Latent Variable Modeling of Paired Comparisons with Application to NCAA Men's Basketball Scores

STATS 700 Final Project Report

Simon Fontaine
Department of Statistics
University of Michigan
Email: simfont@umich.edu

Abstract—**We propose a latent variable model for paired comparisons in the context of Basketball scores. Based on the model proposed by [1], our model assumes multi-dimensional latent skills decomposed into offensive and defensive components as well as team- and conference-specific skills interacting with each other through an inner product model. We propose two inference approaches: maximum likelihood estimation (MLE) and mean-field variational inference (VI). Applied to the 2004-2017 NCAA Men's Basketball season's, the MLE approach yields adequate and interpretable results, but the VI approach was not as successful. Based on the MLE inference, we study the relationship between teams and conferences, we investigate league-wide trends over time and we produce rankings of teams and conferences.**

Contents

I. Introduction

The world of sports leads itself to many statistical analysis tasks: a vast amount of data is collected on various levels—per team, per match, per player, etc.—and learning from it has a wide array of application such as for teams to improve themselves or prepare themselves better against a forthcoming opponent, for leagues to investigate rule changes, for betters and oddsmakers to make better bets or to set accurate odds. In this work, we are interested in evaluating team's performance using matchup results. In particular, we analyze NCAA Men's Basketball matches to produce a predictive model for score outcome as well as a ranking of those teams.

The approach we use is a *latent variable model* where we assume that each team's performance is determined by a set of unobserved quantities. Learning about those latent variables and how they associate with observed scores then enables comparing teams that have yet not played against each other for prediction or ranking purposes. The proposed model is largely inspired from that in [1], which also considers a latent variable model for predicting Basketball scores, but introduces some changes to be discussed.

A main aspect of the model in [1] is the use of *offensive* and *defensive* latent skills. Since Basketball is a possession-based sport where teams exchange possession of the ball, teams alternate between offense and defense. Then, the observed score of a given team depends on how good they are at scoring points and how well the opposing teams is at preventing points. A similar decomposition of skills can also be found in Soccer scores analysis (e.g., [2]).

In the National Collegiate Athletic Association (NCAA), teams are aggregated by conferences of around 10-12 teams and, during the course of a season, teams play the majority of their matches against teams of their own conference. The model in [1] further consider conference-specific offensive and defensive skills in order to assess differences in strength between conferences as well as variation within conferences.

In addition to the offense/defense and team/conference skill decompositions, [1] consider multi-dimensional skills. Using multi-dimensional latent space allows the modeling of *non-transitive* relationships between teams. Indeed, while stronger teams generally perform better than weaker teams, there may be some specific matchups where this is not the case. In particular, certain teams' offensive style may be well-suited against some teams' defensive style, but not for others. Figure Figure 1 on page 2 shows a toy example using the famous *rock-paper-scissors* game of how multidimensional offensive and defensive skills allow for non-transitive comparisons. If closeness of offensive and defensive skills are modeled as producing higher scores, we can see that paper should score high against rock while rock should score low against paper, yielding a paper victory. The same can be said for rock beating scissors and scissors beating paper. Were skills unidimensional, such relationships would not be possible.

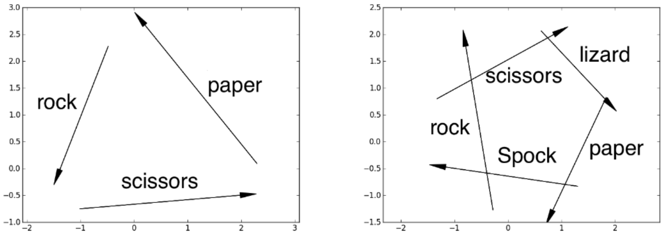

Fig. 1. Non-transitivity in latent spaces. Head of arrows represent offensive skills, tails, defensive skills. Diagram from [3].

A well-known phenomenon in sports is the "home-field advantage" where teams playing at home tend to perform better than playing at a neutral site or at an opponent's home. For example, during the 2017 season, teams playing at home had an overall record of 2930-1573 for a win percentage of 65%, way above an expected 50% under random assignment of location. It is therefore important to include this large effect as part of the model to improve its performance.

A. Goals

The main goal of this research is to construct a predictive model for NCAA Men's Basketball scores that incorporates conference effect and home-field advantage and allows non-transitive comparisons. Starting from the latent variable model described [1], we propose changes fixing some of shortcomings that will be described further:

- Unintuitive relationship between team and conference latent skills;
- Unintuitive home-field advantage;
- Independence between the two teams' scores in a match;
- Weak selection of the latent space dimension;
- Constant skills across seasons.

B. Data Description

We consider season and post-season NCAA Men's Basketball match results from 2004 to 2017 curated in a Kaggle dataset [4]. Each season consists of $N \approx 5000$ regular season matches between $T \approx 350$ teams split across $C \approx 30$ conferences (these values vary between seasons). For each match, we have access to the scores of both team as well as where the match was played (at one of the team's home or on a neutral site.) The post-season, i.e. the March Madness tournament, consists of 67 matches between a selection of the best regular-season teams. The tournament matches thus provides an interesting testing set as these matches are harder to predict: this is particularly true given that tournament matches are played on neutral grounds. For the current analysis, we omit all matches that went to overtime ($\approx 6\%$ each season) as the exposure is increased and scores lie on a different scale.

C. Organization

In Section II, we describe the model in [1] and discuss some of its problems, define the model we propose and discuss the changes made. In Section III, we describe two inference methods for fitting the proposed model: a maximum likelihood method where latent variables are treated as fixed and a variational inference method where they are treated as random. In Section IV, we briefly discuss the implementation of both inference methods. In Section V, we present results when our model and methods is applied to NCAA Men's Basketball data. In Section VI, we discuss some limitations and further improvements to the current research.

II. Model Description

A. Original Model of [1]

The authors of [1] propose the following generative model for the scores. For each team, the offensive and defensive latent positions are sample from K_1 independent Gamma distribution: for all $t \in [T]$,

$$\begin{aligned} T^o_{t,k} &\sim \text{Gamma}\left(s^o_t, r^o_t\right), \qquad k \in [K_1], \\ T^d_{t,k} &\sim \text{Gamma}\left(s^d_t, r^d_t\right), \qquad k \in [K_1], \end{aligned}$$

for some shape and rate hyper-parameters $s^o_t, s^d_t, r^o_t, r^d_t$. Similarly, conferences' offensive and defensive latent positions are sampled from K_2 independent Gamma distributions: for all $c \in [C]$,

$$\begin{aligned} C^o_{c,k} &\sim \text{Gamma}\left(s^o_c, r^o_c\right), \qquad k \in [K_2], \\ C^d_{c,k} &\sim \text{Gamma}\left(s^d_c, r^d_c\right), \qquad k \in [K_2], \end{aligned}$$

for some shape and rate hyper-parameters $s^o_c, s^d_c, r^o_c, r^d_c$. The home coefficient is sampled from a Gamma distribution:

$$H \sim \text{Gamma}\left(s_H, r_H\right),$$

for some shape and rate hyper-parameters s_H, r_H. Then for each match m, we define $h\left(m\right)$ and $a\left(m\right)$ as the team index of the home and the away team, respectively, and $c\left(t\right)$ as the conference index of team t. The scores are then modeled by independent Poisson centered at the sum of inner products of the offensive skill with the adversary's defensive skill for both the team and the conferences. The home-field advantage acts multiplicatively only on the home team:

$$\begin{aligned} Y^h_m &\sim \text{Poisson}\left(H\left[T^{o\top}_{h(m)} T^d_{a(m)} + C^{o\top}_{c(h(m))} C^d_{c(a(m))}\right]\right), \\ Y^a_m &\sim \text{Poisson}\left(T^{o\top}_{a(m)} T^d_{h(m)} + C^{o\top}_{c(a(m))} C^d_{c(h(m))}\right). \end{aligned}$$

