ISSN: 2641-6336
Annals of
Biostatistics & Biometric Applications
DOI: 10.33552/ABBA.2023.05.000605
Iris Publishers

Research Article

Semi-Parametric Bayesian Estimation of Sparse Multinomial Probabilities with An Application to The Modelling of Bowling Performance in T20I Cricket

Lahiru Wickramasinghe[1*], Alexandere Leblanc[2] and Saman Muthukumarana[2]

[1]*Department of Mathematics and Statistics, University of Winnipeg, Winnipeg, Canada*

[2]*Department of Statistics, University of Manitoba, Winnipeg, Canada*

***Corresponding author:** Lahiru Wickramasinghe, Department of Mathematics and Statistics, University of Winnipeg, Winnipeg, Canada

Received Date: November 30, 2022

Published Date: January 23, 2023

Abstract

We consider modeling bowling performance in Twenty20 international cricket using a semi-parametric Bayesian approach. The bowling performance can be represented as a contingency table and typically yield a sparse contingency table due to cells with small counts and/or zeros. This sparsity is common in Twenty20 international cricket when we have many classification statuses with many levels, even when the sample size is large. Using a Dirichlet process in our proposed model, the multinomial probability vectors are supported on a discrete space, which enables the borrowing of information across data while providing a natural clustering mechanism. Another important feature of the approach is that this borrowing of information also allows the resulting estimators to handle sparsity, a common concern in multinomial data with many categories. The performance of the approach is compared against some of the standard methods available in the literature; James-Stein, empirical Bayes, and Bayesian multinomial regression estimation. To illustrate our modelling strategy, we suggest a simple way to assess the bowling performance of 175 world-class bowlers.

Keywords: James-Stein estimator; Empirical Bayes estimator; Dirichlet process; Multinomial regression; cricket; Sparse data

List of Abbreviations: MLE: Maximum Likelihood Estimator; ML: Maximum Likelihood; JS: James-Stein; EB: Empirical Bayes; MSE: Mean Squared Error; BMR: Bayesian Multinomial Regression; DP: Dirichlet Process; OP: Overall Proportion; ICC: International Cricket Council

Introduction

Categorical data are often analyzed using multinomial data, and sparseness in multinomial data is frequently encountered in practice when many cells have small and/or zero counts. Such sparse multinomial data can arise in two ways (1). relatively few observations are dispersed in numerous categories, or (2). cells that are structurally empty, i.e., theoretically impossible to observe. Assume, for instance, that K cells have probabilities $p_1, p_2, ..., p_K$ of occurring. Then, under the first scenario, many cells have a small probability pi relative to the number of observed outcomes leading to small or even zero counts. In this case, increasing the number of effective observations by combining different data sources could help to improve inference. It is the approach we take here. Under the second scenario, however, some cells have a probability $p_i = 0$ of occurrence. Identifying those structural zeros becomes a central part of the inference problem.

In this manuscript, we focus on the case of sparse multinomial data that are due to the observations dispersed in numerous categories or when the number of observations is small relative to the number of categories. The Maximum Likelihood Estimator (MLE) performs very poorly in this setting. Shrinkage estimation is an approach that allows for deriving improved estimators, and in particular, it can handle certain forms of sparsity by borrowing information from other multinomial populations. Our main objective is to develop an improved strategy to jointly estimate the cell probabilities of m multinomial populations in the context of sparse data.

The paper will proceed as follows. In Section 2, we discuss the previously studied statistical models, including James-Stein (JS) estimation, empirical Bayes (EB) estimation, and estimation based on Bayesian multinomial regression. In Section 3, we discuss the proposed statistical model, semi-parametric Bayesian estimation. In Section 4, we apply the methods presented in Sections 2 and 3 to data consisting of the bowling performance of 175 players over ten years. We also conducted a simulation study to compare the methods. We conclude with a short discussion in Section 5 based on the results and methods presented in the paper.

Standard Statistical Models and Estimation

In summarizing the bowling performance of a player in cricket, the total number of wickets taken by a bowler can be divided into K discrete categories (more details will be provided in Section 4). Specifically, assume we have m bowlers and let n_i be the total number of matches the i[th] bowler has played, and X_{ij} be the number of matches in which the i[th] bowler has taken exactly j−1 wickets, j = 1, ..., K. We can naturally model these categorical outcomes using the multinomial distribution as follows;

$$X_i = \left(X_{i1}, X_{i2}, \dots, X_{ik}\right) \sim Multinomial\left(n_i; p_i = \left(p_{i1}, p_{i2}, \dots, p_{ik}\right)\right)$$

given by $\widehat{p}_{ij^{MLE}} = \frac{X_{ij}}{n_i}$ for i = 1, 2, ..., m. Note that the cell counts in some categories will be either small or zero, which will result in a sparse table. In this case, the Maximum Likelihood Estimator (MLE) of p_{ij}, denoted by, $\widehat{p_{ij}}^{MLE}$ performs very poorly and underestimates the true cell probabilities (p_{ij}) due to sparsity. To help to mitigate this problem, the James-Stein approach can be used to estimate p_{ij}.

James-Stein estimation

The concept of shrinkage was first introduced in statistics by [1], and the general principles behind shrinkage estimation were discussed by [2]. The famous James-Stein shrinkage estimator was introduced by [3], which is based on a weighted average of two different models; a high-dimensional model with low bias and high variance and a lower-dimensional model with larger bias but smaller variance.

Suppose that the shrinkage target T is associated with the lower-dimensional model with smaller variance and considerable bias and that U is associated with the high-dimensional model with low bias and high variance. In shrinking, we try to find a compromise between T and U by computing a convex linear combination,

$$U^* = \lambda T + (1-\lambda)U$$

instead of choosing between one of the models. U * is called a shrinkage estimator, and it often outperforms the individual estimators T and U in terms of accuracy and statistical efficiency [4].

Here λ is usually a number between 0 and 1 and is called the shrinkage constant. It measures the weight that is given to the shrinkage target T. If λ = 1, the shrinkage estimate equals the shrinkage target T, whereas, for λ = 0, it equals U. This strategy has been used to estimate the cell probabilities of a multinomial distribution;

$$\widehat{p}_{ij^{JS}} = \lambda_i t_j + \left(1-\lambda_i\right)\widehat{p}_{ij^{MLE}}$$

where t_j is the shrinkage target. Note that $\widehat{p}_{ij^{JS}}$ is improved over MLE by combining the player-specific information given by the MLE with a target t_j that provides "global" information relevant to all populations. The default choice of $t_j = \frac{1}{k}$ is convenient but less than optimal in most cases. A popular shrinkage target (t_j) is the overall proportion of observations in category j. It seems appropriate to choose the population-specific shrinkage constant λ_i in a data-driven fashion by minimizing the mean squared error (MSE) of the resulting estimator. Assuming that the first two moments of the distributions of t_j exist, it can be shown that

$$E\left(\sum_{j=1}^{K}\left(\widehat{p}_{ij}^{JS} - p_{ij}\right)^2\right) = \sum_{j=1}^{K} MSE\left(\widehat{p}_{ij}^{MLE}\right) + \lambda_i^2 \sum_{j=1}^{K} E\left[\left(t_j - \widehat{p}_{ij}^{MLE}\right)^2\right]$$
$$-2\lambda_i \sum_{j=1}^{K}\left[Var\left(\widehat{p}_{ij}^{MLE}\right) - Cov\left(t_j, \widehat{p}_{ij}^{MLE}\right) + Bias\left(\widehat{p}_{ij}^{MLE}\right)\left(E\left[\widehat{p}_{ij}^{MLE} - t_j\right]\right)\right]. \quad (1)$$

Then, the optimal shrinkage constant λ_i^* can be obtained by analytically minimizing this function with respect to λ_i, leading to

$$\lambda_i^* = \frac{\sum_{j=1}^{K}\left[Var\left(\widehat{p}_{ij}^{MLE}\right) - Cov\left(t_j, \widehat{p}_{ij}^{MLE}\right) + E\left[t_j - \widehat{p}_{ij}^{MLE}\right] Bias\left(\widehat{p}_{ij}^{MLE}\right)\right]}{\sum_{j=1}^{K} E\left[\left(t_j - \widehat{p}_{ij}^{MLE}\right)^2\right]}.$$

