Estimation of Sparse Multinomial Cell Probabilities

by

Lahiru Wickramasinghe

A Thesis submitted to the Faculty of Graduate Studies of
The University of Manitoba
In partial fulfilment of the requirements of the Degree of

Doctor of Philosophy

Department of Statistics
University of Manitoba
Winnipeg

Copyright © 2021 by LAHIRU WICKRAMASINGHE

Abstract

The estimation of multinomial cell probabilities is a challenging problem for sparse multinomial data, where many cells have small and/or zero counts. Borrowing information from other multinomial populations is an approach that allows the derivation of improved estimators, and in particular of estimators that can handle the sparsity. We develop three methodologies which use different type of borrowing techniques for sparse multinomial data to improve the estimation of cell probabilities. First, we propose a semi-parametric Bayesian approach that uses the Dirichlet process to borrow information from other multinomial populations using a natural clustering mechanism. Secondly, we use a weighted likelihood approach that borrows information across similar populations. The likelihood weights for the populations are calculated based on the similarity with the target population. The resulting estimator based on the weighted likelihood approach is a James-Stein type shrinkage estimator. Finally, a Bayesian approach using the smoothed Dirichlet distribution is developed. Specifically, using this family of prior distributions, we develop a different type of scaled shrinkage estimator that allows simultaneous inference for many multinomial populations and the sharing of relevant information between those populations, and across cell categories to provide improved inference under sparsity. The proposed methods are applied to the analysis of T20I cricket bowlers data, of batting metrics in major league baseball, and of the age distribution of early cases

of COVID-19 across Canada.

Keywords: Multinomial distribution; Sparse data; Dirichlet process; Weighted likelihood; Smoothed Dirichlet distribution; Cricket; Baseball; COVID-19

Acknowledgment Page

Surviving graduate school would have been an impossible task without the support I've had in my life over the past six years. First, I would like to thank my advisors, Dr. Saman Muthukumarana and Dr. Alexandre Leblanc, for their guidance, invaluable support, patience, and mentorship throughout my Ph.D. research. This dissertation would not have been possible without their supervision and help.

I would also like to thank my committee, Dr. Katherine Davies and Dr. Mike Domaratzki for their time commitment and dedication in serving on my committee. I would like to thank the external committee member Dr. Farouk Nathoo. The suggestions and constructive feedbacks were greatly appreciated and helped make my thesis a success. I would also like to thank Dr. Kevin Fraser who is the chair of my Ph.D. thesis defence.

Many thanks go to the faculty, the staff members, and the colleagues within the Department of Statistics at the University of Manitoba. I also thank the Faculty of Graduate Studies for the financial support through the University of Manitoba Graduate Fellowship (UMGF).

I would like to thank my mother and father, who supported me and believed in me. You both taught me many valuable lessons in my life, and I know that I have got the best mother and father in the world. I love you both a lot.

Last but certainly not least, I would like to thank my wife Suleka for her love, understanding, continuous support, and never letting me give up. Four years ago, you left everything behind to jump into this adventure without any hesitation, and

for that, I am eternally grateful. I'm always grateful for having you in my life, and I love you very much.

Dedication Page

This thesis is dedicated to my wife, mother, father, father-in-law, mother-in-law, sister, and brother-in-law, who have supported me with heart and soul.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

A categorical variable has values that you can put into a countable number of distinct categories based on a characteristic. A classification is exhaustive when it provides sufficient categories to accommodate all possible values, and the categories are mutually exclusive when each observation can be allocated to only one category. Categorical data are often analyzed using multinomial distributions and sparsity, where many cells with small and/or zero counts, is very common in practice. This sparsity can cause difficulties for cell probability estimation. The Maximum Likelihood Estimator (MLE) is the most popular estimator of multinomial cell probabilities. However, it is known that the MLE performs very poorly under sparsity (Hausser and Strimmer (2009)[9]); that is, the MLE underestimates the true cell probabilities of sparse multinomial distribution, and can even be inconsistent in some scenarios.

Our main objective, with this thesis, is to develop several improved strategies to jointly estimate the cell probabilities of m multinomial populations in the context of sparse data. For each such population ($i = 1, 2, \ldots, m$), we assume the existence of an underlying vector $\mathbf{p_i}$ of multinomial cell probabilities that describes how the total

probability is shared between categories. We propose three improved approaches to jointly estimate the cell probabilities of multinomial populations focusing on the problem of sparsity, and we illustrate the proposed approaches through the motivating examples presented in the next section.

1.1 Motivating Examples

1.1.1 Assessment of Bowlers in T20I Cricket

T20I cricket is the most popular form of cricket. It is a short form of cricket where the two teams have a single inning, and each team bowls a limit of 20 overs. The batsmen try to score runs, and the bowlers try to take wickets to slow down the batsmen's aggressiveness. Here, we consider the number of wickets taken in T20I matches by bowlers between January 01, 2010 and March 11, 2020. The data were extracted from the ESPNcricinfo website (www.espncricinfo.com). We consider $m = 175$ bowlers having recorded at least 16 total wickets in total. Table 1.1 provides the wicket distribution for the top 30 ranked bowlers according to ICC rankings on March 11, 2020. Specifically, each row represents data from a multinomial distribution with $K = 7$ categories associated with one bowler. Here, the category 0W is the no wicket category, 1W is one wicket category, and so on. Note that the last three categories haven't been observed frequently. In particular, only 3 bowlers (Deepak Chahar, Ajantha Mendis, and Yuzvendra Chahal) have taken 6-wickets hauls in T20I matches. This suggests that a lot of sparsity exists within this dataset.

Table 1.1: Wicket distribution of top 30 bowlers.

Bowler	Country	Matches	0W	1W	2W	3W	4W	5W	6W	Wickets	Ranking
Rashid Khan	Afghanistan	48	6	15	14	8	3	2	0	89	1
Mujeeb Ur Rahman	Afghanistan	19	4	9	3	2	1	0	0	25	2
Adam Zampa	Australia	29	11	7	7	4	0	0	0	33	3
Ashton Agar	Australia	24	11	6	4	2	0	1	0	25	4
Tabraiz Shamsi	South Africa	22	9	9	4	0	0	0	0	17	5
Mitchell Santner	New Zealand	43	12	15	12	3	1	0	0	52	6
Imad Wasim	Pakistan	42	14	21	3	2	1	1	0	42	7
Adil Rashid	England	36	9	19	5	3	0	0	0	38	8
Shadab Khan	Pakistan	38	9	15	10	3	1	0	0	48	9
Sheldon Cottrell	West Indies	27	6	10	8	2	1	0	0	36	10
Chris Jordan	England	46	16	13	8	7	2	0	0	58	11
Kane Richardson	Australia	18	8	3	5	2	0	0	0	19	12
Jasprit Bumrah	India	49	11	22	11	5	0	0	0	59	13
Andile Phehlukwayo	South Africa	26	6	10	6	3	1	0	0	35	14
Ish Sodhi	New Zealand	44	12	16	11	5	0	0	0	53	15
Tim Southee	New Zealand	60	24	16	11	8	0	1	0	67	16
Pat Cummins	Australia	28	4	14	8	2	0	0	0	36	17
Mark Watt	Scotland	33	9	12	5	6	0	1	0	45	18
Billy Stanlake	Australia	19	4	7	5	2	1	0	0	27	19
Washington Sundar	India	22	8	10	3	1	0	0	0	19	20
Lakshan Sandakan	Sri Lanka	17	7	7	0	2	1	0	0	17	21
Mohammad Nabi	Afghanistan	77	35	24	12	3	3	0	0	69	22
Mitchell Starc	Australia	31	5	14	7	5	0	0	0	43	23
David Willey	England	28	9	10	4	4	1	0	0	34	24
Faheem Ashraf	Pakistan	26	11	9	3	3	0	0	0	24	25
Lasith Malinga	Sri Lanka	63	18	21	15	6	1	2	0	83	26
Tim Curran	England	22	9	5	7	1	0	0	0	22	27
Alasdair Evans	Scotland	25	4	12	5	3	0	1	0	36	28
Yuzvendra Chahal	India	42	12	17	6	4	2	0	1	55	29
Liam Plunkett	England	21	7	7	4	3	0	0	0	24	30

1.1.2 Batting Metrics in Baseball

Baseball is a popular sport, especially in North America. Here, we consider 2018 Major League Baseball (MLB) batting data for all regular season games taking place between March 29, 2018 and October 12, 2018. The data were extracted from the Baseball-Reference website (www.baseball-reference.com). We consider $m = 556$ players with at least 25 plate appearances. A batter is a person who faces the pitcher and a pitcher is a person who throws the baseball towards the batter. A plate appearance refers to a batter's turn at the plate which takes into account every single time a batter comes up. Table 1.2 provides the batting outcomes for the best 30 batters based on ESPN rankings. Each row represents a multinomial distribution with $K = 11$ batting outcomes associated with one batter: SO - strikeout, GO - ground out, AO - air out, SH - sacrifice hit, SF - sacrifice fly, HBP - hit by a pitch,

BB - bases on balls/walk, S - single, D - double, T - triple, and HR - home run. More description about these batting outcomes, and the sport of baseball in general can be found on the MLB website (www.mlb.com/glossary). Our main interest here lies in the estimation of common batting metrics relying on specific cell probabilities for each player. These metrics are routinely reported in the sports media and are used to assess the performance of batters. It is clear from Table 1.2 that the counts for SH, SF, and T are very small even for a large number of plate appearances. This is an other example of sparsity in multinomial data. Table 1.3 provides the batting outcomes for last 8 batters with the fewest number of PA and it is clear that the counts for SH, SF, HBP, D, T and HR are very small.

Table 1.2: Counts for each category of batting outcome for the top 30 batters in the MLB for the 2018 season

Batter	Team	PA	AB	SO	GO	AO	SH	SF	HBP	BB	S	D	T	HR	Ranking
Christian Yelich	MIL	651	574	135	169	83	0	2	7	68	110	34	7	36	1
J.D. Martinez	BOS	649	569	146	126	109	0	7	4	69	106	37	2	43	2
Mookie Betts	BOS	614	520	91	93	156	0	5	8	81	96	47	5	32	3
Jose Ramirez	CLE	698	578	80	133	209	0	6	8	106	75	38	4	39	4
Nolan Arenado	COL	673	590	122	137	156	1	6	3	73	97	38	2	38	5
Alex Bregman	HOU	705	594	85	136	203	0	3	12	96	87	51	1	31	6
Mike Trout	LAA	607	471	124	75	125	0	4	10	122	80	24	4	39	7
Manny Machado	TOT	709	632	104	153	187	0	5	2	70	113	35	3	37	8
Francisco Lindor	CLE	745	661	107	172	199	3	3	8	70	101	42	2	38	9
Freddie Freeman	ATL	707	618	132	147	148	0	6	7	76	120	44	4	23	10
Trevor Story	COL	656	598	168	103	153	0	4	7	47	89	42	6	37	11
Bryce Harper	WSN	695	550	169	124	120	0	9	6	130	69	34	0	34	12
Javier Baez	CHC	645	606	167	151	112	1	4	5	29	93	40	9	34	13
Charlie Blackmon	COL	696	626	134	164	146	1	2	8	59	115	31	7	29	14
Paul Goldschmidt	ARI	689	593	173	121	127	0	0	6	90	99	35	5	33	15
Matt Carpenter	STL	676	564	158	94	167	0	4	6	102	67	42	0	36	16
Andrew Benintendi	BOS	661	579	106	160	145	2	7	2	71	105	41	6	16	17
Anthony Rendon	WSN	597	529	82	108	176	0	8	5	55	93	44	2	24	18
Mitch Haniger	SEA	683	596	148	141	137	0	7	10	70	102	38	4	26	19
Giancarlo Stanton	NYY	705	617	211	124	118	0	10	8	70	91	34	1	38	20
Khris Davis	OAK	654	576	175	113	146	0	7	12	59	65	28	1	48	21
Nick Markakis	ATL	705	623	80	180	178	0	9	1	72	126	43	2	14	22
Whit Merrifield	KCR	707	632	114	134	192	2	6	6	61	134	43	3	12	23
Scooter Gennett	CIN	638	584	125	134	144	3	5	4	42	125	30	3	23	24
Eugenio Suarez	CIN	606	527	142	109	127	0	6	9	64	91	22	2	34	25
Nick Castellanos	DET	678	620	151	126	158	0	3	6	49	111	46	5	23	26
Anthony Rizzo	CHC	665	566	80	148	178	0	9	20	70	105	29	1	25	27
Trea Turner	WSN	740	664	132	187	165	2	0	5	69	128	27	6	19	28
Rhys Hoskins	PHI	659	558	150	101	170	0	5	9	87	65	38	0	34	29
Michael Brantley	CLE	630	570	60	185	149	1	6	5	48	121	36	2	17	30

Table 1.3: Counts for each category of batting outcome for the last 8 batters with the fewest number of plate appearances

Batter	Team	PA	AB	SO	GO	AO	SH	SF	HBP	BB	S	D	T	HR
Boog Powell	OAK	25	24	6	10	4	0	0	0	1	2	1	1	0
Nate Orf	MIL	25	21	8	6	5	0	0	1	3	1	0	0	1
Andrew Susac	BAL	26	26	12	4	7	0	0	0	0	2	1	0	0
Tony Cruz	CIN	26	26	11	5	6	0	0	0	0	2	1	0	1
Brandon Phillips	BOS	27	23	7	10	3	0	0	0	4	2	0	0	1
Jonathan Davis	TOR	27	25	6	10	4	0	0	1	1	4	1	0	0
Raimel Tapia	COL	27	25	7	5	8	0	0	0	2	1	2	1	1
Phil Gosselin	CIN	28	24	8	8	5	0	0	0	4	2	0	0	1

1.1.3 Age Distribution of COVID-19 Cases in Canada

The coronavirus (COVID-19) is a highly infectious disease that has now led to a global pandemic. Here, we consider early COVID-19 cases of infection in Canada, excluding the Northwest Territories. The distributions of COVID-19 cases identified up to June 25, 2020, for different age groups and health regions are provided in Table 1.4. There are 55 health regions, each being associated with a multinomial distribution having $K = 8$ age categories: <20, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80+. Clearly, a few regions have very small cell counts due to a small number of total COVID-19 cases. This is another example of sparsity that we can hope to address by sharing information across health regions.

1.2 Multinomial Distribution

Data in the form of counts occur often in practice. Some examples of count data are the number of university courses a student registers per semester, the number of children a couple has, and the number of doctor visits per year a person makes. The multinomial distribution is the most popular distribution to model counts when

observations can fall into one of K categories, and is commonly used in health and biological count data applications.

Table 1.4: COVID-19 cases by health region and age groups

Province	Health Region	< 20	20-29	30-39	40-49	50-59	60-69	70-79	80+
AB	Calgary Zone	749	717	1029	1147	762	390	180	255
	Central Zone	7	18	12	15	17	13	4	3
	Edmonton Zone	128	178	175	133	132	92	44	32
	North Zone	42	43	44	35	35	31	15	43
	South Zone	206	181	327	341	156	47	19	12
BC	Fraser	77	210	256	223	280	151	131	163
	Interior	4	18	47	43	37	31	17	2
	Northern	3	7	19	10	13	9	4	0
	Vancouver Coastal	21	72	142	129	184	129	86	197
	Vancouver Island	7	16	22	19	20	16	23	8
MB	Interlake-Eastern	1	5	1	1	3	6	3	0
	Northern	0	1	1	0	1	0	0	0
	Prairie Mountain	2	6	9	2	2	4	1	0
	Southern	5	3	4	7	9	8	2	0
	Winnipeg	10	52	48	37	32	29	13	8
ON	Algoma	1	3	1	4	5	6	3	1
	Brant County	9	24	15	25	25	13	10	9
	Chatham-Kent	4	35	49	38	14	13	2	2
	Durham Region	70	171	186	206	309	194	140	397
	Eastern Ontario	7	14	19	19	34	34	15	22
	Grey Bruce	7	11	12	9	20	22	6	15
	Haldimand-Norfolk	11	56	87	91	58	35	16	71
	Haliburton, Kawartha, Pine Ridge	7	13	19	19	30	34	25	46
	Halton Region	42	84	105	118	146	96	43	96
	Hamilton	58	158	95	99	117	93	69	113
	Hastings and Prince Edward Counties	1	6	3	6	8	4	7	9
	Huron Perth	1	6	9	4	10	16	4	8
	Kingston, Frontenac and Lennox & Addington	5	10	8	8	16	9	8	0
	Lambton	14	40	27	25	46	30	32	71
	Leeds, Grenville and Lanark	12	22	25	28	29	40	54	143
	Middlesex-London	27	122	69	66	102	70	52	102
	Niagara	33	111	94	100	89	79	60	174
	North Bay Parry Sound District	2	3	2	3	6	9	1	2
	Northwestern	3	7	9	5	7	4	0	1
	Ottawa	94	257	263	274	293	237	181	466
	Peel	363	1064	785	842	945	640	356	578
	Peterborough	5	18	11	13	15	14	8	11
	Porcupine	4	2	11	7	10	17	9	7
	Region of Waterloo	43	188	182	183	201	131	98	218
	Renfrew County and District	0	3	6	3	5	2	4	4
	Simcoe Muskoka District	40	93	88	78	88	74	58	58
	Southwestern	3	8	12	10	23	12	10	3
	Sudbury and District	4	14	5	6	18	7	11	2
	Thunder Bay District	9	16	12	21	16	9	5	2
	Timiskaming	2	6	2	0	3	2	2	1
	Toronto	522	1699	1781	1838	2081	1436	974	2473
	Wellington-Dufferin-Guelph	24	58	60	81	68	65	50	60
	Windsor-Essex County	38	297	294	226	178	103	67	154
	York Region	117	381	342	390	526	371	241	492
QC	Quebec	4148	7013	7059	8339	8052	4940	4275	11079
NB	New Brunswick	10	28	23	18	30	27	16	13
NS	Nova Scotia	106	276		276		244		159
SK	Saskatchewan	107	260		241		130		21
NL	Newfoundland	22	38		39	58	57	47	
PE	Prince Edward Island	0	10		8		9		0

We assume that we have m multinomial populations, each having the same set of K possible mutually exclusive categories. Now, denote $(X_{i1}, X_{i2}, \ldots, X_{iK})$ the vector of observed counts for all categories within the i^{th} multinomial population with

associated probabilities $\mathbf{p_i} = (p_{i1}, p_{i2}, \ldots, p_{iK})$. Specifically, we write

$$\mathbf{X_i} = (X_{i1}, X_{i2}, \ldots, X_{iK})^t \quad \sim \quad \text{Multinomial}(n_i, \mathbf{p_i}),$$

for $i = 1, 2, \ldots, m$. Here x_{ij} is the number of times an outcome of category j is observed from the i^{th} multinomial population. The parameters of the multinomial distribution are $n_i > 0$, corresponding to the number of trials (trials are the repetitions of an experiment), and the vector of category specific probabilities $\mathbf{p_i} = (p_{i1}, p_{i2}, \ldots, p_{iK})^t$, where $p_{ij} \geq 0$ and $\sum_{j=1}^{K} p_{ij} = 1$. The support of the multinomial distribution is the set of integer vectors $\mathbf{x_i} = (x_{i1}, x_{i2}, \ldots, x_{iK})^t$, subject to the constraint that $0 \leq x_{ij} \leq n_i$ and $\sum_{j=1}^{K} x_{ij} = n_i$. The probability mass function of $\mathbf{X}_i$, the vector of observed counts from the i^{th} multinomial population, is then

$$f(\mathbf{x_i}; n_i, \mathbf{p_i}) = \frac{n_i!}{x_{i1}! \ldots x_{iK}!} p_{i1}^{x_{i1}} \ldots p_{iK}^{x_{iK}}.$$

The implied marginal distribution for X_{ij} of the count in j^{th} category of i^{th} multinomial population, is binomial $\text{B}(n_i, p_{ij})$, with mean and variance given by

$$\text{E}(X_{ij}) = n_i p_{ij} \qquad \text{and} \qquad \text{Var}(X_{ij}) = n_i p_{ij}(1 - p_{ij}),$$

respectively. The maximum likelihood estimator is the most popular estimator of multinomial cell probability vector. For fixed n_i, the MLE of $\mathbf{p}_i$ is $\hat{\mathbf{p}}_i^{\text{MLE}}$, with components given by

$$\hat{p}_{ij}^{\text{MLE}} = \frac{x_{ij}}{n_i} \qquad \text{for } j = 1, 2, \ldots, K.$$

Note that $\hat{\mathbf{p}}_i^{\text{MLE}}$ is an unbiased, and strongly consistent estimator of $\mathbf{p}_i$. It is also a genuine probability assignment among the K categories since $\sum_{j=1}^{K} \hat{p}_{ij}^{\text{MLE}} = 1, \forall i$.

Finally, the interested reader can consult Rao (1957)[11] for a complete overview of the estimation of multinomial probability vector.

1.3 Sparsity

Sparsity is especially common in setups when there are a large number of classification variables, and/or variables with many levels. This is very common when the sample size (n_i) for a population is similar to or smaller than the number of categories (K), in which case many cells are bound to have small or zero counts. This being said, sparsity can happen even when the sample size is large but not relative to the number of cells (for instance, in motivating example 2 for batters with a low number of plate appearances). There are two types of sparsity:

1. when outcomes were not observed for one or more categories due to these outcomes being unobservable (cell probabilities being zero),

2. when outcomes were not observed for one or more categories due to the limited size of the sample and cell probabilities being small, but not actually zero.

The first type of sparsity is often referred to as structural zeros, referring to outcomes whose counts will be always zero, for any sample size n_i. If some or all structural zeros are known in advance, they could be excluded from the model, or equivalently assigned no probability mass although some work has also been done on identifying structural zeros; see Bishop, Fienberg and Holland (1975)[4]. Our focus is on the second type of sparsity and an example for the second type of sparsity is sampling zeros (as well as very low observed counts). In this case, increasing the number of

effective cell counts by combining different sources of data could help to improve inference. Under normal conditions, the MLE of p_{ij} is consistent and efficient: consistency of $\hat{p}_{ij}^{\text{MLE}}$ means that as the sample size gets large the estimates converge in probability to the true cell probabilities and the efficiency of $\hat{p}_{ij}^{\text{MLE}}$ means that no consistent estimator has lower mean squared error (MSE) than $\hat{p}_{ij}^{\text{MLE}}$ as the sample size increases. Indeed, unfortunately the estimator can perform poorly by underestimating the true cell probabilities under sparsity. For sparse multinomial data, the MLE leads to hard-to-interpret zero probability estimates, and fails to satisfy the sparse asymptotic consistency property in some contexts, i.e.,

$$\sup_j \left| \frac{\hat{p}_{ij}^{\text{MLE}}}{p_{ij}} - 1 \right| \neq o_p(1).$$

Fienberg and Holland (1973)[7] showed that the estimator, which is obtained by combining $\hat{p}_{ij}^{\text{MLE}}$ with an informed guess for p_{ij}, can provide an improved performance over $\hat{p}_{ij}^{\text{MLE}}$. Different techniques to construct this informed guess lead to so called shrinkage estimators, which borrow information across other multinomial populations and cell categories. The resulting estimator can have significantly improved performance in some context. In this thesis, we propose such shrinkage techniques that borrow information across other multinomial populations and cell categories using different strategies in order to improve the estimation of p_{ij}'s.

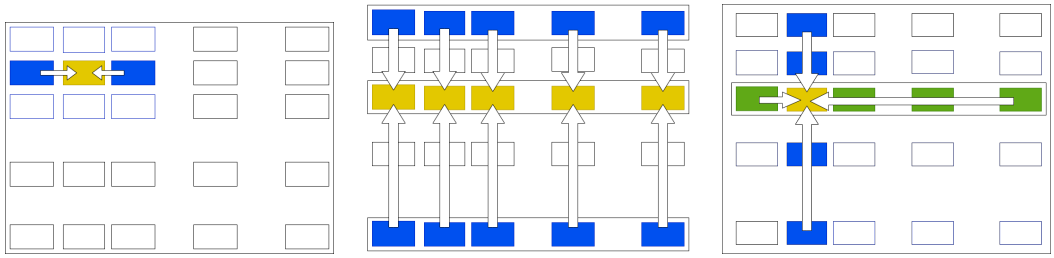
1.4 Different Ways of Borrowing Information

As we mentioned above, different schemes for borrowing information from other available data can improve the estimation of p_{ij}. A first approach, applicable to ordinal categories, is to borrow information from neighboring cells within the same multinomial population; see Figure 1.1(a), to improve the estimation of cell probabilities. This approach is also widely used to borrow information from neighboring cells in a sparse contingency table. Some methodologies have been developed to borrow information from the neighboring cells [see Figure 1.2]. Simonoff (1983)[12] considered an estimator based on a maximum penalized likelihood criterion for sparse multinomial data. Burman (1987)[5] and Hall and Tittering (1987)[8] proposed kernel type estimators for sparse multinomial data, and both of their estimators are sparse asymptotic consistent under some restrictive condition on the true cell probabilities. Dong and Simonoff (1994)[6] used the boundary kernels to relax the some of these conditions. Aerts, Augustyns and Janssen (1997)[1] proposed an estimator based on a local polynomial approach for sparse contingency tables. Albert (2004)[3] shows that the Bayesian paradigm offers a great deal of flexibility to handle the boundary bias and sparseness when analyzing sparse contingency tables. In this approach, the borrowing of information is done within a "block", where the structure of these blocks and the data structure have a huge effect on determining of neighboring cells.

A second approach borrows information across other multinomial populations [see Figure 1.1(b)] instead of borrowing information from a block of neighboring cells to improve the estimation of p_{ij}. In this approach, we identify other multinomial

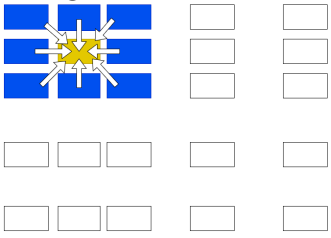
populations that are similar (similar in $\boldsymbol{p}$) to the target population, separately for each category, and borrow information from those similar multinomial populations. Ahmed (2000)[2] demonstrated that the shrinkage estimators are superior to the maximum likelihood estimators of p_{ij}. This second approach is conceptually quite different from the first approach, and we focus on finding different methodologies that borrow information across other multinomial populations. We propose two methodologies that use this approach: one is based on the Dirichlet process, the other on constructing a weighted likelihood scheme.

Figure 1.1: Different ways of borrowing information



(a) Neigbouring cells (b) Across other populations (c) Across other populations and categories

Figure 1.2: Borrowing information in a contingency table



A third approach is combining the first and second approach by borrowing information across both multinomial populations and cell categories [see Figure 1.1(c)] to improve the estimation of multinomial cell probabilities. This approach is applicable and very

useful when the categories are ordinal, i.e., when categories have a natural ordering, guaranteeing that the borrowing information among neighboring cell categories makes sense conceptually. Specifically, we propose a Bayesian approach using the smoothed Dirichlet distribution introduced by Hjort (1996)[10] to specify a prior distribution that naturally leads to the borrowing of information across multinomial populations and cell categories.

1.5 Bayesian Inference

The Bayesian paradigm relax the underlying assumptions and it allows to include the reliable information in addition to the data to produce useful statistical results. There are few advantages of Bayesian methods over classical statistical methods.

1. Relax and clear model assumptions.
2. Ability to update the inference easily when a new observation is received.
3. Systematic incorporation of previous knowledge about the data.
4. Missing information handled easily as part of the estimation process.

Given a statistical model for the data, the Bayesian approach assumes the unknown model parameters have distributions (prior) whereas the classical methods assumes fixed values for the unknown model parameters.

1.5.1 Prior Distribution

The Bayesian methods model the uncertainty about the model parameters using a distribution called prior distribution. The prior is the scientific expert information,

knowledge from the experience and past studies, and one's belief about the unknown parameters before we observe the data. One way to guarantee the unknown model parameters are independent of data is to acquire the prior information before the data have been collected. For example, if you have binary outcomes (yes and no) and if you model the data using a binomial distribution with unknown parameter p, then based on past studies and experiences one can assume a beta distribution as a good choice for the prior distribution. Indeed, the probabilities are in $[0, 1]$ and the support of the beta distribution is also $[0, 1]$.

Informative and Non-Informative Priors

An informative prior expresses specific, definite information about a unknown model parameter. Informative priors are those that deliberately provide information which plays a significant role in the statistical inference. This information is typically derived from previous studies and published works, scientific expert information, and researcher intuition and belief. An example is a prior distribution for the temperature at noon tomorrow. If you know today's noon time temperature, then a reasonable approach is to make the prior a normal distribution with expected value equal to today's noon time temperature, with a small variance.

A non-informative prior provides vague information about a unknown model parameter. Most of the time, we don't have absolutely no subjective knowledge about the unknown model parameters. In such situations, we select non-informative priors and they are generally chosen based on computational convenience. An obvious choice for the non-informative prior is the uniform distribution. Uniform priors are particularly easy to specify in the case of a parameter with bounded support.

Another common non-informative prior is a normal distribution with very large variance.

Conjugate Prior

The concept of conjugacy is a joint property of the prior and data that provides a posterior distribution from the same distributional family as the prior distribution. The primary advantage of using conjugate prior is their mathematical convenience in producing posterior distribution.

Improper Prior

These are the prior distributions that do not add (Probability Mass Function - discrete case) or integrate (Probability Density Function - continuous case) to a finite value. These improper forms typically arise when we use non-informative priors. A very common improper prior is $p(\theta) = c, -\infty < \theta < \infty$ where c is a constant. These improper prior has no preference to any parameter value over any other. The rational behind the improper prior is that if the model is set up so that the data dominates the prior distribution to such an extent that it reduced the complexity of deriving the posterior distribution and the posterior is still proper.

1.5.2 Posterior Distribution

Starting with the prior distribution for unknown model parameters, we update this prior knowledge with the observed data. This revised knowledge about the unknown

model parameters is called posterior distribution. The posterior probability is the probability of the unknown model parameters given the data:

$$\text{Posterior Distribution} \propto \text{Prior Distribution} \times \text{Data}.$$

Specifically, the uncertainty is described by a probability distribution called your prior distribution. The data changes that uncertainty, which is then described by a new probability distribution called your posterior distribution. If you provide a non-informative prior distribution, then the posterior distribution "trust the data" and provides more weight to the data than to the prior information.

In estimation, a Bayes estimator is an estimator which minimizes the posterior expected value of a loss function. The commonly used loss function is the squared error loss function, then the Bayes estimator is simply the mean of the posterior distribution (posterior mean).

1.5.3 Markov Chain Monte Carlo (MCMC) Techniques

To make inference using a Bayesian approach, we need to simulate observation from the posterior distribution. The purpose of the Markov Chain Monte Carlo (MCMC) techniques is to simulate observation from the posterior distribution using an iterative process. The basic principle behind the MCMC technique is to simulate the posterior distribution long enough in an iterative chain. The important property of MCMC is the Markovian property; successive simulated quantities that depend probabilistically only on the values of their immediate predecessor. That means each new observation only depends on the previous observation. There are two main MCMC techniques, Gibbs sampling and Metropolis sampling.

1.6 Main Contributions and Thesis Organization

In the previous section, we discussed different types approaches to improve the estimation of cell probabilities in the context of sparse multinomial data. In this thesis, our main contributions are to propose different techniques to borrow information from other multinomial distributions to improve the estimation of multinomial cell probabilities, each proposed technique achieving this through a different inferential strategy. The following is a summary of the our contributions, which are presented in Chapters 2 to 4.

- In Chapter 2, we introduce a semi-parametric Bayesian approach for simultaneously estimating many multinomial cell probability vectors using a Dirichlet process prior. The Dirichlet process allows to borrow information across multinomial populations while providing a natural clustering mechanism that allows the resulting estimators to handle sparsity. This work was presented in a manuscript that was resubmitted, after a first round of revisions to the Australian and New Zealand Journal of Statistics. This manuscript is the second chapter of this thesis.

- In Chapter 3, we propose an approach to estimate many multinomial probability vectors by using a weighted likelihood methodology. The resulting estimators allow for the borrowing of information across multinomial populations in a category and population specific manner, thus generalizing our previous work. This work was presented in a manuscript has been accepted for publication in the Journal of Applied Statistics and is the third chapter of this thesis.

- In Chapter 4, we develop a Bayesian approach for estimating multinomial cell

probabilities using a smoothed Dirichlet prior. Using a smoothed Dirichlet prior, we develop a different type of scaled shrinkage estimator, which allows simultaneous inference for many multinomial populations and the sharing of relevant information between multinomial populations and across cell categories to provide improved inference under sparsity. This work was presented in a manuscript that was submitted to the Canadian Journal of Statistics and makes up the fourth chapter of this thesis.

Finally, in Chapter 5, we summarize this thesis with some concluding remarks and discussion on future research directions. Note that each chapter is accompanied by its own bibliography. This is a "sandwich thesis", Chapters 2 to 4 consist of papers, which are published or have been submitted for publication. We now briefly outline with a bit more details about the methodological contributions presented in those chapters.

1.6.1 A Semi-parametric Bayesian Approach

In Chapter 2, we introduce a semi-parametric Bayesian approach using the Dirichlet process. The key distinction between classical and Bayesian inference is how uncertainty regarding parameters is treated. In the classical inference, we assume the parameters are unknown and fixed, i.e., repeatable random samples are taken, and the parameters remain constant during this repeatable sampling process. Whereas in the Bayesian inference, we assume the parameters are unknown and random. The randomness of the parameters is associated with using the prior distribution, which encapsulates the uncertainty about the values of the parameters. The prior distribution represents our belief about the different possible values for the parameter

with a probability distribution. We update this prior uncertainty of parameters with current data to produce posterior probability distributions for the parameters that contain less uncertainty. In recent years, there has been a dramatic increase in the application of Bayesian methods, motivated largely by the availability of simple, and efficient methods for posterior computation using the Dirichlet Process. In a semi-parametric Bayesian framework, the Dirichlet process is used as a prior distribution in hierarchical modeling, which naturally borrows information across other multinomial distributions through model-based clustering. The number of clusters is not known a priori and allowing for any finite number of clusters motivates us to use the Dirichlet process. In a Bayesian setting, the terminology *borrowing strength* is referred to as the attempt to improve the precision of estimates by using data from different sources. The Bayesian approach using the Dirichlet process tries to improve the precision of these estimates by borrowing information across other multinomial populations. Also, traditional parametric models use a fixed and finite number of parameters, and the model selection plays a big role in finding the best model. The Bayesian semi-parametric approach is an alternative to these parametric methods, which relaxes some parametric assumptions. In particular, the flexibility of inferences based on the resulting posterior distribution is a really interesting advantage in some setups.

1.6.2 An Approach Based on Weighted Likelihood

The weighted likelihood methodology is presented in Chapter 3. According to the classical likelihood principle, all the information from the data are equally relevant (weighed) to make inferences about the parameters. Whereas in the weighted

likelihood principle, each information is weighted differently and the weights are determined by a suitable weighting function. A biased estimator may sometimes be preferable to the unbiased estimator as the biased one may compensate by featuring a smaller variance than the unbiased estimator. Recently, there has been increasing interest in combining information from different data sources to improve estimation, especially when the sample size is small. For example, when two or more multinomial populations have the cell counts for similar categories, combining the associated datasets through an appropriate weighting scheme may yield more reliable conclusions than those available from a single dataset. To produce inference for a target population, the weighted likelihood methodology allows such borrowing of information across similar populations, and the weights for the populations can be calculated based on similarity with the target population. The proposed estimator based on the weighted likelihood approach is a type of shrinkage estimator that is different from other traditional shrinkage estimators. It looks to improve inference by combining the MLE with another estimator t (shrinkage target) derived from information on all the multinomial populations, but does this in a different way for each category. The likelihood weights adapt by keeping the populations that are similar to the target population and dismissing the ones that are too different. For comparison, both the Bayesian estimator using the Dirichlet process and the estimator based on the weighted likelihood approach borrow information across multinomial distributions to improve the estimation of cell probabilities, and both estimators do this using some form of clustering. In the Dirichlet process, at each iteration of the MCMC algorithm, the clustering assignment will change, and this exploration of the different clustering patterns is used to improve inference. By opposition, the weighted likelihood based approach selects a clustering scheme and

uses that to improve inference. Also, the Bayesian estimator based on the Dirichlet process doesn't have a closed form. The weighted likelihood approach produces a nicely interpretable closed form for the resulting estimator. This estimator is a shrinkage estimator, which allows us to easily see how the borrowing of information is done across other multinomial distributions using the shrinkage targets: t_j (global target), and t_{ij} (population-specific target).

1.6.3 A Bayesian Approach using a Smoothed Dirichlet Prior

In Chapter 4, we introduce a Bayesian approach using a smoothed Dirichlet prior for estimating multinomial cell probabilities. As discussed above, the main advantage of the semi-parametric Bayesian method is that it relaxes some distributional assumptions, naturally adding flexibility to the method for inference. The two methods presented above borrow information across multinomial populations, but do not allow the sharing of information between categories to improve the estimation of cell probabilities. The most important feature of the smoothed Dirichlet prior is that it forces the probabilities of neighboring cells to be closer to each other than under the standard Dirichlet prior. The proposed Bayesian estimator, based on a smoothed Dirichlet prior, is a scaled shrinkage estimator, that allows for simultaneous inference for many multinomial populations by borrowing information across populations and cell categories. The resulting estimator can be viewed as producing double (or two-way) shrinkage in order to provide improved inference. This technique seems especially useful in the context of sparse multinomial data having ordered categories where neighboring categories are considered to be somewhat closer to each other.

Bibliography

[1] M. Aerts, I. Augustyns, and P. Janssen. Smoothing sparse multinomial data using local polynomial fitting. *Journal of Non-parametric Statistics*, 8:127–147, 1997.

[2] S. E. Ahmed. Construction of improved estimators of multinomial proportions. *Communications in Statistics - Theory and Methods*, 29(5-6):1273–1291, 2000.

[3] J. H. Albert. Bayesian methods for contingency tables. *Encylopedia of Biostatistics*, 2004.

[4] Y. Bishop, S. Fienberg, and P. Holland. *Discrete multivariate analysis : theory and practice*. Cambridge, Mass: MIT Press, 1975.

[5] P. Burman. Smoothing sparse contingency tables. *Sankhya: The Indian Journal of Statistics, Series A (1961-2002)*, 49:24–36, 1987.

[6] J. Dong and J. Simonoff. The construction and properties of boundary kernels for smoothing sparse multinomials. *Journal of Computational and Graphical Statistics*, 3(1):57–66, 1994.

[7] S. Fienberg and P. Holland. Simultaneous estimation of multinomial cell probabilities. *Journal of the American Statistical Association*, 68:683–691, 1973.

[8] P. Hall and D. Titterington. On smoothing sparse multinomial data. *Australian Journal of Statistics*, 29:19–37, 1987.

[9] J. Hausser and K. Strimmer. Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. *Journal of Machine Learning Research*, 10:1469–1484, 2009.

[10] N. L. Hjort. *Bayesian Statistics 5*, chapter Bayesian approaches to non- and semiparametric density estimation, pages 223–254. Oxford - University Press, 1996.

[11] C. R. Rao. Maximum likelihood estimation for the multinomial distribution. *Sankhya: The Indian Journal of Statistics (1933-1960)*, 18:139–148, 1957.

[12] J. Simonoff. A penalty function approach to smoothing large sparse contingency tables. *The Annals of Statistics*, 11(1):208–218, 1983.

Chapter 2

Semi-parametric Bayesian Estimation of Sparse Multinomial Probabilities

In this chapter, we propose a semi-parametric Bayesian approach using a Dirichlet process to estimate the cell probabilities of sparse multinomial data. We consider the wicket distribution of the bowler to assess the bowling performance using cell probabilities. Accurately estimating cell probabilities is very important to provide a good comparison of the bowling performance for each bowler. When a player plays a few matches, the estimated cell proportions do not provide the accurate true cell probabilities. The proposed estimation method allows borrowing of information across other players to improve the estimation of cell probabilities using a natural clustering mechanism.

2.1 Introduction

Categorical data are often analyzed using the multinomial distribution, 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 $p_1, p_2, \ldots, p_K$ of occurring. Then, under the first scenario, many cells have small probability p_i 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 sources of data 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 chapter, 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) is known to perform very poorly in this setting. Shrinkage estimation is an approach that allows the derivation of improved estimators, and in particular, it can handle certain forms of sparsity by borrowing information from other mutinomial populations. Our main objective is to develop an improved strategy to jointly estimate the cell probabilities of m multinomial populations in a context of sparse data.

In Section 2.2, we discuss the previously studied statistical models, including James-

Stein (JS) estimation and empirical Bayes (EB) estimation. In Section 2.3, we discuss the proposed statistical models, semi-parametric Bayesian estimation and an estimation based on Bayesian Multinomial regression. In Section 2.4, we apply the methods proposed in Section 2.3 on data consisting of the bowling performance of 175 players over a period of 10 years. We conclude with a short discussion in Section 2.5 on the results and methods presented in this chapter.

2.2 Standard Statistical Models and Estimation

We assume there are m multinomial populations with K categories each. For the i^{th} multinomial population, let X_{ij} be the cell count of category j $(j = 1, 2, \ldots, K)$, and n_i be the total cell counts. Then

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

for $i = 1, 2, \ldots, m$. 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 MLE of p_{ij}, denoted by $\hat{p}_{ij}^{\text{MLE}}(\hat{p}_{ij}^{\text{MLE}} = \frac{X_{ij}}{n_i})$, 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}.

