Expectation Propagation for Rating Players in Sports Competitions

Adriana Birlutiu and Tom Heskes

Institute for Computing and Information Sciences, Radboud University Nijmegen
Toernooiveld 1, 6525 ED Nijmegen, The Netherlands
{adrianab, tomh}@cs.ru.nl

Abstract. Rating players in sports competitions based on game results is one example of paired comparison data analysis. Since an exact Bayesian treatment is intractable, several techniques for approximate inference have been proposed in the literature. In this paper we compare several variants of expectation propagation (EP). EP generalizes assumed density filtering (ADF) by iteratively improving the approximations that are made in the filtering step of ADF. Furthermore, we distinguish between two variants of EP: EP-Correlated, which takes into account the correlations between the strengths of the players and EP-Independent, which ignores those correlations. We evaluate the different approaches on a large tennis dataset to find that EP does significantly better than ADF (iterative improvement indeed helps) and EP-Correlated does significantly better than EP-Independent (correlations do matter).

1 Introduction

Our goal is to develop and evaluate methods for the analysis of paired comparison data. In this paper we illustrate such methods by rating players in sports, in particular in tennis.

We consider the player's strength as a probabilistic variable in a Bayesian framework. Before taking into account the match outcomes, information available about the players can be incorporated in a prior distribution. Using Bayes' rule we compute the posterior distribution over the players' strengths. We take the mean of the posterior distribution as our best estimate of the players' strengths and the covariance matrix as the uncertainty about our estimation.

An exact Bayesian treatment is intractable, even for a small number of players; the posterior distribution cannot be evaluated analytically, and therefore we need approximations for it. Expectation propagation [7] is a popular approximation technique. We will use it in this paper for approximating the posterior distribution over the players' strengths. The question that we want to answer here is: how do different variants of expectation propagation perform for this setting? In particular, does it make sense to perform backward and forward iterations for the approximations and does it help to have a more complicated (full) covariance structure?

The paper is structured as follows: in the next section we introduce the probabilistic framework used to estimate players' strengths; in Section 3 we present algorithms for approximate inference and the way they apply to our setting; in Section 4 we show experimental results for real data, which we use to compare the performance of the algorithms; and in the last section we draw the conclusions.

2 Probabilistic Framework to Estimate Players' Strengths

Let $\boldsymbol{\theta}$ be an n_{players}-dimensional probabilistic variable whose components represent the players' strengths. We define $r_{ij} = 1$ if player i beats player j, and $r_{ij} = -1$ otherwise. For the probability of r_{ij} as a function of the strengths θ_i and θ_j, we take the Bradley-Terry model [2]:

$$p(r_{ij}|\theta_i, \theta_j) = \frac{1}{1 + \exp[-r_{ij}(\theta_i - \theta_j)]} \,. \tag{1}$$

A straightforward method to approximate the players' strengths is to build the likelihood of $\boldsymbol{\theta}$ given R; where R stands for the outcomes of all played matches. We take the maximum of the likelihood as the estimate for the strengths of the players.

The maximum likelihood approach gives a point estimate, the Bayesian approach, on the other hand, yields a whole distribution over the players' strengths. Furthermore, useful sources of information, like results in previous competitions and additional information about the players, can be incorporated in a prior distribution over the strengths. Using Bayes' rule we compute the posterior distribution over the players' strengths:

$$p(\boldsymbol{\theta}|R) = \frac{1}{d} p(R|\boldsymbol{\theta}) p(\boldsymbol{\theta}) = \frac{1}{d} p(\boldsymbol{\theta}) \prod_{i \neq j} p(r_{ij}|\theta_i, \theta_j) \,, \tag{2}$$

where $p(\boldsymbol{\theta})$ is the prior, $p(r_{ij}|\theta_i, \theta_j)$ from (1), and d is a normalization constant.

We take the mean or the mode of the posterior as the best estimate for the players' strengths. While computing the mean of the posterior distribution is computationally intractable, its mode (MAP) can be determined using optimization algorithms. For the MAP estimate the computation time is linear in the number of matches, and the number of iterations needed to obtain convergence. Typically, the number of iterations needed scales linearly with the number of players with a state-of-the-art optimization method such as conjugate gradient.