When the match is played on a neutral site, the factor H is omitted for both teams.

B. *Comments on the Original Model*

We raise a few concerns regarding the previously described generative model which will justify some of the changes we propose.

First, the team and conference's latent variable lie in spaces of different dimensions and are therefore not directly interpretable. In a prediction setting, which is the main goal in [1], this is not a major concern, but if we wish to further analyses the learned latent representation, it is desirable that they at least lie in the same dimensional space. Furthermore, even if the spaces have the same dimensionality, the model would not associate the k-th team component to the k-th conference component.

Second, the contribution of the home-field advantage seems misguided and this can be seen particularly when comparing neutral-site matches to non-neutral site matches. Indeed, the effect is included only for teams playing at home. It seems preferable to include the inverse effect for team playing away so the overall effect agrees with the neutral case where no home-field advantage/disadvantage is added.

Third, basketball being a possession- and pace-based sport, we expect the two scores in a match to be correlated with each other, even given the team and conferences skills. In particular, a match with more possessions will induce higher scores for both team. This model does not allow such correlation and assumes that the latent skills will capture the propensity of matches to be low- or high-scoring.

Fourth, it is not clear that a Poisson model for the score is a reasonable assumption. While scores are non-negative integers, the variance restriction might not agree with observations. The Poisson assumption is partly influenced by a similar modeling strategy for Soccer scores (see, e.g., [2]). It is important to note that Soccer scores typically range between 0 and 5 so a integer-valued distribution is crucial in this case. The choice of Poisson model is also justified by conjugacy with priors as well as with taking inner products: the inference scheme therefore consists of analytical updates.

C. *Proposed Model*

The model we propose (Figure 2) will take a similar form except for a few changes addressing the various issues. We consider standard normal priors for teams' and conferences' offensive and defensive skills, but now choosing $K_1 = K_2 = K$:

$$\begin{aligned} T^o_{t,k} &\sim \mathcal{N}(0,1), \quad k \in [K], t \in [T], \\ T^d_{t,k} &\sim \mathcal{N}(0,1), \quad k \in [K], t \in [T], \\ C^o_{c,k} &\sim \mathcal{N}(0,1), \quad k \in [K], c \in [C], \\ C^d_{c,k} &\sim \mathcal{N}(0,1), \quad k \in [K], c \in [C]. \end{aligned}$$

To insure that both latent space have the same interpretation, we combine team-specific and conference-specific linearly by introducing a parameter $\lambda \in \mathbb{R}$ controlling the relative weight of conferences:

$$S^o_t = T^0_t + \lambda C^o_{c(t)}, \quad S^d_t = T^d_t + \lambda C^d_{c(t)}, \quad t \in [T].$$

Then, for each match m, we define team i's propensity to produce points as the inner product between that team's offensive skill and their adversary's defensive skill. For $t_i(m)$ defined as team's i index in match m, $i = 0, 1$, we define

$$M_{m,i} = S^{o\top}_{t_i(m)} S^d_{t_{1-i}(m)}, \qquad i = 0, 1.$$

Next, we include the home-field advantage additively:

$$\widetilde{M}_{m,i} = M_{m,i} + H h_i(m), \qquad i = 0, 1,$$

where

$$h_i(m) = \begin{cases} +1 & \text{team } i \text{ is playing at home in match } m, \\ 0 & \text{match } m \text{ is played at a neutral site}, \\ -1 & \text{team } i \text{ is playing away in match } m. \end{cases}$$

Finally, we write $\widetilde{M}_m = \left(\widetilde{M}_{m,0} \widetilde{M}_{m,1}\right)^\top$ and model the scores using a bivariate normal distribution:

$$Y_m \mid \widetilde{M}_m \sim \mathcal{N}_2\left(\mu \mathbf{1}_2 + c\widetilde{M}_m, \Sigma\right),$$

where $\mu \in \mathbb{R}$ centers the scores, $c \in \mathbb{R}$ scales the skills and $\Sigma \in \mathbb{R}^{2\times 2}$ is a symmetric positive definite matrix controlling the variance of the scores as well as the correlation between them. Additionally, we impose Σ to have following structure:

$$\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix},$$

where $\sigma^2 > 0$ and $\rho \in (-1, 1)$ are model parameters.

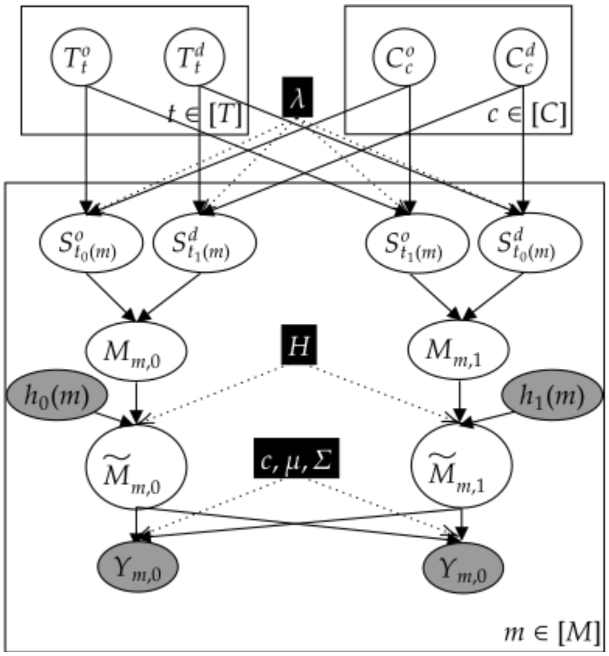

Fig. 2. Proposed model. White nodes are hidden variables; grey nodes are observed; black nodes are model parameters. [3]

D. Comments on the Proposed Model

The scaling parameter λ, together with the fixed scale of the latent variables, allows further estimation of how much a team's skill is determined, or captured, by the strength of their conference. For small values of λ, then most of the variation between team is team-specific; for large values of λ, most of the variation between team is captured by their conference.