2.2.1 James-Stein Estimation

The concept of shrinkage was first introduced in statistics by Stein (1956)[17] and the general principles behind shrinkage estimation were discussed by Ledoit and Wolf (2003)[8]. The famous James-Stein shrinkage estimator was introduced by James and

Stein (1961)[6] and 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. Specifically, they show that the shrinkage estimation improves the MLE of a multivariate normal mean vector if the dimension is at least 3. Now, 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 (Hausser and Strimmer (2009)[5]). Here λ is called the shrinkage constant and measures the weight that is given to the shrinkage target T. The positive-part James-Stein estimator improves the standard James-Stein estimator [see Shao and Strawderman (1994)[15]], by forcing the shrinkage constant to take a value between 0 and 1, i.e., truncate it at zero if it is negative. The choice of the shrinkage target T should involve only a small number of parameters, but also provide some characteristic of the quantity being estimated. If $\lambda = 1$, the shrinkage estimate equals the shrinkage target T whereas for $\lambda = 0$, it equals U. This strategy has been used to estimate the cell probabilities of a multinomial distribution;

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

where t_j is the shrinkage target. Note that $\hat{p}_{ij}^{\text{JS}}$ is improved over the MLE by combining the population 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 distribution of t_j exist, it can be shown that

$$
\begin{aligned}
E\left(\sum_{j=1}^{K}(\hat{p}_{ij}^{\mathrm{JS}} - p_{ij})^2\right) &= \sum_{j=1}^{K} E(\hat{p}_{ij}^{\mathrm{JS}} - p_{ij})^2 \\
&= \sum_{j=1}^{K} E\left(\hat{p}_{ij}^{\mathrm{JS}} - \hat{p}_{ij}^{\mathrm{MLE}} + \hat{p}_{ij}^{\mathrm{MLE}} - p_{ij}\right)^2 \\
&= \sum_{j=1}^{K} E\left(\lambda_i t_j + (1-\lambda_i)\hat{p}_{ij}^{\mathrm{MLE}} - \hat{p}_{ij}^{\mathrm{MLE}} + \hat{p}_{ij}^{\mathrm{MLE}} - p_{ij}\right)^2 \\
&= \sum_{j=1}^{K} E\left(\lambda_i(t_j - \hat{p}_{ij}^{\mathrm{MLE}}) + (\hat{p}_{ij}^{\mathrm{MLE}} - p_{ij})\right)^2 \\
&= \lambda_i^2 \sum_{j=1}^{K} E\left[(t_j - \hat{p}_{ij}^{\mathrm{MLE}})^2\right] + \sum_{j=1}^{K} E(\hat{p}_{ij}^{\mathrm{MLE}} - p_{ij})^2 \\
&\quad + 2\lambda_i \sum_{j=1}^{K} E\left[(t_j - \hat{p}_{ij}^{\mathrm{MLE}})(\hat{p}_{ij}^{\mathrm{MLE}} - p_{ij})\right] \\
&= \lambda_i^2 \sum_{j=1}^{K} E\left[(t_j - \hat{p}_{ij}^{\mathrm{MLE}})^2\right] + \sum_{j=1}^{K} MSE(\hat{p}_{ij}^{\mathrm{MLE}}) \\
&\quad + 2\lambda_i \sum_{j=1}^{K} E\left[(t_j - E(t_j) + E(t_j) - \hat{p}_{ij}^{\mathrm{MLE}})(\hat{p}_{ij}^{\mathrm{MLE}} - p_{ij})\right]
\end{aligned}
$$

$$
\begin{aligned}
=&\lambda_i^2 \sum_{j=1}^{K} E\left[(t_j - \hat{p}_{ij}^{\mathrm{MLE}})^2\right] + \sum_{j=1}^{K} MSE(\hat{p}_{ij}^{\mathrm{MLE}}) \\
&+ 2\lambda_i \sum_{j=1}^{K} E\left[(t_j - E(t_j))(\hat{p}_{ij}^{\mathrm{MLE}} - p_{ij})\right] \\
&+ 2\lambda_i \sum_{j=1}^{K} E\left[(E(t_j) - \hat{p}_{ij}^{\mathrm{MLE}})(\hat{p}_{ij}^{\mathrm{MLE}} - p_{ij})\right] \\
=&\lambda_i^2 \sum_{j=1}^{K} E\left[(t_j - \hat{p}_{ij}^{\mathrm{MLE}})^2\right] + \sum_{j=1}^{K} MSE(\hat{p}_{ij}^{\mathrm{MLE}}) + 2\lambda_i \sum_{j=1}^{K} Cov(t_j, \hat{p}_{ij}^{\mathrm{MLE}}) \\
&+ 2\lambda_i \sum_{j=1}^{K} E\left[(E(t_j) - p_{ij} + p_{ij} - \hat{p}_{ij}^{\mathrm{MLE}})(\hat{p}_{ij}^{\mathrm{MLE}} - p_{ij})\right] \\
=&\lambda_i^2 \sum_{j=1}^{K} E\left[(t_j - \hat{p}_{ij}^{\mathrm{MLE}})^2\right] + \sum_{j=1}^{K} MSE(\hat{p}_{ij}^{\mathrm{MLE}}) + 2\lambda_i \sum_{j=1}^{K} Cov(t_j, \hat{p}_{ij}^{\mathrm{MLE}}) \\
&- 2\lambda_i \sum_{j=1}^{K} E\left[(\hat{p}_{ij}^{\mathrm{MLE}} - p_{ij})^2\right] + 2\lambda_i \sum_{j=1}^{K} E\left[(E(t_j) - p_{ij})(\hat{p}_{ij}^{\mathrm{MLE}} - p_{ij})\right] \\
=&\lambda_i^2 \sum_{j=1}^{K} E\left[(t_j - \hat{p}_{ij}^{\mathrm{MLE}})^2\right] + \sum_{j=1}^{K} MSE(\hat{p}_{ij}^{\mathrm{MLE}}) + 2\lambda_i \sum_{j=1}^{K} Cov(t_j, \hat{p}_{ij}^{\mathrm{MLE}}) \\
&- 2\lambda_i \sum_{j=1}^{K} Var(\hat{p}_{ij}^{\mathrm{MLE}}) + 2\lambda_i \sum_{j=1}^{K} (E(t_j) - p_{ij}) E\left[(\hat{p}_{ij}^{\mathrm{MLE}} - p_{ij})\right] \\
=&\lambda_i^2 \sum_{j=1}^{K} E\left[(t_j - \hat{p}_{ij}^{\mathrm{MLE}})^2\right] + \sum_{j=1}^{K} MSE(\hat{p}_{ij}^{\mathrm{MLE}}) + 2\lambda_i \sum_{j=1}^{K} Cov(t_j, \hat{p}_{ij}^{\mathrm{MLE}}) \\
&- 2\lambda_i \sum_{j=1}^{K} Var(\hat{p}_{ij}^{\mathrm{MLE}}) + 2\lambda_i \sum_{j=1}^{K} (E(t_j) - E(\hat{p}_{ij}^{\mathrm{MLE}})) Bias(\hat{p}_{ij}^{\mathrm{MLE}})
\end{aligned}
$$

$$
\begin{aligned}
=&\lambda_i^2 \sum_{j=1}^{K} E\left[(t_j - \hat{p}_{ij}^{\mathrm{MLE}})^2\right] + \sum_{j=1}^{K} MSE(\hat{p}_{ij}^{\mathrm{MLE}}) + 2\lambda_i \sum_{j=1}^{K} Cov(t_j, \hat{p}_{ij}^{\mathrm{MLE}}) \\
&- 2\lambda_i \sum_{j=1}^{K} Var(\hat{p}_{ij}^{\mathrm{MLE}}) + 2\lambda_i \sum_{j=1}^{K} (E\left[t_j - \hat{p}_{ij}^{\mathrm{MLE}}\right]) Bias(\hat{p}_{ij}^{\mathrm{MLE}}) \\
=&\sum_{j=1}^{K} MSE(\hat{p}_{ij}^{\mathrm{MLE}}) + \lambda_i^2 \sum_{j=1}^{K} E\left[(t_j - \hat{p}_{ij}^{\mathrm{MLE}})^2\right] \\
&- 2\lambda_i \sum_{j=1}^{K} \left[Var(\hat{p}_{ij}^{\mathrm{MLE}}) - Cov(t_j, \hat{p}_{ij}^{\mathrm{MLE}}) + Bias(\hat{p}_{ij}^{\mathrm{MLE}})(E\left[\hat{p}_{ij}^{\mathrm{MLE}} - t_j\right])\right]
\end{aligned}
$$

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

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

Given that $\hat{p}_{ij}^{\mathrm{MLE}}$ is an unbiased estimator for p_{ij} and following Ledoit and Wolf (2003)[8], we can further simplify the above expression to

$$
\lambda_i^* = \frac{\sum_{j=1}^{K} Var(\hat{p}_{ij}^{\mathrm{MLE}}) - Cov(t_j, \hat{p}_{ij}^{\mathrm{MLE}})}{\sum_{j=1}^{K} E[(t_j - \hat{p}_{ij}^{\mathrm{MLE}})^2]},
$$

which can be estimated using its sample counterpart given by:

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

Note that $\widehat{Var}(\hat{p}_{ij}^{\text{MLE}}) = \dfrac{\hat{p}_{ij}^{\text{MLE}}(1-\hat{p}_{ij}^{\text{MLE}})}{n_i - 1}$ and $\widehat{Cov}(t_j, \hat{p}_{ij}^{\text{MLE}}) = 0$ when $t_j = \dfrac{1}{K}$. In the case where $t_j = \frac{1}{n}\sum_{i=1}^{m} n_i \hat{p}_{ij}^{\text{MLE}}$ (overall proportion of outcome j), the independence between players further leads to $\widehat{Cov}(t_j, \hat{p}_{ij}^{\text{MLE}}) = w_i \widehat{Var}(\hat{p}_{ij}^{\text{MLE}})$ where $w_i = \dfrac{n_i}{\sum_{i=1}^{m} n_i}$.

2.2.2 Empirical Bayes Estimation

One can also take an empirical Bayes approach for estimating p_{ij} by following the development in Efron and Morris (1973)[3]. First, assume that each p_i has the same prior distribution

$$\mathbf{p_i} = (p_{i1}, p_{i2}, \ldots, p_{iK})^t \quad \sim \quad \text{Dirichlet}(\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_K)^t).$$

Then, the posterior distribution of $\mathbf{p_i}$, given the observed counts for player i, is

$$\mathbf{p_i}|\mathbf{x_i}; \boldsymbol{\alpha} \sim \text{Dirichlet}(x_{i1} + \alpha_1, x_{i2} + \alpha_2, \ldots, x_{iK} + \alpha_K).$$

Under the squared error loss function, the Bayes estimator of p_{ij} is

$$\hat{p}_{ij}^{\text{Bayes}} = \frac{x_{ij} + \alpha_j}{\left(n_i + \sum_{j=1}^{K} \alpha_j\right)},$$

and can be further written as

$$\hat{p}_{ij}^{\text{Bayes}} \quad = \quad \lambda_i t_j + (1-\lambda_i)\hat{p}_{ij}^{\text{MLE}},$$

which is a shrinkage estimator with $t_j = \left(\frac{\alpha_j}{\sum_{j=1}^{K} \alpha_j} \right)$ 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 by their sample counterparts,

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

in the above expression for the Bayes estimator. The parameter estimates $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_K)^t$ are obtained by using the `dirmult` package in R (Tvedebrink (2010)[22]), relying on maximum likelihood estimation based on the marginal distribution of the data given $\boldsymbol{\alpha}$, and can be interpreted as using data-driven parameters in the prior distribution. The `dirmult` package uses the Fisher scoring algorithm which was proposed by Paul et al. (2005)[12] to obtain the MLE of the Dirichlet-multinomial distribution.

2.3 Proposed Statistical Models and Estimation

2.3.1 Semi-parametric Bayesian Estimator

We now propose a semi-parametric Bayesian estimator based on the Dirichlet process (DP). Dirichlet processes, introduced by Ferguson (1973)[4], 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

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

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

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

for any measurable subset $A \subset \Theta$ (probability space). 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:

$$\mathbf{X_i}|\mathbf{p_i} \sim \text{Multinomial}(n_i, \mathbf{p_i}) \qquad i = 1, \dots m$$

$$\mathbf{p_i}|G \sim G$$

$$G|\alpha_0, G_0 \sim \text{DP}(\alpha_0, G_0)$$

$$G_0 \sim \text{Dirichlet}(\alpha_1, \alpha_2, \dots, \alpha_K),$$

where $\mathbf{X_i} = (X_{i1}, X_{i2}, \dots, X_{iK})^t$ and $\mathbf{p_i} = (p_{i1}, p_{i2}, \dots, p_{iK})^t$. By following Blackwell and MacQueen (1973)[1], the conditional distribution of G given $\mathbf{p_i}$ is also a DP, i.e.:

$$G|\mathbf{p_i}, \mathbf{p_{-i}} \sim \text{DP} \left(\alpha_0 + m - 1, \frac{1}{\alpha_0 + m - 1} \sum_{j \neq i} \delta_{\mathbf{p_j}} + \frac{\alpha_0}{\alpha_0 + m - 1} G_0 \right)$$

$$\sim \text{DP} \left(\alpha_0 + m - 1, \frac{m-1}{\alpha_0 + m - 1} \frac{\sum_{j \neq i} \delta_{\mathbf{p_j}}}{m-1} + \frac{\alpha_0}{\alpha_0 + m - 1} G_0 \right)$$

with base distribution $\frac{1}{\alpha_0 + m - 1} \left(\sum_{j \neq i} \delta_{\mathbf{p_j}} + \alpha_0 G_0 \right)$ and concentration parameter $\alpha_0 + m - 1$. Here $\delta_{\mathbf{p_{n_s}}}$ is the indicator function such that,

$$\delta_{\mathbf{p_{n_s}}} = \begin{cases} 1 & \text{if} \quad \mathbf{p} = \mathbf{p_{n_s}} \\ 0 & \text{otherwise.} \end{cases}$$

The posterior base distribution is a weighted average between the prior base distribution G_0 and the empirical distribution $\frac{\sum_{j \neq i} \delta_{\mathbf{p_j}}}{m-1}$. 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

the DP converges to the true underlying distribution, i.e., the distribution of true cell probabilities. We refer readers to Teh et al. (2006)[21] for further properties of DP formulation and its properties. Note that one can use the stick-breaking construction, as proposed by Sethuraman(1994)[14], in order to draw samples from a DP. In the stick-breaking construction, we draw

$$\beta_k \sim \text{Beta}(1, \alpha_0) \qquad k = 1, 2, \ldots$$

and construct

$$\pi_1 = \beta_1 \qquad \text{and} \qquad \pi_{n_s} = \beta_{n_s} \prod_{k=1}^{n_s-1} (1 - \beta_k) \qquad n_s = 2, 3, \ldots .$$

Intuitively, consider starting with a stick of unit length and breaking a random proportion β_1 of that stick. The length of this piece gives you the first weight π_1. Then, from the remaining stick, break a random portion β_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 $\pi = \{\pi_{n_s}\}_{n_s=1}^{\infty}$ can be generated to an infinite number of clusters n_s (number of breaks in the stick). This weight π_{n_s} defines the probability that different $\mathbf{p_i}$ share the same "atom" $\mathbf{p_{n_s}}$, thus creating the natural clustering mechanism of the DP. When n_s gets larger and larger, the lengths of the pieces of the stick, or the weights, will tend to get smaller and smaller. This motivates to approximate the Dirichlet process using a finite number of clusters which is very useful in computation. 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 $\mathbf{p}$ using Gibbs sampling approach described in Neal (2000)[11].

2.3.2 Bayesian Multinomial Regression Estimation

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

$$\begin{array}{rcl} \mathbf{x_i}|\mathbf{p_i} & \sim & \text{Multinomial}(n_i, \mathbf{p_i}) \\ \mathbf{p_i}|\boldsymbol{\alpha}_i & \sim & \text{Dirichlet}(\boldsymbol{\alpha}_i = (\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{iK})^t). \end{array}$$

Here $\boldsymbol{\alpha}_i$ are positive and a natural link function is the log-link function leading to

$$\boldsymbol{\alpha_i} = \exp(\boldsymbol{\gamma_i}) \qquad \text{where } \boldsymbol{\gamma_i} = (\gamma_{i1}, \gamma_{i2}, \ldots, \gamma_{iK})^t$$

and assume

$$\gamma_{ij} = \beta_{0j} + \sum_{a=1}^{l} \beta_{aj} Y_{ia} \qquad i = 1, 2, \ldots, m; j = 1, \ldots, K.$$

In this setup, β_{aj} captures the effect of the a^{th} covariate on the j^{th} category. Note that in this model, the multinomial populations have the same $\boldsymbol{\beta}$'s but different $\mathbf{p}$'s

because of their covariates taking different values. We assume normal priors on β_{0j} and β_{aj}. Specifically, we use

$$\begin{array}{rcll} \beta_{0j} & \sim & \mathrm{N}\,(0,1) & \\ \beta_{aj} & \sim & \mathrm{N}\,(0,1) & j = 1,\ldots,K \text{ and } a = 1,\ldots,l. \end{array}$$

In what follows, $\mathbf{x} = (\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_m})$, and let $\boldsymbol{\beta} = (\boldsymbol{\beta_0}, \boldsymbol{\beta_1}, \ldots, \boldsymbol{\beta_K})$, where $\boldsymbol{\beta_j} = (\beta_{0j}, \beta_{1j}, \ldots, \beta_{lj})^t$ and $\mathbf{p} = (\mathbf{p_1}, \mathbf{p_2}, \ldots, \mathbf{p_m})$. Then, we have

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

and

$$\pi(\mathbf{p_i}|\boldsymbol{\alpha_i}) = \frac{\Gamma\left(\sum_{j=1}^{K} \alpha_{ij}\right)}{\prod_{j=1}^{K} \Gamma(\alpha_{ij})} \prod_{j=1}^{K} p_{ij}^{\alpha_{ij}-1}, \tag{2.1}$$

where $\pi(.)$ is the probability density function. Substitute $\alpha_{ij} = \exp(\gamma_{ij}) = \exp(\beta_{0j} + \sum_{a=1}^{l} \beta_{aj} Y_{ia})$ into 2.1, then

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

The posterior distribution of $\mathbf{p}$ and $\boldsymbol{\beta}$ is

$$\pi(\mathbf{p}, \boldsymbol{\beta}|\mathbf{x}) \propto \pi(\mathbf{x}|\boldsymbol{p}) \times \pi(\mathbf{p}|\boldsymbol{\beta}) \times \pi(\boldsymbol{\beta})$$

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

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

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

$$\times \ \exp\left(-\sum_{a=0}^{l} \sum_{j=1}^{K} \frac{\beta_{aj}^2}{2}\right). \tag{2.2}$$

Note that by holding $\boldsymbol{\beta}$ fixed,

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

implying conditional independence (given $\boldsymbol{\beta}$) of the $\mathbf{p}_i$'s with marginal PDF given by

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

Letting $\gamma_{ij}^* = x_{ij} + \exp\left(\beta_{0j} + \sum_{a=1}^{l} \beta_{aj} Y_{ia}\right)$ and $\gamma_{\mathbf{i}}^* = (\gamma_{i1}^*, \gamma_{i2}^*, \ldots, \gamma_{iK}^*)^t$, the posterior conditional distribution of $\mathbf{p_i}$ is

$$\mathbf{p_i} | \boldsymbol{\beta}, \mathbf{x_i} \sim \text{Dirichlet}(\gamma_{\mathbf{i}}^*).$$

When holding $\mathbf{p}$ fixed, however,

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

so that the posterior conditional distribution of $\boldsymbol{\beta}$ (given $\mathbf{p}$) does not belong to a standard family of distributions. To perform inference based on the full posterior distribution (2.2), we developed a Metropolis within Gibbs algorithm to generate values from this distribution which is provided in the Appendix A. 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:

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

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

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

and

$$p_{iK} = 1 - (p_{i1} + \ldots + p_{i(K-1)}) = \frac{1}{1 + \sum_{l=1}^{K-1} \exp(\gamma_{il})} \qquad i = 1, 2, \ldots, m.$$

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 $\boldsymbol{p}_i$ instead of calculating p_{ij}'s separately as the multinomial logistic regression model.

2.4 Application to Inference on Bowling Performance

2.4.1 Some Background Information

The game of cricket was first started in the late 16^{th} century and has become popular globally in the 19^{th} and 20^{th} 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 duration of five days and one-day cricket has a duration 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, which has 105 countries as its members. Swartz, Gill and Muthukumarana (2009)[19], Swartz et al. (2017)[18] and van Staden et al. (2017)[24] provide a comprehensive discussion on various aspects of cricket and its recent statistical research directions.

Bowling, batting, and fielding are the three most important aspects of cricket. A bowler is a player who throws (bowls) the cricket ball to a batsman. A batsman/batter

is a player who hits the cricket ball with a bat to score runs. A fielder is a player who collects the ball after it is struck by the batsman. In cricket, an over consists of six consecutive legal deliveries bowled by a bowler. T20I cricket, which was introduced in England in 2003, is the most popular and the most recent form of cricket. It is a short form of cricket where the two teams have single inning and each team bowls a limit of 20 overs. Twenty20 international cricket can be seen as the most aggressive form of cricket, where batsmen try to hit boundaries (4s or 6s) and bowlers try to bowl dot balls (no run) or to take wickets to slow down the batsmen's aggressiveness.

There has been growing interest in T20I cricket performance analysis in recent literature. Silva et al. (2016)[16] provided a comprehensive overview of tactics in T20 international cricket. A simulator for modelling T20I cricket was proposed by Davis, Perera and Swartz (2015)[2]. Note that there are three major components that lead to success in cricket: batting, bowling and fielding. Koulis, Muthukumarana and Briercliffe (2014)[7], Manage, Scariano and Hallum (2013)[10], van Staden (2009)[23] and Lemmer (2004)[9] assessed the batting performance of players in the Indian Premier League (IPL) and one-day international cricket. Koulis, Muthukumarana and Briercliffe (2014)[7] proposed a Bayesian hidden Markov model and Lemmer (2004)[9] 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. Perera, Davis and Swartz (2015)[13] proposed an approach based on random forests to measure the fielding performance in T20I cricket. In what follows, we study bowling performance of players in T20I cricket using the models introduced in Section 2.3.

2.4.2 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. The dismissal of a batsman is known as a wicket. Here we consider a wicket as the dismissal of a batsman for which the bowler receives the credit for dismissing the batsman. 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 time period, and he was the highest-ranked bowler on March 11, 2020, according to ICC.

Let X_{ij} be the number of matches in which the i^{th} bowler has taken exactly $j-1$ wickets $j = 1, 2, \ldots, 7$ and

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

for each of the $m = 175$ bowlers, and that $p_{ij} (j = 1, 2, \ldots, 7)$ denotes the probability that bowler i records $j-1$ wickets. Table 2.1 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 is rarely achieved and there exists sparsity in this dataset. Note that Deepak Chahar, Ajantha Mendis and Yuzvendra Chahal are the only bowlers to have taken 6-wickets hauls (6-wickets in a single inning) in T20I matches. Deepak Chahar has the best bowling figure

(record) in T20I cricket over all players and Ajantha Mendis has two 6-wickets hauls in T20I matches.

Table 2.1: Number of players with non-zero counts for each wicket category.

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

In Table 2.2, we provide summary statistics for the top 30 ranked bowlers (according to ICC rankings). Note that Mendis has retired from T20I cricket so that he is not included in the top 30 bowlers given in the table.

Table 2.2: 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

The James-Stein estimator of p_{ij} is

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

and we considered two choices for the shrinkage target: the uniform target $t_j = \frac{1}{K} = \frac{1}{7} = 0.143$ for $j = 1, 2, \ldots, 7$, and the overall proportion (op) $\bar{p}_j = \dfrac{\sum_{i=1}^{175} x_{ij}}{\sum_{i=1}^{175}\sum_{j=1}^{7} x_{ij}}$

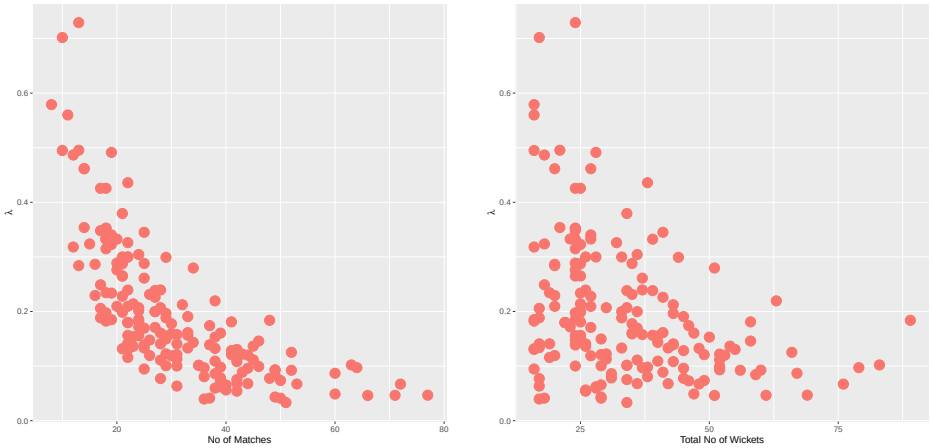
for each category given in Table 2.3. Here, x_{i1} is number of matches in which the i^{th} bowler did not take any wicket, x_{i2} is number of matches in which the i^{th} bowler took exactly one wicket and so on.

Table 2.3: Overall proportion (op) for the 7 wicket categories.

	0W	1W	2W	3W	4W	5W	6W
$\bar{p}_j$	0.350	0.335	0.201	0.087	0.021	0.005	0.001

The James-Stein estimate of the optimal shrinkage constant $\hat{\lambda}_i^*$ for each player is given in Figure 2.1 when the shrinkage target is 0.143. Lungi Ngidi has the maximum optimal shrinkage constant 0.729 and has the highest shrinking towards the shrinkage target. Darren Sammy has the minimum optimal shrinkage constant 0.033 and has the lowest shrinking towards the shrinkage target. The optimal shrinkage constant represents the confidence in our shrinkage target; a low shrinkage constant indicates more confidence on the baseline estimate given by the MLE.

Figure 2.1: Optimal shrinkage constants corresponding to shrinkage target $t_j = \frac{1}{7}$ plotted against (a). the number of matches and (b). the total number of wickets.



(a) (b)

The empirical Bayes estimator of p_{ij} is

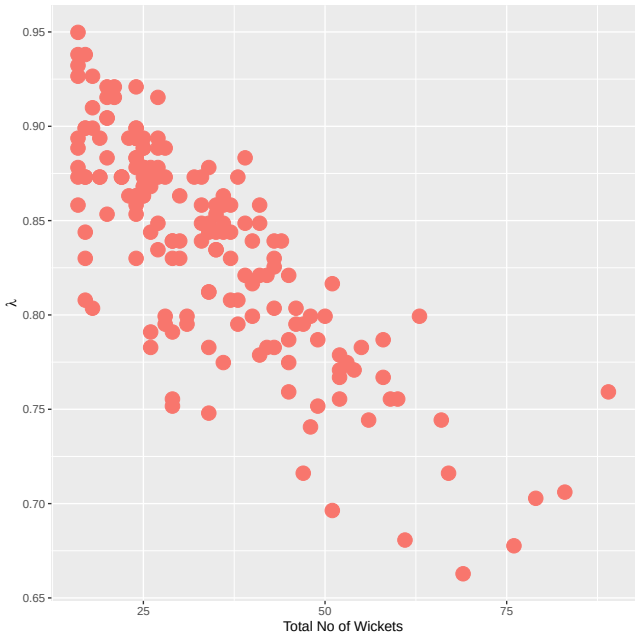
$$\hat{p}_{ij}^{\text{Bayes}} = \hat{\lambda}_i^* \hat{t}_j + (1 - \hat{\lambda}_i^*)\hat{p}_{ij}^{\text{MLE}} \qquad i = 1, 2, \ldots, 175 \text{ and } j = 1, 2, \ldots, 7.$$

Table 2.4 provides the parameter estimates of the concentration parameters $\hat{\boldsymbol{\alpha}}$. While $\hat{\boldsymbol{\alpha}} < \mathbf{1}$ provides a sparse multinomial distribution, $\hat{\boldsymbol{\alpha}} > \mathbf{1}$ provides a smooth multinomial distribution. The values of the concentration parameters for the first five wicket categories are higher than 1 but for the last 2 wicket categories the values are less than 1.

Table 2.4: Estimates of the concentration parameters and the shrinkage targets.

	0W	1W	2W	3W	4W	5W	6W
$\hat{\alpha}_j$	52.27	50.89	30.69	13.35	3.31	0.77	0.10
$\hat{t}_j$	0.345	0.336	0.203	0.088	0.022	0.005	0.001

Figure 2.2: Optimal shrinkage constants for the empirical Bayes method, where the shrinkage target is $\hat{t}_j = \frac{\hat{\alpha}_j}{\sum_{j=1}^{K} \hat{\alpha}_j}$.



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 2.2. Ashok Dinda has the maximum optimal shrinkage constant 0.950 and has the highest shrinkage towards the shrinkage target. Mohammad Nabi has the minimum optimal shrinkage constant 0.663 and has the lowest shrinkage towards the shrinkage target.

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:

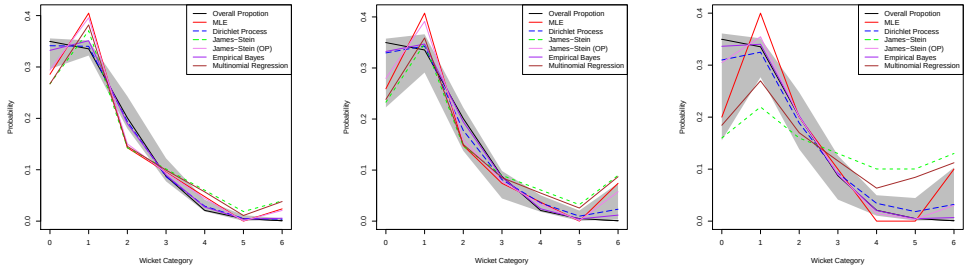
$$G \sim \text{DP}(\alpha_0 = 150, G_0),$$

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

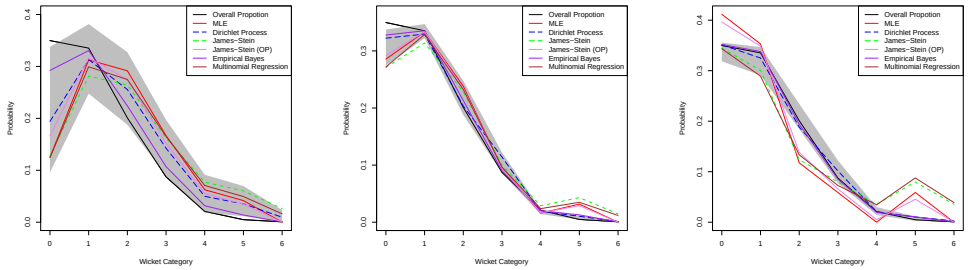
Teh (2010)[20] proposed a formula to find the expected number of atoms/clusters in DP realizations based on α_0 and the number of observations. Based on that formula, we picked $\alpha_0 = 150$ which provides a reasonable number of clusters throughout the posterior simulations. A prior distribution could also have been introduced for α_0 but would have resulted in a considerable increase in model complexity. This increase in complexity is the reason we used Teh's formula to pick a value for α_0 instead.

For implementing the Bayesian regression model, we considered the following covariates: number of overs the bowler bowled, number of runs the bowler conceded in T20 matches from 1$^{\text{st}}$ of January 2010 to 11$^{\text{th}}$ of March 2020, type of bowler (0='Seam', 1='Spin'), the age of the bowler and the economy rate of the bowler. In cricket, the bowler's economy rate is the average number of runs the batsman has scored per over (6 legal consecutive balls) bowled. For the age, we calculate the median age of the bowler for his playing period.

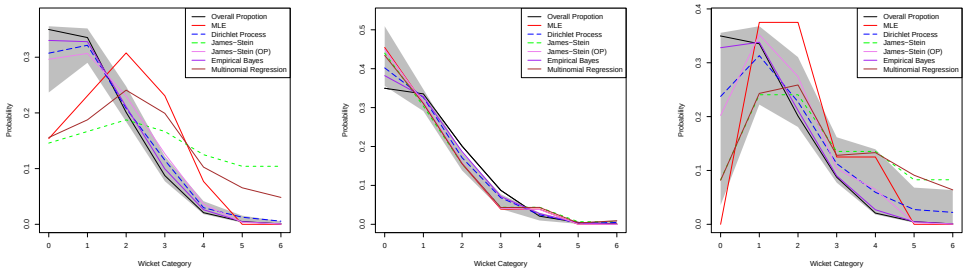
Figure 2.3: Comparison of DP, JS, EB, BMR and ML estimates.



(a) Ajantha Mendis (b) Yuzvendra Chahal (c) Deepak Chahar



(d) Rashid Khan (e) Lasith Malinga (f) Rangana Herath

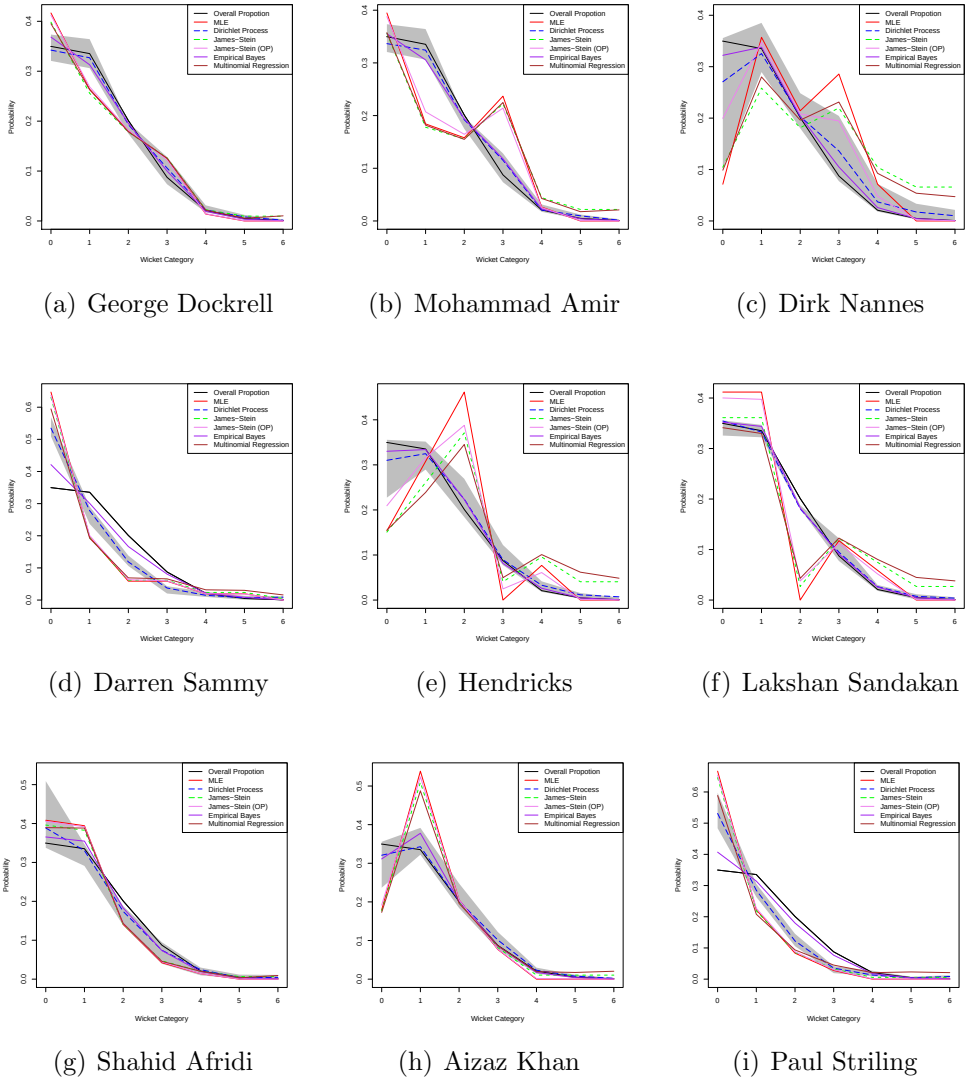


(g) Lungi Ngidi (h) Mohammad Nabi (i) Ashok Dinda

Figure 2.3 and Figure 2.4 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.

Figure 2.4: Comparison of DP, JS, EB, BMR and ML estimates.



(a) George Dockrell (b) Mohammad Amir (c) Dirk Nannes

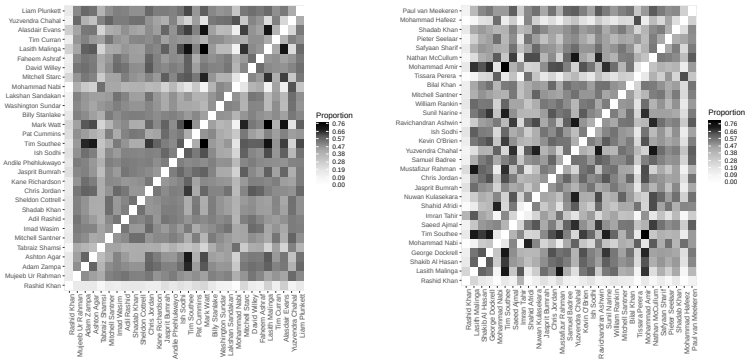
(d) Darren Sammy (e) Hendricks (f) Lakshan Sandakan

(g) Shahid Afridi (h) Aizaz Khan (i) Paul Striling

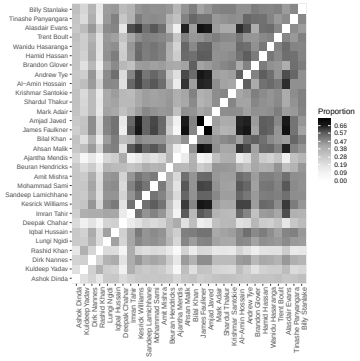
The grey shaded area is the 95% credible interval of Dirichlet Process estimates which shows 95% of Dirichlet Process estimates lie within a particular region. The

reason we included the 95% credible interval of Dirichlet Process estimates is to summarize the uncertainty related to 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 2.3 are for the bowlers who got 6-wickets. We can see clearly, EB estimates, DP estimates are very close to the overall proportion. Rashid Khan and Ashok Dinda have wider 95% credible interval of DP estimates since these two players don't cluster with other players a lot, that means those players have unique cell probabilities which differ from other players.

Figure 2.5: Clustering of players



(a) Top ranked bowlers (b) Top wicket takers



(c) Highest wickets per match

In a single iteration of MCMC, the players are clustered according to whether they share an atom of the current DP realization. In subsequent iterations of MCMC, the cluster membership may differ. With the MCMC output, we are able to calculate the proportion of iterations that any given pair of players cluster together and this provides an estimate of the posterior pairwise probability of clustering. Figure 2.5 is a graphical way of displaying the proportion of clustering of bowlers from DP approach for top ranked bowlers, top wicket takers and highest wicket takers per match. The players with similar performance will cluster with each other most often and the proportion of clustering is large across the posterior simulations. Darker cells represent higher proportion of clustering and lighter cells represent lower proportion of clustering of bowlers.

Table 2.5: Clustering of bowlers based on DP

Ranks	Bowler	1	2	3	4	5
ICC ranking	R Khan	I Tahir (0.399)	K Yadav (0.382)	A Evans (0.274)	L Malinga (0.260)	A Malik (0.251)
	M Ur Rahman	R Ashwin (0.570)	K Ali (0.557)	Y Chahal (0.557)	S Ajmal (0.554)	S Badree (0.550)
	A Zampa	T Southee (0.660)	M Amir (0.630)	G Dockrell (0.618)	U Gul (0.608)	S Narine (0.604)
	A Agar	T Southee (0.680)	M Amir (0.648)	U Gul (0.617)	G Dockrell (0.612)	M Watt (0.611)
	T Shamsi	S Afridi (0.549)	M Nabi (0.546)	N McCullum (0.524)	R Jadeja (0.523)	C Young (0.514)
Top wicket takers	R Khan	I Tahir (0.399)	K Yadav (0.382)	A Evans (0.274)	L Malinga (0.260)	A Malik (0.251)
	L Malinga	M Watt (0.765)	J Faulkner (0.749)	M Rahman (0.718)	T Southee (0.713)	A Malik (0.706)
	S Al Hasan	T Southee (0.665)	M Amir (0.631)	U Gul (0.629)	G Dockrell (0.621)	S Narine (0.609)
	G Dockrell	T Southee (0.722)	M Amir (0.696)	S Narine (0.671)	U Gul (0.645)	K O'Brien (0.635)
	M Nabi	S Afridi (0.699)	M Mortaza (0.648)	R Jadeja (0.627)	N McCullum (0.620)	A Raza (0.618)
Top wicket taker per match	A Dinda	N Kulasekara (0.342)	S Sharif (0.334)	K KC (0.332)	R Ashwin (0.332)	M Santner (0.331)
	K Yadav	I Tahir (0.389)	R Khan (0.382)	L Malinga (0.352)	A Evans (0.340)	A Hossain (0.332)
	D Nannes	M Watt (0.480)	J Faulkner (0.470)	A Javed (0.469)	L Malinga (0.456)	K Williams (0.452)
	R Khan	I Tahir (0.399)	K Yadav (0.382)	A Evans (0.274)	L Malinga (0.260)	A Malik (0.251)
	L Ngidi	M Watt (0.503)	L Malinga (0.498)	T Southee (0.494)	A Javed (0.490)	J Faulkner (0.487)

In Table 2.5, we present the 5 bowlers that are most similar to each of the top 5 ranked bowlers (based on ICC ranking, total wickets and wickets per match)

throughout the iterations of MCMC algorithm, i.e., whether their current vector of outcome probabilities $\mathbf{p}$ are the same atom of the DP for that iteration. Indeed with the MCMC output, we are able to calculate the proportion of iterations that any given pair of bowlers cluster together. This provides an estimate of the posterior pairwise probability that two players share the same vector $\mathbf{p}_i$, suggesting they have the same bowling ability. This posterior probability is provided in Table 2.5 below the bowler's name in brackets. Rashid Khan is the ICC top ranked player and also the top wicket taker. Interestingly, he has smaller estimated posterior pairwise probabilities of clustering compared to other top bowlers, suggesting the DP identifies Rashid Khan as a unique player.

Table 2.6 presents the highest and the lowest posterior pairwise probabilities of clustering in throughout the posterior simulation. From Table 2.6, it is clear that Jean-Paul Duminy, Paul Sterling, Angelo Mathews and Mahmudullah cluster with each other a lot throughout the MCMC iterations, suggesting that these bowlers have very similar bowling ability. Also, Jean-Paul Duminy, Paul Sterling, Angelo Mathews, Mahmudullah and Darren Sammy essentially do not cluster with Rashid Khan. This again points to Rashid Khan having very unique abilities.

Table 2.6: The highest and the lowest posterior pairwise probability of clustering

Highest	Lowest
J Duminy, P Sterling (0.963)	R Khan, J Duminy (0.0006)
A Mathews, P Sterling (0.947)	R Khan, A Mathews (0.0007)
A Mathews, J Duminy (0.946)	R Khan, P Sterling (0.0007)
Mahmudullah, P Sterling (0.931)	R Khan, Mahmudullah (0.0012)
Mahmudullah, J Duminy (0.931)	R Khan, D Sammy (0.0014)

Figure 2.6 is the multi-dimensional scaling (MDS) plot constructed from MLE of all bowlers. The MDS plot is a 2-dimensional graph that helps to visualize high-dimensional data by maintaining the distance between points to help to assess

similarities or dissimilarities between $\mathbf{p}$'s in the data. It is expected that players with smaller dissimilarity may cluster together more often during the posterior simulations. In Figure 2.6, it is clear that Jean-Paul Duminy, Paul Sterling, Angelo Mathews and Mahmudullah (all in the upper left corner of the graph) have small dissimilarity between themselves but all have larger dissimilarity with Rashid Khan (in the lower right corner).