Given that $\widehat{p}_{ij^{MLE}}$ is an unbiased estimator for p_{ij} and following Ledoit and Wolf (2003), we can further simplify the above expression to

$$\lambda_i^* = \frac{\sum_{j=1}^{k} Var\left(\widehat{p}_{ij}^{MLE}\right) - Cov\left(t_j, \widehat{p}_{ij}^{MLE}\right)}{\sum_{j=1}^{k} E\left[\left(t_j - \widehat{p}_{ij}^{MLE}\right)^2\right]}$$

which can be estimated using its sample counterpart given by;

$$\widehat{\lambda}_i^* = \frac{\sum_{j=1}^{k} \widehat{Var}\left(\widehat{p}_{ij}^{MLE}\right) - \widehat{Cov}\left(t_j, \widehat{p}_{ij}^{MLE}\right)}{\sum_{j=1}^{k}\left(t_j - \widehat{p}_{ij}^{MLE}\right)^2}.$$

Note that

$\widehat{Var}\left(\widehat{p}_{ij}^{MLE}\right) = \frac{\widehat{p}_{ij}^{MLE}\left(1-\widehat{p}_{ij}^{MLE}\right)}{n_i - 1}$ and $\widehat{Cov}\left(t_j, \widehat{p}_{ij}^{MLE}\right) = 0$ when $t_j = \frac{1}{k}$. In the case where $t_j = \frac{1}{n}\sum_{i=1}^{m} n_i \widehat{p}_{ij}^{MLE}$ (overall proportion of outcomes of type j), the independence between players further leads to

$$\widehat{Cov}\left(t_j, \hat{p}_{ij}^{MLE}\right) = w_i \widehat{Var}\left(\hat{p}_{ij}^{MLE}\right), w_i = \frac{n_i}{\sum_{i=1}^{m} n_i}.$$

Empirical Bayes estimation

One can also take an empirical Bayes approach for estimating p_{ij} by following the development in [5]. First, assume that each p_i has the same prior distribution;

$$p_i = \left(p_{i1}, p_{i2}, \ldots, p_{ik}\right)^t \sim \text{Dirichlet}\left(\alpha = \left(\alpha_1, \alpha_2, \ldots, \alpha_k\right)^t\right)$$

Then, the posterior distribution of p_i, given the observed counts for player i, is

$$p_i | x_i; \propto \sim \text{Dirichlet}(x_{i1} + \alpha_1, x_{i2} + \alpha_2, \ldots, x_{ik} + \alpha_k)$$

and it can be shown that the Bayes estimator of p_{ij} is

$$\hat{p}_{ij}^{Bayes} = \lambda_i t_j + \left(1 - \lambda_i\right) \hat{p}_{ij}^{MLE}$$

which is a shrinkage estimator with

$$t_j = \left(\frac{\alpha_j}{\sum_{j=1}^{K} \alpha_j}\right) \text{ and } \lambda_i = \frac{\sum_{j=1}^{K} \alpha_j}{\left(n_i + \sum_{j=1}^{K} \alpha_j\right)}. .$$

Here t_j is the shrinkage target corresponding to the prior mean of p_{ij} and $\lambda_i \in (0,1)$ is the shrinkage constant. Then, the empirical Bayes strategy is to replace the shrinkage target and constant with their sample counterparts,

$$\hat{t}_j = \left(\frac{\hat{\alpha}_j}{\sum_{j=1}^{K} \hat{\alpha}_j}\right) \text{ and } \hat{\lambda}_i^* = \frac{\sum_{j=1}^{K} \hat{\alpha}_j}{\left(n_i + \sum_{j=1}^{K} \hat{\alpha}_j\right)}$$

in the above expression for the Bayes estimator. The parameter estimates $\hat{\alpha} = \left(\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_K\right)^t$ are obtained by using the dirmult package in R [6], relying on maximum likelihood estimation based on the marginal distribution of the data given α, and can be interpreted as using data-driven parameters in the prior distribution.

Bayesian multinomial regression estimation

We now describe a Bayesian multinomial regression approach for estimating p_{ij} in the presence of covariates. Let $Y = (Y_1, Y_2, \ldots, Y_l)$ represent an $m \times l$ matrix of l covariates for the m players, where $Y_l = (Y_{1l}, Y_{2l}, \ldots, Y_{ml})^t$. In the Bayesian multinomial regression model formulation, we write our estimation problem as

$$\mathbf{x_i} \mid \mathbf{p_i} \sim \text{Multinominal}\left(n_i, \mathbf{p_i}\right),$$

$$\mathbf{p_i} \mid \boldsymbol{\alpha_i} \sim \text{Dirichlet}\left(\boldsymbol{\alpha_i} = \left(\alpha_{i1}, \alpha_{i1}, \ldots, \alpha_{ik}\right)^t\right).$$

Here α_i is positive, and a natural link function is a log-link function leading to

$$\alpha_i = \exp(\gamma i) \qquad \text{Where } \gamma i = \left(\gamma_{i1}, \gamma_{i2}, \ldots, \gamma_{ik}\right)^t$$

and assume

$$\gamma ij = \beta_{oj} + \sum_{a=1}^{l} \beta_{aj} Y_{ia} \qquad \text{i} = 1,2,\ldots, \text{m; j} = 1,\ldots, \text{K}$$

In this setup, β_{aj} captures the effect of a[th] covariate on the j[th] category. Note that in this model, the players have the same β's but different p's because their covariates take different values. We assume normal priors on β_{0j} and β_{aj} .

Specifically, we use

$$\beta_{oj} \sim \text{N}(0,1).$$

$$\beta_{aj} \sim \text{N}(0,1), \text{j} = 1, \ldots \text{K and a} = 1, \ldots, l$$

Vannucci et al. (2017) introduced a similar model and used the spike-and-slab mixture priors for βs; one of the prior is a Dirac-delta at 0. The Dirac-delta allows zero probabilities, whereas, in our case, the true cell probabilities are small but not actually zero under sparsity. In what follows, $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m)$, let $\boldsymbol{\beta} = (\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$, where $\beta_j = (\beta_{0j}, \beta_{1j}, \ldots, \beta_{lj})^t$ and $\mathbf{p} = (\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_m)$. Then, we have

$$\pi(x \mid p) = \prod_{i=1}^{m} \frac{n_i}{\prod_{j=1}^{K} x_{ij}!} \prod_{j=1}^{K} p_{ij}^{x_{ij}}, \ \pi(\beta) = \prod_{j=1}^{K} \prod_{a=0}^{l} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\beta_{aj}^2}{2}\right) \propto \exp\left(-\sum_{j=1}^{K} \sum_{a=0}^{l} \frac{\beta_{aj}^2}{2}\right)$$

and

$$\pi\left(p_i \mid \alpha_i\right) = \frac{\Gamma\left(\sum_{j=1}^{K} \alpha_{ij}\right)}{\prod_{j=1}^{K} \Gamma\left(\alpha_{ij}\right)} \prod_{j=1}^{K} p_{ij}^{\alpha_{ij}-1}.$$

Substitute

$$\alpha_{ij} = \exp\left(\gamma_{ij}\right) = \exp\left(\beta_{0j} + \sum_{a=1}^{l} \beta_{aj} Y_{ia}\right),$$

$$\pi(p \mid \beta) = \prod_{i=1}^{m} \frac{\Gamma\left(\sum_{j=1}^{K} \exp\left(\beta_{0j} + \sum_{a=1}^{l} \beta_{aj} Y_{ia}\right)\right)}{\prod_{j=1}^{K} \Gamma\left(\exp\left(\beta_{0j} + \sum_{a=1}^{l} \beta_{aj} Y_{ia}\right)\right)} \prod_{j=1}^{k} p_{ij}^{\exp\left(\beta_{0j} + \sum_{a=1}^{l} \beta_{aj} Y_{ia}\right) - 1}.$$