The inner product factor, in conjunction with the Gaussian prior and Gaussian likelihood, breaks the conjugacy of the model and prevents utilizing analytical updates. In Section III, we will propose two inference methods to deal with this non-conjugacy.

To insure that our model in exchangeable within each match, i.e., that the team assignment to position 0 or 1 does not influence inference, we have to impose some structure on the Gaussian likelihood factor. To this effect, we impose scores to be centered at a symmetric mean $\mu\mathbf{1}_2$ and the covariance matrix Σ to be homogeneous.

The full effect of the home-field advantage on the outcome of a match can be understood as a difference of $2mH$ points between the home team and the away team.

In comparison with the Poisson model, we assume a fixed but estimable mean-variance relationship given $\widetilde{M}_m$. Indeed, for any such $\widetilde{M}_m$ we use the same covariance matrix which assumes that low-scoring games have the same variation in scores than higher-scoring games. While this may be an assumption not supported by data, it allows to estimate both the variance—which could be different from that in a Poisson model—as well as the covariance—which cannot be easily modeled in a Poisson model.

We treat $\theta = \left(\lambda, H, c, \mu, \sigma^2, \rho\right)$ as model parameters and therefore do not set prior distribution onto them. This assumption simplifies calculations and is justified by the fact that they are estimated from a large amount of observations so their respective posterior distribution would concentrate greatly.

III. Inference

The likelihood part of the generating model consists only of the Gaussian likelihood; indeed, conditionally on the latent variables, all other hidden variables are deterministic functions of one another. Let $\varphi_p\left(\cdot \mid \mu, \Sigma\right)$ denote the density of a p-variate Gaussian with mean μ and variance Σ and let $Z = \left(T^o, T^d, C^o C^d\right)$ denote all latent variables. Then, the likelihood is given by

$$p\left(Y \mid \theta, Z\right) = \prod_{m=1}^{M} \varphi_2\left(Y_m \mid \mu_m, \Sigma\right),$$

where $\mu_m = \mu\mathbf{1}_2 + c\widetilde{M}_m$. The log-likelihood is then

$$\begin{aligned}\ell\left(\theta, Z \mid Y\right) = &-\frac{M}{2}\log\det\left(2\pi\Sigma\right)\\ &-\frac{1}{2}\sum_{m=1}^{M}\left[Y_m - \mu_m\right]^\top \Sigma^{-1}\left[Y_m - \mu_m\right]. \quad (1)\end{aligned}$$

The prior term is given by

$$\pi\left(Z\right) = \prod_{x\in\{o,d\}}\left[\prod_{t=1}^{T}\varphi\left(T_t^x\right)\prod_{c=1}^{C}\varphi\left(C_c^x\right)\right]$$

A. Maximum Likelihood Estimation

A first approach in estimating the latent variables and the model parameters is through maximum likelihood estimation. In this setting, we consider the latent variables as fixed instead of as random. Now, instead of using the prior to constrain the latent variables, we rather impose a similar yet different constraint. In particular, we force each matrix of latent variable to be orthogonal column-wise; for example, we constrain $T^o \in \mathbb{R}^{T\times K}$ to be such that $T^{o\top}T^o = I_K$. This choice of constraint is motivated by two factors. First, each column corresponds to a latent component so this fixes the variation within each component. The two scaling parameters, λ and c, then adjust the latent positions to an appropriate scale. Note that since there are fewer conferences, the scale of C^o and C^d will be larger than that of T^o and T^d, but the scaling parameters will adjust for that fact. Second, the orthogonality between columns induces components that are different from each other. We then get more meaningful components and avoid repeated or highly correlated components.

The optimization problem can thus be stated as

$$\begin{aligned}\underset{Z,\theta}{\text{argmin}} \quad & -\ell\left(Z, \theta \mid Y\right)\\ \text{subject to} \quad & T^{o\top}T^o = I_K\\ & T^{d\top}T^d = I_K\\ & C^{o\top}C^o = I_K\\ & C^{d\top}C^d = I_K\end{aligned}$$

To obtain estimates, we proceed to a block-wise projected gradient descent algorithm (Algorithm 1). Cycling trough five blocks of parameters, $T^o, T^d, C^o, C^d, \theta$, we perform a gradient step and, if the current block is not θ, we perform a reprojection of the new estimate to the closest orthogonal matrix. This reprojection can be easily performed using a singular value decomposition. Let $X \in \mathbb{R}^{N\times K}$ be an updated latent variable matrix, which may not be column-wise orthogonal. Given the SVD $X = UDV^\top$, where $U \in \mathbb{R}^{N\times K}$ is orthogonal column-wise, $D \in \mathbb{R}^{K\times K}$ is diagonal and $V \in \mathbb{R}^{K\times K}$ is orthogonal, the nearest orthogonal matrix to X is given by setting the singular values to 1, i.e., $\widetilde{X} = UV^\top$.

Once the estimates are obtained, we can perform prediction of a new match's results by computing $\widetilde{M}$ for that match using the estimates and predict the mean scores as $\hat{\mu}\mathbf{1}_2 + \hat{c}\widetilde{M}$. Furthermore, we can predict a win probability: from the Gaussian likelihood, we have

$$Y_{m,0} - Y_{m,1} \sim \mathcal{N}\left(\mu_{m,0} - \mu_{m,1}, 2\sigma^2\left(1-\rho\right)\right),$$

so we can use the cumulative distribution function to compute the predicted probability of $\mu_{m,0} > \mu_{m,1}$, that is, a victory by team 0, using plug-in estimators.

Algorithm 1 Block-wise Projected Stochastic Gradient Descent

```
Input: Observed matches ($t_i(m)$, $h_i(m)$ and $Y_{m,i}$, $i = 0, 1$, $m \in [M]$) and conference assignment $c(t), t \in [T]$
Parameters: Convergence threshold, gradient step-size, mini-batch size
Procedure:
  1) Initialize $Z$ and $\theta$;
  2) Until convergence of the log-likelihood:
     a) Sample a mini-batch of matches;
     b) Choose a block $X \in \{T^o, T^d, C^o, C^d, \theta\}$;
     c) Compute the log-likelihood and its gradient w.r.t. $X$;
     d) Take a gradient step to update $X$;
     e) If $X \neq \theta$, reproject $X$ to the nearest orthogonal matrix.
Output: Estimated $\widehat{Z}$ and $\hat{\theta}$.
```