Figure 2.6: Multi-dimensional scaling of all bowlers made from MLEs

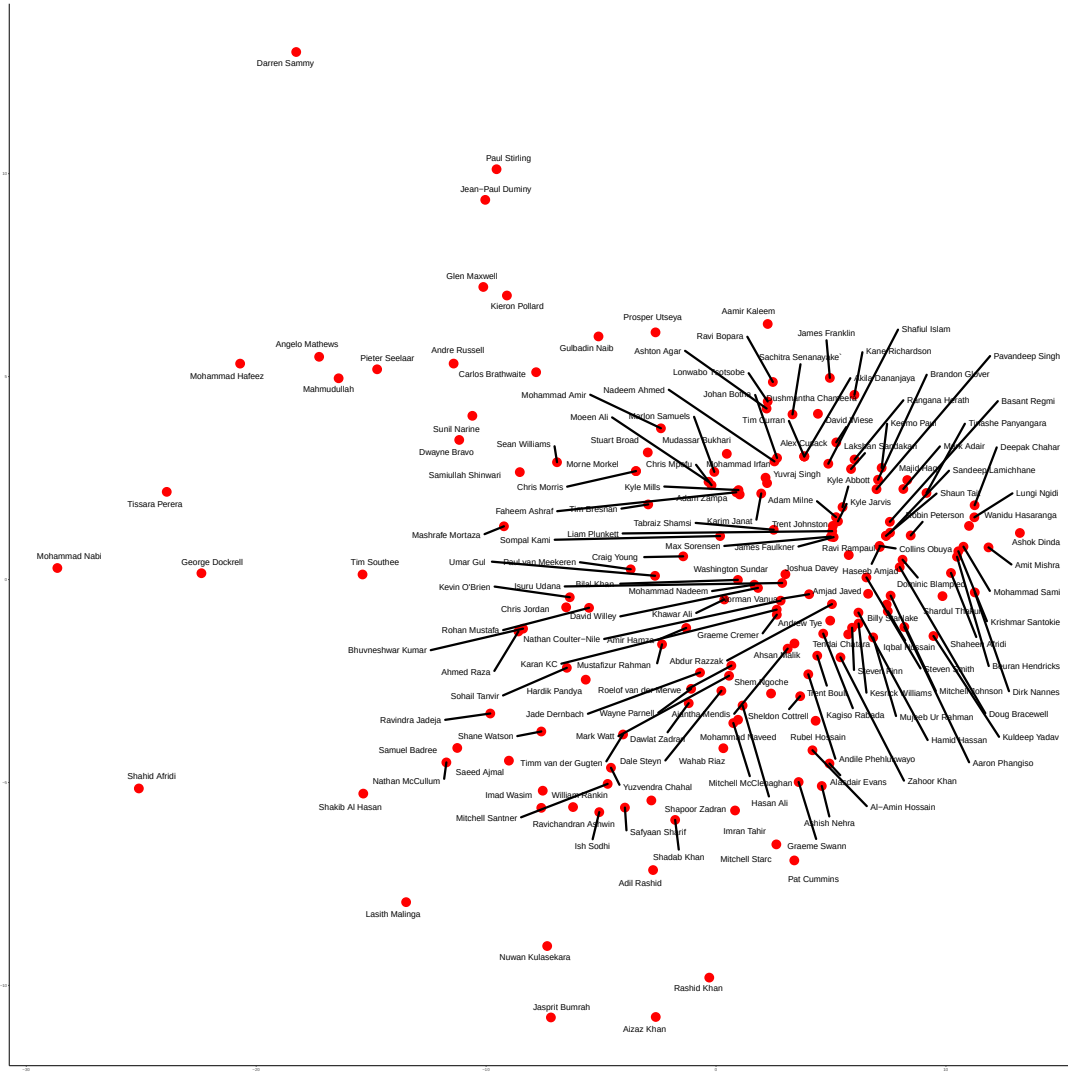


Figure 2.7 represents the 95% posterior credible intervals for β_{aj}'s from the Bayesian multinomial regression model. Posterior means of β_{aj} are distributed mostly around zero and the largest contribution is from β_{26} (runs and 5-wickets). Most of the β_{aj}'s have negative contributions. Whenever a β_{aj} is significant, all the values in the credible interval will be on the same side of zeros (either all positive or all negative) and β_{13}, β_{52}, and β_{56} are the significant β_{aj}'s.

Figure 2.7: 95% credible interval for β_{aj}

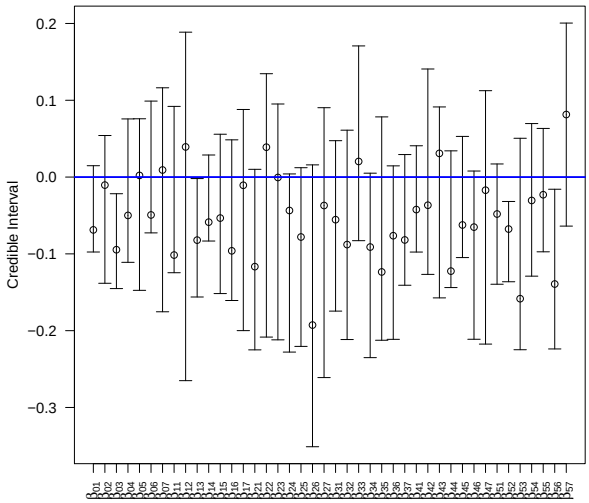


Table 2.7 - 2.9 provide the expected number of wickets per match based on estimates to assess the bowling performances for top ranked bowlers, top wicket takers and the highest wicket takers per match. The expected number of wickets per match (E) is calculated as

$$E_i = 0 \times \hat{p}_{i1} + 1 \times \times \hat{p}_{i2} + \ldots + 6 \times \hat{p}_{i7},$$

where $\hat{p}_{ij}s$ are the cell probability estimates from different methods. The ranks are given from highest to lowest value (the highest is rank 1) and the ties are given to the highest rank using all the bowlers. The first column on Tables 2.8 - 2.9 provide the rank of top wicket takers and the highest wicket takers per match respectively. In Table 2.9, the available ICC ranking is provided inside the brackets in the first column. 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 2.7: Expected number of wickets per match for top ranked bowlers

ICC	Bowler	Original Data		JS		JS (OP)		EB		DP		BMR	
Ranking		Value	Rank	Value	Rank	Value	Rank	Value	Rank	Value	Rank	Value	Rank
1	R Khan	1.85	4	2.06	17	1.72	1	1.30	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	1.40	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	1.20	75	1.14	67	1.17	89	1.40	109
7	I Wasim	1.00	130	1.14	143	1.01	130	1.10	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	1.30	133
9	S Khan	1.26	61	1.49	81	1.24	59	1.15	46	1.18	69	1.50	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	1.20	79	1.35	112	1.20	76	1.14	64	1.20	37	1.38	115
14	A Phehlukwayo	1.35	39	1.73	45	1.29	37	1.16	37	1.18	62	1.70	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	1.20	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	0.90	147	1.09	142	1.12	141	1.39	113
21	L Sandakan	1.00	119	1.38	106	1.02	121	1.11	111	1.14	129	1.56	74
22	M Nabi	0.90	145	0.99	156	0.91	146	1.05	160	1.01	158	1.00	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	1.30	33	1.18	14	1.21	21	1.42	105
27	T Curran	1.00	119	1.36	109	1.02	123	1.11	118.	1.14	126	1.50	84
28	A Evans	1.44	28	1.70	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

Table 2.8: Expected number of wickets per match for top wicket takers

Overall	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	1.30	1	1.68	1	1.97	15
2	L Malinga	1.32	50	1.49	84	1.30	33	1.18	14	1.21	21	1.42	105
3	S Al Hasan	1.23	71	1.41	101	1.22	65	1.16	35	1.18	68	1.34	126
4	G Dockrell	1.06	109	1.19	136	1.06	111	1.10	128	1.17	91	1.17	150
5	M Nabi	0.90	145	0.99	156	0.91	146	1.05	160	1.01	158	1.00	163
6	T Southee	1.12	102	1.28	120	1.12	102	1.12	102	1.19	51	1.25	141
7	S Ajmal	1.27	60.	1.49	86	1.25	57	1.16	32	1.16	112	1.40	110
8	I Tahir	1.66	8	1.95	26	1.54	4	1.23	2	1.33	6	1.81	30
9	S Afridi	0.86	149	0.96	160	0.87	150	1.04	165	1.04	153	0.97	166
10	N Kulasekara	1.22	73	1.39	103	1.21	70	1.15	52	1.16	99	1.39	114
11	J Bumrah	1.20	79	1.35	112	1.20	76	1.14	64	1.20	37	1.38	115
12	C Jordan	1.26	62	1.51	77	1.24	60	1.16	38	1.17	83	1.43	102
13	M Rahman	1.42	31	1.70	51	1.36	22	1.18	9	1.21	26	1.61	68
14	S Badree	1.08	106	1.25	124	1.08	106	1.11	110	1.16	102	1.23	146
15	Y Chahal	1.31	54	1.53	75	1.28	43	1.16	25	1.15	117	1.48	90
16	K O'Brien	1.20	81	1.45	91	1.19	80	1.14	70	1.17	80	1.40	112
17	I Sodhi	1.21	78	1.42	96	1.19	77	1.14	69	1.19	46	1.38	117
18	R Ashwin	1.13	100	1.32	115	1.13	100	1.12	99	1.15	118	1.31	131
19	S Narine	1.06	107	1.24	126	1.07	109	1.11	117	1.16	92	1.24	144
20	W Rankin	1.16	96	1.36	108	1.15	96	1.13	92	1.19	53	1.35	124
21	M Santner	1.21	76	1.43	95	1.20	75	1.14	67	1.17	89	1.40	109
22	B Khan	1.50	16	1.92	30	1.39	13	1.19	6	1.18	67	1.72	46
23	T Perera	0.77	157	0.88	162	0.79	161	1.02	169	0.94	162	0.93	169
24	M Amir	1.32	51	1.57	69	1.28	41	1.16	30	1.19	43	1.55	76
25	N McCullum	0.98	131	1.13	145	0.99	131	1.09	144	1.12	139	1.16	151
26	S Sharif	1.20	84	1.41	99	1.18	81	1.14	75	1.17	87	1.42	108
27	P Seelaar	0.91	143	1.05	151	0.92	145	1.07	153	1.07	147	1.09	155
28	S Khan	1.26	61	1.49	81	1.24	59	1.15	46	1.18	69	1.50	85
29	M Hafeez	0.78	155	0.89	161	0.80	160	1.03	166	0.91	164	0.95	168
30	P van Meekeren	1.21	77	1.49	82	1.19	79	1.14	72	1.16	108	1.45	99

Table 2.9: Expected number of wickets per match for the highest wicket takers per match

Overall	Bowler	Original Data		JS		JS (OP)		EB		DP		BMR	
		Value	Rank	Value	Rank	Value	Rank	Value	Rank	Value	Rank	Value	Rank
1	A Dinda	2.00	1	2.58	3	1.49	5	1.17	21	1.61	3	2.52	1
2 (44)	K Yadav	1.95	2	2.30	7	1.67	2	1.22	3	1.65	2	2.21	5
3	D Nannes	1.93	3	2.42	4	1.55	3	1.19	7	1.44	4	2.29	4
4 (1)	R Khan	1.85	4	2.06	17	1.72	1	1.30	1	1.68	1	1.97	15
5 (72)	L Ngidi	1.85	5	2.69	1	1.31	30	1.18	13	1.30	8	2.29	3
6	I Hussain	1.73	6	2.28	9	1.46	6	1.20	4	1.25	11	2.02	12
7 (43)	D Chahar	1.70	7	2.61	2	1.29	40	1.16	33	1.40	5	2.31	2
8 (39)	I Tahir	1.66	8	1.95	26	1.54	4	1.23	2	1.33	6	1.81	30
9 (37)	K Williams	1.64	9	2.11	14	1.46	7	1.20	5	1.23	16	1.91	19
10 (84)	S Lamichhane	1.62	10	2.14	13	1.43	8	1.18	10	1.22	18	1.98	14
11	M Sami	1.61	11	2.30	6	1.37	19	1.16	29	1.23	14	2.11	7
12	A Mishra	1.60	12	2.29	8	1.36	23	1.15	42	1.24	12	2.19	6
13 (64)	B Hendricks	1.54	13	1.95	25	1.42	10	1.16	36	1.27	9	2.08	10
14	A Mendis	1.52	14	1.85	37	1.43	9	1.18	11	1.27	10	1.78	36
15	A Malik	1.52	15	1.96	21	1.40	12	1.19	8	1.22	19	1.79	34
16 (31)	B Khan	1.50	16	1.92	30	1.39	13	1.19	6	1.18	67	1.72	46
17	J Faulkner	1.50	16	1.96	24	1.38	17	1.17	15	1.23	15	1.83	24
18	A Javed	1.50	16	1.95	27	1.38	16	1.17	18	1.22	20	1.84	23
19 (41)	M Adair	1.50	16	2.00	19	1.37	18	1.16	26	1.20	31	1.95	17
20 (52)	S Thakur	1.50	16	2.03	18	1.36	21	1.16	39	1.20	29	2.02	13
21	K Santokie	1.50	16	2.23	10	1.31	31	1.15	48	1.20	34	2.10	8
22 (49)	A Hossain	1.48	22	1.78	42	1.41	11	1.18	12	1.22	17	1.76	41
23	A Tye	1.48	23	1.88	32	1.38	15	1.17	16	1.20	27	1.78	37
24 (41)	B Glover	1.47	24	2.22	11	1.30	34	1.16	28	1.18	72	1.92	18
25	H Hassan	1.46	25	1.96	22	1.34	25	1.16	23	1.19	44	1.83	25
26 (55)	W Hasaranga	1.46	25	2.32	5	1.26	50	1.14	54	1.20	30	2.09	9
27 (60)	T Boult	1.44	27	1.81	39	1.37	20	1.17	17	1.19	45	1.74	45
28 (28)	A Evans	1.44	28	1.70	50	1.38	14	1.17	19	1.24	13	1.76	40
29	T Panyangara	1.43	29	2.15	12	1.28	44	1.15	50	1.20	40	1.97	16
30 (19)	B Stanlake	1.42	30	1.96	23	1.32	28	1.16	34	1.18	60	1.86	21

2.5 Discussion

In this chapter, we considered four approaches for modelling bowling performance in T20I cricket. The advantage of 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 assignment of players are obtained as a by-product of the posterior simulation of the Dirichlet process and is 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 reduced confidence on raw estimates (estimates obtained from cell counts) based on past data for instance, when based on a small sample size. This being said, confidence in the shrinking target also does play a role here: raw estimates that are in-line with a target tend to be shrunken more than others.

Table 2.10 provides the ranking based on the bowling statistics and estimates for the top 5 expected wicket takers per match. The total number of wickets and number 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. The wicket taking ability is high if the bowler has 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. Askok Dinda has the highest wicket per match but he played very few matches. Since he played 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 other four bowlers. It seems that the Dirichlet process and empirical Bayes approaches rank these five bowlers in a more sensible way compared to the other approaches.

Table 2.10: 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

Bibliography

[1] D. Blackwell and J. B. MacQueen. Ferguson distributions via polya urn schemes. *The Annals of Statistics*, 1(353-355), 1973.

[2] J. Davis, H. Perera, and T.B. Swartz. A simulator for Twenty20 cricket. *The Australian and New Zealand Journal of Statistics*, 57(1):55–71, 2015.

[3] B. Efron and C Morris. Stein's estimation rule and its competitors - an empirical Bayes approach. *Journal of the American Statistical Association*, 68(341):117–130, 1973.

[4] T. S. Ferguson. A Bayesian analysis of some nonparametric problem. *The Annals of Statistics*, pages 209–230, 1973.

[5] J. Hausser and K. Strimmer. Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. *Journal of Machine Learning Research*, 10:1469–1484, 2009.

[6] W. James and C. Stein. Estimation with quadratic loss. volume 1, pages 361–379. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, 1961.

[7] T. Koulis, S. Muthukumarana, and C.D Briercliffe. A Bayesian stochastic model for batting performance evaluation in one-day cricket. *Journal of Quantitative Analysis in Sports*, 10(1-13), 2014.

[8] O. Ledoit and M. Wolf. Improved estimation of the covariance matrix of stock return with an application to portfolio selection. *Journal of Empirical Finance*, 10:603–621, 2003.

[9] H. H. Lemmer. A measure for the batting performance of cricket players. *South African Journal for Research in Sport, Physical Education and Recreation*, 26(55-64), 2004.

[10] A. B. W. Manage, S. M. Scariano, and C. R. Hallum. Performance analysis of T20-world cup cricket 2012. *Sri Lankan Journal of Applied Statistics*, 14:1–12, 2013.

[11] R. M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

[12] S. R. Paul, U. Balasooriya, and T. Banerjee. Fisher information matrix of the Dirichlet-multinomial distribution. *Biometrical Journal*, 47(2):230–236, 2005.

[13] H. Perera, J. Davis, and T.B. Swartz. Assessing the impact of fielding in Twenty20 cricket. *Journal of the Operational Research Society,*, 2015.

[14] J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

[15] P. Y. Shao and W. E. Strawderman. Improving on the James-Stein positive-part estimator. *Annals of Statistics*, 22(3):1517–1538, 1994.

[16] R. Silva, H. Perera, J. Davis, and T.B. Swartz. Tactics for Twenty20 cricket. *South African Statistical Journal*, 20(2):261–271, 2016.

[17] C. Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. volume 1, pages 197–206, Berkeley and Los Angeles, 1956. Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, University of California Press.

[18] T.B. Swartz, J. Albert, M. E. Glickman, and R. H. Koning. *Handbook of statistical methods and analyses in sports*, chapter Research directions in cricket. Chapman & Hall/CRC, 2017.

[19] T.B. Swartz, P. S. Gill, and S. Muthukumarana. Modelling and simulation for one-day cricket. *The Canadian Journal of Statistics*, 37(2):143–160, 2009.

[20] Y. W. Teh. Dirichlet processes. In *Encyclopedia of Machine Learning*. Springer, 2010.

[21] Y.W. Teh, M. I. Jordan, M. Beal, and D. M. Blei. Hierarchical Dirichlet process. *Journal of the American Statistical Association*, 101:1566–1581, 2006.

[22] T. Tvedebrink. Overdispersion in allelic counts and l-correction in forensic genetics. *Theoretical Population Biology*, (200-210), 2010.

[23] P. J. van Staden. Comparison of cricketers' bowling and batting performances using graphical displays. *Current Science*, 96:764–766, 2009.

[24] P. J. van Staden, J. J. Cochran, J. Bennett, and J. Albert. *The oxford anthology of statistics in sports*, volume 1:2000-2004, chapter Cricket. Oxford University Press, 2017.

Chapter 3

Model Based Estimation of Baseball Batting Metrics

In this chapter, we propose a shrinkage estimator using the weighted likelihood methodology to estimate the cell probabilities for sparse multinomial data. To infer about the target population, the weighted likelihood method borrows information across similar populations, and the weights for the populations are calculated based on the similarity with the target population. We consider the distribution of batting outcomes of baseball batters. To predict a good baseball batting metric for each batter, it is very important to estimate the true cell probabilities accurately especially when a batter has a few plate appearances. The weighted likelihood allows the inference on each batter to make use of the batting data from all other batters in the league and, in the process, allows for improved inference.

3.1 Introduction

Major League Baseball (MLB) is a professional baseball organization that was formed in 1903 with the merger of the two U.S. professional baseball leagues: the National

League (NL) and the American League (AL). A total of 30 teams play in MLB: 29 in the United States and 1 in Canada. The teams play 162 games each season and five teams in each league advance to a four-round postseason tournament that culminates in the World Series, a best-of-seven championship series between the two league champions.

Efron and Morris (1975)[6] applied Stein's estimator to a classic example from baseball by considering the player's batting average in the first 50 at bats to predict the player's seasonal batting average. Also, Efron and Morris (1977)[7] clarified why pooling leads to better estimates than taking simple averages using an example from baseball. Albert (2010)[1] introduced three sophisticated baseball databases and conducted some analyses using the R software. Bennett and Flueck (1983)[4] proposed numerous statistics and measures to evaluate offensive performance in baseball. A large part of the literature related to baseball consists of papers related to modeling or predicting the outcomes of a game. For instance, Yang and Swartz (2004)[24] proposed a two-stage Bayesian model which combines the past records, overall batting ability and starting pitcher ERA (Earned Run Average) of two teams to predict the outcomes of a game involving these teams. There are few papers on predicting the batting outcomes and statistics in baseball. For instance, Bailey, Loeppky and Swartz (2018) [3] proposed a linear regression approach to predict the batting average in MLB and Albert (2006)[2] used standard batting statistics and decomposed them into components to make better predictions. In this chapter, we introduce a weighted likelihood and semi-parametric Bayesian approach to model batting outcomes in MLB and use these modeling strategies to propose model-based estimates of common batting metrics.

There are few papers which used the weighted likelihood approach to predict the outcomes of sporting events. Hu and Zidek (2004)[10] used the weighted likelihood approach to forecast NBA basketball playoff outcomes. The weighted likelihood approach was used by Wang and Vandebroek (2013)[19] to design an effective ranking system for soccer teams. To our knowledge, our work presented here is the first that relies on the weighted likelihood methodology to study baseball data.

In Section 3.2, we discuss the batting metrics and the multinomial distribution. The weighted likelihood approach and the shrinkage estimation of multinomial properties using MAMSE (Minimum Averaged Mean Squared Error) weights are discussed in Section 3.3. In Section 3.4, we discuss a semi-parametric Bayesian approach using the Dirichlet process. In Section 3.5, we apply the weighted likelihood approach using MAMSE weights and the Dirichlet process to model batting outcomes for baseball players. We conclude with a short discussion in Section 3.6 of the results and methods presented in this chapter.

3.2 Batting Metrics and the Multinomial Distribution

The outcome of batting in baseball can be divided into discrete categories; this is the basis for constructing metrics that evaluate the batting performance of players. It is also the basis of our analysis. Let x_{ij} be the number of plate appearances in which the batting outcome j occurs for the i^{th} batter ($j = 1, 2, \ldots, K$) and denote the number of plate appearances for the i^{th} batter by n_i. In this chapter, we consider $K = 11$ batting outcomes and the joint distribution of the counts for the K discrete

categories for batter i is given by

$$\mathbf{x_i} = (x_{i1}, x_{i2}, \ldots, x_{iK})^t \quad \sim \quad \text{Multinomial}(n_i, \mathbf{p_i}),$$

where $\mathbf{p_i} = (p_{i1}, p_{i2}, \ldots, p_{iK})^t$ represents the vector of outcome specific probabilities satisfying $\sum_{j=1}^{K} p_{ij} = 1$. The 11 batting outcomes are presented in Table 3.1 and the batting metrics of interest are defined in Tables 3.2 (raw data) and 3.3 (parameter estimates).

Table 3.1: Batting outcomes

Outcome label	Shorthand notation	Outcome category
1	SO	Strike Out
2	GO	Ground Out
3	AO	Air Out
4	SH	Sacrifice Hit
5	SF	Sacrifice Fly
6	HBP	Hit By a Pitch
7	BB	Base on Ball/walk
8	S	Single hit
9	D	Double hit
10	T	Triple hit
11	HR	Home Run

The batting average (BA) is the ratio of the number of hits (S, D, T and HR) to the number of at bats and, in assessing batter performance, assumes all hits have equal weights. In reality, however, home runs are often more important for the outcome of a game. The on-base percentage (OBP) is the rate at which the batter reaches the bases, which includes two extra categories of outcomes: BB and HBP. The slugging percentage (SLG) is the average total bases (TB) per at bat and gives different weight to all hits but doesn't consider BB and HBP. Finally, wOBA combines all the outcomes using yet a different weighted linear combination of each outcome based

on the expected "value" of each type of outcome in terms of producing runs. These weights were obtained from the Fangraphs website (www.fangraphs.com) and were specific to the 2018 season.

Table 3.2: Description of batting statistics based on raw data

Notation	Name	Formula based on raw data
PA	Plate Appearances	$SO + GO + AO + SH + SF + HBP + BB + S + D + T + HR$
AB	At Bats	$SO + GO + AO + S + D + T + HR$
TB	Total Bases	$1 \times S + 2 \times D + 3 \times T + 4 \times HR$
BA	Batting Average	$\frac{S + D + T + HR}{AB}$
OBP	On-Base Percentage	$\frac{S + D + T + HR + BB + HBP}{PA - SH - SF}$
SLG	Slugging Percentage	$\frac{1 \times S + 2 \times D + 3 \times T + 4 \times HR}{AB}$
wOBA	Weighted On-Base Average	$\frac{0.69 \times BB + 0.72 \times HBP + 0.89 \times S + 1.27 \times D + 1.62 \times T + 2.10 \times HR}{PA}$

Table 3.3: Estimation of batting metrics based on multinomial estimates

Notation	Name	Formula based on multinomial probability estimates
$\widehat{BA}$	Batting Average	$\frac{\hat{p}_8 + \hat{p}_9 + \hat{p}_{10} + \hat{p}_{11}}{1 - \hat{p}_4 - \hat{p}_5 - \hat{p}_6 - \hat{p}_7}$
$\widehat{OBP}$	On-Base Percentage	$\frac{\hat{p}_6 + \hat{p}_7 + \hat{p}_8 + \hat{p}_9 + \hat{p}_{10} + \hat{p}_{11}}{1 - \hat{p}_4 - \hat{p}_5}$
$\widehat{SLG}$	Slugging Percentage	$\frac{1 \times \hat{p}_8 + 2 \times \hat{p}_9 + 3 \times \hat{p}_{10} + 4 \times \hat{p}_{11}}{1 - \hat{p}_4 - \hat{p}_5 - \hat{p}_6 - \hat{p}_7}$
w$\widehat{OBA}$	Weighted On-Base Average	$0.69 \times \hat{p}_7 + 0.72 \times \hat{p}_6 + 0.89 \times \hat{p}_8 + 1.27 \times \hat{p}_9 + 1.62 \times \hat{p}_{10} + 2.10 \times \hat{p}_{11}$

3.3 Maximum Weighted Likelihood Estimates

The main purpose of weighted likelihood estimation is to reduce the variance of the traditional maximum likelihood estimator (MLE) in exchange for increasing its bias,

with the goal of reducing the mean squared error (MSE). The choice of weights is a major challenge for maximum weighted likelihood estimation. Wang, van Eeden and Zidek (2002)[21] showed that the maximum weighted likelihood estimator (MWLE) with adaptive weights can have an advantage over the MLE as the estimated MSE for the MWLE can be significantly smaller than that of the MLE in some set ups.

Suppose the data $\mathbf{X_1}, \ldots, \mathbf{X_m}$ come from m distinct populations with probability density functions $f_1(\cdot;\theta_1), \ldots, f_m(\cdot;\theta_m)$ where $\mathbf{X_i} = (X_{i1}, \ldots, X_{in_i})^t$. Further suppose that you want to make inference about population 1, in particular θ_1, an unknown vector of parameters using the other $m-1$ populations. The classical likelihood would be

$$L_1(\theta_1; \mathbf{x_1}) = \prod_{j=1}^{n_i} f_1(x_{1j}; \theta_1),$$

and wouldn't involve any of $\mathbf{X_2}, \ldots, \mathbf{X_m}$, the available data from the other populations. The weighted likelihood relevant to making inference on θ_1 is defined as

$$WL(\theta_1; \mathbf{x}) = \prod_{i=1}^{m} f_1(\mathbf{x_i}; \theta_1)^{w_i} = \prod_{i=1}^{m} \prod_{j=1}^{n_i} f_1(x_{ij}; \theta_1)^{w_i},$$

for fixed $\mathbf{x}$, where $\boldsymbol{w} = (w_1, \ldots, w_m)^t$ is a vector of weights. The weighted likelihood allows the other populations to contribute to the inference on the first population so that the relevant information they contain can be used, acting as if these other populations had the same distributional properties as population 1. These weights can be determined, for instance, based on the similarity of the target population to

each of the other populations. The weighted log-likelihood is then

$$\log\, WL(\theta_1; \mathbf{x}) = \sum_{i=1}^{m} w_i \sum_{j=1}^{n_i} \log\, f_1(x_{ij}; \theta_1),$$

and the maximum weighted likelihood estimator $\hat{\theta}_1$ is a value of θ_1 that maximizes $WL(\theta_1; \mathbf{x})$,

$$\hat{\theta}_1 = \arg\, \sup_{\theta_1 \in \Theta} WL(\theta_1; \mathbf{x}) = \arg\, \sup_{\theta_1 \in \Theta} \sum_{i=1}^{m} w_i \sum_{j=1}^{n_i} \log\, f_1(x_{ij}; \theta_1).$$

Obviously, the choice of the vector of weights $w = (w_1, \ldots, w_m)^t$ is crucial to obtain a resulting estimator with good properties. Interestingly, a few methods have been introduced for determining data-driven likelihood weights. Wang (2001)[20] and Wang and Zidek (2005)[22] introduced cross-validation weights, which were later shown to suffer from instability problems [see Plante (2008)[13]]. In particular, when some of the populations are identical to the target population, the cross-validation weights may not even be defined. Plante (2008, 2009)[13, 14] proposed and studied the minimum averaged mean squared error (MAMSE) weights. Plante's simulation studies showed the WMLE based on MAMSE weights performs better than the unweighted MLE in some contexts. In a fully Bayesian setting, an alternate approach would be to introduce a prior distribution for the weights and consider them as parameters to be estimated. It is unclear however, how successful this approach would be. The advantage of MAMSE weights is that they are adaptive while still ensuring the target population gets a higher weight compared to the other populations.

3.3.1 Minimum Averaged Mean Squared Error Weights

Let F_i be the cumulative density function corresponding to the i^{th} population, and let $\hat{F}_i$ denote the empirical distribution function based on the sample from population i. The weighted empirical distribution is given by

$$\hat{F}_{\boldsymbol{w}} = \sum_{i=1}^{m} w_i \hat{F}_i \qquad \text{with} \qquad w_i \geq 0 \qquad \text{and} \qquad \sum_{i=1}^{m} w_i = 1,$$

and combines data from all populations. One approach to the selection of weights then consists in trying to make $\hat{F}_{\boldsymbol{w}}$ close to F_1 while displaying less variance than $\hat{F}_1$. It is known that inference on F_1 based on $\hat{F}_{\boldsymbol{w}}$ can then outperform inference based on $\hat{F}_1$ for a properly selected set of weights $\boldsymbol{w}$. Plante (2008)[13] suggested to use the minimum averaged mean squared error (MAMSE) weights, which are selected by minimizing the objective function

$$C(\boldsymbol{w}) = \int_{\mathbb{R}} \left[\left(\hat{F}_1(x) - \hat{F}_{\boldsymbol{w}}(x) \right)^2 + \widehat{\text{var}} \left(\hat{F}_{\boldsymbol{w}}(x) \right) \right] \mathrm{d}\hat{F}_1(x),$$

where $\widehat{\text{var}} \left(\hat{F}_{\boldsymbol{w}}(x) \right) = \sum_{i=1}^{m} \frac{w_i^2}{n_i} \hat{F}_i(x)(1 - \hat{F}_i(x))$ subject to the constraints $\{w_i \geq 0, i = 1, \ldots, m\}$ and $\sum_{i=1}^{m} w_i = 1$. The integrand of $C(\boldsymbol{w})$ is composed of two terms: a squared bias term and a variance term. The MAMSE weights calculations can be performed with the MAMSE R package available form the Comprehensive R Archive Network. The asymptotic properties of MAMSE weights are studied by Plante (2009)[14]. One interesting feature of weighted likelihood estimation is that it can lead to James-Stein type shrinkage estimators, e.g. Hu and Zidek (2002)[9]. As a

result, a major benefit of weighted likelihood estimation is that it can significantly reduce the mean squared error as James-Stein type shrinkage estimators often do.

3.3.2 Shrinkage Estimation of Multinomial Cell Probabilities

The distribution of batting outcomes for the i^{th} batter is given by

$$\mathbf{x_i} = (x_{i1}, x_{i2}, \ldots, x_{iK})^t \quad \sim \quad \text{Multinomial}(n_i, \text{prob} = \mathbf{p_i}),$$

where x_{ij} denotes the number of times outcome j was observed for player i and $\mathbf{p_i} = (p_{i1}, p_{i2}, \ldots, p_{iK})^t$ denotes the vector of outcome probabilities for player i. The weighted likelihood for estimating $\mathbf{p_i}$ (using the data available from all players) is given by

$$WL(\mathbf{p_i}|\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_m}) = \prod_{l=1}^{m} L_i(\mathbf{p_i}|\mathbf{x_i})^{w_{il}} \propto \prod_{l=1}^{m}\prod_{j=1}^{K} p_{ij}^{x_{lj}w_{il}} = \prod_{j=1}^{K} p_{ij}^{\sum_{l=1}^{m} x_{lj}w_{il}},$$

where $\mathbf{w_i} = (w_{i1}, w_{i2}, \ldots, w_{im})^t$ are weights assigned to each player when making inference for player i, i.e., on $\mathbf{p_i} = (p_{i1}, p_{i2}, \ldots, p_{iK})^t$. Taking a Bayesian approach, we assume that

$$\mathbf{p_i} = (p_{i1}, p_{i2}, \ldots, p_{iK})^t \quad \sim \quad \text{Dirichlet}(\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_K)^t).$$

Then, the posterior distribution for the i^{th} batter is given by

$$P(\mathbf{p}_i|\mathbf{x}, \alpha, \mathbf{w}_i) \quad \propto \quad \prod_{j=1}^{K} p_{ij}^{\sum_{l=1}^{m} x_{lj}w_{il}+\alpha_j} = \prod_{j=1}^{K} p_{ij}^{x^*_{ij}+\alpha_j},$$

where $x^*_{ij} = \sum_{l=1}^{m} x_{lj} w_{il}$ can be viewed as new re-weighted counts for all j, so that

$$\mathbf{p}_i|\mathbf{x}, \alpha, \mathbf{w}_i \sim \text{Dirichlet}\left(x^*_{i1} + \alpha_1, \ldots, x^*_{iK} + \alpha_K\right).$$

Under the squared error loss function, the Bayes estimator of p_{ij} is

$$\hat{p}_{ij}^{\text{Bayes}} = \frac{\sum_{l=1}^{m} x_{lj} w_{il} + \alpha_j}{\left(\sum_{l=1}^{m} n_l w_{il} + \sum_{j=1}^{K} \alpha_j\right)},$$

and can be further written as

$$\hat{p}_{ij}^{\text{Bayes}} = \lambda_i \hat{p}_{ij}^{\text{MLE}} + (1 - \lambda_i) t_{ij}, \tag{3.1}$$

$$\text{where } \hat{p}_{ij}^{\text{MLE}} = \frac{x_{ij}}{n_i}, t_{ij} = \frac{\left(\sum_{\substack{l=1 \\ l \neq i}}^{m} n_l w_{il} \hat{p}_{lj}^{\text{MLE}} + \alpha_j\right)}{\sum_{j=1}^{K} \left(\sum_{\substack{l=1 \\ l \neq i}}^{m} n_l w_{il} \hat{p}_{lj}^{\text{MLE}} + \alpha_j\right)} \text{ and } \lambda_i = \left(\frac{n_i w_{ii}}{\sum_{l=1}^{m} n_l w_{il} + \sum_{j=1}^{K} \alpha_j}\right). \tag{3.2}$$

The derivation is provided in the Appendix B. Note that in Chapter 2, λ_i is the weight for the shrinkage target t_j. Here t_{ij} is referred to as the shrinkage target and λ_i as the shrinkage constant. Note that the shrinkage target and constant are both player specific, but the shrinkage constant is shared across all categories, i.e., the estimates of all categories are shrunk by the same amount. Also, when estimating $\mathbf{p_i}$, the shrinkage target uses the information of all the batters except that of batter i. When the number of plate appearances of the i^{th} batter (n_i) is large, then λ_i

tends to be large. This means the shrinkage estimator "trusts" the MLE over the shrinkage target and hence gives higher weight to the MLE in this case. Note that both the shrinkage target and constant depend on the choice of the prior parameters $\alpha_1, \alpha_2, \ldots, \alpha_K$. Using the empirical Bayes approach, we suggest to estimate these parameters by maximizing the resulting marginal weighted likelihood. This results in replacing λ_i and t_{ij} with their sample counterparts:

$$\hat{t}_{ij} = \frac{\left(x^*_{ij} + \hat{\alpha}_j - w_{ii}x_{ij}\right)}{\sum_{j=1}^{K}\left(x^*_{ij} + \hat{\alpha}_j - w_{ii}x_{ij}\right)} \text{ and } \hat{\lambda}_i = \left(\frac{n_i w_{ii}}{\sum_{l=1}^{m} n_l w_{il} + \sum_{j=1}^{K}\hat{\alpha}_j}\right),$$

where the parameter estimates $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_k)^t$ can be obtained using the R package `dirmult`[18]. The `dirmult` package uses the Fisher scoring algorithm to obtain the MLE using the Dirichlet-multinomial distribution with the weighted counts.

Alternatively, we can rewrite $\hat{p}_{ij}^{\text{Bayes}}$ as

$$\hat{p}_{ij}^{\text{Bayes}} = \lambda_{1i}\hat{p}_{ij}^{\text{MLE}} + \lambda_{2i}t_j + (1 - \lambda_{1i} - \lambda_{2i})t_{ij}, \tag{3.3}$$

$$\text{where } t_j = \left(\frac{\alpha_j}{\sum_{j=1}^{K}\alpha_j}\right), t_{ij} = \left(\frac{\sum_{\substack{l=1 \\ l\neq i}}^{m} n_l w_{il}\hat{p}_{lj}^{\text{MLE}}}{\sum_{\substack{l=1 \\ l\neq i}}^{m} n_l w_{il}}\right), \lambda_{1i} = \frac{n_i w_{ii}}{\left(\sum_{l=1}^{m} n_l w_{il} + \sum_{j=1}^{K}\alpha_j\right)}$$

$$\text{and } \lambda_{2i} = \frac{\sum_{j=1}^{K} \alpha_j}{\left(\sum_{l=1}^{m} n_l w_{il} + \sum_{j=1}^{K} \alpha_j \right)}. \tag{3.4}$$

Here t_j is a global target (prior mean for that category), t_{ij} is a player-specific target, λ_{1i} is the weight for the MLE and λ_{2i} is the weight for t_j. The global target is a category specific baseline that is shared by all the batters (depending only on the prior distribution) and, as in (3.1), the shrinkage target uses the information of all the batters except that of the batter under study. Note, however, that the roles of the prior mean of each category and of the data-based player specific target are now clearly separated. This was not the case in (3.1). See the Appendix B for the derivation of (3.3). If $\sum_{j=1}^{K} \alpha_j$ is greater than n_i, then the category specific target gets higher weight than the MLE. We can rewrite (3.3) as

$$\hat{p}_{ij}^{\text{Bayes}} = \lambda_{1i} \hat{p}_{ij}^{\text{MLE}} + \lambda_{2i} t_j + \lambda_{3i} t_{ij}, \tag{3.5}$$

where $\lambda_{3i} = (1 - \lambda_{1i} - \lambda_{2i})$. Then, as above, we propose to estimate λ_{1i}, λ_{2i}, t_j and t_{ij} by using their sample counterparts:

$$\hat{t}_j = \left(\frac{\hat{\alpha}_j}{\sum_{j=1}^{K} \hat{\alpha}_j} \right), \hat{t}_{ij} = \left(\frac{\sum_{\substack{l=1 \\ l \neq i}}^{m} n_l w_{il} \hat{p}_{lj}^{\text{MLE}}}{\sum_{\substack{l=1 \\ l \neq i}}^{m} n_l w_{il}} \right), \hat{\lambda}_{1i} = \frac{n_i w_{ii}}{\left(\sum_{l=1}^{m} n_l w_{il} + \sum_{j=1}^{K} \hat{\alpha}_j \right)}$$

$$\text{and } \hat{\lambda}_{2i} = \frac{\sum_{j=1}^{K} \hat{\alpha}_j}{\left(\sum_{l=1}^{m} n_l w_{il} + \sum_{j=1}^{K} \hat{\alpha}_j \right)}, \tag{3.6}$$

where the estimates $\hat{\alpha}_j$ are again obtained through marginal maximum likelihood estimation following the empirical Bayes paradigm. In chapter 2, the Bayes estimator is a weighted average between $\hat{p}_{ij}^{\text{MLE}}$ and t_j. The main reason to have this (3.6) setting is that we want to examine the effect of t_{ij}. In our implementation of the weighted likelihood methodology, we used the MAMSE weights. For practical purposes, the MAMSE weights for each batter are calculated by using a subset of the batters instead of considering all the batters. Details of these calculations are provided in Section 3.5.

3.4 Semi-parametric Bayesian Estimator

We now outline a semi-parametric Bayesian estimation approach based on the Dirichlet process (DP) introduced in Chapter 2. Dirichlet processes, introduced by Ferguson (1973)[8], are a family of stochastic processes whose realizations are probability distributions. These can be seen as distributions over distributions as each draw from a Dirichlet process is itself a distribution. They have been called Dirichlet processes because they form a generalization of the Dirichlet distribution to an infinite number of dimensions, to model the weights of their 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.

We denote the Dirichlet Process by

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

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

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

for any measurable subset $A \subset \Theta$, where Θ is some probability space. When $\alpha_0 \to 0$, all the realizations become concentrated at a single value, while in the limit as $\alpha_0 \to \infty$, the realizations become continuous. We formulate our semi-parametric Bayesian model as follows:

$$\begin{aligned}
\mathbf{x_i}|\mathbf{p_i} &\sim \mathrm{Multinomial}(n_i, \mathbf{p_i}) \qquad i = 1, \ldots m, \\
\mathbf{p_i}|G_i &\sim G_i \\
G_i|\alpha_0, G_0 &\sim \mathrm{DP}(\alpha_0, G_0) \\
G_0 &\sim \mathrm{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_K)
\end{aligned}$$

where $\mathbf{x_i} = (x_{i1}, x_{i2}, \ldots, x_{iK})^t$ again denote the counts of each outcome for player i and $\mathbf{p_i} = (p_{i1}, p_{i2}, \ldots, p_{iK})^t$ is the probability of each batting outcome for player i. By following Blackwell and MacQueen (1973)[5], the conditional distribution of G_i given $\mathbf{p_i}$ is also a DP, i.e.,

$$G_i|\mathbf{p_i}, \mathbf{p_{-i}} \sim \mathrm{DP}\left(\alpha_0 + m - 1, \frac{1}{\alpha_0 + m - 1} \sum_{j \neq i} \delta_{\mathbf{p_j}} + \frac{\alpha_0}{\alpha_0 + m - 1} G_0\right)$$

with base distribution $\frac{1}{\alpha_0 + m - 1}\left(\sum_{j \neq i} \delta_{\mathbf{p_j}} + \alpha_0 G_0\right)$ and concentration parameter $\alpha_0 + m - 1$. Here $\mathbf{p_{-i}}$ considers all $\mathbf{p}$'s excluding $\mathbf{p}_i$, and δ_y stands for a point mass (degenerate distribution) at y. Thus, the posterior base distribution is a weighted average between the prior base distribution G_0 and the empirical distribution $\frac{\sum_{j \neq i} \delta_{\mathbf{p_j}}}{m-1}$. The weights are controlled by the concentration parameter α_0. The larger α_0, 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 DP converges to the empirical distribution of all $\mathbf{p}$'s excluding $\mathbf{p}_i$. More properties of DPs can be found in Teh et al. (2006)[17]. Note that one can use the stick-breaking construction as proposed by Sethuraman (1994)[15] in order to draw samples from a DP.

