Penalty Specialists Among Goalkeepers: A Nonparametric Bayesian Analysis of 44 Years of German Bundesliga

Björn Bornkamp, Arno Fritsch, Oliver Kuss, and Katja Ickstadt

Abstract Penalty saving abilities are of major importance for a goalkeeper in modern football. However, statistical investigations of the performance of individual goalkeepers in penalties, leading to a ranking or a clustering of the keepers, are rare in the scientific literature. In this paper we will perform such an analysis based on all penalties in the German Bundesliga from 1963 to 2007. A challenge when analyzing such a data set is the fact that the counts of penalties for the different goalkeepers are highly imbalanced, leading to the question on how to compare goalkeepers who were involved in a disparate number of penalties. We will approach this issue by using Bayesian hierarchical random effects models. These models shrink the individual goalkeepers estimates towards an overall estimate with the degree of shrinkage depending on the amount of information that is available for each goalkeeper. The underlying random effects distribution will be modelled nonparametrically based on the Dirichlet process. Proceeding this way relaxes the assumptions underlying parametric random effect models and additionally allows to find clusters among the goalkeepers.

1 Introduction

In modern football, penalty shots are of vital importance. The world cup finals in 1990, 1994, and 2006, for example, were all decided by penalties. Nevertheless, scientific investigations of penalty conversions or savings are rare. Shooting techniques and tactics, ball speed, anticipation of the keeper, stress management of the shooter, or empirical investigation of penalty myths have been the objects of investigation ([8, 12, 16, 15, 13, 21, 9]). However, we are not aware of studies which try to find rankings or clusters of successful penalty scorers or savers.

Björn Bornkamp
Fakultät Statistik, Technische Universität Dortmund, D-44221 Dortmund, Germany
bornkamp@statistik.tu-dortmund.de

This is astonishing as the perception of especially skilled goalkeepers seems to be commonplace. For example, the English Wikipedia page for 'Penalty kick' (http://en.wikipedia.org/wiki/Penalty_kick, accessed 15/04/2008) shows a list of goalkeepers (Carlo Cudicini, Peter Schmeichel, Pepe Reina, Oliver Kahn, Ricardo, Francesco Toldo, Brad Friedel, Artur Boruc, Jens Lehmann, Edwin van der Sar and Mark Schwarzer) who are 'noted for their penalty-saving capabilities', but there is no quantitative evidence to support the claim of existence of a group of 'penalty specialists'. The German Wikipedia page on the penalty (http://de.wikipedia.org/wiki/Elfmeter, accessed 15/04/2008) asserts that there are some goalkeepers being able to save more penalties than the average goalkeeper and gives a ranking of the German goalkeepers with the largest number of saved penalties. It is interesting from a statistical viewpoint that this ranking contains only the absolute number of saved penalties, not accounting for the number of potentially savable penalties for the respective goalkeeper.

In this paper we approach the problem of ranking and clustering goalkeepers for their penalty-saving capabilities in a statistically more valid way. Our data set includes all 3,768 penalties from August 1963 to May 2007 from the German Bundesliga. Data were collected from three different sources. All penalties from August 1963 to May 1993 were taken from [7]. Penalties from August 1993 to February 2005 were made available by IMP AG, München (www.impire.de), a German company that collects data for a commercial football data base. The remaining penalties were found by a systematic internet search, their completeness was checked via the aggregated data published by the "kicker" (the leading German football magazine) in its annual review of the Bundesliga season. As we are focusing on the goalkeeper's ability to save penalties, we removed all penalties that missed the goal or hit goal-post or crossbar. This resulted in 261 deletions with 3,507 penalties remaining for final analysis. Out of these 3,507 penalties 714 were saved by the goalkeeper corresponding to a rate of 20.4%. The following additional information was available for each penalty: goalkeeper, goalkeeper's team, scorer, scorer's team, experience of goalkeeper and scorer (in terms of penalties), home advantage, day and year of season, and, of course, successful conversion or saving of the penalty. In total 273 goalkeepers were involved in the 3,507 penalties, many of them having been faced only with a small number of penalties (94 were involved in three or less penalties, see also Fig. 1 (i)). Figure 1 (ii) shows the relative frequencies of saved penalties for all goalkeepers. The modes of the density at 0 and 1 are due to the goalkeepers that were involved in very few penalties and saved none or all. It is intuitively clear that a goalkeeper who was involved in only one single penalty during his career and saved this, should not be considered the best penalty saver despite his 100% saving rate. Consequently, the relative frequency of saved penalties is a bad estimator of the "true" ability of the goalkeeper, motivating the use of more sophisticated statistical procedures.

That is, we are faced with two main statistical challenges:

(i) How to derive a sound statistical model, which will produce more "reasonable" estimates for the goalkeepers effect than simple relative frequencies?

