Advances in Survey Methodology and Sports Science

by

Kyle C. Burris

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Peter Hoff, Supervisor

_____
Jerome P. Reiter

_____
Mine Çetinkaya-Rundel

_____
Rebecca Steorts

_____
L. Gregory Appelbaum

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Statistical Science
in the Graduate School of Duke University
2019

Abstract

Advances in Survey Methodology and Sports Science

by

Kyle C. Burris

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Peter Hoff, Supervisor

_____
Jerome P. Reiter

_____
Mine Çetinkaya-Rundel

_____
Rebecca Steorts

_____
L. Gregory Appelbaum

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistical Science in the Graduate School of Duke University
2019

Abstract

This thesis develops statistical methodology for efficient uncertainty quantification in the presence of small sample sizes and/or missing data. These methods have a wide range of potential applications, though they are particularly relevant for the analysis of cross-sectional survey data.

In the analysis of survey data it is frequently of interest to estimate and quantify uncertainty about means or totals for each of several non-overlapping subpopulations, or areas. Sometimes there are areas with small sample sizes under the survey design, which can result in wide confidence intervals. While some model-based methods have been developed to reduce interval width by utilizing data from other areas, these interval procedures do not have the nominal frequentist coverage rate for all values of the target quantity. We develop an alternative model-based confidence interval procedure that leverages data from other areas to reduce expected interval width. Importantly, our procedure maintains the nominal frequentist coverage rate for all values of the target quantity and is coverage-robust to model misspecification.

Missing data values are also pervasive in survey samples. Imputing multiple completed datasets is a principled way to avoid removing observations with incomplete values while simultaneously accounting for the uncertainty involved in the imputation procedure. The quality of imputations can be improved when the support of the data is known *a priori*. We develop methodology for multiple imputation of mixed data when the support is known *a priori* to be a subset of the data product space.

This can improve the quality of the resulting imputed data sets, resulting in more efficient statistical inference.

In addition to its contributions in the field of survey methodology, this thesis also contributes to the sports science literature by developing Bayesian latent variable models for the analysis of visual-motor expertise. In particular, we consider a multivariate dataset consisting of visual-motor assessments for 2317 athletes, including 252 professional baseball players. We quantify the variation in visual-motor expertise in athletes by level of expertise, gender, and sport type. Moreover, we examine the dependence among the battery of assessments and their relationship to on-field performance in professional baseball. We find significant positive associations between performance on the assessment battery and measures of baseball performance, particularly those that involve plate discipline.

Ad majorem Dei gloriam

Contents

List of Tables

List of Figures

List of Abbreviations and Symbols

Symbols

$\mathbb{R}$	The set of real numbers
$\boldsymbol{Y}$	A matrix, either random or observed
$\boldsymbol{y}$	A column vector, either random or observed
y	A scalar, either random or observed
Φ	Normal cumulative distribution function
N	Normal distribution
G	Gamma distribution

Abbreviations

EB	Empirical Bayes
FAB	Frequentist, Assisted by Bayes
MI	Multiple Imputation
PMM	Predictive Mean Matching
ERL	Extended Rank Likelihood
JTM	Joint Transformation Model
CJTM	Constrained Joint Transformation Model
ACS	American Community Survey
MCMC	Markov chain Monte Carlo
MAP	Maximum *a posteriori*

OBP	On-Base Percentage
BB%	Walk Rate
K%	Strikeout Rate
SLG	Slugging Percentage
FIP	Fielder-Independent Pitching
MLB	Major League Baseball
VC	Visual Clarity
CS	Contrast Sensitivity
DP	Depth Perception
NFQ	Near-Far Quickness
TC	Target Capture
PS	Perception Span
EHC	Eye-Hand Coordination
GNG	Go/No-Go
RXN	Reaction Time

Acknowledgements

The past four years have been incredibly challenging and rewarding. Conducting research is a highly non-linear process; there are periods of substantial progress followed by periods of stagnation and frustration. I am incredibly grateful for the continual support of many people throughout this journey, and I would like to express my gratitude to them here.

First and foremost, I would like to thank God, for his sanctifying grace is what enables me to develop, both professionally and personally. Secondly, I would like to thank my family for their presence, love, and encouragement. In particular, my wife Rachel Burris has been a bedrock of support throughout my time at Duke. By example, she consistently challenges me to become more thoughtful, caring, and diligent. I have also been blessed with wonderful parents, who have invested substantial time, effort, and resources in me over the years. They have always been a constant source of encouragement, believing in me even when I doubted myself.

Next, I would like to thank the countless faculty who have contributed to my academic development over the past four years. I have nothing but admiration and respect for my advisor, Peter Hoff. Peter is an excellent researcher, lecturer, and mentor. He adeptly straddles the line between patient and motivational, encouraging me to express myself clearly and precisely. Though I traversed a highly curved research path, he has been very supportive. I would also like to thank my preliminary advisor, Jerry Reiter. Jerry is kind, patient, and wonderful to work with. He cares

about each of his students and works hard to partner with them to achieve their goals. Mine Çetinkaya-Rundel has greatly contributed to my development as a teacher. She has not only profoundly influenced my teaching philosophies and practices, but also provided numerous teaching and course development opportunities. Her dedication and hard work is appreciated by thousands of students, both at Duke and around the world. I would also like to thank Rebecca Steorts for her helpful and insightful feedback on earlier drafts of this document on her willingness to do so on short notice. I am very grateful for Greg Appelbaum's enthusiasm and willingness to allow me to work in his lab the last few years. His optimism and passion for his research continues to inspire me in my work. Many thanks go to professors David Dunson, Merlise Clyde, Robert Wolpert, Mike West, Surya Tokdar, Li Ma, Fan Li, Scott Schmidler, David Carlson, and Colin Rundel for their instruction and support.

My fellow graduate students have been instrumental in keeping me grounded during this time. In particular, my office-mates Jake Coleman, Lindsay Berry, Liz Lorenzi, Phil White, Michael Jauch, Abbas Zaidi, and Austin Talbot always help me gain perspective. I give Jake much of the credit for exposing me to baseball analytics my first year of graduate school, which has helped spur my interest in statistical applications to sports.

The Cleveland Indians organization has been an incredible partner throughout the last few years. They have worked to develop me both as a person and as a statistician and I am thankful to have the opportunity to join their front office.

I gratefully acknowledge funding from the National Science Foundation (DMS-1505136, SES-1131897), the Duke Graduate School, and the U.S. Army.

1

Introduction

Statistical inference is fundamentally about learning the properties of an underlying process based on limited observations of that process. In many cases, this process is an environment in which decisions must be made. In order to make optimal decisions in such an unknown environment, it is essential to characterize not only the most likely state, but also a set of plausible states of the decision-making environment.

Decisions can be made with more environmental certainty when the set of plausible processes is small. In general, the size of this set can be reduced by restricting attention to certain types of processes, imposing a stricter definition of plausibility, and/or utilizing more information about the underlying process. Chapters 2 and 3 of this thesis provide novel methodology for utilizing additional information in a data set to reduce the set of plausible generating processes. As a result, these methods make uncertainty quantification about the process of interest more efficient.

Much of statistical practice involves describing an underlying process via a probabilistic model, which embodies a set of mathematical assumptions made about the process. As put by Box and Draper (1986), "all models are wrong, but some are useful." Models can be useful for their ability to describe salient features of the under-

lying process and predict future observations from the process. Chapter 4 develops and examines classes of statistical models designed to provide insight into the underlying process of athletic performance. The remainder of this chapter provides some general background on statistical concepts that form the backbone of the research topics that follow.

1.1 Background

This section contains some necessary background material on confidence intervals, multiple imputation, and copulas. The methodology proposed in this thesis relies heavily on these concepts.

1.1.1 Confidence Intervals

Suppose that a random vector $\boldsymbol{y}$ is assumed to be generated from a probability distribution P, indexed by parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. A confidence region procedure is a function $\mathcal{C} : \mathcal{Y} \to \mathcal{P}(\boldsymbol{\Theta})$, where $\mathcal{Y}$ is the support of $\boldsymbol{y}$ and $\mathcal{P}(\boldsymbol{\Theta})$ is the power set of $\boldsymbol{\Theta}$. Moreover, the coverage probability of $\mathcal{C}$ at $\boldsymbol{\theta}$ is given by

$$P(\boldsymbol{\theta} \in \mathcal{C}(\boldsymbol{y}) \mid \boldsymbol{\theta}). \tag{1.1}$$

Here, note that the confidence region $\mathcal{C}(\boldsymbol{y})$ is a random quantity, not the parameter. A confidence region procedure maintains $1 - \alpha$ frequentist coverage if

$$\inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} P(\boldsymbol{\theta} \in \mathcal{C}(\boldsymbol{y}) \mid \boldsymbol{\theta}) \geqslant 1 - \alpha. \tag{1.2}$$

A confidence region procedure that maintains $1 - \alpha$ frequentist coverage is desirable in many applications, as there is a known lower bound on the proportion of random samples for which the corresponding confidence region will contain the true parameter. Moreover, a confidence region obtained via this procedure corresponds to an acceptance region of a level-α hypothesis test.

Confidence interval procedures are special cases of confidence region procedures, where $\boldsymbol{\Theta} = \mathbb{R}$ and $\mathcal{C}(\boldsymbol{y})$ can be expressed as $[L(\boldsymbol{y}), U(\boldsymbol{y})]$, where $L(\boldsymbol{y}) \leqslant U(\boldsymbol{y})$ for all $\boldsymbol{y} \in \mathcal{Y}$. As an example, consider a situation in which $y_i \sim N(\theta, \sigma^2)$, $i = 1, \cdots, n$, with the variance σ^2 known. Consider the direct confidence interval procedure

$$\mathcal{C}(y_1, \cdots, y_n) = \left\{ \theta : \bar{y} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2} < \theta < \bar{y} + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} \right\}, \tag{1.3}$$

where $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ and z_p is the pth quantile of the standard normal distribution. It is easy to show that this procedure maintains $1 - \alpha$ frequentist coverage and has width $\frac{\sigma}{\sqrt{n}}(z_{1-\alpha/2} - z_{\alpha/2})$. In general, for a given level of frequentist coverage, it is preferable to have narrower confidence intervals because this reduces the uncertainty about θ. As $n \to \infty$, the width of (1.3) goes to zero. However, if n is small and/or σ^2 is large, the interval can be quite wide.

1.1.2 Multiple Imputation

Consider a dataset collected by a researcher or statistical organization that consists of n units, for which up to p numeric variables may be recorded. Suppose that we wish to use this dataset to perform inference on a target quantity Q. Depending on the context, Q may be a function of complete data from a finite population (e.g., population mean) or a superpopulation parameter (e.g., regression coefficient). Inference on Q may be complicated by item nonresponse, in which some of the values in the data are unobserved.

To more succinctly describe the problems that item nonresponse presents, we introduce some mathematical notation. We define the data matrix $\boldsymbol{Y}$ to be the $n \times p$ matrix representing the values of the p numeric variables for all n units in the sample. Let $\boldsymbol{y}_i$ be the ith row of $\boldsymbol{Y}$, $\boldsymbol{y}_{(j)}$ be the jth column of $\boldsymbol{Y}$ and y_{ij} be the value of jth numeric variable for the ith unit, $i = 1, \cdots, n$, $j = 1, \cdots p$. Furthermore, we define the $n \times p$ response matrix $\boldsymbol{R}$, where the element $r_{ij} = 1$ if y_{ij} is observed and zero

otherwise. Finally, let $\boldsymbol{Y}_{obs} = \{y_{ij} : r_{ij} = 1\}$ and $\boldsymbol{Y}_{mis} = \{y_{ij} : r_{ij} = 0\}$ be the set of observed and unobserved values in the data, respectively. The actually observed data is $(\boldsymbol{Y}_{obs}, \boldsymbol{R})$.

Suppose that when all of the data values are observed (i.e., $r_{ij} = 1$ for all i, j), there exist standard statistical procedures for valid point and interval estimation of Q, but that these procedures are not applicable when some data values are unobserved. When this is the case, it is of interest to obtain a dataset without any missing values so that these procedures can be applied.

One way to obtain such a dataset is complete cases analysis, in which all units with any missing values are discarded. However, valid inference can only be obtained via complete cases analysis if the data are missing completely at random (MCAR), such that

$$p(\boldsymbol{R} \mid \boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}, \boldsymbol{\psi}) = p(\boldsymbol{R} \mid \boldsymbol{\psi}), \tag{1.4}$$

where $\boldsymbol{\psi}$ indexes a model for the response mechanism. Even when the data are MCAR, complete cases analysis is inefficient, since data that could provide information about Q are discarded prior to statistical analysis (Little and Rubin, 1986).

Imputation is another common technique for handling missing data; missing values are filled in and then treated as known quantities when performing the complete-data procedures. Many imputation procedures make the assumption that the data are missing at random (MAR), such that

$$p(\boldsymbol{R} \mid \boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}, \boldsymbol{\psi}) = p(\boldsymbol{R} \mid \boldsymbol{Y}_{obs}, \boldsymbol{\psi}). \tag{1.5}$$

When data are MAR, it is unnecessary to model the response mechanism when imputing the missing data values (Rubin, 1987), as

$$p(\boldsymbol{Y}_{mis} \mid \boldsymbol{Y}_{obs}, \boldsymbol{R}) = p(\boldsymbol{Y}_{mis} \mid \boldsymbol{Y}_{obs}). \tag{1.6}$$

Despite this, imputing $\boldsymbol{Y}_{mis}$ is challenging because $p(\boldsymbol{Y}_{mis} \mid \boldsymbol{Y}_{obs})$ is unknown and must be modeled. If the model for $p(\boldsymbol{Y}_{mis} \mid \boldsymbol{Y}_{obs})$ is properly specified, complete data

inference for Q based on a single simulation of $\boldsymbol{Y}_{mis}$ generally leads to unbiased point estimation but invalid interval estimation, as the uncertainty inherent in the missing data values is not accounted for (Reiter and Raghunathan, 2007).

Multiple imputation (Rubin, 1987) has become an increasingly popular way to handle item nonresponse due to its ability to accommodate this additional source of uncertainty. Rather than simply filling in the missing values and analyzing a single completed dataset, one simulates M completed datasets based on a statistical model, performs separate inferences on Q based on each of the completed datasets, and combines these inferences. In particular, multiple imputation can motivated under a Bayesian approach, for

$$p(\theta \mid \boldsymbol{Y}_{obs}) = \int p(\theta \mid \boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis})\ p(\boldsymbol{Y}_{mis} \mid \boldsymbol{Y}_{obs})\ d\boldsymbol{Y}_{mis}, \tag{1.7}$$

$$\mathrm{E}\left[\theta \mid \boldsymbol{Y}_{obs}\right] = \mathrm{E}\left[\mathrm{E}\left[\theta \mid \boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}\right] \mid \boldsymbol{Y}_{obs}\right], \tag{1.8}$$

$$\mathrm{Var}\left[\theta \mid \boldsymbol{Y}_{obs}\right] = \mathrm{Var}\left[\mathrm{E}\left[\theta \mid \boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}\right] \mid \boldsymbol{Y}_{obs}\right] + \mathrm{E}\left[\mathrm{Var}\left[\theta \mid \boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}\right] \mid \boldsymbol{Y}_{obs}\right]. \tag{1.9}$$

If the imputation model $p(\boldsymbol{Y}_{mis} \mid \boldsymbol{Y}_{obs})$ is well specified, approximate samples from $p(\boldsymbol{Y}_{mis} \mid \boldsymbol{Y}_{obs})$ can be obtained by sampling approximately from the joint posterior distribution $p(\boldsymbol{Y}_{mis}, \boldsymbol{\psi} \mid \boldsymbol{Y}_{obs})$. Multiple imputation involves taking these samples and using them to construct Monte Carlo estimates of (1.8) and (1.9). Specifically, suppose that we obtain K approximate samples $\{\boldsymbol{Y}_{mis}^{(k)}\}$, $k = 1, \cdots, K$, from $p(\boldsymbol{Y}_{mis} \mid \boldsymbol{Y}_{obs})$. Let $\hat{\theta}(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis})$ be an estimator of θ. Define the estimates $\hat{\theta}_k = \hat{\theta}(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}^{(k)})$, $k = 1, \cdots, K$, with associated variance estimates $\hat{W}_k$. These K estimates can be

combined to create an uncertainty interval for θ as follows. Let

$$\bar{\theta}_K = \frac{1}{K}\sum_{k=1}^{K}\hat{\theta}_k. \tag{1.10}$$

$$\overline{W}_K = \frac{1}{K}\sum_{k=1}^{K}\hat{W}_k. \tag{1.11}$$

$$B_K = \frac{1}{K-1}\sum_{k=1}^{K}(\hat{\theta}_k - \bar{\theta}_K)^2. \tag{1.12}$$

$$T_K = \overline{W}_K + \frac{K+1}{K}B_K. \tag{1.13}$$

$$\nu_K = (K-1)\left(1 + \frac{\overline{W}_K}{(K+1)B_K}\right)^2. \tag{1.14}$$

When the number of observations n is large and K is modest, the quantity $(\bar{\theta}-\theta)$ is approximately t-distributed with variance T_K and degrees of freedom ν_K. From this, confidence intervals can be constructed for θ with approximately $1-\alpha$ frequentist coverage. While such an interval procedure is not guaranteed to maintain $1-\alpha$ frequentist coverage due to the bias in T_K and variations in the imputation model, inferences based on multiple imputation procedures are sensible in a variety of applications (Rubin, 2003). Adaptations of multiple imputation in the case of small sample sizes and multivariate target quantities are discussed by Reiter and Raghunathan (2007).

1.2 Research Topics and Principal Contributions

Below, we provide summaries of the research topics and contributions that the work contained in this thesis makes to the academic literature. This thesis draws some of its material verbatim from other papers written by its author, including Burris et al. (2018), Burris et al. (2019), Burris and Hoff (2019a), and Burris and Hoff (2019b).

1.2.1 Exact adaptive confidence intervals for small areas

In the analysis of survey data, it is frequently of interest to estimate and quantify uncertainty about means or totals for each of several non-overlapping subpopulations, or areas. In order to reduce interval width, practitioners often utilize multilevel models in order to borrow strength across areas, resulting in intervals centered around shrinkage estimators. However, such intervals only have the nominal coverage rate on average across areas under the assumed model for across-area heterogeneity. The coverage rate for a given area depends on the actual value of the area mean, and can be nearly zero for areas with means that are far from the across-group average. As such, the use of uncertainty intervals centered around shrinkage estimators are inappropriate when area-specific coverage rates are desired. In Chapter 2, we propose an alternative confidence interval procedure for area means and totals under normally distributed sampling errors. This procedure not only has constant $1 - \alpha$ frequentist coverage for all values of the target quantity, but also uses auxiliary information to borrow strength across areas. Because of this, the corresponding intervals have shorter expected lengths than standard confidence intervals centered on the unbiased direct estimator. Importantly, the coverage of the procedure does not depend on the assumed model for across-area heterogeneity. Rather, improvements to the model for across-area heterogeneity result in reduced expected interval width.

1.2.2 Bayesian hot deck imputation for multivariate numeric data

Many datasets collected by organizations contain observations for which some fields are missing. To handle missing data, organizations extensively use multiple imputation, generating multiple completed datasets that can be combined when performing inference. Despite the popularity of multiple imputation for handling missing data, few approaches exist for performing multiple imputation when datasets contain a combination of continuous, discrete, ordered categorical, and binary variables. The

approaches that have been developed suffer from important theoretical and/or practical limitations.

In Chapter 3, we present a new method to flexibly model the joint distribution of mixed continuous, discrete, ordered categorical, and binary data. Our approach assumes that the observed data values are transformations of underlying continuous latent variables, which are modeled via a truncated Dirichlet process mixture of multivariate normal distributions. We propose a Bayesian procedure for model estimation that enables straightforward multiple imputation. Under this procedure, each missing value is filled in with an observed data value, resulting in a Bayesian "hot deck" multiple imputation engine. The proposed method is flexible, straightforward to implement, and suitable for moderately sized, cross-sectional datasets. We extend the method to perform constrained multiple imputation, which is desirable when *a priori* information about the support of the data is available. We illustrate our approach via simulation studies and the analysis of 2017 American Community Survey microdata.

1.2.3 Visual-motor expertise in athletes

Elite athletes not only run faster, hit harder, and jump higher, but also see and react better. However, the specific visual-motor skills that differentiate high-achieving athletes are still not well understood. In Chapter 4, we examine 2317 athletes (1871 male) tested on the Nike SPARQ Sensory Station, a digital test battery measuring visual, perceptual and motor skills relevant for sports performance. We develop a multivariate Gaussian transformation model to robustly estimate visual-motor differences by level, gender, and sport type. Results demonstrate that visual-motor performance is superior for athletes at higher levels, with males faster at near-far eye movements and females faster at eye-hand reaction times. Interestingly, athletes who play interceptive sports such as baseball and tennis exhibit better visual sensitivity

and simple reaction times, while athletes from strategic sports like soccer and basketball have higher perception spans. These findings provide quantitative evidence of domain-specific visual expertise in athletes.

In addition, we consider the subset of 252 professional baseball players in an effort to evaluate the links between sensorimotor skills and on-field performance. For this purpose, we develop a series of Bayesian hierarchical latent variable models enabling the comparison of statistics across professional baseball leagues. Within this framework, we find that sensorimotor abilities are significant predictors of on-base percentage, walk rate and strikeout rate, accounting for age, position, and league. We find no such relationship for either slugging percentage or fielder-independent pitching. The pattern of results suggests performance contributions from both visual-sensory and visual-motor abilities and indicates that sensorimotor screenings may be useful for player scouting.

2

Exact Adaptive Confidence Intervals for Small Areas

2.1 Introduction

Studies that gather data from non-overlapping areas (subpopulations) are common in a variety of disciplines, including ecology (Brewer and Nolan, 2007), education (Wall, 2004), epidemiology (Ghosh et al., 1999), public policy (Maples, 2017), and sports (Efron and Morris, 1977). As policy interventions have become more targeted, the demand for precise estimates of population characteristics of these areas has increased. To estimate target quantities, sample surveys may use "direct" estimators, which are only based on the area-specific sample data. Direct estimators typically utilize survey weights, with corresponding inferences made based on the sampling design (Rao and Molina, 2015). When the direct estimates are area-specific sample averages (possibly weighted), the central limit theorem justifies the area-specific *sampling model* $y_j \sim N(\theta_j, \sigma_j^2),\ j = 1, \dots m$, where y_j is a design-unbiased and consistent direct estimate of θ_j, the jth area mean, and σ_j^2 is the variance of the direct estimate under the sampling design. If additionally the survey data are sam-

pled independently across areas, the joint sampling model for the area-specific direct estimates is

$$\boldsymbol{y} \sim N(\boldsymbol{\theta}, \boldsymbol{D}), \tag{2.1}$$

where $\boldsymbol{y} = (y_1, \ldots, y_m)$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$, and $\boldsymbol{D}$ a diagonal matrix with elements $\{\sigma_1^2, \ldots, \sigma_m^2\}$.

For a specific area j, when σ_j^2 is assumed known, the classical "direct" $1 - \alpha$ confidence interval for θ_j is

$$C_D^j(\boldsymbol{y}) = \left\{\theta : y_j + \sigma_j z_{\alpha/2} < \theta < y_j + \sigma_j z_{1-\alpha/2}\right\}, \tag{2.2}$$

where z_p is the pth quantile of the standard normal distribution. This direct confidence interval has the important property of *area-specific coverage* under the sampling model (2.1), since

$$\Pr(\theta_j \in C_D^j(\boldsymbol{y}) \mid \boldsymbol{\theta}) = 1 - \alpha, \tag{2.3}$$

for all $\boldsymbol{\theta}$ and $j = 1, \ldots, m$.

However, it is sometimes the case that there are areas with small sample sizes under the survey design, resulting in unacceptably wide direct confidence intervals (Pfeffermann, 2013). When additional interval precision is needed, model-based estimators are used to borrow information from other areas and utilize area-level auxiliary covariates. A statistical model for across-area heterogeneity is referred to as a *linking model* in the small area estimation literature. For example, the popular Fay-Herriot model (Fay and Herriot, 1979) posits that $\theta_j \sim N(\boldsymbol{x}_j^\top \boldsymbol{\beta}, \tau^2)$, independently across areas, where $\boldsymbol{x}_j$ is a vector of observed area-specific covariates. If the values of $\boldsymbol{\psi} = \{\boldsymbol{\beta}, \tau^2\}$ are known, then the Fay-Herriot linking model corresponds to a $N(\boldsymbol{x}_j^\top \boldsymbol{\beta}, \tau^2)$ prior over θ_j. Given this prior, Bayes' rule could be used to obtain the conditional distribution of θ_j given y_j. From this distribution, one could compute a

Bayesian credible interval

$$C_B^j(\boldsymbol{y}) = \left\{\theta : \check{\mu}_j + \check{\tau}_j z_{\alpha/2} < \theta < \check{\mu}_j + \check{\tau}_j z_{1-\alpha/2}\right\}, \tag{2.4}$$

where $\check{\mu}_j$ and $\check{\tau}_j^2$ are the conditional mean and variance of θ_j given y_j, respectively.

In practice, appropriate values for the linking model parameters $\boldsymbol{\psi}$ are unknown. A fully Bayesian approach is to place a prior distribution on $\boldsymbol{\psi}$, from which the joint posterior distribution of $\theta_1, \ldots, \theta_m$ may be obtained (You and Chapman, 2006). A more common approach is an empirical Bayes strategy, whereby "plug-in" estimates of $\boldsymbol{\psi}$ are obtained from the marginal likelihood of $\boldsymbol{\psi}$, which is itself obtained by integrating the density of the sampling model (2.1) for $\boldsymbol{y}$ over the values of $\boldsymbol{\theta}$ with respect to the linking model. Given such an estimate $\hat{\boldsymbol{\psi}}$ of $\boldsymbol{\psi}$, the empirical Bayes confidence interval is given by

$$C_{EB}^j(\boldsymbol{y}) = \left\{\theta : \hat{\mu}_j + \hat{\tau}_j z_{\alpha/2} < \theta < \hat{\mu}_j + \hat{\tau}_j z_{1-\alpha/2}\right\}, \tag{2.5}$$

where $\hat{\mu}_j$ and $\hat{\tau}_j^2$ are the conditional mean and variance of θ_j, given y_j and using $\hat{\boldsymbol{\psi}}$ as the parameters in the linking model.

The Bayesian credible interval procedure C_B^j has the property of population-level coverage, in the sense that the coverage level is $1-\alpha$ *on average* with respect to the linking model. Specifically,

$$\int \Pr(\theta_j \in C_B^j(\boldsymbol{y})|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\psi})d\theta = 1-\alpha, \tag{2.6}$$

where $\pi(\boldsymbol{\theta}|\boldsymbol{\psi})$ is the probability density of $\boldsymbol{\theta}$ under the linking model. The empirical Bayes confidence interval procedure C_{EB}^j has this property asymptotically in the number of groups, as long as $\hat{\boldsymbol{\psi}}$ is a consistent estimator of $\boldsymbol{\psi}$. This property only holds when the linking model is correctly specified. However, neither C_B^j nor C_{EB}^j have $1-\alpha$ area-specific coverage, as defined in (2.3). This is because they are

centered around a biased estimator of θ_j, given θ_j. To illustrate this lack of area-specific coverage, consider the Fay-Herriot linking model

$$\theta_j \sim N(\boldsymbol{x}_j^\top \boldsymbol{\beta}, \tau^2), \qquad j = 1, \ldots, m \tag{2.7}$$

where $\boldsymbol{x}_j$ is a vector of covariates for area j. Standard conditional probability calculations (provided in the A.1) give that the area-specific coverage of C_B^j is a function of $\theta_j - \boldsymbol{x}_j^\top \boldsymbol{\beta}$ and can be expressed as

$$\begin{aligned} &\Phi\left(\sigma_j(\theta_j - \boldsymbol{x}_j^\top \boldsymbol{\beta})/\tau^2 + z_{1-\alpha/2}\sqrt{1 + \sigma_j^2/\tau^2}\right) \\ &\quad - \Phi\left(\sigma_j(\theta_j - \boldsymbol{x}_j^\top \boldsymbol{\beta})/\tau^2 + z_{\alpha/2}\sqrt{1 + \sigma_j^2/\tau^2}\right), \end{aligned} \tag{2.8}$$

where Φ is the standard normal cumulative distribution function. In general, the coverage probability for a given area will be higher than the nominal level when θ_j is close to $\boldsymbol{x}_j^\top \boldsymbol{\beta}$ and lower when θ_j is far away from $\boldsymbol{x}_j^\top \boldsymbol{\beta}$, a relationship that is visualized in Figure 2.1. This difference is amplified when the linking model variance τ^2 is small relative to the sampling variance σ_j^2.

In many applications, policy decisions and interventions are frequently targeted at outlying groups or areas. In these cases, it is important that uncertainty intervals have area-specific coverage, so that the study has sufficient power to detect extreme values of the target quantity, regardless of what it may be. If area-specific coverage is desired, neither the C_B nor C_{EB} interval procedures can be recommended, as their coverage levels will vary as a function of the target quantity θ_j. However, intervals generated by the direct interval procedure C_D may also be unsatisfactory, since they may be too wide to be useful when area sample sizes are small, as they do not make use of information across areas.

In this chapter, we propose a confidence interval procedure for small area analysis that maintains exact area-specific coverage, while also allowing for information sharing across areas via a linking model. Because they borrow information from

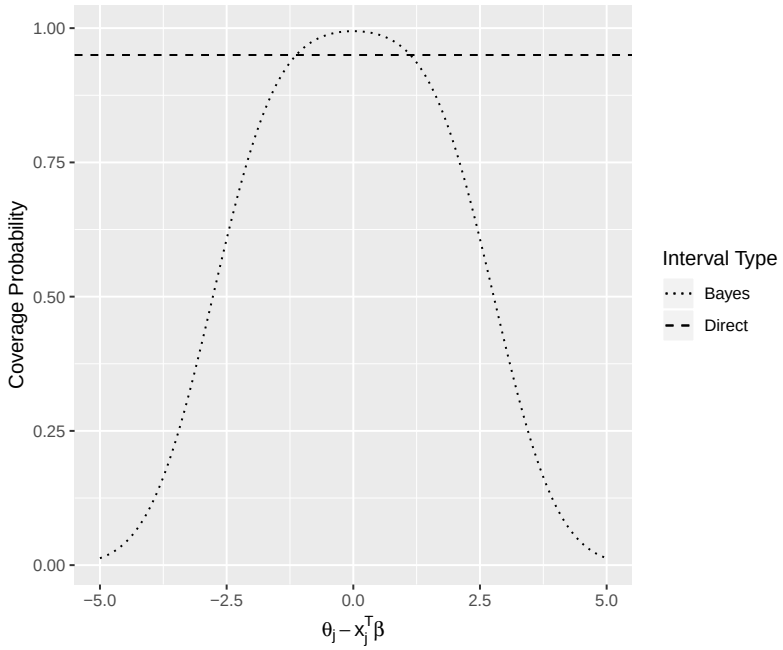


FIGURE 2.1: Area-specific coverage probability for C_D and C_B at $\alpha = 0.05$ under the linking model $\theta_j \sim N(0, 1)$ and sampling model $y_j \sim N(\theta_j, 1)$. Although C_B (and asymptotically C_{EB}) obtains 95% coverage probability on average across values of $\theta_j - \boldsymbol{x}_j^\top \boldsymbol{\beta}$, there will be some areas that have much less than the nominal coverage probability. In contrast, the direct interval maintains $1-\alpha$ area-specific coverage for all areas regardless of the value of $\boldsymbol{\theta}$.

other areas, these intervals are narrower on average than corresponding direct intervals. Importantly, unlike the Bayes and empirical Bayes procedures, our procedure is appropriate for area-level inference in that it maintains area-specific coverage rates. Moreover, like direct confidence intervals, these intervals have exact $1-\alpha$ area-specific coverage under the sampling model (2.1), regardless of whether or not a particular linking model is misspecified, making them *coverage-robust*.