For this, consider starting with a stick of unit length. Then, break a random proportion $\beta_1 \sim \text{Beta}(1, \alpha_0)$ of the stick. The length of this piece gives you the first weight π_1. Then, from the remaining stick, we again break a random portion $\beta_2 \sim \text{Beta}(1, \alpha_0)$. The length of the second piece gives you the second weight $\pi_2 = \beta_2(1 - \beta_1)$. Now, recursively break the other portions to obtain $\pi_3 = \beta_3(1 - \beta_1)(1 - \beta_2)$ and so forth. Using this construction, an infinite sequence of weights $\pi = \{\pi_n\}_{n=1}^{\infty}$ can be generated, satisfying

$$\pi_1 = \beta_1 \qquad \text{and} \qquad \pi_n = \beta_n \prod_{k=1}^{n-1}(1 - \beta_k) \qquad n \geq 2.$$

When n gets larger and larger, the stick lengths, or the weights, will tend to get smaller and smaller. The stick lengths are influenced by the concentration parameter

α_0. For small α_0, only the first few sticks will have larger length, the remaining sticks all having very small length. On the other hand, for large α_0, the stick lengths will tend to be more uniform and all small. Then the realized discrete random probability distribution is

$$G = \sum_{n=1}^{\infty} \pi_n \delta_{\mathbf{p_n}},$$

which can be used to obtain a random draw for $\mathbf{p}$. As a result, sampling from the posterior distribution of $\mathbf{p}$ following from the above model can be accomplished by using an approach based on Gibbs sampling [see Neal (2000)[12]].

3.5 Data Analysis

The data used to build the model consists of 2018 Major League Baseball (MLB) batting data from the Baseball Reference website (www.baseball-reference.com). We consider data for all the regular season games taking place between March 29, 2018 and October 12, 2018 excluding the league championship series and world series. Our analysis includes $m = 556$ batters with at least 25 plate appearances each. Our main interest is in analyzing the batting outcomes of batters, excluding pitches. Most of the players who have a batting record with less than 25 plate appearances, were pitchers. For that reason we selected the batters with at least 25 plate appearances and still the dataset is very sparse. For comparison, the MLB batting rankings were obtained from the ESPN website (www.espn.com) on October 12, 2018 for 315 batters with the most at bat (these are all the available rankings). Table 3.4 displays the outcomes for the best 30 batters based on the ESPN rankings.

The last 30 batters are the ones with the smallest number of plate appearances among the 556 players considered.

Table 3.4: Counts for each category of batting outcome for the top 30 batters (according to their ESPN ranking) in the MLB for the 2017/18 season

Batter	Team	PA	AB	SO	GO	AO	SH	SF	HBP	BB	S	D	T	HR	Ranking
Christian Yelich	MIL	651	574	135	169	83	0	2	7	68	110	34	7	36	1
J.D. Martinez	BOS	649	569	146	126	109	0	7	4	69	106	37	2	43	2
Mookie Betts	BOS	614	520	91	93	156	0	5	8	81	96	47	5	32	3
Jose Ramirez	CLE	698	578	80	133	209	0	6	8	106	75	38	4	39	4
Nolan Arenado	COL	673	590	122	137	156	1	6	3	73	97	38	2	38	5
Alex Bregman	HOU	705	594	85	136	203	0	3	12	96	87	51	1	31	6
Mike Trout	LAA	607	471	124	75	125	0	4	10	122	80	24	4	39	7
Manny Machado	TOT	709	632	104	153	187	0	5	2	70	113	35	3	37	8
Francisco Lindor	CLE	745	661	107	172	199	3	3	8	70	101	42	2	38	9
Freddie Freeman	ATL	707	618	132	147	148	0	6	7	76	120	44	4	23	10
Trevor Story	COL	656	598	168	103	153	0	4	7	47	89	42	6	37	11
Bryce Harper	WSN	695	550	169	124	120	0	9	6	130	69	34	0	34	12
Javier Baez	CHC	645	606	167	151	112	1	4	5	29	93	40	9	34	13
Charlie Blackmon	COL	696	626	134	164	146	1	2	8	59	115	31	7	29	14
Paul Goldschmidt	ARI	689	593	173	121	127	0	0	6	90	99	35	5	33	15
Matt Carpenter	STL	676	564	158	94	167	0	4	6	102	67	42	0	36	16
Andrew Benintendi	BOS	661	579	106	160	145	2	7	2	71	105	41	6	16	17
Anthony Rendon	WSN	597	529	82	108	176	0	8	5	55	93	44	2	24	18
Mitch Haniger	SEA	683	596	148	141	137	0	7	10	70	102	38	4	26	19
Giancarlo Stanton	NYY	705	617	211	124	118	0	10	8	70	91	34	1	38	20
Khris Davis	OAK	654	576	175	113	146	0	7	12	59	65	28	1	48	21
Nick Markakis	ATL	705	623	80	180	178	0	9	1	72	126	43	2	14	22
Whit Merrifield	KCR	707	632	114	134	192	2	6	6	61	134	43	3	12	23
Scooter Gennett	CIN	638	584	125	134	144	3	5	4	42	125	30	3	23	24
Eugenio Suarez	CIN	606	527	142	109	127	0	6	9	64	91	22	2	34	25
Nick Castellanos	DET	678	620	151	126	158	0	3	6	49	111	46	5	23	26
Anthony Rizzo	CHC	665	566	80	148	178	0	9	20	70	105	29	1	25	27
Trea Turner	WSN	740	664	132	187	165	2	0	5	69	128	27	6	19	28
Rhys Hoskins	PHI	659	558	150	101	170	0	5	9	87	65	38	0	34	29
Michael Brantley	CLE	630	570	60	185	149	1	6	5	48	121	36	2	17	30

In our analysis, we consider $K = 11$ possible outcomes to batting: SO - strikeout, GO - ground out , AO - air out, SH - sacrifice hit , SF - sacrifice fly, HBP - hit by a pitch, BB - bases on balls/walk, S - single, D - double, T - triple and HR - home run. There are many batting metrics to measure the performance of batters. Among those are the Batting Average (BA), the On-Base Percentage (OBP), the Slugging Percentage (SLG), the Weighted On-Base Average (wOBA) introduced in Section 3.2. In Table 3.6, we provide these metrics for the top 30 batters.

Table 3.5 provides the MLB-wide overall proportion (op) $\bar{p}_j = \frac{\sum_{i=1}^{556} x_{ij}}{\sum_{i=1}^{556}\sum_{j=1}^{11} x_{ij}}$ for

the 11 outcomes. We can clearly see that the outcome with the highest percentage is ground outs (GO) (22.7%) and with the lowest is Sacrifice Hits (SH) (0.2%).

Table 3.5: Overall league proportion (op) for the 11 outcomes

	SO	GO	AO	SH	SF	HBP	BB	S	D	T	HR
$\bar{p}_j$	0.216	0.227	0.225	0.002	0.007	0.011	0.086	0.144	0.046	0.005	0.031

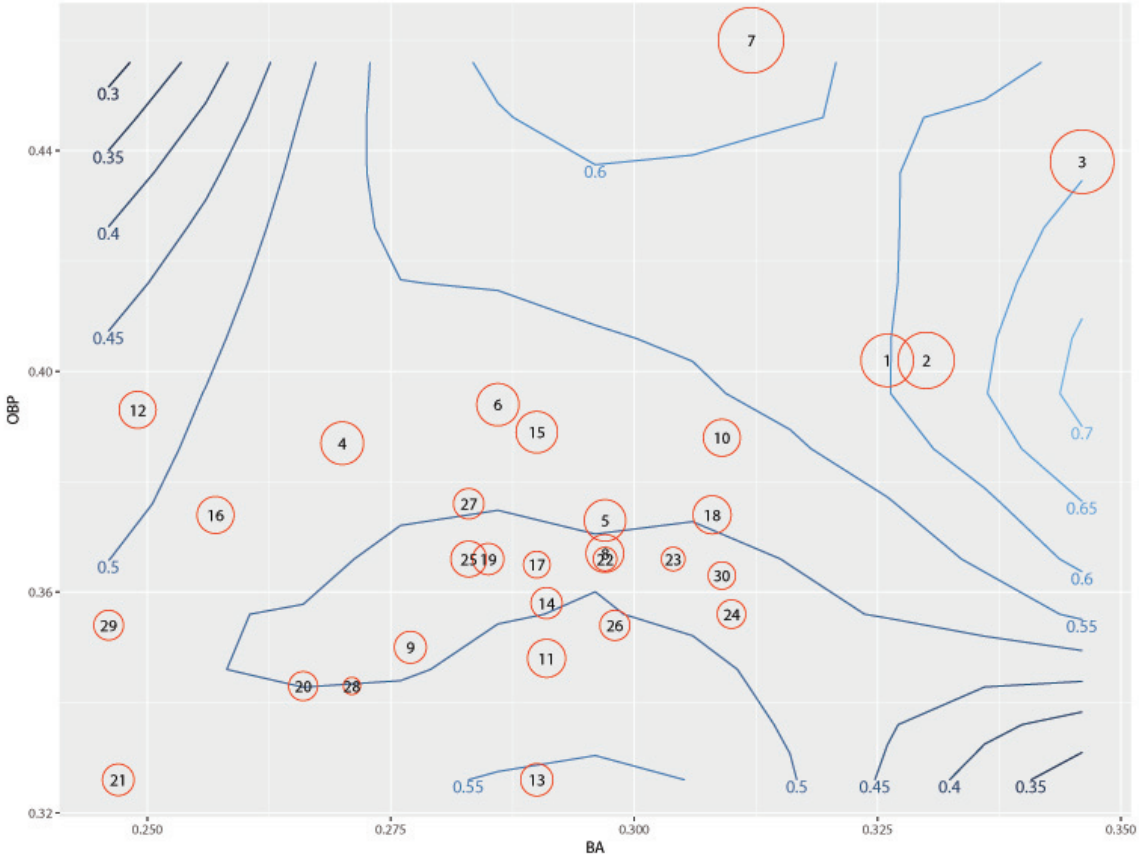


Figure 3.1: Comparison of BA, OBP, SLG and wOBA for the top 30 batters in the MLB for the 2017/18 season.

Figure 3.1 compares the batting performance of the top 30 batters from the 2017/18 MLB season. The top 30 batters are plotted on a scatter plot, with the X-axis

as BA and the Y-axis as OBP. Then, the plot was augmented by adding contours that represent SLG. To take account for wOBA, circles were added to represent each observation: the radius is proportional to wOBA. The number inside the circle represents the ESPN ranking of the batter. The better batters will tend to have higher values of BA, OBP, SLG and wOBA, and hence they should appear towards the upper right-hand corner of the graph with relatively larger circles, whereas poor batters will tend to have lower values of these metrics, and hence they should appear towards the lower left-hand corner of the graph with relatively smaller circles. As one would expect, there are no batters in the higher left-hand corner or lower right-hand corner of the graph as these would correspond to unusual batting patterns, at least among the best batters. Each batting metric does however give weight to each type of offensive contribution differently, and, in particular, wOBA is the only metric considering runs scored.

Among the top 30 batters, Mookie Betts (ranking 3) has the highest BA and SLG. Mike Trout (ranking 7) has the highest OBP and wOBA among the considered batters, which is due to his very large number of BB. Among the considered batters, Khris Davis (ranking 21) has the lowest OBP and the second lowest BA. Javier Baez (ranking 13) has the lowest OBP because he has the smallest number of BB. Trea Turner (ranking 28) has the smallest wOBA because he has the largest single percentage out of hits (71%). For this season, he also has the lowest SLG among top 30 batters. Khris Davis (ranking 21) has the highest HR percentage and Whit Merrifield (ranking 23) has the lowest HR percentage per PA or AB among top 30 batters. ESPN includes runs, RBIs (Runs Batted In) and team win percentage in their evaluation of batters and hence, in producing their rankings. Mookie Betts

(ranking 3) and Mike Trout (ranking 7) have higher BB and a lower number of runs batted in (RBIs) when compared to the best batters; this could explain why those two players aren't at the top 2 ESPN ranks although they have the highest OBP and very high BA.

Table 3.6: Empirical or "raw" batting metrics for the top 30 batters in the MLB for the 2017/18 season (according to ESPN)

Batter	BA	OBP	SLG	wOBA
Christian Yelich	0.326	0.402	0.598	0.430
J.D. Martinez	0.330	0.402	0.629	0.440
Mookie Betts	0.346	0.438	0.640	0.459
Jose Ramirez	0.270	0.387	0.552	0.404
Nolan Arenado	0.297	0.373	0.561	0.401
Alex Bregman	0.286	0.394	0.532	0.403
Mike Trout	0.312	0.460	0.628	0.464
Manny Machado	0.297	0.367	0.538	0.391
Francisco Lindor	0.277	0.350	0.519	0.376
Freddie Freeman	0.309	0.388	0.505	0.389
Trevor Story	0.291	0.348	0.567	0.392
Bryce Harper	0.249	0.393	0.496	0.389
Javier Baez	0.290	0.326	0.554	0.377
Charlie Blackmon	0.291	0.358	0.502	0.374
Paul Goldschmidt	0.290	0.389	0.533	0.401
Matt Carpenter	0.257	0.374	0.523	0.389
Andrew Benintendi	0.290	0.365	0.465	0.362
Anthony Rendon	0.308	0.374	0.535	0.392
Mitch Haniger	0.285	0.366	0.493	0.374
Giancarlo Stanton	0.266	0.343	0.509	0.368
Khris Davis	0.247	0.326	0.549	0.375
Nick Markakis	0.297	0.366	0.440	0.354
Whit Merrifield	0.304	0.366	0.438	0.354
Scooter Gennett	0.310	0.356	0.490	0.367
Eugenio Suarez	0.283	0.366	0.526	0.386
Nick Castellanos	0.298	0.354	0.500	0.371
Anthony Rizzo	0.283	0.376	0.470	0.372
Trea Turner	0.271	0.343	0.416	0.337
Rhys Hoskins	0.246	0.354	0.496	0.370
Michael Brantley	0.309	0.363	0.468	0.364

To calculate MAMSE weights for each player, instead of considering all the batters, we considered the contribution of 10 batters: the batter at hand and a group of 9 others selected based on a clustering argument. We restricted ourselves to clustering the batters who have similar "pattern" of cell probabilities to obtain MAMSE

weights to avoid the convergence and instability issues that arise when we use all the batters. More about these issues is presented in the discussion section below. We use clustering to identify batters which are somehow similar to each other. To measure similarity, it is common to refer to dissimilarity and/or distance. Two batters are then considered similar or close when their dissimilarity or distance is small or their similarity large. For our baseball data, we measure the dissimilarity between the cell probabilities of batters i and l $(\mathbf{p_i}, \mathbf{p_l})$, i.e., our targets for inference, but covariates could be used. We created three categories, SO, GO and AO for outs, which isn't important for all the metrics, but in used in the clustering step when relying on batting pattern as we do here. A variety of measures are available but the most popular dissimilarity measures are distance measures. The Euclidean Distance (ED) and the Kullback-Leibler Distance (KLD) are the most commonly used distance measures for comparing discrete probability distributions and are what we suggest to do here. The Euclidean Distance between the probability vectors $\mathbf{p_i}$ and $\mathbf{p_l}$ is given by

$$d_{ED,il} = \left[\sum_{j=1}^{K}(p_{ij} - p_{lj})^2\right]^{1/2} \tag{3.7}$$

and the Kullback Leibler Distance between the probability vectors $\mathbf{p_i}$ and $\mathbf{p_l}$ is calculated by

$$d_{KLD,il} = \sum_{j=1}^{K} p_{ij} \times \log\left(\frac{p_{ij}}{p_{lj}}\right). \tag{3.8}$$

The ED measures the shortest distance between the cell probabilities of batter i and l but it wouldn't classify the two batters are either being too close or too far away statistically. On the other hand, the KLD has a statistical meaning as it measures

the ratio between the likelihood and unlikelihood that batter l is similar to batter i. Then, the matrix of dissimilarities, or dissimilarity matrix D, is created by using the dissimilarity measures. The clustering of the batters are based on D, and the order the batters cluster is based on the dissimilarity measures; lowest values cluster first. Then, when considering the MAMSE weights $w_{i1}, \ldots, w_{im}$ for inference on the i^{th} batter, we select a cluster of 9 other batters (from the 556) based on minimizing dissimilarity (3.7) and (3.8). We calculated the re-weighted counts x^*_{ij} for each batter based on the weighted likelihood approach using MAMSE weights. We denoted the clustering based on the Euclidean Distance as EDC and the clustering based on the Kullbac-Leibler Distance as KLDC. There are two approaches to select a cluster of batters based on EDC and KLDC. The first approach is by determining a pre-defined threshold for clustering. If the dissimilarity between batter i and another batter is below or equal to a threshold, then that batter will cluster with batter i and be used for the corresponding inference. For the threshold, we experimented with 0.15 and 0.2. The second approach is to select a fixed number of batters to cluster with player i. Based on argument made by Plante (2008) [13], we need to select less than $K-1$ batters in each cluster to get a unique set of MAMSE weights. There are here $K = 11$ outcomes and, to get a unique solution for MAMSE weights, we select less than 10 batters per cluster. We considered 5, 7 and 9 other batters per cluster. In this chapter, we report the results using the second approach with 9 batters per cluster. It would make sense to use other variables such as age, field position of the batter or any other covariate to measure the dissimilarity between batters. Note that we created three categories, SO, GO and AO, for outs. This distinction isn't important for the considered batting metrics, but it plays a role in the clustering of

players by allowing to identify different patterns in batting outcomes.

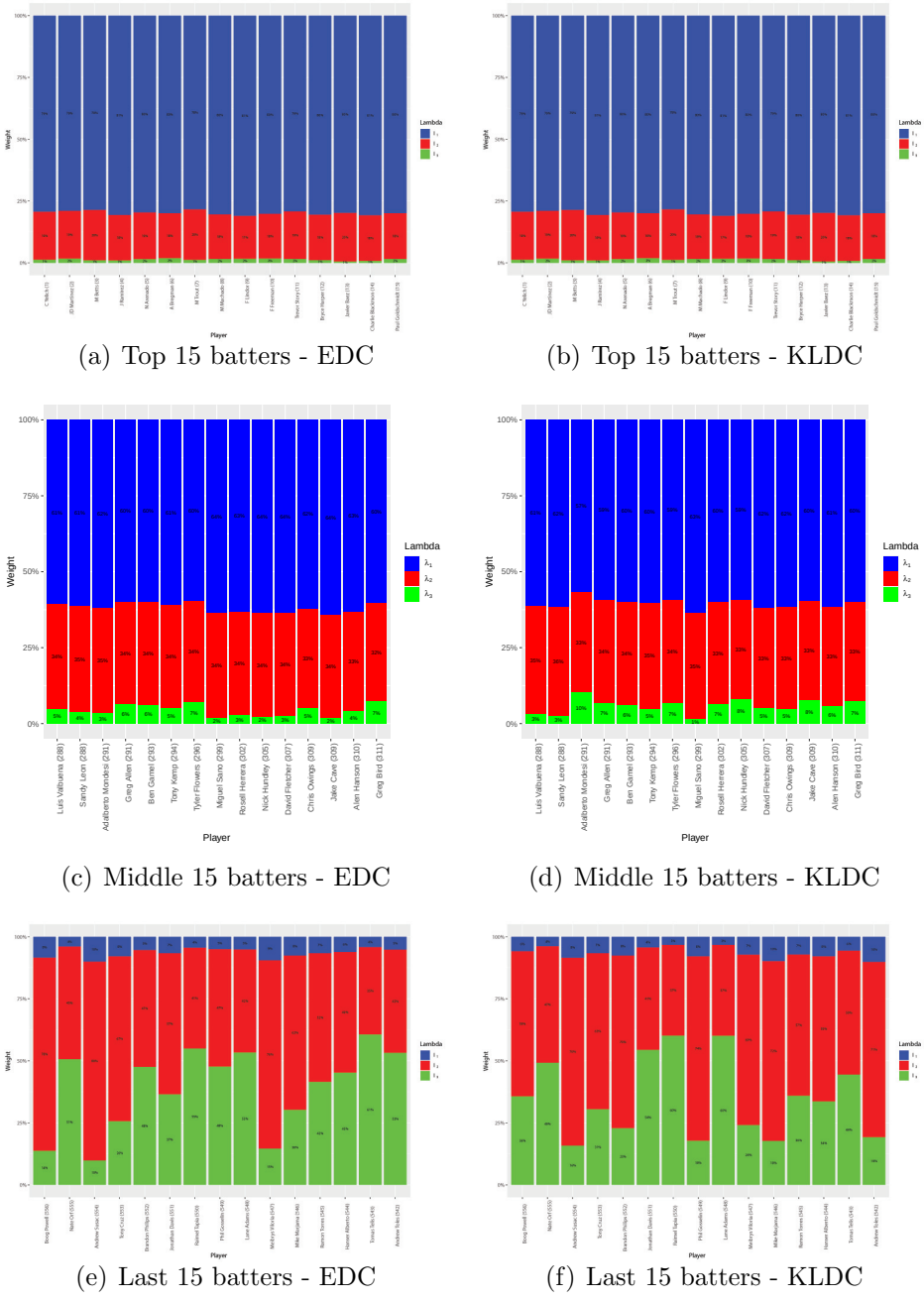
Table 3.7: Estimates of the concentration parameters $\hat{\alpha}_j$.

	SO	GO	AO	SH	SF	HBP	BB	S	D	T	HR
WL estimates (ED)	34.35	35.43	34.79	0.40	1.28	1.78	13.18	22.43	7.32	0.87	4.76
WL estimates (KLD)	34.78	36.51	36.07	0.41	1.29	1.79	13.61	23.29	7.52	0.85	4.76

Table 3.7 provides the parameter estimates of the concentration parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{11})$. The estimates of the concentration parameters based on maximizing the marginal weighted likelihood (as outlined at the end of Section 3.2) depend on which measure of distance is used. Interestingly, the estimates obtained from using EDC and KLDC have the same pattern, although the estimates based on KLDC are higher compared than those based on EDC.

Figure 3.2 provides the weight $\hat{\lambda}_{1i}$, $\hat{\lambda}_{2i}$ and $\hat{\lambda}_{3i}$ based on (3.5) and (3.6) that are used to estimate $\mathbf{p_i}$. For the top 15 batters, it is clear that the weight for the MLE ($\hat{\lambda}_{1i}$) is very high and the weight for the player specific shrinkage target ($\hat{\lambda}_{3i}$) is very small. On the other hand, for the last 15 batters, the weights for the outcome specific target ($\hat{\lambda}_{2i}$) and shrinkage target ($\hat{\lambda}_{3i}$) are higher than the weight of the MLE ($\hat{\lambda}_{1i}$). For the middle 15 batters, compared to the top 15 batters, the weight for the MLE ($\hat{\lambda}_{1i}$) decreased and $\hat{\lambda}_{2i}$ and $\hat{\lambda}_{3i}$ increased. This behaviour was to be expected, given the top 15 batters have a large number of plate appearances, in their case, so the MLE already provides good information and the shrinkage estimator trusts the data more. For the last 15 batters, that only have a small number of plate appearances, the shrinkage estimator trusts the shrinkage target (t_{ij}) and outcome specific target (t_j) more. Notice that the patterns for the weights $\hat{\lambda}_{1i}$, $\hat{\lambda}_{2i}$ and $\hat{\lambda}_{3i}$ are essentially

the same for both the EDC and KLDC approaches.



(a) Top 15 batters - EDC

(b) Top 15 batters - KLDC

(c) Middle 15 batters - EDC

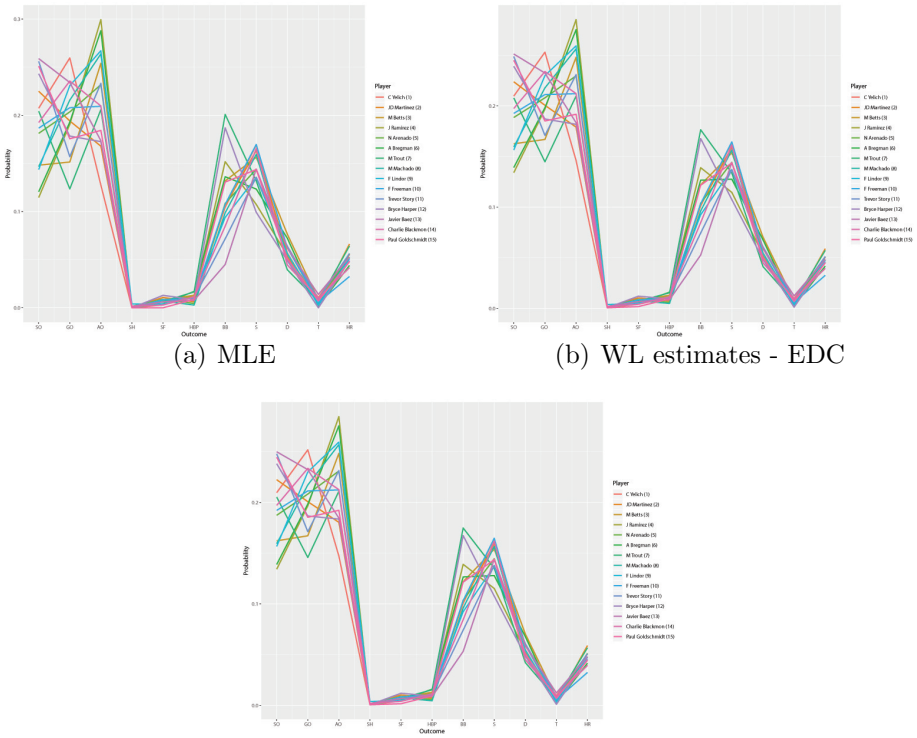
(d) Middle 15 batters - KLDC

(e) Last 15 batters - EDC

(f) Last 15 batters - KLDC

Figure 3.2: Weight comparison for the EDC and the KLDC approaches for the top 15, 15 batters in the middle of the dataset, and 15 batters with fewest number of plate appearances (based on ESPN rankings)

Figure 3.3 shows the estimates of the cell probabilities for the top 15 batters based on the weighted likelihood approach with EDC and KLDC are very close to the MLE. As we mentioned above, the top 15 batters have a large number of plate appearances and, because of that, the shrinkage estimator trusts the MLE (data) more and the WL estimates are close to the MLE. Figure 3.4 shows the WL estimates for the last 15 batters based on the weighted likelihood approach with EDC and KLDC are very close each other. Interestingly, the last 15 batters have only a few cell counts for each outcome and, as a result, the shrinkage estimators trust the targets (t_{ij} and t_j) more and the WL estimates are quite different from the MLE.

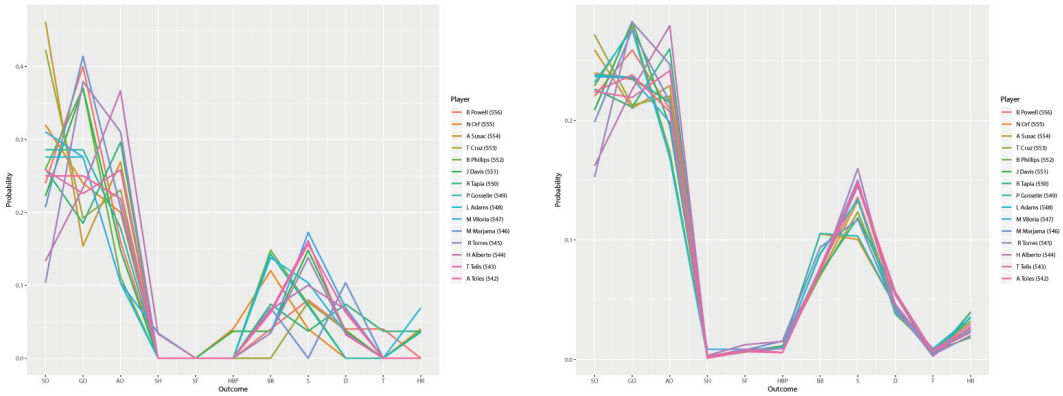


(a) MLE

(b) WL estimates - EDC

(c) WL estimates - KLDC

Figure 3.3: Comparison of WL estimates for both the EDC and the KLDC approaches with MLE (observed proportions from raw data) for the top 15 batters (based on ESPN rankings)



(a) MLE

(b) WL estimates - EDC

(c) WL estimates - KLDC

Figure 3.4: Comparison of WL estimates for both the EDC and the KLDC approaches with MLE (observed proportions from raw data) for the last 15 batters

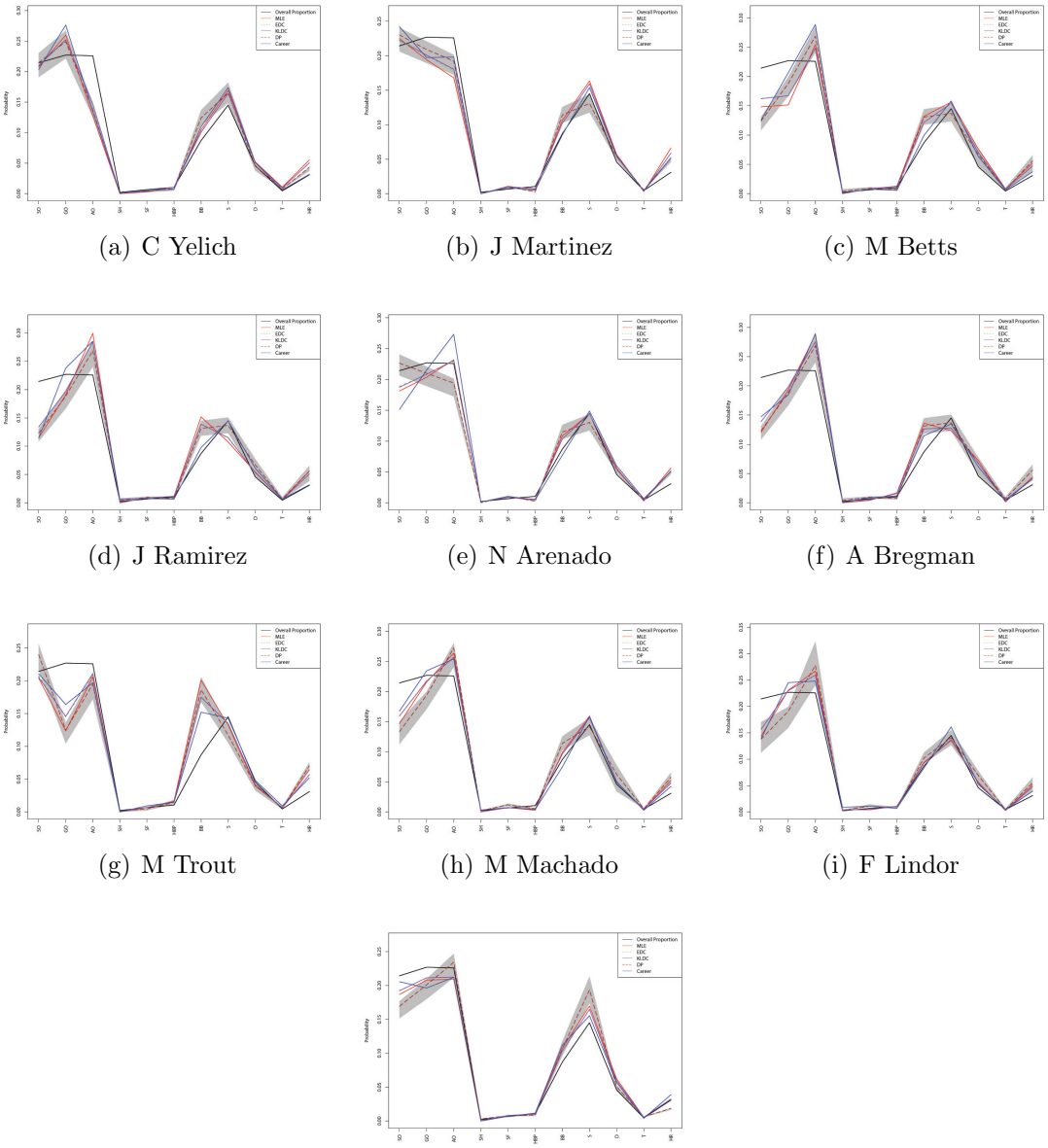
In the semi-parametric Bayesian approach, 60,000 draws were taken from a DP using Gibbs sampling with a burn-in of 50,000 draws under the following choice of hyper-parameters for the hierarchical prior distribution:

$$G \sim \text{DP}(\alpha_0 = 200, G_0),$$

$$G_0 \sim \text{Dirichlet}\left(\alpha_1 = \alpha_2 = \ldots = \alpha_{11} = 1\right).$$

Teh (2010)[16] proposed a formula to find the expected number of clusters based on α_0 and the number of observations. Based on that formula we picked $\alpha_0 = 200$,

which we found to provide a reasonable number of clusters. As an alternative, one can introduce a prior on α_0.



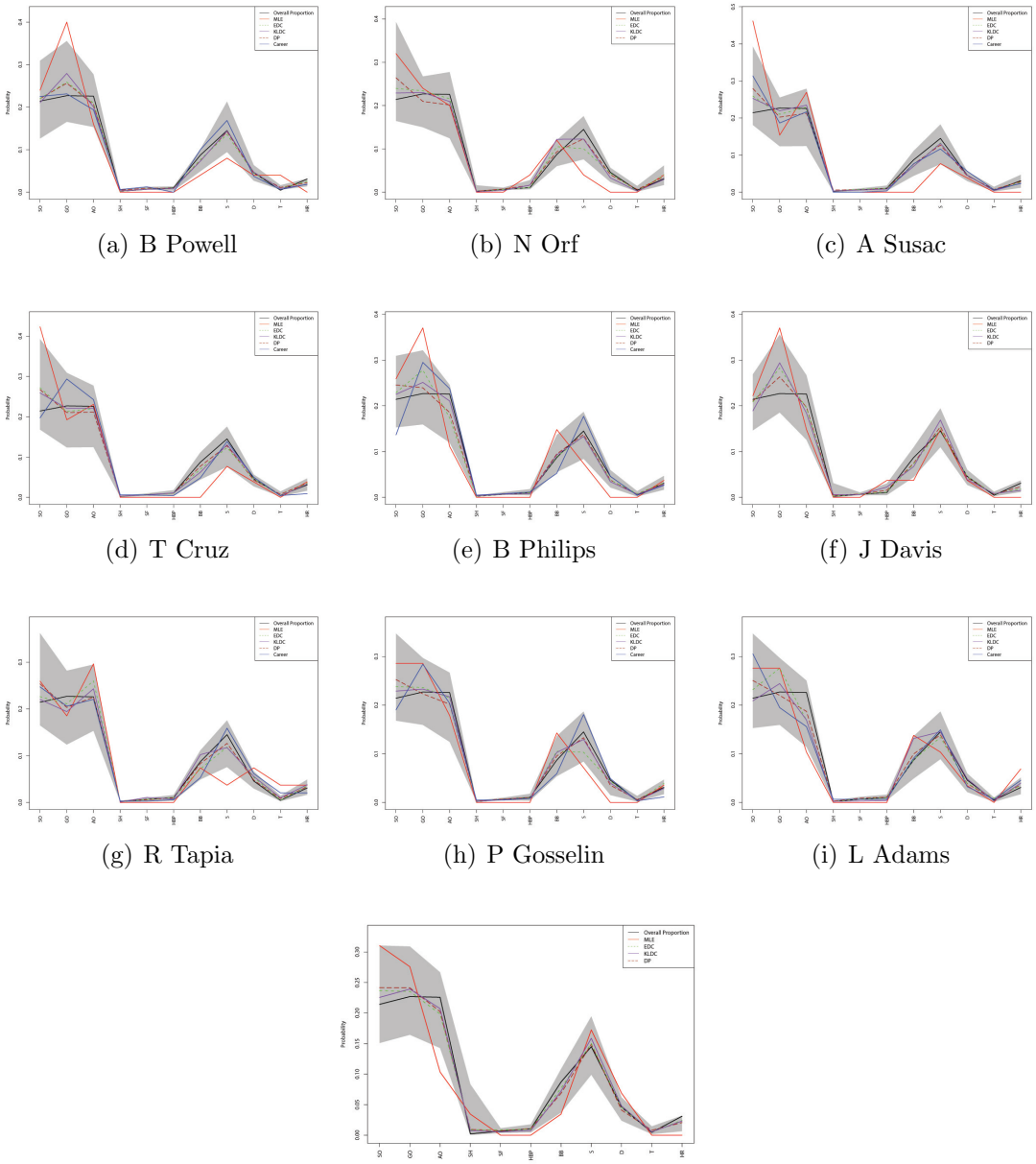
(a) C Yelich (b) J Martinez (c) M Betts

(d) J Ramirez (e) N Arenado (f) A Bregman

(g) M Trout (h) M Machado (i) F Lindor

(j) F Freeman

Figure 3.5: Comparison of MLE, WL estimates for both the EDC and the KLDC approaches, DP estimates with 95% credible interval and overall proportion of outcomes for the top 10 batters



(a) B Powell (b) N Orf (c) A Susac

(d) T Cruz (e) B Philips (f) J Davis

(g) R Tapia (h) P Gosselin (i) L Adams

(j) M Viloria

Figure 3.6: Comparison of MLE, WLE for both the EDC and the KLDC approaches, DP estimates with 95% credible interval and overall proportion of outcomes for the last 10 batters with fewest number of plate appearances.

Figure 3.5 displays the weighted likelihood (WL) estimates with MAMSE weights

using EDC, the WL estimates with MAMSE weights using KLDC, the MLE, the semi-parametric Bayes estimator based on the Dirichlet process (DP) and overall proportions of batting outcomes for the top 10 batters. The grey shaded area is the 95% credible interval based on the output of the posterior simulation and the DP estimates. It is clear that WL, ML and DP estimates are very close to each other in the case of these players, all having a large number of plate appearances. Figure 3.6 displays the same estimates but for the last 10 batters, all having a small number of PA. In this figure, the WL estimates are very close to the overall proportions of batting outcomes. When $\mathbf{p}_i$ is far away from the overall proportions of batting outcomes ($\bar{\mathbf{p}}$), the DP estimates tend to be closer to the MLE. In cases where cell counts are zero, the MLE is zero, but the other estimates generally have a small positive value, which seems like a desirable behavior given the nature of the sport. We also include the career proportions: that is the proportions of batting outcomes for all the seasons the player has played. It is quite interesting to note that, in many cases, the MLE based on only a few outcomes, is quite far from career proportions. The estimates based on all the methods presented here correct this a nice way in most cases. Note that, this was the first season N. Orf, J. Davis and M. Viloria played in MLB, making it clear that career based statistics don't have the same references in their case.

Table 3.8 and 3.9 provide the composition of the clusters used for WL inference based on EDC and KLDC. In Table 3.10, we present the batters that cluster with each other most often throughout the iterations of MCMC used for inference based on the DP. In a single iteration of MCMC, the batters are clustered according to whether their current vector of outcome probabilities $\mathbf{p}$ are the same atom of the DP

for that iteration. With the MCMC output, we are able to calculate the proportion of iterations that any given pair of batters cluster together and this provides an estimate of the posterior pairwise probability that two players share the same vector $\mathbf{p_i}$, suggesting they have the same batting ability. This posterior probability is provided in Table 3.10 after the batter name. The color 'Red' shows the batters who clustered in both EDC and KLDC, the color 'Blue' the batters who clustered in both EDC and DP, the color 'Teal' the batters who clustered in both KLDC and DP, the color 'Black' the batters who clustered only by one approach and '*' the batters who clustered with all the approaches.