For making predictions and estimating the confidence of these predictions, we need the whole posterior distribution over the players' strengths. The posterior obtained using Bayes' rule in equation (2) cannot be evaluated analytically, hence we need to make approximations for it. For this task, sampling methods are very costly because of the high-dimensionality of the sampling space: the dimension is equal to the number of players. Therefore, for rating players, we here focus on deterministic approximation techniques, in particular expectation propagation and variants of it.

3 Expectation Propagation

Expectation propagation (EP) [7] is an approximation technique which tunes the parameter of a simpler approximate distribution, to match the exact posterior distribution of the model parameters given the data.

Assumed Density Filtering ADF is an approximation technique in which the terms of the posterior distribution are added one at a time, and in each step the result of the inclusion is projected back into the assumed density. As the assumed density we take the Gaussian, to which we will refer below as q.

The first term which is included is the prior, $q(\boldsymbol{\theta}) = p(\boldsymbol{\theta})$; then we add terms one at a time $\tilde{p}(\boldsymbol{\theta}) = \Psi_{ij}(\theta_i, \theta_j)q(\boldsymbol{\theta})$, where $\Psi_{ij}(\theta_i, \theta_j) = p(r_{ij}|\theta_i, \theta_j)$; and at each step we approximate the resulting distribution as closely as possible by a Gaussian $q^{\text{new}}(\boldsymbol{\theta}) = \text{Project}\{\tilde{p}(\boldsymbol{\theta})\}$. Using the Kullback-Leibler (KL) divergence as the measure between the non-Gaussian $\tilde{p}$ and the Gaussian approximation, projection becomes moment matching: the result q^{new} of the projection is the Gaussian that has the first two moments, mean and covariance, the same as $\tilde{p}$.

After we add a term and project, the Gaussian approximation changes. We call the quotient between the new and old Gaussian approximation a *term approximation*.

Iterative Improvement EP generalizes ADF by performing backward-forward iterations to refine the term approximations until convergence. The final approximation will be independent of the order of incorporating the terms. The algorithm performs the following steps.

1. Initialize the term approximations $\tilde{\Psi}_{ij}(\theta_i, \theta_j)$, e.g., by performing ADF; and compute the initial approximation
$$q(\boldsymbol{\theta}) = p(\boldsymbol{\theta}) \prod_{i \neq j} \tilde{\Psi}_{ij}(\theta_i, \theta_j) \,.$$
2. Repeat until all $\tilde{\Psi}_{ij}$ converge:
   (a) Remove a term approximation $\tilde{\Psi}_{ij}$ from the approximation, yielding
   $$q^{\setminus ij}(\boldsymbol{\theta}) = \frac{q(\boldsymbol{\theta})}{\tilde{\Psi}_{ij}(\theta_i, \theta_j)} \,.$$
   (b) Combine $q^{\setminus ij}(\boldsymbol{\theta})$ with the exact factor $\Psi_{ij} = p(r_{ij}|\theta_i, \theta_j)$ to obtain
   $$\tilde{p}(\boldsymbol{\theta}) = \Psi_{ij}(\theta_i, \theta_j) q^{\setminus ij}(\boldsymbol{\theta}) \,. \tag{3}$$
   (c) Project $\tilde{p}(\boldsymbol{\theta})$ into the approximation family
   $$q^{\text{new}}(\boldsymbol{\theta}) = \underset{q \in Q}{\text{argmin}}\, KL[\tilde{p}||q] \,.$$
   (d) Recompute the term approximation through the division
   $$\tilde{\Psi}_{ij}^{\text{new}}(\theta_i, \theta_j) = \frac{q^{\text{new}}(\boldsymbol{\theta})}{q^{\setminus ij}(\boldsymbol{\theta})} \,.$$