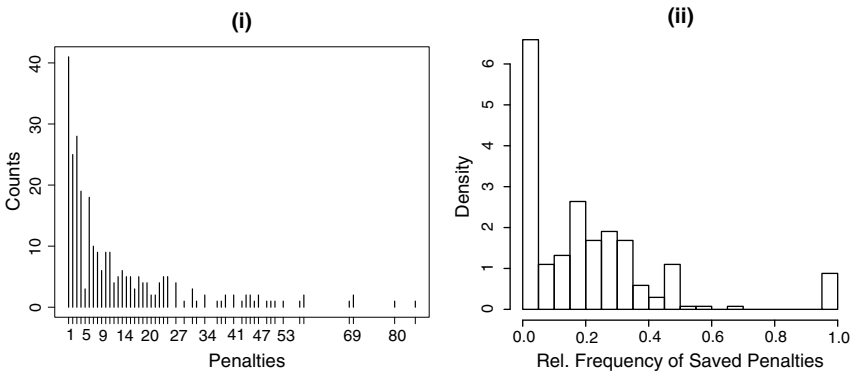

Fig. 1 (i) Counts of penalties per goalkeeper and (ii) histogram of relative frequencies of saved penalties per goalkeeper

(ii) How to investigate whether the population of goalkeepers can be grouped into clusters containing, for example, a group of 'penalty specialists' and a group of 'penalty losers'?

In Section 2 we will introduce the statistical methods, which will allow us to approach (i) and (ii), while Section 3 is devoted to the analysis of the data. Final conclusions will be drawn in Section 4.

2 Statistical Methodology: Hierarchical Models and Bayesian Nonparametrics

In the first two parts of this section we will describe the statistical methodology, while the third part deals with the actual model and priors we will use for the analysis in Section 3. The material in this section is mainly based on [4], who provides a recent review of nonparametric modeling of random effects distributions in Bayesian hierarchical models, and [14], who also illustrate how to implement a related model in BUGS.

2.1 Hierarchical Models

An appropriate tool to approach problem (i) from a statistical perspective is the hierarchical model. In its most simple form it can be described as follows: Suppose we observe a normally distributed random variable Y_i once for each of n subjects. The model for the data would then be

$$Y_i = \mu_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ and μ_i some unknown *constant* effect of the ith subject. In a classical model, the maximum likelihood estimate for each of the subject effects μ_i would equal y_i. Typically, this will lead to highly variable estimates for the subject's effect as there are as many parameters as observations. However, if it is known (or at least reasonable to assume) that the subjects belong to the same population, a different approach is more appropriate. In this case one would model the subject effects μ_i as realizations from an unknown population (i.e., random effects) distribution P, rather than as unrelated constants. Consequently, in this model all realizations y_i are used in estimating the random effects distribution P, which in turn leads to estimates of the individual effects μ_i. However, these would be shrunken towards each other. That is, hierarchical models allow for sharing information across subjects, rather than treating subjects as completely unrelated. A Bayesian analysis with an implementation via Gibbs and general Markov chain Monte Carlo (MCMC) sampling is particularly suited for the analysis of more complex hierarchical models (while the standard frequentist approaches become infeasible). Such a Bayesian approach is taken in this article.

2.2 The Dirichlet Process

Reformulating question (ii) statistically we would like to investigate, whether the random effects distribution P of the goalkeeper effects is multimodal. Figure 1 (ii) suggests that this might be the case, even when ignoring the modes at 0 and 1. For this reason we base the analysis in this article on Bayesian nonparametric methodologies, as they allow to model a multimodal random effects distribution. Specifically, we will model the random effects distribution P as a (location) mixture of normal distributions and assume a nonparametric prior for the mixing distribution. The motivation for using mixtures of normal distributions stems from the fact that any distribution on the real line can be approximated arbitrarily well by a mixture of normals ([2]). We hence model the density of the random effects distribution P as $\int N(x|\theta, \sigma^2) Q(d\theta)$, where $N(.|\theta, \sigma^2)$ is a normal density with mean θ and variance σ^2 and $Q(d\theta)$ is a discrete mixing distribution. The main issue in this kind of Bayesian analysis is which prior to assume for the unknown discrete mixing distribution $Q(d\theta)$. A flexible and convenient solution is to use the Dirichlet process, dating back to [5]. The Dirichlet process is a random discrete probability measure, i.e., a stochastic process that realizes discrete probability measures. It is characterized by two parameters: A base probability measure F_0 and a positive real number α. A random probability measure Q follows a Dirichlet process prior if the distribution of $(Q(S_1), \ldots, Q(S_k))'$ for a partition $S_1, \ldots, S_k$ of the underlying sample space (in our case $\mathbb{R}$) has a Dirichlet distribution with parameter $(\alpha F_0(S_1), \ldots, \alpha F_0(S_k))'$. Hence F_0 is the underlying prior mean distribution (i.e., $\mathbb{E}(Q(S_i)) = F_0(S_i)$), while α is a precision parameter (for $\alpha \to \infty$ the realizations will be more and more similar to F_0). The main reason for the popularity of the Dirichlet process for Bayesian nonparametric applications is the fact that it has an important conjugacy property,