B. Variational Inference

A second estimation approach is to approximate the posterior distribution using variational inference. The posterior distribution, for fixed model parameters, is proportional to

$$p_\theta(Z \mid Y) \propto p_\theta(Z, Y) = \pi(Z)\, p(Y \mid Z, \theta),$$

where the proportionality constant is the marginal likelihood (or *model evidence*),

$$p_\theta(Y) = \int p_\theta(Z, Y)\, \mathrm{d}Z.$$

The model evidence can be used as a model selection criterion: larger model evidence is associated with better-fitting models. Variational inference relies on the following identity: for any distribution q with the same support as Z, and using Jensen's inequality,

$$\begin{aligned}\log p_\theta(Y) &= \log \int p_\theta(Z, Y) \frac{q(Z)}{q(Z)}\, \mathrm{d}Z \\ &\geqslant \mathbb{E}_q \left\{ \log \frac{p_\theta(Z, Y)}{q(Z)} \right\} \\ &= \mathbb{E}_q \{ \log p_\theta(Z \mid Y) \} + \mathrm{KL}(q \,||\, \pi) \\ &=: \mathrm{ELBO}_\theta(q),\end{aligned}$$

where $\mathrm{ELBO}_\theta(q)$ is known as the *evidence lower bound* under q for model parameters θ and where $\mathrm{KL}(q \,||\, p) = \mathbb{E}_{X \sim q} \left\{ \log \frac{q(X)}{p(X)} \right\}$ is the Kullback-Leibler divergence. Then, maximizing this lower bound is related to maximizing the model evidence. Note that the gap between the evidence and the lower bound is given by

$$\log p_\theta(Y) - \mathrm{ELBO}_\theta(q) = \mathrm{KL}(q \,||\, p_\theta(\cdot \mid Y)), \tag{2}$$

which implies that maximizing the ELBO corresponds to minimizing that gap and thus minimizing the KL divergence between the true posterior $p_\theta(\cdot \mid Y)$ and the approximation q.

Now, the gap (2) is minimized by choosing $q = p_\theta(\cdot \mid Y)$, but this requires us to compute the true posterior. Instead, we choose q to be in a sufficiently simple family within which we can find the best approximation. In this case, we consider a fully-factorized (*mean-field*) Gaussian family,

$$q_\phi(Z) = \prod_{x \in \{o,d\}} \left[\prod_{t=1}^{T} \varphi\left(T_t^x \mid \mu_{T_t^x}, \sigma^2_{T_t^x}\right) \prod_{c=1}^{C} \varphi\left(C_c^x \mid \mu_{C_c^x}, \sigma^2_{C_c^x}\right) \right],$$

where ϕ denote the set of variational parameters, i.e., the set of all means and variances. The goal is to find the best such approximation, that is, the optimal variational parameters ϕ:

$$\begin{aligned}\hat{\phi} &= \operatorname*{argmin}_{\phi} \mathrm{KL}(q_\phi \,||\, p_\theta(\cdot \mid Y)) \\ &= \operatorname*{argmax}_{\phi} \mathrm{ELBO}_\theta(q_\phi).\end{aligned}$$

Furthermore, we optimize over model parameter to get the optimization problem

$$\left(\hat{\theta}, \hat{\phi}\right) = \operatorname*{argmax}_{\theta, \phi} \mathrm{ELBO}_\theta(q_\phi). \tag{3}$$

In order to solve (3), we consider a Stochastic Gradient Variational Bayes (SGVB) algorithm [5] summarized in Algorithm 2. We note that the ELBO decomposes as the expected likelihood and the KL term. The KL term has a closed form expression: by the fully factorized form of both the approximated posterior and the prior, we get

$$\begin{aligned}\mathrm{KL}(q_\phi \,||\, \pi) &= \mathbb{E}_{q_\phi} \left\{ \sum_{z \in Z} \log \frac{\varphi(z \mid \mu_z, \sigma_z^2)}{\varphi(z)} \right\} \\ &= \frac{1}{2} \sum_{z \in Z} - \log \sigma_z^2 - 1 + \mu_z^2 + \sigma_z^2.\end{aligned}$$

This expression is differentiable w.r.t. ϕ so we are able to compute gradients. For the expected likelihood, we find

$$\mathbb{E}_{q_\phi} \{ \log p_\theta(Z \mid Y) \} = \sum_{m=1}^{M} \mathbb{E}_q \{ \log \varphi_2(Y_m \mid \mu_m, \Sigma) \}.$$

However, the distribution of μ_m under the variational approximation is non-trivial because of the inner product. Hence, we resort to a Monte Carlo (MC) estimate of the expected likelihood:

$$\mathbb{E}_q \{ \log p_\theta(Z \mid Y) \} \approx \frac{1}{B} \sum_{b=1}^{B} \log p_\theta\left(Z^{(b)} \mid Y\right),$$

where the $Z^{(b)}$'s are independent samples from the current posterior approximation. Similarly, the gradient w.r.t. ϕ could be estimated by the same MC sample,

$$\nabla_\phi \mathbb{E}_{q_\phi} \{ \log p_\theta(Z \mid Y) \} \approx \frac{1}{B} \sum_{b=1}^{B} \nabla_\phi \log p_\theta\left(Z^{(b)} \mid Y\right),$$

but sampling $Z^{(b)}$ loses the dependency of the variational parameters. We thus resort to the reparameterization trick and

Algorithm 2 Stochastic Gradient Variational Bayes with MC approximation under reparamaterization trick

```
Input: Observed matches ($t_i(m)$, $h_i(m)$ and $Y_{m,i}$, $i = 0, 1$, $m \in [M]$) and conference assignment $c(t), t \in [T]$
Parameters: Convergence threshold, gradient step-size, mini-batch size, MC sample size
Procedure:
  1) Initialize $\phi$ and $\theta$;
  2) Until convergence of the ELBO:
       a) Sample a mini-batch of matches;
       b) Compute KL $(q_\phi \,||\, \pi)$ and its gradient w.r.t. $\phi$;
       c) Sample $\boldsymbol{\varepsilon}^{(b)}$ for $b \in [B]$;
       d) Compute the MC estimate of the expected likelihood and its gradient w.r.t. $(\theta, \phi)$;
       e) Take a gradient step.
Output: Estimated $\widehat{\phi}$ and $\hat{\theta}$.
```

rather work with $Z_z^{(b)} = \mu_z + \sigma_z \varepsilon_z^{(b)}$ for iid $\varepsilon_z^{(b)} \sim \mathcal{N}(0, 1)$ in which case $\nabla_\phi \log p_\theta \left(Z^{(b)} \mid Y\right)$ can be computed explicitly.

To predict new matches, we proceed similarly. Given the estimated variational and model parameters, we sample latent variables and feed them to the decoder to find the corresponding $\widetilde{M}$ and then $\hat{\mu}\mathbf{1}_n + \hat{c}\widetilde{M}$ is the predicted scores. Averaging over independent samples of latent variables yields our predicted score.

IV. Implementation Details

The implementation of both inference methods, as well as the code reproducing the results, methods is available at https://github.com/fontaine618/700-Project. All code is in `Python` and relies on the `pyTorch` library [6] for automatic differentiation and optimization.

A. Maximum Likelihood Estimation

For the gradient step, we use adaptive step-size computed by the Adam optimization method [7]. The variance parameter σ^2 in the Gaussian likelihood is stored and updated on the log scale; the correlation parameter ρ is stored and updated on the hyperbolic tangent scale. The mini-batch size is set to be 500 where the number of matches is around 5000.