The posterior distribution

$$\pi(p, \beta \mid x) \propto \pi(x \mid p) \times \pi(p \mid \beta) \times \pi(\beta)$$

$$= \prod_{i=1}^{m} \frac{n_i}{\prod_{j=1}^{K} x_{ij}!} \prod_{j=1}^{K} p_{ij}^{x_{ij}} \times \frac{\Gamma\left(\sum_{j=1}^{K} \exp\left(\beta_{0j} + \sum_{a=1}^{l} \beta_{aj} Y_{ia}\right)\right)}{\prod_{j=1}^{K} \Gamma\left(\exp\left(\beta_{0j} + \sum_{a=1}^{l} \beta_{aj} Y_{ia}\right)\right)}$$

$$\times \prod_{j=1}^{k} p_{ij}^{\exp\left(\beta_{0j} + \sum_{a=1}^{l} \beta_{aj} Y_{ia}\right) - 1} \times \prod_{j=1}^{K} \prod_{a-0}^{l} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\beta_{aj}^2}{2}\right)$$

$$\propto \prod_{i=1}^{m} \frac{\Gamma\left(\sum_{j=1}^{K} \exp\left(\beta_{0j} + \sum_{a=1}^{l} \beta_{aj} Y_{ia}\right)\right)}{\prod_{j=1}^{K} \Gamma\left(\exp\left(\beta_{0j} + \sum_{a=1}^{l} \beta_{aj} Y_{ia}\right)\right)} \prod_{j=1}^{K} p_{ij}^{x_{ij} + \exp\left(\beta_{0j} + \sum_{a=1}^{l} \beta_{aj} Y_{ia}\right) - 1}$$

$$\times \exp\left(-\sum_{a=0}^{l} \sum_{j=1}^{K} \frac{\beta_{aj}^2}{2}\right). \quad (2)$$

Note that by holding β fixed,

$$\pi(p, \beta \mid x) \propto \prod_{i=1}^{m} \prod_{j=1}^{K} p_{ij}^{x_{ij} + \exp\left(\beta_{0j} + \sum_{a=1}^{l} \beta_{aj} Y_{ia}\right) - 1},$$

implying conditional independence (given β) of the p_i's with marginal PDF given by

$$\pi(p_i \mid \beta, x_i) \propto \prod_{j=1}^{K} p_{ij}^{x_{ij} + \exp\left(\beta_{0j} + \sum_{a=1}^{l} \beta_{aj} Y_{ia}\right) - 1}$$

Letting

$$\gamma_{ij}^* = x_{ij} + \exp\left(\beta_{0j} + \sum_{a=1}^{l} \beta_{aj} Y_{ia}\right) \text{ and } \tilde{\mathbf{a}}_{\mathbf{i}}^* = \left(\gamma_{i1}^*, \gamma_{i2}^*, \ldots\ldots, \gamma_{iK}^*\right)^t$$

the posterior conditional distribution of p_i is

$$\mathbf{p_i} \mid \boldsymbol{\beta}, \mathbf{x_i} \sim Dirichlet\left(\boldsymbol{\gamma}_{\mathbf{i}}^*\right).$$

When holding p fixed, however,

$$\pi(\boldsymbol{\beta} \mid \mathbf{p}, \mathbf{x}) \propto \prod_{i=1}^{m} \frac{\Gamma\left(\sum_{j=1}^{K} \exp\left(\beta_{0j} + \sum_{a=1}^{l} \beta_{aj} Y_{ia}\right)\right)}{\prod_{j=1}^{K} \Gamma\left(\exp\left(\beta_{0j} + \sum_{a=1}^{l} \beta_{aj} Y_{ia}\right)\right)} \prod_{j=1}^{K} p_{ij}^{\exp\left(\beta_{0j} + \sum_{a=1}^{l} \beta_{aj} Y_{ia}\right) - 1} \exp\left(-\sum_{a=0}^{l} \sum_{j=1}^{K} \frac{\beta_{aj}^2}{2}\right)$$

So that the posterior conditional distribution of β (given p) does not belong to a standard family of distributions. To perform inference based on the full posterior distribution (2), we developed a Metropolis within Gibbs algorithm to generate values from this distribution. Alternatively, Bayesian multinomial logistic regression could have been used in the current context. In multinomial logistic regression, we nominate one of the categories to be a baseline or reference category (usually the last category), calculate log odds for all other categories relative to the baseline, and let the log odds be a linear function of the predictors as follows:

$$\gamma_{ij} = \log\left(\frac{p_{ij}}{p_{iK}}\right) = \beta_{0j} + \sum_{a=1}^{l} \beta_{aj} Y_{ia} \qquad \text{i} = 1,2,\ldots\ldots, \text{m}; \text{j} = 1,\ldots\ldots, \text{K-1}$$

Note that we need only K – 1 equations to compute p_{ij}s such that,

$$\mathrm{p}_{ij} = \frac{\exp(\gamma_{ij})}{1 + \sum_{l=1}^{K-1} \exp(\gamma_{il})} \qquad \text{i} = 1,2,\ldots\ldots, \text{m}; \text{j} = 1,\ldots\ldots, \text{K-1}$$

And

$$\mathrm{p}_{iK} = 1 - \left(p_{i1} + \ldots + p_{i(K-1)}\right) = \frac{1}{1 + \sum_{l=1}^{K-1} \exp(\gamma_{il})} \qquad \text{i} = 1, 2, \ldots\ldots, \text{m}$$

In the Bayesian setting, we put priors on β_{0j}s and β_{aj}s. One important advantage of our proposed multinomial regression model is that it derives a distribution for p_i instead of calculating p_{ij}'s separately as the multinomial logistic regression model.

Proposed Semi-parametric Bayesian Estimation

Our focus is mainly on data sparsity, which means cell counts are small for one or more categories due to the small sample size and the cell probabilities are small but not actually zero. However, the conventional multinomial models perform very poorly under sparsity. We aim to develop an approach to derive improved estimators that can handle sparsity by borrowing information from other multinomial populations. The Dirichlet process is one of the most popular Bayesian non-parametric models. The primary motivation to use the Dirichlet process as a prior distribution is that it gives a large space over which we make our inferences as well as a tractable posterior distribution [Teh (2010)]. Another advantage of the Dirichlet process is the clustering property which clusters the populations without specifying the number of clusters in advance. This clustering property borrows information from similar populations, which helps to handle data sparsity. With those advantages, we propose a semi-parametric Bayesian estimator based on the Dirichlet process (DP).

Dirichlet processes, introduced by [9], are a family of stochastic processes whose realizations are probability distributions. These can be seen as a distribution over distributions as each draw from a Dirichlet process is itself a distribution. It is called a Dirichlet process because it is a generalization of the Dirichlet distribution to an infinite number of dimensions to model the weights of these components. A Dirichlet process is completely specified by two components; an underlying base distribution (G_0) and a positive real number (α_0) called the concentration parameter and is denoted by

$$\mathrm{G} \sim \mathrm{DP}(\alpha_0, G_0).$$

If the base distribution is continuous, then G is a discrete distribution made up of a countably infinite number of point masses. The concentration parameter α_0 is also called the strength parameter as it specifies how "strong" this discretization actually is. It can be shown that

$$E(G(A)) = G_0(A) \qquad \mathrm{Var}(G(A)) = \frac{G_0(A(1 - G_0(A)))}{\alpha_0 + 1}.$$