Computational Complexity When minimizing the KL divergence in step (c) we can take advantage of the locality property of EP [10]. From equation (3), because the term Ψ_{ij} does not depend on $\boldsymbol{\theta}^{\setminus ij}$, we can rewrite $\tilde{p}$ as:

$$\tilde{p}(\boldsymbol{\theta}) = \tilde{p}(\boldsymbol{\theta}_{\setminus ij}|\theta_i, \theta_j)\tilde{p}(\theta_i, \theta_j) = \tilde{p}(\theta_i, \theta_j)q^{\setminus ij}(\boldsymbol{\theta}_{\setminus ij}|\theta_i, \theta_j)\,.$$

Furthermore we obtain:

$$\begin{aligned} KL[\tilde{p}(\boldsymbol{\theta})||q(\boldsymbol{\theta})] = {} & KL[\tilde{p}(\theta_i, \theta_j)||q(\theta_i, \theta_j)] \\ & + E_{\tilde{p}(\theta_i,\theta_j)}[KL[q^{\setminus ij}(\boldsymbol{\theta}_{\setminus ij}|\theta_i, \theta_j)||q(\boldsymbol{\theta}_{\setminus ij}|\theta_i, \theta_j)]]\,. \end{aligned} \tag{4}$$

The two terms on the right-hand side can be minimized independently. Minimization of the second term gives:

$$q^{\text{new}}(\boldsymbol{\theta}_{\setminus ij}|\theta_i, \theta_j) = q^{\setminus ij}(\boldsymbol{\theta}_{\setminus ij}|\theta_i, \theta_j)\,. \tag{5}$$

Minimizing the KL divergence for the first term in the right-hand side in (4) reduces to matching the moments, mean and covariance, between the 2-dimensional distributions $\tilde{p}(\theta_i, \theta_j)$ and $q(\theta_i, \theta_j)$.

Exploiting this locality property, we managed to go from n_{players}-dimensional integrals to 2-dimensional integrals, which can be further reduced to 1 dimension, by rewriting them in the following way (see e.g., the appendix of [1]):

$$\langle \Psi(\theta_i, \theta_j)\rangle_{\mathcal{N}(\boldsymbol{m},\boldsymbol{C})} = \langle F(\boldsymbol{a}\boldsymbol{\theta}_{ij})\rangle_{\mathcal{N}(\boldsymbol{m},\boldsymbol{C})} = \langle F(\theta\sqrt{\boldsymbol{a}^T\boldsymbol{C}\boldsymbol{a}} + \boldsymbol{a}^T\boldsymbol{m})\rangle_{\mathcal{N}(0,1)}$$

where $\boldsymbol{a}$ is the vector $[-1,\ 1]$ if player i is the winner, or $\boldsymbol{a} = [1,\ -1]$ if player j is the winner, $\boldsymbol{\theta}_{ij} = [\theta_i,\ \theta_j]$, F is defined through equation (1), and $\mathcal{N}(\boldsymbol{m}, \boldsymbol{C})$ stands for a Gaussian with mean $\boldsymbol{m}$ and covariance matrix $\boldsymbol{C}$. Substituting the solution (5), we see that the term approximation, in step (d) of the algorithm, indeed only depends on θ_i and θ_j.

We can simplify the computations by using the canonical form of the Gaussian distribution. Because, when projecting, we need the moment form of the distribution, we go back and forth between distributions in terms of moments and in terms of canonical parameters. For a Gaussian, this requires computing the inverse of the covariance matrix, which is of the order n^3_{players}. Since the covariance matrix, when refining the term corresponding to the game between players i and j, changes only for the elements corresponding to players i and j, we can use the Woodbury formula [8] to reduce the cubic complexity of the matrix inversion to a quadratic one. Thus, the complexity of EP is:

$$\mathcal{C}(\text{EP}) = \mathcal{O}(n_{\text{iterations}} \times n^2_{\text{players}} \times n_{\text{matches}})$$

where $n_{\text{iterations}}$ is the number of iterations back and forth in refining the term approximations. In practice, the number of iterations to converge seems largely independent of the number of players or matches. In our experiments, we needed $n_{\text{iterations}} \approx 5$ to converge.

We will refer to this version of EP as EP-Correlated: by projecting into a non-factorized Gaussian, it takes into account the correlations between the players' strengths.