which allows for an efficient exact implementation in many cases (see [5] for details). Another reason for the popularity of Dirichlet process priors is the constructive *stick-breaking* representation of the Dirichlet process given by [17]. Sethuraman showed that Q has a Dirichlet process prior with parameters α and F_0 iff

$$Q(d\theta) = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_h}(d\theta), \quad \text{with } \theta_h \overset{iid}{\sim} F_0, \tag{1}$$

where δ_θ denotes the probability measure degenerated at θ and $\pi_h = V_h \prod_{l<h}(1 - V_h)$ with $V_h \overset{iid}{\sim} \text{Beta}(1,\alpha)$. The terminology *stick-breaking* is used, because starting with a probability stick of length one, V_1 is the proportion of the stick broken off and allocated to θ_1, V_2 is the proportion of the remaining $1 - V_1$ stick length allocated to θ_2, and so on (see also [6] for details on the general class of stick-breaking priors). From this stick-breaking representation it becomes obvious that the precision parameter α also determines the clustering properties of the Dirichlet process. For small α, most probability mass will be distributed on the first realizations of F_0 leading to a clustering of observations. On the other hand for $\alpha \to \infty$ there will be many clusters and a specific realization of Q will be more similar to F_0. For a review of Bayesian clustering procedures, including those based on the Dirichlet process see, for example, [10]. For a random sample of size n from a probability distribution realized by a Dirichlet process [1] has shown that the prior density of the number of distinct values (clusters/components) k in n realizations is

$$p(k|\alpha,n) = c_n(k) n! \alpha^k \frac{\Gamma(\alpha)}{\Gamma(\alpha+n)} \quad , \tag{2}$$

where $c_n(k) = \frac{\left[{n \atop k}\right]}{\sum_{j=0}^{n} \left[{n \atop j}\right]}$, and $\left[{n \atop j}\right]$ denotes a Stirling number of the first kind (to approximate Stirling numbers for large n methods introduced by [20] can be used). The expected number of clusters k in n realizations is given by

$$E(k|\alpha,n) = \sum_{i=1}^{n} \frac{\alpha}{\alpha + i - 1}. \tag{3}$$

Both formulas play an important role for the prior elicitation of the parameter α.

The stick-breaking representation of the Dirichlet process is also useful because it directly leads to good finite dimensional approximations for the Dirichlet process by truncation of the sum (1). A finite dimensional approximation to the Dirichlet process is given by

$$Q(d\theta) = \sum_{h=1}^{N} \pi_h \delta_{\theta_h}(d\theta), \quad \text{with } \theta_h \overset{iid}{\sim} F_0$$

where $\pi_h = V_h \prod_{l<h}(1 - V_h)$, $V_h \overset{iid}{\sim} \text{Beta}(1,\alpha), h = 1,\ldots,N-1$, and $\pi_N = 1 - \sum_{h=1}^{N-1} \pi_h$. N is a truncation parameter, which is chosen large enough to obtain a good

approximation. For small values of α a relatively small N is sufficient to approximate the underlying Dirichlet process well. We refer to [14] for a detailed discussion of this aspect. In the following we will abbreviate the truncated prior distribution induced for the weights as $Seth_N(\alpha)$, i.e., $\boldsymbol{\pi} \sim Seth_N(\alpha)$.

2.3 Model and Prior Specification

We will model the jth observed penalty of the ith goalkeeper as a realization of a Bernoulli random variable with probability ρ_{ij} that the goalkeeper saves the penalty. This probability ρ_{ij} is hence modeled as a function of the ith goalkeeper and some additional covariates $\boldsymbol{x}_{ij}$. That is, we assume the model

$$\text{logit}(\rho_{ij}) = \gamma_i + \boldsymbol{\beta}' \boldsymbol{x}_{ij}, \quad i = 1,\ldots,273,\ j = 1,\ldots,n_i,$$

where γ_i is the random effect of the ith goalkeeper, n_i is the number of penalties the ith goalkeeper was involved in, and $\boldsymbol{\beta}$ is the vector of regression coefficients for the covariates.