For any measurable subset $A \subset \Theta$. When $\alpha_0 \to 0$, all the realizations are concentrated at a single value, while in the limit of $\alpha_0 \to \infty$, the realizations become continuous. We formulate our semi-parametric Bayesian model as follows;

$$\mathbf{x_i} \mid \mathbf{p_i} \sim Multinomial\left(n_i,\ \mathbf{p_i}\right) \text{ i} = 1,\ldots., \text{m}$$
$$p_i \mid G \sim G$$
$$G \mid \alpha_0,\ \mathbf{G_0} \sim Dp\left(\alpha_0, \mathbf{G_0}\right)$$
$$\mathbf{G_0} \sim Dirichlet\left(\alpha_1,\ \alpha_2,\ldots..,\ \alpha_K\right),$$

where $X_i = (X_{i1}, X_{i2},..., X_{iK})^t$ and $p_i = (p_{i1}, p_{i2},..., p_{iK})^t$. By following [10], the successive conditional distribution of *pi* given $p_1, p_2,..., p_{i-1}, p_{i+1},...,pm$ has the following form when G has been integrated out:

$$p_i \mid \mathrm{p}_1,\ \mathrm{p}_2,\ldots,\ \mathrm{p}_{i-1},\ \mathrm{p}_{i+1},\ldots,\ \mathrm{p}_m, \alpha_0, G_0 \sim \frac{1}{\alpha_0+m-1}\sum_{l\neq i}\delta_{pl} + \frac{\alpha_0}{\alpha_0+m-1}G_0$$
$$\sim \left(\frac{m-1}{\alpha_0+m-1}\right)\frac{\sum_{l\neq 1}\delta_{pl}}{m-1} + \left(\frac{\alpha_0}{\alpha_0+m-1}\right)G_0$$

Here $\delta_{\mathbf{p_{n_s}}}$ is the indicator function such that,

$$\delta_{\mathbf{p_{n_s}}} = \begin{cases} 1 & \text{if } \mathbf{P} = \mathbf{P}_{n_s} \\ 0 & \text{otherwise} \end{cases}$$

The posterior base distribution is a weighted average between the prior base distribution G_0 and the empirical distribution $\frac{\sum_{l\neq i}\delta_{pl}}{m-1}$. The weights are controlled by the concentration parameter α_0. The larger the α_0 value, the larger the weight for G_0 in comparison to the weight for the empirical distribution and vice versa. For $m >> \alpha_0$, the empirical distribution will dominate. For $m \to \infty$, the posterior of the DP converges to the true underlying distribution over p. We refer readers to [11] for other properties of DP formulation and its properties. Note that one can use the stick-breaking construction proposed by [12] to draw samples from a DP. In the stick-breaking construction, we draw

$\xi_k \sim \text{Beta}(1, \alpha_0)$ k = 1, 2,...

and construct

$$\pi_1 = \xi_1 \text{ and } \pi_{n_s} = \xi_{n_s}\prod_{k=1}^{n_s-1}(1-\xi_k) \quad \text{n}_s = 2,3\ldots\ldots$$

Intuitively, consider starting with a stick of unit length and breaking a random proportion ξ_1 of that stick. The length of this piece gives you the $\pi = \left\{\pi_{n_s}\right\}_{n_s=1}^{\infty}$ first weight π_1. Then, break a random portion from the remaining stick ξ_2. The length of the second piece gives you the second weight π_2. Now, continue breaking the remaining portions of the stick to obtain π_3, π_4, and so forth. Using this construction, an infinite sequence of weights can be generated. When ns gets larger and larger, the lengths of the pieces of the stick, or the weights, will tend to get smaller and smaller. The lengths of the pieces are determined by the concentration parameter α_0. For small α_0, only the first few pieces will have significant lengths, the remaining pieces having very small lengths. On the other hand, for large α_0, the lengths will tend to be more uniform. Then, the discrete random probability distribution is

$$G = \sum_{n_s=1}^{\infty}\pi_{n_s}\delta_{\mathbf{p_{n_s}}}$$

and we sample from the posterior distribution of p using the Gibbs sampling approach described in [13].

Data Analysis

Application to Inference on Bowling Performance

Some background information: The game of cricket was first started in the late 16th century and has become popular globally in the 19th and 20th centuries due to the introduction of various formats, including Twenty20 international cricket (T20I cricket). Cricket has multiple formats depending on the desired length of a typical match. Test cricket has a duration of five days, and one-day cricket has a period of one day, whereas T20I matches are completed in roughly three hours, with each inning lasting about 75-90 minutes. The International Cricket Council (ICC) governs the body of cricket, with 105 countries as its members. [14-16] provides a comprehensive discussion on various aspects of cricket and its recent research directions.

There has been growing interest in T20I cricket performance analysis in recent literature. [17] provided a comprehensive overview of tactics in T20 international cricket. A simulator for modeling T20I cricket was proposed by [18]. Note that there are three major components that lead to success in cricket: batting, bowling, and fielding. [19-22] assessed the batting performance of players in the Indian Premier League (IPL) and one-day international cricket. [19] proposed a Bayesian hidden Markov model, and [22] proposed a performance measure combining batting average, batsman's consistency, and strike rate. However, we remark that there has been little attention on bowling or fielding performance. [23] proposed an approach based on random forests to measure the fielding performance in T20I cricket. In what follows, we study the bowling performance of players in T20I cricket using the models introduced in Section 3.

About Bowling Performance in T20I Cricket: Here, we consider the number of wickets taken in T20 international matches by bowlers between 1st of January 2010 and 11th of March 2020. Our analysis includes m = 175 bowlers with at least 16 total wickets. Details of these T20I matches can be found in the Archive section of the ESPNcricinfo website (www.espncricinfo.com), and the T20I bowler rankings were obtained from the ICC cricket website (www.icc-cricket.com) on March 11, 2020. The number of matches ranged from 8 to 77 matches for each bowler. Rashid Khan is the highest wicket-taker with 89 total wickets for the given period, and he was the highest-ranked bowler on March 11, 2020, according to ICC.

Recall that the basic assumption here is that

$\mathbf{X_i}|\mathbf{p_i} \sim \text{Multinomial}(n_i;\ \mathbf{p_i} = (p_{i1}, p_{i2}, ..., p_{iK})),$

for each of the m = 175 bowlers, and that p_{ij} (j = 1, 2, ...,7) denotes the probability that player i records j – 1 wickets.

Table 1 provides the number of players (out of 175) who have non-zero counts for each wicket category. The highest number of wickets recorded in a single match by one bowler in T20 international matches was 6 wickets over the considered time period. It is clear that 6 wickets are very rarely achieved, and there exists data sparsity for this dataset. Note that Deepak Chahar, Ajantha Mendis, and Yuzvendra Chahal are the only bowlers to have taken 6-wickets hauls in T20I matches. Deepak Chahar has the best bowling figure in T20I cricket over all players, and Ajantha Mendis has two 6-wicket hauls in T20I matches.

Table 1: No of players with non-zero counts for each wicket category.

Wickets	0W	1W	2W	3W	4W	5W	6W
No of players	174	175	174	157	87	23	3

In Table 2, we provide summary statistics for the top 30 ranked bowlers according to ICC rankings. Note that Mendis has retired from T20I cricket, so he is not included in the top 30 bowlers given in the table. The James-Stein estimator of p_{ij} is

$$\hat{p}_{ij}^{JS} = \lambda_i t_j + \left(1-\lambda_i\right)\hat{p}_{ij}^{MLE},$$

and we considered two choices for the shrinkage target: the uniform target $t_j = \frac{1}{K} = \frac{1}{7} = 0.143$ for j = 1, 2, ...,7, and the overall pro portion (op) $\overline{p_j} = \frac{\sum_{i=1}^{175} x_{ij}}{\sum_{i=1}^{175}\sum_{j=1}^{7} x_{ij}}$ for each category given in Table 3.

Here, x_{i1} is the number of matches in which the i[th] bowler did not take any wickets, x_{i2} is the number of matches in which the i[th] bowler took exactly one wicket, and so on.

Table 2: Summary Statistics of bowlers.