Our proposed interval procedure extends that of Yu and Hoff (2018), who developed an adaptive procedure with area-specific coverage using an exchangeable linking model. In this chapter, we extend this idea to the types of linking models often used for small area analysis, including models that allow for area-specific features and spatial or temporal correlation between area means. In Section 2.2, we

briefly review the interval procedure first developed by Pratt (1963), and extended by Yu and Hoff (2018) to include the case of unknown sampling variances. We also demonstrate how to apply these ideas to the analysis of small areas, using the spatial Fay-Herriot linking model as a running example. Section 2.3 describes a simulation study designed to compare interval procedures under a variety of linking models. In Section 2.4, we apply our methodology to estimate household radon levels in 196 U.S. counties. A discussion follows in Section 2.5.

2.2 Methods

2.2.1 The FAB interval procedure

We first consider constructing a $1-\alpha$ confidence interval procedure for a specific group j, based on the sampling model (2.1), where for now we assume σ_j^2 to be known. Let s_j be any function mapping $\mathbb{R}$ to the unit interval $[0,1]$, possibly depending on data from other areas, that is, $\boldsymbol{y}_{-j} = \{y_i : i \neq j\}$. Then, assuming the sampling model, it is easily verified that

$$C_{s_j}^j = \left\{\theta : y_j + \sigma_j z_{\alpha(1-s_j(\theta))} < \theta < y_j + \sigma_j z_{1-\alpha s_j(\theta)}\right\} \tag{2.9}$$

is a valid $1-\alpha$ frequentist confidence region, satisfying the area-specific coverage property (2.3). All $1-\alpha$ frequentist confidence intervals are elements in this class and have the property of area-specific coverage. The standard direct interval corresponds to $s_j(\theta) = 1/2$.

Now suppose that, based on $\boldsymbol{y}_{-j}$ and a linking model, we believe θ_j is likely to be near some value μ_j. We encode this belief with a normal probability distribution $\theta_j \sim N(\mu_j, \tau_j^2)$. For example, μ_j and τ_j^2 might be the conditional expectation and variance of θ_j, given $\boldsymbol{y}_{-j}$ and the linking model. Given such information, we may prefer an area-specific interval procedure that, relative to the direct interval, is more precise (has shorter expected width) for values of θ_j near μ_j, at the expense of having

longer expected width for values of θ_j deemed unlikely by the linking model. We may then wish to use the area-specific interval procedure that minimizes the expected width, relative to the linking model.

The minimizer of this expected width among all $1-\alpha$ frequentist intervals can be obtained using results of Pratt (1963), who considered frequentist interval construction for a single mean parameter with a normal prior distribution. The $1-\alpha$ frequentist interval that has minimum width, on average with respect to a $N(\mu_j, \tau_j^2)$ distribution for θ_j, can be shown to be given by (2.9) with

$$\begin{aligned} s_j(\theta) &= g^{-1}(2\sigma_j(\theta-\mu_j)/\tau_j^2) \\ g(\omega) &= \Phi^{-1}(\alpha\omega) - \Phi^{-1}(\alpha(1-\omega)). \end{aligned} \tag{2.10}$$

s_j in (2.10) depends on θ, a hypothesized value of the jth population mean, not θ_j, which is the true value of the jth population mean. Following Yu and Hoff (2018), we refer to confidence intervals constructed in this way as FAB intervals because, thinking of the conditional distribution of θ_j given $\boldsymbol{y}_{-j}$ as a prior distribution, they are "frequentist, assisted by Bayes". Importantly, even if θ_j is located in a region of low probability under the linking model, a FAB interval will still maintain $1-\alpha$ area-specific coverage for θ_j. As such, the FAB interval procedure is coverage-robust to misspecification of the linking model. In contrast, the Bayes and empirical Bayes interval procedures do not maintain constant area-level coverage rates even if the linking model perfectly describes the across-area distribution of $\boldsymbol{\theta}$ (unless all area-specific means are the same). In terms of precision, FAB intervals represent significant improvements in expected interval width over direct intervals if 1) data from other areas provide information about the area mean and 2) the specified linking model can capture this information (Figure 2.2).

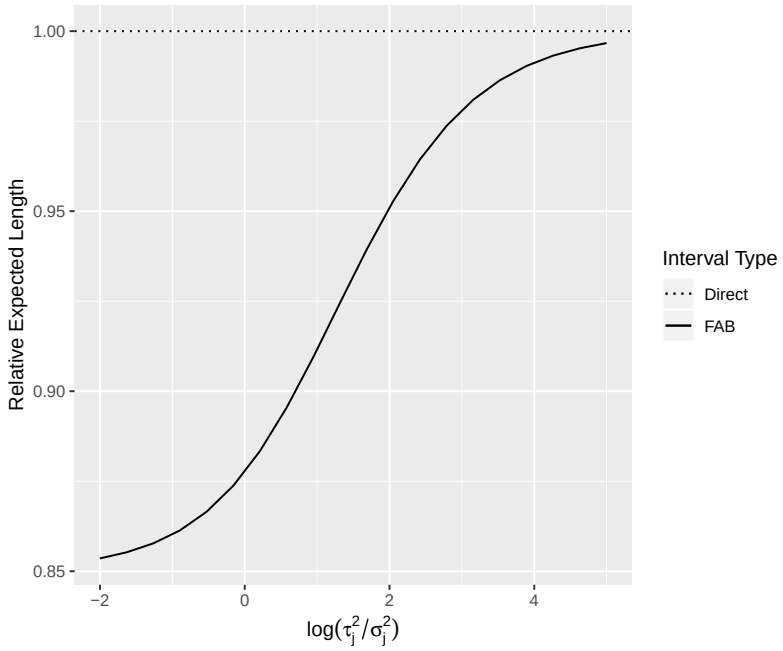


FIGURE 2.2: The expected relative improvement of the 95% FAB $z-$interval over the direct interval. When the prior variance is of a similar magnitude to the sampling variance or smaller, there can be a substantial reduction in expected interval width. However, there appear to be diminishing returns as the prior variance τ_j^2 decreases, due to the constraint of constant $1-\alpha$ frequentist coverage.

2.2.2 FAB intervals for the spatial Fay-Herriot model

The spatial Fay-Herriot model is frequently employed by researchers and statistical agencies due to the abundance of cross-sectional survey data that come from non-overlapping geographic areas such as counties, neighborhoods, school districts, and electoral precincts. Area-level direct estimates from this type of data typically exhibit high spatial autocorrelation, in which areas closer together tend to have similar values for their target quantities, even after accounting for the auxiliary covariates.

The spatial Fay-Herriot model includes the sampling model (2.1) which we assume to be correct, and a spatial linking model for across-unit heterogeneity of the θ_j's, which we do not assume is correct. The linking model can be written as

$$\boldsymbol{\theta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \qquad \boldsymbol{u} \sim N(\boldsymbol{0}, \boldsymbol{G}(\boldsymbol{\psi})) \tag{2.11}$$

where $\boldsymbol{\psi} = \{\tau^2, \rho\}$ parameterizes the dispersion and spatial relationship of the random effects. The conditional autoregressive (CAR) model and the simultaneous autoregressive (SAR) model are two of the main approaches for structured covariance modeling of spatially autocorrelated areal data (Banerjee et al., 2014). Following Singh et al. (2005) and Pratesi and Salvati (2008), we consider the SAR model

$$\boldsymbol{u} = \rho \boldsymbol{W} \boldsymbol{u} + \boldsymbol{v} \qquad \Rightarrow \qquad \boldsymbol{u} = (\boldsymbol{I} - \rho \boldsymbol{W})^{-1} \boldsymbol{v}, \tag{2.12}$$

where $\boldsymbol{W}$ is a $m \times m$ neighborhood proximity matrix, ρ a spatial relationship parameter, and $\boldsymbol{v}$ a $m \times 1$ mean-zero random vector with independent normal entries, each with variance τ^2. A binary continguity neighborhood matrix is often chosen for $\boldsymbol{W}$, in which $W_{ij} = 1$ if areas i and j are neighbors and zero otherwise. Regardless of the choice of $\boldsymbol{W}$, it is typically first row-standardized to make the row elements sum to one. When the proximity matrix is standardized in this way, $\boldsymbol{I} - \rho \boldsymbol{W}$ is non-singular when $\rho \in (-1, 1)$, and ρ can be treated as a spatial autocorrelation parameter.

Combining the above equations, the linking model for $\boldsymbol{\theta}$ becomes

$$\boldsymbol{\theta} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \tau^2[(\boldsymbol{I} - \rho \boldsymbol{W})(\boldsymbol{I} - \rho \boldsymbol{W}^T)]^{-1}) \tag{2.13}$$

Our proposed confidence interval for a small area mean θ_j is obtained by first using data $\boldsymbol{y}_{-j} = (y_1, \ldots, y_{j-1}, y_{j+1}, \ldots, y_m)$ from the other groups, along with the linking model (2.13) to obtain a mean μ_j and variance τ_j^2 that describe the likely values of θ_j, and then using these values to construct the FAB interval given by (2.9) and (2.10). Recall that the resulting confidence interval has exact $1 - \alpha$ coverage for θ_j, regardless of the value of θ_j or the accuracy of the linking model, as long as the sampling model is correct and the values of μ_j and τ_j^2 are chosen independently of the value of y_j.

A fully Bayesian approach to obtaining values of μ_j and τ_j^2 would be to take them to be the conditional mean and variance of θ_j given $\boldsymbol{y}_j$, under a suitable prior

distribution for the parameters $\{\boldsymbol{\beta}, \tau^2, \rho\}$ of the linking model, and computed using a MCMC approximation algorithm. However, this can be prohibitively computationally costly, as a separate approximation would need to be run for each area. As a more feasible alternative, we suggest an empirical Bayes approach in which $\{\boldsymbol{\beta}, \tau^2, \rho\}$ are first estimated from the marginal distribution of $\boldsymbol{y}_{-j}$, which are then used to obtain empirical Bayes estimates of the $\boldsymbol{\theta}_{-j}$'s. The resulting conditional mean and variance of θ_j, using "plug-in" values of $\boldsymbol{\theta}_{-j}$ and $\{\boldsymbol{\beta}, \tau^2, \rho\}$, are given by

$$\mu_j = \mathrm{E}\left[\theta_j \mid \boldsymbol{\theta}_{-j} = \hat{\boldsymbol{\theta}}_{-j}, \boldsymbol{\beta} = \hat{\boldsymbol{\beta}}, \rho = \hat{\rho}, \tau^2 = \hat{\tau}^2\right] \tag{2.14}$$

$$= \boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}} + \hat{\boldsymbol{G}}_{j,-j}\hat{\boldsymbol{G}}_{-j,-j}^{-1}\left(\hat{\boldsymbol{\theta}}_{-j} - \boldsymbol{X}_{-j}\hat{\boldsymbol{\beta}}\right) \tag{2.15}$$

$$\tau_j^2 = \mathrm{Var}\left[\theta_j \mid \boldsymbol{\theta}_{-j} = \hat{\boldsymbol{\theta}}_{-j}, \boldsymbol{\beta} = \hat{\boldsymbol{\beta}}, \rho = \hat{\rho}, \tau^2 = \hat{\tau}^2\right] \tag{2.16}$$

$$= \hat{\boldsymbol{G}}_{j,j} - \hat{\boldsymbol{G}}_{j,-j}\hat{\boldsymbol{G}}_{-j,-j}^{-1}\hat{\boldsymbol{G}}_{-j,j}, \tag{2.17}$$

where $\hat{\boldsymbol{G}} = \hat{\tau}^2\left(\left(\boldsymbol{I} - \hat{\rho}\boldsymbol{W}\right)\left(\boldsymbol{I} - \hat{\rho}\boldsymbol{W}\right)\right)^{-1}$

In sum, the steps to obtain a FAB interval are

1. Estimate linking model parameters using data from all counties other than j. Details of maximum likelihood estimation for the spatial Fay-Herriot are provided in A.3.

2. Obtain a normal prior distribution for θ_j using plug-in estimates from the fitted linking model. In the case of the spatial Fay-Herriot model, the prior mean μ_j and prior variance τ_j^2 are given by (2.14).

3. Obtain the optimal s-function for area j given prior information about θ_j, as described in Section 2.2.

4. Construct the z-interval $\left\{\theta : y_j + \sigma_j z_{\alpha(1-s_j(\theta))} < \theta < \bar{y}_j + \sigma_j z_{1-\alpha s_j(\theta)}\right\}$.

2.2.3 Unknown within-area variances

The procedure detailed above assumes that the sampling variance σ_j^2 is known (or known with a high degree of accuracy). However, in practice the variance of the direct estimate of each area is rarely known, and only consistent estimates $\hat{\sigma}_j^2$ are available. Under the assumption that the response is normally distributed within area j,

$$q_j \hat{\sigma}_j^2 / \sigma_j^2 \sim \chi_{q_j}^2, \tag{2.18}$$

where q_j is the effective number of degrees of freedom for area j implied by the sampling design (Cochran, 1977). Yu and Hoff (2018) extended Pratt's original z-interval to the case of an unknown sampling variance as follows: If the sample statistics y_j and $\hat{\sigma}_j^2$ are independent, where $y_j \sim N(\theta_j, \sigma_j^2)$ and $q_j \hat{\sigma}_j^2 / \sigma_j^2 \sim \chi_{q_j}^2$, then for any nondecreasing function $s_j : \mathbb{R} \to [0, 1]$,

$$C_{s_j}^j(y, \hat{\sigma}_j) = \{\theta : y_j + \hat{\sigma}_j t_{\alpha(1-s_j(\theta)),q_j} < \theta < y_j + \hat{\sigma}_j t_{1-\alpha s_j(\theta),q_j}\}, \tag{2.19}$$

where t_{p,q_j} is the pth quantile of the t distribution with q_j degrees of freedom, is a valid $1 - \alpha$ confidence interval with area-specific coverage. The function $s_j(\theta)$ can be selected on the basis of prior information about not only the target quantity θ_j, but also the sampling variance σ_j^2. If this prior information can be summarized by a normal distribution for θ_j and an inverse-gamma distribution for σ_j^2, it is possible to obtain the function s_j that minimizes the prior expected length of the interval (2.19) via numerical methods described in Yu and Hoff (2018). To obtain this prior information, we recommend specifying a linking model for both $\boldsymbol{\theta}$ and $\boldsymbol{D}$, possibly allowing for the presence of auxiliary covariates in the model for $\boldsymbol{D}$. As before, parameters of the linking model can be estimated and moment-matching used to obtain a normal distribution for θ_j and an inverse-gamma distribution for σ_j^2, which represent the indirect information about θ_j with which a FAB t-interval may be

constructed. We provide an empirical example of the FAB t-interval procedure in Section 2.4.

2.3 Simulation study

To compare the properties of FAB intervals and direct intervals, we constructed a simulation study in which area means may exhibit spatial autocorrelation and/or association with an explanatory variable. We aimed to quantify the reduction in expected interval width obtained via the FAB interval procedure relative to the direct interval procedure. Throughout the study, we assumed the sampling model $y_j \sim N(\theta_j, \sigma_j^2)$ with $\sigma_j^2 = 1$ known for all areas j, yielding the direct confidence interval $C_D^j = y_j \pm z_{1-\alpha/2}$.

Forty-nine areas were located on a 7×7 lattice. For each of 5000 datasets, we simulated area means under the following procedure:

1) Draw $u_j \sim U(0,1),\ j = 1, \ldots, m$

2) Set $x_j = \frac{u_j - \bar{u}}{s_u}, \qquad \bar{u} = \frac{1}{m}\sum_{j=1}^{m} u_j, \qquad s_u = \frac{1}{m-1}\sum_{j=1}^{m}(u_j - \bar{u})^2,$
   and $\boldsymbol{X} = (x_1, \cdots, x_m)^T$

3) Draw $\boldsymbol{\theta} \sim N\left(\boldsymbol{X}\beta, \tau^2[(\boldsymbol{I} - \rho\boldsymbol{W})(\boldsymbol{I} - \rho\boldsymbol{W}^T)]^{-1}\right)$

4) Draw $y_j \sim N(\theta_j, 1)$,

This data generating procedure was repeated eight times, one for each setting of $\rho \in \{0, 0.9\}$, $\tau^2 \in \{0.5, 5\}$ and $\beta \in \{0, 10\}$. In each repetition, the neighborhood matrix $\boldsymbol{W}$ was assumed to be a row standardized binary contiguity matrix (a binary contiguity matrix is defined such that the i, jth entry equals 1 if areas i and j border each other, and 0 otherwise).

2.3.1 Intervals with Area-Specific Coverage

For each area in a dataset, we constructed five types of 95% confidence intervals that have area-specific coverage. These consist of the direct interval and four different FAB intervals based on maximum likelihood estimation of linking models ranging in complexity. The linking models considered were

1) The exchangeable model: $\theta_j \sim N(0, \tau^2)$ independently across areas $j = 1, \cdots, m$.

2) The covariate model: $\theta_j \sim N(x_j\beta, \tau^2)$ independently across areas $j = 1, \ldots, m$.

3) The spatial model: $\boldsymbol{\theta} \sim N\left(\mathbf{0}, \tau^2[(\boldsymbol{I} - \rho\boldsymbol{W})(\boldsymbol{I} - \rho\boldsymbol{W}^T)]^{-1}\right)$.

4) The full model: $\boldsymbol{\theta} \sim N\left(\boldsymbol{X}\beta, \tau^2[(\boldsymbol{I} - \rho\boldsymbol{W})(\boldsymbol{I} - \rho\boldsymbol{W}^T)]^{-1}\right)$.

Under each data generating process, average interval lengths over all simulations for these five confidence interval procedures were calculated. Average lengths relative to the direct interval are given in Table 2.1 for each of the four FAB procedures. For the simulations in which the data were generated with strong spatial autocorrelation $\rho = 0.9$, the spatial FAB intervals outperformed their non-spatial counterparts in terms of average interval width. Similarly, the non-spatial FAB intervals are slightly narrower than their spatial counterparts when the data is generated without spatial autocorrelation due to the increased uncertainty that comes with estimating ρ. For lower values of the random effect variance τ^2, the FAB intervals are significantly narrower due to the increased precision of the available indirect information. When a covariate is a strong predictor of the area mean, FAB intervals estimated under a linking model with a covariate were narrower than those without a covariate. Most importantly, no matter the linking model, FAB intervals were narrower on average than intervals based on direct estimates alone. The percentage decrease in interval length ranged from 0.4% to 13.2%.

Table 2.1: Average confidence interval length relative to the direct interval by simulation, each with 5000 datasets. Since each column represents a separate data generating process, interval widths should only be compared across columns. FAB intervals are narrower on average than direct intervals and are narrower when the linking model appropriately models the data generating process.

	$\tau^2 = 1/2$				$\tau^2 = 5$			
	$\beta = 0$		$\beta = 10$		$\beta = 0$		$\beta = 10$	
Linking Model	$\rho = 0$	$\rho = 0.9$	$\rho = 0$	$\rho = 0.9$	$\rho = 0$	$\rho = 0.9$	$\rho = 0$	$\rho = 0.9$
Exchangeable	0.868	0.901	0.995	0.996	0.938	0.976	0.996	0.996
Covariate	0.869	0.901	0.869	0.901	0.939	0.977	0.939	0.976
Spatial	0.868	0.877	0.996	0.996	0.939	0.939	0.996	0.996
Full	0.869	0.878	0.869	0.878	0.940	0.940	0.940	0.940

It is important to note that a given FAB interval is not guaranteed to be narrower than the corresponding direct interval. Rather, FAB intervals will be narrower on average than direct intervals. Table 2.2 details the percentage of areas with shorter FAB intervals than direct intervals by simulation.

Table 2.2: Percentage of areas for which the FAB interval is narrower than the corresponding direct interval, by simulation. For a vast majority of the areas, any FAB interval will be narrower than the direct interval, regardless of the linking model chosen. However, there are more areas that demonstrate improvements when the linking model appropriately models the data generating process.

	$\tau^2 = 1/2$				$\tau^2 = 5$			
	$\beta = 0$		$\beta = 10$		$\beta = 0$		$\beta = 10$	
Linking Model	$\rho = 0$	$\rho = 0.9$	$\rho = 0$	$\rho = 0.9$	$\rho = 0$	$\rho = 0.9$	$\rho = 0$	$\rho = 0.9$
Exchangeable	96.7%	91.9%	81.3%	81.5%	86.8%	83.6%	81.7%	81.9%
Covariate	96.5%	91.4%	96.5%	91.5%	86.1%	82.7%	86.0%	82.6%
Spatial	96.6%	95.5%	79.2%	79.4%	85.9%	88.4%	79.6%	80.2%
Full	96.4%	95.2%	96.4%	95.2%	85.1%	87.5%	85.0%	87.5%

2.3.2 Comparison to Empirical Bayes

In addition to the direct interval and four FAB intervals, we also calculated empirical Bayes (EB) intervals based on the four linking models detailed above. Because empirical Bayes intervals are not constrained to have area-specific coverage, they are able to be narrower than FAB intervals, particularly when each area-level mean is well-predicted by the linking model (e.g., τ^2 is small). However, as shown in Table 2.3, empirical Bayes and FAB intervals have increasingly similar average widths as τ^2 increases.

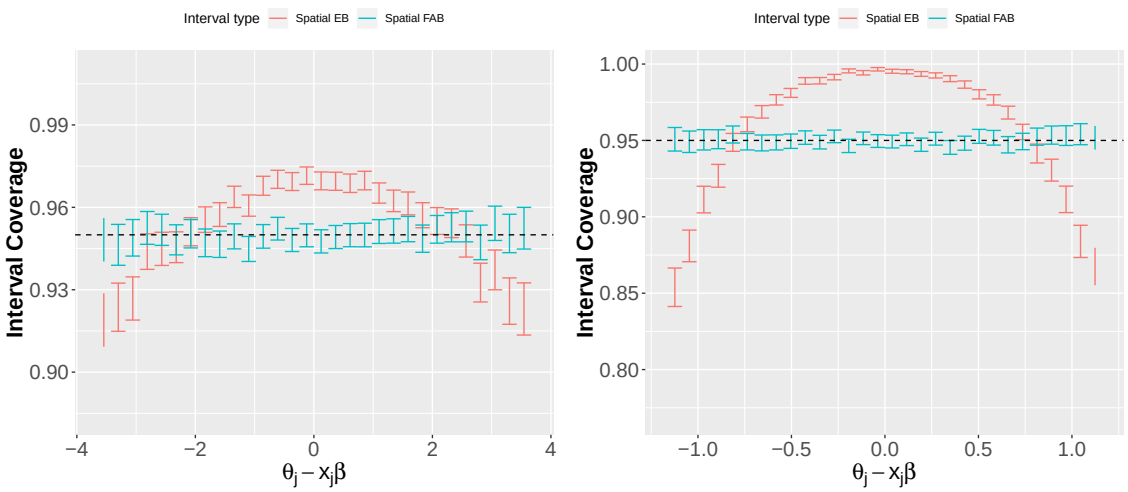
Table 2.3: Average lengths of FAB and empirical Bayes (EB) confidence intervals by simulation. In general, when the across-area heterogeneity is small, empirical Bayes intervals are able to be much narrower than FAB intervals.

		$\tau^2 = 1/2$				$\tau^2 = 5$			
		$\beta = 0$		$\beta = 10$		$\beta = 0$		$\beta = 10$	
Type	Linking Model	$\rho = 0$	$\rho = 0.9$	$\rho = 0$	$\rho = 0.9$	$\rho = 0$	$\rho = 0.9$	$\rho = 0$	$\rho = 0.9$
EB	Exchangeable	2.387	3.182	3.903	3.903	3.601	3.816	3.903	3.905
EB	Covariate	2.445	3.203	2.447	3.195	3.609	3.818	3.608	3.818
EB	Spatial	2.511	2.682	3.904	3.904	3.621	3.597	3.904	3.905
EB	Full	2.576	2.726	2.580	2.719	3.633	3.609	3.632	3.610
FAB	Exchangeable	3.402	3.530	3.902	3.902	3.679	3.826	3.903	3.905
FAB	Covariate	3.405	3.533	3.405	3.531	3.682	3.828	3.682	3.828
FAB	Spatial	3.403	3.440	3.903	3.903	3.682	3.681	3.904	3.905
FAB	Full	3.405	3.443	3.406	3.442	3.686	3.685	3.686	3.685

However, Table 2.3 does not tell the full story. Although the empirical Bayes confidence intervals approximately achieve $1 - \alpha$ coverage on average across areas, the area-specific coverage rate depends on the value of unknown target quantity θ_j. For values of θ_j that are close to their predicted means under the linking model, the EB interval has greater than $1 - \alpha$ coverage. For values much farther away, it has much less (Figure 2.3), since each EB interval is centered around a biased estimate of the target quantity. This pattern of area-specific coverage variation exhibited by Empirical Bayes intervals is evident in all simulations and linking model specifications. In fact, with the exception of two points, the frequentist coverage of the EB interval is unequal to $1 - \alpha$ for all values of θ_j. Unlike the EB interval, the FAB interval shares the property of constant coverage with the direct interval.

2.4 Empirical example: Household radon levels

Between 1987 and 1988, the U.S. Environmental Protection Agency collected household-level data on radon concentration as part of its State Residential Radon Survey (SRRS). The data consist of a stratified random sample of 12,777 homes, each located in one of 472 counties in nine different states. We examine a subset of the SRRS data, concentrating on four of the nine states in the study: Minnesota, Wisconsin, Michigan, and Indiana. These states are geographically close and demographically



(a) $\rho = 0$, $\beta = 10$, $\tau^2 = 5$ (b) $\rho = 0$, $\beta = 0$, $\tau^2 = 0.5$

FIGURE 2.3: Estimated coverage rate and 95% confidence interval for binned values of $\theta_j - \boldsymbol{x}_j^\top \boldsymbol{\beta}$. It is readily apparent that there are values of θ_j for which the empirical Bayes interval has far less than $1 - \alpha$ coverage. In contrast, the FAB interval has area-specific coverage.

similar to one another, so patterns of radon concentration may be similar across this region. Within these four states, there are 3,767 household measurements, located in 209 distinct counties. One of the primary goals of the study was to "provide the best estimate and uncertainty quantification of a county's true geometric mean of radon screening measurements" (Price et al., 1996).

Price et al. (1996) analyzed the subset of the SRRS data from the state of Minnesota, developing a linear mixed model to construct 95% Bayesian credible intervals for county-specific geometric mean radon levels. However, these intervals do not have 95% coverage at the county level. In particular, as outlined in Section 2.1, they will suffer from undercoverage for counties with exceptionally high or low true means. Such systematic undercoverage can be dangerous, because counties with extremely high radon levels present significant public health risks to their communities and need to be detected to necessitate appropriate policy action. Moreover, the link-

ing model may be misspecified, so Empirical Bayes intervals may not even maintain $1-\alpha$ frequentist coverage on average. For these reasons, it is desirable to use interval procedures that maintain known, constant county-specific coverage rates.

Of the 209 counties in the data, 124 of them have fewer than ten sampled households, and 72 have fewer than five sampled households. For these counties, direct confidence intervals will be extremely wide, which can limit their usefulness in practice. However, the average precision of these intervals can be improved by using FAB intervals to borrow information across counties. In this section, we compare direct intervals to several FAB interval procedures corresponding to different linking models for the county-specific means.

Following Price et al. (1996), we make a small empirical adjustment to the radon concentration values to mitigate the impact of very low concentration measurements that arise as a byproduct of measurement error. In particular, letting $\tilde{r}_{i,j}$ be the raw measured radon concentration for household i in county j, we define the adjusted log-radon concentration as

$$r_{i,j} = \frac{\tilde{r}_{i,j} + \sqrt{\tilde{r}_{i,j}^2 + 1}}{2}. \tag{2.20}$$

In addition, we also follow the authors in assuming that adjusted radon concentrations within counties follow a roughly log-normal distribution, which appears warranted by exploratory data analysis. Letting $y_{i,j} = \log r_{i,j}$, we assume the within-county sampling model $y_{1,j}, \ldots, y_{n_j,j} \sim N(\theta_j, \omega_j^2)$, where θ_j is the unknown true geometric mean radon concentration for county j and ω_j^2 is the unknown variance of log radon measurements in county j. Under the assumption of random sampling within counties, the county sample mean is distributed as $\bar{y}_j \sim N(\theta_j, \sigma_j^2)$, where $\sigma_j^2 = \omega_j^2/n_j$ is the variance of the sample mean. We define $\hat{\sigma}_j^2 = \hat{\omega}_j^2/n_j$, where $\hat{\omega}_j^2$ is the sample variance of log-radon measurements within county j. $\hat{\sigma}_j^2$ is an unbiased

and consistent estimate of σ_j^2.

We illustrate the use of FAB intervals for this small area analysis by considering several linking models for across-county heterogeneity, of which the most general is the spatial Fay-Herriot model. This model uses county-level surficial radium content (ppm), measured by the National Uranium Resource Evaluation (NURE), as an area-level predictor in a linear model. Under this model, $\theta_1, \cdots, \theta_m$ are jointly normally distributed, with $\mathrm{E}[\theta_j] = \mu + \beta_1 x_j$, where x_j is the measured surficial radium content for area j. $\mathrm{Cov}[\boldsymbol{\theta}]$ is defined as in (2.13), where the proximity matrix $\boldsymbol{W}$ represents the row-standardized squared exponential distance between county centroids (measured via longitude and latitude), since no counties in Minnesota and Wisconsin are first-order neighbors of counties in Michigan or Indiana. Explicitly,

$$W_{ij} = \frac{e^{-d_{ij}^2}}{\left(\sum_{j \neq i} e^{-d_{ij}^2}\right)} \qquad i \neq j, \tag{2.21}$$

where d_{ij} represents the distance between the centroids of county i and county j. The diagonal elements of $\boldsymbol{W}$ are equal to zero.