Table 3.8: Clustering of batters based on EDC: the 9 batters considered to be most similar to each of the top 10 players used for constructing the MAMSE weights

Batter	1	2	3	4	5	6	7	8	9
Yelich	Martini	Freese	Munoz	Holt	Ramos	Soto	Mazara	Peralta	Choo
Martinez	Ohtani*	Schebler	Acura	Aguilar	Pham	Suarez	Conforto	Haniger	Goldschmidt
Betts	Stewart	Turner	Grossman	Rendon	Arenado	Lowrie	Bregman*	Merrifield	Machado
Ramirez	Wieters	Turner	Gregorius	Shaw	Rendon	Kepler	Rizzo	Santana*	Bregman*
Arenado	White	Pearce	Gyorko	Puig	Hernandez	Camargo	Realmuto	Bogaerts	Freeman
Bregman	Turner	Rendon	Kepler	Betts*	Santana	Rizzo	Ramirez*	Machado	Lindor
Trout	Iannetta	Bautista	Cervelli	Descalso	Muncy*	Hoskins	Carpenter	Goldschmidt	Harper
Machado	Seager	Hernandez*	Reddick	Span	Gregorius*	Rizzo	Arenado	Semien	Lindor
Lindor	Seager	Cozart	Hernandez	Reddick	Span	Gregorius*	Profar	Rizzo	Machado
Freeman	Casali	Narvaez	Easton	Gyorko*	Camargo	Springer	Arenado	Haniger	Blackmon

Table 3.9: Clustering of batters based on KLDC: the 9 batters considered to be most similar to each of the top 10 players used for constructing the MAMSE weights

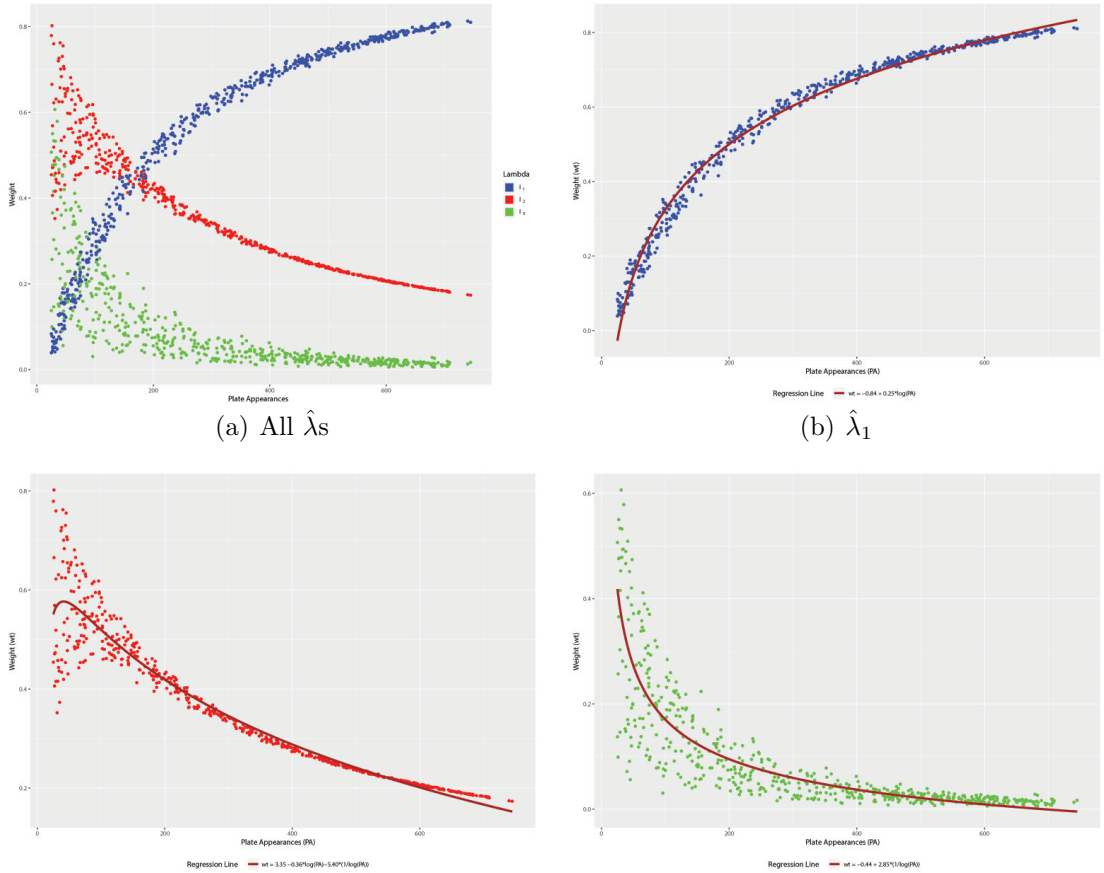
Batter	1	2	3	4	5	6	7	8	9
Yelich	Freese	Culberson	Mazara	Pham	Peralta	Desmond	Mancini	Martinez	Blackmon
Martinez	Choi	Ohtani*	Acuna	Camargo	Suarez	Conforto	Arenado	Haniger	Stanton
Betts	Pearce	Turner	Bogaerts	Rendon	Arenado	Lowrie	Ramirez	Bregman*	Freeman
Ramirez	Turner	Gregorius	Rendon	Kepler	Betts	Santana*	Bregman*	Machado	Lindor
Arenado	Pederson	Puig	Hernandez	Hicks	Kipnis	Benintendi	Blackmon	Machado	Lindor
Bregman	Pearce	Turner	Rendon	Kipnis	Kepler	Betts*	Rizzo	Ramirez*	Lindor
Trout	Iannetta	Granderson	Descalso	Muncy*	Grandal	Hicks	Suarez	Betts	McCutchen
Machado	Pearce	Gyorko	Puig	Hernandez*	Gregorius*	Rendon	Moustakas	Arenado	Lindor
Lindor	Frazier	Pederson	Hernandez	Ahmed	Gregorius*	Krpnis	Arenado	Albied	Blackmon
Freeman	Pearce	Gyorko*	Camargo	Bogaerts	Crawford	Arenado	McCutchen	Haniger	Blackmon

Table 3.10: Clustering of batters based on DP: the 9 batters considered to be most similar to each of the top 10 players most often throughout the iterations of the MCMC algorithm

Batter	1	2	3	4	5	6	7	8	9
Yelich	Soto (0.14)	Pham (0.03)	Martini (0.02)	Flaherty (0.02)	Reyes (0.02)	Munoz (0.02)	Nimmo (0.02)	Contreras (0.02)	Hosmer (0.02)
Martinez	Arenado (0.04)	Smoak (0.04)	Donaldson (0.04)	Ohtani* (0.03)	White (0.03)	Myers (0.03)	Grandal (0.03)	Aguilar (0.03)	O'Brien (0.03)
Betts	Bregman* (0.36)	Ramirez (0.40)	Santana (0.09)	Alberto (0.01)	Stewart (0.01)	Duvall (0.01)	Belt (0.01)	Moustakas (0.01)	Olson (0.01)
Ramirez	Bregman* (0.44)	Betts (0.40)	Santana* (0.06)	Powell (0.01)	Orf (0.01)	Alberto (0.01)	Stewart (0.01)	Chisenhall (0.01)	Seager (0.01)
Arenado	Grandal (0.05)	Martinez (0.04)	Anguilar (0.04)	Donaldson (0.03)	Reynolds (0.03)	Myers (0.03)	Ohtani (0.03)	Moreland (0.03)	Smoak (0.03)
Bregman	Ramirez* (0.44)	Betts* (0.36)	Santana (0.08)	Torres (0.01)	Alberto (0.01)	Telis (0.01)	Robinson (0.01)	Robles (0.01)	Chisenhall (0.01)
Trout	Muncy* (0.70)	Holiday (0.11)	Bautista (0.05)	Orf (0.01)	Gosselin (0.01)	Wisdom (0.01)	Stewart (0.01)	Voit (0.01)	O'Hearn (0.01)
Machado	Hernandez* (0.29)	Redrick (0.29)	Gregorius* (0.18)	Santana (0.18)	Brito (0.01)	Stewart (0.01)	Luplow (0.01)	Wieters (0.01)	Murphy (0.01)
Lindor	Moustakas (0.18)	Gregorius* (0.16)	Escobar (0.11)	Rendon (0.11)	Kinsler (0.09)	Wieters (0.09)	Lowrie (0.04)	Flores (0.03)	Higashioka (0.01)
Freeman	Gyorko* (0.07)	Benintendi (0.07)	Votto (0.07)	Merrifield (0.06)	Polanco (0.05)	Winker (0.05)	Grossman (0.05)	Cain (0.06)	Semien (0.05)

Figure 3.7 and Figure 3.8 provide the weight comparisons for all the batters. We can see clearly that there is an upward trend with plate appearances for $\hat{\lambda}_{1i}$ and downward trends for $\hat{\lambda}_{2i}$ and $\hat{\lambda}_{3i}$. Given what was discussed previously, this behavior was to be expected; when you increase the number of plate appearances, the MLE provides more reliable information and the shrinkage estimator trusts the data more. The MLE weight ($\hat{\lambda}_{1i}$) is increasing at a logarithmic rate in terms of plate appearance (PA). As a result, when you decrease the number of plate appearances, the shrinkage estimator trusts the shrinkage target (t_{ij}) and outcome specific target (t_j) more. The player specific target weight ($\hat{\lambda}_{3i}$) is decreasing at a inverse logarithmic rate in terms of PA. For large but achievable plate appearances, $\hat{\lambda}_{1i} \to 0.8$ and $\hat{\lambda}_{3i} \to 0.2$. It is interesting to note that, in principle, one can get the approximate values for $\hat{\lambda}_{1i}$, $\hat{\lambda}_{2i}$ and $\hat{\lambda}_{3i}$ for all the batters based on their plate appearances using the fitted curves given in Figure 3.7 and 3.8, thus saving a considerable amount of computational

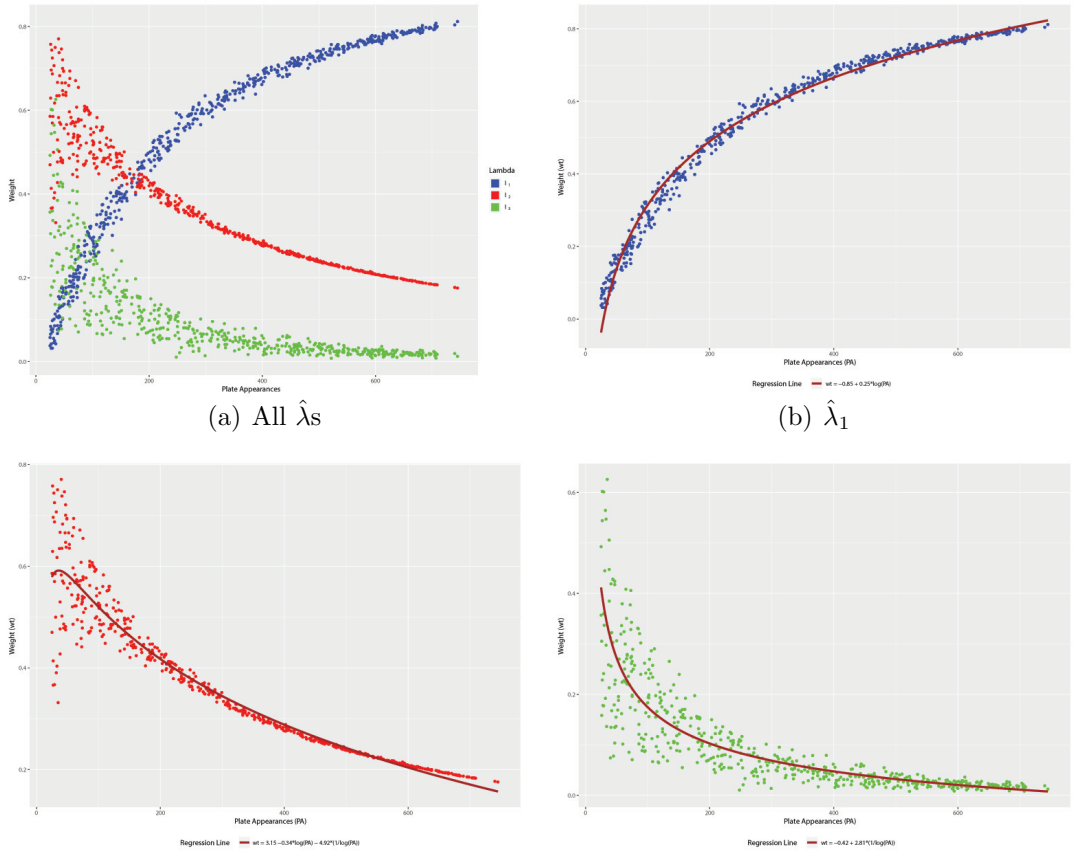
time. Here R^2 measures how close the data are to the fitted regression line. In this case, however, one still needs to obtain the player specific targets which is a numerically involved task to complete.



(a) All $\hat{\lambda}$s (b) $\hat{\lambda}_1$

(c) $\hat{\lambda}_2$ (d) $\hat{\lambda}_3$

Figure 3.7: Weight comparison of $\hat{\lambda}_1, \hat{\lambda}_2$ and $\hat{\lambda}_3$ for the EDC for all the batters. R^2 for fitted curves are $(b)R^2 = 0.987$, $(c)R^2 = 0.901$, $(d)R^2 = 0.767$.



(a) All $\hat{\lambda}$s (b) $\hat{\lambda}_1$

(c) $\hat{\lambda}_2$ (d) $\hat{\lambda}_3$

Figure 3.8: Weight comparison of $\hat{\lambda}_1, \hat{\lambda}_2$ and $\hat{\lambda}_3$ for the KLDC for all the batters. R^2 for fitted curves are $(b)R^2 = 0.988$, $(c)R^2 = 0.910$, $(d)R^2 = 0.781$.

3.6 Discussion

In this chapter, we introduced a new approach based on weighted likelihood for modeling batting outcomes in baseball and used it to estimate commonly used metrics used for assessing batters. Our weighted likelihood methodology borrows information from other batters to improve the estimation of batter specific parameters, whereas the widely used classical likelihood method utilizes only the information from the current batter. The MLE could provide quite misleading results due to a insufficient

sample size, i.e., for players having a very small number of plate appearances. The main advantage of the weighted likelihood approach is that it may be possible to obtain better estimation by sharing information among batters in a straight forward manner and in a nice intuitive way. For comparison, we also presented a second approach based on the Dirichlet process, which naturally borrows information across similar batters and clusters them together. The cluster assignment of players is an interesting characteristic of this approach which is significantly more computationally intensive.

Some computational challenges arose when calculating $\hat{\boldsymbol{\alpha}}$. When using the original counts, the R package **dirmult**[18] and the fixed-point iteration of Minka (2000)[11] given by

$$\hat{\alpha}_j^{\text{new}} = \hat{\alpha}_j^{\text{old}} \frac{\sum_{i=1}^{m} \psi(n_{ij} + \alpha_j) - \psi(\alpha_j)}{\sum_{i=1}^{m} \psi\left(n_i + \sum_{j=1}^{k} \alpha_j\right) - \psi\left(\sum_{j=1}^{k} \alpha_j\right)},$$

provide the same estimates for $\boldsymbol{\alpha}$ where $\psi(x) = \dfrac{d \log\Gamma(x)}{dx}$ is the digamma function. However, when considering the MAMSE weights based on all players and using these to compute the new counts (x_{ij}^*), then the estimates obtained from these two approaches disagree and the MAMSE weights run into identifiability problems. Possible solutions these problems are to use:

1. a player specific subset of players when obtaining MAMSE weights for each batter (which is what we did above);

2. a pre-defined value for $A = \sum_{j=1}^{K} \alpha_j$.

Yu and Shaw (2014)[25] proposed different methods to estimate $\boldsymbol{\alpha}$ using a pre-defined value A. We found both approaches to work well in terms of computations with player specific clusters of 9 batters. For the second approach, the main issue is then to select A properly in a data-driven way.

Table 3.11: Comparison of raw season metrics, estimated metrics using EDC and KLDC and career metric for last 10 batters with fewer number of plate appearances and top 5 batters

Player	BA				OBP				SLG			
	Raw	EDC	KLDC	Career	Raw	EDC	KLDC	Career	Raw	EDC	KLDC	Career
Boog Powell	0.167	0.227	0.236	0.262	0.200	0.279	0.296	0.333	0.292	0.311	0.362	0.383
Nate Orf	0.095	0.179	0.184	0.095	0.240	0.291	0.330	0.240	0.238	0.377	0.313	0.238
Andrew Susac	0.115	0.140	0.143	0.221	0.115	0.173	0.160	0.283	0.154	0.240	0.184	0.373
Tony Cruz	0.154	0.182	0.211	0.216	0.154	0.206	0.237	0.256	0.308	0.364	0.400	0.308
Brandon Phillips	0.130	0.248	0.232	0.275	0.259	0.319	0.306	0.320	0.261	0.396	0.375	0.420
Jonathan Davis	0.200	0.240	0.257	0.200	0.259	0.298	0.320	0.259	0.240	0.349	0.327	0.240
Raimel Tapia	0.200	0.217	0.251	0.280	0.259	0.303	0.336	0.322	0.480	0.408	0.498	0.456
Phil Gosselin	0.125	0.182	0.209	0.265	0.250	0.289	0.330	0.314	0.250	0.380	0.367	0.363
Lane Adams	0.240	0.261	0.291	0.263	0.345	0.374	0.403	0.333	0.520	0.437	0.517	0.467
Meibrys Viloria	0.259	0.260	0.276	0.259	0.286	0.322	0.304	0.286	0.333	0.376	0.395	0.333
Christian Yelich	0.326	0.326	0.326	0.299	0.402	0.403	0.402	0.380	0.598	0.599	0.598	0.480
J.D. Martinez	0.330	0.330	0.330	0.293	0.402	0.402	0.402	0.354	0.629	0.628	0.629	0.533
Mookie Betts	0.346	0.346	0.346	0.300	0.438	0.438	0.438	0.371	0.640	0.642	0.642	0.513
Jose Ramirez	0.270	0.271	0.272	0.277	0.387	0.387	0.388	0.351	0.552	0.550	0.555	0.469
Nolan Arenado	0.297	0.297	0.296	0.294	0.374	0.374	0.373	0.349	0.561	0.562	0.560	0.544

We note again, by looking at Table 3.11, that for the last 10 batters, except for N. Orf, J. Davis and M. Viloria as discussed above, (this was their first season), the estimated metrics are close to career metrics compared to raw season metrics. On the other hand, for Brandon Phillips has about 8000 career plate appearances, we can see clearly the raw season metrics are far away from career metrics but the estimated metrics (EDC and KLDC) improved the estimation by borrowing information from other players and are very close to career metrics which seems desirable. As we expected, the top 10 batters' raw season metrics are very close to estimated metrics. As these batters have a large number of plate appearances, the raw season metrics

provide good estimates and small weight is given to the borrowed information from other players. Finally, throughout our analyses, we used non-informative priors as baseline distributions. One could also use informative priors in another attempt to improve the estimation. Such an informative prior could be based, for instance, on results obtained in previous seasons or player specific career statistics, although many options could be considered.

Bibliography

[1] J. Albert. Baseball data at season, play-by-play, and pitch-by-pitch levels. *Journal of Statistics Education*, 18(3):1–27, 2010.

[2] J. Albert. Improved component predictions of batting measures. *Journal of Quantitative Analysis in Sports*, 12:73–85, 2016.

[3] S.R. Bailey, J. Loeppky, and T.B. Swartz. The prediction of batting averages in major league baseball. *Stats*, 3:84–93, 2020.

[4] J. Bennett and J. Flueck. An evaluation of major league baseball offensive performance models. *The American Statistician*, 37(1):76–82, 1983.

[5] D. Blackwell and J. B. MacQueen. Ferguson distributions via Polya urn schemes. *The Annals of Statistics*, 1(353-355), 1973.

[6] B. Efron and C. Morris. Data analysis using Stein's estimatior and its generalizations. *Journal of the American Statistical Association*, pages 311–319, 1975.

[7] B. Efron and C. Morris. Stein's paradox in statistics. *Scientific American*, 256(5):119–127, 1977.

[8] T. S. Ferguson. A Bayesian analysis of some nonparametric problem. *The Annals of Statistics*, pages 209–230, 1973.

[9] F. Hu and J. Zidek. The weighted likelihood. *The Canadian Journal of Statistics*, 30(3):347–371, 2002.

[10] F. Hu and J. Zidek. Forecasting NBA basketball playoff outcomes using the weighted likelihood. *Lecture Notes-Monograph Series*, 45:385–395, 2004.

[11] T. Minka. Estimating a Dirichlet distribution. *Technical Report - M.I.T.*, 2000.

[12] R. M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

[13] J. Plante. Nonparametric adaptive likelihood weights. *The Canadian Journal of Statistics*, 36:443–461, 2008.

[14] J. Plante. Asymptotic properties of the MAMSE adaptive likelihood weights. *Journal of Statistical Planning and Inference*, 139(7):2147–2161, 2009.

[15] J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

[16] Y. W. Teh. Dirichlet processes. In *Encyclopedia of Machine Learning*. Springer, 2010.

[17] Y. W. Teh, M. I. Jordan, M. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101:1566–1581, 2006.

[18] T. Tvedebrink. Dirmult: estimation in Dirichlet-multinomial distribution, R Package Version 0.1.3. 2009.

[19] C. Wang and M. Vandebroek. A model based ranking system for soccer teams. *Technical Report - KU Leuven*, 2013.

[20] X. Wang. *Maximum weighted likelihood estimation*. PhD thesis, University of British Columbia, 2001.

[21] X. Wang, C. van Eeden, and J. Zidek. Technical report on weighted likelihood estimation and asymptotic properties of the weighted likelihood estimators. *Technical report no 201*, Online, 2002.

[22] X. Wang and J. Zidek. Selecting likelihood weights by cross-validation. *The Annals of Statistics*, 33:463–501, 2005.

[23] L. Wickramasinghe, A. Leblanc, and S. Muthukumarana. Semi-parametric Bayesian estimation of sparse multinomial probabilities with an application to the modelling of bowling performance in T20I cricket. *Under Review*.

[24] T. Y. Yang and T. Swartz. A two-stage Bayesian model for predicting winners in major league baseball. *Journal of Data Science*, pages 61–73, 2004.

[25] P. Yu and C.A. Shaw. An efficient algorithm for accurate computation of the Dirichlet-multinomial log-likelihood function. *Bioinformatics*, 30:1547–1554, 2014.

Chapter 4

Bayesian Inference on Sparse Multinomial Data Using a Smoothed Dirichlet Prior

In this chapter, we develop a Bayesian approach for estimating multinomial cell probabilities using a smoothed Dirichlet prior. The most important feature of the smoothed Dirichlet prior is that it forces the probabilities of neighboring cells to be closer to each other than under the standard Dirichlet prior. Using this Bayesian approach, we propose a shrinkage type of estimators to estimate multinomial cell probabilities under sparsity. We demonstrate the proposed approach using early cases of COVID-19 data and estimate the distribution across age groups for Canadian health regions where the age groups are arranged in order. This proposed estimator allows us to borrow information across other health regions and age groups to improve the estimation of cell probabilities. Our approach especially improves the estimation in smaller health regions where not many cases have been observed.

4.1 Introduction

The sparseness in multinomial data is frequently encountered in practice when many cells have small and/or zero counts. The sparse multinomial data can arise in many ways and we focus on the case of sparse multinomial data that are due to the observations dispersed in numerous categories. In this chapter, the main objective is to develop an improved strategy to jointly estimate the cell probabilities of multinomial populations in a context of sparse data.

Suppose there are m multinomial populations with K discrete categories. We assume the existence of an underlying vector $\mathbf{p_i}$ of multinomial cell probabilities that describes how the total probability is shared between categories for the i^{th} multinomial population. Specifically, the joint distribution of the counts for the K discrete categories of the i^{th} multinomial population is given by

$$\mathbf{X_i} = (X_{i1}, X_{i2}, \ldots, X_{iK})^t \sim \text{Multinomial}(n_i, \mathbf{p_i}),$$

where $\mathbf{p_i} = (p_{i1}, p_{i2}, \ldots, p_{iK})^t$ represents the vector of outcome specific probabilities satisfying $\sum_j p_{ij} = 1$. Note that the cell counts in some categories will be either small or zero, which will result a sparse dataset. The Maximum Likelihood Estimator (MLE) is known to perform very poorly in these setups (Hausser and Strimmer (2009)[3]). Shrinkage estimation is an approach that allows the derivation of improved estimators by borrowing information across other multinomial populations and cell categories to improve the estimation of cell probabilities. We proposed a Bayesian shrinkage type of estimator using smoothed Dirichlet distribution to estimate multinomial cell probabilities under sparsity.

In Chapter 2, we developed a semi-parametric Bayesian estimator based on the Dirichlet process (DP) using the stick-breaking construction, which was proposed by Sethuraman (1994)[7]. This Dirichlet process approach naturally borrows information across a similar population and clusters them together. Also, in Chapter 3 we proposed an approach to model the batting outcomes of baseball batters based on the weighted likelihood approach. The weighted likelihood allows the sharing of relevant information among batters. Both of these two methods borrow information only across other multinomial populations to improve the estimation of cell probabilities. Crucially, both methods do not allow the sharing of information between cells.

The chapter will proceed as follows. In Section 4.2, we discuss the previously studied statistical models, including empirical Bayes (EB) estimation. In Section 4.3, we introduce the smoothed Dirichlet distribution and its properties. The proposed statistical models: the Bayes estimation and estimation based on Bayesian Multinomial regression are discussed in Section 4.4. In Section 4.5, we apply the methods proposed in Section 4.4 on the COVID-19 dataset. We conclude with a short discussion in Section 4.6 based on the results and methods we discuss in this chapter.

4.2 Existing Statistical Model and Estimation

The concept of *shrinkage* was first introduced by Stein (1956)[8] and the general principles behind shrinkage estimation were discussed by Ledoit and Wolf (2003)[6]. The rationale behind shrinkage is that a baseline estimate is improved by combining it with other information. Optimal shrinkage estimation usually reduces variance

at the cost of bias. Shrinkage estimation is an approach that allows the derivation of improved estimators, and in particular, it can handle certain forms of sparsity. The James-Stein (JS) shrinkage estimator (U^*), was introduced by James and Stein (1961)[5] in a context of estimating normal means and is based on a weighted average of two different estimators; the first, U, coming from a high-dimensional model with low bias and high variance, and the second, T, coming from a lower-dimensional model with larger bias but smaller variance,

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

The typical approach to derive JS shrinkage estimators is to use the Bayesian paradigm that supplies extra information through a prior distribution and that may lead to vast improvements over the MLE (Agresti and Hitchcock (2005)[1]). Fienberg and Holland (1973)[2] showed that empirical Bayes estimators are often superior to the MLE in sparse multinomial setups under squared-error loss function.

4.2.1 Empirical Bayes Estimator

The distribution of cell counts ($\mathbf{x_i}$) for i^{th} multinomial population is given by

$$\mathbf{x_i} = (x_{i1}, x_{i2}, \ldots, x_{iK})^t \quad \sim \quad \text{Multinomial}(n_i, \text{prob} = \mathbf{p_i}),$$

where $\mathbf{p_i} = (p_{i1}, p_{i2}, \ldots, p_{iK})^t$ denotes the vector of cell probabilities for i^{th} multinomial population. Taking a Bayesian approach, we assume that

$$\mathbf{p_i} = (p_{i1}, p_{i2}, \ldots, p_{iK})^t \quad \sim \quad \text{Dirichlet}(\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_K)^t).$$

The Dirichlet distribution is a conjugate prior distribution for the vector of multinomial cell probabilities ($\mathbf{p_i}$). Then, the posterior distribution of $\mathbf{p_i}$, given the observed

counts for i^{th} multinomial population is

$$\mathbf{p_i}|\mathbf{x_i};\boldsymbol{\alpha} \sim \text{Dirichlet}(x_{i1}+\alpha_1, x_{i2}+\alpha_2, \ldots, x_{iK}+\alpha_K)$$

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

$$\hat{p}_{ij}^{\text{Bayes}} = \lambda_i t_j + (1-\lambda_i)\hat{p}_{ij}^{\text{MLE}}, \tag{4.1}$$

which is a shrinkage estimator with $t_j = \left(\dfrac{\alpha_j}{\sum_{j=1}^{K} \alpha_j} \right)$ and $\lambda_i = \dfrac{\sum_{j=1}^{K} \alpha_j}{\left(n_i + \sum_{j=1}^{K} \alpha_j \right)}$. One main advantage of the empirical Bayes estimator is its ability to account for prior information to somehow keep your estimates under control (suitable range), while obtaining the prior empirically from the data itself. In this case, $\boldsymbol{\alpha}$ is obtained by ML estimation based on the Dirichlet-multinomial distribution. The resulting Bayes estimator can be written as a shrinkage estimator where the shrinkage target (t_j) is the Dirichlet prior mean and the shrinkage constant (λ_i) is a combination of sample size and prior mean. A very common approach is then to use the empirical Bayes paradigm and estimate λ_i and t_j using their sample counterparts:

$$\hat{t}_j = \left(\frac{\hat{\alpha}_j}{\sum_{j=1}^{K} \hat{\alpha}_j} \right) \qquad \text{and} \qquad \hat{\lambda}_i = \frac{\sum_{j=1}^{K} \hat{\alpha}_j}{\left(n_i + \sum_{j=1}^{K} \hat{\alpha}_j \right)}. \tag{4.2}$$

The parameter estimates $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_K)^t$ are obtained by using the **dirmult** package in R (Tvedebrink (2010)[10]). These are based on maximum likelihood estimation and can be interpreted as using data-driven parameters in the prior

distribution.

In many situations, there is some knowledge of the smoothness of the underlying vector $\mathbf{p_i}$ of multinomial cell probabilities that describe how the total probability is shared between categories for the i^{th} multinomial population. It is clear that the ordering of the p_{ij}'s is immaterial under a standard Dirichlet prior. The most important feature of the smoothed Dirichlet prior is that it forces the probabilities of neighboring cells to be closer to each other than under the standard Dirichlet prior. Details about the smoothed Dirichlet distribution and its properties are provided in Section 4.3.

4.3 Smoothed Dirichlet Distribution

The smoothed Dirichlet distribution was suggested by Hjort (1996)[4] as a variation to the Dirichlet distribution that forces the successive cell probabilities to be closer to each other. Let $\mathbf{p}_i = (p_{i1}, p_{i2}, \ldots, p_{iK})^t$ be a vector with K components where $p_{ij} \geq 0$ for $j = 1, 2, \ldots, K$ and $\sum_{j=1}^{K} p_{ij} = 1$. Also, let $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_K)^t$, where $\alpha_j > 0$ for each j and $\delta > 0$. The smoothed Dirichlet (SD) probability density function is

$$f(\mathbf{p}_i|\boldsymbol{\alpha}, \delta) = C(\boldsymbol{\alpha}, \delta) \prod_{j=1}^{K} p_{ij}^{\alpha_j - 1} \exp(-\delta \Delta(\mathbf{p}_i)),$$

where $\exp(-\delta\Delta(\mathbf{p}_i))$ is a penalty function which forces successive p_{ij}'s to be close to each other with higher probability than under the standard Dirichlet distribution

satisfying $\sum_{j=1}^{K} p_{ij} = 1$. The constant $C(\boldsymbol{\alpha}, \delta)$ can be written as

$$C(\boldsymbol{\alpha}, \delta) = \frac{C_2(\boldsymbol{\alpha})}{E_{\boldsymbol{\alpha}}[\exp(-\delta\Delta(\mathbf{p}_i))]} \text{ with } \mathbf{p}_i \sim \text{Dirichlet}(\boldsymbol{\alpha}) \text{ and } C_2(\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{j=1}^{K}\alpha_j)}{\prod_{j=1}^{K}\Gamma(\alpha_j)}. \quad (4.3)$$

Let $\mathbf{P} = (P_1, P_2, \ldots, P_K)^t \sim \text{SD}(\boldsymbol{\alpha}, \delta, \Delta)$. The mean of the smoothed Dirichlet distribution is

$$E(P_i) = \frac{\alpha_i}{\sum_{j=1}^{K}\alpha_j} \times \frac{E_{\boldsymbol{\alpha}+\boldsymbol{\gamma_i}}[\exp(-\delta\Delta(\mathbf{P}))]}{E_{\boldsymbol{\alpha}}[\exp(-\delta\Delta(\mathbf{P}))]}, \quad (4.4)$$

where $\gamma_{il} = 0, \forall, l \neq i$ and $\gamma_{ii} = 1$. More generally, the moments of the smoothed Dirichlet distribution random variables can be expressed as

$$E(P_i^n) = \frac{\Gamma(\sum_{j=1}^{K}\alpha_j)}{\Gamma(\sum_{j=1}^{K}\alpha_j + n)} \times \frac{\Gamma(\alpha_i + n)}{\Gamma(\alpha_i)} \times \frac{E_{\boldsymbol{\alpha}+n\boldsymbol{\gamma_i}}[\exp(-\delta\Delta(\mathbf{P}))]}{E_{\boldsymbol{\alpha}}[\exp(-\delta\Delta(\mathbf{P}))]}, \quad (4.5)$$

where γ_{il} is defined as above. Then the variance of the smoothed Dirichlet distribution is

$$\begin{aligned}\text{Var}(P_i) = &\frac{\alpha_i}{\left(\sum_{j=1}^{K}\alpha_j\right)^2\left(\sum_{j=1}^{K}\alpha_j + 1\right)E^2_{\boldsymbol{\alpha}}[\exp(-\delta\Delta(\mathbf{P}))]} \\ &\times \left((\alpha_i+1)\left(\sum_{j=1}^{K}\alpha_j\right)E_{\boldsymbol{\alpha}+2\boldsymbol{\gamma_i}}[\exp(-\delta\Delta(\mathbf{P}))]E^2_{\boldsymbol{\alpha}}[\exp(-\delta\Delta(\mathbf{P}))] - \alpha_i\left(\sum_{j=1}^{K}\alpha_j + 1\right)E^2_{\boldsymbol{\alpha}+\boldsymbol{\gamma_i}}[\exp(-\delta\Delta(\mathbf{P}))]\right).\end{aligned}$$

where γ_{il} is defined as above. The derivations are provided in the Appendix. Note that when $\delta \to 0$,

$$E(P_i) = \frac{\alpha_i}{\sum_{j=1}^{K}\alpha_j} \quad \text{and} \quad \text{Var}(P_i) = \frac{\alpha_i\left(\sum_{j=1}^{K}\alpha_j - \alpha_i\right)}{\left(\sum_{j=1}^{K}\alpha_j\right)^2\left(\sum_{j=1}^{K}\alpha_j + 1\right)}.$$

The roles of δ and Δ are very important when constructing the smoothed Dirichlet distribution. Some examples of Δ functions that we consider are

$$\sum_{j=1}^{K-1}(p_{j+1}-p_j)^2, \quad \sum_{j=1}^{K-1}(p_{j+1}-2p_j+p_{j-1})^2, \quad \text{and} \quad \sum_{j=1}^{K-1}(\log p_{j+1}-\log p_j)^2.$$

The penalty parameter (δ) dictates the extent to which cell probabilities of neighboring categories have to be similar. We use the smoothed Dirichlet distribution as a prior distribution, which forces the successive p_{ij}'s to be closer to each other with higher probability than under standard Dirichlet distribution. To find $E(P_i)$ and $\text{Var}(P_i)$, first we need to calculate $\dfrac{E_{\boldsymbol{\alpha}+\boldsymbol{\gamma_i}}[\exp(-\delta\Delta(\mathbf{P}))]}{E_{\boldsymbol{\alpha}}[\exp(-\delta\Delta(\mathbf{P}))]}$ and $\dfrac{E_{\boldsymbol{\alpha}+2\boldsymbol{\gamma_i}}[\exp(-\delta\Delta(\mathbf{P}))]}{E_{\boldsymbol{\alpha}}[\exp(-\delta\Delta(\mathbf{P}))]}$. It is hard to calculate the above two ratios numerically, so we attempt to find the approximate bounds for $\dfrac{E_{\boldsymbol{\alpha}+\boldsymbol{\gamma_i}}[\exp(-\delta\Delta(\mathbf{P}))]}{E_{\boldsymbol{\alpha}}[\exp(-\delta\Delta(\mathbf{P}))]}$ and $\dfrac{E_{\boldsymbol{\alpha}+2\boldsymbol{\gamma_i}}[\exp(-\delta\Delta(\mathbf{P}))]}{E_{\boldsymbol{\alpha}}[\exp(-\delta\Delta(\mathbf{P}))]}$. The different approaches to estimate this ratio using sampling techniques are discussed in the next section. For our analysis, we focus on the penalty function $\Delta = \sum_{j=1}^{K-1}(p_{j+1}-p_j)^2$.

When $p_1 = p_2 = \ldots = p_K = \dfrac{1}{K}$ then

$$\Delta = \left(\frac{1}{K}-\frac{1}{K}\right)^2 + \left(\frac{1}{K}-\frac{1}{K}\right)^2 + \ldots + \left(\frac{1}{K}-\frac{1}{K}\right)^2 = 0.$$

This is the minimum value of this penalty function $\Delta = \sum_{j=1}^{K-1}(p_{j+1}-p_j)^2$. Also, the maximum value of Δ is attained when one of the p's is 1. Then the maximum

value of Δ is

$$\begin{aligned}
\Delta &= \sum_{j=1}^{K-1}(p_{j+1}-p_j)^2, \\
&\leq \sum_{j=1}^{K-1} p_{j+1}^2 + \sum_{j=1}^{K-1} p_j^2, \\
&\leq \sum_{j=1}^{K-1} p_{j+1} + \sum_{j=1}^{K-1} p_j, \\
&\leq 1+1, \\
&= 2.
\end{aligned}$$

Then

$$\exp(-2\delta)\left(\frac{\alpha_i}{\sum_{j=1}^{K}\alpha_j}\right) \leq E(P_i) \leq \exp(2\delta)\left(\frac{\alpha_i}{\sum_{j=1}^{K}\alpha_j}\right). \tag{4.6}$$

Note that when $\delta \to 0$, $E(P_i) \to \dfrac{\alpha_i}{\sum_{j=1}^{K}\alpha_j}$ which is the expected value of the Dirichlet distribution. We can also calculate the lower limit of $\mathrm{Var}(P_i)$ (denoted by $\mathrm{Var}_{LL}(P_i)$) and the upper limit of $\mathrm{Var}(P_i)$ (denoted by $\mathrm{Var}_{UL}(P_i)$) separately. Then

$$\begin{aligned}
\mathrm{Var}_{LL}(P_i) = &\left(\frac{\alpha_i\left(\sum_{j=1}^{K}\alpha_j-\alpha_i\right)}{\left(\sum_{j=1}^{K}\alpha_j\right)^2\left(\sum_{j=1}^{K}\alpha_j+1\right)}\right) \\
&-\left(\frac{2\delta\alpha_i}{\left(\sum_{j=1}^{K}\alpha_j\right)^2\left(\sum_{j=1}^{K}\alpha_j+1\right)}\right)\left(2\delta\alpha_i\left(\sum_{j=1}^{K}\alpha_j+1\right)+(1+\alpha_i)\sum_{j=1}^{K}\alpha_j+2\alpha_i\left(\sum_{j=1}^{K}\alpha_j+1\right)\right).
\end{aligned} \tag{4.7}$$

and

$$\begin{aligned}
\mathrm{Var}_{UL}(P_i) = &\left(\frac{\alpha_i\left(\sum_{j=1}^{K}\alpha_j-\alpha_i\right)}{\left(\sum_{j=1}^{K}\alpha_j\right)^2\left(\sum_{j=1}^{K}\alpha_j+1\right)}\right) \\
&-\left(\frac{2\delta\alpha_i}{\left(\sum_{j=1}^{K}\alpha_j\right)^2\left(\sum_{j=1}^{K}\alpha_j+1\right)}\right)\left(2\delta\alpha_i\left(\sum_{j=1}^{K}\alpha_j+1\right)-(1+\alpha_i)\sum_{j=1}^{K}\alpha_j-2\alpha_i\left(\sum_{j=1}^{K}\alpha_j+1\right)\right).
\end{aligned} \tag{4.8}$$

Note that $\lim_{\delta\to 0}\text{Var}_{LL}(P_i) = \lim_{\delta\to 0}\text{Var}_{UL}(P_i) \to \dfrac{\alpha_i\left(\sum_{j=1}^{K}\alpha_j - \alpha_i\right)}{(\sum_{j=1}^{K}\alpha_j)^2(\sum_{j=1}^{K}\alpha_j + 1)}$, which is the variance of the Dirichlet distribution. The derivation of (4.6) - (4.8) are provided in the Appendix C. In Section 4.4, we propose two statistical methods using smoothed Dirichlet distribution to estimate multinomial cell probabilities under sparsity.

4.4 Proposed Statistical Model and Estimation

We now outline how to estimate empirical Bayes estimator for $\mathbf{p}_i$ using smoothed Dirichlet distribution.

4.4.1 Bayes Estimation using Smoothed Dirichlet Prior with Fixed δ

We suggest using the smoothed Dirichlet distribution as the prior distribution for $\mathbf{p_i}$,

$$\mathbf{p}_i = (p_{i1}, p_{i2}, \ldots, p_{iK})^t \sim \text{SD}(\boldsymbol{\alpha}, \delta, \Delta).$$

Like the Dirichlet distribution, the smoothed Dirichlet distribution is also a conjugate prior for the vector of multinomial cell probabilities. Then, the posterior distribution of $\mathbf{p}_i$, given the observed counts for the i^{th} multinomial population, is

$$\mathbf{p}_i|\mathbf{x}_i; \boldsymbol{\alpha} \sim \text{SD}(\boldsymbol{\alpha} + \boldsymbol{x}_i, \delta, \Delta).$$

Then following (4.4), the Bayes estimator of p_{ij} is

$$\hat{p}_{ij}^{\text{Bayes}} = \frac{x_{ij} + \alpha_i}{\sum_{j=1}^{K}(x_{ij} + \alpha_j)} \times \frac{E_{\boldsymbol{\alpha}+\boldsymbol{x}_i+\boldsymbol{\gamma_i}}[\exp(-\delta\Delta(\mathbf{P}_i))]}{E_{\boldsymbol{\alpha}+\boldsymbol{x}_i}[\exp(-\delta\Delta(\mathbf{P}_i))]},$$

where $\gamma_{il} = 0, \forall l, l \neq i$ and $\gamma_{ii} = 1$. Further, it can be shown that the Bayes estimator of p_{ij} can also be written as

$$\hat{p}_{ij}^{\text{Bayes}} = \epsilon_{1,ij}(\delta)t_j + \epsilon_{2,ij}(\delta)\hat{p}_{ij}^{\text{MLE}}, \tag{4.9}$$

where

$$\epsilon_{1,ij}(\delta) = \frac{E_{\boldsymbol{\alpha}+\boldsymbol{x}_i+\boldsymbol{\gamma_i}}[\exp(-\delta\Delta(\mathbf{P}_i))]}{E_{\boldsymbol{\alpha}+\boldsymbol{x}_i}[\exp(-\delta\Delta(\mathbf{P}_i))]} \times \frac{\sum_{j=1}^K \alpha_j}{n_i + \sum_{j=1}^K \alpha_j} \text{ and } \epsilon_{2,ij}(\delta) = \frac{E_{\boldsymbol{\alpha}+\boldsymbol{x}_i+\boldsymbol{\gamma_i}}[\exp(-\delta\Delta(\mathbf{P}_i))]}{E_{\boldsymbol{\alpha}+\boldsymbol{x}_i}[\exp(-\delta\Delta(\mathbf{P}_i))]} \times \frac{n_i}{n_i + \sum_{j=1}^K \alpha_j}.$$

As in (4.1), t_j is the shrinkage target corresponding to the prior mean of p_{ij} and $(\epsilon_{1,ij}, \epsilon_{2,ij})^t$ are the shrinkage constants. Using an empirical Bayes approach, we estimate t_j, $\epsilon_{1,ij}$ and $\epsilon_{2,ij}$ using their sample counterparts,

$$\hat{t}_j = \left(\frac{\hat{\alpha}_j}{\sum_{j=1}^K \hat{\alpha}_j} \right), \hat{\epsilon}_{1,ij}(\delta) = \frac{E_{\hat{\boldsymbol{\alpha}}+\boldsymbol{x_i}+\boldsymbol{\gamma_i}}[\exp(-\delta\Delta(\mathbf{P}_i))]}{E_{\hat{\boldsymbol{\alpha}}+\boldsymbol{x_i}}[\exp(-\delta\Delta(\mathbf{P}_i))]} \times \frac{\sum_{j=1}^K \hat{\alpha}_j}{n_i + \sum_{j=1}^K \hat{\alpha}_j} \text{ and } \hat{\epsilon}_{2,ij}(\delta) = \frac{E_{\hat{\boldsymbol{\alpha}}+\boldsymbol{x_i}+\boldsymbol{\epsilon_i}}[\exp(-\delta\Delta(\mathbf{P}_i))]}{E_{\hat{\boldsymbol{\alpha}}+\boldsymbol{x_i}}[\exp(-\delta\Delta(\mathbf{P}_i))]} \times \frac{n_i}{n_i + \sum_{j=1}^K \hat{\alpha}_j}.$$

Then, we can write the empirical Bayes approximation to (4.9) as a different type of scaled shrinkage estimator,

$$\hat{p}_{ij}^{\text{EBayes}} = \hat{c}_{ij}(\delta)\left[\hat{\lambda}_i t_j + (1-\hat{\lambda}_i)\hat{p}_{ij}^{\text{MLE}}\right] \text{ where } \hat{c}_{ij}(\delta) = \frac{E_{\hat{\boldsymbol{\alpha}}+\boldsymbol{x}_i+\boldsymbol{\gamma_i}}[\exp(-\delta\Delta(\mathbf{P}_i))]}{E_{\hat{\boldsymbol{\alpha}}+\boldsymbol{x}_i}[\exp(-\delta\Delta(\mathbf{P}_i))]}, \tag{4.10}$$

with $\hat{\lambda}_i$ defined as in (4.2). It can be seen that $c_{ij}(\delta)$ is a population and cell-specific scale factor which will inflate or deflate the shrinkage estimator. The proposed scaled shrinkage estimator in (4.10) allows for simultaneous inference for many multinomial populations and the sharing of relevant information between multinomial populations and across cell categories to provide improved inference under sparsity. For a given set of observed data, there are three approaches to approximate $\hat{p}_{ij}^{\text{EBayes}}$, which cannot be computed exactly, and the advantages and disadvantages of each approach are

briefly discussed below.