Bowler	Country	Matches	Wickets	0W	1W	2W	3W	4W	5W	6W	Ranking
Rashid Khan	Afghanistan	48	89	6	15	14	8	3	2	0	1
Mujeeb Ur Rahman	Afghanistan	19	25	4	9	3	2	1	0	0	2
Adam Zampa	Australia	29	33	11	7	7	4	0	0	0	3
Ashton Agar	Australia	24	25	11	6	4	2	0	1	0	4
Tabraiz Shamsi	South Africa	22	17	9	9	4	0	0	0	0	5
Mitchell Santner	New Zealand	43	52	12	15	12	3	1	0	0	6
Imad Wasim	Pakistan	42	42	14	21	3	2	1	1	0	7
Adil Rashid	England	36	38	9	19	5	3	0	0	0	8
Shadab Khan	Pakistan	38	48	9	15	10	3	1	0	0	9
Sheldon Cottrell	West Indies	27	36	6	10	8	2	1	0	0	10
Chris Jordan	England	46	58	16	13	8	7	2	0	0	11
Kane Richardson	Australia	18	19	8	3	5	2	0	0	0	12
Jasprit Bumrah	India	49	59	11	22	11	5	0	0	0	13
Andile Phehlukwayo	South Africa	26	35	6	10	6	3	1	0	0	14
Ish Sodhi	New Zealand	44	53	12	16	11	5	0	0	0	15
Tim Southee	New Zealand	60	67	24	16	11	8	0	1	0	16
Pat Cummins	Australia	28	36	4	14	8	2	0	0	0	17
Mark Watt	Scotland	33	45	9	12	5	6	0	1	0	18
Billy Stanlake	Australia	19	27	4	7	5	2	1	0	0	19
Washington Sundar	India	22	19	8	10	3	1	0	0	0	20
Lakshan Sandakan	Sri Lanka	17	17	7	7	0	2	1	0	0	21
Mohammad Nabi	Afghanistan	77	69	35	24	12	3	3	0	0	22
Mitchell Starc	Australia	31	43	5	14	7	5	0	0	0	23
David Willey	England	28	34	9	10	4	4	1	0	0	24
Faheem Ashraf	Pakistan	26	24	11	9	3	3	0	0	0	25
Lasith Malinga	Sri Lanka	63	83	18	21	15	6	1	2	0	26
Tim Curran	England	22	22	9	5	7	1	0	0	0	27
Alasdair Evans	Scotland	25	36	4	12	5	3	0	1	0	28

Yuzvendra Chahal	India	42	55	12	17	6	4	2	0	1	29
Liam Plunkett	England	21	24	7	7	4	3	0	0	0	30

Table 3: Overall proportion (op) for the 7 wicket categories.

Wickets	0W	1W	2W	3W	4W	5W	6W
$\bar{p}_j$	0.35	0.335	0.201	0.087	0.021	0.005	0.001

The James-Stein estimate of the optimal shrinkage constant for each $\hat{\lambda}_i^*$ player is given in Figure 1 when the shrinkage target is 0.143. Lungi Ngidi has the maximum optimal shrinkage constant of 0.729 and has the highest shrinking towards the shrinkage target. Darren Sammy has the minimum optimal shrinkage constant of 0.033 and has the lowest shrinking toward the shrinkage target. The optimal shrinkage constant represents the confidence in our shrinkage target; a low shrinkage constant indicates more confidence in the baseline estimate given by the MLE.

The empirical Bayes estimator p_{ij} is

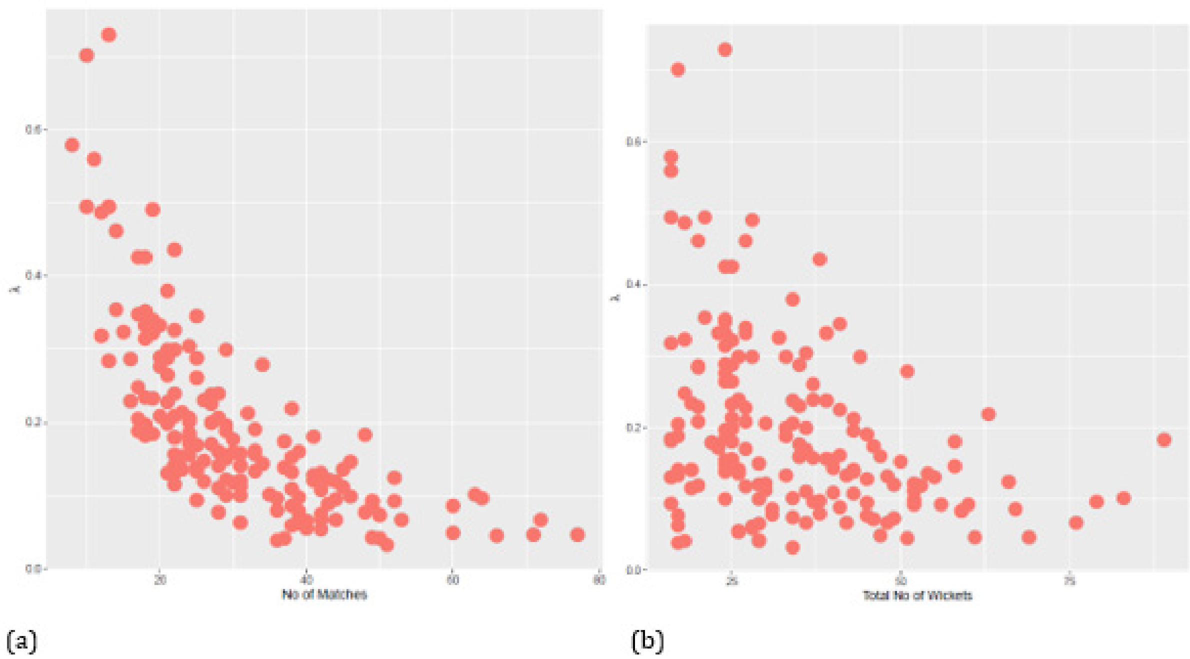

Figure 1: Optimal shrinkage constants corresponding to shrinkage target plotted against $t_j = \frac{1}{7}$ (a). the number of matches, and (b). the total number of wickets.

$$\hat{p}_{ij}^{Bayes} = \hat{\lambda}_i^* \hat{t}_j + \left(1 - \hat{\lambda}_i^*\right) \hat{p}_{ij}^{MLE} \quad i = 1,2,...., \ 175 \text{ and } j = 1,2,...., 7$$

Table 4 provides the parameter estimates of the concentration parameters $\hat{\alpha}$. While $\hat{\alpha} < 1$, provides sparse multinomial distribution, while $\hat{\alpha} > 1$, provides smooth multinomial distribution. The values of the concentration parameters for the first five-wicket categories are higher than one, but for the last two wicket categories, the values are less than 1.

Table 4: Estimates of the concentration parameters and the shrinkage targets ($\hat{t}_j$).

Wickets	0W	1W	2W	3W	4W	5W	6W
$\hat{\alpha}_j$	52.27	50.89	30.69	13.35	3.31	0.77	0.1
	0.345	0.336	0.203	0.088	0.022	0.005	0.001

$\hat{t}_j$

The estimates of the optimal shrinkage constants for the empirical Bayes method, where the shrinkage target is, $\hat{t}_j = \frac{\hat{\alpha}_j}{\sum_{j=1}^{K} \hat{\alpha}_j}$ are given in Figure 2. Ashok Dinda has the maximum optimal shrinkage constant of 0.950 and has the highest shrinking towards the shrinkage target. Mohammad Nabi has the minimum optimal shrinkage constant of 0.663 and has the lowest shrinking toward the shrinkage target.

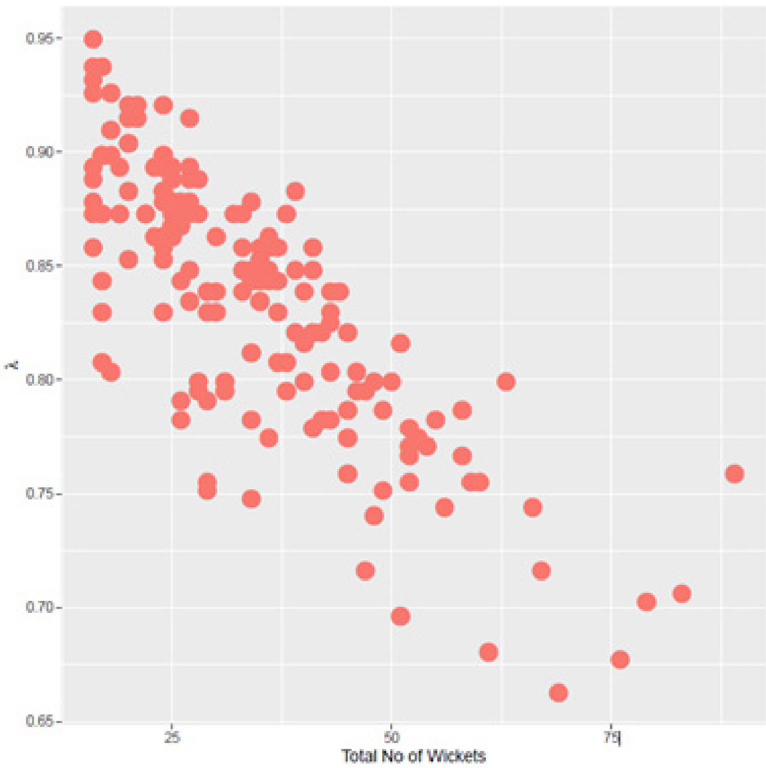

Figure 2: Optimal shrinkage constants for the empirical Bayes method, where the shrinkage target is $\hat{t}_j = \frac{\hat{\alpha}_j}{\sum_{j=1}^{K} \hat{\alpha}_j}$