We also consider three simplifications of this model, corresponding to assumptions that either the regression coefficient $\beta_1 = 0$ and/or the spatial autocorrelation $\rho = 0$. Let the matrix $\boldsymbol{X}$ be an $m \times 2$ matrix consisting of a column of ones and the column vector $(x_1, \cdots, x_m)^\top$ and let $\boldsymbol{\beta} = (\mu, \beta_1)^\top$. Specifically, the four linking models examined are

1. full model: $\boldsymbol{\theta} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \tau^2[(\boldsymbol{I} - \rho\boldsymbol{W})(\boldsymbol{I} - \rho\boldsymbol{W}^T)]^{-1})$;
2. spatial model: $\boldsymbol{\theta} \sim N(\mathbf{1}\mu, \tau^2[(\boldsymbol{I} - \rho\boldsymbol{W})(\boldsymbol{I} - \rho\boldsymbol{W}^T)]^{-1})$;
3. covariate model: $\boldsymbol{\theta} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \tau^2\boldsymbol{I})$;
4. exchangeable model: $\boldsymbol{\theta} \sim N(\mathbf{1}\mu, \tau^2\boldsymbol{I})$.

Unlike in the simulation study, here we treat the county-level variance parameters σ_j^2 as unknown, resulting in t-intervals instead of z-intervals. We model the sampling variance parameters as $1/\omega_1^2, \ldots, 1/\omega_m^2 \sim$ i.i.d. $G(a, b)$ and estimate the hyperparameters a and b via marginal maximum likelihood. Details of this procedure are provided in A.4. Given estimates $\hat{a}$ and $\hat{b}$ based on data from other areas, we obtain prior information $1/\sigma_j^2 \sim IG(\hat{a}, n_j\hat{b})$ that is used to construct the FAB t-interval. For computational convenience, we obtain prior information for θ_j separately, using plug-in estimates $\hat{\sigma}_j^2$ when estimating $\{\mu, \boldsymbol{\beta}, \tau^2, \rho\}$. This procedure is analogous to that detailed in Sections 2.2 and A.

Because the county-level variances are unknown, we are able to construct confidence intervals with constant coverage for the 196 of the 209 counties with a sample size of at least two. For each of these counties, we construct FAB intervals for a specific county j under the four linking models specified above via the following process:

1. Estimate linking model parameters using data from all counties other than j.

2. Obtain prior distributions for both θ_j and σ_j^2 using plug-in estimates from the fitted linking model. This yields a normal distribution for θ_j and an inverse-gamma distribution for σ_j^2.

3. Obtain the optimal s-function for county j given prior information about θ_j and σ_j^2 (obtained using data not from j), as described in Section 2.2.

4. Construct the interval $\left\{\theta : \bar{y}_j + \hat{\sigma}_j t_{\alpha(1-s_j(\theta))} < \theta < \bar{y}_j + \hat{\sigma}_j t_{1-\alpha s_j(\theta)}\right\}$.

where the quantiles correspond to a those from a t-distribution with $n_j - 1$ degrees of freedom.

As visualized in Figure 2.4 and depicted numerically in Table 2.4, FAB intervals under each of the four linking models are significantly narrower than the direct in-

terval, representing a 23-26% improvement in average interval width. Incorporating a spatial linking model significantly reduces interval width, and including covariate information does not appear to have much of an impact on average interval width. Although a specific FAB interval is not guaranteed to be narrower than the corresponding direct interval, the vast majority of the FAB intervals represented improvements. The proportion of counties with narrower FAB intervals varied from 89 to 96 percent of the counties, depending on the quality of the chosen linking model.

Table 2.4: Average 95% confidence interval width, width ratio relative to the direct interval, and percentage of counties for which the FAB intervals are narrower than the direct intervals across the 196 Midwestern counties in the SRRS dataset.

Type	Linking Model	Mean Width	Rel. Width	% Narrower
Direct	-	1.701	1.000	-
FAB	Exchangeable	1.312	0.771	89.8%
FAB	Covariate	1.312	0.771	88.8%
FAB	Spatial	1.257	0.739	96.4%
FAB	Full	1.256	0.739	95.5%

In general, EB confidence intervals for county-specific radon levels are narrower than those constructed via the FAB procedure, although this is not always the case. Under the full linking model, the empirical Bayes interval is narrower than the corresponding FAB interval for 128 out of the 196 counties. The differences are most pronounced in the counties for which the combination of small sample size and high sampling variance is present (2.5). Regardless, EB intervals lack county-specific coverage, which limits their use in making county-specific inferences.

2.5 Discussion

In the field of small area analysis, researchers typically use confidence interval procedures that either have constant coverage across areas but do not share information, or utilize shared information but lack constant coverage. Although the empirical Bayes procedures commonly used in the literature have $1 - \alpha$ coverage on average

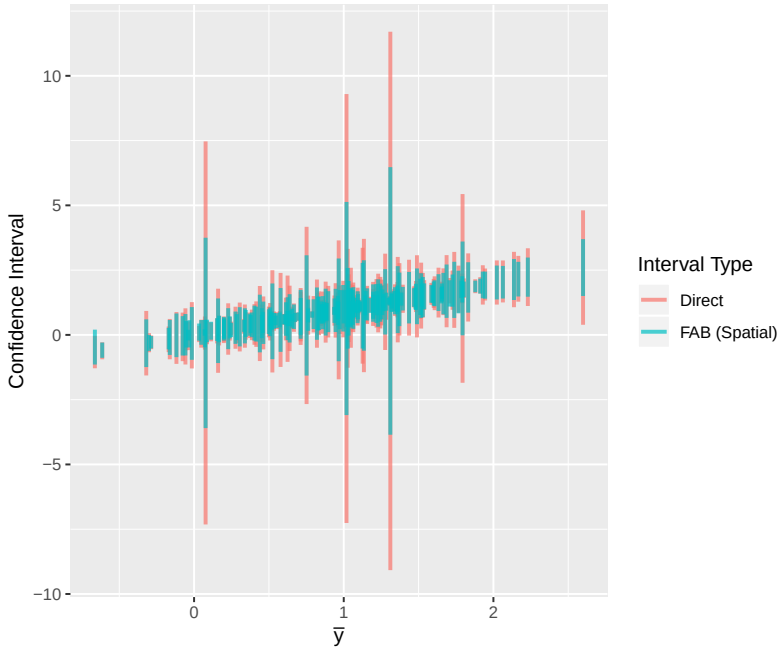


FIGURE 2.4: The FAB intervals based on the spatial Fay-Herriot model are substantially narrower than the direct intervals, on average across counties in the SRRS dataset.

across groups, the actual coverage rate may differ substantially for some values of θ_j, calling into question the resulting area-specific inferences. The FAB procedures developed by Yu and Hoff (2018) and outlined in this chapter have constant $1-\alpha$ coverage for each area regardless of what the true area-level means are, and are valid for all linear mixed models with normal sampling variances. This class of models is very flexible, enabling researchers accommodate auxiliary covariates, as well as spatial and temporal autocorrelation.

Importantly, although the empirical Bayes confidence interval procedure is guaranteed to have asymptotic $1-\alpha$ marginal coverage on average if and only if the linking model is true, the FAB procedure will always have $1-\alpha$ constant coverage, regardless of the chosen linking model. This is not to say that the linking model is unimportant; a properly specified linking model can substantially reduce expected FAB interval width, as evidenced by the simulation study and empirical example.

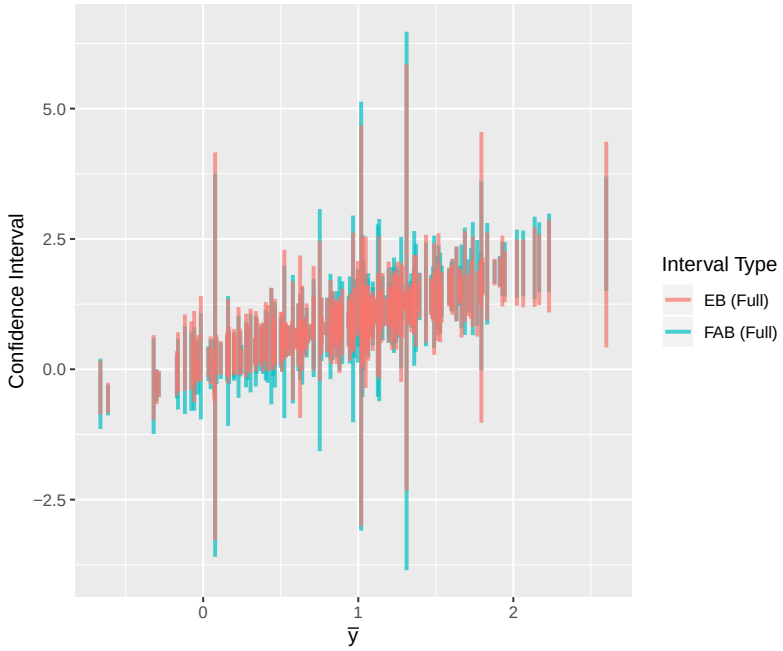


FIGURE 2.5: The EB intervals are generally narrower than FAB intervals, though this is not always the case. However, EB intervals lack area-specific coverage, which makes their use in area-specific inference unclear, particularly when the linking model is misspecified.

FAB intervals are somewhat more computationally demanding to calculate than direct confidence intervals or empirical Bayes intervals since m model estimations must occur to obtain confidence intervals for m areas. This can be burdensome under complex linking models, such as the spatial Fay-Herriot model, when the number of areas is large. Since FAB intervals will always have constant coverage, regardless of whether the linking model or the estimation procedure is correct, computational shortcuts can be taken to significantly reduce the burden, if necessary. For example, when the number of areas is large, one possibility is to separate the areas into k heterogeneous clusters and construct prior distributions for areas belonging to a given cluster based the direct estimates from other clusters. This means that only k models must be estimated, instead of m, resulting in computational gains. When the number of areas is prohibitively large, we recommend simply estimating the hy-

perparameters of the linking model once and then using those estimates to calculate all FAB intervals. Although this will violate the condition of independence necessary to guarantee $1 - \alpha$ area-specific coverage, the influence of a single area on model estimates is likely to be small in such a context, so the FAB intervals will have very close to $1 - \alpha$ coverage for all areas.

One area of future work is to extend the FAB procedure to generalized linear mixed models by constructing FAB intervals for target quantities when responses are discrete or categorical. This has significant applications in the small area estimation literature for applications such as disease mapping, where researchers are often interested in inferring area-level relative risks.

The FAB procedure for constructing confidence intervals with area-specific coverage can be implemented using a variety of software packages for estimating small area estimation models, such as `sae` (Molina and Marhuenda, 2015) or `lme4` (Bates et al., 2015). The only additional computational functionality needed is the Bayes optimal s-function, which we have implemented in R and made available in the `fabCI` R package on CRAN. Replication code is provided at `https://github.com/burrisk/fabci`. Data for the empirical example can be found at `http://www.stat.columbia.edu/~gelman/arm/examples/radon/`.

3

Bayesian Hot Deck Multiple Imputation for Multivariate Numeric Data

3.1 Introduction

Statistical organizations frequently collect datasets that contain missing values. Over the past few decades, multiple imputation methodology for statistical inference (Rubin, 1987) has become an increasingly popular way to handle missing values. This approach involves filling in the missing values multiple times using a probability model, analyzing each completed dataset using standard statistical methods, and then pooling inferences using well-known techniques. By generating multiple completed datasets, this procedure attempts to account for the additional uncertainty associated with the missing values so that confidence intervals approximately maintain nominal coverage (Reiter and Raghunathan, 2007).

Performing multiple imputation is challenging when the dataset of interest consists of mixed binary, discrete, ordered categorical, and continuous variables, each of which may have some values that are missing. Some multiple imputation methods treat all of these numeric variables as continuous, which may result in imputations

that are marginally infeasible. For example, if an individual's number of children is not treated as a discrete variable, a non-integer number of children may be imputed. This may impact the reliability of subsequent inference and/or reduce confidence in the quality of the imputed datasets. As such, the most common approaches for performing multiple imputation of numeric data tend to involve hot deck imputation, in which in which a missing value is replaced by an observed value from a similar observation. Filling in missing values in this way ensures that all imputed values are marginally realistic (Andridge and Little, 2010).

Predictive mean matching (Little, 1988) is arguably the most popular method of performing multiple imputation on multivariate numeric data, in no small part due to its default implementation in the `mice` R software package (van Buuren and Groothuis-Oudshoorn, 2011). Predictive mean matching relies on specifying a sequence of fully conditional univariate models (e.g., linear regression) to define a distance metric between observations. Missing values are then imputed by randomly sampling observed values from "similar" observations (Morris et al., 2014). Despite its popularity, the theoretical properties of predictive mean matching are unclear, as a series of full conditional distributions does not necessarily correspond to a valid joint distribution (Arnold et al., 2001).

The primary alternative to specifying a series of full conditional models involves constructing a joint model for the data. While more theoretically grounded, this approach has proven less popular due to the difficulty involved in specifying a reasonable model for mixed numeric data. Although the extended rank likelihood developed by Hoff (2007) alleviates some of this difficulty by modeling the marginal distributions of the variables nonparametrically, it imposes the assumption of linear dependence, which may be too restrictive in some contexts.

To address these deficiencies, we propose a new method for modeling multivariate numeric data of mixed types. Our approach is a latent variable model, in which

each observation is associated with an unobserved, latent variable. The proposed method generalizes that of Hoff (2007) by modeling the joint distribution of the latent variables far more flexibly. Estimating this model via a pseudo-Bayesian approach allows practitioners to obtain multiple imputed datasets where each imputed value is equal to at least one observed data value. In this way, our model produces hot deck imputations.

In addition, we demonstrate how our method can be adapted for jointly modeling multivariate numeric data when the support of the data is a strict subset of the data product space. As examples, a child cannot be older than his father and males cannot be pregnant. When this model is estimated using a pseudo-Bayesian approach, practitioners can obtain multiple hot deck imputations that are guaranteed to satisfy joint feasibility constraints.

In Section 3.2, we provide a brief review of multiple imputation and further motivate our proposed approach. We present our joint modeling approach in Section 3.3, extending it to constrained multiple imputation in Section 3.4. Section 3.5 demonstrates the approach in simulation studies, and Section 3.6 applies it to the analysis of American Community Survey microdata. A discussion follows in Section 3.7.

3.2 Literature Review

3.2.1 Multiple Imputation of Multivariate Numeric Data

Model-based multiple imputation methods can generally be classified into one of two categories: fully conditional models and joint models. Fully conditional modeling approaches involve specifying p full conditional distributions $p(y_j \mid y_{(-j)})$, which may also depend on model parameters. In order to obtain multiple imputed datasets, missing values are first filled in using a simple method. Then, for each variable with missing values, missing values are drawn from the corresponding full conditional

distribution, where the conditioning is on the observed and imputed values of the other variables (Raghunathan et al., 2000). The procedure iterates until apparent convergence.

Although it is possible to specify separate conditional models with each potentially having different support, predictive mean matching is by far the most popular way for generating multiple imputations of mixed numeric data. To impute a missing value for a variable using predictive mean matching, an imputer first defines the distance between observations to be the absolute difference in expectations of the variable under the conditional model. The imputer then chooses a set of "nearby" observations with observed values for the variable of interest and randomly samples one of them, substituting its observed value for the missing value. This enables the imputation procedure to be more robust to model misspecification (van Buuren, 2018). For example, predictive mean matching can use conditional linear models to impute binary data.

One of the main drawbacks of predictive mean matching is that there is no mathematical theory to justify its use. Important considerations such as the number of nearby observations to consider are generally arbitrary and justified by performance in simulation studies. As there is no guarantee that predictive mean matching even corresponds to a valid model for $p(\boldsymbol{Y}_{mis} \mid \boldsymbol{Y}_{obs})$, caution must be taken when interpreting subsequent inference.

In contrast to fully conditional models, joint modeling approaches assume a model $p(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis} \mid \boldsymbol{\theta})$. Bayesian procedures for estimating joint probability models provide a natural mechanism for imputing missing data values while also accounting for model uncertainty (Murray, 2018). In particular, under a Bayesian imputation procedure, imputations are generated from

$$p(\boldsymbol{Y}_{mis} \mid \boldsymbol{Y}_{obs}) = \int p(\boldsymbol{Y}_{mis} \mid \boldsymbol{Y}_{obs}, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{Y}_{obs}) d\boldsymbol{\theta} \tag{3.1}$$

Although the theoretical properties and assumptions of joint models are more transparent, it is far more challenging to develop a joint model for multivariate data than it is to develop a sequence of univariate models, particularly when the data contains a mixture of continuous, binary, ordinal, and discrete variables.

To make modeling the joint distribution of numeric data easier, Hoff (2007) proposed a semiparametric Gaussian copula model. This approach is a latent variable model, in which the observed data values are assumed to be transformations of a continuous underlying latent variable. The distribution of the latent variables is assumed to follow a multivariate normal distribution and the transformations are modeled nonparametrically. Model estimation is based on the *extended rank likelihood*, depending on only the ranks of the observed data values. One of the applications of this joint modeling approach is that it can be used to perform multiple imputation under the assumption that the data is missing at random. However, the assumption of multivariate normality of the latent variables imposes a restrictive assumption of linear dependence, which may not be realistic for complex survey datasets.

In sum, multiple imputation methods based on fully conditional models tend to be the predominant approach for handling data assumed to be missing at random, in no small part due to their flexibility. Although methods based on assuming a joint distribution for the data are more theoretically motivated and make their assumptions more explicit, their lack of flexibility hinders their use, particularly for mixed numeric data.

3.2.2 Multiple Imputation Quality

In many situations, the individuals who generate the imputed datasets and the individuals who analyze the completed datasets are not the same and do not frequently communicate. For example, the United States Census Bureau disseminates multiply imputed survey datasets to be used widely by the general public. When specifying

the imputation model, the researchers at the Census Bureau do not know the quantities that will be of interest to future researchers. Similarly, researchers who access the data are unlikely to know much about the method used to perform multiple imputation.

In this vein, when specifying an imputation model, it may be important to generate imputed datasets that not only yield approximately valid inference for a variety of target quantities, but also can be trusted by an outsider. For example, if a person in a publicly released American Community Survey dataset has a non-integer imputed number of children, confidence in the reliability of the imputed data is likely to be low, even if the imputation method yields asymptotically valid inference for the population mean number of children.

Both predictive mean matching and the extended rank likelihood approach of Hoff (2007) are examples of hot deck imputation procedures. In hot deck imputation, each missing value is replaced by an observed value from a similar observation. One of the main appeals of hot deck imputation is that all imputations are marginally plausible, since each potential donor value is observed at least once.

However, there is no guarantee that hot deck imputations are jointly plausible. For example, it is possible for a completed dataset to contain pregnant males or individuals who work full-time yet make zero income. While such infeasible imputations are uncommon under a well-specified imputation model, they can still occur so long as the model has support over the data product space. Because statistical agencies sometimes release imputed datasets to the public, it can be important to ensure that imputed observations satisfy joint feasibility constraints to avoid undermining public confidence in the disseminated data (Kim et al., 2015).

A variety of strategies for constrained multiple imputation have been proposed in the literature (De Waal, 2017). Many statistical agencies employ a two-step process, in which data values are first imputed via hot deck imputation and then minimally

adjusted to satisfy the constraints (Pannekoek and Zhang, 2015). Such procedures have been implemented in a number of software packages developed by statistical agencies, such as the Census Bureau's SPEER (Draper and Winkler, 1997) and Statistics Canada's GEIS (Kovar and Whitridge, 1990).

Because the theoretical properties of these methods are unclear, a variety of approaches begin by first specifying a multivariate distribution with support on the data product space, and then assuming that the data is constrained to lie in a defined feasible region (Geweke, 1991). For example, Tempelman (2007) assumes that the distribution of the data can be described by a constrained multivariate normal distribution. Kim et al. (2014) and Kim et al. (2015) provide additional flexibility, assuming that the distribution of the data can be well described by a constrained Dirichlet process mixture of multivariate normal distributions. Such an approach is only applicable for specific constraints involving continuous variables and cannot be used when some variables are discrete. To our knowledge, no approach has been proposed for performing constrained multiple imputation of mixed numeric data.

3.3 Joint hot deck multiple imputation

3.3.1 Model specification

In this section, we propose a flexible joint model for numeric data that generalizes that of Hoff (2007). Moreover, we propose a pseudo-Bayesian procedure for estimating this model. When this estimation procedure is used, multiple imputed datasets can be obtained under a missing-at-random assumption. As each imputed value is identical to at least one observed value, our proposed procedure produces hot deck imputations.

We assume that elements of the observed data vector $\boldsymbol{y}_i$ for observation i are elementwise monotonic transformations of a continuous latent vector $\boldsymbol{z}_i$, $i = 1, \cdots, n$.

We specify a transformation model in order to accommodate binary, ordinal, discrete, and continuous variables. In many contexts, ordinal variables can be treated as binned representations of underlying continuous variables (e.g., Likert scale). The latent vectors are modeled as independent samples from a probability distribution p, indexed by parameter vector $\boldsymbol{\theta}$. Mathematically,

$$\begin{aligned} \boldsymbol{z}_i &\sim p(\cdot \mid \boldsymbol{\theta}), & \boldsymbol{\theta} &\in \boldsymbol{\Theta}, \\ \boldsymbol{y}_i &= g(\boldsymbol{z}_i) & i &= 1, \cdots, n, \\ g(\boldsymbol{z}) &= (g_1(z_1), \cdots, g_p(z_p)) \end{aligned} \tag{3.2}$$

for a sequence of p monotone non-decreasing functions $\{g_j : \mathbb{R} \to \mathcal{Y}_j\}$, where $\mathcal{Y}_j$ is the marginal support of the jth numeric variable.

Such a model is highly general and related to a number of existing approaches for modeling multivariate data. Many of these approaches assume a multivariate normal model for the $\boldsymbol{z}_i$'s, letting $\boldsymbol{\theta}$ represent a p-dimensional mean vector and a $p \times p$ covariance matrix. Under this assumption, the model (3.2) corresponds to a multivariate probit model (Chib and Greenberg, 1998) when applied to binary data. Pitt et al. (2006) and Smith and Khaled (2012) attempt to model the transformations $\{g_j\}$ parametrically, while Hoff (2007) models the transformations nonparametrically.

The multivariate normal model for the latent variables can be restrictive in some contexts. Kottas et al. (2005) proposed a Dirichlet process mixture model for the latent variables, which is far more flexible. However, they assume that the transformations are known in advance or fixed prior to model estimation. Fixing the transformations not only forces an unnecessary degree of complexity in the latent model, but also precludes its use for performing hot deck imputation of continuous variables.

The Dirichlet process mixture model allows for a potentially infinite number of mixture components as $n \to \infty$. However, since sample sizes are finite, a truncated

Dirichlet process mixture model based on a large, but finite number of mixture components can serve as a reasonable approximation. As such, we choose to model p via a finite mixture of K multivariate normal distributions, indexed by parameters $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})$. The parameters $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K)$, $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_1, \cdots, \boldsymbol{\Sigma}_K)$, and $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_K)$ respectively represent the mean vectors, covariance matrices, and weights of the K mixture components. Under this model,

$$p(\boldsymbol{z}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k N(\boldsymbol{z}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \qquad i = 1, \cdots, n. \tag{3.3}$$

The finite mixture model can also be represented by introducing latent variables $H_i \in \{1, \cdots, K\}$ representing the component indicator of unit i, $i = 1, \cdots, n$. Under this representation,

$$\begin{aligned} H_i &\sim \text{Multinomial}(\boldsymbol{\pi}) \\ \boldsymbol{z}_i \mid \boldsymbol{\mu}_{H_i}, \boldsymbol{\Sigma}_{H_i} &\sim N(\boldsymbol{\mu}_{H_i}, \boldsymbol{\Sigma}_{H_i}), \qquad i = 1, \cdots n. \end{aligned} \tag{3.4}$$

Other than ensuring that each marginal transformation $g_j : \mathbb{R} \to \mathcal{Y}_j$ is non-decreasing, we impose no further assumptions on $\{g_j\}$. As a result, the model defined by (3.2) and (3.4) is a semiparametric transformation model, displayed graphically in Figure 3.1.

3.3.2 Model estimation

The semiparametric transformation model defined above is nonidentifiable. Instead of imposing restrictions on either the marginal transformations or finite mixture model parameters, we propose a two-stage estimation procedure. First, based on the observed data values, the empirical marginal CDFs of each of the p numeric variables are calculated and used to obtain plug-in estimates $\{\hat{g}_j\}$ of the marginal transformations. Then, assuming that the transformations are equal to their plug-in estimates, we perform Bayesian inference on the mixture model parameters $\boldsymbol{\theta}$

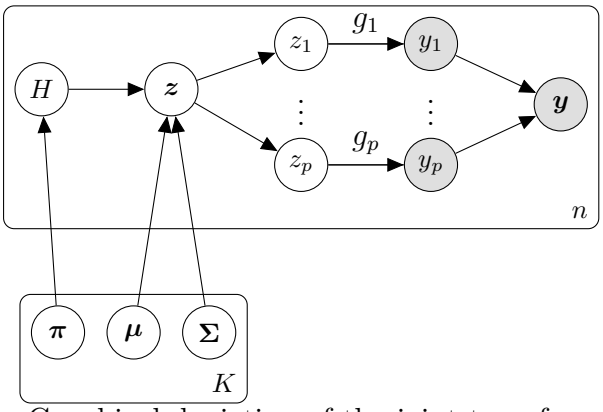


FIGURE 3.1: Graphical depiction of the joint transformation model

and the latent variable matrix $\boldsymbol{Z}$ after specifying a suitable prior distribution for $\boldsymbol{\theta}$. Approximate samples from the joint posterior distribution of $(\boldsymbol{\theta}, \boldsymbol{Z})$ can be combined with $\{\hat{g}_j\}$ to obtain multiple completed datasets that reflect uncertainty about both the missing values and model parameters.

Estimation of marginal transformations

To obtain plug-in estimates $\{\hat{g}_j\}$ after observing $\boldsymbol{Y}$, we first calculate the empirical marginal CDFs of the observed data values for each of the p numeric variables:

$$\hat{F}_j(y) = \frac{\sum_{i=1}^{n} I(y_{ij} \leqslant y, r_{ij} = 1)}{\sum_{i=1}^{n} r_{ij}}, \qquad j = 1, \cdots, p. \tag{3.5}$$

As $\hat{F}_j$ is not strictly increasing, it does not have an inverse function. Nevertheless, we define the empirical quantile function

$$\hat{Q}_j(u) = \inf_y \hat{F}_j(y) \leqslant u, \qquad j = 1, \cdots, p. \tag{3.6}$$

Lastly, we estimate the jth marginal transformation as

$$\hat{g}_j(z) = \hat{Q}_j(\Phi^{-1}(z)), \qquad j = 1, \cdots, p, \tag{3.7}$$

where Φ^{-1} is the inverse-cdf of the standard normal distribution. We then use the estimates $\{\hat{g}_j\}$ as plug-in estimates for $\{g_j\}$ when estimating $\boldsymbol{\theta}$.

This approach for estimating the marginal transformations has a number of practical advantages. Simultaneously estimating $\{g_j\}$ and $\boldsymbol{\theta}$ is very challenging, since given an observed data value y_{ij}, the transformation g_j and latent variable z_{ij} are highly related. Moreover, when the support of the data is not a product space (see Section 3.4), using plug-in estimates of the marginal transformations induces a known support for the latent variables ($\boldsymbol{y}_i \in \mathcal{A} \Leftrightarrow \boldsymbol{z}_i \in \mathcal{D}$, $i = 1, \cdots, n$).

In addition, under our definition of $\{\hat{g}_j\}$, the marginal support of each of the p numeric variables does not have to be known or specified by the imputer ahead of time. Any approach that either fixes g_j in advance (Kottas et al., 2005) or models it parametrically (Pitt et al., 2006; Smith and Khaled, 2012) must ensure that $g_j : \mathbb{R} \to \mathcal{Y}_j$, $j = 1, \cdots, p$, which can be cumbersome in high dimensions. Furthermore, defining $\{\hat{g}_j\}$ in this way reduces the burden placed on the components of the mixture model (see Section 3.5.1). In particular, if the data values are missing completely at random, $\hat{F}_j$ converges almost surely to the cumulative distribution function of the jth variable by the strong law of large numbers. When $\hat{F}_j$ is a good approximation to the jth empirical CDF, the induced distribution of the latent variables $(z_{1j}, \cdots, z_{nj})$ is approximately standard normal. As such, when the number of mixture components $K = 1$, the approach reduces to a Gaussian copula model with the marginal distributions assumed known and equal to the empirical margins of the observed data (Hoff, 2007).

Estimation of mixture model parameters

Given plug-in estimates $\{g_j\}$ of the marginal transformations, we now turn our attention to estimation of $\boldsymbol{\theta}$. In an attempt to quantify uncertainty about the parameters of the mixture model, we propose a procedure for performing approximate Bayesian inference on $\boldsymbol{\theta}$ using Markov Chain Monte Carlo (MCMC) techniques. To do so, we need to specify a prior distribution for $\boldsymbol{\theta}$.

In specifying priors for the component means $\boldsymbol{\mu}$ and covariance matrices $\boldsymbol{\Sigma}$, we follow Kim et al. (2014) in assuming a hierarchical prior:

$$\boldsymbol{\mu}_k \sim N(\boldsymbol{\mu}_0, h^{-1}\boldsymbol{I}). \tag{3.8}$$

$$\boldsymbol{\Sigma}_k \sim \text{Inverse-Wishart}(\nu, \boldsymbol{\Sigma}_0). \tag{3.9}$$

Assuming that our estimates $\{\hat{F}_j\}$ reasonably represent the marginal distributions of the variables in our dataset, we set the hyperparameters $\boldsymbol{\mu}_0 = \mathbf{0}$, $\nu = p + 5$, $h = 4/3$, and $\boldsymbol{\Sigma}_0 = \boldsymbol{I}$. Under these choices of hyperparameters, the latent variables z_{ij} are marginally standard normal with a balance of heterogeneity within classes and between classes.

We place a truncated stick breaking process prior on the component weights (Sethuraman, 1994; Ishwaran and James, 2001):

$$\pi_k = v_k \prod_{g<k}(1 - v_g). \tag{3.10}$$

$$v_k \sim \text{Beta}(1, \alpha), \qquad k \in \{1, \cdots, K-1\}. \tag{3.11}$$

$$v_K = 1. \tag{3.12}$$

$$\alpha \sim \text{Gamma}(a_\alpha, b_\alpha). \tag{3.13}$$

We set $a_\alpha = b_\alpha = 0.5$ to represent a vague prior specification for α. The stick breaking prior given in (3.10) - (3.13) encourages π_k to decrease as k increases. In particular, when α is small, most of the probability mass is contained in the first few mixture components.