B. Variational Inference

To implement the SGVB, we use a Variational Auto-Encoder scheme. The encoder splits into 4 encoders, one for each of the four latent variables. Each of them consists of two parallel fully connected linear layers without bias, with input size equal to the number of teams T or conferences C and with output size K. The inputs are one-hot encodings of the team or conference index so that the weights corresponds to the variational parameters. Again, the variance parameters are stored and updated on the log scale. Then, a sampling step, using the reparameterization trick, returns sampled latent variables to be fed to the decoder. The decoder then takes latent variables and computes the $\widetilde{M}_m$'s deterministically before comparing them to the observed scores in the Gaussian likelihood.

Again, we use the Adam optimizer [7] for adaptive step-sizes in the gradient step and reparameterize σ^2 and ρ as previously. The MC sample size for estimating the expected likelihood and its gradient is chosen to be 1.

V. Results

In this section, we fit our model using the MLE to NCAA Men's Basketball scores from 2004 to 2017; results for the variational inference method will be discussed in Section V-E.

Contrarily to the analysis in [1], we fit our model on each season independently. While teams have relatively constant performance across years, the actual composition of each teams, in terms of players, changes rapidly. Indeed, players can only play up to five years and the players getting the largest share of play time tend to play even less than that (e.g., freshman play less and seniors tend to play more). In particular, good teams suffer from the *one-and-done* phenomenon were very talented player only play for a year before going to the NBA. In [1], they consider only four seasons so the variation between years will be lesser than over our 14 seasons, so it seems less of a problem in their analysis, but our longer time span warranted this distinction. We discuss in Section VI-B potential ways to model skills through time.

A. Selecting the Latent Space Dimension

One modeling aspect of [1] that was overlooked was the selection of the dimension space. The authors consider a few different latent dimension spaces and do not observe significant difference between them. In particular, they consider $(K_1, K_2) \in \{(1,0), (1,1), (10,10), (5,15)\}$ as well as a average of 10 models with varying dimensions, but very little improvement in performance can be seen by increasing the dimensionality.

In this section, we use the MLE method in order to select an appropriate latent space dimension favoring smaller, more interpretable latent spaces. To this end, we fit our model for each of the 14 seasons and for $K \in [10]$, and compute training (regular-season matches) and testing (tournament matches) metrics. In particular, we consider the log-likelihood of the scores and their mean squared error (MSE) as well as prediction accuracy and the binary cross-entropy of the predicted win probabilites.

The results can be found in Figure 3. As expected, training metrics improve as K increases, but not much improvement can be observed beyond $K = 2$ or $K = 3$. Indeed, the largest improvement in performance is between $K = 1$ and $K = 2$. As for testing metrics, we find that the performance decreases for K larger than 2, especially for the two metrics on scores (log-likelihood and MSE); testing metrics on win probabilities seem relatively constant with K. For these reasons, we choose to consider a model with $K = 2$ latent dimensions: adjustment to training data seems as good as it can be, performance on test matches seems better and it has the convenience of being easier to analyze visually.

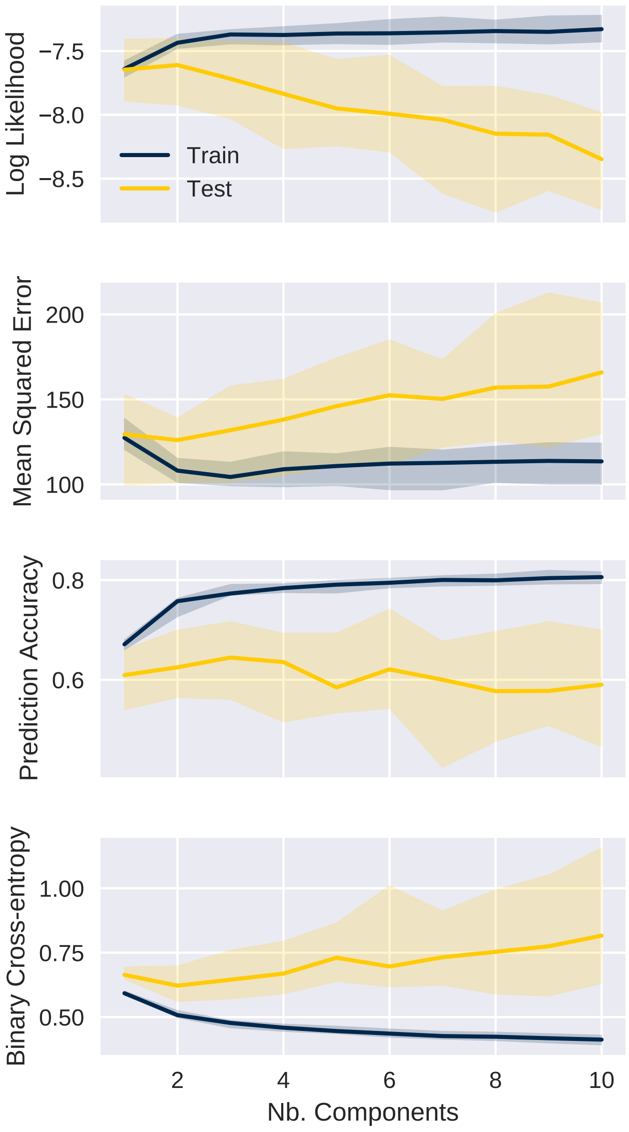

Fig. 3. Training and testing metrics under MLE for varying latent dimension K. Lines represent the median across the 14 seasons; bands represent the minimum and maximum range across the 14 seasons.

We note that win prediction is much harder on the testing set than on the training set because matches are played on neutral grounds which removes the home-field advantage that would ease predictions. Together with the general understanding that out-of-sample prediction is harder, this seems to explain the worse win prediction metrics on the testing set.

B. League-wide Variations over Time

In each of the 14 seasons considered, the model estimates not only the latent variables, but also some global model parameters. In this section, we investigate the variation over time of the estimated parameters in order to extract league-wide trends. Using the MLE method on a model with $K = 2$ latent dimensions, we obtain the results in Figure 4.

In the case of the home-field advantage, whose effect is computed as $2Hc$ corresponding to the difference between expected scores, we find that playing at home has the effect of adding between 2.5 and 4.5 points to a team's score, depending on the season. We also find that this effect has recently diminished, especially starting from the 2014 season. These results seem to agree with other analyzes: for example, [8] finds that per-team home-field advantage range roughly between 1 and 7 points with a national average of 3.5 points.

The overall mean μ score appears to remained relatively constant throughout the years with a slight increase in the last four seasons. The unexplained variation in the scores, captured by the global parameter σ^2, remained fairly constant with the seasons with indications of a decrease starting from the 2011 season. The estimated value of σ^2 ranges between 101 and 116 inducing a standard deviation of 10 to 11 points around the fitted means. As for the correlation between two teams' scores ρ, we find a steady estimated value around 0.4, which indicates moderate correlation. The inclusion of this correlation in the model was justified from intuition, but the data seem to corroborate our *a priori*.