In the semi-parametric Bayesian approach, 100,000 draws were taken from a DP using Gibbs sampling with a burn-in of 50,000 under the following parameters:

$G \sim DP\ (\alpha_0 = 150, G_0),$

$G_0 \sim \text{Dirichlet}\ (\alpha_{0W} = \alpha_{1W} = \dots = \alpha_{6W} = 1).$

Teh [8] proposed a formula to find the expected number of atoms in DP realizations based on α0 and the number of observations. Based on that formula, we picked $\alpha_0 = 150$, which provides reasonable no of clustering throughout the posterior simulations. For implementing the Bayesian regression model, we considered the following covariates: the number of overs the bowler bowled; the number of runs the bowler conceded in T20 matches from 1st of January 2010 to 11th of March 2020; type of bowler (0='Seam', 1='Spin'); the age of the bowler and the economy rate of the bowler. For the age, we calculate the median age of the bowler for his playing period.

Figure 3 and Figure 4 provide the comparison of James-Stein (JS) shrinkage estimates with shrinkage target $t_j = \frac{1}{7}$ James-Stein (JS) shrinkage estimates with overall proportion (OP) as the shrinkage target, empirical Bayes (EB) estimates, maximum likelihood (ML) estimates, Dirichlet process (DP) estimates, Bayesian multinomial regression (BMR) estimates and overall proportions of wicket categories for different bowlers. The grey shaded area is the 95% credible interval of Dirichlet Process estimates. Here JS estimates are close to BMR estimates, whereas EB estimates are close to DP estimates, JS (OP) estimates, and ML estimates. The plots in the first row of Figure 3 are for the bowlers who got 6 wickets. We can see clearly, EB estimates and DP estimates are very close to the overall proportion. Rashid Khan and Ashok Dinda have wider 95% credible intervals of DP estimates since these two players don't cluster with other players a lot, meaning those players have unique cell probabilities that differ from others.

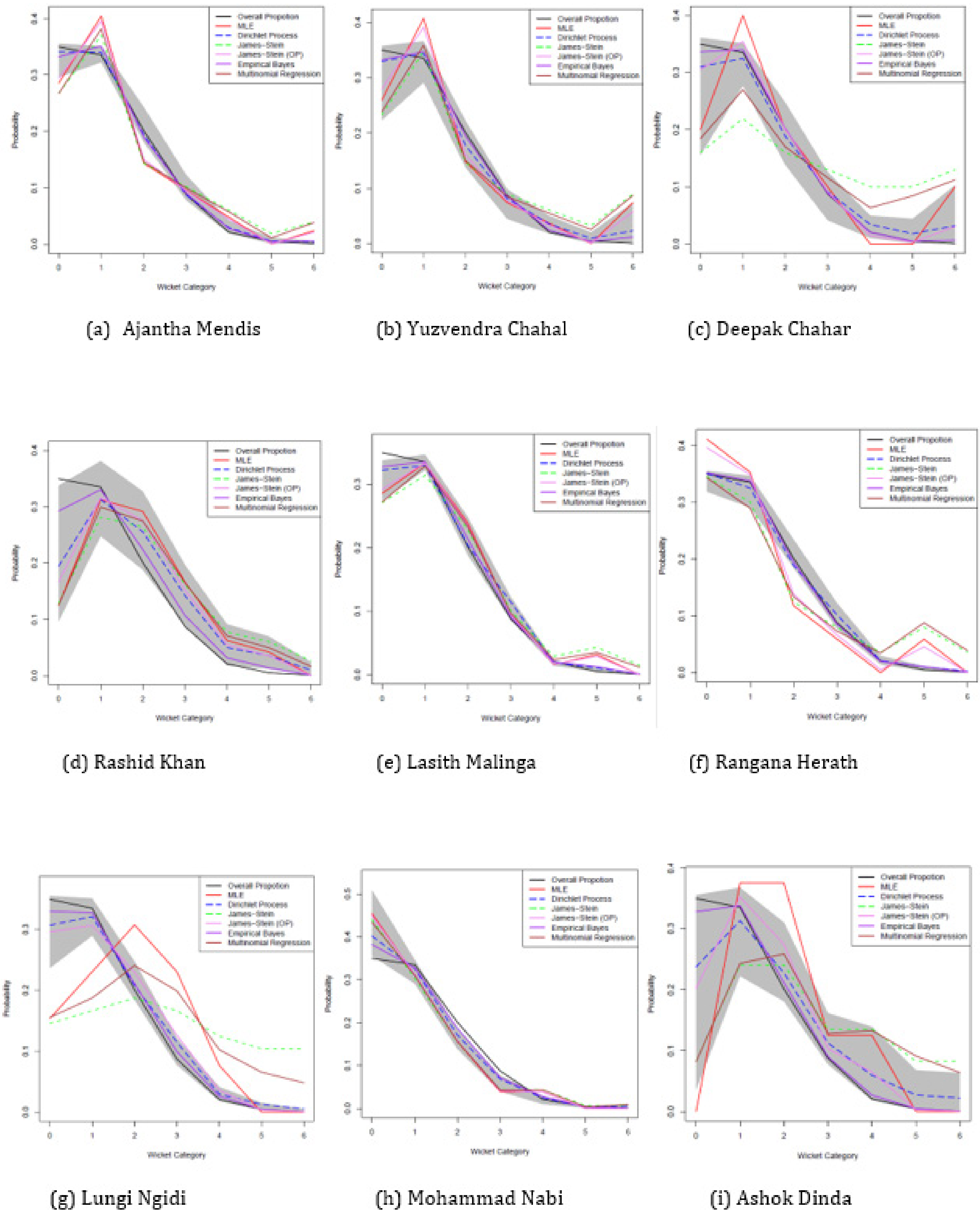

Figure 3: Comparison of DP, JS, EB, BMR, and ML estimates.