Under the prior specification introduced above, a modified blocked Gibbs sampler (Ishwaran and James, 2001) can be used to obtain approximate samples of $\boldsymbol{\theta}$ from its posterior distribution, given $\boldsymbol{Y}_{obs}, \boldsymbol{R}, \{g_j\} = \{\hat{g}_j\}$. The computational details of how the parameters are updated in each iteration of the Gibbs sampler are provided

in Algorithm 1.

Given current values of model parameters;
for $k = 1, \cdots, K$ **do**
  Set $N_k = \sum_{i=1}^{n} I(H_i = k)$;
  Set $\bar{\boldsymbol{z}}_k = \sum_{i:H_i=k} \boldsymbol{z}_i / N_k$;
  Set $\boldsymbol{S}_k = \sum_{i:H_i=k} (\boldsymbol{z}_i - \bar{\boldsymbol{z}}_k)(\boldsymbol{z}_i - \bar{\boldsymbol{z}}_k)^\top$;
  Set $\boldsymbol{\Omega}_k = (N_k \boldsymbol{\Sigma}_k^{-1} + h\boldsymbol{I})^{-1}$;
  Sample $\boldsymbol{\mu}_k \sim N\left(\boldsymbol{\Omega}_k(N_k \boldsymbol{\Sigma}_k^{-1} \bar{\boldsymbol{z}} + h\boldsymbol{\mu}_0), \boldsymbol{\Omega}_k\right)$;
  Sample $\boldsymbol{\Sigma}_k \sim$ Inverse-Wishart$(\nu + N_k, \boldsymbol{\Sigma}_0 + \boldsymbol{S}_k)$;
  if $k \neq K$ **then**
    Sample $v_k \sim \text{Beta}(1 + N_k, \alpha + \sum_{\ell > k} N_\ell)$;
  else
    Set $v_k = 1$;
  end
  Set $\pi_k = v_k \prod_{\ell < k} (1 - v_k)$;
end
Sample $\alpha \sim \text{Gamma}(a_\alpha + K - 1, b_\alpha - \log \pi_k)$;
for $i = 1, \cdots, n$ **do**
  for $k = 1, \cdots, K$ **do**
    Set $\phi_{ik} = \pi_k p(\boldsymbol{z}_i \mid H_i = k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$;
  end
  Sample $H_i \sim \text{Multinomial}\left(\frac{\phi_{i1}}{\sum_{k=1}^K \phi_{ik}}, \cdots, \frac{\phi_{iK}}{\sum_{k=1}^K \phi_{ik}}\right)$;
end
for $i = 1, \cdots, n$ **do**
  for $j = 1, \cdots, p$ **do**
    Set m_{ij} equal to the conditional mean of z_{ij}, given $z_{i,-j}$, $\boldsymbol{\mu}_{H_i}$, $\boldsymbol{\Sigma}_{H_i}$;
    Set σ_{ij}^2 equal to the conditional variance of z_{ij}, given $z_{i,-j}$, $\boldsymbol{\Sigma}_{H_i}$;
    if $r_{ij} = 1$ **then**
      Set $a = \inf_z \hat{g}_j(z) = y_{ij}$;
      Set $b = \sup_z \hat{g}_j(z) = y_{ij}$;
      Sample $z_{ij} \sim \text{TN}(m_{ij}, \sigma_{ij}^2, (a, b))$;
    else
      Sample $z_{ij} \sim N(m_{ij}, \sigma_{ij}^2)$;
      Set $y_{ij} = \hat{g}_j(z_{ij})$;
    end
  end
end

Algorithm 1: Gibbs sampling iteration to update parameters for the joint transformation model

At a high level, in each iteration of the sampler we update the latent variables z_{ij} one at a time by first finding the univariate normal distribution that results from conditioning on the other coordinates. If y_{ij} is observed, we sample from this univariate distribution, truncated to the interval $\{z : \hat{g}_j(z) = y_{ij}\}$. If instead y_{ij} is missing, we sample z_{ij} from the unconstrained normal distribution and set $y_{ij} = \hat{g}_j(z_{ij})$. Given the values of the latent variables, updates for the other parameters in the model can be obtained using standard procedures. See, for example, Kim et al. (2014) or Murray and Reiter (2016).

After a suitable burn-in period, The Gibbs sampler outlined above can iterate T times to obtain T correlated samples from the posterior distribution of the model parameters, given the observed data and plug-in estimates of the marginal transformations. We then select M samples from the posterior distribution spaced far enough apart to be treated as approximately independent. The samples of the missing values of $\boldsymbol{Y}$ can be treated as multiple imputed datasets. Based on our specification of $\{\hat{g}_j\}$ an imputed values for a variable is identical to at least one observed value for that variable. As a result, estimation of the model above yields hot deck imputations that correspond to a valid joint model for the data.

3.4 Constrained hot deck multiple imputation

The approach outlined in the previous section produces imputations that are marginally feasible, but not necessarily jointly feasible. To extend the approach developed in Section 3.3 for jointly feasible hot deck multiple imputation, we adapt the joint transformation model so that each $\boldsymbol{y}_i$ has support on $\mathcal{A}$, a strict subset of the product space $\mathcal{Y}_1 \times \cdots, \times \mathcal{Y}_p$. Specifically, given parameters $(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ and transformations $\{g_j\}$, we consider the alternative model

$$
\begin{aligned}
g(\boldsymbol{z}) &= (g_1(z_1), \cdots, g_p(z_p)) \\
\mathcal{D} &= \{\boldsymbol{z} : g(\boldsymbol{z}) \in \mathcal{A}\} \\
p(\boldsymbol{z}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) &= \frac{\sum_{k=1}^{K} \pi_k N(\boldsymbol{z}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) I(\boldsymbol{z}_i \in \mathcal{D})}{\mathcal{Z}(\mathcal{D}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})}, \qquad i = 1, \cdots, n. \\
\mathcal{Z}(\mathcal{D}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) &= \int_{\mathcal{D}} \sum_{k=1}^{K} \pi_k N(\boldsymbol{z}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) d\boldsymbol{z}_i \\
\boldsymbol{y}_i &= g(\boldsymbol{z}_i).
\end{aligned}
\tag{3.14}
$$

The departure from the previous model is that each $\boldsymbol{z}_i$ now has support on the set $\mathcal{D}$, a strict subset of $\mathbb{R}^p$.

Estimating the parameters of this model is far more challenging than for the model described in Section 3.3, as $\mathcal{D}$ is unknown. The difficulty surrounding the unknown support of the latent variables can be resolved if the marginal transformations are assumed equal to the plug-in estimates calculated via (3.5) - (3.7). However, even if the marginal transformations and $\mathcal{D}$ are known, performing Bayesian inference on the mixture model parameters is still not straightforward. This is because exact evaluation of $p(\boldsymbol{z}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})$ involves the intractable normalizing constant $\mathcal{Z}(\mathcal{D}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})$.

Using the plug-in estimates $\{\hat{g}_j\}$ and the priors for $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})$ outlined in Section 3.3, we propose a data augmentation strategy for performing Gibbs sampling updates of $\boldsymbol{\theta}$, given values of $\boldsymbol{z}_i$, $i = 1, \cdots, n$ (Rao et al., 2016). Define

$$
q(\boldsymbol{z}_i \mid \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k N(\boldsymbol{z}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{3.15}
$$

such that

$$
p(\boldsymbol{z}_i \mid \boldsymbol{\theta}) = \frac{q(\boldsymbol{z}_i \mid \boldsymbol{\theta}) I(\boldsymbol{z}_i \in \mathcal{D})}{\mathcal{Z}(\mathcal{D}, \boldsymbol{\theta})}. \tag{3.16}
$$

It is clear from (3.16) that $p(\boldsymbol{z}_i \mid \boldsymbol{\theta})$ has an associated rejection sampling algorithm (Robert and Casella, 2005). In particular, we can obtain samples from p by repeatedly sampling from q, rejecting all samples that lie outside the set $\mathcal{D}$.

Typically, the set of rejected proposals $\{\boldsymbol{z}_i^c\}$ are simply discarded. However, by instantiating the rejected proposals, it becomes straightforward to perform Bayesian inference on $\boldsymbol{\theta}$. To see this, let $W_i = |\{\boldsymbol{z}_i^c\}|$, the number of rejected samples between the $(i-1)$th accepted sample and the ith accepted sample. Note that W_i is an independent geometrically distributed random variable with probability of success $\mathcal{Z}(\mathcal{D}, \boldsymbol{\theta})$. The joint distribution of $\boldsymbol{z}_i$, W_i, and $\{\boldsymbol{z}_i^c\}$ is given by

$$p(\boldsymbol{z}_i, W_i, \{\boldsymbol{z}_i^c\} \mid \boldsymbol{\theta}) = q(\boldsymbol{z}_i \mid \boldsymbol{\theta}) I(\boldsymbol{z}_i \in \mathcal{D}) \prod_{w=1}^{W_i} q(\boldsymbol{z}_{i(w)}^c \mid \boldsymbol{\theta}) I(\boldsymbol{z}_{i(w)} \notin \mathcal{D}), \tag{3.17}$$

which does not involve the intractable normalizing constant $\mathcal{Z}(\cdot)$. This suggests an approach that generates from $q(\cdot)$ until n acceptances have been achieved, storing the rejected proposals, and then updating $\boldsymbol{\theta}$ given an augmented dataset consisting of $\boldsymbol{z}_i$ and $\{\boldsymbol{z}_i^c\}$, $i = 1, \cdots, n$.

This approach has been used for performing multiple imputation when the data values are subject to inequality constraints (Kim et al., 2014, 2015) and structural zeroes (Manrique-Vallier and Reiter, 2017; Akande et al., 2018). However, these authors consistently note that this can be computationally expensive, particularly when a mixture model is utilized as the imputation engine. Specifically, the number of rejected proposals prior to achieving n acceptances can grow quite large. Although Akande et al. (2018) proposes approximations that involve either capping the number of rejected proposals or weighting a fewer number of rejected proposals, these methods tend to introduce non-negligible approximation error.

We suggest an alternative approach that involves generating rejected proposals $\{\boldsymbol{z}_i^c\}$ conditional on the component indicator H_i. Conditional on $H_i = k$, the joint

distribution of $\boldsymbol{z}_i$ is a constrained multivariate normal distribution

$$\begin{aligned} p(\boldsymbol{z}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}, H_i = k) &= \frac{N(\boldsymbol{z}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) I(\boldsymbol{z}_i \in \mathcal{D})}{\mathcal{Z}_k(\mathcal{D}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}. \\ \mathcal{Z}_k(\mathcal{D}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) &= \int N(\boldsymbol{z}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) I(\boldsymbol{z}_i \in \mathcal{D}) \, d\boldsymbol{z}_i. \end{aligned} \tag{3.18}$$

Such a formulation is equivalent to expressing the model for $\boldsymbol{z}_i$ in terms of a mixture of constrained multivariate normal distributions, rather than a constrained mixture of multivariate normal distributions. To see why this may be advantageous, note that

$$p(\boldsymbol{z}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \frac{\pi_k \mathcal{Z}_k(\cdot) p(\boldsymbol{z}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, H_i = k)}{\left(\sum_{k=1}^{K} \pi_k \mathcal{Z}_k(\cdot) \right)}. \tag{3.19}$$

When $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are such that significant probability mass is placed outside the feasible region $\mathcal{A}$, the normalizing constant $\mathcal{Z}(\cdot)$ is small. Using weights proportional to $\pi_k \mathcal{Z}(\cdot)$ rather than π_k places less weight on components with significant infeasible probability mass, resulting in fewer observations with rejected proposals. As such, capping the number of rejected proposals for each observation introduces non-negligible approximation error only when estimating the parameters of components with low probability mass. This is an improvement over the approximation proposed in Akande et al. (2018), where inferences on all mixture components were affected by limiting the number of rejected proposals.

As such, we can express the conditional distribution

$$\begin{aligned} p(\boldsymbol{z}_i, W_i, \{\boldsymbol{z}_i^c\} \mid \cdot, H_i = k) &= N(\boldsymbol{z}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) I(\boldsymbol{z}_i \in \mathcal{D}) \\ &\quad \times \prod_{w=1}^{W_i} N(\boldsymbol{z}_{i(w)}^c \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) I(\boldsymbol{z}_{i(r)}^c \notin \mathcal{D}), \end{aligned} \tag{3.20}$$

which involves no intractable normalizing constants. Based upon these results, we can develop a Gibbs sampling algorithm for sampling from the posterior distribution

of the mixture model parameters and subsequently generating imputations. The full details of the procedure are provided in Algorithm 3.

To reduce the computational burden of the sampler, the following modifications can be made to Algorithm 3. First, the latent variables corresponding to unobserved data values can be updated one at a time, rather than jointly. This avoids the additional rejection sampling step involved with sampling these variables at the possible expense of slower mixing. In addition, a cap on the number of rejected proposals for a given observation may be considered, as occasionally there can be one or two observations near the constraint boundary for which the number of rejected proposals can be unacceptably large. It is important to note that the corresponding Markov chain does not have the target stationary distribution if a cap is imposed. Nevertheless, we have found that similar results can be attained after imposing a cap, with significant computational gains.

3.5 Simulation Studies

In this section, we perform three simulation studies in an effort to further understand the properties of the proposed joint and provide insight about its performance relative to existing methods. The first compares our proposed method of obtaining plug-in estimates with an alternative approach that attempts to mimic the procedure of Kottas et al. (2005). The second compares the performance of the model described in Section 3.3 with predictive mean matching and the extended rank likelihood under repeated sampling from an actual dataset. The third simulates data with support on a non-product space and compares the constrained model developed in Section 3.4 with the three methods examined the second simulation study.

For each of the simulation studies and empirical examples that follow, we assume a maximum of $K = 15$ mixture components. For each of the Bayesian imputation procedures used in the second and third simulation studies, we use a burn-in period of

100 iterations and sample for 1000 iterations, generating a total of $M = 10$ imputed datasets.

3.5.1 Plug-in estimation procedures

In this study, we compare two different methods of obtaining plug-in estimates for the transformations $\{\hat{g}_j\}$. Suppose that n_j unique values of the jth numeric variable are observed. Let $y_{(\ell),j}$ correspond to the ℓth smallest observed value of the jth variable, $\ell = 1, \cdots, n_j$. The first plug-in estimate uses unit-spaced cutoffs, similar to those used by Kottas et al. (2005), such that

$$\hat{g}_j(z) = \begin{cases} y_{(1),j} & z \leqslant -(\ell-1)/2 \\ y_{(2),j} & -(\ell-1)/2 < z \leqslant -(\ell-1)/2 + 1 \\ \vdots & \\ y_{(\ell),j} & z > (\ell-1)/2 \end{cases} \tag{3.21}$$

We refer to the plug-in estimation procedure given by (3.21) as the unit-spaced procedure transformations and the procedure given by equations (3.5) - (3.7) as the empirical CDF procedure. For each of these choices of plug-in estimation procedures, we use the simulated data to estimate the joint transformation model described in Section 3.3. Because we have less prior information about the cluster location and scale parameters when using the unit-spaced spaced procedure, we specify weakly informative priors for the mixture model parameters when this procedure is used, setting $h = 0.0001$ and $\nu = 3$.

We assume a data generating process for two variables that are marginally Poisson and Chi-Square respectively, tied together via a Gaussian with a moderate degree of dependence ($\rho = 0.5$). As such, we simulate a dataset of 1,000 observations of two variables as follows:

Here, F_1^{-1} is the pseudo-inverse CDF of a Poisson distribution with mean equal

for $i \in \{1, \cdots, n\}$ **do**
  Draw $\boldsymbol{z}_i \sim N(\boldsymbol{0}, \rho \boldsymbol{1}\boldsymbol{1}^\top + (1-\rho)\boldsymbol{I})$;
  Set $y_{i1} = F_1^{-1}(\Phi(z_{i1}) \mid \lambda)$, $y_{i2} = F_2^{-1}(\Phi(z_{i2}) \mid a, b)$;
  Set $\boldsymbol{y}_i = (y_{i1}, y_{i2})$
end
Set $\boldsymbol{Y} = (\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n)^\top$;

Algorithm 2: Bivariate data generating process for the simulation study

to 40 and F_2^{-1} is the pseudo-inverse CDF of a Gamma distribution with shape equal to 1 and rate equal to 0.01. This data generating process results in a mixed dataset, with substantial skew in the second variable. A bivariate scatterplot of the simulated data is provided in Figure 3.2.

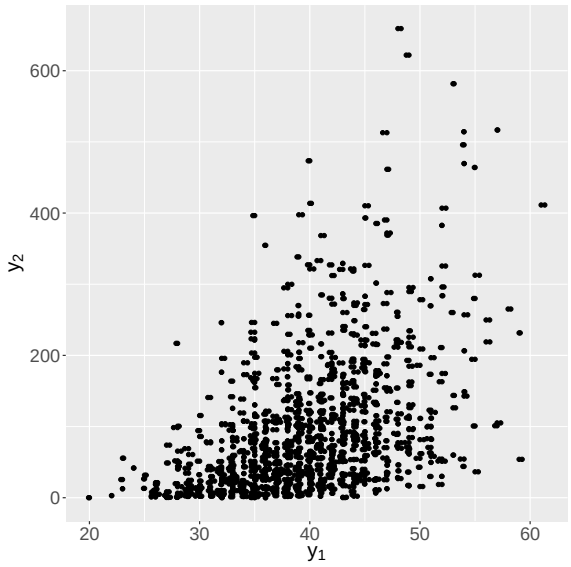


FIGURE 3.2: Simulated bivariate data

Under each procedure, approximate samples from the posterior distribution of the finite mixture model parameters were obtained via Gibbs sampling, ran for 1,000 iterations after a burn-in of 100 iterations. To assess model performance, we examined the samples of the mixture model parameters, and simulated $1,000$ observations from the posterior predictive distribution under each model and also recorded their cluster memberships. In general, a joint model performs well when the distribution

of observations simulated from the posterior predictive distribution is similar to the distribution of the observed data (Gelman et al., 2014).

Looking at Figure 3.3, we see that the joint model with unit-spaced transformations places too much probability mass on the minimum and maximum observed values of the variables. Moreover, the unit-spaced transformation model places high probability mass on three clusters.

In contrast, the joint model with empirical transformations appears to fit the data quite well. Very high probability mass is placed on the first cluster, which is not only consistent with the data generating process, but also more interpretable. From this simulation, it appears that using the empirical transformations of the observed data can result in a better fitting, more interpretable model than other methods of obtain plug-in estimates $\{\hat{g}_j\}$. In particular, fixing the transformations prior to model estimation appears to force additional complexity in the latent variable model, resulting in a higher number of mixture components that are needed to adequately capture the joint distribution.

3.5.2 Repeated sampling study

To assess the performance of our proposed method under repeated sampling compared to existing methods for performing multiple imputation of numeric data, we consider a dataset of 53,940 round-cut diamonds, available in the `ggplot2` R package (Wickham, 2016). For each diamond in the dataset, a total of ten continuous and ordinal variables are recorded, for which five of the most relevant are summarized in Table 3.1.

Table 3.1: Description of five variables included in the diamonds dataset

Variable	Description
Carat	Size of the diamond
Cut	Quality of the cut, ranging from Fair to Ideal
Color	Color of the diamond, ranging from D (colorless) to J (near colorless)
Clarity	Clarity of the diamond, ranging from I1 (included) to IF (flawless)
Price	Price of the diamond, in U.S. dollars

Treating this dataset as our population of interest, we generate simple random samples of size 1000. We repeat this procedure to obtain 500 distinct samples of size 1000. For each of the datasets, we introduce missingness under a missing at random mechanism. In particular, letting y_{i1} denote the size of diamonds i in carats, we set

$$\Pr(r_{ij} = 0) = \frac{1}{1 + \exp(y_{i1})}, \qquad j = 2, \cdots 10. \tag{3.22}$$

Under this response mechanism, the size of each diamond is always observed, but the values of the other variables may be unobserved. Missing values tend to be more prevalent for smaller diamonds.

We then impute these values using three distinct imputation procedures: predictive mean matching as implemented in `mice` (van Buuren and Groothuis-Oudshoorn, 2011), the semiparametric Gaussian copula model in the `sbgcop` R package (Hoff, 2018), and the model proposed in Section 3.4. For brevity, we refer to these procedures via the acronyms PMM, ERL, and JTM respectively. The sampling procedure for the extended rank likelihood failed for four of the 500 datasets, and we based our assessments of its performance by aggregating results for the 496 datasets for which the algorithm succeeded.

To assess the performance of these imputation procedures, we consider a linear regression model for log price, with carat, color, cut, and clarity as covariates. We include main effects terms for each variable as well as all two-way interaction terms that involve carat. In all, a total of 36 regression coefficients are calculated on

the basis of the full dataset. For each of the $M = 10$ datasets imputed under the three methods, we use multiple imputation combining rules to obtain point estimates and standard errors of the regression coefficients of interest. We then compare the inferences made under each of the imputation models in terms of absolute raw bias and confidence interval coverage.

Based on Figure 3.4, inferential performance based on datasets imputed via JTM was significantly better than that based on datasets imputed via PMM. In particular, for the same target quantities, inference based on the JTM imputations had lower absolute raw bias and higher confidence interval coverage. Both methods did not maintain nominal coverage for the quantities involving the clarity term. Despite this, the undercoverage for the JTM imputation engine was less severe.

The Gaussian copula model estimated via the extended rank likelihood and the joint transformation model performed very similarly, with ERL generating slightly narrower uncertainty intervals. This is not surprising, particularly because the joint transformation model tends to place very high weight on exactly one mixture component. This provides evidence that the dependence structure of the numeric variables in the diamonds dataset is nearly linear.

3.5.3 Imputation under feasibility constraints

To assess whether the constrained joint imputation model can present inferential advantages, we perform a simulation study involving three variables: one binary, one discrete, and one continuous. We begin by independently simulating a population of 100,000 observations under the following data generating process:

$$
\begin{aligned}
&X_1 \sim \text{Bernoulli}(p = 0.5), \qquad X_2 \sim \text{Pois}(\lambda = 4), \qquad X_3 \sim \text{Gamma}(1, 0.2) \\
&Y_1 = X_1 \\
&Y_2 = 3X_1 + X_2 \\
&Y_3 = 3X_1 + X_2 + X_3
\end{aligned}
\tag{3.23}
$$

Under this data generating process, Y_3 must be greater than Y_2 and Y_2 must be greater than or equal to 3 if $Y_1 = 1$.

We create 500 independent simple random samples from this population, each with sample size equal to 100. In order to present a strong challenge for the imputation procedures, we randomly blank half of the values in each of the samples completely at random. Once the missing values are introduced, we perform multiple imputation under PMM, ERL, JTM, and the constrained joint transformation model (CJTM) presented in Section 3.4, where joint feasibility constraints of the three variables are included in the probability model for the data.

Table 3.2 provides the marginal proportion of imputed observations that satisfy the joint feasibility constraints. As the CJTM accounts for the joint feasibility constraints, all of the imputed values are jointly feasible. Although the difference in the overall proportion of infeasible imputations between JTM and PMM is less than 1%, the difference is statistically significant based on a paired hypothesis test ($p \approx 0.01$).

Table 3.2: Proportion of imputed observations that satisfy the feasibility constraints 1) $y_3 > y_2$ and 2) $y_2 \geqslant 3$ if $y_1 = 1$.

Method	% Feasible
PMM	90.8%
ERL	90.6%
JTM	91.3%
CJTM	100.0%

Furthermore, consider the estimation of the regression coefficients of the model

$Y_3 = \beta_0 + \beta_1 X_1 + \beta_2 X_2$. Based on the data generating process, the population values are $\beta_0 = 5$, $\beta_1 = 0$, and $\beta_2 = 1$. From Table 3.3 and Figure 3.5, we see that the inference based on the CJTM yields higher confidence interval coverage rates and lower raw bias than the other methods. Note that the nominal coverage rate is 95%.

Table 3.3: 95% confidence interval coverage rates for model parameters in the simulation study. Inference based on the CJTM appears to have overcoverage, whereas regression coefficient inference based on PMM suffers from undercoverage.

Method	β_0	β_1	β_2
PMM	94.6%	91.8%	92.8%
ERL	94.4%	95.4%	96.2%
JTM	94.4%	96.0%	95.4%
CJTM	97.4%	97.0%	97.2%

Although the CJTM exhibited superior inferential performance than the other methods considered, this simulation study is a rather extreme example. Taking the support of the data into account when estimating the joint model tends to be most useful when the sample size is small and the proportion of missing values is high. In this example, the extended rank likelihood performs better than the unconstrained joint transformation model and predictive mean matching. This is because the data generating process created linear dependence among the variables. Predictive mean matching yields higher absolute raw bias, wider confidence intervals, and lower confidence interval coverage rates than the joint models.

3.6 Empirical Example

3.6.1 American Community Survey microdata

The American Community Survey (ACS) is a survey administered by the United States Census Bureau to collect information about households, individuals, and the workforce in the United States. We consider a 0.1% public-use subsample of the 2017 ACS for individuals in the state of Alabama. The ACS survey design is highly

complex, involving survey weights and replicates. For simplicity, we assume that all individuals in our subsample have identical weights, though all methods considered can accommodate survey weights in post-imputation inference.

For this analysis, we consider a total of seven variables for each individual. After preprocessing, there are three binary variables (Medicare, Medicaid, Disability), three numeric variables (Age, Income, Hours Worked), and one ordered categorical variable (Education), summarized in Table 3.4. A value of 0 for education means that the individual is between 0 and 3 years old and has never attended school, whereas a value of 24 for education means that the individual has obtained a doctoral degree.

Table 3.4: Description of variables selected from the ACS data

Variable	Description
Age	Age of the individual, in years
Medicare	Does the individual receive Medicare?
Medicaid	Does the individual receive Medicaid?
Disability	Is the individual disabled?
Education	Education level (0-24)
Income	Annual income, in US dollars
Hours Worked	Number of hours worked per week

There are some combinations of values for two or more of these variables that are impossible for an individual to attain. Impossible combinations of values violate one or more feasibility constraints, which we enumerate below. A handful of observations do not satisfy these constraints (possibly due to the imputation methods employed by the Census Bureau), and we remove them from the dataset so that all observations in the dataset are feasible.

1) Education constraints

- Individuals under the age of three cannot attend kindergarten or above.

- Individuals cannot be under 12 years old and have graduated from high school.
- Individuals cannot be under 16 years old and have graduated from college
- Individuals under the age of 20 cannot have a doctorate.

2) Medicare eligibility

- An individual can only receive Medicare if he/she is either at least 65 years old or disabled.

3) Medicaid eligibility

- An individual cannot receive Medicaid and also make more than 100,000 dollars per year.

4) Minimum wage

- Individuals must have a positive income if they are employed.

5) Child labor laws

- Individuals under the age of 14 must work zero hours per week.
- Individuals under the age of 18 cannot work more than 40 hours per week.

We delete 20% of the values completely at random in order to create a strong challenge for each imputation model. We proceed to generate $M = 10$ imputed datasets using PMM, ERL, JTM, and CJTM. Only the latter approach guarantees that imputed observations satisfy feasibility constraints.

Table 3.5 provides the average proportion of observations in the imputed datasets that satisfy feasibility constraints, under each of the four imputation methods. Under PMM, ERL, and JTM, each of the five types of constraints are occasionally violated.

Interestingly, the joint transformation model proposed in Section 3.3 generates fewer infeasible imputed observations than PMM or ERL, perhaps indicating that our proposed model is better able to capture the joint distribution of the observed data. As expected, the constrained joint transformation model ensures that all feasible observations satisfy feasibility constraints.

Table 3.5: Percentage of imputed American Community Survey responses that satisfy constraints. Both PMM, ERL, and JTM consistently imputes a small proportion of infeasible observations.

Constraint Type	PMM	ERL	JTM	CJTM
All	96.6%	94.3%	98.5%	100.0%
Education	99.5%	97.9%	99.8%	100.0%
Medicare Eligibility	98.9%	98.8%	99.4%	100.0%
Medicaid Eligibility	99.9%	99.9%	99.9%	100.0%
Minimum Wage	98.2%	97.9%	99.3%	100.0%
Child Labor	99.6%	98.0%	99.8%	100.0%

In addition, after performing imputation under each of the four approaches, we fit the regression model

$$\begin{aligned}\log(\text{Income})_i = \beta_0 + \beta_1 \text{Age}_i + \beta_2 \text{Disability}_i + \beta_3 \text{Educ}_i \\ + \beta_4 \text{HrsWorked}_i + \beta_5 \text{Medicaid}_i + \beta_6 \text{Medicare}_i + \epsilon_i,\end{aligned} \tag{3.24}$$

and pool all inferences using multiple imputation combining rules. The models were trained on all observations for which was income was positive. Figure 3.6 displays the coefficient estimates and 95% confidence intervals obtained under each of the four imputation methods, along with those obtained when using the actual data. A table with the exact values is provided in the Appendix.

From these results, we see that the estimated coefficients of numeric variables tends to be biased towards zero when PMM or ERL is used and estimated coefficients of binary variables tend to be biased away from zero. In contrast, both the JTM and CJTM models produce estimates much more in line with those obtained from the actual data and have lower standard errors. We do not see a substantial difference

between the results obtained using the joint transformation model and the results obtained using the constrained joint transformation model.

3.7 Discussion

In this chapter, we presented a novel method for modeling the joint distribution of multivariate numeric data that generalizes that introduced by Hoff (2007). We demonstrated how to use a pseudo-Bayesian estimation procedure to generate multiply imputed datasets. We then extended our approach to perform jointly feasible multiple imputation by developing a model that placed nonzero probability mass strictly on the support of the data.

The proposed approach has a number of advantages. First, it is a multiple imputation engine based on a coherent and explicit joint model for the data. As a result, it is much easier to theoretically study than fully conditional models or predictive mean matching. Our procedure is much more flexible than the Gaussian copula model developed by Hoff (2007) and can be used for a wider variety of numeric datasets. In particular, the Gaussian copula model performed quite poorly in the empirical example discussed in Section 3.6. Our procedure for estimating the transformation model places far less of a burden on the individual mixture components than that proposed by Kottas et al. (2005) and can be applied to continuous, ordinal, discrete, and binary data.

