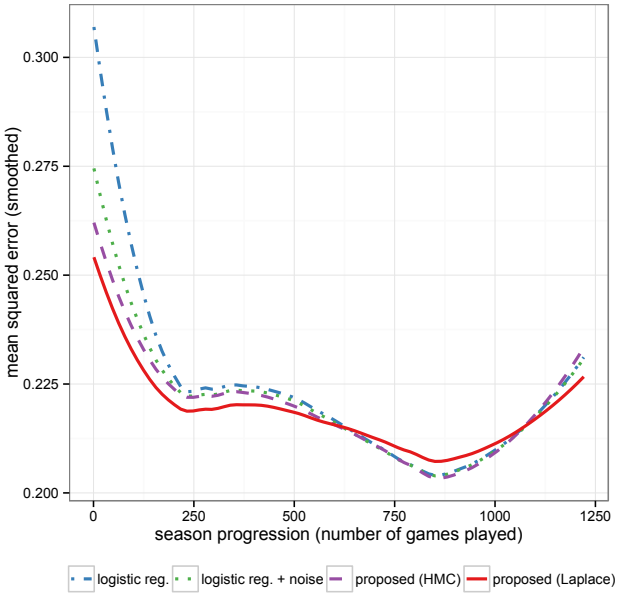

Figure 2: Prediction errors over the course of a season, averaged across all seasons in the NBA data set. For all models, the predictions are least accurate at the beginning of the season and improve as more information on teams' characteristics become available. Note that the decrease in accuracy at the end of the season corresponds to the playoffs stage of an NBA season, where only the top teams participate, making outcomes more difficult to predict.

4.1 Basketball. The structural approximation variant of the proposed model outperforms all other models (see Figure 1). These differences are not only statistically discernible, but also practically relevant (see [15] and references therein).

Although the HMC-based approximation yields relatively good predictions, it is discernibly worse than the structural approximation. This can, at least in part, be explained by the inferior accuracy of the HMC-based approximation due to slow mixing of the MCMC method. The effective sample sizes for the β regression coefficients were, on average, $\sim 20\%$ of the total number of

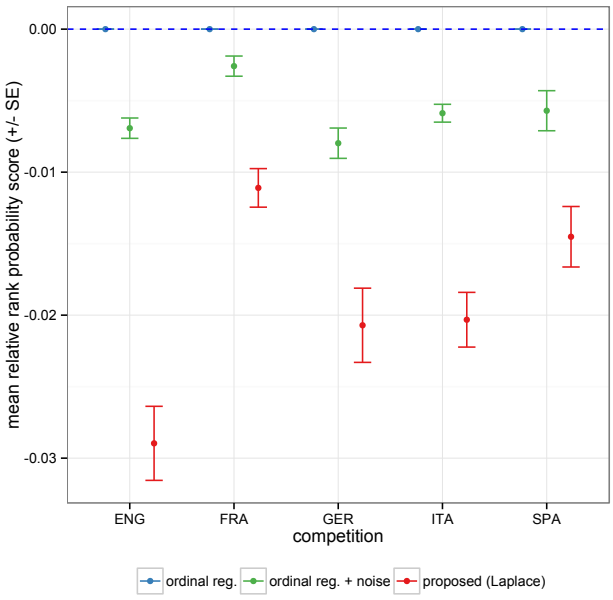

Figure 3: Estimated prediction errors for each football league and across all train-test season pairs in that league. Similar to Figure 1, the errors are relative to the baseline comparison, the ordinal logistic regression model.

iterations and at 1000 iterations, the estimated MCMC sampling error was, on average, approximately 10% of posterior standard deviation.

As we anticipated, the superior predictive accuracy of the proposed models is mostly due to better performance in the early stages of the season, where the uncertainty in the inputs is the highest (see Figure 2).

4.2 Football. Results on the football data sets are similar to those on the basketball data sets, however, the proposed model outperforms the other models even more convincingly across all football competitions (see Figure 3). This is not surprising, because, compared to NBA basketball, football seasons are much shorter (in terms of matches per team) and there is more uncertainty in the inputs derived from match statistics.

Again, similar to the binary case, the proposed model excels at the beginning of each season and the differences between models' prediction errors decrease as the season progresses and input variables become more certain (see Figure 4). Compared to NBA basketball, football seasons are much shorter (in terms of number of matches played), so the entire season has relative uncertainty in the inputs, making the differences between the proposed model and the other models more pronounced.

5 Conclusion.

Our hypothesis that ignoring the uncertainty in the input variables in sports will lead to less accurate predictions was correct. The model we developed to deal with this issue substantially outperformed typically-used models in this application domain. As expected, the improvement is most substantial at the beginning of the season, when the uncertainty is the highest. This was true for both basketball and football across all seasons and leagues in our experimental data.

Including uncertain inputs leads to more complex models and, in the case of sampling-based approximation, slow mixing. Therefore, an efficient approximation was necessary to achieve not only substantial improvement in predictive accuracy, but also sufficient efficiency for practical use.