C. Exploratory Analysis of the Latent Variables

Using MLE inference on a model with $K = 2$ components, we analyze the latent variables for the 2014 season. For $K = 2$ components, latent skills are determined by 2 offensive skills and 2 defensive skills: we can therefore plot the 4 latent skills in a 2-dimensional plot we drawing the oriented line segment from defensive skills to offensive skills. As mentioned in the introduction, we can understand this type of plot by comparing heads of arrows (offense) to tails of arrows (defense): when they are close to each other, their inner product will be large and induce larger mean score and when they are far apart, this results in small or even negative inner products indicating smaller mean score.

In Figure 5, we plot the 33 conferences' latent skill and highlight two conferences, the Big Ten conference (blue) and the Big Sky conference (yellow). Comparing their respective skill arrows, we see that the Big Ten would have a large mean score because of the positive inner product and that the Big Sky would have a small mean score because of the negative inner product. More generally, we see that better defenses are in the upper left quadrant as they are far from most offenses, worse defenses on the upper right quadrant since they intersect with some offenses, which coincides with better offenses, and worse offenses tend to be in the lower right quadrant.

Next, we consider the team latent skills as computed from the weighted sum of team-specific and conference-specific skill $S_t = T_t + \lambda C_{c(t)}$. Figure 6 depicts all 351 teams in the 2014 season, but highlight teams from the Big Ten and the Big Sky conferences. By our modeling choices, we see that the conference-specific and team-specific skills have the same meaning by the additivity. We find that the conference effect is much larger than the team effect as the variation within conferences is smaller than the variation between conferences. These results seem to confirm one of the prior model assumption that

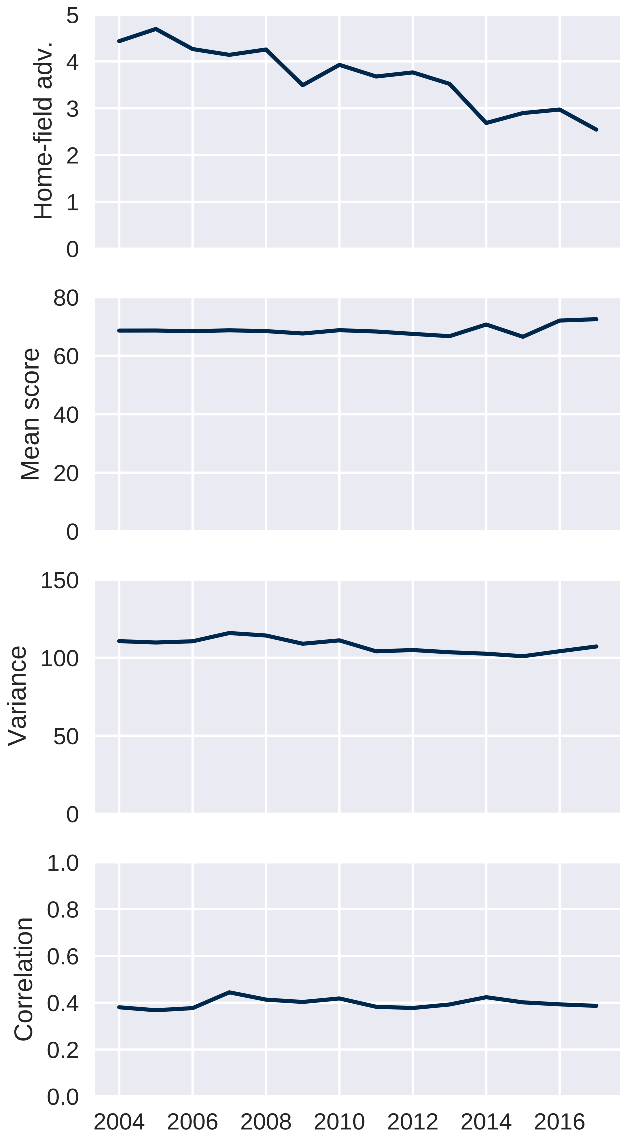

Fig. 4. Estimated global parameters using the MLE method on a model with $K = 2$ latent dimensions. The home-field advantage refers to $2Hc$, the mean score to μ, the variance to σ^2 and the correlation to ρ.

we should include the conference effect. An observation that can be made from the stronger effect of conferences is that conferences with longer skill arrows will have smaller mean score when two teams from the same conference face off. Indeed, the team-specific effect being smaller, the conference effect dominates leading to smaller inner products and thus smaller mean scores. Alternatively, conferences with shorter skill arrows will have offensive skills closer to their defensive skills, leading to larger mean scores.

D. Team and Conference Rankings

An important observation to be made about Figures 5 and 6 is that teams and conferences compare relatively transitively. Indeed, the distribution of skill vectors does not go "full circle" as in the *rock-paper-scissors* example (Figure 1). This suggest that we can produce an overall ordering of teams and conferences. The authors of [1] propose a ranking based on the score difference when facing an "average" opponent. While this approach is sensible, we rather produce a ranking based on how many opponent a given team or conference would be favored by our model. Using our estimated latent skills and model parameters, we predict the result of all pairwise comparisons and compute the umber of predicted wins.

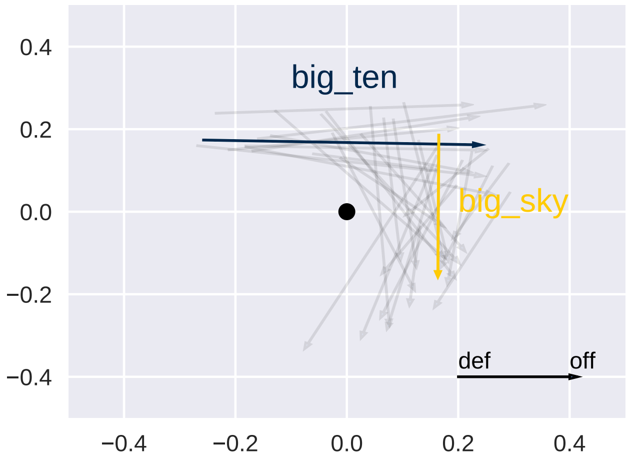

Fig. 5. Conference latent skills in the 2014 season.