A first simple approach is to approximate the empirical Bayes estimator based on (4.10). Specifically, in this expression for the Bayes estimator, only $\hat{c}_{ij}(\delta)$ is not known. This can be easily approximated using two independent Monte Carlo simulations. For the denominator, generate

$$\mathbf{p}_i \sim \text{Dirichlet}(\hat{\boldsymbol{\alpha}} + \boldsymbol{x}_i),$$

and for the numerator,

$$\mathbf{p}_i \sim \text{Dirichlet}(\hat{\boldsymbol{\alpha}} + \boldsymbol{x}_i + \boldsymbol{\gamma}_j).$$

Then, we approximate the scale factor $\hat{c}_{ij}(\delta)$ by

$$\hat{c}_{ij}(\delta) \approx \frac{\hat{E}_{\hat{\boldsymbol{\alpha}}+\boldsymbol{x}_i+\boldsymbol{\gamma}_i}[\exp(-\delta\Delta(\mathbf{P}_i))]}{\hat{E}_{\hat{\boldsymbol{\alpha}}+\boldsymbol{x}_i}[\exp(-\delta\Delta(\mathbf{P}_i))]},$$

where $\hat{E}$ denotes the average of the sampled vectors. The main advantage of this approach is that $\hat{c}_{ij}(\delta)$ can then be approximated for any values of δ using the same set of randomly generated vectors. Also, this approach is easily implemented and computationally efficient. Unfortunately, this approach suffers from what we consider to be a major issue: the resulting approximate empirical Bayes estimates usually do not sum to 1 $\left(\sum_{j=1}^{K} \hat{p}_{ij}^{\text{EBayes}} \neq 1\right)$ and the behaviors of the approximate scale factor $\hat{c}_{ij}(\delta)$ tend to be problematic, especially when δ is very large. In particular, the MC approximation of $\hat{c}_{ij}(\delta)$ does not seem to stabilize even when the number of randomly generated vectors are very large.

As an alternative to the approach presented above, a second approach to approximate the Bayes estimator would be to directly generate random vectors according to

$$\mathbf{p}_i \sim \text{SD}(\hat{\boldsymbol{\alpha}} + \boldsymbol{x}_i; \delta; \Delta),$$

and take their average using a direct Monte Carlo method. For this, we use the acceptance-rejection method with the ordinary Dirichlet distribution as the proposal distribution. One advantage of this approach is that the resulting approximate empirical Bayes estimates will from a proper distribution by appropriately summing to units ($\sum_{j=1}^{K} \hat{p}_{ij}^{\text{EBayes}} = 1$). One important disadvantage of this approach, however, is very low acceptance rate of the Dirichlet proposals for moderate and large δ values. In fact, this issue makes this approach totally impractical for large values of δ. Given the issues raised above with the first two approaches to approximating the $\hat{p}_{ij}^{\text{EBayes}}$, we recommend a third approach based on importance sampling. This third approach is computationally efficient in terms of time and allows one to recycle the generated random vectors and use them for any values of δ, and leads to approximate estimates that leads to a proper probability distribution by satisfying $\sum_{j=1}^{K} \hat{p}_{ij}^{\text{EBayes}} = 1$. The key step here is to rewrite the numerator of $\hat{c}_{ij}(\delta)$ as

$$E_{\hat{\boldsymbol{\alpha}}+\boldsymbol{x}_i+\boldsymbol{\gamma}_i}[\exp(-\delta\Delta(\mathbf{P}_i))] = \frac{\sum_{j=1}^{K} \hat{\alpha}_{ij} + n_i}{\hat{\alpha}_{ij} + x_{ij}} E_{\hat{\boldsymbol{\alpha}}+\boldsymbol{x}_i}[p_{ij}\exp(-\delta\Delta(\mathbf{P}_i))].$$

Then,

$$\hat{c}_{ij}(\delta) = \left(\frac{\sum_{j=1}^{K} \hat{\alpha}_{ij} + n_i}{\hat{\alpha}_{ij} + x_{ij}} \right) \times \frac{E_{\hat{\boldsymbol{\alpha}}+\boldsymbol{x}_i}[p_{ij}\exp(-\delta\Delta(\mathbf{P}_i))]}{E_{\hat{\boldsymbol{\alpha}}+\boldsymbol{x}_i}[\exp(-\delta\Delta(\mathbf{P}_i))]},$$

with both expectations being taken under the same Dirichlet distribution (that is independent of δ). The approximation for $\hat{c}_{ij}(\delta)$ is then computed using the importance sampling approach to Monte Carlo and it is given by

$$\hat{c}_{ij}(\delta) \approx \left(\frac{\sum_{j=1}^{K} \hat{\alpha}_{ij} + n_i}{\hat{\alpha}_{ij} + x_{ij}} \right) \times \frac{\frac{1}{n}\sum_{l=1}^{n} p_{ijl}\exp(-\delta\Delta(\mathbf{p}_{il}))}{\frac{1}{n}\sum_{l=1}^{n}\exp(-\delta\Delta(\mathbf{p}_{il}))}$$

where the random vectors $\mathbf{p}_{i1}, \mathbf{p}_{i2}, \ldots, \mathbf{p}_{in}$ are generated from the regular Dirichlet distribution with parameter $\hat{\boldsymbol{\alpha}} + \boldsymbol{x}_i$. Then, the empirical Bayes estimator of p_{ij} can be approximated using (4.10) and the above approximation for $\hat{c}_{ij}(\delta)$. In Section 5, we use this approach to find the estimates of the proposed shrinkage estimator of p_{ij}. More details of the acceptance-rejection method and the importance sampling method are provided in the Appendix C. In the following section, we consider the use of smoothed Dirichlet distributions for Bayesian multinomial regression, which allows to include covariates in the estimation problem.

4.4.2 Bayesian Multinomial Regression Estimation

We now describe a Bayesian multinomial regression approach for estimating p_{ij} in the presence of covariates. Let $\mathbf{Y}$ represent an array with l covariates for the m populations and K categories. In the Bayesian multinomial regression model formulation, we write our estimation setting as

$$\begin{array}{lcll} \mathbf{x_i}|\mathbf{p_i} & \sim & \text{Multinomial}(n_i, \mathbf{p_i}), & \\ \mathbf{p_i}|\boldsymbol{\alpha}_i & \sim & \text{SD}(\boldsymbol{\alpha}_i, \delta, \Delta), & \\ \boldsymbol{\alpha_i} & = & \exp(\boldsymbol{\gamma_i}) & \text{where } \boldsymbol{\gamma_i} = (\gamma_{i1}, \gamma_{i2}, \ldots, \gamma_{iK})^t, \end{array}$$

and assume

$$\gamma_{ij} = \mu_i + \sum_{a=1}^{l} \beta_{aj} Y_{aij} \qquad a = 1, 2, \ldots, l, i = 1, 2, \ldots, m \text{ and } j = 1, \ldots, K.$$

In this setup, β_{aj} captures the effect of a^{th} covariate on the j^{th} category and μ_i captures the effect of i^{th} population. We assume normal priors on β_{aj}'s and μ_i's. Specifically, we use

$$\begin{array}{rcll} \beta_{aj} & \sim & \text{N}\ (0,1) & a = 1, 2, \ldots, l \text{ and } j = 1, 2, \ldots, K. \\ \mu_i & \sim & \text{N}\ (0, \sigma^2) & i = 1, 2, \ldots, m \end{array}$$

In what follows, $\mathbf{x} = (\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_m})$, let $\boldsymbol{\beta} = (\boldsymbol{\beta_1}, \ldots, \boldsymbol{\beta_K})$, where $\boldsymbol{\beta_j} = (\beta_{1j}, \beta_{2j}, \ldots, \beta_{lj})^t$, $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_m)^t$ and $\mathbf{p} = (\mathbf{p_1}, \mathbf{p_2}, \ldots, \mathbf{p_m})$. Then, we have

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

and

$$\pi(\mathbf{p_i}|\boldsymbol{\alpha_i}) = C(\boldsymbol{\alpha_i}, \delta) \prod_{j=1}^{K} p_{ij}^{\alpha_{ij}-1} \exp(-\delta \Delta(\mathbf{p}_i)),$$

where $\pi(.)$ is the probability density function. We substitute $\alpha_{ij} = \exp(\gamma_{ij}) = \exp(\mu_i + \sum_{a=1}^{l} \beta_{aj} Y_{aij})$ to get

$$\pi(\mathbf{p}_i|\boldsymbol{\beta}, \mu_i) = C(\boldsymbol{\beta}, \mu_i, \boldsymbol{Y}_i, \delta) \prod_{j=1}^{K} p_{ij}^{\exp\left(\mu_i + \sum_{a=1}^{l} \beta_{aj} Y_{aij}\right) - 1} \exp(-\delta \Delta(\mathbf{p}_i)).$$

where $C(\boldsymbol{\beta}, \mu_i, \boldsymbol{Y}_i, \delta)$ is the normalizing constant after substituting $\alpha_{ij} = \exp(\mu_i + \sum_{a=1}^{l} \beta_{aj} Y_{aij})$. The posterior distribution is

$$\pi(\mathbf{p}, \boldsymbol{\beta}, \boldsymbol{\mu}|\mathbf{x}) \propto \pi(\mathbf{x}|\boldsymbol{p}) \times \pi(\mathbf{p}|\boldsymbol{\beta}) \times \pi(\boldsymbol{\beta})$$

$$= \prod_{i=1}^{m} \frac{n_i}{\prod_{j=1}^{K} x_{ij}!} \prod_{j=1}^{K} p_{ij}^{x_{ij}} \times C(\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{Y_i}, \delta) \prod_{j=1}^{K} p_{ij}^{\exp\left(\mu_i + \sum_{a=1}^{l} \beta_{aj} Y_{aij}\right) - 1} \exp(-\delta \Delta(\mathbf{p}_i))$$

$$\times \prod_{a=1}^{l} \prod_{j=1}^{K} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\beta_{aj}^2}{2}\right) \times \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\mu_i^2}{2\sigma^2}\right)$$

$$\propto \prod_{i=1}^{m} \prod_{j=1}^{K} p_{ij}^{x_{ij} + \exp\left(\mu_i + \sum_{a=1}^{l} \beta_{aj} Y_{aij}\right) - 1} \exp(-\delta \Delta(\mathbf{p}_i)) \times \exp\left(-\sum_{a=1}^{l} \sum_{j=1}^{K} \frac{\beta_{aj}^2}{2}\right)$$

$$\times \exp\left(-\sum_{i=1}^{m} \frac{\mu_i^2}{2\sigma^2}\right). \tag{4.11}$$

Now by holding $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$ fixed,

$$\pi(\mathbf{p}|\boldsymbol{\beta}, \boldsymbol{\mu}, \mathbf{x}) \propto \prod_{i=1}^{m} \prod_{j=1}^{K} p_{ij}^{x_{ij} + \exp\left(\mu_i + \sum_{a=1}^{l} \beta_{aj} Y_{aij}\right) - 1} \exp(-\delta \Delta(\mathbf{p}_i)),$$

implying conditional independence (given $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$) of the $\mathbf{p}_i$'s with marginal PDF given by

$$\pi(\mathbf{p_i}|\boldsymbol{\beta}, \mu_i, \mathbf{x_i}) \propto \prod_{j=1}^{K} p_{ij}^{x_{ij} + \exp\left(\mu_i + \sum_{a=1}^{l} \beta_{aj} Y_{aij}\right) - 1} \exp(-\delta \Delta(\mathbf{p}_i)),$$

Letting $\gamma_{ij}^* = x_{ij} + \exp\left(\mu_i + \sum_{a=1}^{l} \beta_{aj} Y_{aij}\right)$ and $\gamma_{\mathbf{i}}^* = (\gamma_{i1}^*, \gamma_{i2}^*, \ldots, \gamma_{iK}^*)^t$, the posterior conditional distribution of $\mathbf{p_i}$ is

$$\mathbf{p_i}|\boldsymbol{\beta}, \boldsymbol{\mu}, \mathbf{x_i} \sim \mathrm{SD}(\gamma_{\mathbf{i}}^*, \delta, \Delta).$$

When holding $\mathbf{p}$ and $\boldsymbol{\mu}$ fixed, however,

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

$$\times \exp\left(-\sum_{a=1}^{l}\sum_{j=1}^{K} \frac{\beta_{aj}^2}{2}\right) \times \frac{1}{E_{\boldsymbol{\alpha_i}}[\exp(-\delta\Delta(\mathbf{P}_i))]},$$

so that the posterior conditional distribution of $\boldsymbol{\beta}$ (given $\mathbf{p}$ and μ) does not belong to a standard family of distributions. To perform inference based on the full posterior distribution (4.11), we developed a Metropolis within Gibbs algorithm to generate values from this distribution. Similarly, when holding $\mathbf{p}$ and $\boldsymbol{\beta}$ fixed,

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

$$\times \frac{1}{E_{\boldsymbol{\alpha_i}}[\exp(-\delta\Delta(\mathbf{P}_i))]},$$

and this posterior conditional distribution of $\boldsymbol{\mu}$ (given $\mathbf{p}$ and $\boldsymbol{\beta}$) does not belong to a standard family of distributions. Metropolis within Gibbs algorithm is used to generate values from this posterior distribution.

4.5 Data Analysis

4.5.1 Dataset

Coronavirus disease (COVID-19) is an infectious disease caused by a newly discovered coronavirus (SARS-CoV-2) that was first identified in Wuhan, China in December 2019. COVID-19 was declared a global pandemic in March 2020. The coronavirus is mainly transmitted through droplets generated when an infected person coughs, sneezes, or exhales.

Here, we consider early COVID-19 positive cases up to June 25, 2020, in Canada, excluding Northwest Territories. COVID-19 positive cases are extracted based on age groups and health regions from provincial websites. In our analysis, we consider $K = 8$ age groups; <20, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80+. The distributions of COVID-19 positive cases based on age groups and health regions are not publicly available for the provinces Quebec, New Brunswick, Nova Scotia, Saskatchewan, Newfoundland and Labrador, and Prince Edward Island. For those provinces, the provincial level COVID-19 positive cases are extracted based on age groups and we represent them as one health region. The COVID-19 positive cases by the health regions and the age groups are provided in the Appendix C. Also, for some of the age groups of these provinces, the combined COVID-19 positive cases are publicly available, and we split those combined cells using a data augmentation technique.

4.5.2 Issues Around Missing Cell Counts

The data augmentation techniques can be used to impute the missing cell counts of the split categories when the combined cell count is known. We proposed a fully Bayesian, joint modeling approach to impute cell counts involving data augmentation (Tanner and Wong (1987)[9]) and Markov chain Monte Carlo (MCMC) sampling. Suppose that we have m multinomial population with K categories. Each multinomial population (i) has a vector of cell counts $\mathbf{x_i} = (x_{i1}, x_{i2}, \ldots, x_{iK})^t$ with the probability vector $\mathbf{p_i} = (p_{i1}, p_{i2}, \ldots, p_{iK})^t$. For simplicity, let's assume that the last two categories $x_{i(K-1)}$ and x_{iK} are combined with the cell count x for the i^{th} multnomial population. Now using the data augmentation technique, we want to split the cell count x to $x_{i(K-1)}$ and x_{iK}. In the Bayesian imputation method, we assume

$$\begin{aligned} x_{i(K-1)}|p_{i(K-1)}, p_{iK} &\sim \text{Binomial}\left(x, \frac{p_{i(K-1)}}{p_{i(K-1)} + p_{iK}}\right), \\ \frac{p_{i(K-1)}}{p_{i(K-1)} + p_{iK}}|c, d &\sim \text{Beta}(c, d). \end{aligned}$$

Here the posterior distribution of $X_{i(k-1)}$ is a beta-binomial distribution:

$$\pi(x_{i(K-1)} = z|c, d) = \binom{x}{z} \frac{B(z + c, x - z + d)}{B(c, d)},$$

where $B(a, b)$ is the beta function. We used this technique to impute the cell counts of the health regions for the combined age groups.

4.5.3 Shrinkage Estimation

Table 4.1 provides the posterior expected values of the imputed COVID-19 positive cases by the health regions and the age groups for the combined cells using the EM

algorithm which is provided later. The imputed cell counts are displayed by an asterisk after the cell count.

Now, there are 55 health regions, and we assume that the distribution of COVID-19 positive cases based on age groups for each health region has a multinomial distribution. Recall that,

$$\boldsymbol{X}_i|\boldsymbol{p}_i \sim \text{Multinomial } (n_i; \boldsymbol{p}_i = (p_{i1}, p_{i2}, ..., p_{iK})),$$

for each of the $m = 55$ health regions and that $p_{ij}(j = 1, 2, \ldots, 8)$ denotes the cell probability of COVID-19 positive cases for the health region i and age group j. Here the age groups are arranged in order where age group 1 represents that the age is less than 20 (<20), and age group 8 represents that the age is 80 or more (80+).

Table 4.1: Posterior expected values of the imputed COVID-19 positive cases by the health regions and the age groups

Province	Health Region	< 20	20-29	30-39	40-49	50-59	60-69	70-79	80+
NS	Nova Scotia	106	137.6*	138.4*	128.3*	147.7*	147.4*	96.6*	159
SK	Saskatchewan	107	127.7*	132.3*	111.8*	129.2*	81.9*	48.1*	210
NL	Newfoundland	22	18.7*	10.3*	39	58	57	22.6*	24.4*
PE	Prince Edward Island	0	5.5*	4.5*	3.3*	4.7*	5.5*	3.5*	0

Table 4.2: Overall proportion of the COVID-19 positive cases by age group

	<20	20-29	30-39	40-49	50-59	60-69	70-79	80+
$\bar{p}_j^*$	0.071	0.136	0.139	0.153	0.153	0.099	0.075	0.174

The overall proportion (op) $\bar{p}_j^* = \dfrac{\sum_{i=1}^{55} x_{ij}^*}{\sum_{i=1}^{55}\sum_{j=1}^{8} x_{ij}^*}$ for each age group given in Table 4.2 including the imputed cell counts from Nova Scotia, Saskatchewan, Newfoundland and Labrador and Prince Edward Island by splitting the combined age groups. Here x_{ij}^* is the imputed (split the combined categories) or original cell count of the i^{th}

health region and the j^{th} age category. We can clearly see that the lowest overall proportion is for the age group 1 (<20) and the highest overall proportion is for the age group 8 (80+).

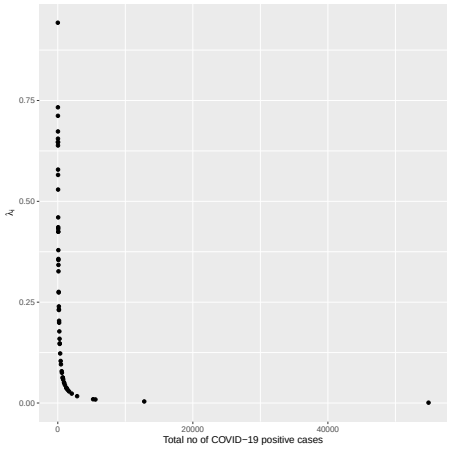


Figure 4.1: Optimal shrinkage constants ($\hat{\lambda}_i$)

The scaled shrinkage estimator (4.10) is

$$\hat{p}_{ij}^{\text{EBayes}} = \hat{c}_{ij}(\delta)\left[\hat{\lambda}_i t_j + (1-\hat{\lambda}_i)\hat{p}_{ij}^{\text{MLE}}\right] \text{ where } \hat{c}_{ij}(\delta) = \frac{E_{\hat{\boldsymbol{\alpha}}+\boldsymbol{x}_i+\boldsymbol{\gamma_i}}[\exp(-\delta\Delta(\mathbf{P}_i))]}{E_{\hat{\boldsymbol{\alpha}}+\boldsymbol{x}_i}[\exp(-\delta\Delta(\mathbf{P}_i))]}, i = 1,\ldots,55 \text{ and } j = 1,\ldots,8.$$

Table 4.3: Estimates of the concentration parameters and the shrinkage targets ($\hat{t}_j$).

	< 20	20-29	30-39	40-49	50-59	60-69	70-79	80+
$\hat{\alpha}_j$	3.201	7.241	7.526	7.300	8.468	6.557	4.146	5.007
$\hat{t}_j$	0.065	0.146	0.152	0.148	0.171	0.133	0.084	0.101

Table 4.3 provides the parameter estimates of the concentration parameters $\hat{\boldsymbol{\alpha}}$ and the estimates of the optimal shrinkage constants for the health regions are given in Figure 4.1. Northern Manitoba health region has the highest optimal shrinkage constant at 0.943 and Quebec has the lowest optimal shrinkage constant at 0.001.

It is clear that Northern Manitoba heath region has the lowest total number of COVID-19 positive cases whereas Quebec has the highest total number of COVID-19 positive cases in Canada. $\hat{\boldsymbol{\alpha}}$ is obtained by ML estimation based on the Dirichlet-multinomial distribution using the EM algorithm, which is provided below. The EM algorithm is widely used to obtain ML estimates for incomplete data, and it has two steps: the expectation step (E-step) uses current estimates of the parameters to find (expectation of) complete data, and the maximization step (M-step) uses the updated data from the E-step to find the ML estimates of the parameters.

EM Algorithm to find MLE of the Dirichlet-multinomial

The goal of EM algorithm is to maximize the likelihood from the observed data and the resulting observed log-likelihood is

$$l(\boldsymbol{\alpha}; \boldsymbol{x}) = \log\, f(\boldsymbol{x}; \boldsymbol{\alpha}),$$

$$= m\, \log\, \Gamma\left(\sum_{j=1}^{K} \alpha_j\right) - m \sum_{j=1}^{K} \log\, \Gamma(\alpha_j) + \sum_{j=1}^{K} \left[(\alpha_j - 1) \sum_{i=1}^{m} \log\, p_{ij}\right],$$

where p_{ij} is the posterior expectation of

$$\mathbf{p}_i \sim \text{Dirichlet}(\boldsymbol{\alpha} + \boldsymbol{x}_i).$$

Define

$$Q(\boldsymbol{\alpha}; \boldsymbol{\alpha}^{(k)}) = \text{E}\left[l(\boldsymbol{\alpha}; \boldsymbol{x}) | \boldsymbol{x}, \boldsymbol{\alpha}^{(k)}\right],$$

$$= m\, \log\, \Gamma\left(\sum_{j=1}^{K} \alpha_j\right) - m \sum_{j=1}^{K} \log\, \Gamma(\alpha_j) + \sum_{j=1}^{K} \left[(\alpha_j - 1) \sum_{i=1}^{m} \text{E}(\log\, p_{ij} | \boldsymbol{x}_i, \boldsymbol{\alpha}^{(k)})\right],$$

where $\text{E}(\log\, p_{ij} | \boldsymbol{x}_i, \boldsymbol{\alpha}^{(k)})$ is the conditional expectation of the logarithm of p_{ij} given observed data and $\boldsymbol{\alpha}^{(k)}$. Let's assume that the last two categories $x_{i(K-1)}$ and x_{iK}

are combined with the cell count x for the i^{th} multnomial population. For the v^{th} iteration, recall that we split the combined cell counts using

$$x^{*(v)}_{i(K-1)} | p^{(v)}_{i(K-1)}, p^{(v)}_{iK} \sim \text{Binomial}\left(x, \frac{p^{(v)}_{i(K-1)}}{p^{(v)}_{i(K-1)} + p^{(v)}_{iK}}\right).$$

Hence

$$x^{*(v)}_{i(K-1)} = x \times \frac{p^{(v)}_{i(K-1)}}{p^{(v)}_{i(K-1)} + p^{(v)}_{iK}}.$$

The EM algorithm proceeds in the following way.

1. Set $v = 0$, and find an appropriate starting values $\boldsymbol{\alpha}^{(0)}$.

2. **E-step**: Impute $\boldsymbol{x}^{*(v)}$ using $\boldsymbol{\alpha}^{(v)}$ and construct

$$Q(\boldsymbol{\alpha}; \boldsymbol{\alpha}^{(v)}) = \text{E}\left[l(\boldsymbol{\alpha}; \boldsymbol{x}^{*(v)}) | \boldsymbol{x}^{*(v)}, \boldsymbol{\alpha}^{(v)}\right].$$

3. **M-step**: Calculate the candidate $\boldsymbol{\alpha}^{(v+1)}$ by solving

$$\boldsymbol{\alpha}^{(v+1)} = \arg\max Q(\boldsymbol{\alpha}; \boldsymbol{\alpha}^{(v)}).$$

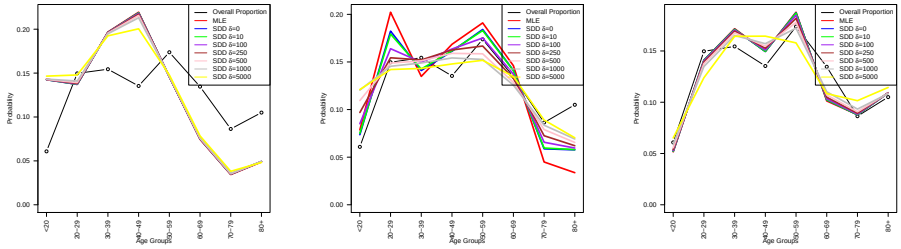
4. If convergence criteria is satisfied, return $\boldsymbol{\alpha}^{(v+1)}$ as root; otherwise, set $v = v+1$ and return to step 2.

The convergence is assessed by verifying that

$$||\boldsymbol{\alpha}^{(v+1)} - \boldsymbol{\alpha}^{(v)}|| < \epsilon,$$

for some small predefined tolerance level $\epsilon > 0$.

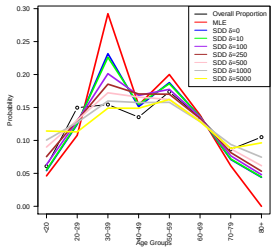
Figure 4.2: Comparison of ML estimates and scaled shrinkage estimates with different δ



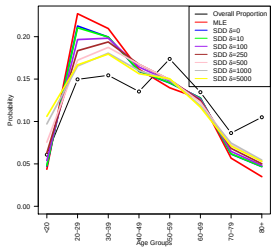
(a) AB - Calgary Zone

(b) AB - Central Zone

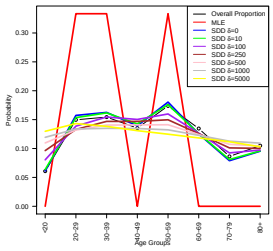
(c) BC - Fraser



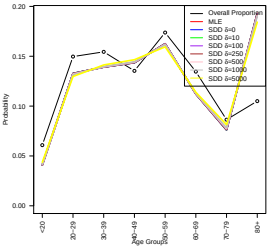
(d) BC - Northern



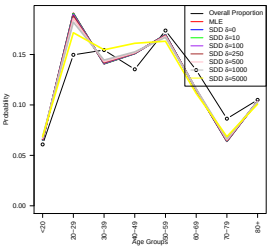
(e) MB - Winnipeg



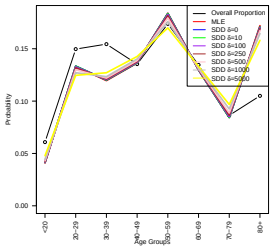
(f) MB - Northern



(g) ON - Toronto



(h) ON - Peel

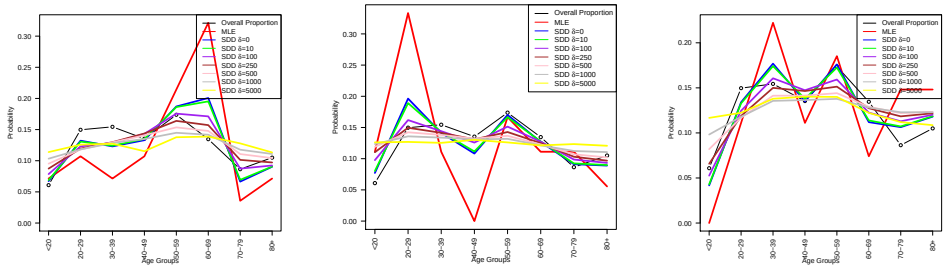


(i) ON - York

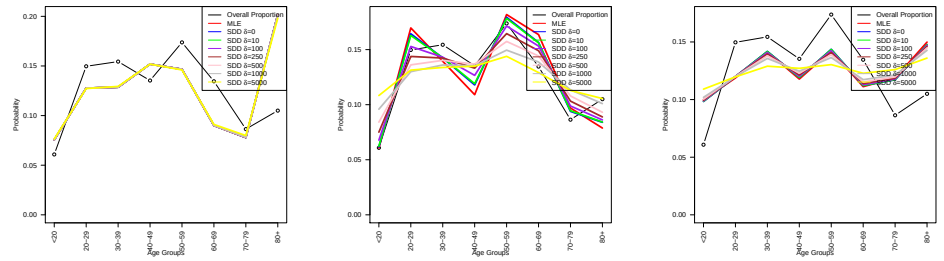
Figure 4.2 and Figure 4.3 provide the comparison of maximum likelihood (ML) estimates and scaled shrinkage estimates with different δ. When the cell counts (or row totals) are large, the effect of δ is very small. It is clear that, for Quebec, Toronto, and Peel, the scale shrinkage estimates have less shrinking and trust the MLE more. But, this is opposite for the health regions with small cell counts and selecting a

suitable δ is very important. When the cell counts are small (Northern Manitoba, North Bay Parry, and Prince Edward Island), the scaled shrinkage estimates have less trust in MLE. Also, when δ increases, the scaled shrinkage estimates get closer to the equal proportions ($p = 1/8$). Note that, when $\delta = 0$, we get the Bayes estimator discussed in Chapter 2 with the Dirichlet prior.

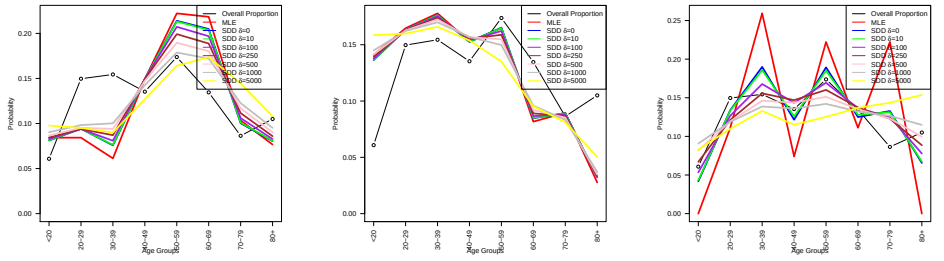
Figure 4.3: Comparison of ML estimates and scaled shrinkage estimates with different δ



(a) ON - North Bay Parry (b) ON - Timiskaming (c) ON - Renfrew



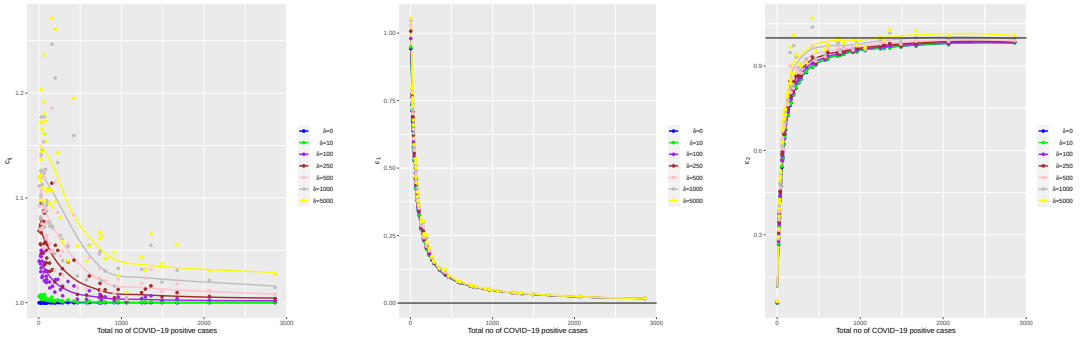
(d) Quebec (e) Nova Scotia (f) New Brunswick



(g) Newfoundland (h) Saskatchewan (i) PEI

Selecting a suitable δ value is very important to control the effect of the penalty term. We picked a δ value by visually examining Figure 4.4. The effect of the penalty is very small for the health regions with a large total number of COVID-19 positive cases even when δ is very large. Figure 4.4 provides the smoothed lines of the average $\hat{c}_{ij}$, $\hat{\epsilon}_{1,ij}$ and $\hat{\epsilon}_{2,ij}$ of the empirical Bayes estimates for the health regions with a small total number of COVID-19 positive cases (<3000) with different δ leaving Quebec, Toronto, York, Peel and Calgary out of the graphing regions. It is clear that there is no impact for $\hat{\epsilon}_{1,ij}$ by changing δ value. We picked $\delta = 500$, which is the highest δ value where the smoothed line for $\hat{\epsilon}_{2,ij}$ is below 1, implying there is no over shrinking.

Figure 4.4: $\hat{c}_{ij}$, $\hat{\epsilon}_{1,ij}$ and $\hat{\epsilon}_{2,ij}$ with different δ



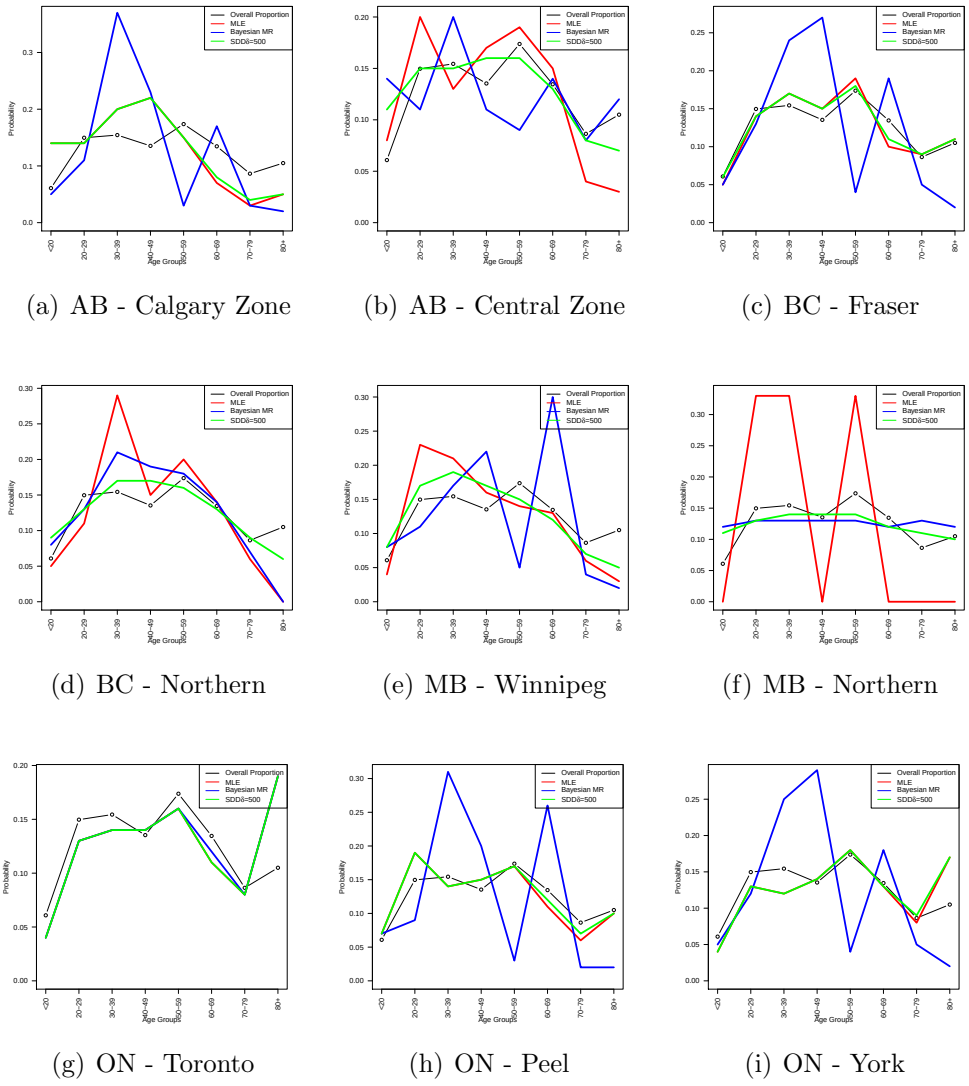
(a) $\hat{c}_{ij}(\delta)$ (b) $\hat{\epsilon}_{1,ij}(\delta)$ (c) $\hat{\epsilon}_{2,ij}(\delta)$

4.5.4 Inference with Covariates

For the Bayesian multinomial regression model, we considered the following covariates: number of COVID-19 positive cases per 1000 COVID-19 tests, number of COVID-19 tests per 1000 people, type of health region (0='Rural', 1='Urban'), and the population density of the health region. The geographical maps for these covariates

are provided in the Appendix C.

Figure 4.5: Comparison of ML and BMR estimates



(a) AB - Calgary Zone (b) AB - Central Zone (c) BC - Fraser

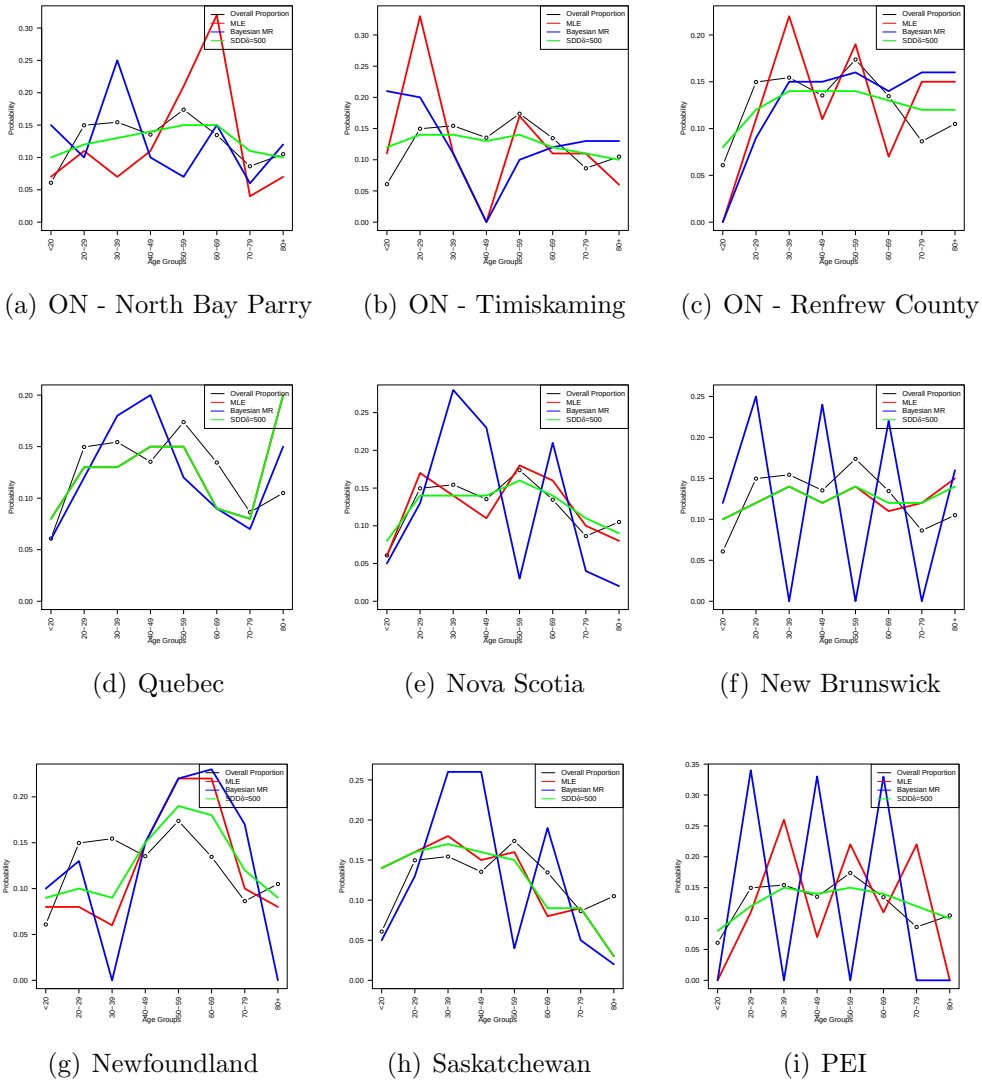
(d) BC - Northern (e) MB - Winnipeg (f) MB - Northern

(g) ON - Toronto (h) ON - Peel (i) ON - York

A health region is considered to be an 'Urban' health region if it has at least one large urban population center. According to STATCAN (https://www.statcan.gc.ca), the large urban population center has a population of 100,000 or more and a population density of 400 persons or more per square kilometer. The provinces

(Quebec, Saskatchewan, etc.) were considered 'Urban' health regions. The population density calculates the population per square kilometer and we used 2019 population counts extracted from provincial/health regions websites. There is no collinearity between the covariates.