The γ_i are modeled as iid realizations of a random effect distribution P, which in turn is modeled as a location mixture of normal distributions

$$\int N(x|\theta,\sigma^2) Q(d\theta) = \sum \pi_h N(x|\theta_h,\sigma^2),$$

and the Dirichlet process will be used as a prior for $Q(d\theta)$. The parameter α of the Dirichlet process will be chosen equal to $1/3$. Using (3) it can be seen that this leads to a prior mean of ≈ 2.91 clusters/components for a sample of size 273. Calculation of (2) shows (see also Fig. 2), that the prior density for the number of components has peaks at 2 and 3 and then decreases rapidly, leaving virtually no probability mass for $k > 8$, which seems reasonable for our penalty data. As the expected number of components is relatively small it is sufficient to select the truncation parameter N equal to 20. As the base measure F_0 of the Dirichlet process we will use a normal distribution with parameters 0 and variance 3.289. F_0 is chosen such that it is approximately equal to a uniform distribution on the probability scale. For the precision σ^{-2} of the normal densities in the mixture we will use an exponential prior distribution with mean 16. The prior distribution for $\boldsymbol{\beta}$, the coefficients of the covariates, are chosen as vague uniform distributions. A concise summary of the model and its different hierarchies is given in Table 1.

To assess the merit of a nonparametric model of the random effects distribution via the proposed Dirichlet process model, we compare it to two less flexible models via the deviance information criterion (DIC) [18]. The DIC is similar to AIC or BIC but more suitable for hierarchical models. Defining $\boldsymbol{\rho}$ as the vector containing the probabilities ρ_{ij} the deviance is in our case given by

Table 1 Hierarchical model used for analysis

Level	Parameters
I	$Y_{ij} \sim \text{Bernoulli}(\rho_{ij})$, $i = 1,\ldots,273$ $j = 1,\ldots,n_i$
II	$\text{logit}(\rho_{ij}) = \gamma_i + \boldsymbol{\beta}' \boldsymbol{x}_{ij}$
III	$\gamma_i \overset{iid}{\sim} \sum_{h=1}^{20} \pi_h f(x, \theta_h, \sigma^2)$, $i = 1,\ldots,273$ $\boldsymbol{\beta} \sim U([-10,10]^p)$
IV	$\sigma^{-2} \sim Exp(1/16), \boldsymbol{\pi} \sim Seth_{20}(\alpha = 1/3)$ $\theta_h \overset{iid}{\sim} \mathcal{N}(0, 3.289)$, $h = 1,\ldots,20$

$$D(\boldsymbol{\rho}|y) = -2\sum_{i=1}^{273}\sum_{j=1}^{n_i} y_{ij}\log(\rho_{ij}) + (1-y_{ij})\log(1-\rho_{ij}).$$

The DIC is then defined as $\overline{D(\boldsymbol{\rho}|y)} + p_D$, where $\overline{D(\boldsymbol{\rho}|y)}$ is the average deviance over the MCMC draws measuring the model fit and $p_D = \overline{D(\boldsymbol{\rho}|y)} - D(\bar{\boldsymbol{\rho}}|y)$ is an estimate of the "effective" number of parameters penalizing the model complexity ($\bar{\boldsymbol{\rho}}$ is the average of $\boldsymbol{\rho}$ over the MCMC iterations). For more details on the DIC we refer to [18]. The first model that we will use for comparison, is a model that does not allow for individual goalkeeper effects at all, leading to $\text{logit}(\rho_{ij}) = \mu_0 + \boldsymbol{\beta}' \boldsymbol{x}_{ij}$, with a fixed common intercept μ_0. Hence, by comparing this model with the Dirichlet process model in terms of the DIC we will be able to quantify the improvement of modeling individual goalkeeper effects. The second model we use for a comparison is a parametric normal random effects model, which can be obtained by setting $\gamma_i \overset{iid}{\sim} \mathcal{N}(\mu_0, \sigma_0^2)$ in level III of Table 1, and using suitable vague hyper-priors for μ_0 and σ_0^2 (here we use $\mu_0 \sim \mathcal{N}(0, 3.289)$ and $\sigma_0^2 \sim U([0,3])$). By comparing the Dirichlet process model with this parametric model we will be able to quantify the improvement of a nonparametric modeling of the random effects distribution. Subsequently the two restricted models will be referred to as 'Intercept' and 'Normal', our proposed model will be termed the 'Dirichlet' model.

2.4 Choice of Covariates

The main aim of this analysis is to model the goalkeepers effect on the probability of saving a penalty kick, but the effect of the scorer should also be taken into account. The logarithm of the number of taken penalties provides a good fit in an univariate logistic regression and is chosen to represent the penalty takers effect. For better interpretability the logarithm of base 2 is chosen. As home field advantage has an effect in many sports, the home field advantage of the goalkeeper is included as a binary covariate. To see whether there is a general time trend in the probability of saving a penalty, year is included as a covariate. "Year" here refers to a football season, which starts at the end of summer. A year effect could be due to improved

techniques for saving or taking a penalty. In addition the day of the season is included as a covariate to account for possible time trends within a season. For model fitting all covariates are scaled to lie between 0 and 1.