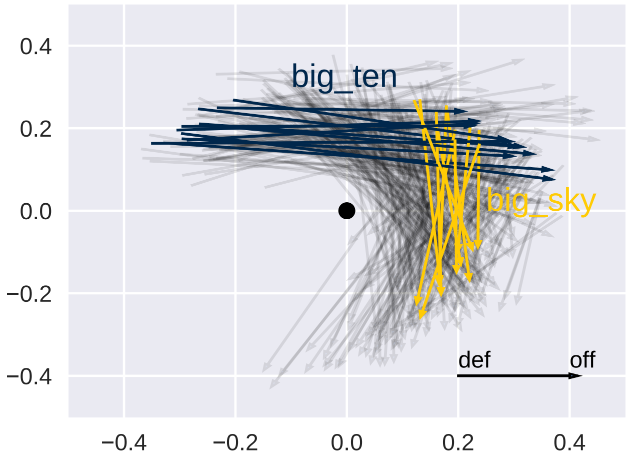

Fig. 6. Team latent skills in the 2014 season.

Table I contains the conference ranking for the 2014 season where our ranking is compared to [1]'s ranking using 2010-2014 data. First, we note that there is few instances of non-transitivity; indeed, except for middle-of-the-pack conferences, the predicted number of wins follows exactly the reverse rankings. In particular, the highest-ranked conference, the Big East conference, is predicted to beat any other conference. Our ranking differs slightly from [1]'s ranking mostly because of the different time span. For example, the large increase in rankings for the Big East and Southeastern conferences follows from the fact that these conferences add especially strong teams in 2014 (the Big East conference had 3 teams ranked in the Top 25 AP poll [9] at the end of the season, the Southeastern conference had the best ranked team, Florida).

TABLE I
CONFERENCE RANKINGS FOR THE 2014 SEASON.

	Conference	Pred. wins	[1] rank*
1	Big East	32	7
2	Big 12	31	4
3	Big Ten	30	2
4	Pacific-12	29	1
5	Atlantic Coast	28	3
6	Southeastern	27	13
7	American Athletic	26	8
8	Atlantic 10	25	5
9	Mountain West	24	11
10	West Coast	23	12
11	Missouri Valley	22	15
12	Mid-American	21	16
13	Horizon League	20	14
14	Conference USA	19	9
15	Big West	18	18
16	Sun Belt	17	20
17	Colonial Athletic Association	15	6
—	Summit League	15	25
19	Ivy League	13	22
—	Western Athletic	13	32
21	Metro Atlantic Athletic	12	33
—	Big Sky	12	19
23	Ohio Valley	11	23
24	Patriot League	9	31
25	Southland	8	29
26	Atlantic Sun	7	34
27	Southern	6	21
28	Big South	5	10
29	Southwest Athletic	4	27
30	Northeast	3	24
31	Mid-Eastern Athletic	2	30
32	America East	1	35
33	Independent	0	—

**based on the 2010-2014 seasons.*

We proceed similarly with teams by simulating all possible matches and recording the predicted winner. Table II contains the top 25 along with the predicted number of wins, the AP Poll ranking [9] and their end-of-season record. Our model predicts that Arizona should win against any of the other 350 teams and there seem to be very few instances of non-transitivity. The AP Poll ranking [9] lists the top 25 teams according to a panel of 65 "experts" which are trying to asses the strength of teams from regular-season results and taking into account the strength of their opponents. In comparison to that ranking, our ranking seem to disagree: teams such as UCLA and Creighton are ranked better under how model and teams such as Michigan and Iowa St. are ranked lower than the expert opinion. We notice a few interesting patterns with respect to conferences. The Big East conference being ranked so high induces high rankings for their teams: Creighton, Providence, Xavier and St John's all get promoted under our model. Conversely, the American Athletic conference conference is ranked slightly lower than other power conferences and this affects its teams' ranking: Louisville gets a lower ranking and teams such as Cincinnati and Connecticut are excluded from the top 25 compared to the AP Poll.

TABLE II
TEAM RANKINGS FOR THE 2014 SEASON.

	Team	Pred. wins	AP rank [9]	Record
1	Arizona	350	4	33-5
2	Villanova	349	6	29-5
3	UCLA	347	20	28-9
—	Creighton	347	16	27-8
5	Florida	346	1	36-3
6	Kansas	345	10	25-10
7	Virginia	344	3	30-7
8	St John's	343	—	20-13
9	Duke	341	8	26-9
10	Gonzaga	340	—	29-7
11	Oklahoma	339	21	23-10
12	Louisville	338	5	31-6
13	Oregon	337	—	24-10
—	Oklahoma St.	337	—	21-13
15	Baylor	336	23	26-12
16	Wisconsin	335	12	30-8
17	Michigan St.	334	11	29-9
18	Michigan	332	7	28-9
—	Tennessee	332	—	24-13
20	Providence	329	—	23-12
—	Utah	329	—	21-12
—	Xavier	329	—	31-13
23	Ohio St.	328	22	25-10
24	Iowa St.	327	9	28-8
25	Kentucky	326	—	29-11

By highlighting the best and worse teams in the latent space, we can get a confirmation of our interpretation of the latent positions. In Figure 7, we depict all 351 skill vectors and show the best and worst 25 teams according to our ranking. We see a clear clustering pattern where offensive and defensive skills tend to lie in the same region within each subgroup. In particular, we find that the best offenses are very close to the worst defenses (large positive inner products) and that the best defenses are almost diametrically opposed to the worst offenses (large negative inner products).

E. Variational Inference

Experimentation under the variational inference method described in Section III-B did not yield convincing results. In particular, the algorithm converges to mean values: team and conference effects are set to 0 and only the home-field advantage influences the predictions. Hence, we do not recover meaningful latent variable estimates and the training and testing metrics are significantly worse than using the MLE approach (Figure 8). For example, the training prediction accuracy agrees with predicting the team playing at home all the time (65%) and the testing accuracy is around 50% since all matches are played on a neutral site. Furthermore, we do not observe any improvement nor any deterioration with the latent dimension K.

Different attempts at improving on those results were considered, but none were particularly effective. The initialization of the latent variables' mean was taken to be either random

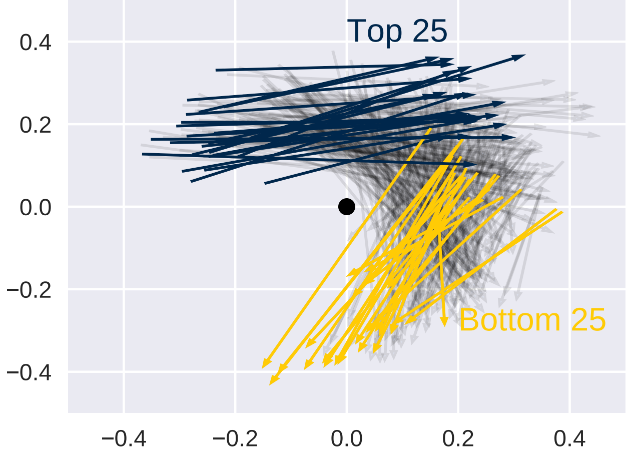

Fig. 7. Team latent skills in the 2014 season, best and worse team highlighted.

or from the MLE estimate for a model with the same dimension: both cases yielded similar results. Varying the MC approximation sample size did not improve the fit either. Block updates of the variables and changing other optimization details (non-adaptive step-sizes, step-size, momentum, mini-batch size, etc.) were not successful approaches either.