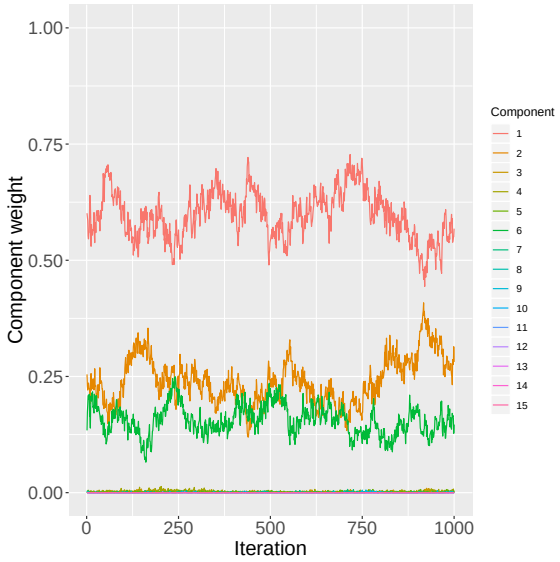
Using this model to perform multiple imputation is particularly appealing for organizations who wish to release multiply imputed datasets satisfying feasibility constraints. For one, this method produces hot deck imputations, ensuring that all imputed values are marginally reasonable. In addition, the model proposed in Section 3.4 can be used to impute data values satisfying general feasibility constraints, not simply those that can be defined by a system of linear inequalities.

Despite its appealing properties and strong empirical performance, the proposed

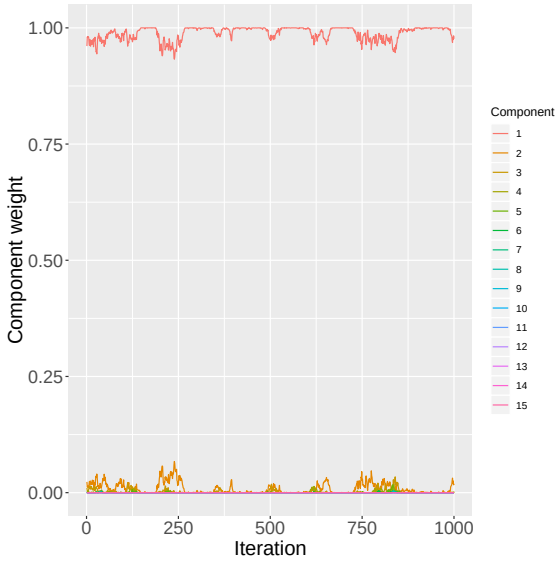
methodology has a number of limitations. For one, the method is only suitable for imputation of numeric data. In its current form, any categorical variables in the dataset are not accounted for when performing imputations, discarding potentially valuable information about the numeric variable of interest. In addition, the model tends to be highly sensitive to outliers in the data, as imputations are invariant to the scale of the data. As an extreme example, consider a situation in which the tallest person in a dataset is actually 7 feet, 0 inches tall but is mistakenly recorded as 70 feet, 0 inches tall. Regardless of the mistake, our method will impute a height of 70 feet just as frequently as it would have imputed a height of 7 feet without the error. As a result, it is especially important to make sure that all potential donor observations are of high quality prior to performing imputation using the joint transformation model presented here.

Replication code is available via the R software package `midamix`, hosted on GitHub at `https://github.com/burrisk/midamix`.

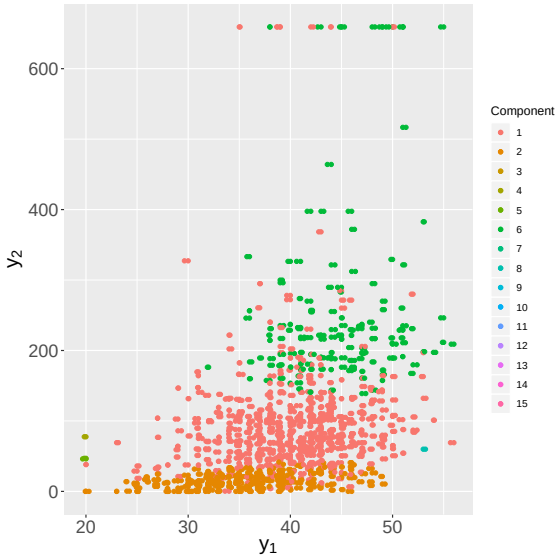
(a) Cluster weights (Unit-Spaced)



(b) Cluster weights (Empirical CDF)



(c) Posterior Predictive (Unit-Spaced)



(d) Posterior Predictive (Empirical CDF)

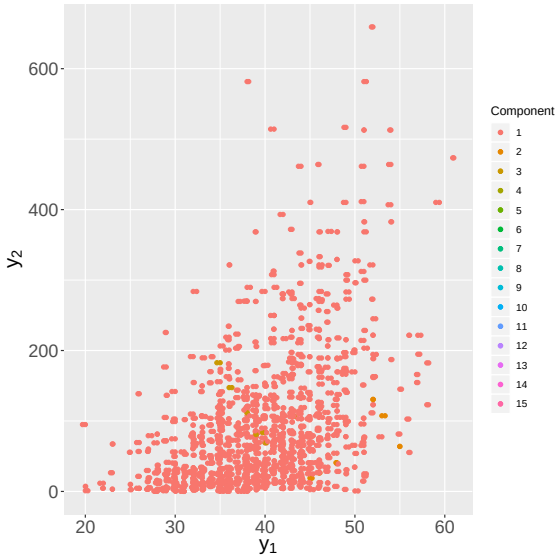


FIGURE 3.3: Traceplots of component weights and draws from the posterior predictive distribution under the unit-spaced procedure ((a) and (c)) and the empirical CDF procedure ((b) and (d)). The empirical CDF procedure not only results in a better fit, but also reduces the number of clusters, aiding interpretability.

(a) Absolute raw bias comparison

(b) CI coverage rate comparison

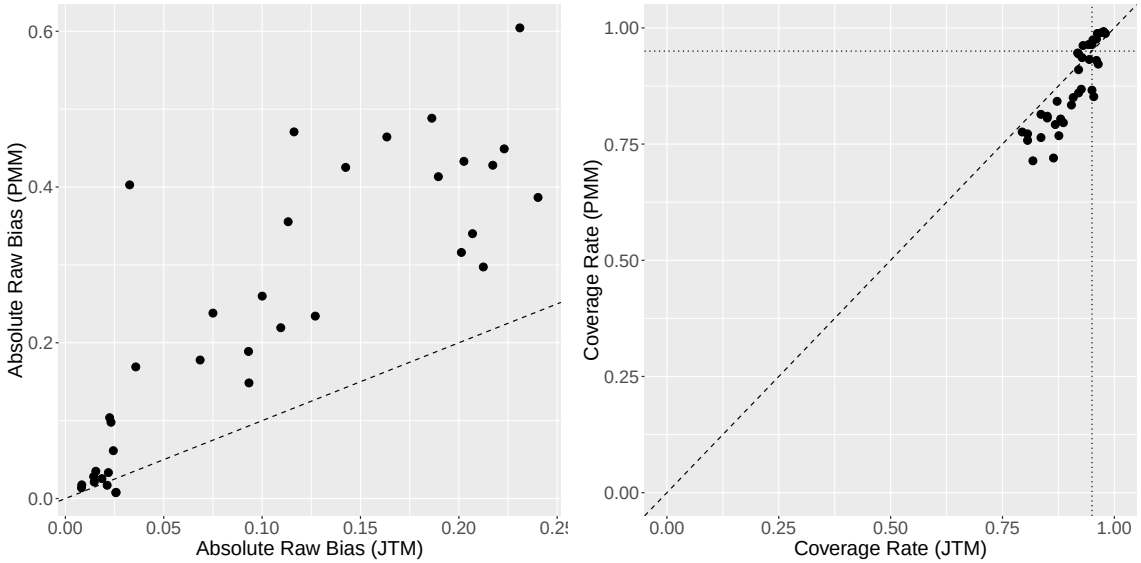


FIGURE 3.4: Comparison of inference based on imputed datasets generated by predictive mean matching (PMM) and the joint transformation model (JTM). For the same target quantities, JTM tends to have lower absolute raw bias and higher coverage rates than PMM.

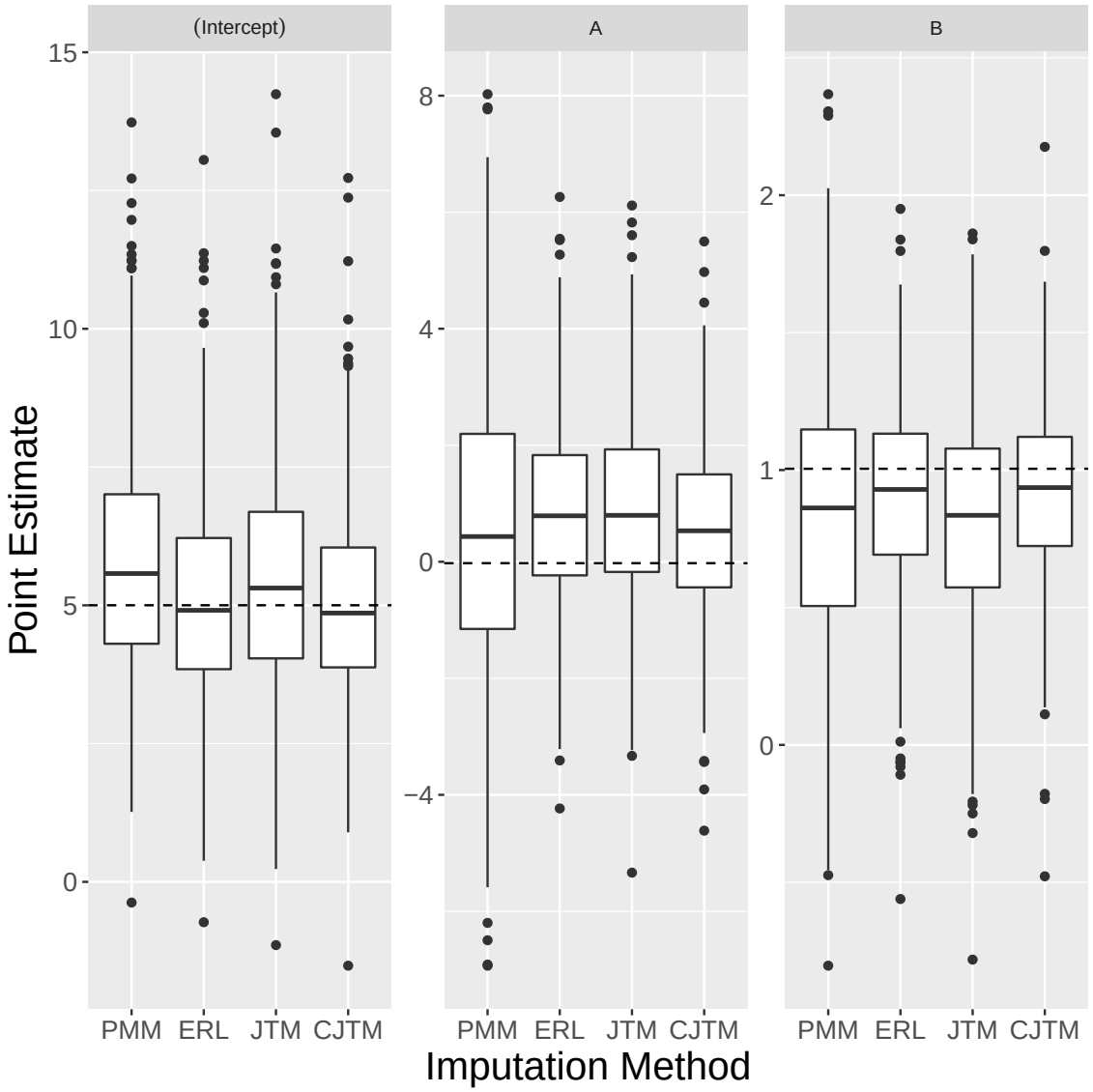


FIGURE 3.5: Boxplots of the point estimates of the regression coefficients obtained from the 500 datasets. The bias and uncertainty of inferences obtained via CJTM imputations appears to be lower than those based on the other three methods.

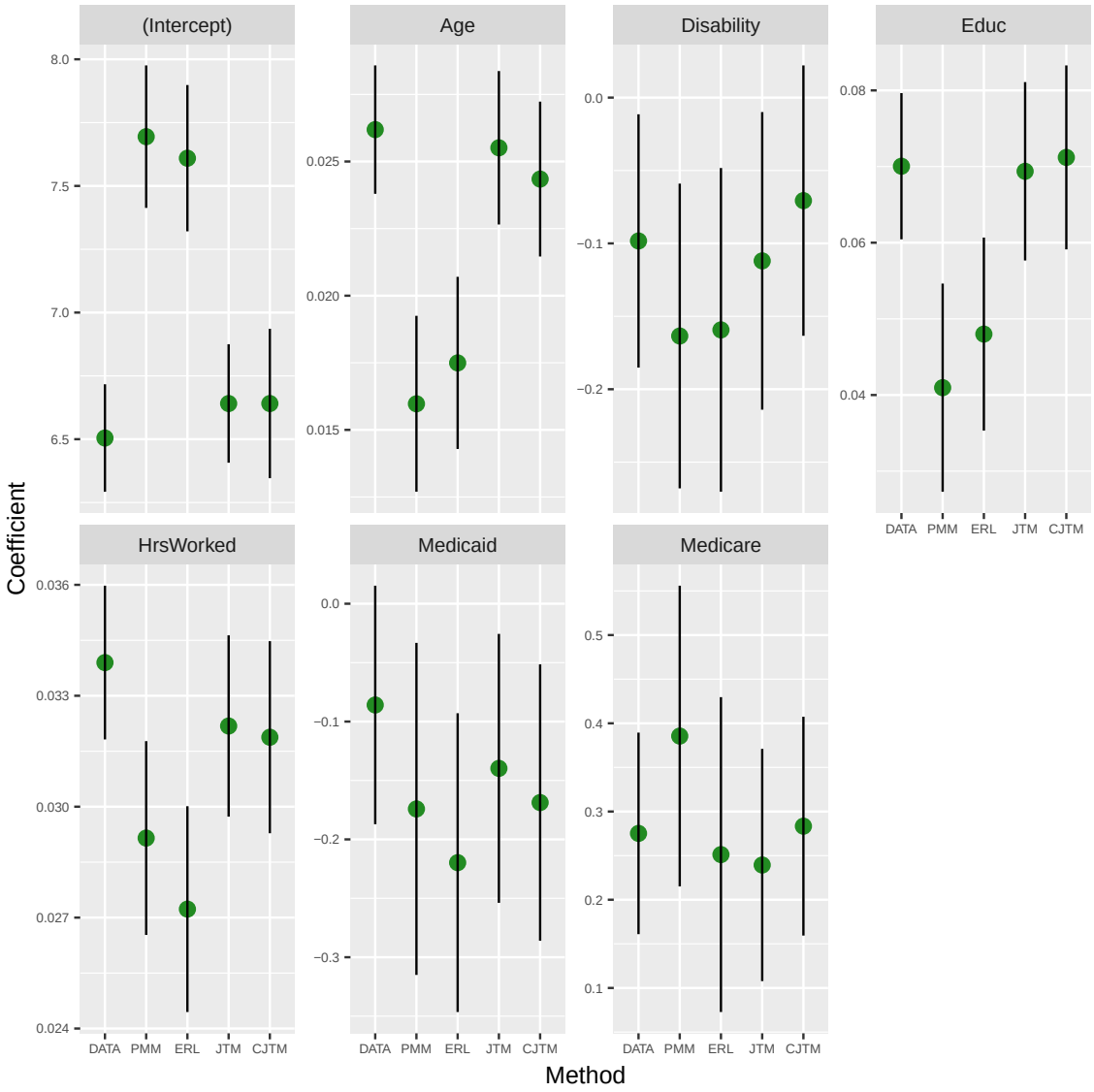


FIGURE 3.6: Point estimates, along with 95% confidence intervals for the regression coefficients based on the data (DATA), predictive mean matching (PMM), extended rank likelihood (ERL), joint transformation model (JTM), and constrained joint transformation model (CJTM)

.

4

Visual-Motor Expertise in Athletes

4.1 Introduction

4.1.1 Motivation

Sports place incredible demands on the human visual system. Hitting a baseball, returning a serve, or blocking a shot on goal all require an athlete to see and react with great efficiency and precision. Over the last several decades, scientists have attempted to understand how the eyes and the visual brain contribute to athletic expertise (Gregory, 1997; Yarrow et al., 2009). As captured in two recent meta-analyses (Mann et al., 2007; Voss et al., 2010), higher-achieving athletes are better at detecting perceptual cues, efficiently moving their eyes, processing information quickly, and maintaining attention. Across this literature, elite athletes tend to outperform sub-elite athletes and non-athletes in both sport-specific tasks and component-skill tasks that tap into broad, fundamental visual (Hitzeman and Beckerman, 1993; Laby et al., 1996) and perceptual-cognitive mechanisms (Starkes and Ericsson, 2003; Casanova et al., 2009; Williams and Ericsson, 2005; Williams and Ford, 2008). In addition, these expertise-related benefits are largely reflective of the types of demands that

are required by the specific roles athletes play. For example, experts in sports that require a greater horizontal distribution of attention (e.g., hockey) demonstrate a greater horizontal breadth of attention than athletes whose sports require more vertical attention (e.g., volleyball), and vice versa (Huttermann et al., 2014). Collectively, the literature indicates that high-level athletes may be experts at processing some types of visual information (Klemish et al., 2017), making athletes a valuable population to help understand the limits of visual-motor abilities in individuals with considerable training.

While studies of visual-motor expertise represent a fruitful domain for exploring models of learning and determining the limits of human performance, research with experts has been severely limited in impact because of small sample sizes and heterogeneous psychological constructs. In particular, it is difficult to obtain access to high-achieving individuals for research purposes, and therefore, inference on experts is often inconclusive due to the large degree of uncertainty inherent in small sample sizes [1]. In addition, each study uses different tests to measure visual-motor abilities, which has resulted in a disparate set of findings that are difficult to aggregate (Eccles et al., 2006; Voss et al., 2010). It is also not always clear how performance on controlled laboratory tests maps onto real-world achievement, therefore creating a fundamental disconnect between research and practical applications. Despite these limitations, there remains tremendous interest in understanding the visual, perceptual, cognitive, and motor faculties that differentiate experts from non-experts, and in revealing how these differences manifest themselves in competition.

In particular, the sport of baseball is a particularly appealing domain for the study of visual-motor expertise. Major League Baseball (MLB) pitches move at speeds near the processing limits of the vestibular-ocular tracking (Bahill and LaRitz,

[1] The average sample size across the 62 studies included in the meta-analyses of Mann and Voss consists of 15.39 athletic experts each, for a total of 954 athletic experts across the whole literature.

1984; Schalen, 1980), leaving a baseball batter with mere milliseconds to decipher the pitch, project its trajectory, decide to swing, and coordinate the timing and path of a 2.25-inch diameter bat. The immense difficulty of this task is underscored by the fact that players who hit successfully on less than a third of their at-bats can receive hundred million dollar contracts in today's free-agent market. Pitching, while equally demanding, draws upon a fundamentally different skill set. Pitchers attempt to deny batters effective contact with the ball while projecting it through the strike zone 60 feet away. Despite the need to visualize the strike zone, it has been argued that motor demands, such as controlling the speed, spin, and location of the ball are more important for pitching success than visual requirements (Molia et al., 1998).

Given the substantial role of visual and motor demands in baseball, there has been a concerted effort to determine which elements of the perception-action cycle contribute to successful baseball performance. However, the combination of game statistics with high sampling variability and costly sample acquisition makes inferring meaningful relationships difficult. Although a small number of studies have reported links between superior baseball statistical production and better visual reaction times (Classe et al., 1997), dynamic stereoacuity (Solomon et al., 1988), binocular divergence (Spaniol et al., 2015), and visual recognition (Reichow et al., 2011; Szymanski et al., 2015)), their inferences are generally based on small sample sizes. A more common approach for inferring the visual-motor abilities important for baseball performance involves studying the difference between professionals, amateurs, and non-athletes. This literature, and the larger debate across all sports, centers on the question of whether athletes possess inherently better visual-system physiology (so-called, visual hardware), or if differences are restricted to enhanced perceptual-cognitive abilities that can be shaped through practice (so-called visual software; (Elmurr, 2011). Some studies have found that expert baseball players possess superior visual acuity (Laby et al., 1996), enhanced contrast sensitivity (Hoffman et al.,

1984), better peripheral vision (Kato and Fukuda, 2002), and better visual tracking abilities (Uchida et al., 2012) than non-athlete controls. While these studies indicate that superior batters possess superior visual hardware, the preponderance of evidence in the literature concludes that, in the absence of hardware differences, expert performance is subserved by superior visual software. For example, past research with baseball and cricket batters found that expert athletes demonstrated more adept anticipation, pattern recognition, and visual search skills than non-experts (Abernethy et al., 1994b; Eccles et al., 2006); (Helsen and Starkes, 1999; Ward and Williams, 2003; Williams et al., 2011). Nevertheless, given the challenges inherent in doing research with high-level athlete populations, the contribution of hardware and software to expert performance remains an open question.

4.1.2 Data

In 2011, Nike Incorporated (Inc.) launched the SPARQ Sensory Stations as a tool to quantitatively evaluate athlete visual-motor abilities. The Sensory Stations include of a battery of nine psychometric tasks (Table 4.1) administered under standardized conditions with video instructions by trained and certified administrators. Prior to testing, participants complete a registry of information that reports demographic (e.g., age and height), sport (e.g., primary sport and position), concussion history (number and recency), and vision (e.g., eye dominance and eye care history) characteristics (Wang et al., 2018). The Sensory Stations operated for four years until 2015 and all assessment data were maintained on a central database. This information was used to provide sport-specific normative information to individuals about their abilities compared to their specific athletic cohort and to monitor learning when coupled with visual-motor training interventions.

Past research with the Sensory Stations has demonstrated that the battery of tests is reliable (Erickson et al., 2011; Wang et al., 2015), with some tasks demon-

strating linear improvements with practice over multiple sessions (Krasich et al., 2016). Improved performance on this battery has been seen following sports vision training interventions (Appelbaum et al., 2016). Collectively, past research (reviewed by Appelbaum and Erickson (2016)) suggests that this battery may serve as a useful tool for understanding human performance, warranting further investigation into the visual-motor characteristics of athletes and their relation to performance outcomes.

Table 4.1: Brief descriptions of the Nike Sensory Station tasks

Task	Label	Description
Visual Clarity	VC	Measures visual acuity for fine details at a distance
Contrast Sensitivity	CS	Measures the minimum resolvable difference in contrast at a distance
Depth Perception	DP	Measures how quickly and accurately participants are able to detect differences in depth at a distance using liquid crystal glasses
Near-Far Quickness	NFQ	Measures the number of near and far targets that can be correctly reported in a 30 second time interval
Target Capture	TC	Measures the speed at which participants can shift attention and recognize peripheral targets
Perception Span	PS	Measures the ability to remember and recreate visual patterns
Eye-Hand Coordination	EHC	Measures the speed at which participants can make visually-guided hand responses to rapidly changing targets
Go/No-Go	GNG	Measures the ability to execute and inhibit visually guided hand responses in the presence of go and no-go stimuli
Reaction Time	RXN	Measures how quickly participants react and respond to a simple visual stimulus

In this chapter, we analyze the Nike visual-motor assessment data from 2317 distinct athletes, a particularly large sample compared to others in the literature. All data were shared under a secondary-data protocol approved by the Duke University Institutional Review Board [IRB B0706]. Under this protocol, all data were collected

for real world use, without informed consent, and shared with the research team after removal of all protected health information (PHI). As such, these data conform to U.S. Department of Health and Human Services', Regulatory considerations regarding classification of projects involving real world data, (DHHS, 2015).

The Nike database consists of the task scores, as well as athlete characteristics such as level of expertise, primary sport, and gender for 2317 athletes (1871 male, 446 female). Tables 4.2 and 4.3 describe the distribution of gender, sport type, and level of expertise for athletes who completed a Nike Sensory Station evaluation.

Table 4.2: Distribution of athlete level and sport type for male athletes

	Male Athlete Level				
Sport Type	Middle School	High School	College	Pro	Total
Strategic	122	222	459	358	1161
Interceptive	75	123	111	401	710
Total	197	345	570	759	1871

Table 4.3: Distribution of athlete level and sport type for female athletes

	Female Athlete Level				
Sport Type	Middle School	High School	College	Pro	Total
Strategic	40	61	86	57	244
Interceptive	39	55	105	3	202
Total	79	116	191	60	446

Raw scores for each of the tasks are recorded on different scales; for example, Visual Clarity scores are expressed in log units and can take on negative values, whereas typical scores for Eye-Hand Coordination are reported in milliseconds and range between 40,000 and 70,000. In addition, lower raw scores are better for some tasks, such as Reaction Time, but worse for others, such as Perception Span. For interpretability, we center and scale the responses such that a higher standardized score corresponds with superior performance on the task. Figure 4.1 illustrates the standardized marginal distributions for each of the task scores. The distributions of

task scores tend to be left-skewed, with a few outliers in the lower tail. Although all task scores are numeric, many are discrete (e.g. Contrast Sensitivity, Perception Span), while others are continuous (e.g. Reaction Time, Eye-Hand Coordination). In addition, three of the tasks have missing values; Visual Clarity is missing 15, Contrast Sensitivity is missing 6, and Depth Perception is missing 104.

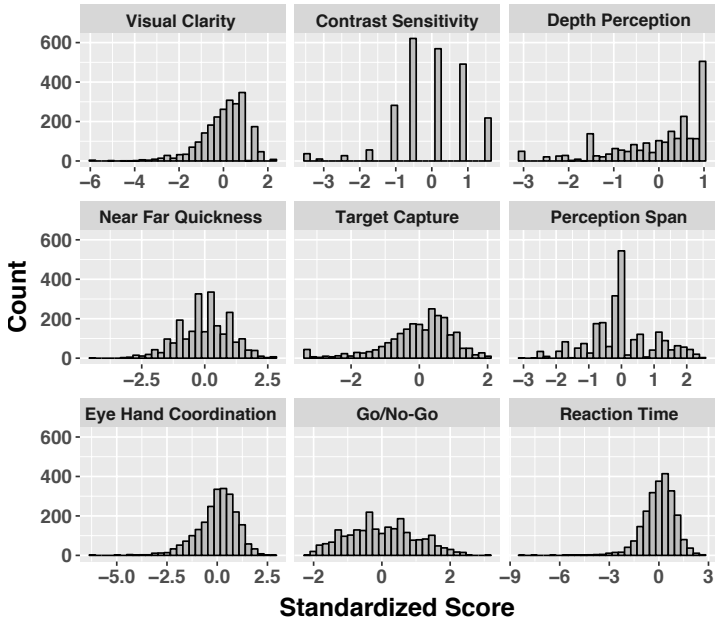


FIGURE 4.1: Distribution of task scores for the athletes in the Nike database

In analyzing this multivariate data, we have two primary objectives: first, we aim to evaluate the relationship between visual-motor abilities and athletic expertise across levels of achievement and sport type, for both men and women. For this purpose, we develop a Bayesian semiparametric transformation model in Section 4.2. Second, we aim to quantify the relationship between visual-motor abilities and on-field performance in professional baseball by considering a subset of 252 professional baseball players who completed Sensory Station assessments. We develop a series of Bayesian latent variable models that enable comparison of player performances in different contexts in Section 4.3. A discussion of our conclusions and contributions to the literature is provided in Section 4.4.

4.2 Visual-motor variation in the Nike database

In this section, we analyze 2317 athletes tested on the Sensory Station to evaluate the relationship between visual-motor abilities and athletic expertise across levels of achievement and sport type, for both men and women. To do so, we develop a Bayesian semiparametric transformation model for the analysis of multivariate data. This model has particular value in that it can be applied generally to datasets that contain missing values, negative values, mixed data types, and non-normal marginal distributions. Model estimation is based only on the ranks of the data, which enables robust inference that is invariant to monotonic transformations of the data. Using this methodology, we model athlete scores for seven of the nine visual-motor tasks, excluding Depth Perception and Go/No-Go. We do this due to technological malfunctions associated with the Depth Perception data in some testing centers and the known limitations of Go/No-Go as a task (Wang et al., 2015). We consider covariates such as the level of athletic expertise, primary sport type, and gender. Following past research (Mann et al., 2007; Voss et al., 2010), sports are classified as *interceptive* if the primary athletic actions require coordination between an athlete's body, body parts, or a held implement, and an object in the environment (e.g., tennis, baseball) (Davids, 2002). Conversely, a sport is classified as *strategic* if it is important to divide attention in order to monitor the location of teammates, opponents, and projectiles on the field (e.g., soccer, basketball) (Singer, 2000). We fit two separate models: one including only the main effects for interpretability and the other with all two-way interactions to capture heterogeneous relationships across combinations of level, sport type, and gender.

4.2.1 Methods

In this section, we aim to estimate and quantify the uncertainty about the joint dependence of the task scores, as well as the conditional dependence between performance on the tasks and covariates such as gender, level, and sport type. To do so, we extend the general method of modeling the joint distribution of mixed data introduced by Hoff (2007) to multivariate regression. The key idea is to use a copula to separate the dependence structure from the marginal distributions of the task scores. We model the associations among the task scores and covariates parametrically, and leave the marginal distributions arbitrary and unspecified. As such, the ranks of the observations rather than their actual values are used to estimate the association parameters. This approach produces results that are not based on any marginal distributional assumptions, invariant under any monotonic transformation of the data, and interpretable across covariates and responses. Moreover, this method can be trivially extended to accommodate data that are missing-at-random (MAR).