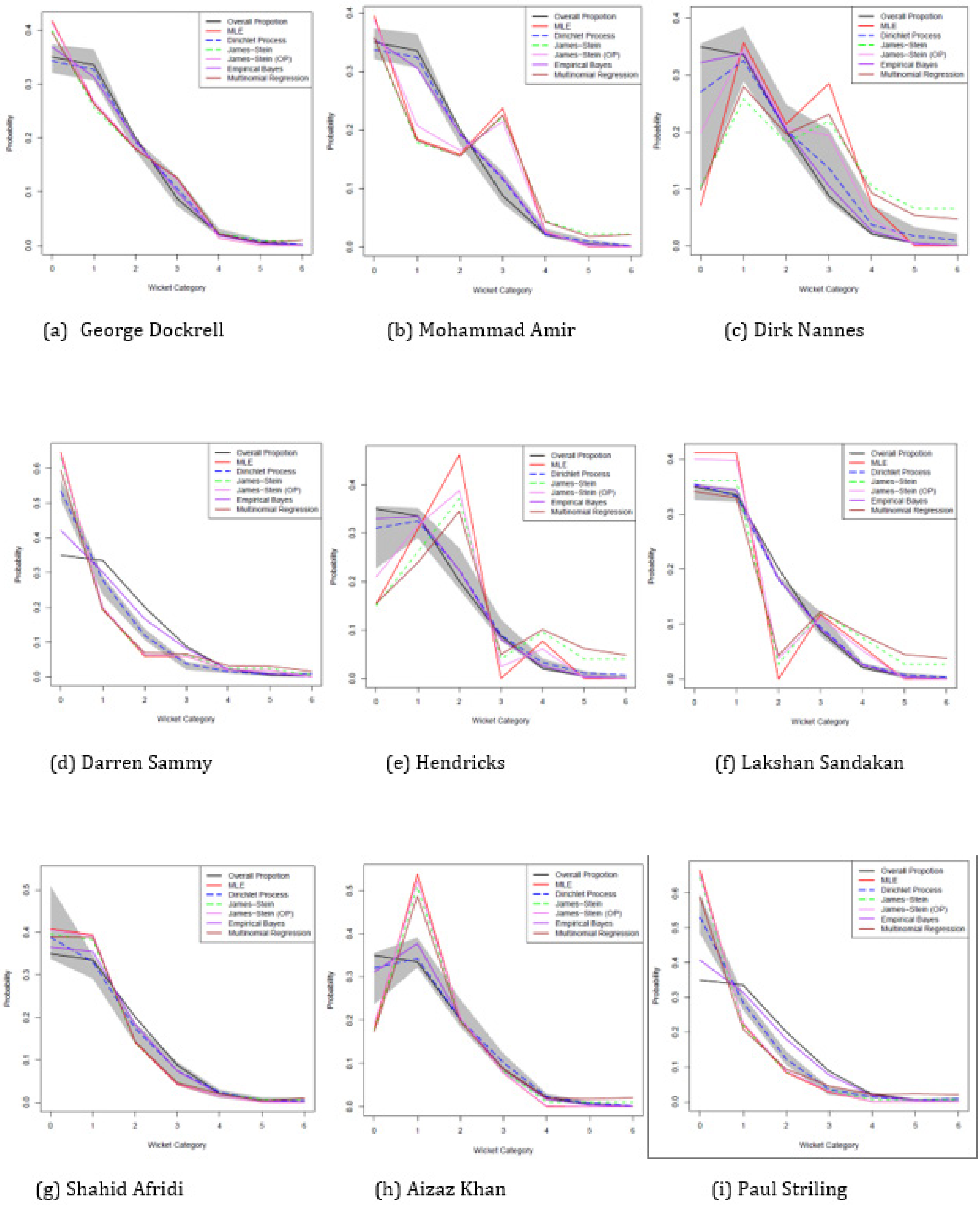

Figure 4: Comparison of DP, JS, EB, BMR, and ML estimates.

Table 5 provides the expected number of wickets per match based on estimates to assess the bowling performances of top-ranked bowlers. The ranks are given from the highest to the lowest value (the highest is rank 1), and the ties are given the highest rank using all the bowlers. Rashid Khan is the ICC top-ranked player and also the top wicket-taker who has the highest expected number of wickets per match (rank 1) based on estimates of James-Stein (JS) shrinkage estimates with overall proportion (OP) as the shrinkage target, empirical Bayes (EB) estimates and Dirichlet process (DP) estimates. However, Rashid Khan is given rank 4 based on the maximum likelihood (ML) estimates.

Table 5: Expected number of wickets per match for top-ranked bowlers.

ICC Ranking	Bowler	Original Data		JS		JS (OP)		EB		DP		BMR	
		Value	Rank	Value	Rank	Value	Rank	Value	Rank	Value	Rank	Value	Rank
1	R Khan	1.85	4	2.06	17	1.72	1	1.3	1	1.68	1	1.97	15
2	M Ur Rahman	1.32	52	1.71	49	1.27	47	1.14	61	1.18	70	1.79	35
3	A Zampa	1.14	98	1.49	85	1.13	98	1.12	98	1.18	75	1.46	96
4	A Agar	1.04	114	1.4	102	1.05	114	1.11	108	1.16	114	1.46	95
5	T Shamsi	0.77	157	1.07	149	0.82	156	1.08	148	1.06	150	1.26	138
6	M Santner	1.21	76	1.43	95	1.2	75	1.14	67	1.17	89	1.4	109
7	I Wasim	1	130	1.14	143	1.01	130	1.1	134	1.15	121	1.22	147
8	A Rashid	1.06	109	1.21	133	1.06	110	1.11	113	1.18	74	1.3	133
9	S Khan	1.26	61	1.49	81	1.24	59	1.15	46	1.18	69	1.5	85
10	S Cottrell	1.33	43	1.67	58	1.29	39	1.16	40	1.18	65	1.65	56
11	C Jordan	1.26	62	1.51	77	1.24	60	1.16	38	1.17	83	1.43	102
12	K Richardson	1.06	109	1.51	78	1.07	108	1.12	107	1.16	107	1.61	67
13	J Bumrah	1.2	79	1.35	112	1.2	76	1.14	64	1.2	37	1.38	115
14	A Phehlukwayo	1.35	39	1.73	45	1.29	37	1.16	37	1.18	62	1.7	48
15	I Sodhi	1.21	78	1.42	96	1.19	77	1.14	69	1.19	46	1.38	117
16	T Southee	1.12	102	1.28	120	1.12	102	1.12	102	1.19	51	1.25	141
17	P Cummins	1.29	55	1.48	88	1.27	48	1.15	51	1.2	28	1.63	62
18	M Watt	1.36	38	1.68	55	1.32	29	1.17	22	1.21	23	1.62	65
19	B Stanlake	1.42	30	1.96	23	1.32	28	1.16	34	1.18	60	1.86	21
20	W Sundar	0.86	147	1.16	141	0.9	147	1.09	142	1.12	141	1.39	113
21	L Sandakan	1	119	1.38	106	1.02	121	1.11	111	1.14	129	1.56	74
22	M Nabi	0.9	145	0.99	156	0.91	146	1.05	160	1.01	158	1	163
23	M Starc	1.39	36	1.61	64	1.35	24	1.17	20	1.21	22	1.66	54
24	D Willey	1.21	75	1.58	67	1.19	78	1.14	77	1.17	84	1.55	78
25	F Ashraf	0.92	142	1.23	129	0.95	142	1.09	137	1.14	131	1.38	119
26	L Malinga	1.32	50	1.49	84	1.3	33	1.18	14	1.21	21	1.42	105
27	T Curran	1	119	1.36	109	1.02	123	1.11	118	1.14	126	1.5	84
28	A Evans	1.44	28	1.7	50	1.38	14	1.17	19	1.24	13	1.76	40
29	Y Chahal	1.31	54	1.53	75	1.28	43	1.16	25	1.15	117	1.48	90
30	L Plunkett	1.14	97	1.64	61	1.14	97	1.12	97	1.18	73	1.58	71

Simulation Study

Although an in-depth simulation study is beyond what we originally set out to accomplish, it would be helpful to demonstrate and compare how the proposed and existing estimators perform over simulated datasets considering different scenarios. Nevertheless, we report here on a brief simulation study conducted using 10000 Monte Carlo simulations in two scenarios where the true cell probabilities are known. We report the Mean Squared Error (MSE) and compare the estimators below. Also, we varied the number of populations (m - 100, 200, and 500) and the number of categories (K5, 10, and 15) to explore the performance of the estimators. We generated data from multinomial distributions with sample sizes (n_i) ranging from 15 to 75.

Scenario 1

In this first scenario, the true cell probabilities are strictly decreasing, but the differences between successive p's are the same. For example, when K = 5, the true cell probabilities are

$p=\left(\frac{5}{15},\frac{4}{15},\frac{3}{15},\frac{2}{15},\frac{1}{15}\right)'$ and that is used to generate the counts for all the populations. Table 6 provides the mean squared error values for each estimator based on Scenario 1. When m is fixed and K increases, the MSE decreases. The MSE also decreases when K is fixed and m increases. The proposed semi-parametric Bayesian estimator performs better than the existing estimators and MLE. This scenario is very similar to the cricket application, where the true cell probabilities of the wicket categories for each player generally have a decreasing pattern.

Table 6: MSE values for scenario 1.