In this paper we focused on prediction with structured uncertainty in the inputs and put less emphasis on the estimation of the uncertainty. Our modelling of measurement error was based on two assumptions: the Poisson rates of each variable are constant throughout a season and independent of each other. Future work could entail developing a model that allows the Poisson rates to vary over time and takes into account the possible correlations between them, e.g. by using spline interpolation or Gaussian processes and enforcing some covariance structure. The approximation approach we used for Bayesian binary and proportional-odds ordinal logistic regression could also be extended to non-proportional odds Bayesian ordinal regression and multinomial logistic regression, which is another model commonly used in sports analytics.

References

[1] C. M. Bishop, *Pattern recognition and machine learning*, springer, 2006.

[2] S. Brooks, A. Gelman, G. Jones, and X.-L. Meng, *Handbook of Markov Chain Monte Carlo*, CRC press, 2011.

[3] R. J. Carroll, D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu, *Measurement error in nonlinear models: a modern perspective*, CRC press, 2006.

[4] S. Chib and E. Greenberg, *Understanding the Metropolis-Hastings algorithm*, The American Statistician, 49 (1995), pp. 327–335.

[5] W. A. Fuller, *Measurement error models*, vol. 305, John Wiley & Sons, 2009.

[6] T. Gneiting and A. E. Raftery, *Strictly proper scoring rules, prediction, and estimation*, Journal of the American Statistical Association, 102 (2007), pp. 359–378.

[7] M. D. Hoffman and A. Gelman, *The no-U-turn sampler: Adaptively setting path lengths in Hamilto-*

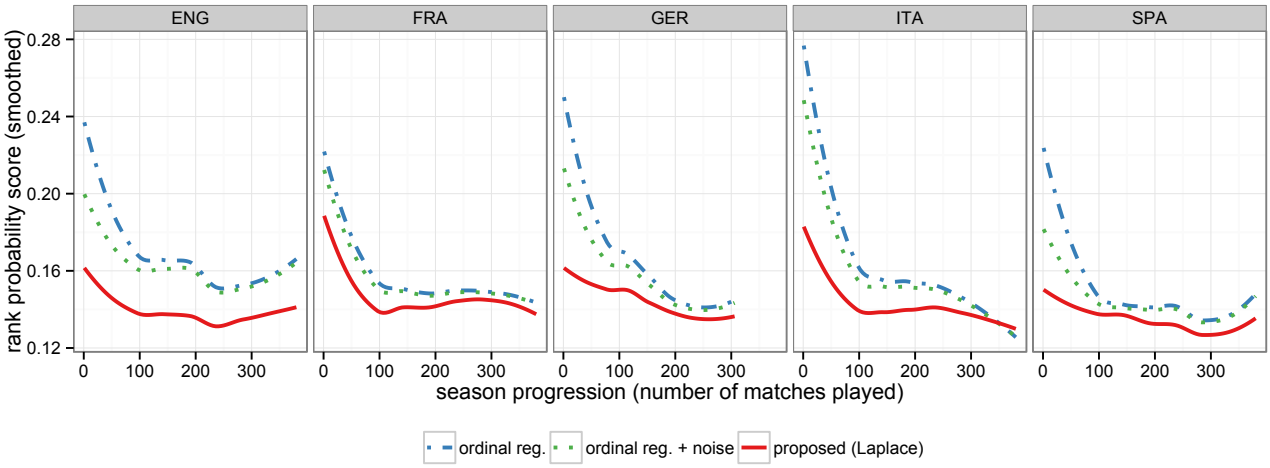

Figure 4: Prediction errors over the course of a season, averaged across all seasons and for each football competition separately.

nian Monte Carlo, The Journal of Machine Learning Research, 15 (2014), pp. 1593–1623.

[8] J. Kubatko, D. Oliver, K. Pelton, and D. T. Rosenbaum, *A starting point for analyzing basketball statistics*, Journal of Quantitative Analysis in Sports, 3 (2007).

[9] P. McCullagh, *Regression models for ordinal data*, Journal of the royal statistical society. Series B (Methodological), (1980), pp. 109–142.

[10] B. Peterson and F. E. Harrell Jr, *Partial proportional odds models for ordinal response variables*, Applied statistics, (1990), pp. 205–217.

[11] R. P. Schumaker, O. K. Solieman, and H. Chen, *Predictive modeling for sports and gaming*, Springer, 2010.

[12] O. Solieman, *Data mining in sports: A research overview*, Dept. of Management Information Systems, (2006).

[13] Stan developement team, *Stan Modeling Language Users Guide and Reference Manual, Version 2.8.0*, 2015.

[14] H. O. Stekler, D. Sendor, and R. Verlander, *Issues in sports forecasting*, International Journal of Forecasting, 26 (2010), pp. 606–621.

[15] E. Štrumbelj and P. Vračar, *Simulating a basketball match with a homogeneous Markov model and forecasting the outcome*, International Journal of Forecasting, 28 (2012), pp. 532–542.