Specifically, we propose a semiparametric transformation model for the data, in which the dependence structure of the tasks is described by a Gaussian copula and the marginal distribution F_j of each task is left unspecified. Specifically,

$$\begin{aligned} \boldsymbol{Z} &\sim \text{Normal}\left(\boldsymbol{X}\boldsymbol{B}, \boldsymbol{C} \otimes \boldsymbol{I}_n\right) \\ \boldsymbol{U}_{ij} &= \Phi\left(\boldsymbol{Z}_{ij}\right) \\ \boldsymbol{Y}_{ij} &= F_j^{-1}\left(\boldsymbol{U}_{ij}\right), \end{aligned} \tag{4.1}$$

where $\boldsymbol{Y} \in \mathbb{R}^{n \times p}$ is a matrix of p different task scores for n athletes, $\boldsymbol{X} \in \mathbb{R}^{n \times q}$ a design matrix of q covariates, $\boldsymbol{B} \in \mathbb{R}^{q \times p}$ a matrix of regression coefficients, $\boldsymbol{C} \in \mathbb{R}^{p \times p}$ a positive-definite correlation matrix describing the correlation between the tasks, $\boldsymbol{I}$ the identity matrix, and the symbol $\otimes$ the Kronecker product. In context, $\boldsymbol{B}_{ij}$ represents the contribution of the ith covariate to the mean of the jth task score. The covariates $\boldsymbol{X} \in \mathbb{R}^{n \times q}$ are centered to ensure identifiability and $F_j^{-1}(t) = \inf\{y \in \mathbb{R} : F_j(y) \geqslant t\}$ is

the pseudo-inverse of F_j. This is a highly general model, with standard multivariate regression and multivariate probit models as special cases. Because the transformation F_j^{-1} is non-decreasing, observing $\boldsymbol{Y}$ gives us substantial information about the latent variable matrix $\boldsymbol{Z}$, even though the marginal distributions are unknown. In particular,

$$\boldsymbol{Y}_{ij} < \boldsymbol{Y}_{i'j} \Rightarrow \boldsymbol{Z}_{ij} < \boldsymbol{Z}_{i'j}. \tag{4.2}$$

As such, observing $\boldsymbol{Y}$ means that $\boldsymbol{Z}$ must lie in the set

$$D = \left\{ \boldsymbol{Z} \in \mathbb{R}^{n \times p} : \max \left\{ \boldsymbol{Z}_{i'j} : \boldsymbol{Y}_{i'j} < \boldsymbol{Y}_{ij} \right\} < \boldsymbol{Z}_{ij} < \min \left\{ \boldsymbol{Z}_{i'j} : \boldsymbol{Y}_{ij} < \boldsymbol{Y}_{i'j} \right\} \right\}, \tag{4.3}$$

so the likelihood can be expressed as

$$\begin{aligned} p\left(\boldsymbol{Y} \mid \boldsymbol{B}, \boldsymbol{C}, F_1, \cdots, F_p\right) &= p\left(\boldsymbol{Y}, \boldsymbol{Z} \in D \mid \boldsymbol{B}, \boldsymbol{C}, F_1, \cdots, F_p\right) \\ &= p\left(\boldsymbol{Z} \in D \mid \boldsymbol{B}, \boldsymbol{C}, F_1, \cdots, F_p\right) \\ &\quad \times p\left(\boldsymbol{Y} \mid \boldsymbol{Z} \in D, \boldsymbol{B}, \boldsymbol{C}, F_1, \cdots, F_p\right) \\ &= p\left(\boldsymbol{Z} \in D \mid \boldsymbol{B}, \boldsymbol{C}\right) \times p\left(\boldsymbol{Y} \mid \boldsymbol{Z} \in D, \boldsymbol{B}, \boldsymbol{C}, F_1, \cdots, F_p\right). \end{aligned} \tag{4.4}$$

Here inference is performed on $\boldsymbol{\theta} = \{\boldsymbol{B}, \boldsymbol{C}\}$ based on the extended rank likelihood $p\left(\boldsymbol{Z} \in D \mid \boldsymbol{\theta}\right)$, which does not depend on the nuisance parameters $F_1, \cdots, F_p$ (Hoff, 2007). Although obtaining the maximum likelihood estimate of $\boldsymbol{\theta}$ is difficult, performing Bayesian inference on $\boldsymbol{\theta}$ is straightforward via Gibbs sampling. Since

$$p\left(\boldsymbol{\theta} \mid \boldsymbol{Z} \in D\right) \propto p\left(\boldsymbol{Z} \in D \mid \boldsymbol{\theta}\right) \times p\left(\boldsymbol{\theta}\right), \tag{4.5}$$

the posterior distribution $p\left(\boldsymbol{\theta} \mid \boldsymbol{Z} \in D\right)$ can be empirically approximated by samples from the posterior $p\left(\boldsymbol{\theta}, \boldsymbol{Z} \mid \boldsymbol{Z} \in D\right)$, obtained by the Gibbs sampler that iterates as follows:

1. Simulate $\boldsymbol{\theta} \sim p\left(\boldsymbol{\theta} \mid \boldsymbol{Z}\right)$.

2. Simulate $\boldsymbol{Z}_{ij} \sim p\left(\boldsymbol{Z}_{ij} \mid \boldsymbol{Z} \in D, \boldsymbol{Z}_{-[i,j]}, \boldsymbol{\theta}\right)$.

Using the methodology presented above and in Section C.2, we model athlete scores for $p = 7$ visual-motor tasks, excluding Depth Perception and Go/No-Go. The design matrix $\boldsymbol{X}$ includes centered indicators for athlete level of expertise, primary sport type, and gender with no intercept to ensure identifiability. We fit two separate models: one including only the main effects for interpretability and another with all two-way interactions to capture heterogeneous relationships across combinations of level, gender, and sport type.

For each model, we draw a total of 100,000 samples after a burn-in of 1,000 iterations, storing the values of $\boldsymbol{Z}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ every tenth iteration. All athletes in the data, including those with missing values for Visual Clarity and Contrast Sensitivity, were included due to the fact that the missing values were sampled jointly with the parameters of interest. The model was validated by standard convergence diagnostics and posterior predictive checks. Summaries of the samples from the joint posterior distribution of $\boldsymbol{B}$ and $\boldsymbol{C}$ are described in Section 4.2.2. Details of the Gibbs sampling approach, including the prior distributions that we used in obtaining the results presented in this study, are given in Section C.2.

4.2.2 Results

Main Effects Model

Throughout this section, we adopt a convention that a regression coefficient is *significant* if the corresponding symmetric 95% credible interval does not contain zero. This indicates that there is at least a 95% probability that the estimated direction of the association is correct, based on our posterior beliefs after seeing the data. Similarly, we call the difference between two groups significant if the 95% credible interval of the difference in predicted group means under the model does not include zero. We subjectively refer to a significant group difference for a given task as *substantial*

if the corresponding magnitude of the posterior mean is high, relative to those of other tasks and groups.

Using the model described in Section 4.2.1, we use Markov Chain Monte Carlo sampling to estimate the posterior distribution of the model parameters. Posterior means of the matrix of regression coefficients are given in Table 4.4 and significant coefficients are bolded. The relative magnitudes of the coefficients are further visualized in Figure 4.2, where the baseline group is a middle school male who plays a strategic sport.

Table 4.4: Posterior mean coefficients for the main effects model. Coefficients for which the 95% credible intervals do not contain zero are bolded.

	VC	CS	NFQ	TC	PS	EHC	RXN
High School	0.103	0.049	**0.534**	0.002	**0.429**	**0.982**	0.158
College	**0.273**	0.099	**0.799**	0.113	**0.453**	**1.380**	**0.489**
Pro	**0.320**	**0.215**	**0.744**	0.018	**0.436**	**1.664**	**0.553**
Interceptive	**0.187**	**0.123**	**0.121**	0.001	**-0.086**	0.001	**0.143**
Female	-0.002	-0.042	**-0.128**	0.078	**0.105**	**0.193**	0.020

Overall, results indicated that athletes with higher levels of expertise perform better for all Nike Sensory Station tasks, with the exception of Target Capture. Interestingly, the largest and most compelling differences were exhibited in tasks that demand greater motor control, such as Eye-Hand Coordination, Near Far Quickness, and Reaction Time. In these tasks, substantial differences existed between middle school and high school athletes, as well as between high school and college athletes. While athlete level differences were much smaller in the Visual Clarity and Contrast Sensitivity tasks, these measures of visual sensitivity also showed small gradations.

Athletes who play primarily interceptive sports such as tennis and baseball scored significantly better on measures of visual sensitivity such as Visual Clarity and Contrast Sensitivity. In addition, interceptive sport athletes also had significantly higher Near-Far Quickness scores and Reaction Times than strategic sport athletes, the

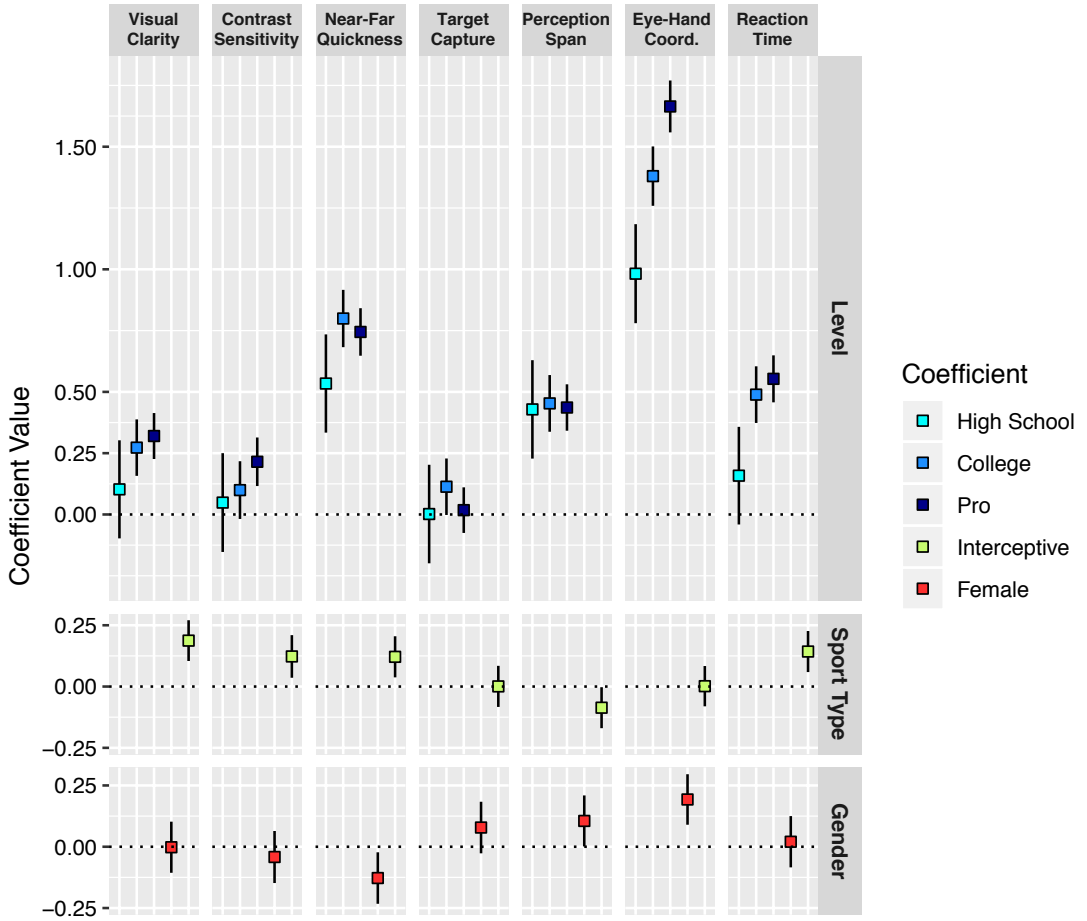


FIGURE 4.2: Posterior means, along with 95% credible intervals for the main effects model coefficients of sport level, sport type, and gender. The baseline, marked by a dotted line, corresponds to a middle school male who plays a strategic sport. Our model enables comparison of magnitudes across both coefficients and tasks.

latter of which was also noted by Mann et al. (2007) in their meta-analysis of the literature. Interestingly, athletes who play strategic sports tended to do better at Perception Span, which measures the ability to store and recreate visual patterns. This may indicate that spatial working memory (e.g., mentally representing and maintaining the location of teammates and opponents) is of primary importance in strategic sports.

Although the magnitudes of the visual-motor differences between genders were substantially less than those across levels of expertise, significant gender differences in

task performance were present. In particular, holding sport type and level of expertise constant, males typically ranked higher than females at Near-Far Quickness, while females typically ranked higher at Eye-Hand Coordination and Perception Span.

Interaction Model

To provide a more systematic look at group-level differences, a second multivariate regression model was fit with both main effects and all two-way interactions. Table 4.5 presents a summary table of posterior means for this model, and Figure 4.3 visualizes the estimated task score percentile for a typical athlete under each combination of level, sport type, and gender, a table for which is provided in Section C.3.

Table 4.5: Posterior means for the interaction model. Coefficients for which the 95% credible intervals do not contain zero are bolded. The baseline is identical to that used in Table 4 and Figure 1., as before, corresponds to a middle school who plays a strategic sport.

	VC	CS	NFQ	TC	PS	EHC	RXN
High School	0.254	0.121	**0.544**	0.023	0.389	**1.086**	0.198
College	**0.343**	0.199	**0.847**	0.140	**0.380**	**1.458**	**0.565**
Pro	**0.393**	**0.322**	**0.736**	0.027	**0.364**	**1.727**	**0.442**
Interceptive	**0.403**	0.336	0.108	-0.035	-0.233	-0.072	-0.037
Female	0.131	-0.122	0.041	0.010	0.261	**0.597**	0.200
High School × Interceptive	**-0.248**	**-0.211**	0.066	-0.038	**0.269**	0.064	0.101
College × Interceptive	-0.106	**-0.357**	0.066	-0.077	**0.206**	**0.186**	0.060
Pro × Interceptive	**-0.203**	**-0.255**	0.043	0.028	**0.186**	0.077	**0.352**
High School × Female	-0.196	0.058	-0.156	-0.033	**-0.276**	**-0.494**	**-0.289**
College × Female	-0.034	0.113	**-0.279**	-0.050	-0.012	**-0.539**	**-0.446**
Pro × Female	0.081	-0.039	0.084	0.008	-0.043	**-0.336**	0.196
Interceptive × Female	**-0.226**	0.087	-0.075	**0.259**	**-0.184**	-0.036	**0.199**

One finding from the interaction model outcomes was that the sport type differences in measures of visual sensitivity and spatial working memory are compressed at higher levels of athletic expertise. Specifically, interceptive sport athletes scored higher than strategic sport athletes on Visual Clarity and Contrast Sensitivity, with the most substantial differences exhibited at the middle school level. The same pattern held for Perception Span, though strategic sport athletes outperformed inter-

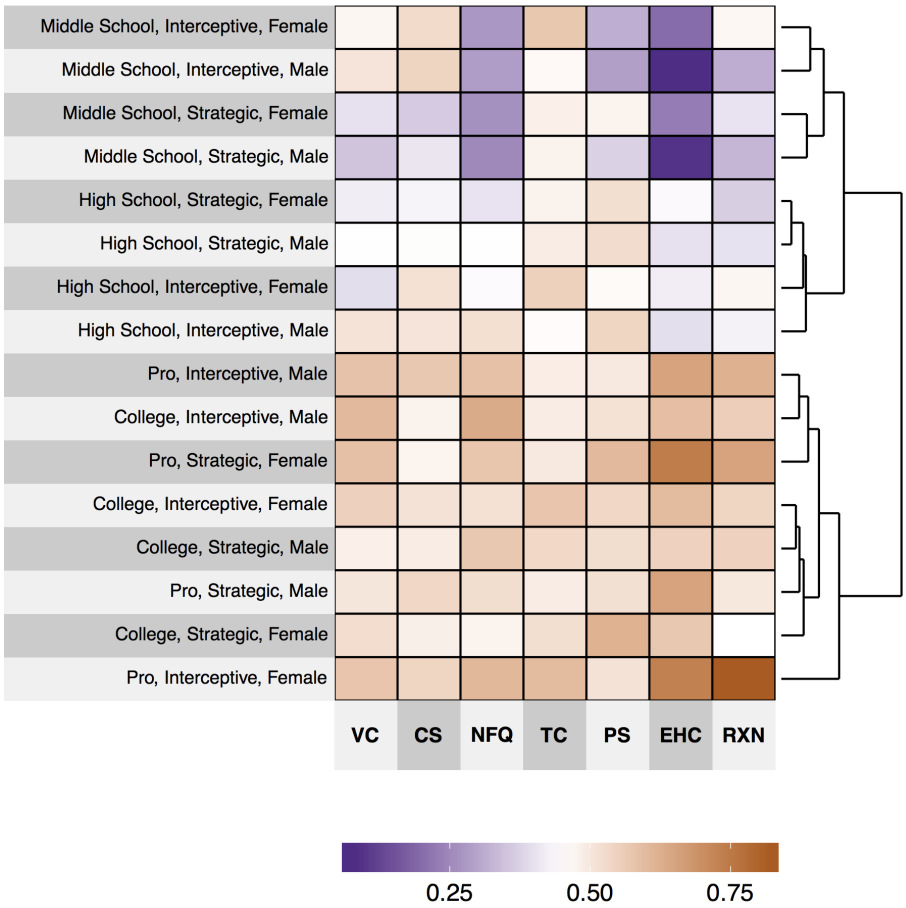


FIGURE 4.3: Heat map of the estimated percentiles of task performance for each of the 16 groups of athletes. The groups are clustered hierarchically in the row dendrogram. Note that middle school athletes and high school athletes form their own clusters, subdivided by sport type. College and professional athletes are more mixed, and typically subdivided by gender.

ceptive sport athletes at that task. However, the sport type differences in measures of visual-motor control either remained approximately the same or were amplified at higher levels of athletic expertise. In particular, there was a clear difference in professional athlete reaction times by sport type, with interceptive athletes typically demonstrating quicker responses.

When the groups were clustered hierarchically on the basis of the estimated percentiles across the seven tasks, the groups were primarily divided by level of expertise, with the greatest division occurring between high school and college. The younger

cohorts (middle school and high school) separated cleanly into their own clusters and subdivided by sport type. There was much more mixing within the collegiate and professional cohorts.

Relationship between tasks

The tasks comprising the Sensory Station battery were designed to measure sensory and motor abilities using a combination of psychometric methods (e.g. adaptive staircase procedures, speeded reaction tasks, etc.). To explore how these tasks relate to one another, and derive a more holistic picture of the constructs tested in this battery, the posterior distribution of the task correlation matrix $\boldsymbol{C}$ was evaluated. Figure 4.4 illustrates the location of the tasks along the first two eigenvectors of the maximum a posteriori (MAP) estimate of $\boldsymbol{C}$. The first eigenvector can be interpreted as capturing overall visual-motor ability, whereas the second eigenvector captures differences between the tasks that measure visual sensitivity (Visual Clarity and Contrast Sensitivity) and the other tasks. An examination of the conditional dependence structure of the task scores is provided in Section C.4.

4.3 The relationship between visual-motor abilities and on-field performance in professional baseball

In addition to quantifying variation in visual-motor expertise among athletes in the Nike database, we also attempt to identify the components of visual-motor ability that are most relevant for on-field performance in professional baseball. In this section, we analyze the Sensory Station assessments from 252 professional baseball players collected in 2012 and 2013 were compared to game statistics to evaluate the relationship between visual-motor skills and baseball production. For each player, game statistics from the season after testing were acquired along with information about his league(s) of participation and his primary position.

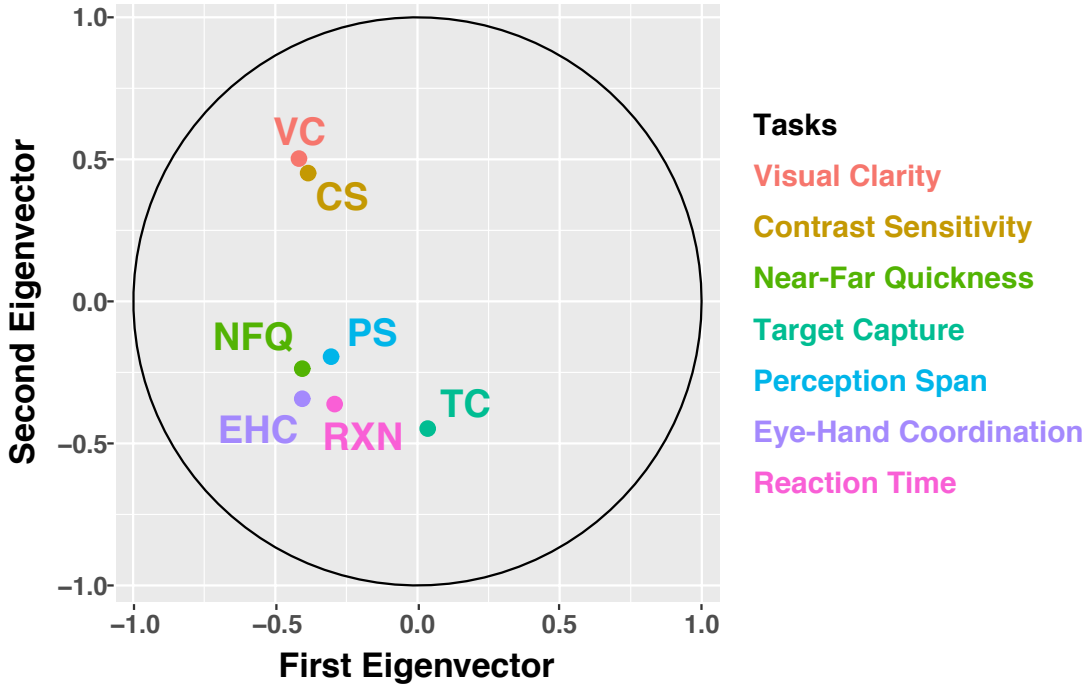


FIGURE 4.4: The location of each of the tasks along the first and second eigenvectors of the MAP correlation matrix $\boldsymbol{C}$. Most of the tasks are the same sign along the first eigenvector, whereas VC and CS are different signs than the other tasks along the second eigenvector.

4.3.1 Methods

Response Variables

A player's Nike Sensory Station assessment serves as our best measurement of a player's underlying visual-motor abilities. Similarly, a player's game statistics are the best indicators of a player's baseball ability and his contributions to his team. Here, we use on-base percentage (OBP), walk rate (BB%), strikeout rate (K%), and slugging-percentage (SLG) to measure the performance of batters. In addition, we use fielder-independent pitching (FIP) to measure the performance of pitchers. Below are brief descriptions of each of these statistics and our motivation for using

them as response variables in our models.

On-Base Percentage measures a player's propensity to reach base. On-base percentage is defined as

$$\text{OBP} = \frac{\text{Hits} + \text{Walks} + \text{Hit By Pitch}}{\text{At-Bats} + \text{Walks} + \text{Hit By Pitch} + \text{Sacrifice Flies}} \tag{4.6}$$

On-base percentage is a simple and widely used metric for player evaluation, since frequently reaching base gives the offense more opportunities to score runs. Players with high on-base percentages consistently make effective contact with the ball and draw walks. Since projecting the location of a pitch through the strike zone is important for making good contact and determining the difference between a ball and a strike, we hypothesize that this batting statistic may modulate with measures of visual-motor control.

Walk Rate measures a players propensity to draw walks. Walk rate is defined as

$$\text{BB\%} = \frac{\text{Walks}}{\text{Plate Appearances}} \tag{4.7}$$

Players who routinely draw walks generally differentiate well between balls and strikes, forcing the pitcher to throw pitches that are easier to hit. Walk rate can also provide information about a hitter's underlying approach at the plate.

Strikeout Rate measures a players propensity to strike out. Strikeout rate is defined as

$$\text{K\%} = \frac{\text{Strikeouts}}{\text{Plate Appearances}} \tag{4.8}$$

Strikeouts are an unequivocally negative outcome for the offense and should be avoided in an at-bat. Although some successful players have high strikeout rates, a high strikeout rate indicates that a batter struggles recognizing pitches or making contact with the ball. A player who strikes out frequently and walks rarely typically has a dim future in baseball.

Slugging Percentage measures a player's propensity to hit for power. Slugging percentage is defined as

$$\text{SLG} = \frac{\text{Total Bases}}{\text{At-Bats}} \tag{4.9}$$

Slugging percentage makes use of the fact that not all hits are equally valuable. Although it is an imperfect metric (e.g. doubles are not worth twice as much as singles), it does a decent job of quantifying batting power. Visual-motor abilities may have different effects on a batter's ability to hit for contact and ability to hit for power.

Fielder-Independent Pitching measures a pitchers run prevention, independent of the ability of the defense behind him. FIP is defined in terms of only variables that cannot be affected by the ability of the defense behind the pitcher.

$$\begin{aligned}\text{FIP} = &\frac{13 \cdot \text{Home Runs} + 3 \cdot (\text{Walks} + \text{Hit By Pitch}) - 2 \cdot \text{Strikeouts}}{\text{Innings Pitched}} \\ &+ \text{FIP Constant}\end{aligned} \tag{4.10}$$

Fielder independent pitching is a lower variance estimator of a pitcher's latent ability than ERA because it does not generally factor in outcomes that arise from balls in play, which are highly variable and debatably outside the pitcher's control (Davies and Basco, 2010). It is generally more stable than ERA, since it is a measurement that cancels out the effects of defense and luck. Visual-motor abilities may be related to pitcher performance, and FIP represents one of the best metrics for quantifying pitcher performance in a game setting.

Sample Characteristics

Although 308 professional baseball players (149 batters, 159 pitchers) completed assessments, we only examine data for the players with more than 30 at-bats or more than 30 innings pitched to mitigate the statistical noise associated with low sample

sizes. This yields a final analyzed data set of 252 players (141 batters, 111 pitchers). Table 4.6 reports the distribution of age and positional category in this sample. In general, most of the players in the sample are young prospects between 20-25 years old. However, there are older players in the sample who disproportionately play in the Major Leagues.

Not all professional leagues are equal. The level of competition in Major League baseball significantly outclasses that of AA baseball, for example. Players in our sample played in leagues ranging from Rookie League to Major League Baseball, which makes player comparison more challenging. Table 4.7 displays the number of players in our sample who play in each league. Note that some players play in more than one league.

Table 4.6: Age and positional characteristics of professional baseball players in the sample

	Batters	**Pitchers**
Age		
Mean (SD)	22.7 (3.9)	23.7 (3.6)
Min-Max	17-37	18-39
Position		
# Catchers	19	
# Infielders	65	
# Outfielders	57	

Table 4.7: Distribution of leagues by player type

	Rookie	A	Adv. A	AA	AAA	Majors
Batters	63	18	33	17	23	14
Pitchers	29	17	24	21	17	13

We fit separate models for each of the five response variables. The models use a common set of predictors. For each model, all parameters are estimated using only the data for that model. For convenience, we use a common notation across models when describing the models, so α_j in the OBP model will be distinct from α_j in the

SLG model, for example.

Binomial Response

Since OBP, BB%, and K% are defined as the number of successes divided by the number of opportunities, we use a binomial response for these three variables. Without loss of generality, we present the model for OBP. Let OB_{ij} denote the number of times that player i reached base in league j out of N_{ij} opportunities between 2012 and 2013. We treat each OB_{ij} as a realization of a random sample, with the player's true on base percentage equal to p_{ij}. Each p_{ij} is a function of the degree of difficulty of getting on base in league j, as well as the player's latent on-base ability parameter A_i. Each A_i is a function of variables $\boldsymbol{x}_i$ that include the player's scores on the Nike Sensory station tasks, a set of indicator variables for position, and age. Putting it together, we have the Bayesian multilevel model (Gelman and Hill, 2006)

$$\begin{aligned} \text{OB}_{ij} &\sim \text{Binomial}(N_{ij}, p_{ij}) \\ \text{logit}(p_{ij}) &\sim N(\alpha_j + \gamma_j A_i, \tau^{-1}) \\ A_i &= \boldsymbol{x}_i^\top \boldsymbol{\beta} \end{aligned} \tag{4.11}$$

Here, α_j represents the degree of difficulty in league j, and γ_j represents the impact of ability on performance in league j. Accounting for league differences in this way enables us to compare, for example, a 0.400 OBP player in AAA ball to a 0.320 OBP player in the MLB. We constrain γ_j to be positive, so that a higher latent ability level corresponds to a higher probability of reaching base. We use $\tau^{-1} > 0$ to allow for additional player heterogeneity when modeling p_{ij}. We include all of the visual-motor variables in $\boldsymbol{x}_i$ with the exception of Go/No-Go, since it is highly correlated with the Eye-Hand Coordination task and has limitations as a task (Wang et al., 2015). Note that we include Depth Perception in our model, since the data from the MLB team stations was deemed reliable. We transform Depth

Perception to the log scale as it is highly right-skewed and model the performance effects of age as linear. Diagnostics indicated that modeling age linearly fits the data reasonably well for all outcome models. We note that our findings were robust to both non-linear models for age and a maximum age threshold, mainly because age and visual-motor tasks have weak associations in our sample (see also (Klemish et al., 2017)). In addition to standardized age, $\boldsymbol{x}_i$ includes an indicator for catcher and an indicator for infielder. Hence, interpretations of all positional coefficients are with respect to outfielders. Ultimately, we are interested in performing inference on the posterior distribution of $\boldsymbol{\beta}$, which represents the impact of visual-motor abilities on A_i. We use non-informative normal and gamma priors on $\boldsymbol{\beta}$ and τ respectively; see Section C.6 for details.

Normal Response

SLG and FIP are long-run statistical averages over the number of at-bats and innings pitched, respectively. We present the model for SLG below; the model for FIP uses the same format. Let SLG_{ij} be the slugging percentage for player i in league j in N_{ij} at-bats for player i in league j. By the central limit theorem, as N_{ij} increases, the sampling distribution of SLG_{ij} approaches a normal distribution with mean μ_{ij} and variance σ^2/N_{ij}. Because we only included players for which $N_{ij} > 30$, the assumption of normality is plausible. We then specify a Bayesian multilevel model conditional on the slugging percentage ability parameters and league adjustment parameters. We have

$$\begin{aligned} \text{SLG}_{ij} &\sim N(\mu_{ij}, \sigma^2/N_{ij}) \\ \mu_{ij} &\sim N(\alpha_j + \gamma_j A_i, \tau^{-1}) \\ A_i &= \boldsymbol{x}_i^\top \boldsymbol{\beta} \end{aligned} \tag{4.12}$$

The procedure for estimating this model is analogous to the binomial response

case, but with an inverse-gamma prior distribution for σ^2. The prior specifications are provided in Section C.6.

Model Estimation

The models outlined above are not identifiable since α_j, γ_j, and A_i are unknown and depend upon each other. We overcome this problem by imposing highly concentrated priors on α_j and γ_j, obtained by modeling the game statistics of all professional baseball players between 2012 and 2013 who played in multiple leagues. Details about these league equivalence models are available in Section C.5. The prior means of the league effect parameters α_j and γ_j obtained via the model of game statistics with all professional players are summarized in Tables 4.8 and 4.9. In particular, Table 4.8 illustrates that there are two significant jumps in difficulty in professional baseball. There is a sizable increase in the quality of competition between Rookie baseball and non-rookie minor league baseball (A-AAA). In addition, there is an immense gap between AAA and the Major Leagues. Our model was unable to differentiate significantly between the non-rookie minor leagues. From Table 4.9, the impacts of ability are consistent across leagues, with the exception of the Major Leagues. With some statistics, such as OBP, latent ability matters less in the Major Leagues than it does in others. With others, such as FIP, it matters much more.

Table 4.8: Posterior Means for α_j displaying the inverse-logit of the means for OBP, BB%, and K% for interpretability. In context, we project that an average professional player will obtain a 0.358 OBP in Rookie ball and a 0.292 OBP in the MLB.