EP-Independent The complexity of the EP algorithm can be reduced further if we keep track only of the diagonal elements of the covariance matrix, ignoring the correlations. The matrix inversion has in this case linear complexity. The algorithm is faster and requires less memory.

4 Experiments

We applied the approximation algorithms, presented in the previous section, to the analysis of a real dataset. The dataset consists of results of 38538 tennis matches played on ATP events among 1139 players between 1995 and 2006. The goal was to compute ratings for the players based on the match outcomes. The methods described yield a Gaussian distribution of the players' strengths; the mean of the distribution represents our estimate of the players' strengths, the rating, and the variance relates to the uncertainty. Furthermore, we predict results of future games, and estimate the confidence of our predictions. We take as the prior a Gaussian distribution with mean zero and covariance equal to the identity matrix.

Figure 1 shows the empirical distribution of the players' strengths (means of the posterior distribution) in comparison with the average width of the posterior for an individual player. It can be seen that the uncertainty for individual players is comparable to the diversity between players.

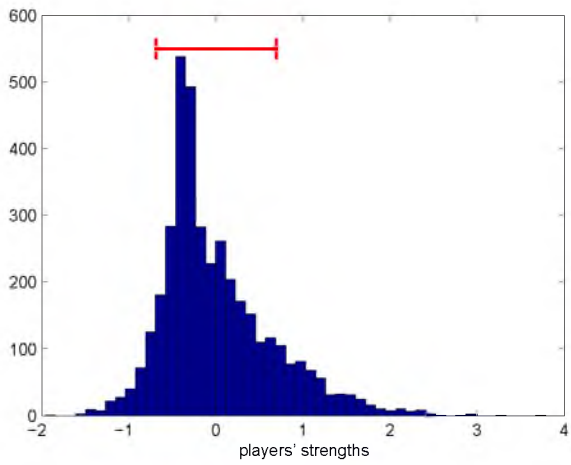

Fig. 1. A histogram of the players' strengths (means of the posterior distribution) for all years. The bar indicates the average width of the posterior distribution for each of the individual players. The results shown are for EP-Correlated.

4.1 Accuracy

We computed the ratings for the players at the end of each year, based on the matches from that year. Furthermore, based on these ratings we made predictions for matches in the next year: in a match we predicted the player with the highest rating to win.

EP-Correlated versus ADF We compared the accuracy of the predictions based on EP-Correlated ratings with the ones based on ADF ratings. We divided all joint predictions into 4 categories as shown in Table 1. We applied a binomial test on the matches for which the two algorithms gave different predictions to check the significance of the difference in performance [9]. The p-value obtained for this one-sided binomial test is 3×10^{-14}, which indicates that the difference is highly significant: EP-Correlated performs significantly better than ADF.

EP-Correlated versus EP-Independent The same type of comparison was performed between EP-Correlated and EP-Independent, the results are shown in Table 1. As for the previous comparison, the p-value is very small, 3×10^{-7}: the binomial test suggests that the difference between the two algorithms is again highly significant.

Table 1. Comparison between EP-Correlated, ADF and EP-Independent based on the number of matches correctly/incorrectly predicted.

	ADF		EP-Independent	
	correct	incorrect	correct	incorrect
EP-Correlated				
correct	16636 (54.48%)	2395 (7.81%)	17857 (58.46%)	1174 (3.83%)
incorrect	1902 (6.21%)	9620 (31.50%)	945 (3.09%)	10577 (34.62%)

EP-Correlated versus Laplace and ATP Rating We compared Laplace and EP-Correlated to find out that EP-Correlated does slightly, but not significantly better (p-value is 0.3). They disagree on only 0.2% of all matches.

We also compared the accuracy of the predictions based on the EP ratings with the accuracy of the predictions obtained using the ATP ratings at the end of the year. The ATP rating system gives points to players according to the type of the tournament and how far in the tournament they reached. Averaged over all the years, both EP and ATP ratings, give similar accuracy of predictions for the next, about 62%.

4.2 Confidence

With a posterior probability over the players' strengths we can compute the confidence of the predictions.