Figure 4.6: Comparison of ML and BMR estimates



(a) ON - North Bay Parry (b) ON - Timiskaming (c) ON - Renfrew County

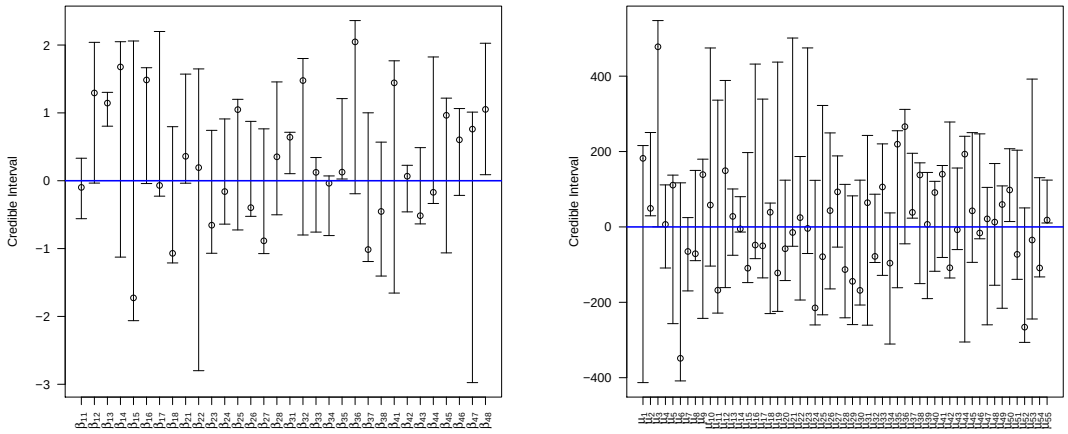
(d) Quebec (e) Nova Scotia (f) New Brunswick

(g) Newfoundland (h) Saskatchewan (i) PEI

Figure 4.5 and Figure 4.6 provide the comparison of maximum likelihood (ML) estimates and Bayesian multinomial regression (BMR) estimates with $\delta = 500$. In the Bayesian multinomial regression model, 25,000 draws were taken with a burn-in of 20,000 under the following parameters:

$\beta_{aj} \sim \text{N}\ (0, 1)$ and $\mu_i \sim \text{N}\ (0, 10)$ where $i = 1, 2, \ldots, 55; j = 1, 2, \ldots, 8$ and $a = 1, 2, 3, 4$.

Figure 4.7: 95% credible interval



(a) β_{aj}s (b) μ_is

The trace plots for some of the β_{ij}'s are provided in the Appendix C. Figure 4.7 represents the 95% posterior credible intervals for β_{aj}'s and μ_i's from the Bayesian multinomial regression model. The posterior means of β_{aj}'s fluctuate between -3 and 2, and all of the significant β_{aj}'s have a positive contribution. The significant β_{aj}'s are β_{13}, β_{31}, β_{35}, and β_{48}. The posterior means of μ_i's fluctuate between -400 and 400. The significant positive contributions are μ_2, μ_{37}, μ_{50}, and μ_{55}.

4.6 Discussion

In this chapter, we introduced a Bayesian approach using smoothed Dirichlet distribution. This proposed approach allows us to borrow information across other health regions and age groups to improve the estimation of cell probabilities. The main advantage of the proposed approach is that it may be possible to obtain a better estimation by sharing information among health regions and age groups in a meaningful manner.

We remark that choosing a suitable δ for smoothed Dirichlet distribution is challenging. We used a range from 0 to 10,000, and we picked the δ by considering the amount of shrinking. Also, the suitable δ value is data specific, and depends on the data. To overcome this issue, one can introduce a prior distribution for δ and use MCMC to estimate δ. $\hat{\boldsymbol{\alpha}}$ is obtained by ML estimation based on the Dirichlet-multinomial distribution using the EM algorithm. Also, one can use the empirical Bayes approach to obtain ML estimation based on the Dirichlet-multinomial distribution and $\hat{\boldsymbol{\alpha}}$ estimates are very close in these two approaches. The acceptance-rejection rates for β's and μ's were low and we used many techniques to improve the acceptance-rejection rates for β's and μ's, and the method we discussed here provides the highest rates out of all those techniques.

Bibliography

[1] A. Agresti and D. Hitchcock. Bayesian inference for categorical analysis. *Statistical Methods and Applications*, 14:297–330, 2005.

[2] S. Fienberg and P. Holland. Bayesian inference for categorical analysis. *Journal of the American Statistical Association*, 68:683–691, 1973.

[3] J. Hausser and K. Strimmer. Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. *Journal of Machine Learning Research*, 10:1469–1484, 2009.

[4] N. L. Hjort. *Bayesian Statistics 5*, chapter Bayesian approaches to non- and semiparametric density estimation, pages 223–254. Oxford - University Press, 1996.

[5] W. James and C. Stein. Estimation with quadratic loss. volume 1, pages 361–379. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, 1961.

[6] O. Ledoit and M. Wolf. Improved estimation of the covariance matrix of stock return with an application to portfolio selection. *Journal of Empirical Finance*, 10:603–621, 2003.

[7] J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

[8] C. Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. volume 1, pages 197–206, Berkeley and Los Angeles, 1956. Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, University of California Press.

[9] M. Tanner and W. Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 398(82):528–540, 1987.

[10] T. Tvedebrink. Overdispersion in allelic counts and l-correction in forensic genetics. *Theoretical Population Biology*, (200-210), 2010.

Chapter 5

Conclusion

5.1 Summary of Main Contributions

In this thesis, we presented three improved estimation methods to estimate multinomial cell probabilities applicable in a context of sparsity. Each approach borrows information from other multinomial populations but does so by proceeding in a different way.

In Chapter 2, we developed a semi-parametric Bayesian approach for estimating a vector $\mathbf{p}$ of multinomial cell probabilities using a Dirichlet process prior. The semi-parametric Bayesian approach relaxes some parametric assumptions while naturally borrowing information across other multinomial distributions through the Dirichlet prior. In particular, the Dirichlet process enables model-based clustering of the multinomial distributions via MCMC exploration: throughout the iteration of the MCMC implementation, the clustering assignments will change.

Our weighted likelihood methodology presented in Chapter 3, borrows information from other similar multinomial populations to make inference about a target multi-

nomial population. The estimator based on the weighted likelihood approach is a different type of shrinkage estimator, and it is constructed by combining the MLE with another estimator t (shrinkage target) derived from information associated with all the multinomial populations. An important aspect of how this is achieved is that the shrinkage target differs for each population for which inference is performed. This is a major difference with traditional shrinkage estimators.

Finally, both the semi-parametric Bayesian estimator using the Dirichlet process and the estimator based on the weighted likelihood approach borrow the information across other multinomial distributions to improve the estimation of cell probabilities, but do not allow information sharing across neighboring cells of the same multinomial population. In Chapter 4, we developed a Bayesian approach using smoothed Dirichlet prior distributions to borrow information across other multinomial populations but more importantly, also across neighboring ordinal categories to improve the estimation of cell probabilities. This approach allows to make improved simultaneous inference for many multinomial populations under sparsity.

For each proposed estimator, we considered different data applications. An in-depth simulation study would be useful to demonstrate and compare how the proposed and existing estimators perform over a simulated datasets and considering different scenarios. Nevertheless, we performed a brief simulation study using 10000 Monte Carlo simulations, knowing the true cell probabilities. We report the Mean Squared Error (MSE) and compare the estimators below. Also, we varied the number of populations (N - 100, 200 and 500) and number of categories (K- 5, 10 and 15) to explore the performance of the estimators. We generated data from multinomial

distributions with sample size ranged from 15 to 75. We considered three scenarios for the true cell probabilities.

Scenario 1

In this scenario, the true cell probabilities are strictly decreasing but the differences between successive $\mathbf{p}$'s are the same. For example, when $K = 5$, the true cell probabilities are $\mathbf{p} = \left(\frac{5}{15}, \frac{4}{15}, \frac{3}{15}, \frac{2}{15}, \frac{1}{15}\right)^t$.

Table 5.1: MSE values for scenario 1

Estimator	N=100			N=200			N=500		
	$K = 5$	$K = 10$	$K = 15$	$K = 5$	$K = 10$	$K = 15$	$K = 5$	$K = 10$	$K = 15$
MLE	0.1240	0.0561	0.0456	0.1227	0.0544	0.0454	0.1226	0.0542	0.0453
James-Stein	0.1120	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	0.0530	0.0437
Dirichlet Process	0.1102	0.0501	0.0425	0.1086	0.0472	0.0422	0.1085	0.0471	0.0422
Weighted Likelihood	0.0973	0.0444	0.0413	0.0946	0.0412	0.0412	0.0940	0.0412	0.0411
Smoothed Dirichlet	0.0969	0.0442	0.0411	0.0944	0.0411	0.0411	0.0938	0.0410	0.0411

Table 5.1 provides the mean squared error values for each estimator based on Scenario 1. When N is fixed and K increases the MSE decreases. The MSE also decreases when K is fixed and N increases. The proposed estimators perform better compared to the existing estimators and MLE. The Bayesian shrinkage estimator based on a smoothed Dirichlet prior is the best estimator based on the MSE.

Scenario 2

In this scenario, the true cell probabilities are strictly increasing but the difference between the successive $\mathbf{p}$'s are again the same. For example, when $K = 5$, the true cell probabilities are $\mathbf{p} = \left(\frac{1}{15}, \frac{2}{15}, \frac{3}{15}, \frac{4}{15}, \frac{5}{15}\right)^t$.

Table 5.2: MSE values for scenario 2

Estimator	N=100			N=200			N=500		
	$K=5$	$K=10$	$K=15$	$K=5$	$K=10$	$K=15$	$K=5$	$K=10$	$K=15$
MLE	0.1281	0.0590	0.0468	0.1243	0.0571	0.0465	0.1240	0.0568	0.0463
James-Stein	0.1169	0.0553	0.0452	0.1165	0.0527	0.0450	0.1163	0.0525	0.0450
Empirical Bayes	0.1171	0.0559	0.0457	0.1168	0.0532	0.0453	0.1164	0.0531	0.0453
Dirichlet Process	0.1153	0.0523	0.0439	0.1116	0.0493	0.0438	0.1114	0.0491	0.0438
Weighted Likelihood	0.0995	0.0461	0.0423	0.0970	0.0431	0.0422	0.0966	0.0431	0.0421
Smoothed Dirichlet	0.0987	0.0458	0.0420	0.0963	0.0427	0.0419	0.0960	0.0426	0.0419

Table 5.2 provides the mean squared error values for each estimator based on Scenario 2. As in the previous scenario, when N is fixed and K increases the MSE decreases. The MSE also decreases when K is fixed and N increases. The proposed estimators perform better compared to the existing estimators and MLE. Once again, the Bayesian shrinkage estimator based on the smoothed Dirichlet prior is the best estimator based on the MSE.

Scenario 3

In this scenario, the true cell probabilities are increasing and decreasing (zig zag pattern). For example, when $K = 5$, the true cell probabilities are $\mathbf{p} = \left(\frac{1}{23}, \frac{10}{23}, \frac{1}{23}, \frac{10}{23}, \frac{1}{23}\right)^t$.

Table 5.3: MSE values for scenario 3

Estimator	N=100			N=200			N=500		
	$K=5$	$K=10$	$K=15$	$K=5$	$K=10$	$K=15$	$K=5$	$K=10$	$K=15$
MLE	0.1314	0.0710	0.0636	0.1299	0.0698	0.0690	0.1295	0.0695	0.0689
James-Stein	0.1278	0.0670	0.0614	0.1263	0.0662	0.0608	0.1261	0.0661	0.0607
Empirical Bayes	0.1297	0.0690	0.0625	0.1283	0.0680	0.0675	0.1280	0.0678	0.0674
Dirichlet Process	0.1243	0.0651	0.0602	0.1231	0.0645	0.0600	0.1229	0.0644	0.0600
Weighted Likelihood	0.1212	0.0623	0.0591	0.1201	0.0617	0.0589	0.1199	0.0616	0.0588
Smoothed Dirichlet	0.1288	0.0674	0.0618	0.1275	0.0668	0.0614	0.1272	0.0666	0.0614

Table 5.3 provides the mean squared error values for each estimator based on Scenario 3. As previous scenarios, when N is fixed and K increases the MSE decreases. The

MSE also decreases when K is fixed and N increases. The proposed estimators perform better compared to the existing estimators and MLE. The estimator based on weighted likelihood approach is the best estimator based on the MSE for this scenario. The Bayes estimator based on the smoothed Dirichlet prior is not doing well in this case, which is not surprising given it isn't designed to handle this case where successive cell probabilities are different.

5.2 Future Works

Multinomial data often arise in the form of a contingency table. A contingency table is a type of table which displays the frequency distribution of categorical variables, and it was first used by Pearson (1904)[7]. A typical multinomial data is a $1 \times K$ contingency table with neighboring cells to the left and to the right. A contingency table is still multinomial data, but the structure is different with neighboring cells also above and below on other rows. The cells come from considering combinations of categories and often one's goal is to test for independence of these categories. Contingency tables with small cell counts are said to be sparse. Sparse contingency tables commonly occur when the sample sizes of the populations are small. The number of cells in a contingency table increases multiplicatively when the dimension of the table increases. So higher-dimensional tables are more likely to be sparse even when the sample sizes are large. The methods we discussed in Chapters 2 - 4 may be generalized to contingency tables and, in particular, can help to handle sparsity in these tables. Various methodologies have been developed to measure the association between the ordinal categories in ordinal contingency tables, most of them struggling in the presence of sparsity. This being said, some methods are specifically designed

to handle this problem. For instance, Dong and Simonoff (1995)[1] proposed a technique based on the geometric combination estimators, and Koehler (1986)[5] proposed a goodness-of-fit test of association applicable to sparse multinomial data.

Adaptive weights play a big role in weighted likelihood estimation. We use MAMSE weights, and we can try other data-based weights to compare the estimation. To compute MAMSE weights for each multinomial population, instead of considering all the multinomial populations, we consider a cluster of the multinomial populations with fixed cluster size. One possible future extension is to develop a mechanism to determine the cluster size based on characteristics of the population, the cluster size potentially differing from population to population. For example, if the sample size is small, it might be better to have a large cluster size that helps borrow information from a large set of populations thus using more data for the inferential problem. To cluster the multinomial populations, we use distance measures, and one may use other clustering measures as well. In particular, an interesting and promising idea is to do this clustering based on relevant covariates. For the baseball problem, factors such as league, age, field position, and hitter type (left-handed, right-handed, switch hitter) seem like obvious natural choices. The estimator based on the weighted likelihood approach is a shrinkage estimator and the shrinking weights depend on the parameter estimates $\hat{\alpha} = (\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_K)^t$ associate to the underlying Dirichlet prior distribution. Yu and Shaw (2014)[8] suggested to estimate α using a pre-defined value $A = \sum_{j=1}^{K} \hat{\alpha}_j$ and select A properly in a data-driven way. We briefly experimented with this idea ourselves and, although we could not nail down all the required details, it seemed like a quite promising approach.

The smoothed Dirichlet distribution is a distribution which has very interesting properties. Although it was introduced almost 30 years ago [see Hjort (1996)[3]], little work has been done towards estimating its parameters. An important parameter is the penalty δ which dictates the extent to which cell probabilities of neighboring categories have to be similar. In our work, we have used a fixed value for δ. However, instead of using a fixed δ, it would seem natural to use a fully Bayesian estimation approach and to introduce a prior distribution on δ within a hierarchical modeling framework. The Bayesian shrinkage estimator using the smoothed Dirichlet distribution has two scaling factors $\hat{\epsilon}_{1,ij}(\delta)$ and $\hat{\epsilon}_{2,ij}(\delta)$ for t_j and $\hat{p}_{ij}^{\text{MLE}}$, respectively. A theoretical examination of the behavior of $\hat{\epsilon}_{1,ij}(\delta)$ and $\hat{\epsilon}_{2,ij}(\delta)$ will be undertaken with the goal of better understanding how the sharing of information across populations and categories occurs in sparse multinomial data. The behavior of the shrinkage weights is the key to obtaining improvements in the considered inferential problem. More work on this issue will allow to further enhance the current technique, but will also potentially bring a new understanding to the general shrinkage estimation problem.

We proposed three improved estimation methods to estimate the sparse multinomial cell probabilities. Each approach borrows information from other multinomial populations, but the borrowing is done differently, but all of these techniques could be used in testing for goodness-of-fit. Indeed, Pearson (1990)[6] proposed the Pearson's chi-squared test (X^2) for goodness-of-fit testing for multinomial distribution, given by

$$X^2 = \sum_{i=1}^{m} \sum_{j=1}^{K} \frac{(x_{ij} - n_i \hat{p}_{O,ij})^2}{n_i \hat{p}_{O,ij}},$$

where $\hat{p}_{O,ij}$ is an hypothesized value. However, the assumption that the expected cell counts become asymptotically large, crucial for inference based on the chi-square distribution, is not reasonable for analyzing sparse multinomial distribution. Zelterman (1987)[9] proposed a statistic D^2 for testing goodness-of-fit for sparse multinomial distribution where

$$D^2 = \sum_{i=1}^{m}\sum_{j=1}^{K} \frac{(x_{ij} - n_i\hat{p}_{O,ij})^2 - x_{ij}}{n_i\hat{p}_{O,ij}} = X^2 - \sum_{i=1}^{m}\sum_{j=1}^{K} \frac{x_{ij}}{n_i\hat{p}_{O,ij}}.$$

Johnson (2004)[4] introduced a Bayesian chi-squared test for goodness-of-fit, and we can expand this methodology for testing goodness-of-fit for sparse multinomial distribution. Gelman (2013)[2] proposed a posterior predictive p-value method, which can also be used to compare our proposed approaches.

Bibliography

[1] J. Dong and J. S. Simonoff. A geometric combination estimator for $d-$dimesional ordinal sparse contingency tables. *The Annals of Statistics*, 23:1143–1159, 1995.

[2] A. Gelman. Two simple examples for understanding posterior p-values whose distributions are far from uniform. *Electronic Journal of Statistics*, 7:2595–2602, 2013.

[3] N. L. Hjort. *Bayesian Statistics 5*, chapter Bayesian approaches to non- and semiparametric density estimation, pages 223–254. Oxford - University Press, 1996.

[4] V. E. Johnson. A Bayesian χ^2 test for goodness-of-fit. *Annals of Statistics*, 32(6):2361–2384, 2004.

[5] K. J. Koehler. Goodness of fit test for log-linear models in sparse contingency tables. *Journal of the American Statistical Association*, 81:483–492, 1986.

[6] K. Pearson. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900.

[7] K. Pearson. *On the theory of contingency and its relation to association and normal correlation*. London, Dulau and Co., 1904.

[8] P. Yu and C.A. Shaw. An efficient algorithm for accurate computation of the Dirichlet-multinomial log-likelihood function. *Bioinformatics*, 30:1547–1554, 2014.

[9] D. Zelterman. Goodness-of-fit tests for large sparse multinomial distributions. *Journal of the American Statistical Association*, 82(398):624–629, 1987.

Appendix A

For Chapter 2

Metropolis within Gibbs algorithm

Here we used the acceptance-rejection method to generate values from the above posterior distribution.

1. Choose $\boldsymbol{\beta^*_{00}} = (\boldsymbol{\beta^*_{01}}, \boldsymbol{\beta^*_{02}}, \ldots, \boldsymbol{\beta^*_{0K}})$ where $\boldsymbol{\beta^*_{0j}} = (\beta^*_{00j}, \beta^*_{01j}, \beta^*_{02j}, \beta^*_{03j}, \beta^*_{0lj})^t$ is a l x K matrix of initial values. For the proposal distribution we use a multivariate normal distribution centered at $\boldsymbol{\beta}$.

2. For each iteration t, generate the candidate $\boldsymbol{\beta^*} = (\boldsymbol{\beta^*_1}, \boldsymbol{\beta^*_2}, \ldots, \boldsymbol{\beta^*_K})$ and calculate the acceptance ratio $\alpha = \min\left(1, \frac{\pi(\boldsymbol{\beta^*}|\mathbf{p}, \mathbf{x})}{\pi(\boldsymbol{\beta^*_{tt}}|\mathbf{p}, \mathbf{x})}\right)$ which will be used to decide whether to accept or reject the candidate. Generate a uniform random number u on $[0, 1]$.

   - If $u \leq \alpha$ accept the candidate by setting $\boldsymbol{\beta^*_{(t+1)(t+1)}} = \boldsymbol{\beta^*}$.
   - Else $\boldsymbol{\beta^*_{(t+1)(t+1)}} = \boldsymbol{\beta^*_{tt}}$.

To reduce the computational complexity of the ratio of $\frac{\pi(\boldsymbol{\beta}^*|\mathbf{p},\mathbf{x})}{\pi(\boldsymbol{\beta}^*_{\boldsymbol{tt}}|\mathbf{p},\mathbf{x})}$ is calculated by taking the logarithm of $\pi(\boldsymbol{\beta}|\mathbf{p},\mathbf{x})$

$$\begin{aligned}\log\left(\pi(\boldsymbol{\beta}|\mathbf{p},\mathbf{x})\right) \quad \propto \quad & \sum_{i=1}^{m}\log\left[\Gamma\left(\sum_{j=1}^{K}\exp\left(\beta_{0j}+\sum_{a=1}^{l}\beta_{aj}Y_{ia}\right)\right)\right] \\ & -\sum_{i=1}^{m}\sum_{j=1}^{K}\log\left[\Gamma\left(\exp\left(\beta_{0j}+\sum_{a=1}^{l}\beta_{aj}Y_{ia}\right)\right)\right] \\ & +\sum_{i=1}^{m}\sum_{j=1}^{K}\left[\exp\left(\beta_{0j}+\sum_{a=1}^{l}\beta_{aj}Y_{ia}\right)\right]\log(p_{ij}) \\ & +\sum_{i=1}^{m}\sum_{a=0}^{l}\sum_{j=1}^{K}\frac{\beta_{aj}^2}{2}.\end{aligned}$$

To find a good proposal distribution, tuning the multivariate normal proposal distribution is very important to get a high acceptance rate and good mixing. We considered a vector of zeros for the mean vector and a diagonal matrix with a large value (100) for the variance-covariance matrix.

Appendix B

For Chapter 3

We first derive (3.1). In this case, the posterior distribution of $\mathbf{p}_i$ is

$$P(\mathbf{p}_i|\mathbf{x},\alpha,\mathbf{w}) \;\propto\; \sum_{j=1}^{K} p_{ij}^{\sum_{l=1}^{m} x_{lj}w_{il}+\alpha_j},$$

so that $\mathbf{p}_i$ admits a posterior Dirichlet distribution given by

$$\mathbf{p}_i|\mathbf{x},\alpha,\mathbf{w} \;\sim\; \text{Dirichlet}\left(\sum_{l=1}^{m} x_{l1}w_{il}+\alpha_1,\ldots,\sum_{l=1}^{m} x_{lk}w_{il}+\alpha_k\right).$$

Then

$$\hat{p}_{ij}^{\text{Bayes}} \;=\; \int\cdots\int_{\Delta_K} p_{ij}\frac{\Gamma\left(\sum_{a=1}^{K}\left(\sum_{l=1}^{m} x_{la}w_{il}+\alpha_a\right)\right)}{\prod_{a=1}^{K}\Gamma\left(\sum_{l=1}^{m} x_{la}w_{il}+\alpha_a\right)}\prod_{a=1}^{K} p_{ia}^{\sum_{l=1}^{m} x_{la}w_{il}+\alpha_a-1}dp_{i1}\ldots dp_{jK},$$

where Δ_K denotes the K-dimensional simplex. Solving this K-dimensional integral can be done in closed-form and leads to:

$$\begin{aligned}\hat{p}_{ij}^{\text{Bayes}} &= \frac{\left(\sum_{l=1}^{m} x_{lj} w_{il} + \alpha_j\right)}{\left(\sum_{j=1}^{K}\sum_{l=1}^{m} x_{lj} w_{il} + \sum_{j=1}^{K} \alpha_j\right)} \\ &= \frac{\left(\sum_{l=1}^{m} n_l w_{il} \hat{p}_{lj}^{\text{MLE}} + \alpha_j\right)}{\left(\sum_{l=1}^{m} n_l w_{il} + \sum_{j=1}^{K} \alpha_j\right)},\end{aligned}$$

since $\sum_{j=1}^{K} x_{lj} = n_l$. Now, to obtain the wanted result, it suffices to further breakdown the numerator of the above expression to get

$$\hat{p}_{ij}^{\text{Bayes}} = \frac{n_i w_{ii}}{\left(\sum_{l=1}^{m} n_l w_{il} + \sum_{j=1}^{K} \alpha_j\right)} \hat{p}_{ij}^{\text{MLE}} + \frac{\left(\sum_{l\neq i} n_l w_{il} + \sum_{j=1}^{K} \alpha_j\right)}{\left(\sum_{l=1}^{m} n_l w_{il} + \sum_{j=1}^{K} \alpha_j\right)} \frac{\left(\sum_{\substack{l=1 \\ l\neq i}}^{m} n_l w_{il} \hat{p}_{lj}^{\text{MLE}} + \alpha_j\right)}{\left(\sum_{l\neq i} n_l w_{il} + \sum_{j=1}^{K} \alpha_j\right)}. \tag{B.1}$$

Hence, we indeed have that

$$\hat{p}_{ij}^{\text{Bayes}} = \lambda_i \hat{p}_{lj}^{\text{MLE}} + (1 - \lambda_i) t_{ij},$$

where the shrinkage target t_{ij} and the shrinkage constant λ_i are defined as in (3.2). We now move to the proof of (3.3).

For this, we rewrite $\hat{p}_{ij}^{\text{Bayes}}$ given in (B.1), as

$$\begin{aligned}
\hat{p}_{ij}^{\text{Bayes}} &= \frac{n_i w_{ii}}{\left(\sum_{l=1}^{m} n_l w_{il} + \sum_{j=1}^{K} \alpha_j\right)} \hat{p}_{ij}^{\text{MLE}} + \frac{\alpha_j}{\left(\sum_{l=1}^{m} n_l w_{il} + \sum_{j=1}^{K} \alpha_j\right)} + \frac{\sum_{\substack{l=1 \\ l\neq i}}^{m} n_l w_{il} \hat{p}_{lj}^{\text{MLE}}}{\left(\sum_{l=1}^{m} n_l w_{il} + \sum_{j=1}^{K} \alpha_j\right)} \\
&= \frac{n_i w_{ii}}{\left(\sum_{l=1}^{m} n_l w_{il} + \sum_{j=1}^{K} \alpha_j\right)} \hat{p}_{ij}^{\text{MLE}} + \frac{\sum_{j=1}^{K} \alpha_j}{\left(\sum_{l=1}^{m} n_l w_{il} + \sum_{j=1}^{K} \alpha_j\right)} \left(\frac{\alpha_j}{\sum_{j=1}^{K} \alpha_j}\right) \\
&\quad + \frac{\sum_{\substack{l=1 \\ l\neq i}}^{m} n_l w_{il}}{\left(\sum_{l=1}^{m} n_l w_{il} + \sum_{j=1}^{K} \alpha_j\right)} \frac{\sum_{\substack{l=1 \\ l\neq i}}^{m} n_l w_{il} \hat{p}_{lj}^{\text{MLE}}}{\sum_{\substack{l=1 \\ l\neq i}}^{m} n_l w_{il}} \\
&= \lambda_{1i} \hat{p}_{ij}^{\text{MLE}} + \lambda_{2i} t_j + (1 - \lambda_{1i} - \lambda_{2i}) t_{ij}
\end{aligned}$$

where $t_{ij}, t_j, \lambda_{1i}$ and λ_{2i} are defined as in (3.4).

Appendix C

For Chapter 4

The derivation of $E(P_i)$ in (4.4)

$$\begin{aligned} E(P_i) &= \int \dots \int p_i f(\mathbf{p}|\boldsymbol{\alpha}, \lambda) dp_1 dp_2 \dots dp_K, \\ &= \int \dots \int p_i C(\boldsymbol{\alpha}, \delta) \prod_{j=1}^{K} p_j^{\alpha_j - 1} \exp(-\delta \Delta(\mathbf{p})) dp_1 dp_2 \dots dp_K, \\ &= C(\boldsymbol{\alpha}, \delta) \int \dots \int p_i^{\alpha_i} \prod_{j \neq i} p_j^{\alpha_j - 1} \exp(-\delta \Delta(\mathbf{p})) dp_1 dp_2 \dots dp_K. \end{aligned}$$

Let $\gamma_{il} = 0, \forall l, l \neq i$ and $\gamma_{ii} = 1$. Then

$$\begin{aligned} E(P_i) &= \frac{C(\boldsymbol{\alpha}, \delta)}{C(\boldsymbol{\alpha} + \boldsymbol{\gamma_i}, \delta)} \int \dots \int C(\boldsymbol{\alpha} + \boldsymbol{\gamma_j}, \delta) \prod_{j=1}^{K} p_j^{\alpha_j + \gamma_{ij} - 1} \exp(-\delta \Delta(\mathbf{p})) dp_1 dp_2 \dots dp_K, \\ &= \frac{C(\boldsymbol{\alpha}, \delta)}{C(\boldsymbol{\alpha} + \boldsymbol{\gamma_i}, \delta)}, \end{aligned}$$

where $C(\boldsymbol{\alpha} + \boldsymbol{\gamma_i}, \delta) = \dfrac{1}{\int \dots \int \prod_{j=1}^{K} p_j^{\alpha_j + \gamma_{ij} - 1} \exp(-\delta \Delta(\mathbf{p})) dp_1 dp_2 \dots dp_K}$.

Using (4.3),

$$E(P_i) = \frac{\dfrac{C_2(\boldsymbol{\alpha})}{E_{\boldsymbol{\alpha}}[\exp(-\delta\Delta(\mathbf{P}))]}}{\dfrac{C_2(\boldsymbol{\alpha}+\boldsymbol{\gamma_i})}{E_{\boldsymbol{\alpha}+\boldsymbol{\gamma_i}}[\exp(-\delta\Delta(\mathbf{P}))]}},$$

$$= \frac{C_2(\boldsymbol{\alpha})}{C_2(\boldsymbol{\alpha}+\boldsymbol{\gamma_i})} \times \frac{E_{\boldsymbol{\alpha}+\boldsymbol{\gamma_i}}[\exp(-\delta\Delta(\mathbf{P}))]}{E_{\boldsymbol{\alpha}}[\exp(-\delta\Delta(\mathbf{P}))]}.$$

Here

$$C_2(\boldsymbol{\alpha}+\boldsymbol{\gamma_i}) = \frac{\Gamma(\sum_{j=1}^{K}(\alpha_j+\gamma_{ij}))}{\prod_{j=1}^{K}\Gamma(\alpha_j+\gamma_{ij})},$$

$$= \frac{\Gamma(\sum_{j=1}^{K}\alpha_j+1)}{\prod_{j\neq i}^{K}\Gamma(\alpha_j)\Gamma(\alpha_i+1)},$$

$$= \frac{\Gamma(\sum_{j=1}^{K}\alpha_j)\sum_{j=1}^{K}\alpha_j}{\prod_{j\neq i}^{K}\Gamma(\alpha_j)\Gamma(\alpha_i)\alpha_i},$$

$$= \frac{\sum_{j=1}^{K}\alpha_j}{\alpha_i} \times \frac{\Gamma(\sum_{j=1}^{K}\alpha_j)}{\prod_{j=1}^{K}\Gamma(\alpha_j)},$$

$$= \frac{\sum_{j=1}^{K}\alpha_j}{\alpha_i} C_2(\boldsymbol{\alpha}). \tag{C.1}$$

Substituting in (C.1),

$$E(P_i) = \frac{\alpha_i}{\sum_{j=1}^{K}\alpha_j} \times \frac{E_{\boldsymbol{\alpha}+\boldsymbol{\gamma_i}}[\exp(-\delta\Delta(\mathbf{P}))]}{E_{\boldsymbol{\alpha}}[\exp(-\delta\Delta(\mathbf{P}))]}.$$

Next we derive $E(P_i^n)$ in (4.5). Similarly, we obtain

$$\begin{aligned} E(P_i^n) &= \int \dots \int p_i^n f(\mathbf{p}|\boldsymbol{\alpha}, \delta) dp_1 dp_2 \dots dp_K, \\ &= \int \dots \int p_i^n C(\boldsymbol{\alpha}, \delta) \prod_{j=1}^{K} p_j^{\alpha_j - 1} \exp(-\delta \Delta(\mathbf{p})) dp_1 dp_2 \dots dp_K, \\ &= C(\boldsymbol{\alpha}, \delta) \int \dots \int p_i^{\alpha_i + n - 1} \prod_{j \neq i} p_j^{\alpha_j - 1} \exp(-\delta \Delta(\mathbf{p})) dp_1 dp_2 \dots dp_K. \end{aligned}$$

Then

$$\begin{aligned} E(P_i^n) &= \frac{C(\boldsymbol{\alpha}, \delta)}{C(\boldsymbol{\alpha} + n\boldsymbol{\gamma}_i, \delta)} \int \dots \int C(\boldsymbol{\alpha} + n\boldsymbol{\gamma}_i, \delta) \prod_{j=1}^{K} p_j^{\alpha_j + n\gamma_{ij} - 1} \exp(-\delta \Delta(\mathbf{p})) dp_1 dp_2 \dots dp_K, \\ &= \frac{C(\boldsymbol{\alpha}, \delta)}{C(\boldsymbol{\alpha} + n\boldsymbol{\gamma}_i, \delta)}, \end{aligned}$$

where $C(\boldsymbol{\alpha} + n\boldsymbol{\gamma}_i, \delta) = \dfrac{1}{\int \dots \int \prod_{j=1}^{K} p_j^{\alpha_j + n\gamma_{ij} - 1} \exp(-\delta \Delta(\mathbf{p})) dp_1 dp_2 \dots dp_K}$.

Using (4.3)

$$\begin{aligned} E(P_i^n) &= \frac{\dfrac{C_2(\boldsymbol{\alpha})}{\overline{E_{\boldsymbol{\alpha}}[\exp(-\delta \Delta(\mathbf{P}))]}}}{\dfrac{C_2(\boldsymbol{\alpha} + n\boldsymbol{\gamma}_i)}{\overline{E_{\boldsymbol{\alpha} + n\boldsymbol{\gamma}_j}[\exp(-\delta \Delta(\mathbf{P}))]}}}, \\ &= \frac{C_2(\boldsymbol{\alpha})}{C_2(\boldsymbol{\alpha} + n\boldsymbol{\gamma}_i)} \times \frac{E_{\boldsymbol{\alpha} + n\boldsymbol{\gamma}_i}[\exp(-\delta \Delta(\mathbf{P}))]}{E_{\boldsymbol{\alpha}}[\exp(-\delta \Delta(\mathbf{P}))]}. \end{aligned}$$

Here

$$
\begin{aligned}
C_2(\boldsymbol{\alpha}+n\boldsymbol{\gamma}_i) &= \frac{\Gamma(\sum_{j=1}^{K}(\alpha_j+n\gamma_{ij}))}{\prod_{j=1}^{K}\Gamma(\alpha_j+n\gamma_{ij})},\\
&= \frac{\Gamma(\sum_{j=1}^{K}\alpha_j+n)}{\prod_{j\neq i}^{K}\Gamma(\alpha_j)\Gamma(\alpha_i+n)},\\
&= \frac{\Gamma(\sum_{j=1}^{K}\alpha_j+n)}{\prod_{j\neq i}^{K}\Gamma(\alpha_j)\Gamma(\alpha_i+n)}\times\frac{\Gamma(\alpha_i)}{\Gamma(\alpha_i)}\times\frac{\Gamma(\sum_{j=1}^{K}\alpha_j)}{\Gamma(\sum_{j=1}^{K}\alpha_j)},\\
&= \frac{\Gamma(\sum_{j=1}^{K}\alpha_j+n)}{\Gamma(\sum_{j=1}^{K}\alpha_j)}\times\frac{\Gamma(\alpha_i)}{\Gamma(\alpha_i+n)}\times\frac{\Gamma(\sum_{j=1}^{K}\alpha_j)}{\prod_{j=1}^{K}\Gamma(\alpha_j)},\\
&= \frac{\Gamma(\sum_{j=1}^{K}\alpha_j+n)}{\Gamma(\sum_{j=1}^{K}\alpha_j)}\times\frac{\Gamma(\alpha_i)}{\Gamma(\alpha_i+n)}\times C_2(\boldsymbol{\alpha}). \qquad \text{(C.2)}
\end{aligned}
$$

Substituting in (C.2),

$$
E(P_i^n) = \frac{\Gamma(\sum_{j=1}^{K}\alpha_j)}{\Gamma(\sum_{j=1}^{K}\alpha_j+n)}\times\frac{\Gamma(\alpha_i+n)}{\Gamma(\alpha_i)}\times\frac{E_{\boldsymbol{\alpha}+n\boldsymbol{\gamma}_i}[\exp(-\lambda\Delta(\mathbf{P}))]}{E_{\boldsymbol{\alpha}}[\exp(-\lambda\Delta(\mathbf{P}))]}. \qquad \text{(C.3)}
$$

More generally, moments of smoothed Dirichlet distribution random variables can be expressed as

$$
E\left(\prod_{i=1}^{K}P_i^{n_i}\right) = \frac{\Gamma(\sum_{j=1}^{K}\alpha_j)}{\Gamma(\sum_{j=1}^{K}(\alpha_j+n_j))}\times\left[\prod_{i=1}^{K}\frac{\Gamma(\alpha_i+n_i)}{\Gamma(\alpha_i)}\right]\times\frac{E_{\boldsymbol{\alpha}+\boldsymbol{n}}[\exp(-\lambda\Delta(\mathbf{P}))]}{E_{\boldsymbol{\alpha}}[\exp(-\lambda\Delta(\mathbf{P}))]},
$$

where $\boldsymbol{\alpha}+\boldsymbol{n} = (\alpha_1+n_1, \alpha_2+n_2, \ldots, \alpha_K+n_K)^t$

Now we derive $\mathrm{Var}(P_i)$. From (C.3), when $n = 2$,

$$E(P_i^2) = \frac{\Gamma(\sum_{j=1}^K \alpha_j)}{\Gamma(\sum_{j=1}^K \alpha_j + 2)} \times \frac{\Gamma(\alpha_i + 2)}{\Gamma(\alpha_i)} \times \frac{E_{\boldsymbol{\alpha}+2\boldsymbol{\gamma}_i}[\exp(-\lambda\Delta(\mathbf{P}))]}{E_{\boldsymbol{\alpha}}[\exp(-\lambda\Delta(\mathbf{P}))]},$$

$$= \frac{\alpha_i}{\sum_{j=1}^K \alpha_j} \times \frac{\alpha_i + 1}{\sum_{j=1}^K \alpha_j + 1} \times \frac{E_{\boldsymbol{\alpha}+2\boldsymbol{\gamma}_i}[\exp(-\lambda\Delta(\mathbf{P}))]}{E_{\boldsymbol{\alpha}}[\exp(-\lambda\Delta(\mathbf{P}))]}.$$

Then we can derive

$$\mathrm{Var}(P_i) = E(P_i^2) - (E(P_i))^2,$$

$$= \frac{\alpha_i}{\sum_{j=1}^K \alpha_j} \times \frac{\alpha_i + 1}{\sum_{j=1}^K \alpha_j + 1} \times \frac{E_{\boldsymbol{\alpha}+2\boldsymbol{\gamma}_i}[\exp(-\delta\Delta(\mathbf{P}))]}{E_{\boldsymbol{\alpha}}[\exp(-\delta\Delta(\mathbf{P}))]} - \left(\frac{\alpha_i}{\sum_{j=1}^K \alpha_j} \times \frac{E_{\boldsymbol{\alpha}+\boldsymbol{\gamma}_i}[\exp(-\lambda\Delta(\mathbf{P}))]}{E_{\boldsymbol{\alpha}}[\exp(-\delta\Delta(\mathbf{P}))]} \right)^2,$$

$$= \frac{\alpha_i}{(\sum_{j=1}^K \alpha_j)^2 (\sum_{j=1}^K \alpha_j + 1) E_{\boldsymbol{\alpha}}^2[\exp(-\delta\Delta(\mathbf{P}))]}$$

$$\times \left((\alpha_i + 1) \left(\sum_{j=1}^K \alpha_j \right) E_{\boldsymbol{\alpha}+2\boldsymbol{\gamma}_i}[\exp(-\delta\Delta(\mathbf{P}))] E_{\boldsymbol{\alpha}}^2[\exp(-\delta\Delta(\mathbf{P}))] - \alpha_i \left(\sum_{j=1}^K \alpha_j + 1 \right) E_{\boldsymbol{\alpha}+\boldsymbol{\gamma}_i}^2[\exp(-\delta\Delta(\mathbf{P}))] \right).$$

Next we derive the upper and lower limit for $E(P_i)$ in (4.6). We know that for the penalty function $\Delta(\mathbf{p}) = \sum_{j=1}^{K-1}(p_{j+1} - p_j)^2$, the values are in $[0, 2]$. Then

$$0 \le \Delta(\mathbf{p}) \le 2$$

$$-2\lambda \le -\delta\Delta(\mathbf{p}) \le 0$$

$$\exp(-2\delta) \le \exp(-\delta\Delta(\mathbf{p})) \le 1$$

$$\int \cdots \int \frac{\Gamma(\sum_{j=1}^K \alpha_j)}{\prod_{j=1}^K \Gamma(\alpha_j)} \prod_{j=1}^K p_j^{\alpha_j - 1} \exp(-2\delta) dp_1 dp_2 \ldots dp_K \le \int \cdots \int \frac{\Gamma(\sum_{j=1}^K \alpha_j)}{\prod_{j=1}^K \Gamma(\alpha_j)} \prod_{j=1}^K p_j^{\alpha_j - 1} \exp(-\delta\Delta(\mathbf{p})) dp_1 dp_2 \ldots dp_K$$

$$\le \int \cdots \int \frac{\Gamma(\sum_{j=1}^K \alpha_j)}{\prod_{j=1}^K \Gamma(\alpha_j)} \prod_{j=1}^K p_j^{\alpha_j - 1} dp_1 dp_2 \ldots dp_K,$$

$$\exp(-2\delta) \int \cdots \int \frac{\Gamma(\sum_{j=1}^K \alpha_j)}{\prod_{j=1}^K \Gamma(\alpha_j)} \prod_{j=1}^K p_j^{\alpha_j - 1} dp_1 dp_2 \ldots dp_K \le \int \cdots \int \exp(-\delta\Delta(\mathbf{p})) \frac{\Gamma(\sum_{j=1}^K \alpha_j)}{\prod_{j=1}^K \Gamma(\alpha_j)} \prod_{j=1}^K p_j^{\alpha_j - 1} dp_1 dp_2 \ldots dp_K$$

$$\le \int \cdots \int \frac{\Gamma(\sum_{j=1}^K \alpha_j)}{\prod_{j=1}^K \Gamma(\alpha_j)} \prod_{j=1}^K p_j^{\alpha_j - 1} dp_1 dp_2 \ldots dp_K,$$

$$\exp(-2\delta) \leq E_{\boldsymbol{\alpha}}[\exp(-\delta\Delta(\mathbf{P}))] \leq 1. \tag{C.4}$$

Similarly,

$$\int\cdots\int p_i \frac{\Gamma(\sum_{j=1}^K \alpha_j)}{\prod_{j=1}^K \Gamma(\alpha_j)} \prod_{j=1}^K p_j^{\alpha_j-1} \exp(-2\delta) dp_1 dp_2 \ldots dp_K \leq \int\cdots\int p_i \frac{\Gamma(\sum_{j=1}^K \alpha_j)}{\prod_{j=1}^K \Gamma(\alpha_j)} \prod_{j=1}^K p_j^{\alpha_j-1} \exp(-\delta\Delta(\mathbf{p})) dp_1 dp_2 \ldots dp_K$$

$$\leq \int\cdots\int p_i \frac{\Gamma(\sum_{j=1}^K \alpha_j)}{\prod_{j=1}^K \Gamma(\alpha_j)} \prod_{j=1}^K p_j^{\alpha_j-1} dp_1 dp_2 \ldots dp_K,$$

$$\exp(-2\delta) \int\cdots\int p_i \frac{\Gamma(\sum_{j=1}^K \alpha_j)}{\prod_{j=1}^K \Gamma(\alpha_j)} \prod_{j=1}^K p_j^{\alpha_j-1} dp_1 dp_2 \ldots dp_K$$

$$\leq \frac{\Gamma(\sum_{j=1}^K \alpha_j)}{\prod_{j=1}^K \Gamma(\alpha_j)} \frac{\prod_{j=1}^K \Gamma(\alpha_j+\gamma_{ij})}{\Gamma(\sum_{j=1}^K \alpha_j+\gamma_{ij})} \int\cdots\int \exp(-\delta\Delta(\mathbf{p})) \frac{\Gamma(\sum_{j=1}^K \alpha_j+\gamma_{ij})}{\prod_{j=1}^K \Gamma(\alpha_j+\gamma_{ij})} \prod_{j=1}^K p_j^{\alpha_j+\gamma_{ji}-1} dp_1 dp_2 \ldots dp_K$$

$$\leq \int\cdots\int p_i \frac{\Gamma(\sum_{j=1}^K \alpha_j)}{\prod_{j=1}^K \Gamma(\alpha_j)} \prod_{j=1}^K p_j^{\alpha_j-1} dp_1 dp_2 \ldots dp_K,$$

$$\exp(-2\delta) E_{\boldsymbol{\alpha}}(P_i) \leq \frac{\Gamma(\sum_{j=1}^K \alpha_j)}{\prod_{j=1}^K \Gamma(\alpha_j)} \frac{\prod_{j\neq i} \Gamma(\alpha_j)\Gamma(\alpha_i+1)}{\Gamma(\sum_{j=1}^K \alpha_j+1)} E_{\boldsymbol{\alpha}+\boldsymbol{\gamma}_i}[\exp(-\delta\Delta(\mathbf{P}))] \leq E_{\boldsymbol{\alpha}}(P_i),$$

$$\exp(-2\delta) E_{\boldsymbol{\alpha}}(P_i) \leq \frac{\alpha_i}{\sum_{j=1}^K \alpha_j} E_{\boldsymbol{\alpha}+\boldsymbol{\gamma}_i}[\exp(-\delta\Delta(\mathbf{P}))] \leq E_{\boldsymbol{\alpha}}(P_i). \tag{C.5}$$

By using (C.4) and (C.5),

$$\exp(-2\delta) E_{\boldsymbol{\alpha}}(P_i) \leq \left(\frac{\alpha_i}{\sum_{j=1}^K \alpha_j}\right) \frac{E_{\boldsymbol{\alpha}+\boldsymbol{\gamma}_i}[\exp(-\delta\Delta(\mathbf{P}))]}{E_{\boldsymbol{\alpha}}[\exp(-\delta\Delta(\mathbf{P}))]} \leq \frac{1}{\exp(-2\delta)} E_{\boldsymbol{\alpha}}(P_i).$$

We know that $E_{\boldsymbol{\alpha}}(P_i) = \dfrac{\alpha_i}{\sum_{j=1}^K \alpha_j}$. Then,

$$\exp(-2\delta) \left(\frac{\alpha_i}{\sum_{j=1}^K \alpha_j}\right) \leq \left(\frac{\alpha_i}{\sum_{j=1}^K \alpha_j}\right) \frac{E_{\boldsymbol{\alpha}+\boldsymbol{\gamma}_i}[\exp(-\delta\Delta(\mathbf{P}))]}{E_{\boldsymbol{\alpha}}[\exp(-\delta\Delta(\mathbf{P}))]} \leq \frac{1}{\exp(-2\delta)} \left(\frac{\alpha_i}{\sum_{j=1}^K \alpha_j}\right),$$

$$\exp(-2\delta) \left(\frac{\alpha_i}{\sum_{j=1}^K \alpha_j}\right) \leq E(P_i) \leq \exp(2\delta) \left(\frac{\alpha_i}{\sum_{j=1}^K \alpha_j}\right).$$

Note that when $\delta \to 0$, $E(P_i) \to \dfrac{\alpha_i}{\sum_{j=1}^{K} \alpha_j}$ which is the expected value of the Dirichlet distribution. Then we derive the upper and lower limit for $\mathrm{Var}(P_i)$ (4.7 and 4.8). We know that

$$\exp(x) = \sum_{k=0}^{\infty} \frac{x^k}{k!}.$$

This is also called the Maclaurin series. Then,

$$\begin{aligned}
\exp(-2\delta) &= 1 + \frac{(-2\delta)^1}{1!} + \frac{(-2\delta)^2}{2!} + \dots \\
&= 1 - 2\delta + 2\delta^2 - \frac{4}{3}\delta^3 + \dots \\
&= 1 - 2\delta + O(\delta^2).
\end{aligned}$$

Similarly, $\exp(2\delta) = 1 + 2\delta + O(\delta^2)$.