Attribute	Rookie	A	Adv. A	AA	AAA	MLB
logit^{-1}(OBP)	0.358	0.327	0.327	0.322	0.329	0.292
logit^{-1}(BB%)	0.108	0.093	0.096	0.093	0.089	0.071
logit^{-1}(K%)	0.170	0.188	0.184	0.192	0.195	0.232
SLG	0.432	0.384	0.379	0.376	0.397	0.351
FIP	3.013	3.517	3.377	3.613	3.782	4.279

Table 4.9: Posterior Means for γ_j. Higher values indicated higher relative impact of ability on the corresponding game statistic, given the model. These values should not be compared across statistics, since they are on different scales.

Attribute	Rookie	A	Adv. A	AA	AAA	MLB
OBP	0.110	0.118	0.109	0.101	0.103	0.060
BB%	0.316	0.304	0.300	0.320	0.305	0.275
K%	0.327	0.356	0.335	0.346	0.345	0.341
SLG	0.059	0.050	0.045	0.041	0.047	0.027
FIP	0.356	0.405	0.383	0.343	0.442	0.556

Once strong prior information on α_j and γ_j is obtained, we estimate the models detailed in above, restricting our attention to the seasons of 141 batters and 111 pitchers in our sample with greater than 30 at-bats or innings pitched in each league. While it is reasonable to include data from all players when estimating the binomial response models, we elect to use the same player pool in all our models for consistency. To facilitate efficient Gibbs sampling and generate comparable coefficients, we standardize all variables in the construction of $\boldsymbol{x}_i$ with the exception of the position dummy variables. Although measurements of Depth Perception are missing for four batters and four pitchers, the missing values are sampled as part of the Gibbs sampler used to estimate the model (Plummer, 2003) with an independent standard normal prior placed on each of the missing values. We ran the model for three chains of 10,000 iterations after a 1000 iteration burn-in period, and validated it using Markov Chain Monte Carlo diagnostics and posterior predictive checks.

4.3.2 Results

To start off the analysis, we check to see if performance on the battery of visual-motor tasks predicts on-field performance. In doing so, for each response variable, we fit two separate models: a full model with the both task scores and control variables included and a reduced model with only age and position included as control variables. If visual-motor abilities predict on-field performance, the full model should

significantly outperform the reduced model. Based on this, we report the individual coefficients for each of the models in which task scores added predictive power beyond the control variables.

WAIC

The Watanabe-Akaike Information Criterion (WAIC) is a useful way to compare two different Bayesian models of a particular response. It uses the log-posterior predictive density as the primary measure of accuracy, with a correction based upon the effective number of parameters in the model (Gelman et al., 2013). Asymptotically, it can be shown that WAIC approaches the results obtained via leave-one-out cross-validation (Vehtari et al., 2016). For each of the five models, we use WAIC to compare the full model with the visual-motor task results included in the design matrix to the reduced model that only accounts for position and age. If visual-motor variables add predictive power above and beyond that of the control variables, then the WAIC of the full model should be lower than that of the reduced model. Table 4.10 compares the WAIC of the full model to that of the reduced model for each of the five response variables. As indicated by the lower values for the Full, relative to the Reduced model, performance on the Nike Sensory Station tasks is predictive of OBP, BB%, and K%. However, visual-motor abilities do not predict either SLG or FIP. We therefore present coefficient summaries for OBP, BB%, and K% in Table 4.11. Summaries for SLG and FIP can be found in C.7.

Table 4.10: WAIC Model Comparison. Lower values for the full models relative to the reduced OBP, BB%, and K% models indicate that the added variables in the full models add meaningful predictive power.

	OBP	BB%	K%	SLG	FIP
Full Model	1210.8	1075.8	1276.4	403.4	363.8
Reduced Model	1226.4	1084.4	1284.6	403.1	361.9

On-Base Percentage The posterior means, standard deviations, and 95% credible

intervals for the coefficients $\boldsymbol{\beta}$ are presented in Table 6 for the full OBP, BB%, K% models. The control covariates Age and Position are included in both the full and reduced models. Variables for which 0 falls outside the 95% credible intervals are bolded. In general, bolded positive coefficients indicate that there is greater than 95% probability that the visual-motor ability measured in the task has a linear association with the game statistic of interest. To illustrate the posterior tail probabilities, a heat map of the z-scored coefficients for OBP, BB% and K% is given in Figure 4.5.

Table 4.11: Mean coefficients, standard deviations, and 95% credible intervals for each model variable are shown for (A) on-base percentage (OBP), (B) walk rate (BB%), and (C) strikeout rate (K%). Values for which the 95% credible interval excludes zero are bolded.

	(A) OBP				(B) BB%				(C) K%			
	Mean	SD	2.5%	97.5%	Mean	SD	2.5%	97.5%	Mean	SD	2.5%	97.5%
Visual Clarity	-0.24	0.17	-0.59	0.10	-0.15	0.10	-0.35	0.05	-0.08	0.07	-0.21	0.06
Contrast Sensitivity	0.13	0.16	-0.18	0.45	0.04	0.09	-0.14	0.23	**0.14**	**0.06**	**0.02**	**0.26**
Depth Perception	0.19	0.16	-0.12	0.50	**0.21**	**0.10**	**0.02**	**0.40**	-0.12	0.07	-0.26	0.02
Near-Far Quickness	-0.02	0.15	-0.32	0.28	-0.05	0.09	-0.23	0.14	**0.21**	**0.07**	**0.09**	**0.34**
Target Capture	0.15	0.16	-0.16	0.47	-0.16	0.09	-0.35	0.01	**0.16**	**0.06**	**0.04**	**0.29**
Perception Span	**0.64**	0.17	**0.31**	**0.99**	0.15	0.10	-0.04	0.34	**0.34**	**0.07**	**0.21**	**0.47**
Eye-Hand Coordination	0.22	0.17	-0.11	0.56	**0.46**	**0.10**	**0.26**	**0.67**	**-0.19**	**0.07**	**-0.32**	**-0.06**
Reaction Time	0.21	0.17	-0.11	0.55	**0.23**	**0.11**	**0.03**	**0.44**	0.12	0.07	-0.02	0.26
Age	**0.66**	**0.17**	**0.34**	**1.00**	**0.53**	**0.09**	**0.36**	**0.71**	**0.22**	**0.06**	**0.09**	**0.34**
Infield	-0.53	0.31	-1.15	0.08	0.05	0.19	-0.33	0.43	0.65	0.13	0.40	0.91
Catcher	**-1.28**	**0.49**	**-2.25**	**-0.35**	0.15	0.29	-0.40	0.72	0.26	0.19	-0.12	0.64
Intercept	-0.13	0.23	-0.57	0.31	**-0.84**	**0.14**	**-1.12**	**-0.57**	**-0.52**	**0.09**	**-0.71**	**-0.34**

From the OBP model results, we observe that performance on the Perception Span task, which measures the ability to remember and recreate visual patterns, is associated with an increased ability to reach base. Moreover, the size of the coefficient is comparable to that of age (SD = 3.8 years), a remarkable result considering it is well known that older players tend to perform better than younger players in professional baseball due to survivorship bias. For interpretation, suppose player X is a 23-year-old outfielder with completely average abilities as a professional baseball player. The model predicts his OBP in MLB to be 0.292. We expect that a similar player who scores one standard deviation higher on the Perception Span task would have an OBP of 0.300, a nontrivial difference. While the coefficients of the other tasks

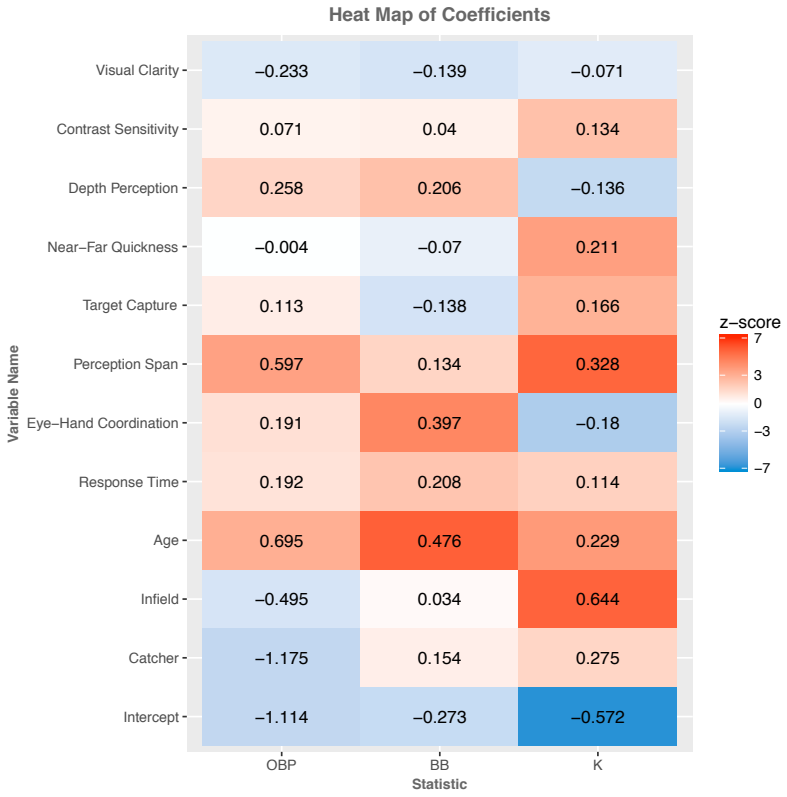


FIGURE 4.5: Heat map of coefficients $\boldsymbol{\beta}$. The darker the color, the closer the posterior tail probability gets to zero (indicating evidence of a linear association).

trend positive, there is simply not enough data to draw strong conclusions about them. Walk rate and strikeout rate are both components of on-base percentage but capture different features of a batter. As such, the visual-motor abilities that affect a player's ability to draw walks should differ from those that affect a player's ability to avoid strikeouts. The results presented in Table 4.11 bear this out. Superior performance on the tasks that measure a player's ability to quickly identify and react to visual stimuli, Eye-Hand Coordination and Reaction Time, were found to be associated with an increased ability to draw walks. For example, our model

predicts player X to obtain a walk rate of 7.1% in the MLB, but predicts a similar player with a one standard deviation superior score on the hand-eye coordination task to have a walk rate of 8.0%. On the other hand, superior performance on tasks that measure a player's spatial recognition and memory, such as Near-Far Quickness, Target Capture, and Perception Span, was found to be associated with an increased ability to avoid strikeouts. In context, our model predicts the strikeout rate of player X to be 23.2% in the MLB. A player similar to player X who scores one standard deviation better on the Perception Span task is predicted to have a strikeout rate of 21.2%. It is surprising that Eye-Hand Coordination was found to be significant in the opposite direction than we would expect *a priori*, which motivates further study.

4.4 Discussion

In this chapter, we explored variation in visual-motor abilities across a sample of 2317 athletes tested on the Nike Sensory Station. Each athlete completed a standardized assessment battery designed to capture distinct visual-motor abilities, such as visual-motor control, visual sensitivity, and eye quickness (Wang et al., 2015). By using principled, Bayesian approaches to model this data, a number of substantial findings emerge.

Of fundamental interest, both the main effects and interaction models in Section 4.2 consistently reveal that visual-motor abilities, especially those with strong visual-motor control demands, are greater at higher levels of athletic achievement. This gradation of expertise level is unique in the literature, given that past research predominantly involves comparing small numbers of athletes to non-athletes on divergent choices of tasks (see Mann et al. (2007) and Voss et al. (2010)). By estimating the performance difference on an identical task battery from athletes ranging from middle school to high school, to college, and to professional level, our study provides a new perspective for understanding visual-motor variability across individuals of

different athletic experience levels.

Within the subdomain of professional baseball, we find that increased levels of many visual-motor abilities are associated with superior measures of on-field performance such as on-base percentage, walk rate, and strikeout rate, as described in Section 4.3. This observation is largely intuitive since it is expected that players draw on these skills to project the location of the pitch through the strike zone and decide whether to swing or not. Conversely, the ability to hit for power, captured by slugging percentage, should have more to do with strength, bat speed, and swing plane than visual-motor abilities. Pitchers rely on a strong arm, consistent mechanics, and a varied repertoire to prevent runs, attributes that are superficially unrelated to visual-motor abilities.

Among the individual tasks tested, Perception Span's relationship with on-field performance produced the strongest relationships, with better scores strongly associated with both increased on-base percentages and reduced strikeout rates. In addition, performance on the Perception Span task exhibits some association with both higher walk rates and increased slugging percentages, though the evidence is not conclusive. This task measures the ability to remember and recreate visual patterns and may reflect visual recognition abilities that have previously been tied to batting performance in small samples of players (Reichow et al., 2011; Szymanski et al., 2015). For reference, across the tested models we find expected age and position effects with stronger batting performance for outfields and older batters. The fact that the Perception Span effects were within the range of magnitudes for age and position effects indicates the relatively strong contribution that visual recognition abilities may play in batting performance. A number of other tasks correlate with higher walk rates and lower strikeout rates as well, including Depth Perception, Eye-Hand Coordination and Reaction Time, though the Eye-Hand Coordination task has a negative estimated strikeout rate coefficient. Overall, while the results suggest per-

formance contributions from both aspects of visual hardware (Contrast Sensitivity and Depth Perception) and visual software (Target Capture, Near-Far Quickness, Perception Span, Eye-Hand Coordination, Reaction Time), on balance it appears that tasks measuring visual software are more predictive of on-field performance in professional baseball, therefore providing further evidence of the importance of these abilities towards on-field achievement in baseball.

The results from Section 4.2 indicate that the visual abilities relevant for on-field performance my differ by sport type. The distinction between strategic and interceptive sports has been made in several meta-analytic reviews (Mann et al., 2007; Voss et al., 2010; Lebeau et al., 2016), reflecting a strong research interest in understanding how competition demands are reflected in athletes underlying abilities. Our findings indicate that athletes playing interceptive sports exhibit better Visual Clarity, Contrast Sensitivity, Near-Far Quickness, and Reaction Times than those playing strategic sports. In contrast, athletes playing strategic sports tend to score higher on the Perception Span task. This suggests that different visual-motor abilities are engaged by the situational demands of each sport type. Specifically, for interceptive sports, the importance of interacting with a fast-moving object may demand an enhanced ability to see the object, distinguish it from its environment, and react to its movement (Davids, 2002). In strategic sports such as soccer, athletes must simultaneously maintain an array of information about teammates, opponents, and the ball. As such, players with a high ability in Perception Span can quickly code and preserve spatial information, obtaining a performance advantage in pattern recognition and recall (Abernethy et al., 1994a), decision-making (Starkes and Ericsson, 2003), and development of team mental model (Mohammed and Dumville, 2001).

The current findings indicate that female athletes are better, on average, at Perception Span and Eye-Hand Coordination than their male counterparts, holding

constant level of expertise and sport type. This is particularly true at the middle school level. The female advantage in Perception Span may be attributed to the combination of rapid development of spatial working memory during the middle school years (Gathercole et al., 2004) and earlier developmental acceleration in females. The female advantage in Eye-Hand Coordination is surprising, though previous studies have found that women are faster at programming a sequence of manual movements (Nicholson and Kimura, 1996) and more accurate at controlling arm movements under time pressure (Liu et al., 2015). Nevertheless, more research is needed to gain a better understanding of these gender differences.

The methods presented in this chapter have many strengths, but also a few limitations. First, while our conclusion are based on analysis a very large sample of athletes, tested under natural settings, the Nike Sensory Stations were not available to all athletes. In addition, the models developed make simplifying assumptions about the data generating process that do not actually hold. For example, Section 4.2.1 assumes that the dependence structure of the tasks can be described by a Gaussian copula, which fails to account for possible tail dependence in the data. Nevertheless, these findings provide quantitative evidence, from a very large, real-world test battery, of domain-specific visual expertise in athletes. By analyzing data collected on athletes, spanning from developing adolescents to many of the most elite professional athletes in the world, this chapter provides a unique lens into the visual and motor capabilities that differentiate individuals with different levels of expertise and types of athletic experience, particularly within the sport of baseball. While we do not attempt to infer causal relationships, these findings do open intriguing questions about the influences of nature and nurture on athletic achievement and may provide useful metrics for the scouting of players through their developmental trajectory.

5

Conclusions

The contributions of this thesis are primarily made in two sub-disciplines: survey methodology and sports science. Chapter 2 extends the confidence interval methodology developed by Yu and Hoff (2018) to the types of statistical models used in small-area estimation in order to borrow information across groups. By taking into account covariate and/or spatiotemporal autocorrelation, practitioners can use this methodology to further reduce confidence interval width while maintaining the property of $1 - \alpha$ frequentist coverage for all possible values of the target quantity.

Although the proposed procedure is applicable when the sampling model is assumed to be normal, it is not when the response variable is discrete. Because count data is pervasive in the small-area estimation literature, our work can be further extended by developing interval methodology for discrete observations that maintains nominal area-specific coverage.

In Chapter 3, we develop methodology for multiple imputation of mixed numeric data using a joint statistical model. We also extend the approach to a situation in which the support of the data is known to be a strict subset of the data product space. Unlike other multiple imputation methods, our approach guarantees that

completed observations consisting of binary, ordinal, and continuous variables will satisfy expert-defined constraints. This is particularly important for statistical agencies who wish to make their data available for public consumption, as infeasible data values may reduce public trust of the released data.

Future research is needed to extend the method to incorporate and/or impute categorical variables, which can help make the proposed approach more broadly applicable. As the model is particularly sensitive to the existence of marginal outliers, a framework to simultaneously detect and impute erroneous data values (Kim et al., 2015; Manrique-Vallier and Reiter, 2017) could be of additional use to a statistical organization. Finally, it could be of use to examine how survey weights could be incorporated into model estimation, not simply post-imputation inference.

In Chapter 4, we analyze a dataset of visual-motor assessments completed by athletes, of which nearly 2,000 are collegiate or professional athletes. We begin by examining the relationships between the tasks, finding that they cleanly separate into visual hardware and visual software, consistent with the demarcation present in the literature. Moreover, systematic differences in task performance are present by athlete level, gender, and primary sport type. For the subset of professional baseball players in the sample, we also collected game statistics in the season after testing in an effort to explore associations between visual-motor ability and on-field performance in professional baseball. In particular, we find positive relationships between many of the task scores and measures of plate discipline, such as strikeout rate, walk rate, and on-base percentage, holding constant position, age, and level of competition. We do not find meaningful relationships between the task scores and other measures of baseball performance, such as slugging percentage and fielder-independent pitching.

Our lab is continuing efforts to further understand the importance of visual-motor expertise in the perception-action cycle and how that translates into performance in physically demanding activities. In particular, we are currently leveraging data

obtained eye-tracking technology to obtain greater insights about the specific reactionary abilities that elite baseball players need to possess. We are also conducting studies with collegiate baseball teams to assess performance gains that can be attributed to visual-motor training interventions. Some of the conclusions drawn from this work may be applicable in other related domains, such as training in other athletic sports and the military.

In sum, this thesis not only provides a contribution to the sports vision literature, but also provides general statistical methodology for other practitioners to use as they seek to perform inferences and make decisions in the face of uncertainty.

Appendix A

Appendix to Chapter 2

A.1 Credible interval coverage rates for the Fay Herriot model

Under the sampling model $y_j \sim N(\theta_j, \sigma_j^2)$ and prior $\theta_j \sim N(\boldsymbol{x}_j^\top \boldsymbol{\beta}, \tau^2)$, the posterior distribution of θ_j is

$$\theta_j \mid y_j \sim N\left(\frac{\tau^2 y_j + \sigma_j^2 \boldsymbol{x}_j^\top \boldsymbol{\beta}}{\sigma_j^2 + \tau^2}, \frac{\sigma_j^2 \tau^2}{\sigma_j^2 + \tau^2}\right).$$

Accordingly, the $1-\alpha$ symmetric credible interval C_B^j can be expressed as

$$C_B^j(\boldsymbol{y}) = \left\{\theta : \frac{\tau^2 y_j + \sigma^2 \boldsymbol{x}_j^\top \boldsymbol{\beta}}{\sigma_j^2 + \tau^2} + \frac{\sigma_j \tau}{\sqrt{\sigma_j^2 + \tau^2}} z_{\alpha/2} < \theta \right.$$
$$\left. < \frac{\tau^2 y_j + \sigma_j^2 \boldsymbol{x}_j^\top \boldsymbol{\beta}}{\sigma_j^2 + \tau^2} + \frac{\sigma_j \tau}{\sqrt{\sigma_j^2 + \tau^2}} z_{1-\alpha/2} \right\}.$$

For a given value of θ_j, the coverage probability is

$$\begin{aligned}
\Pr(\theta_j \in C_B^j \mid \theta_j) &= \Pr\left(\frac{\tau^2 y_j + \sigma_j^2 \boldsymbol{x}_j^\top \boldsymbol{\beta}}{\sigma_j^2 + \tau^2} + \frac{\sigma_j \tau}{\sqrt{\sigma_j^2 + \tau^2}} z_{\alpha/2} < \theta_j \right. \\
&\qquad \left. < \frac{\tau^2 y_j + \sigma_j^2 \boldsymbol{x}_j^\top \boldsymbol{\beta}}{\sigma_j^2 + \tau^2} + \frac{\sigma_j \tau}{\sqrt{\sigma_j^2 + \tau^2}} z_{1-\alpha/2} \right) \\
&= \Pr\left(\frac{\sigma_j(\theta_j - \boldsymbol{x}_j^\top \boldsymbol{\beta})}{\tau^2} + z_{\alpha/2}\sqrt{1 + \sigma_j^2/\tau^2} < \frac{y_j - \theta_j}{\sigma_j} \right. \\
&\qquad \left. < \frac{\sigma_j(\theta_j - \boldsymbol{x}_j^\top \boldsymbol{\beta})}{\tau^2} + z_{1-\alpha/2}\sqrt{1 + \sigma_j^2/\tau^2} \right) \\
&= \Phi\left(\frac{\sigma_j(\theta_j - \boldsymbol{x}_j^\top \boldsymbol{\beta})}{\tau^2} + z_{1-\alpha/2}\sqrt{1 + \sigma_j^2/\tau^2} \right) \\
&\qquad - \Phi\left(\frac{\sigma_j(\theta_j - \boldsymbol{x}_j^\top \boldsymbol{\beta})}{\tau^2} + z_{\alpha/2}\sqrt{1 + \sigma_j^2/\tau^2} \right),
\end{aligned}$$

where Φ is the standard normal cumulative distribution function.

A.2 Computation of FAB intervals

Suppose that we have prior information about a parameter θ_j, encoded by a normal probability distribution $\theta_j \sim N(\mu_j, \tau_j^2)$. For all increasing functions $s_j : \mathbb{R} \to [0, 1]$ the interval $C_{s_j}^j$ given by (2.9) corresponds to a valid $1 - \alpha$ frequentist confidence region. We wish to choose s_j in a way that minimizes the Bayes risk, which we define as the prior expected interval width:

$$R(s_j \mid \sigma_j^2, \mu_j, \tau_j^2) = \int_{\mathbb{R}} p(y_j \mid \theta_j, \sigma_j^2)\ p(\theta_j \mid \mu_j, \tau_j^2)\ \left[\int_{\mathbb{R}} I(\theta \in C_{s_j}^j(y_j))\ d\theta \right] d\theta_j.$$

It is important to note the distinction between θ, a hypothesized value for the jth area mean, and θ_j, the actual mean for the jth area, as s_j depends on θ and not

on θ_j. The prior predictive distribution for y_j is $N(\mu_j, \tau_j^2 + \sigma_j^2)$. From this, we can rewrite the above expression as

$$R(s_j \mid \sigma_j^2, \mu_j, \tau_j^2) = \int_{\mathbb{R}} p(y_j \in A(s_j, \theta) \mid \sigma_j^2, \mu_j \tau_j^2) d\theta$$

$$A(s_j, \theta) = \left\{ y_j : z_{\alpha(1-s_j(\theta))} < \frac{\theta - y_j}{\sqrt{\sigma_j^2 + \tau_j^2}} < z_{1-\alpha s_j(\theta)} \right\}.$$

Expanding, we have that

$$R(s_j \mid \sigma_j^2, \mu_j, \tau_j^2) = \int_{\mathbb{R}} \Phi\left(\frac{\theta - \mu_j - \Phi^{-1}(\alpha(1 - s_j(\theta)))}{\sqrt{\sigma_j^2 + \tau_j^2}} \right)$$

$$- \Phi\left(\frac{\theta - \mu_j - \Phi^{-1}(1 - \alpha s_j(\theta))}{\sqrt{\sigma_j^2 + \tau_j^2}} \right) d\theta.$$

For a fixed θ, let $\omega = s_j(\theta)$ and let $H'(\omega)$ be the derivative of the integrand with respect to ω. Then a critical point satisfies

$$\frac{2\sigma_j(\theta - \mu_j)}{\tau_j^2} = \Phi^{-1}(\alpha s_j(\theta)) - \Phi^{-1}(\alpha(1 - s_j(\theta))).$$

For each θ, there is only one possible s_j that satisfies the above equation, given by (2.10). This s_j is an increasing function on θ, which means that $C^j_{s_j}$ is a confidence interval. From this, the lower and upper bounds of the FAB z-interval can be found by solving two non-linear equations:

$$\theta^U = \frac{y_j + \sigma_j \Phi^{-1}(1 - \alpha + \Phi(\frac{y_j - \theta^U}{\sigma_j}))}{1 + 2\sigma_j^2/\tau_j^2} + \mu_j \frac{2\sigma_j^2/\tau_j^2}{1 + 2\sigma_j^2/\tau_j^2}$$

$$\theta^L = \frac{y_j + \sigma_j \Phi^{-1}(\alpha - \Phi(\frac{\theta^L - y_j}{\sigma_j}))}{1 + 2\sigma_j^2/\tau_j^2} + \mu_j \frac{2\sigma_j^2/\tau_j^2}{1 + 2\sigma_j^2/\tau_j^2}.$$

Calculation of the FAB t-interval is similar, but is more complicated since the optimal s_j cannot be found analytically. Given a $N(\mu_j, \tau_j^2)$ prior on θ_j and a $IG(a, b)$ prior on σ^2, we find s_j that minimizes the Bayes risk

$$R(s_j \mid \mu_j, \tau_j^2, a, b) = \int_{\mathbb{R}} \left[\int_{\mathbb{R}} p(y_j \mid \theta_j, \sigma_j^2)\ p(\theta_j \mid \mu_j, \tau_j^2)\ p(\sigma_j^2 \mid a, b)\ \ d\theta_j\ d\sigma_j^2 \right] \times I(\theta \in C_{s_j}^j(y_j)) d\theta.$$

Evaluation and optimization of the above expression involves numerical integration with respect to σ_j^2. Once the optimal s_j is found, the lower and upper bounds of the confidence interval are found by solving the two non-linear equations

$$\theta^U = F\left(\frac{y_j - \theta^U}{s_j/\sqrt{n_j}}\right) = \alpha s_j(\theta^U)$$

$$\theta^L = F\left(\frac{y_j - \theta^U}{s_j/\sqrt{n_j}}\right) = 1 - \alpha(1 - s_j(\theta^L)),$$

where F corresponds to the CDF of a t distribution with $n_j - 1$ degrees of freedom.

A.3 ML estimation of spatial Fay-Herriot hyperparameters

To estimate the hyperparameters $\{\boldsymbol{\beta}, \rho, \tau^2\}$ based on data from a subset of areas S, where $j \notin S$, we recommend using either ML or REML procedures based on the data from all areas in S. We provide the details for ML estimation below, although REML estimation is straightforward, using transformed data $\boldsymbol{y}_S^* = \boldsymbol{F}^T \boldsymbol{y}_S$, where $\boldsymbol{F}$ is a $(m-1) \times (m-p)$ matrix that is orthogonal to $\boldsymbol{X}_S$. For more details about REML estimation for the spatial Fay-Herriot model, see Pratesi and Salvati (2008). Both ML and REML estimation of the spatial Fay-Herriot model are implemented in the `sae` R package and we also provide an implementation in the replication code.

Defining the marginal variance $\boldsymbol{V}_S = \boldsymbol{D}_S + \boldsymbol{G}_S$, where $\boldsymbol{G}_S = \tau^2[(\boldsymbol{I} - \rho \boldsymbol{W}_S)(\boldsymbol{I} - \rho \boldsymbol{W}_S^T)]^{-1}$, the log-likelihood function is given by

$$\ell(\boldsymbol{\beta}, \rho, \tau^2) = \text{const} - \frac{1}{2} \log |\boldsymbol{V}_S| - \frac{1}{2} (\boldsymbol{y}_S - \boldsymbol{X}_S \boldsymbol{\beta})^T \boldsymbol{V}_S^{-1} (\boldsymbol{y}_S - \boldsymbol{X}_S \boldsymbol{\beta}).$$

The MLE $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is of a familiar form, with

$$\hat{\boldsymbol{\beta}}(\rho, \tau^2) = (\boldsymbol{X}_S^T \boldsymbol{V}_S^{-1} \boldsymbol{X}_S)^{-1} \boldsymbol{X}_S^T \boldsymbol{V}_S^{-1} \boldsymbol{y}_S.$$

The partial derivatives with respect to τ^2 and ρ are given by $s(\boldsymbol{\beta}, \tau^2, \rho)$, where

$$\begin{aligned}
s_{\tau^2}(\boldsymbol{\beta}, \tau^2, \rho) &= \frac{\partial \ell}{\partial \tau^2} \\
&= -\frac{1}{2}\mathrm{tr}(\boldsymbol{V}_S^{-1}\boldsymbol{C}_S^{-1}) + \frac{1}{2}(\boldsymbol{y}_S - \boldsymbol{X}_S\boldsymbol{\beta})^T(\boldsymbol{V}_S^{-1}\boldsymbol{C}_S^{-1}\boldsymbol{V}_S^{-1})(\boldsymbol{y}_S - \boldsymbol{X}_S\boldsymbol{\beta}) \\
s_{\rho}(\boldsymbol{\beta}, \tau^2, \rho) &= \frac{\partial \ell}{\partial \rho} \\
&= -\frac{1}{2}\mathrm{tr}(\tau^2\boldsymbol{V}_S^{-1}(\boldsymbol{C}_S^{-1}[\boldsymbol{W}_S + \boldsymbol{W}_S^T - 2\rho\boldsymbol{W}_S\boldsymbol{W}_S^T]\boldsymbol{C}_S^{-1})) \\
&\quad + \frac{\tau^2}{2}(\boldsymbol{y}_S - \boldsymbol{X}_S\boldsymbol{\beta})^T(\boldsymbol{V}_S^{-1}(\boldsymbol{C}_S^{-1}[\boldsymbol{W}_S + \boldsymbol{W}_S^T - 2\rho\boldsymbol{W}_S\boldsymbol{W}_S^T]\boldsymbol{C}_S^{-1})\boldsymbol{V}_S^{-1}) \\
&\quad \times (\boldsymbol{y}_S - \boldsymbol{X}_S\boldsymbol{\beta}),
\end{aligned}$$

where $\boldsymbol{C}_S = (\boldsymbol{I} - \rho\boldsymbol{W}_S)(\boldsymbol{I} - \rho\boldsymbol{W}_S^T)$. We can then use these to calculate the Fisher information matrix, which is the matrix of expected second derivatives of $-\ell$.

$$\mathcal{I}(\tau^2, \rho) = \begin{bmatrix} \frac{1}{2}\mathrm{tr}\left(\boldsymbol{V}_S^{-1}\boldsymbol{C}_S^{-1}\boldsymbol{V}_S^{-1}\boldsymbol{C}_S^{-1}\right) & \frac{1}{2}\mathrm{tr}(\boldsymbol{V}_S^{-1}\boldsymbol{C}_S^{-1}\boldsymbol{V}_S^{-1}\boldsymbol{A}_S) \\ \frac{1}{2}\mathrm{tr}(\boldsymbol{V}_S^{-1}\boldsymbol{C}_S^{-1}\boldsymbol{V}_S^{-1}\boldsymbol{A}_S) & \frac{1}{2}\mathrm{tr}(\boldsymbol{V}_S^{-1}\boldsymbol{A}_S\boldsymbol{V}_S^{-1}\boldsymbol{A}_S) \end{bmatrix}$$

where $\boldsymbol{A}_S = \tau^2\boldsymbol{C}_S^{-1}[\boldsymbol{W}_S + \boldsymbol{W}_S^T - 2\rho\boldsymbol{W}_S]\boldsymbol{C}_S^{-1}$. From this, we can solve for the maximum likelihood estimates of τ^2 and ρ by Fisher's scoring.

$$[\tau^2, \rho]^{(t+1)} = \mathcal{I}^{-1}\left([\tau^2, \rho]^{(t)}\right) \cdot s\left(\hat{\boldsymbol{\beta}}([\rho, \tau^2]^{(t)}), [\tau^2, \rho]^{(t)}\right)$$

where s is the 2×1 matrix of first partial derivatives with respect to τ^2 and ρ. Since τ^2 and ρ are constrained to lie in the intervals $(0, \infty)$ and $(-1, 1)$, we reduce the step size if the proposed Fisher scoring step violates one or more constraints. The algorithm iterates until convergence.