2.5 Computation

The models described in Section 2.3 are fit to the data using the OpenBUGS software version 2.10. Further analysis is done in R 2.6.2 using the interface provided by the R2WinBUGS package [19].

For each model the MCMC sampler is run with two independent chains with a burn-in of 50,000 iterations followed by 100,000 iterations of which every 20th is kept. Trace plots of parameters did not indicate problems with convergence of the chains and the results of the independent chains are similar. The results presented are based on the pooled draws of the independent chains, leading to a total number of 10,000 draws for each model.

3 Results

First the overall fit of the models is compared with the DIC criterion. Table 2 shows the DIC and its components for the three models considered. Both the Normal and the Dirichlet model improve on the model with only an intercept, indicating some gain with the inclusion of a random effects distribution. The improvement is not very large, indicating that the probability of saving a penalty does not vary too much between goalkeepers. As it is more flexible, the Dirichlet model has a lower average deviance than the Normal model but also a larger number of effective parameters leading to a DIC that is only slightly lower.

To answer the question whether there are distinct clusters of goalkeepers with differing abilities we compare the posterior distribution of the number of distinct components $p(k|y,\alpha,n)$ to the prior computed via (2). Barplots of both distributions are shown in Fig. 2 (i). One can see that the posterior puts less mass on a higher number of components than the prior, with one single component having the highest posterior probability. The posterior mean is 1.98 components compared to the prior mean 2.91. Observing the posterior expectation of the random effects distributions

Table 2 Average deviance, effective number of parameters and DIC for the different models

Model	$\overline{D(\rho\|y)}$	p_D	DIC
Intercept	3,453.8	5.0	3,458.8
Normal	3,422.5	31.1	3,453.6
Dirichlet	3,414.8	36.8	3,451.6

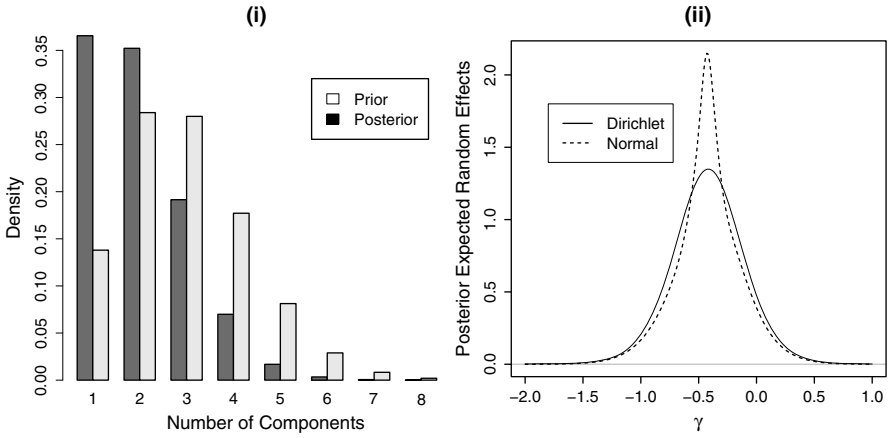

Fig. 2 (i) Distribution of the number of distinct components k for the Dirichlet model. (ii) Posterior expected random effects distribution γ for the Normal and Dirichlet model

shown in Fig. 2 (ii), there is no sign of multimodality. Thus there is not much support for different clusters in the data. In the Dirichlet model even for parameter draws with several distinct components, the resulting distribution tended to be unimodal (a mixture of normal distribution does not have to be multimodal). However, the more flexible Dirichlet model leads to a distribution with heavier tails than the one resulting from the Normal model.

Next we take a look at the estimates for the goalkeepers' probabilities to save a penalty that can be derived from the models. For this we consider

$$E\left(\frac{\exp(\gamma_i+\boldsymbol{\beta}'\boldsymbol{x}_{med})}{1+\exp(\gamma_i+\boldsymbol{\beta}'\boldsymbol{x}_{med})}\Big|y\right),\ i=1,\ldots,273, \tag{4}$$

the posterior expectation of the goalkeepers' probabilities to save a penalty kick when the covariates take their respective median values $\boldsymbol{x}_{med}$. The median values stand for a scorer with 10 taken penalties, the season 1983/84 and the 17th day of the season. The binary variable home field advantage is set to 0, representing no home field advantage for the goalkeeper. Figure 3 shows the posterior mean probabilities of the goalkeepers (from (4)) for all goalkeepers smoothed by a kernel density estimate. Comparing Fig. 3 (i) to the distribution of the relative frequencies in Fig. 1 (ii) it can be seen that the probabilities are considerably shrunken towards each other. The range of estimates is only about 0.1. Figure 3 (ii) shows a close-up look at the distribution in (i), and as for the random effects distribution it can be seen that the estimates of the Normal and Dirichlet model differ mainly in the tails, with the Dirichlet model leading to more pronounced tails.