Then,

$$(1 - 2\delta)\left(\frac{\alpha_i}{\sum_{j=1}^{K} \alpha_j}\right) \leq E(P_i) \leq (1 + 2\delta)\left(\frac{\alpha_i}{\sum_{j=1}^{K} \alpha_j}\right).$$

Also,

$$\int \dots \int p_i^2 \frac{\Gamma(\sum_{j=1}^{K} \alpha_j)}{\prod_{j=1}^{K} \Gamma(\alpha_j)} \prod_{j=1}^{K} p_j^{\alpha_j - 1} \exp(-2\delta) dp_1 dp_2 \dots dp_K \leq \int \dots \int p_i^2 \frac{\Gamma(\sum_{j=1}^{K} \alpha_j)}{\prod_{j=1}^{K} \Gamma(\alpha_j)} \prod_{j=1}^{K} p_j^{\alpha_j - 1} \exp(-\delta\Delta(\mathbf{x})) dp_1 dp_2 \dots dp_K$$

$$\leq \int \dots \int p_i^2 \frac{\Gamma(\sum_{j=1}^{K} \alpha_j)}{\prod_{j=1}^{K} \Gamma(\alpha_j)} \prod_{j=1}^{K} p_j^{\alpha_j - 1} dp_1 dp_2 \dots dp_K,$$

$$\exp(-2\delta) \int \dots \int p_i^2 \frac{\Gamma(\sum_{j=1}^{K} \alpha_j)}{\prod_{j=1}^{K} \Gamma(\alpha_j)} \prod_{j=1}^{K} p_j^{\alpha_j - 1} dp_1 dp_2 \dots dp_K$$

$$\leq \frac{\Gamma(\sum_{j=1}^{K} \alpha_j)}{\prod_{j=1}^{K} \Gamma(\alpha_j)} \frac{\prod_{j=1}^{K} \Gamma(\alpha_j + 2\gamma_{ij})}{\Gamma(\sum_{j=1}^{K} \alpha_j + 2\gamma_{ij})} \int \dots \int \exp(-\delta\Delta(\mathbf{x})) \frac{\Gamma(\sum_{j=1}^{K} \alpha_j + 2\gamma_{ij})}{\prod_{j=1}^{K} \Gamma(\alpha_j + 2\gamma_{ij})} \prod_{j=1}^{K} p_j^{\alpha_j + 2\gamma_{ji} - 1} dp_1 dp_2 \dots dp_K$$

$$\leq \int \dots \int p_i^2 \frac{\Gamma(\sum_{j=1}^{K} \alpha_j)}{\prod_{j=1}^{K} \Gamma(\alpha_j)} \prod_{j=1}^{K} p_j^{\alpha_j - 1} dp_1 dp_2 \dots dp_K,$$

$$\exp(-2\delta)E_{\boldsymbol{\alpha}}(P_i^2) \leq \frac{\Gamma(\sum_{j=1}^K \alpha_j)}{\prod_{j=1}^K \Gamma(\alpha_j)} \frac{\prod_{j\neq i}\Gamma(\alpha_j)\Gamma(\alpha_i+2)}{\Gamma(\sum_{j=1}^K \alpha_j+2)} E_{\boldsymbol{\alpha}+2\boldsymbol{\gamma}_i}[\exp(-\delta\Delta(\mathbf{P}))] \leq E_{\boldsymbol{\alpha}}(P_i^2),$$

$$\exp(-2\delta)E_{\boldsymbol{\alpha}}(P_i^2) \leq \left(\frac{\alpha_i}{\sum_{j=1}^K \alpha_j}\right)\left(\frac{\alpha_i+1}{\sum_{j=1}^K \alpha_j+1}\right) E_{\boldsymbol{\alpha}+2\boldsymbol{\gamma}_i}[\exp(-\delta\Delta(\mathbf{P}))] \leq E_{\boldsymbol{\alpha}}(P_i^2). \tag{C.6}$$

By using (C.4) and (C.6),

$$\exp(-2\delta)E_{\boldsymbol{\alpha}}(P_i^2) \leq \left(\frac{\alpha_i}{\sum_{j=1}^K \alpha_j}\right)\left(\frac{\alpha_i+1}{\sum_{j=1}^K \alpha_j+1}\right) \frac{E_{\boldsymbol{\alpha}+2\boldsymbol{\gamma}_i}[\exp(-\delta\Delta(\mathbf{p}))]}{E_{\boldsymbol{\alpha}}[\exp(-\delta\Delta(\mathbf{P}))]} \leq \frac{1}{\exp(-2\delta)} E_{\boldsymbol{\alpha}}(P_i^2). \tag{C.7}$$

Also,

$$\begin{aligned} E_{\boldsymbol{\alpha}}(P_i^2) &= \int\cdots\int p_i^2 \frac{\Gamma(\sum_{j=1}^K \alpha_j)}{\prod_{j=1}^K \Gamma(\alpha_j)} \prod_{j=1}^K p_j^{\alpha_j-1} dp_1 dp_2 \ldots dp_K, \\ &= \frac{\Gamma(\sum_{j=1}^K \alpha_j)}{\prod_{j=1}^K \Gamma(\alpha_j)} \frac{\prod_{j=1}^K \Gamma(\alpha_j+2\gamma_{ij})}{\Gamma(\sum_{j=1}^K \alpha_j^+ 2\gamma_{ij})} \int\cdots\int \frac{\Gamma(\sum_{j=1}^K \alpha_j+2\gamma_{ij})}{\prod_{j=1}^K \Gamma(\alpha_j+2\gamma_{ij})} \prod_{j=1}^K p_j^{\alpha_j+2\gamma_{ji}-1} dp_1 dp_2 \ldots dp_K, \\ &= \left(\frac{\alpha_i}{\sum_{j=1}^K \alpha_j}\right)\left(\frac{\alpha_i+1}{\sum_{j=1}^K \alpha_j+1}\right). \end{aligned} \tag{C.8}$$

Substituting (C.7) to (C.8),

$$\begin{aligned} \exp(-2\delta)\left(\frac{\alpha_i}{\sum_{j=1}^K \alpha_j}\right)\left(\frac{\alpha_i+1}{\sum_{j=1}^K \alpha_j+1}\right) &\leq \left(\frac{\alpha_i}{\sum_{j=1}^K \alpha_j}\right)\left(\frac{\alpha_i+1}{\sum_{j=1}^K \alpha_j+1}\right) \frac{E_{\boldsymbol{\alpha}+2\boldsymbol{\gamma}_i}[\exp(-\delta\Delta(\mathbf{P}))]}{E_{\boldsymbol{\alpha}}[\exp(-\lambda\Delta(\mathbf{P}))]} \\ &\leq \exp(2\delta)\left(\frac{\alpha_i}{\sum_{j=1}^K \alpha_j}\right)\left(\frac{\alpha_i+1}{\sum_{j=1}^K \alpha_j+1}\right), \\ (1-2\delta)\left(\frac{\alpha_i}{\sum_{j=1}^K \alpha_j}\right)\left(\frac{\alpha_i+1}{\sum_{j=1}^K \alpha_j+1}\right) &\leq \left(\frac{\alpha_i}{\sum_{j=1}^K \alpha_j}\right)\left(\frac{\alpha_i+1}{\sum_{j=1}^K \alpha_j+1}\right) \frac{E_{\boldsymbol{\alpha}+2\boldsymbol{\gamma}_i}[\exp(-\delta\Delta(\mathbf{P}))]}{E_{\boldsymbol{\alpha}}[\exp(-\delta\Delta(\mathbf{P}))]} \\ &\leq (1+2\delta)\left(\frac{\alpha_i}{\sum_{j=1}^K \alpha_j}\right)\left(\frac{\alpha_i+1}{\sum_{j=1}^K \alpha_j+1}\right). \end{aligned}$$

We know that, $\text{Var}(P_i) = E(P_i^2) - (E(P_i))^2$. We will calculate the lower limit of $\text{Var}(P_i)(\text{Var}_{LL}(P_i))$ and the upper limit of $\text{Var}(P_i)(\text{Var}_{UL}(P_i))$ separately.

$$\text{Var}_{LL}(P_i) = (1-2\delta)\left(\frac{\alpha_i}{\sum_{j=1}^K \alpha_j}\right)\left(\frac{\alpha_i+1}{\sum_{j=1}^K \alpha_j+1}\right) - \left[(1+2\lambda)\left(\frac{\alpha_i}{\sum_{j=1}^K \alpha_j}\right)\right]^2,$$

$$= \left(\frac{\alpha_i}{\left(\sum_{j=1}^K \alpha_j\right)^2\left(\sum_{j=1}^K \alpha_j+1\right)}\right)\left[(1-2\delta)(\alpha_i+1)\left(\sum_{j=1}^K \alpha_j\right) - \alpha_i(1+2\delta)^2\left(\sum_{j=1}^K \alpha_j+1\right)\right],$$

$$= \left(\frac{\alpha_i}{\left(\sum_{j=1}^K \alpha_j\right)^2\left(\sum_{j=1}^K \alpha_j+1\right)}\right)\left[\left(\sum_{j=1}^K \alpha_j - \alpha_i\right) - 2\delta\left(2\delta\alpha_i + \sum_{j=1}^K \alpha_j + 3\alpha_i\sum_{j=1}^K \alpha_j + 2\delta\alpha_i\sum_{j=1}^K \alpha_j + 2\alpha_i\right)\right],$$

$$= \left(\frac{\alpha_i\left(\sum_{j=1}^K \alpha_j - \alpha_i\right)}{\left(\sum_{j=1}^K \alpha_j\right)^2\left(\sum_{j=1}^K \alpha_j+1\right)}\right)$$

$$- \left(\frac{2\delta\alpha_i}{\left(\sum_{j=1}^K \alpha_j\right)^2\left(\sum_{j=1}^K \alpha_j+1\right)}\right)\left(2\delta\alpha_i\left(\sum_{j=1}^K \alpha_j+1\right) + (1+\alpha_i)\sum_{j=1}^K \alpha_j + 2\alpha_i\left(\sum_{j=1}^K \alpha_j+1\right)\right).$$

$$\text{Var}_{UL}(P_i) = (1+2\delta)\left(\frac{\alpha_i}{\sum_{j=1}^K \alpha_j}\right)\left(\frac{\alpha_i+1}{\sum_{j=1}^K \alpha_j+1}\right) - \left[(1-2\delta)\left(\frac{\alpha_i}{\sum_{j=1}^K \alpha_j}\right)\right]^2,$$

$$= \left(\frac{\alpha_i}{\left(\sum_{j=1}^K \alpha_j\right)^2\left(\sum_{j=1}^K \alpha_j+1\right)}\right)\left[(1+2\delta)(\alpha_i+1)\left(\sum_{j=1}^K \alpha_j\right) - \alpha_i(1-2\delta)^2\left(\sum_{j=1}^K \alpha_j+1\right)\right],$$

$$= \left(\frac{\alpha_i}{\left(\sum_{j=1}^K \alpha_j\right)^2\left(\sum_{j=1}^K \alpha_j+1\right)}\right)\left[\left(\sum_{j=1}^K \alpha_j - \alpha_i\right) - 2\delta\left(2\delta\alpha_i - \sum_{j=1}^K \alpha_j - 3\alpha_i\sum_{j=1}^K \alpha_j + 2\delta\alpha_i\sum_{j=1}^K \alpha_j - 2\alpha_i\right)\right],$$

$$= \left(\frac{\alpha_i\left(\sum_{j=1}^K \alpha_j - \alpha_i\right)}{\left(\sum_{j=1}^K \alpha_j\right)^2\left(\sum_{j=1}^K \alpha_j+1\right)}\right)$$

$$- \left(\frac{2\delta\alpha_i}{\left(\sum_{j=1}^K \alpha_j\right)^2\left(\sum_{j=1}^K \alpha_j+1\right)}\right)\left(2\delta\alpha_i\left(\sum_{j=1}^K \alpha_j+1\right) - (1+\alpha_i)\sum_{j=1}^K \alpha_j - 2\alpha_i\left(\sum_{j=1}^K \alpha_j+1\right)\right).$$

Note that when $\delta \to 0$, both $\text{Var}_{LL}(P_i)$ and $\text{Var}_{UL}(P_i) \to \dfrac{\alpha_i\left(\sum_{j=1}^K \alpha_j - \alpha_i\right)}{(\sum_{j=1}^K \alpha_j)^2(\sum_{j=1}^K \alpha_j+1)}$

which is the variance of the Dirichlet distribution.

Approaches to approximate the Bayes estimator

Acceptance-rejection method

To generate random sample from the smoothed Dirichlet distribution with parameters $\boldsymbol{\alpha}+\boldsymbol{x}_i, \delta$, and Δ, we use the acceptance-rejection method. In the acceptance-rejection sampling method, we generate sampling values from a target distribution $\boldsymbol{Z}$ by using a proposal distribution $\boldsymbol{Y}$. We generate the values from $\boldsymbol{Y}$ instead of $\boldsymbol{Z}$ and accept the values of $\boldsymbol{Y}$ if $f(\boldsymbol{y}) \leq Cg(\boldsymbol{y})$ where f and g are probability density functions of $\boldsymbol{Z}$ and $\boldsymbol{Y}$, and C is a constant. We consider the Dirichlet distribution as the target distribution. Here are the step of the acceptance-rejection algorithm.

1. Generate a set of random samples $\boldsymbol{y}_{i1}, \boldsymbol{y}_{i2}, \ldots, \boldsymbol{y}_{iM}$ from the Dirichlet distribution with given $\boldsymbol{\alpha} + \boldsymbol{x}_i$.

2. Compute $E_{\boldsymbol{\alpha}+\boldsymbol{x}_i}[\exp(-\delta\Delta(\mathbf{y}))]$ using all the generated random samples.

3. Generate a random number u from Uniform(0,1), set $k = 1$ and $t = 0$.

4. If $u < C = \dfrac{\exp(-\delta\Delta(\mathbf{y}_{ik}))}{E_{\boldsymbol{\alpha}+\boldsymbol{x}_i}[\exp(-\delta\Delta(\mathbf{y}_{ik}))]}$ accept $\boldsymbol{y}_{ik}$, set $t = t + 1$ and $\boldsymbol{z}_i^t = \boldsymbol{y}_{ik}$ and otherwise reject $\boldsymbol{y}_{ik}$, set $k = k + 1$ and repeat the steps again until $k = M$.

Then

$$\hat{p}_{ij}^{\text{Bayes}} = \frac{1}{t}\sum_{l=1}^{t} z_{ij}^{l}.$$

This approach will work since we use the Dirichlet distribution as the proposal, and the resulting approximate empirical Bayes estimates will sum to units ($\sum_{j=1}^{K} \hat{p}_{ij}^{\text{Bayes}} =$

1). However, there is very low acceptance rate of the Dirichlet proposals for moderate and large δ values. One may use the adaptive rejection sampling method to increase the acceptance rate for moderate and large δ values.

Importance sampling method

Let $I_1 = E_{\boldsymbol{\alpha}+\boldsymbol{x_i}+\boldsymbol{\gamma_i}}[\exp(-\delta\Delta(\mathbf{p}_i))]$ and $I_0 = E_{\boldsymbol{\alpha}+\boldsymbol{x_i}}[\exp(-\delta\Delta(\mathbf{p}_i))]$. We want to compute

$$\hat{c}_{ij}(\delta) = \frac{E_{\hat{\boldsymbol{\alpha}}+\boldsymbol{x_i}+\boldsymbol{\gamma_i}}[\exp(-\delta\Delta(\mathbf{p}_i))]}{E_{\hat{\boldsymbol{\alpha}}+\boldsymbol{x_i}}[\exp(-\delta\Delta(\mathbf{p}_i))]} = \frac{\hat{I}_1}{\hat{I}_0}.$$

We can re-write $\hat{I}_1$ as $\frac{\sum_{j=1}^{K}\hat{\alpha}_{ij}+n_i}{\hat{\alpha}_{ij}+x_{ij}} E_{\hat{\boldsymbol{\alpha}}+\boldsymbol{x}_i}[p_{ij}\exp(-\delta\Delta(\mathbf{P}_i))]$. Then

$$\hat{c}_{ij}(\delta) = \left(\frac{\sum_{j=1}^{K}\hat{\alpha}_{ij}+n_i}{\hat{\alpha}_{ij}+x_{ij}}\right) \times \frac{E_{\hat{\boldsymbol{\alpha}}+\boldsymbol{x}_i}[p_{ij}\exp(-\delta\Delta(\mathbf{P}_i))]}{E_{\hat{\boldsymbol{\alpha}}+\boldsymbol{x}_i}[\exp(-\delta\Delta(\mathbf{P}_i))]}$$

$$= \left(\frac{\sum_{j=1}^{K}\hat{\alpha}_{ij}+n_i}{\hat{\alpha}_{ij}+x_{ij}}\right) \times \frac{\int h_1(\mathbf{p}_i)g(\mathbf{p}_i;\hat{\boldsymbol{\alpha}}+\boldsymbol{x}_i)d\mathbf{p}_i}{\int h_0(\mathbf{p}_i)g(\mathbf{p}_i;\hat{\boldsymbol{\alpha}}+\boldsymbol{x}_i)d\mathbf{p}_i}$$

where $h_1(\mathbf{p}_i) = p_{ij}\exp(-\delta\Delta(\mathbf{p}_i))$, $h_0(\mathbf{p}_i) = \exp(-\delta\Delta(\mathbf{p}_i))$ and $g(\mathbf{p}_i;\hat{\boldsymbol{\alpha}}+\boldsymbol{x}_i)$ is the density of the dirichlet distribution with the parameter vector $\hat{\boldsymbol{\alpha}}+\boldsymbol{x}_i$. We compute an approximation for $\hat{c}_{ij}(\delta)$ using Monte Carlo integration based on importance

sampling method and it is given below.

$$\hat{c}_{ij}(\delta) \approx \left(\frac{\sum_{j=1}^{K} \hat{\alpha}_{ij} + n_i}{\hat{\alpha}_{ij} + x_{ij}} \right) \times \frac{\frac{1}{n}\sum_{l=1}^{n} p_{ijl}\exp(-\delta\Delta(\mathbf{p}_{il}))}{\frac{1}{n}\sum_{l=1}^{n}\exp(-\delta\Delta(\mathbf{p}_{il}))}$$

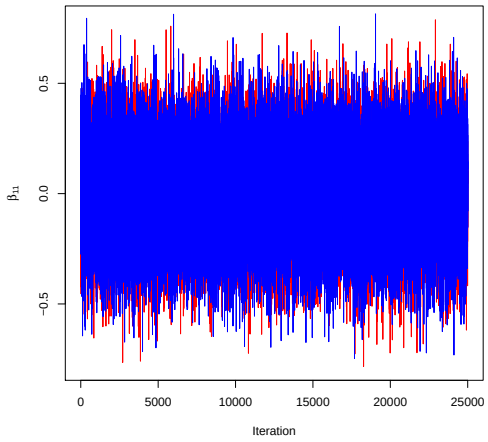
where the random variables $\mathbf{p}_{i1}, \mathbf{p}_{i2}, \ldots, \mathbf{p}_{in}$ are generated from the distribution with the density $g(\mathbf{p}_i; \hat{\boldsymbol{\alpha}} + \boldsymbol{x}_i)$. The estimator for $c_{ij}(\delta)$ is

$$\hat{c}_{ij}(\delta) = \left(\frac{\sum_{j=1}^{K} \hat{\alpha}_{ij} + n_i}{\hat{\alpha}_{ij} + x_{ij}} \right) \left(\frac{\sum_{l=1}^{n} p_{ijl}\exp(-\delta\Delta(\mathbf{p}_{il}))}{\sum_{l=1}^{n}\exp(-\delta\Delta(\mathbf{P}_{il}))} \right).$$

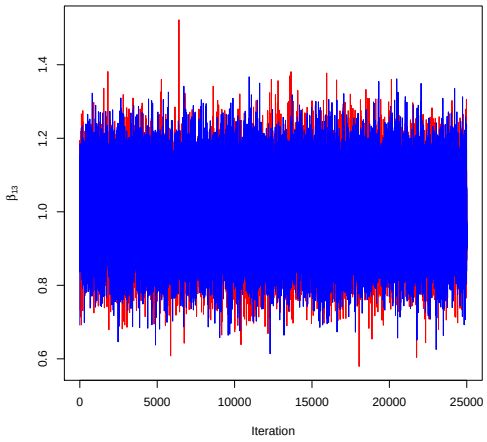
Assessing for convergence

The trace plots are often used to informally assess convergence. In general, we look for a trace plot with good mixing. If the points are randomly scattered and bounced from one point to another quickly, indicates good mixing. The good mixing suggests that the model has converged to a target posterior distribution very quickly. Figure C.1 shows the trace plots with good mixing. Two chains were run (red and blue) to check whether they are converging to a similar posterior distribution. Each chain started with different initial values and they mixed with one another very well.

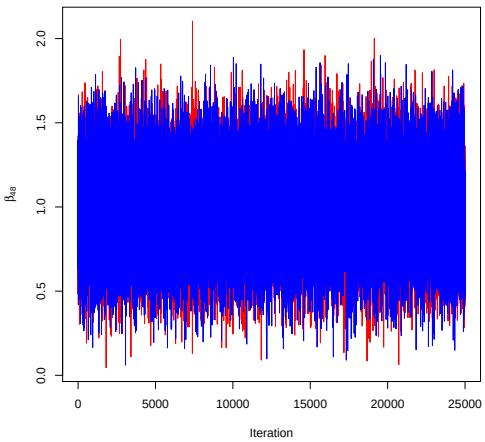
Figure C.1: Trace plots



(a) β_{11}



(b) β_{13}



(c) β_{48}

Table C.1: COVID-19 positive cases by the health regions and the age groups

Province	Health Region	< 20	20-29	30-39	40-49	50-59	60-69	70-79	80+
AB	Calgary Zone	749	717	1029	1147	762	390	180	255
	Central Zone	7	18	12	15	17	13	4	3
	Edmonton Zone	128	178	175	133	132	92	44	32
	North Zone	42	43	44	35	35	31	15	43
	South Zone	206	181	327	341	156	47	19	12
BC	Fraser	77	210	256	223	280	151	131	163
	Interior	4	18	47	43	37	31	17	2
	Northern	3	7	19	10	13	9	4	0
	Vancouver Coastal	21	72	142	129	184	129	86	197
	Vancouver Island	7	16	22	19	20	16	23	8
MB	Interlake-Eastern	1	5	1	1	3	6	3	0
	Northern	0	1	1	0	1	0	0	0
	Prairie Mountain	2	6	9	2	2	4	1	0
	Southern Health	5	3	4	7	9	8	2	0
	Winnipeg	10	52	48	37	32	29	13	8
ON	Algoma	1	3	1	4	5	6	3	1
	Brant County	9	24	15	25	25	13	10	9
	Chatham-Kent	4	35	49	38	14	13	2	2
	Durham	70	171	186	206	309	194	140	397
	Eastern Ontario	7	14	19	19	34	34	15	22
	Grey Bruce	7	11	12	9	20	22	6	15
	Haldimand-Norfolk	11	56	87	91	58	35	16	71
	Haliburton, Kawartha, Pine Ridge	7	13	19	19	30	34	25	46
	Halton	42	84	105	118	146	96	43	96
	Hamilton	58	158	95	99	117	93	69	113
	Hastings and Prince Edward	1	6	3	6	8	4	7	9
	Huron Perth	1	6	9	4	10	16	4	8
	Kingston, Frontenac and Lennox	5	10	8	8	16	9	8	0
	Lambton	14	40	27	25	46	30	32	71
	Leeds, Grenville and Lanark	12	22	25	28	29	40	54	143
	Middlesex-London	27	122	69	66	102	70	52	102
	Niagara	33	111	94	100	89	79	60	174
	North Bay Parry Sound	2	3	2	3	6	9	1	2
	Northwestern	3	7	9	5	7	4	0	1
	Ottawa	94	257	263	274	293	237	181	466
	Peel	363	1064	785	842	945	640	356	578
	Peterborough	5	18	11	13	15	14	8	11
	Porcupine	4	2	11	7	10	17	9	7
	Region of Waterloo	43	188	182	183	201	131	98	218
	Renfrew	0	3	6	3	5	2	4	4
	Simcoe Muskoka	40	93	88	78	88	74	58	58
	Southwestern	3	8	12	10	23	12	10	3
	Sudbury	4	14	5	6	18	7	11	2
	Thunder Bay	9	16	12	21	16	9	5	2
	Timiskaming	2	6	2	0	3	2	2	1
	Toronto	522	1699	1781	1838	2081	1436	974	2473
	Wellington-Dufferin-Guelph	24	58	60	81	68	65	50	60
	Windsor-Essex	38	297	294	226	178	103	67	154
	York	117	381	342	390	526	371	241	492
QC	Quebec	4148	7013	7059	8339	8052	4940	4275	11079
NB	New Brunswick	10	28	23	18	30	27	16	13
NS	Nova Scotia	106	276		276		244		159
SK	Saskatchewan	107	260		241		130		21
NL	Newfoundland	22	38		39	58	57	47	
PE	Prince Edward Island	0	10		8		9		0

Figure C.2: Number of positive cases among 1000 tests

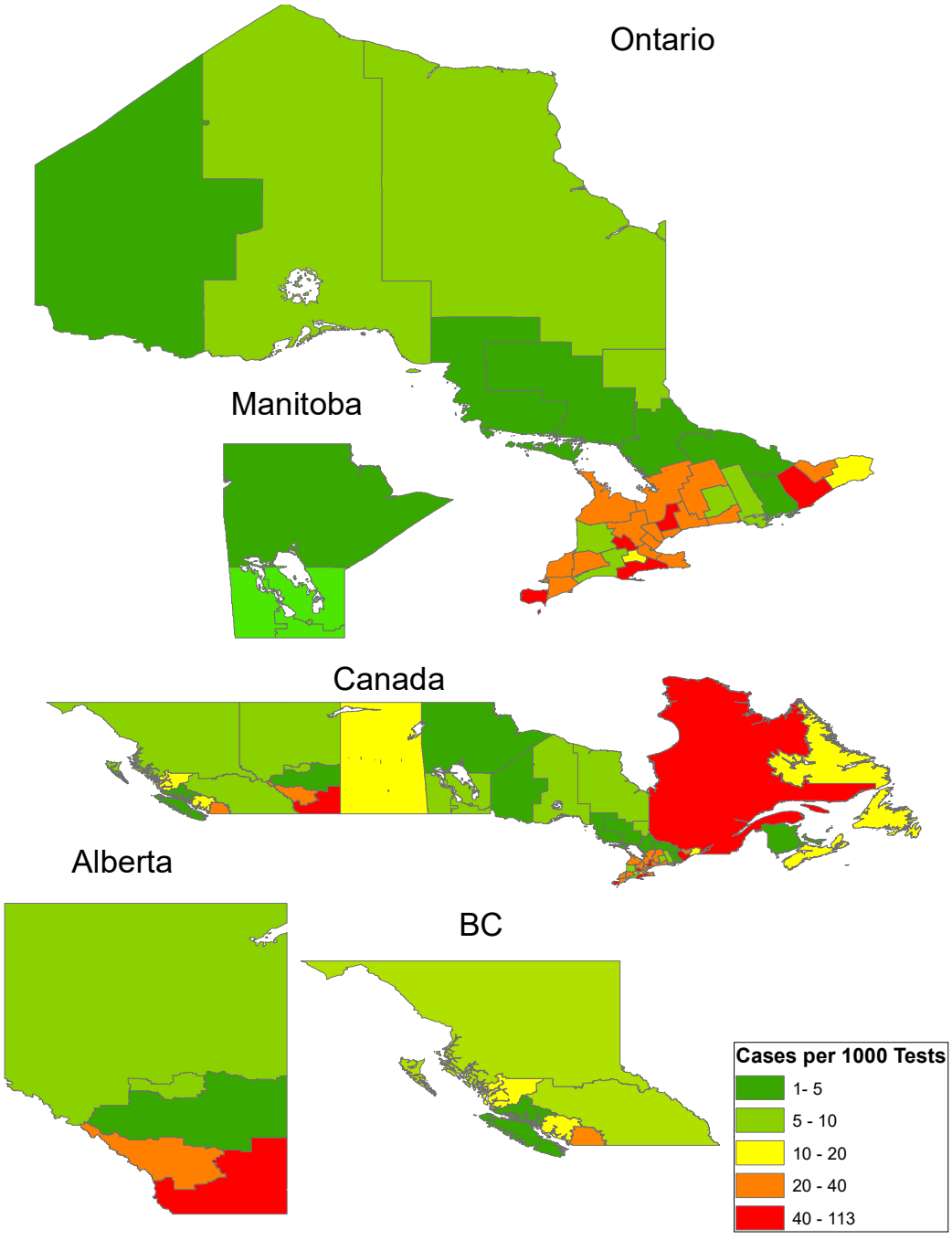


Figure C.3: Number of tests perfomed per 1000 people

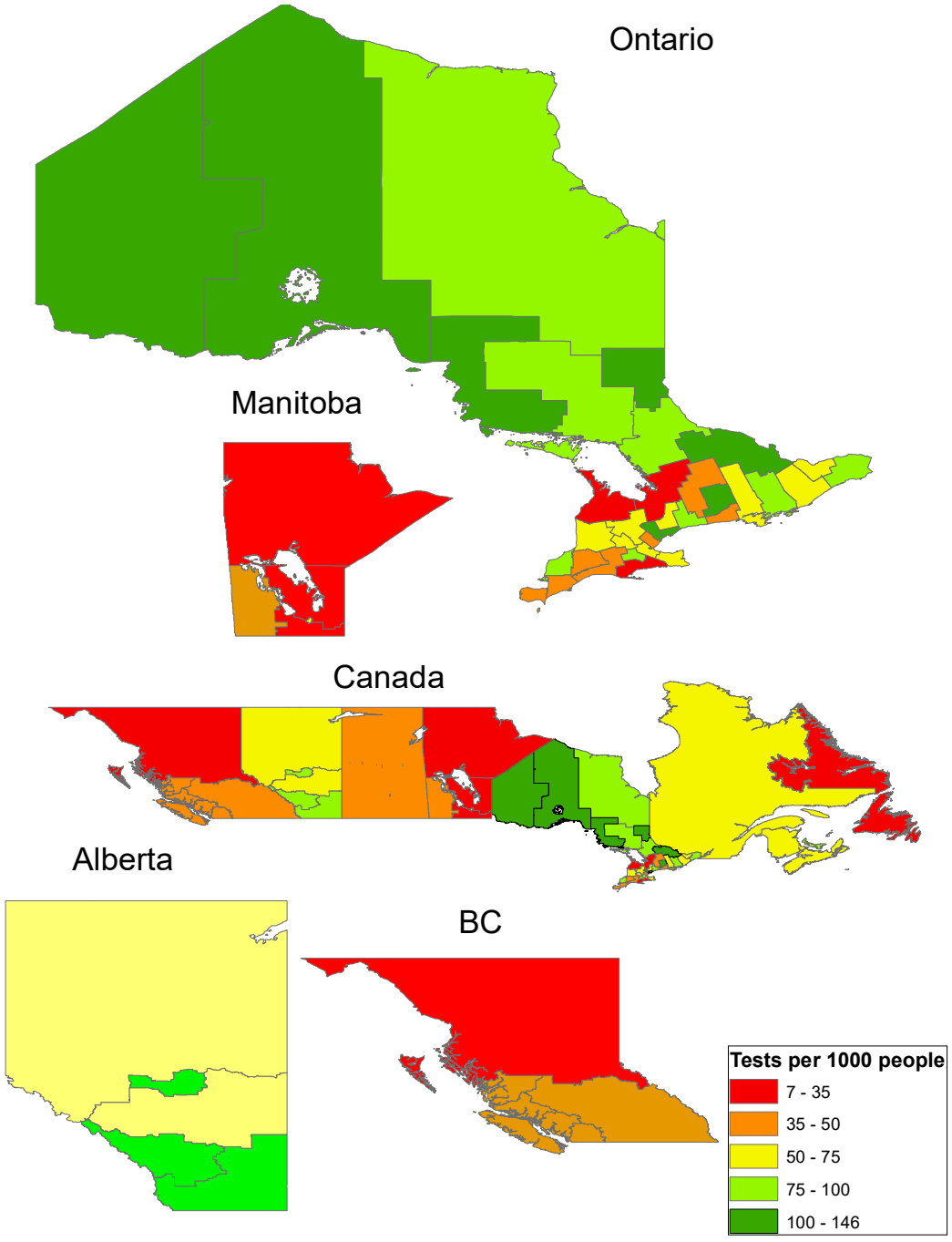


Figure C.4: Population per square km (Population Desnity)

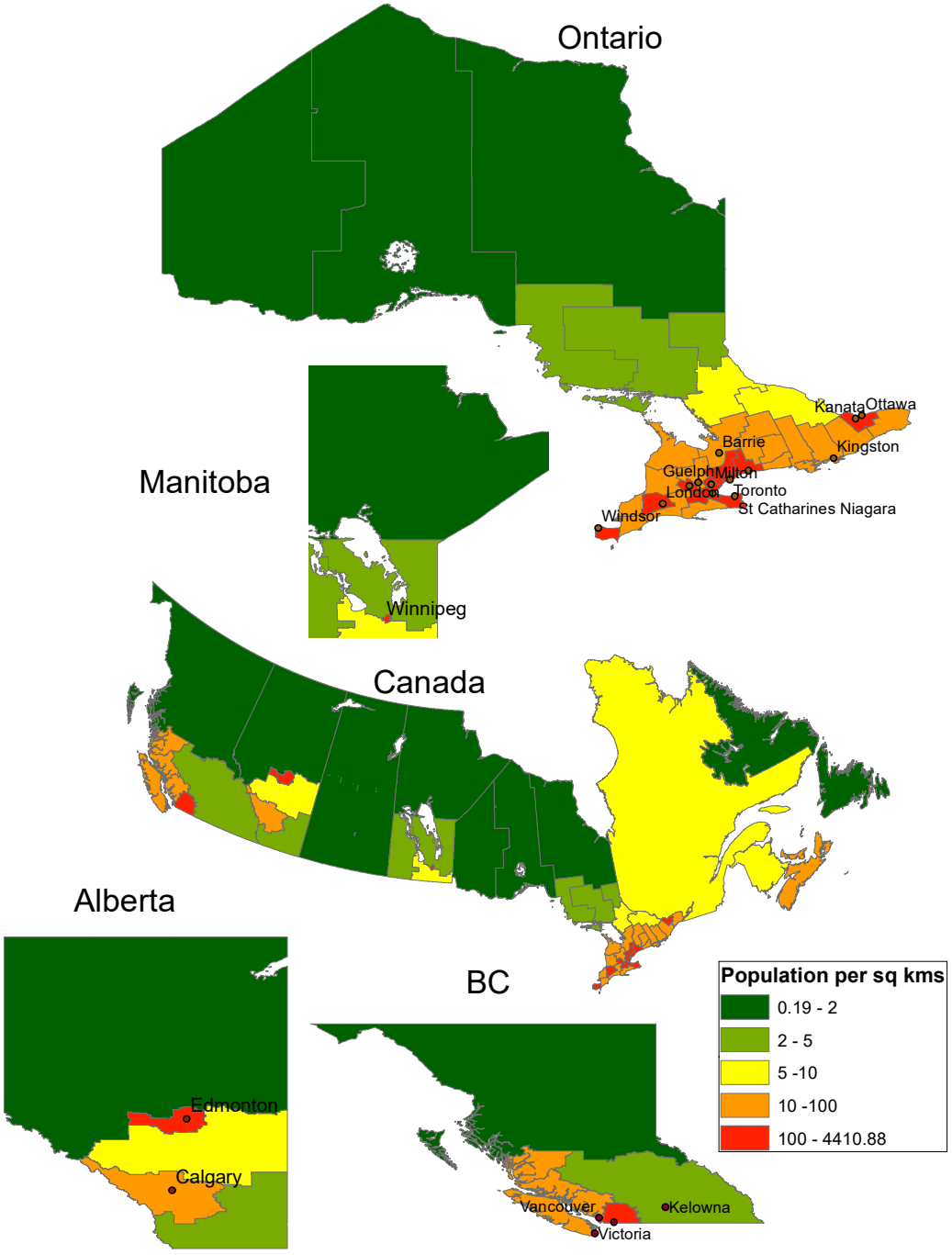


Figure C.5: Major cities in Canada