The algorithms presented perform about the same in estimating the confidence. However, they all tend to be overconfident, in the sense that the actual fraction of correctly predicted matches is smaller than the predicted confidence, as indicated by the solid line in the left plot of Figure 2. We can correct this by adding noise to the players' strengths, to account for the fact that a player's strength changes over time:

$$\theta_{t+1} = \theta_t + \epsilon$$

where ϵ has mean zero and variance σ^2. To evaluate the confidence estimation, we plot on the right side of Figure 2 the Brier score [3] for different values of σ. The optimum is obtained for $\sigma = 1.4$, which then yields the dashed line in the left plot of Figure 2.

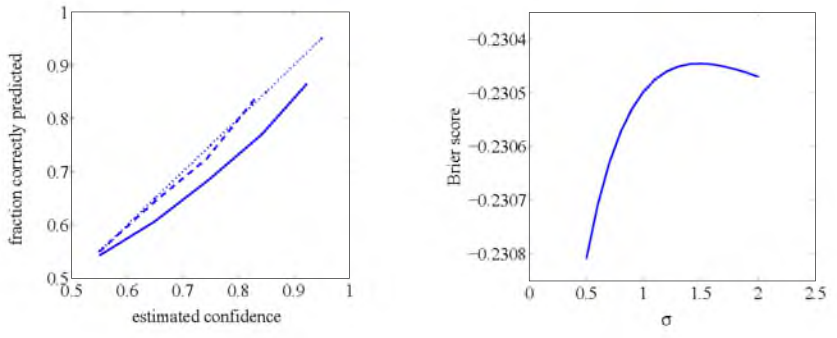

Fig. 2. Left: the actual fraction of correctly predicted matches as a function of the predicted confidence; without added noise (solid line) and with noise of standard deviation 1.4 added (dashed line); the dotted line represents the ideal case and is drawn for reference. Right: the Brier score for the confidence of the predictions as a function of the standard deviation of the noise added to each player's strength.

5 Conclusions

Based on the experimental results reported in this study we draw the conclusion that EP-Correlated performs better in doing predictions for this type of dataset than its modified versions, ADF and EP-Independent. Further experiments should reveal whether this also applies to other types of data.

Our results are generalizable to more complex models, e.g. including dynamics over time, which means that a players rating in the present is related to his performance in the past [4]; and team effects: a player's rating is inferred from

team performance [5, 6]. Specifically for tennis, the more complex models should also incorporate the effect of surface because the performance of tennis players in a match is influenced by the type of surface they play on (grass, clay, hard court, indoor). In this paper we considered the most basic probabilistic rating model; this model performs as good as the ATP ranking system. We would expect that the more complex models could outperform ATP.

Acknowledgments

The statistical information contained in the tennis dataset has been provided by and is being reproduced with the permission of ATP Tour, Inc., who is the sole copyright owner of such information. We would like to thank Franc Klaassen for pointing us in the right direction. This research is supported by the Dutch Technology Foundation STW, applied science division of NWO and the Technology Program of the Ministry of Economic Affairs.

References

1. D. Barber and C. Bishop. Ensemble learning in Bayesian neural networks. *Neural Networks and Machine Learning*, 1998.
2. R.A. Bradley and M.E. Terry. Rank analysis of incomplete block designs: I, the method of paired comparisons. *Biometrika*, 1952.
3. Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 1950.
4. Mark Glickman. *Paired Comparison Models with Time Varying Parameters*. PhD thesis, Harvard University, 1993.
5. Ralf Herbrich, Tom Minka, and Thore Graepel. TrueSkill: A Bayesian skill rating system. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 569–576. MIT Press, Cambridge, MA, 2007.
6. Tzu-Kuo Huang, Chih-Jen Lin, and Ruby C. Weng. A generalized Bradley-Terry model: From group competition to individual skill. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 601–608. MIT Press, Cambridge, MA, 2005.
7. T. P. Minka. *A Family of Algorithms for Approximate Bayesian Inference*. PhD thesis, M.I.T., 2001.
8. William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, 1992.
9. Steven L. Salzberg. On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery*, 1(3):317–328, 1997.
10. M. Seeger. Notes on Minka's expectation propagation for Gaussian process classification. Technical report, University of Edinburgh, 2002.