Regarding the question of identifying the best and worst keepers, the tails of the distribution are of importance. As the Dirichlet model is more flexible in the tails it is used to determine a ranking of the keepers. In performing the ranking (see Table 3)

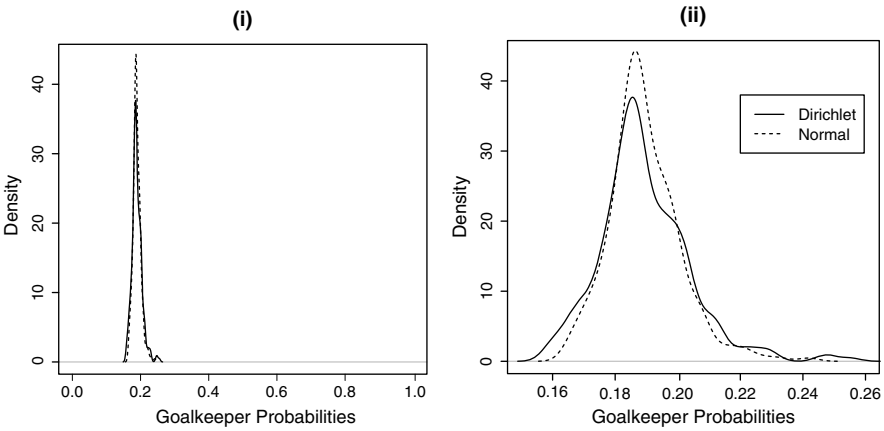

Fig. 3 Posterior expected probabilities of saving a penalty for the Normal and Dirichlet model; (i) on the range [0,1] and (ii) on the range [0.15, 0.26]

we rely on the recommendations of [11] who argue that ranking should be based on the posterior expectation of the rank rather than the posterior expected effect. This explains the fact that in some cases a goalkeeper with a higher rank nevertheless has a higher posterior expected probability of saving a penalty.

Several interesting observations arise from the ranking in Table 3. Goalkeepers estimated saving probabilities are not really different, with the best keeper having 25.5% and the worst keeper having 16.0%, yielding only a 10%-points difference. Moreover, the credible intervals for the saving probabilities are seen to be pretty large, credible intervals for the best and the worst keeper overlap considerably. As such, saving capabilities are rather similar across goalkeepers, reflecting the fact that no explicit clusters of goalkeepers could be found in our analysis. It is nevertheless surprising, that the two German goalkeepers who are thought to be penalty specialists (Oliver Kahn and Jens Lehmann) rank relatively low, indicating that both of them perform rather badly in penalty saving. This is probably due to the perception of the German expertise in penalty shoot-outs in recent tournaments, with Kahn and Lehmann playing prominent roles on these occasions. The degree of shrinking from the Dirichlet model is quite impressive. To demonstrate this, we consider Michael Melka and Gerhard Teupel as two representatives of the goalkeepers who were faced with only one single penalty during their career in the German Bundesliga. Michael Melka who saved this single penalty (thus having an observed 100% saving rate), has an estimated saving probability of only 20.2%. Gerhard Teupel, not saving this single penalty (resulting in an observed 0% saving rate) estimated saving probability is 18.6%, not very different from Melka's probability. Another peculiarity might be the fact that 3 goalkeepers of Bayern München (Manfred Müller, Walter Junghans, and Sepp Maier, having played 588 games or more that 17 seasons for the team altogether) are among the worst 5 penalty savers. This is in strict contrast to the fact that Bayern München is the most successful team in the German Bundesliga. It is

Table 3 Ranking of goalkeepers based on the average rank

Goalkeeper	Rank	$P(\text{Saving}\|y)$ with 95% CI	% Saved	# Saved	# Penalties
Kargus, Rudolf	1	0.255 [0.183, 0.354]	0.329	23	70
Enke, Robert	2	0.248 [0.162, 0.418]	0.500	9	18
Pfaff, Jean–Marie	3	0.247 [0.155, 0.483]	0.545	6	11
Köpke, Andreas	4	0.228 [0.159, 0.324]	0.317	13	41
Radenkovic, Petar	5	0.229 [0.158, 0.331]	0.353	12	34
⋮	⋮	⋮	⋮	⋮	⋮
Melka, Michael	54	0.202 [0.121, 0.317]	1.000	1	1
⋮	⋮	⋮	⋮	⋮	⋮
Teupel, Gerhard	154	0.186 [0.107, 0.285]	0.000	0	1
⋮	⋮	⋮	⋮	⋮	⋮
Kahn, Oliver	224	0.178 [0.120, 0.245]	0.172	10	58
⋮	⋮	⋮	⋮	⋮	⋮
Lehmann, Jens	228	0.178 [0.115, 0.248]	0.176	6	34
⋮	⋮	⋮	⋮	⋮	⋮
Schmadtke, Jörg	269	0.162 [0.098, 0.227]	0.098	4	41
Müller, Manfred	270	0.160 [0.082, 0.232]	0.042	1	24
Junghans, Walter	271	0.158 [0.083, 0.230]	0.042	1	24
Rynio, Jürgen	272	0.159 [0.088, 0.226]	0.074	2	27
Maier, Sepp	273	0.160 [0.104, 0.218]	0.130	9	69