Estimator	*m*=100			*m*=200			*m*=500		
	K = 5	*K* = 10	*K* = 15	*K* = 5	*K* = 10	*K* = 15	*K* = 5	*K* = 10	*K* = 15
MLE	0.124	0.0561	0.0456	0.1227	0.0544	0.0454	0.1226	0.0542	0.0453
James-Stein	0.112	0.0536	0.0437	0.1116	0.0518	0.0436	0.1114	0.0516	0.0436
Empirical Bayes	0.1121	0.0552	0.0439	0.1119	0.0531	0.0437	0.1118	0.053	0.0437
Dirichlet Process	0.1102	0.0501	0.0425	0.1086	0.0472	0.0422	0.1085	0.0471	0.0422

Scenario 2

In the third scenario, the true cell probabilities increase and decrease (zig-zag pattern).

For example, when K = 5, the true cell probabilities are $p=\left(\frac{1}{23},\frac{10}{23},\frac{1}{23},\frac{10}{23},\frac{1}{23}\right)'$

Table 7 provides the mean squared error values for each estimator based on Scenario 2. As in previous scenarios, when m is fixed and K increases, the MSE decreases. The MSE also decreases when K is fixed and m increases. Once again, the proposed semi-parametric Bayesian estimator performs better than the existing estimators and MLE.

Table 7: MSE values for scenario 2.

Estimator	*m*=100			*m*=200			*m*=500		
	K = 5	*K* = 10	*K* = 15	*K* = 5	*K* = 10	*K* = 15	*K* = 5	*K* = 10	*K* = 15
MLE	0.1314	0.071	0.0636	0.1299	0.0698	0.069	0.1295	0.0695	0.0689
James-Stein	0.1278	0.067	0.0614	0.1263	0.0662	0.0608	0.1261	0.0661	0.0607
Empirical Bayes	0.1297	0.069	0.0625	0.1283	0.068	0.0675	0.128	0.0678	0.0674
Dirichlet Process	0.1243	0.0651	0.0602	0.1231	0.0645	0.06	0.1229	0.0644	0.06

Discussion

In this paper, we considered four approaches for modeling bowling performance in T20I cricket. The advantage of the semi-parametric Bayesian approach is that it can accommodate complex and heterogeneous patterns in bowler performance. In particular, the Dirichlet process naturally borrows information across similar players and clusters them together. The cluster assignments of players are obtained as a by-product of the posterior simulation of the Dirichlet process. They are done in such a way that players have identical characteristics within clusters. The DP also seems to help in handling sparsity in the data as estimates for categories with zero counts for most players seem to behave appropriately.

We remark that choosing a suitable shrinkage target (t_j) for James-Stein estimation is challenging. Two shrinkage targets we considered here were the discrete uniform distribution $\frac{1}{K}$ for all categories and the overall proportions. Data analysis suggests that shrinking towards $\frac{1}{K}$ is often less efficient than shrinking towards the overall mean proportion. Note that high shrinkage should generally be interpreted as a greater need to improve the MLE or as a greater lack of confidence in raw estimates based on past data, for instance, when based on small sample size. This being said, confidence in the shrinking target also plays a role here: raw estimates that align with a target tend to be shrunken more than others.

Table 8 provides the ranking based on the bowling statistics and estimates for the top 5 expected wicket takers per match. The total no of wickets and no of matches played are given after the bowler's name in brackets, separated by a comma. The rank based on the expected wickets per match is given. Here we considered three bowling statistics; economy rate, bowling average, and strike rate, which are the popular statistics used to rank bowlers by ICC. The wicket-taking ability is high if the bowler has a lower economy rate, bowling average, and strike rate. The ranks for the bowling statistics are given from lowest to highest value considering all the bowlers. For example, Rashid Khan has the 2nd best bowling average out of all bowlers. Ashok Dinda has the highest wicket per match, but he played very few matches. Since he played a few games, it is clear that the ranking penalizes for the uncertainty, especially EB, through his performance being shrunk more towards the global shrinkage target. Rashid Khan has the highest rank for economy rate and bowling average compared to the other four bowlers. It seems that the Dirichlet process and empirical Bayes approaches rank these five bowlers in a more sensible way than the other approaches.

Table 8: Ranking based on bowling statistics and estimates for the top 5 expected wicket takers per match.

Bowler	A Dinda (8, 16)	K Yadav (20, 39)	D Nannes (14, 27)	R Khan (48, 89)	L Ngidi (13, 24)
Economy Rate	118	67	105	5	162
Bowling Average	3	4	9	2	15
Strike Rate	3	5	6	7	4
DP	3	2	4	1	8
EB	21	3	7	1	13
JS (OP)	5	2	3	1	30
BMR	1	5	2	15	3
JS	3	7	4	17	1

Availability of Data and Materials

All data generated or analyzed during this study are included in this published article.

Conflict of Interest

The authors declare that they have no competing interests.

References

1. Stein C (1956) Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution, University of California Press, Berkeley and Los Angeles (1956), Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability 1: 197-206.
2. Ledoit O, Wolf M (2003) Improved estimation of the covariance matrix of stock return with an application to portfolio selection. Journal of Empirical Finance 10(5): 603-621.
3. James W, Stein C (1961) Estimation with Quadratic Loss, Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, pp. 361-379.
4. Hausser J, Strimmer K (2009) Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. Journal of Machine Learning Research 10: 1469-1484.
5. Efron B, Morris C (1973) Stein's estimation rule and its competitors-an empirical Bayes approach. Journal of the American Statistical Association 68(341): 117-130.
6. Tvedebrink T (2010) Overdispersion in allelic counts and l-correction in forensic genetics. Theoretical Population Biology 200-210.s
7. Wadsworth WD, Argiento R, Guindani M, Galloway-PenaJ, Shelburne SA, et al. (2017) An integrative Bayesian Dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data. BMC Bioinformatics 18: 1-12.
8. Teh YW (2010) Dirichlet processes. Encyclopedia of Machine Learning.
9. Ferguson TS (1973) A Bayesian analysis of some nonparametric problem. The Annals of Statistics, pp. 209-230.
10. Blackwell D, Mac Queen JB (1973) Ferguson distributions via Polya urn schemes. The Annals of Statistics 1(2): 353-355.
11. Teh YW, Jordan MI, Beal M, Blei DM (2006) Hierarchical Dirichlet process. Journal of the American Statistical Association 101: 1566-1581.
12. Sethuraman J (1994) A constructive definition of Dirichlet priors. Statistica Sinica 4: 639-650.
13. Neal RM (2000) Markov chain sampling methods for Dirichlet process mixture models. Journal of Computational and Graphical Statistics 9(2): 249-265.
14. Swartz TB, Gill PS, Muthukumarana S (2009) Modelling and simulation for one-day cricket. The Canadian Journal of Statistics 37(2): 143-160.
15. Swartz TB, Albert J, Glickman ME, Koning RH (2017) Research directions in cricket. Handbook of statistical methods and analyses in sports.
16. Van Staden PJ, Cochran JJ, Bennett J, Albert J (2017) Cricket. The oxford anthology of statistics in sports 1: 2000-2004.
17. Silva R, Perera H, Davis J, Swartz TB (2016) Tactics for Twenty20 cricket. South African Statistical Journal 20(2): 261-271.
18. Davis J, Perera H, Swartz TB (2015) A simulator for Twenty20 cricket. The Australian and New Zealand Journal of Statistics 57(1): 55-71.
19. Koulis T, Muthukumarana S, Briercliffe CD (2014) A Bayesian stochastic model for batting performance evaluation in one-day cricket. Journal of Quantitative Analysis in Sports 10: 1-13.
20. Manage ABW, Scariano SM, Hallum CR (2013) Performance analysis of T20-world cup cricket 2012. Sri Lankan Journal of Applied Statistics 14: 1-12.
21. Van Staden PJ (2009) Comparison of cricketers' bowling and batting performances using graphical displays. Current Science 96: 764-766.
22. Lemmer HH (2004) A measure for the batting performance of cricket players. South African Journal for Research in Sport, Physical Education and Recreation 26: 55-64.
23. Perera H, Davis J, Swartz TB (2015) Assessing the impact of fielding in Twenty20 cricket. Journal of the Operational Research Society 69(8): 1335-1343.