VI. Discussion

A. Limitations

One important observation emerging from our results is that our model and the MLE inference seem to put a lot of weight on conferences compared to teams. While domain knowledge indicate that conferences have different strengths, we would expect variation within conferences to be larger than what was obtained. Additionally, the constraint that latent variable matrices are orthogonal does not impose conference skills to be the average skill in a conference.

The variational inference framework yielded rather underwhelming results and the reason for it is yet unknown.

B. Potential Improvements

A way to model the relationship between teams and conferences differently would be to consider a distance model where, instead of an inner product, we would use a $\mathcal{L}_2$ distance between team skills:

$$M_{m,i} = \left\| S^o_{t_i(m)} - S^d_{t_{1-i}(m)} \right\|_2^2, \qquad i = 0, 1,\ m \in [M].$$

This alternative model would produce intra-conference means that are more consistent with intuition. Indeed, we can write

$$\begin{aligned} &S^o_{t_i(m)} - S^d_{t_{1-i}(m)} \\ &\quad = \left(T^o_{t_i(m)} - T^d_{t_{1-i}(m)} \right) + \lambda \left(C^o_{c(t_i(m))} - C^d_{c(t_{1-i}(m))} \right), \end{aligned}$$

where we note that the second term will be a constant effect for any two teams in a conference. This approach was initially considered, but the inner product model was preferred because it allows for positive and negative values of $M_{m,i}$ in which case the overall mean μ has the expected interpretation instead of being the minimum score.

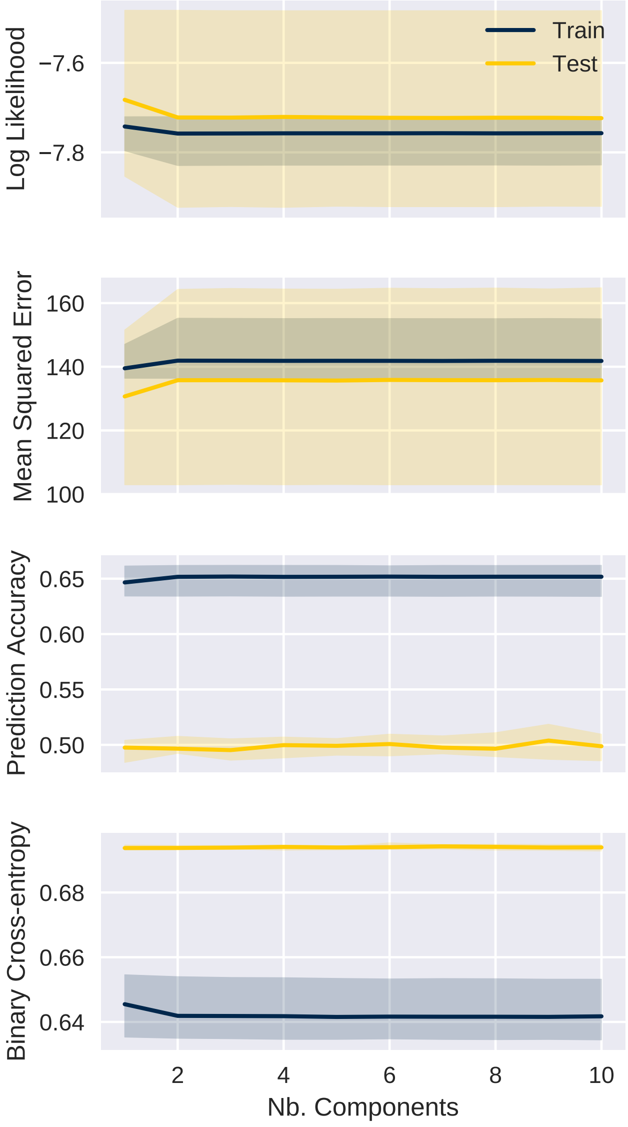

Fig. 8. Training and testing metrics under variational inference for varying latent dimension K. Lines represent the median across the 14 seasons; bands represent the minimum and maximum range across the 14 seasons.

The model proposed by [1] assumes constant skills across years; our model assumes independent skills across years. A modeling choice that would assign different but related skills across years would be to set a constraint between consecutive years. In particular, we could the skills follow a random walk, e.g.,

$$T^o_t\left[y+1\right] \mid T^o_t\left[y\right] \sim \mathcal{N}_K\left(T^o_t\left[y\right], \tau^2 I_K\right).$$

In the MLE inference scheme, this would simply add a regularization term

$$\left\| T^o_t\left[y+1\right] - T^o_t\left[y\right] \right\|_2^2$$

to the objective function.

In additional to scores, the dataset [4] contains more detailed match statistics, known as the *box score*. For example, we have access to the number of rebounds, turnovers and blocks for each team. These finer statistics are often associated with specific offensive or defensive play styles and are also relatively associated with the score. Hence, using the box score to improve our estimation of the latent variables could serve two purposes. First, if these statistics are related to performance, then the estimated latent skills could be more accurate. Second, these statistics could add meaning to the latent variables. As an example, some offensive component could be associated with getting more rebounds as well as the score. A potential way to induce such meaning would be to start from the skill difference

$$S^o_{t_i(m)} - S^d_{t_{1-i}(m)}$$

and suppose each team statistics is, say, Gaussian with mean equal to a linear combination of that difference. Further imposing sparsity in the linear combination could induce a meaning to the skills. In a similar fashion, we could perform correlation analysis between our estimated latent skills and the box score to label the latent components.

References

[1] F. J. R. Ruiz and F. Perez-Cruz, "A generative model for predicting outcomes in college basketball," *Journal of Quantitative Analysis in Sports*, vol. 11, no. 1, pp. 39–52, 2015. [Online]. Available: https://www.degruyter.com/view/journals/jqas/11/1/article-p39.xml

[2] G. Baio and M. Blangiardo, "Bayesian hierarchical model for the prediction of football results," *Journal of Applied Statistics*, vol. 37, no. 2, pp. 253–264, 2010.

[3] S. Chen and T. Joachims, "Modeling intransitivity in matchup and comparison data," in *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, ser. WSDM '16. New York, NY, USA: Association for Computing Machinery, 2016, pp. 227–236. [Online]. Available: https://doi-org.proxy.lib.umich.edu/10.1145/2835776.2835787

[4] Kaggle. (2018) Google Cloud & NCAA® ML Competition 2018-Men's. [Online]. Available: https://www.kaggle.com/c/mens-machine-learning-competition-2018

[5] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[6] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[7] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[8] J. Boyd. Best home court advantage in college basketball by points. [Online]. Available: https://www.boydsbets.com/college-basketball-home-court-advantage/

[9] ESPN. (2014) Men's college basketball rankings - ap top 25 postseason. [Online]. Available: https://www.espn.com/mens-college-basketball/rankings/_/year/2014/poll/1/week/1/seasontype/3