Table 4 Estimated odds ratios with 95% credible intervals in the Dirichlet model. For the penalty taker odds ratio is for a scorer with twice the number of penalties. The odds ratio for year compares the last to the first year, which is also the case for day of the season

Covariate	OR with 95% CI
Scorer	0.754 [0.711, 0.798]
Home Field Advantage	0.956 [0.789, 1.145]
Year	0.894 [0.637, 1.222]
Day of Season	0.894 [0.674, 1.166]

also astonishing that Sepp Maier ranks the worst. After all, he was the goalkeeper of the German team winning the 1974 world cup, and is still the German goalkeeper with the most international matches (N = 95) up to now.

Finally, we consider the effects of the covariates. Since a logistic regression model is fitted, $\exp(\beta_k)$ can be interpreted as the change in the odds of the event, if the kth covariate is risen by 1. Table 4 shows the estimated odds ratios for the Dirichlet model. As the credible interval for the odds ratio of the scorer effect does not contain 1 there is strong evidence that a scorer that has taken more penalties reduces the goalkeeper's probability of saving the penalty. This is a reasonable result, since players that are known to be good penalty takers are probably chosen more often to take a penalty kick. As the scorer effect is given on the log2 scale, we can interpret the odds ratio as follows: Faced with a scorer that scored twice as

many penalties, the goalkeeper's odds of saving is multiplied by 0.754. For all the other covariates, 1 is clearly inside the credible interval. This implies that there is no evidence for a home field advantage for the goalkeeper. Additionally, evidence can neither be found for an overall time trend or a time trend within seasons. These conclusions are also obtained for the other two models.

4 Final Remarks and Outlook

In this article we analyzed the penalty saving abilities of goalkeepers in the first 44 years of the German Bundesliga. As is typical for such a data set, many goalkeepers were involved only in a few penalties. This poses the question on how to derive reasonable estimates for those keepers and how to compare keepers with a highly disparate number of penalties. We approached this issue by using Bayesian hierarchical models, i.e., the goalkeepers are modeled as realizations from a common random effects distribution P. This naturally allows for borrowing strength and hence shrinkage between the goalkeepers individual effect estimates. A major impetus for studying the data was to investigate whether there are certain groups of goalkeepers, such as 'penalty specialists' and 'penalty losers'. This motivated the use of Bayesian nonparametric approaches to model the random effects, as these techniques allow for modelling multimodal random effects distributions.

In the analyses we conducted in Section 3 we did not find any hint for multimodality. On the contrary, a-posteriori there was evidence that the number of components/clusters in the normal mixture model is even smaller than assumed a-priori (see Fig. 2 (i)). We also produced a ranking of the goalkeepers based on the average rank encountered during the MCMC runs. One observation is central: there is no strong evidence in the data that the different goalkeepers are highly different, for example, the credibility intervals for the goalkeeper ranking first (Rudolf Kargus) and last (Sepp Maier) overlap considerably.

From an application viewpoint it is somewhat surprising to see well-known goalkeepers like Sepp Maier ranking so low. This is a direct consequence of the shrinkage effect of the random effects model: As can be seen in Table 3, only goalkeepers who were involved in many penalties can rank at the top or the bottom of the list, while the goalkeepers with fewer penalties are all in the middle of the ranking. This is reasonable from a statistical point of view, as we can only make statistically accurate estimates for keepers with many penalties, while those with few penalties are shrunken towards the overall mean. This shrinkage effect should be kept in mind, when interpreting the ranking of goalkeepers from an application viewpoint. As can be seen in the tails of the random effects distribution and the estimated individual effects (Figs. 2 (ii) and 3 (ii)) the Dirichlet model already allows for a more realistic and flexible type of shrinkage than the normal model. There are however opportunities to model the random effects distribution even more flexible. The Dirichlet process may be replaced by another stochastic process, *e.g.*, a (normalized) α-stable process or the used normal kernel may be replaced by a t-density. Both approaches

would lead to even heavier tails of the random effects distribution and thus to a model representing less shrinkage.

In our analysis only the covariate we used as a substitute for the scorer effect seems to have an important effect. This motivates a further study, where the penalty scorer effect is also modeled by a random effects distribution instead of a simple fixed covariate. This might lead to a more realistic model and would allow for a ranking of the scorers as well. For the Dirichlet model a complication arises, however, if a second random effect is to be included. Then it is necessary to center the random effects distributions to have mean zero. Simply setting the mean of the base probability measure F_0 to zero is not sufficient to achieve zero mean of the random effects distribution, and more sophisticated procedures need to be applied such as the centered Dirichlet process [3], which we plan to do in future research.

Acknowledgement We are grateful to IMP AG, München, for providing parts of the underlying data set and Holger Rahlfs and Jörn Wendland of IMP AG for their kind cooperation. Mareike Kunze (IMEBI Halle) was very helpful in data entry, processing, and management. The work of Björn Bornkamp is supported by the Research Training Group "Statistical modelling" of the German Research Foundation (DFG).

References

[1] Antoniak, C.E.: Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. Ann. Stat. **2**, 1152–1174 (1974)

[2] Diaconis, P., Ylvisaker, D.: Quantifying prior opinion. In: Bernardo, J., DeGroot, M., Lindley, D., Smith, A. (eds.) Bayesian Statistics 2, pp. 133–156. Elsevier, Amsterdam (1985)

[3] Dunson, D.B., Yang, M., Baird, D.: Semiparametric Bayes hierarchical models with mean and variance constraints. Technical Report 2007-08, Department of Statistical Science, Duke University, Durham, NC, USA (2007)

[4] Dunson, D.B.: Nonparametric Bayes applications to biostatistics. Technical Report 2008-06, Department of Statistical Science, Duke University, Durham (2008)

[5] Ferguson, T.S.: A Bayesian analysis of some nonparametric problems. Ann. Stat. **1**, 209–230 (1973)

[6] Ishwaran, H., James, L.F.: Gibbs sampling methods for stick-breaking priors. J. Am. Stat. Assoc. **96**, 161–173 (2001)

[7] Kropp, M., Trapp, A.: 35 Jahre Bundesliga-Elfmeter 1963–1999. Agon Statistics 36, 2nd edn. Agon-Sportverlag, Kassel (1999)

[8] Kuhn, W.: Penaltykick strategies for shooters and goalkeepers. In: Reilly, K.D.T., Lees, A., Murphy, W. (eds.) Science and football, pp. 489–492. E & FN Spon, London (1988)

[9] Kuss, O., Kluttig, A., Stoll, O.: The fouled player should not take the penalty himself: An empirical investigation of an old German football myth. J. Sport. Sci. **25**, 963–967 (2007)
[10] Lau, J.W., Green, P.J.: Bayesian model-based clustering procedures. J. Comput. Graph. Stat. **16**, 526–558 (2007)
[11] Lin, R., Louis, T.A., Paddock, S.M., Ridgeway, G.: Loss function based ranking in two-stage hierarchical models. Bayesian Anal. **1**, 915–946 (2006)
[12] Loy, R.: Handlungsstrategien von Torhütern und Schützen in der Strafstoßsituation des Fußballsports. In: Bäumler, G., Bauer, G. (eds.) Sportwissenschaft rund um den Fußball, pp. 67–78. Schriften der Deutschen Vereinigung für Sportwissenschaft 96, Sankt Augustin (1988)
[13] Morya, E., Ranvaud, R., Pinheiro, W.: Dynamics of visual feedback in a laboratory simulation of a penalty kick. J. Sport. Sci. **21**, 87–95 (2003)
[14] Ohlssen, D.I., Sharples, L.D., Spiegelhalter, D.J.: Flexible random-effects models using Bayesian semi-parametric models: Applications to institutional comparisons. Stat. Med. **26**, 2088–2112 (2007)
[15] Savelsbergh, G., van der Kamp, J., Williams, A., Ward, P.: Anticipation and visual search behaviour in expert soccer goalkeepers. Ergonomics **48**, 1686–1697 (2005)
[16] Savelsbergh, G., Williams, A., van der Kamp, J., Ward, P.: Visual search, anticipation and expertise in soccer goalkeepers. J. Sport Sci. **20**, 279–287 (2002)
[17] Sethuraman, J.: A constructive definition of Dirichlet priors. Stat. Sin. **4**, 639–650 (1994)
[18] Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A.: Bayesian measures of model complexity and fit. J. Roy. Stat. Soc. B **64**, 583–639 (2002)
[19] Sturtz, S., Ligges, U., Gelman, A.: R2WinBUGS: A package for running WinBUGS from R. J. Stat. Softw. **12**, 1–16 (2005)
[20] Temme, N.M.: Asymptotic estimates of Stirling numbers. Stud. Appl. Math. **89**, 233–243 (1993)
[21] Van der Kamp, J.: A field simulation study of the effectiveness of penalty kick strategies in soccer: Late alterations of kick direction increase errors and reduce accuracy. J. Sport. Sci. **24**, 467–477 (2006)