To obtain estimates of the subset of area means $\boldsymbol{\theta}_S$, we find their conditional means given $\boldsymbol{y}_S$, $\hat{\boldsymbol{\beta}}$, and $\hat{\boldsymbol{\psi}}$ under the sampling and linking model. These can be expressed as

$$\hat{\boldsymbol{\theta}}_S(\hat{\boldsymbol{\psi}}) = \boldsymbol{X}_S\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\psi}}) + \boldsymbol{G}_S(\boldsymbol{\psi})\boldsymbol{V}_S(\boldsymbol{\psi})^{-1}(\boldsymbol{y}_S - \boldsymbol{X}_S\hat{\boldsymbol{\beta}}(\boldsymbol{\psi})).$$

A.4 ML Estimation of sampling variance hyperparameters

Suppose that the sampling model for the unbiased direct estimates of the area-specific sampling variances is

$$\frac{(n_j-1)\hat{\omega}_j^2}{\omega_j^2} \sim \chi^2_{n_j-1},$$

where $\hat{\omega}_j^2$ is an unbiased and consistent estimate of the sampling variance ω_j^2, based on a sample of n_j observations. We model the variances of log-radon levels hierarchically, under the assumption that

$$1/\omega_j^2 \sim G(a,b), \qquad j = 1, \cdots, m$$

For each area j, we are interested in estimating a and b via maximum likelihood, based on data from a subset of areas S, where $j \notin S$. Then the log-likelihood is

$$\begin{aligned}
\ell(a,b) &= \sum_{k\in S} \log\left(\int p(\hat{\omega}_k^2 \mid \phi_k)\ p(\phi_k \mid a,b)\ d\phi_k\right) \\
&= \text{const} + \sum_{k\in S} \log \int \left(\frac{b^a}{\Gamma(a)}\phi_k^{\frac{n_k-1}{2}+a-1}\exp\left(-\phi_k\left(\frac{n_k-1}{2}\hat{\omega}_k^2 + b\right)\right)\ d\phi_k\right) \\
&= \text{const} + |S|(a\log(b) - \log\Gamma(a)) \\
&\qquad + \sum_{k\in S}\left[\log\Gamma\left(\frac{n_k-1}{2}+a\right) - \left(\frac{n_k-1}{2}+a\right)\log\left(\frac{n_k-1}{2}\hat{\omega}_k^2 + b\right)\right],
\end{aligned}$$

where $\phi_k = 1/\omega_k^2$, $|S|$ is the cardinality of S and $\Gamma(\cdot)$ is the Gamma function.

The partial derivatives of the log-likelihood with respect to a and b are

$$\frac{\partial \ell}{\partial a} = |S|(\log(b) - \psi(a)) + \sum_{k \in S} \left[\psi\left(\frac{n_k - 1}{2} + a \right) - \log\left(\frac{n_k - 1}{2} \hat{\omega}_k^2 + b \right) \right]$$

$$\frac{\partial \ell}{\partial b} = \frac{|S|a}{b} - \sum_{k \in S} \left[\frac{\frac{n_k - 1}{2} + a}{\frac{n_k - 1}{2} \hat{\omega}_k^2 + b} \right],$$

where ψ is the digamma function, the derivative of the log-gamma function. In Section 2.4, we use the L-BFGS optimization algorithm with the box constraint $\{(0, \infty) \times (0, \infty\}$ to find $\hat{a}$ and $\hat{b}$, the maximum likelihood estimates of a and b. Due to the low dimensionality of the problem, second order information can be utilized to speed up convergence, and the second order partial derivatives are

$$\frac{\partial^2 \ell}{\partial^2 a} = \sum_{k \in S} \psi^{'} \left(\frac{n_k - 1}{2} + a \right) - |S| \psi^{'}(a)$$

$$\frac{\partial^2 \ell}{\partial a \partial b} = \frac{|S|}{b} - \sum_{k \in S} \frac{1}{\frac{n_k - 1}{2} s_k^2 + b}$$

$$\frac{\partial^2 \ell}{\partial^2 b} = \sum_{k \in S} \left[\frac{\frac{n_k - 1}{2} + a}{\left(\frac{n_k - 1}{2} s_k^2 + b \right)^2} \right] - \frac{|S|a}{b^2},$$

where $\psi^{'}$ is the trigamma function. An optimization algorithm that uses first-order information about a and b is implemented in the replication code.

Appendix B

Appendix to Chapter 3

B.1 Gibbs Sampling Steps for the Constrained Joint Transformation Model

Given current values of model parameters;
for $k = 1, \cdots, K$ **do**
  Set $N_k = \sum_{i=1}^{n} I(H_i = k)(1 + W_i)$;
  Set $\bar{\boldsymbol{z}}_k = \sum_{i:H_i=k}(\boldsymbol{z}_i + \sum_{w=1}^{W_i} \boldsymbol{z}^c_{i(w)})/N_k$;
  Set $\boldsymbol{S}_k = \sum_{i:H_i=k}(\boldsymbol{z}_i - \bar{\boldsymbol{z}}_k)(\boldsymbol{z}_i - \bar{\boldsymbol{z}}_k)^\top + \sum_{w=1}^{W_i}(\boldsymbol{z}^c_{i(w)} - \bar{\boldsymbol{z}}_k)(\boldsymbol{z}^c_{i(w)} - \bar{\boldsymbol{z}}_k)^\top$;
  Set $\boldsymbol{\Omega}_k = (N_k \boldsymbol{\Sigma}_k^{-1} + h\boldsymbol{I})^{-1}$;
  Sample $\boldsymbol{\mu}_k \sim N\left(\boldsymbol{\Omega}_k(N_k \boldsymbol{\Sigma}_k^{-1} \bar{\boldsymbol{z}} + h\boldsymbol{\mu}_0), \boldsymbol{\Omega}_k\right)$;
  Sample $\boldsymbol{\Sigma}_k \sim$ Inverse-Wishart$(\nu + N_k, \boldsymbol{\Sigma}_0 + \boldsymbol{S}_k)$;
  if $k \neq K$ **then**
    Sample $v_k \sim$ Beta$(1 + \sum_{i=1}^{n} I(H_i = k), \alpha + \sum_{\ell>k} \sum_{i=1}^{n} I(H_i = \ell))$;
  else
    Set $v_k = 1$;
  end
  Set $\pi_k = v_k \prod_{\ell<k}(1 - v_k)$;
end
Sample $\alpha \sim$ Gamma$(a_\alpha + K - 1, b_\alpha - \log \pi_k)$;

for $i = 1, \cdots, n$ **do**
  for $k = 1, \cdots, K$ **do**
    Set $\phi_{ik} = \pi_k N(\boldsymbol{z}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \prod_{w=1}^{W_i} N(\boldsymbol{z}_{i(w)}^c \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$;
  end
  Sample $H_i \sim \text{Multinomial}\left(\frac{\phi_{i1}}{\sum_{k=1}^K \phi_{ik}}, \cdots, \frac{\phi_{iK}}{\sum_{k=1}^K \phi_{ik}}\right)$;
end
for $i = 1, \cdots, n$ **do**
  Set accept = False;
  Set $W_i = 0$;
  while *accept = False* **do**
    Sample $\boldsymbol{z}^* \sim N(\boldsymbol{z} \mid \boldsymbol{\mu}_{H_i}, \boldsymbol{\Sigma}_{H_i})$;
    if $\boldsymbol{z}^* \in \mathcal{D}$ **then**
      accept = True;
    else
      Set $\boldsymbol{z}_{i(w)}^c = \boldsymbol{z}^*$;
      Set $W_i = W_i + 1$;
    end
  end
end
for $i = 1, \cdots, n$ **do**
  accept = False;
  Calculate $\boldsymbol{m}_{i,mis}$ and $\boldsymbol{\Omega}_{i,mis}$, the conditional mean and variance of the missing values given $\boldsymbol{\mu}_{H_i}$, $\boldsymbol{\Sigma}_{H_i}$, and $\boldsymbol{z}_{i,obs}$;
  while *accept = False* **do**
    Sample $\boldsymbol{z}_{i,mis} \sim N(\boldsymbol{m}_{i,mis}, \boldsymbol{\Omega}_{i,mis})$;
    if $(\boldsymbol{z}_{i,obs}, \boldsymbol{z}_{i,mis}) \in \mathcal{D}$ **then**
      accept = True;
    end
  end
  for $j = 1, \cdots, p$ **do**
    if $r_{ij} = 1$ **then**
      Set m_{ij} equal to the conditional mean of z_{ij}, given $z_{i,-j}$, $\boldsymbol{\mu}_{H_i}$, $\boldsymbol{\Sigma}_{H_i}$;
      Set σ_{ij}^2 equal to the conditional variance of z_{ij}, given $z_{i,-j}$, $\boldsymbol{\Sigma}_{H_i}$;
      Set $a = \inf_z \hat{g}_j(z) = y_{ij}$;
      Set $b = \sup_z \hat{g}_j(z) = y_{ij}$;
      Sample $z_{ij} \sim \text{TN}(m_{ij}, \sigma_{ij}^2, (a, b))$;
    end
    Set $y_{ij} = \hat{g}_j(z_{ij})$;
  end
end

Algorithm 3: Gibbs sampling iteration to update parameters for the constrained joint transformation model

B.2 ACS Regression Results

Table B.1: Linear model results with log(Income) as the response variable. Regression coefficients estimated after imputation via predictive mean matching appear to suffer from attenuation bias and have higher standard errors than when based on the JTM and CJTM imputation methods.

	Actual Data est/se	PMM est/se	JTM est/se	CJTM est/se
Intercept	6.5047 (0.1082)	7.6900 (0.1440)	6.6400 (0.1190)	6.6401 (0.1504)
Age	0.0262 (0.0012)	0.0160 (0.0017)	0.0255 (0.0015)	0.0243 (0.0015)
Disability	−0.0983 (0.0443)	−0.1630 (0.0533)	−0.1120 (0.0521)	−0.0706 (0.0473)
Education Level	0.0700 (0.0049)	0.0410 (0.0067)	0.0694 (0.0060)	0.0712 (0.0061)
Hours Worked	0.0339 (0.0011)	0.0292 (0.0013)	0.0322 (0.0013)	0.0319 (0.0013)
Medicaid	−0.0859 (0.0516)	−0.1740 (0.0719)	−0.140 (0.0582)	−0.1690 (0.0598)
Medicare	0.2752 (0.0583)	0.3860 (0.0870)	0.2390 (0.0672)	0.2834 (0.0633)

Appendix C

Appendix to Chapter 4

C.1 Visual-Motor Assessments

The Sensory Stations consist of a battery of nine computerized sensorimotor tasks (illustrated in Figure C.1), each designed to evaluate a specific facet of a participants visual-motor abilities. The first five tasks were completed using a handheld Apple iPod Touch, standing 4.9 m from the station. The last four tasks were completed at arms length from the touchscreen monitor. Four of the tasks Visual Clarity, Contrast Sensitivity, Depth Perception, and Target Capture operated on staircase schedules in which subsequent stimulus difficulty increased following a correct response and decreased following an incorrect response. For these tasks, scores were calculated as the final step according to response accuracy on the staircase schedule. All tasks were preceded by video instructions. Procedures and descriptions for each task are provided below, and detailed descriptions can be found in Erickson et al. (2011) and Wang et al. (2015).

- The Visual Clarity task measures visual acuity for fine details at a distance using a black Landolt ring an incomplete ring with a small gap oriented in one

of the four cardinal directions. Participants were asked to swipe on the iPod in the direction that corresponded to the orientation of the gap in the ring. The task was completed in three separate rounds: one with an occluder covering the right eye, then the left, then a final round with both eyes uncovered. Visual Clarity scores were taken as the average of these three conditions.

- The Contrast Sensitivity task measures the minimum resolvable difference in contrast at a distance. Participants were presented with four black rings on a light gray background and asked to indicate which ring contained a pattern of dark gray concentric circles by swiping on the iPod in the direction corresponding to the patterned ring.

- The Depth Perception task measures how quickly and accurately participants are able to detect differences in depth at a distance using liquid crystal glasses. Here, four black rings were presented and participants were asked to swipe in the direction of the ring that appeared to have depth. The task was completed three times: once facing towards the screen, once facing to the left and looking over the right shoulder, and once facing right and looking over the left shoulder. Depth Perception scores were taken as the average of these three conditions.

- The Near-Far Quickness task measures the number of near and far targets that can be correctly reported in 30 seconds. Participants aligned the top of the iPod with the bottom edge of the large monitor then swiped in the direction of the gap in the Landolt ring that appeared on either the iPod or larger monitor screen. Participants were instructed to respond as quickly as possible, and the ring only moved from one screen to another following a correct response. Participants continued to respond until they answered correctly or ran out of time. Near-Far Quickness scores were the total number of correct responses

made in 30-seconds.

- The Target Capture task measures the speed at which participants can shift attention and recognize peripheral targets. A small black Landolt ring was briefly presented in one of the four corners of the monitor, and participants were asked to swipe on the iPod in the direction corresponding to the gap in the ring. Following a correct answer, the ring was presented for a shorter duration, and for a longer duration following an incorrect answer, per the staircase procedure. Because this task was performed on a duration staircase, the final accuracy step reflected the minimum stimulus duration according to accuracy on the staircase schedule.

- The Perception Span task measures the capacity of spatial working memory. As participants stood at arms length from the monitor, a grid of empty black circles was presented, and a subset was filled briefly with green dots that disappeared after 100 milliseconds. Participants were asked to recreate the pattern on each trial by touching the circles that had previously contained the green dots. There were eleven total possible pseudo-randomized trials with increased grid sizes and increasing number of green dots presented at each level. Perception Span scores were computed as the total number of correctly identified dots minus the number of missed or falsely identified dots across all of the trials.

- The Eye Hand Coordination task measures the speed at which participants can make visually-guided hand responses to rapidly changing targets. A grid of 48 evenly spaced black rings was presented on the screen. When a green dot appeared in one of the rings, participants touched the dot as quickly as possible. The dot then relocated to another ring for a total succession of 96 dots. The score for Eye Hand Coordination was the total time it took to complete the

sequence.

- The Go/No-Go task measures the ability to execute and inhibit visually guided hand responses in the presence of go and no-go stimuli. Similar to the previous task, a grid of 48 rings was presented; however, in this task the dots could appear either green or red. Participants tried to touch the green dots as quickly as possible while avoiding red dots. 96 dots were presented for 500 milliseconds each before disappearing, and the total score was calculated as the number of green dots touched minus the number of red dots touched.

- The Response Time task measures how quickly participants react and respond to a simple visual stimulus. Two rings were shown on each side of the large monitor. Participants began with their dominant hand in the starting ring, while their body was oriented in front of the landing ring on the opposite side of the screen. When the landing ring turned green, participants moved their hand from the starting ring to the landing ring as quickly and accurately as possible. A total of seven separate trials were completed, and participants had the opportunity to repeat up to two of these trials if any were slower than two standard deviations from the mean. Response Time scores were taken as the average of the seven best trials.

C.2 Gibbs Sampling for the Extended Rank Likelihood

As described in Section 4.2.1, we can use the extended rank likelihood $p(\boldsymbol{Z} \in D \mid \boldsymbol{B}, \boldsymbol{C})$ to perform Bayesian inference on the copula parameters $\boldsymbol{B}$ and $\boldsymbol{C}$. This can be done by constructing a Markov chain with stationary distribution equal to $p(\boldsymbol{B}, \boldsymbol{C} \mid \boldsymbol{Z} \in D) = p(\boldsymbol{Z} \in D \mid \boldsymbol{B}, \boldsymbol{C}) \times p(\boldsymbol{B}, \boldsymbol{C})$. However, there is not a conjugate prior for $\boldsymbol{B}, \boldsymbol{C}$, so we follow Hoff (2007); Murray et al. (2013); Wang et al. (2017) in employing a parameter expansion approach introduced by Liu and Wu (1999).

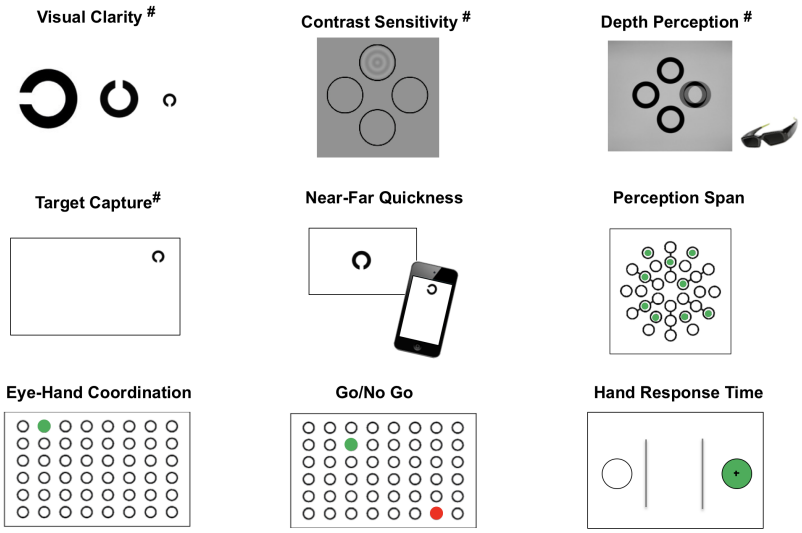


FIGURE C.1: Illustrations of the nine perceptual and visual-motor tasks included in the Nike SPARQ Sensory Station battery. # indicates tasks that performed under a staircase schedule.

Specifically, we put a matrix-normal/inverse-Wishart prior on unscaled parameters $\tilde{\boldsymbol{B}}$, $\boldsymbol{\Sigma}$ and then rescale them at each iteration such that $\boldsymbol{\Sigma}$ becomes a correlation matrix $\boldsymbol{C}$. The parameter-expanded model can be conceptualized as follows:

$$\begin{aligned}
\boldsymbol{\Sigma} &\sim \text{Inverse-Wishart}\left(\nu_0, \nu_0 \boldsymbol{\Sigma}_0\right) \\
\tilde{\boldsymbol{B}} \mid \boldsymbol{\Sigma} &\sim \text{Normal}\left(\boldsymbol{B}_0, \boldsymbol{\Sigma} \otimes \boldsymbol{\Lambda}_0^{-1}\right) \\
\tilde{\boldsymbol{Z}} &\sim \text{Normal}\left(\boldsymbol{X}\tilde{\boldsymbol{B}}, \boldsymbol{\Sigma} \otimes \boldsymbol{I}\right) \\
\boldsymbol{Z}_{ij} &= \tilde{\boldsymbol{Z}}_{ij} / \sqrt{\boldsymbol{\Sigma}_{jj}} \\
\boldsymbol{Y}_{ij} &= F_j^{-1}\left(\Phi\left(\boldsymbol{Z}_{ij}\right)\right)
\end{aligned} \tag{C.1}$$

It is readily seen that under this model,

$$\begin{aligned}
\boldsymbol{Z} &\sim \text{Normal}\left(\boldsymbol{X}\boldsymbol{B}, \boldsymbol{C} \otimes \boldsymbol{I}\right), \\
\boldsymbol{C}_{ij} &= \frac{\boldsymbol{\Sigma}_{ij}}{\sqrt{\boldsymbol{\Sigma}_{ii}\boldsymbol{\Sigma}_{jj}}} \\
\boldsymbol{B}_{ij} &= \frac{\tilde{\boldsymbol{B}}_{ij}}{\sqrt{\boldsymbol{\Sigma}_{jj}}}.
\end{aligned} \tag{C.2}$$

As such, posterior samples of $\boldsymbol{B}$ and $\boldsymbol{C}$ can be obtained by converting corresponding samples of $\boldsymbol{\Sigma}$, $\tilde{\boldsymbol{B}}$, $\tilde{\boldsymbol{Z}}$. Notice that the ordering D of $\tilde{\boldsymbol{Z}}$ must be the same as that of $\boldsymbol{Z}$. Using conjugacy results from the matrix-normal/Inverse-Wishart distributions (Rossi et al., 2005), we can construct a Gibbs sampler, the steps of which are outlined in Algorithm 4.

Specify initial values $\boldsymbol{\Sigma}^{(0)}$, $\tilde{\boldsymbol{B}}^{(0)}$;
for $t = 1, \cdots, T$ **do**
  Set $\tilde{\boldsymbol{Z}}^{(t)} = \tilde{\boldsymbol{Z}}^{(t-1)}$;
  for $j = 1, \cdots, p$ **do**
    Compute $\sigma_j^2 = \boldsymbol{\Sigma}^{(t-1)}_{[j,j]} - \boldsymbol{\Sigma}^{(t-1)}_{[j,-j]} \left(\boldsymbol{\Sigma}^{(t-1)}_{[-j,-j]}\right)^{-1} \boldsymbol{\Sigma}^{(t-1)}_{[-j,j]}$;
    for *each* $y \in$ *unique*$(\boldsymbol{Y}_{1j}, \cdots, \boldsymbol{Y}_{nj})$ **do**
      Compute $z_l = \max\left\{\tilde{\boldsymbol{Z}}^{(t-1)}_{ij} : \boldsymbol{Y}_{ij} < y\right\}$ and
       $z_u = \min\left\{\tilde{\boldsymbol{Z}}^{(t-1)}_{ij} : y < \boldsymbol{Y}_{ij}\right\}$;
      for *each* i *such that* $\boldsymbol{Y}_{ij} = y$ **do**
        Compute
         $\mu_{ij} = \boldsymbol{X}\boldsymbol{B}_{ij} + \left(\boldsymbol{Z}^{(t)}_{[i,-j]} - \boldsymbol{X}\boldsymbol{B}_{[i,-j]}\right) \boldsymbol{\Sigma}^{(t-1)}_{[j,-j]} \left(\boldsymbol{\Sigma}^{(t-1)}_{[-j,-j]}\right)^{-1}$;
        Sample $\boldsymbol{U}_{ij}$ uniformly from $\left[\Phi\left(\frac{z_l - \mu_{ij}}{\sigma_j}\right), \Phi\left(\frac{z_u - \mu_{ij}}{\sigma_j}\right)\right]$;
        Set $\boldsymbol{Z}^{(t)}_{ij} = \mu_{ij} + \sigma_j \times \Phi^{-1}(\boldsymbol{U}_{ij})$;
      end
    end
  end
  Draw $\boldsymbol{\Sigma}^{(t)} \sim$
   Inverse-Wishart $\left(n + \nu_0, \nu_0\boldsymbol{\Sigma}_0 + \left(\boldsymbol{Z}^{(t)} - \boldsymbol{X}\tilde{\boldsymbol{B}}^{(t-1)}\right)^T \left(\boldsymbol{Z}^{(t)} - \boldsymbol{X}\tilde{\boldsymbol{B}}^{(t-1)}\right)\right.$
   $\left. + \left(\tilde{\boldsymbol{B}}^{(t-1)} - \boldsymbol{B}_0\right)^T \boldsymbol{\Lambda}_0 \left(\tilde{\boldsymbol{B}}^{(t-1)} - \boldsymbol{B}_0\right)\right)$;
  Draw $\tilde{\boldsymbol{B}}^{(t)} \sim$
   Normal $\left((\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{\Lambda}_0)^{-1}(\boldsymbol{X}^T\boldsymbol{Z}^{(t)} + \boldsymbol{\Lambda}_0\boldsymbol{B}_0), \boldsymbol{\Sigma}^{(t)} \otimes (\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{\Lambda}_0)^{-1}\right)$;
  Set $\boldsymbol{C}^{(t)}_{ij} = \boldsymbol{\Sigma}^{(t)}_{ij} / \sqrt{\boldsymbol{\Sigma}^{(t)}_{ii}\boldsymbol{\Sigma}^{(t)}_{jj}}$;
  Set $\boldsymbol{B}^{(t)}_{ij} = \tilde{\boldsymbol{B}}^{(t)}_{ij} / \sqrt{\boldsymbol{\Sigma}^{(t)}_{jj}}$;
end

Algorithm 4: Gibbs sampling procedure for multivariate regression via the extended rank likelihood

Under this paradigm, values of $\boldsymbol{Y}_{ij}$ that are missing at random can be trivially

accommodated by simply sampling $\boldsymbol{Z}_{ij}^{(t)}$ from a unconstrained normal distribution at each iteration. A similar approach has been used for multiple imputation of mixed data (Wang et al., 2017), under which the empirical marginal distributions $\hat{F}_j$ are used to impute missing values of $\boldsymbol{Y}_{ij}$, given samples of $\boldsymbol{Z}_{ij}$.

For each model, we place a weakly informative prior on $\boldsymbol{\Sigma}$ concentrated near the identity, such that $\boldsymbol{\Sigma}_0 = \boldsymbol{I}$ and $\nu_0 = p + 2$. In addition, we place a form of Zellner's g-prior on $\tilde{\boldsymbol{B}}$ (Zellner, 1986), such that parameter estimation is invariant to changes in the scale of the regressors (Hoff, 2009). Specifically, letting $g = n$, we have that $\boldsymbol{B}_0 = \boldsymbol{0}$ and $\boldsymbol{\Lambda}_0 = (\boldsymbol{X}^T\boldsymbol{X})/n$.

C.3 Estimated Task Score Percentiles by Group

Table C.1: Estimated percentile of task performance based on the posterior predictive means obtained via the two-way interaction model. In context, we expect a middle school male baseball player to score better than 7.8% of all athletes at Eye-Hand Coordination. By contrast, we expect a professional female soccer player to score better than 73.9% of all athletes at the same task.

Level	Sport_Type	Gender	VC	CS	NFQ	TC	PS	EHC	RXN
Middle School	Interceptive	Male	0.505	0.539	0.283	0.462	0.287	0.078	0.313
High School	Interceptive	Male	0.507	0.504	0.515	0.456	0.538	0.394	0.425
College	Interceptive	Male	0.598	0.477	0.633	0.487	0.510	0.589	0.555
Pro	Interceptive	Male	0.580	0.566	0.582	0.484	0.495	0.650	0.620
Middle School	Strategic	Male	0.348	0.406	0.248	0.476	0.371	0.089	0.326
High School	Strategic	Male	0.445	0.454	0.446	0.485	0.523	0.397	0.400
College	Strategic	Male	0.481	0.485	0.566	0.532	0.520	0.544	0.545
Pro	Strategic	Male	0.501	0.534	0.522	0.487	0.514	0.648	0.496
Middle School	Interceptive	Female	0.467	0.526	0.272	0.569	0.313	0.196	0.465
High School	Interceptive	Female	0.392	0.513	0.439	0.550	0.458	0.420	0.469
College	Interceptive	Female	0.547	0.508	0.511	0.574	0.535	0.597	0.536
Pro	Interceptive	Female	0.574	0.537	0.601	0.593	0.508	0.729	0.816
Middle School	Strategic	Female	0.397	0.360	0.261	0.480	0.472	0.227	0.401
High School	Strategic	Female	0.420	0.429	0.401	0.476	0.517	0.437	0.366
College	Strategic	Female	0.519	0.481	0.472	0.516	0.618	0.567	0.447
Pro	Strategic	Female	0.584	0.470	0.572	0.494	0.600	0.739	0.650

C.4 Conditional Dependence Structure of Task Scores

Based on the Gibbs sampling algorithm detailed previously, we obtain correlated samples of $\boldsymbol{C}$ from its posterior distribution. From this, it is possible to examine the posterior distribution of $\boldsymbol{C}_{[j,-j]}\boldsymbol{C}^{-1}_{[-j,-j]}$ to obtain estimates of the conditional dependencies between the underlying processes that give rise to the score on task j and the scores on the other tasks. Based upon the distribution of these quantities, an undirected graph can be constructed to describe the conditional dependence structure of the task scores (Figure C.2). For example, given Visual Clarity and Near-Far Quickness, there is insufficient evidence to conclude that Target Capture is monotonically associated with performance on any of the other tasks, since the VC and NFQ nodes block all paths from TC to the other nodes (d-separation). By contrast, even given the other task scores, Visual Clarity and Contrast Sensitivity have a strongly positive monotonic association.

C.5 League Equivalence Models

To do estimate α_j and γ_j, we first scrape publicly available minor and major league play-by-play data between 2012 and 2013 from MLB.com using the pitchRx package in the R language (Sievert, 2015). We collate the statistics of all players (1695 batters, 1053 pitchers) who played in multiple leagues during that period. By examining the difference in player performance between the leagues, we quantify the degree of difficulty of each league via a Bayesian model. For example, if the Major Leagues is more difficult than AAA, we should expect a player who plays in both leagues to register a lower on-base percentage in the Major Leagues than in AAA.

For each of the five game statistic variables, we estimate the corresponding model detailed in Section 4.3, using the data of all players who played in multiple leagues between 2012 and 2013. Since we do not have sensorimotor measurements for these

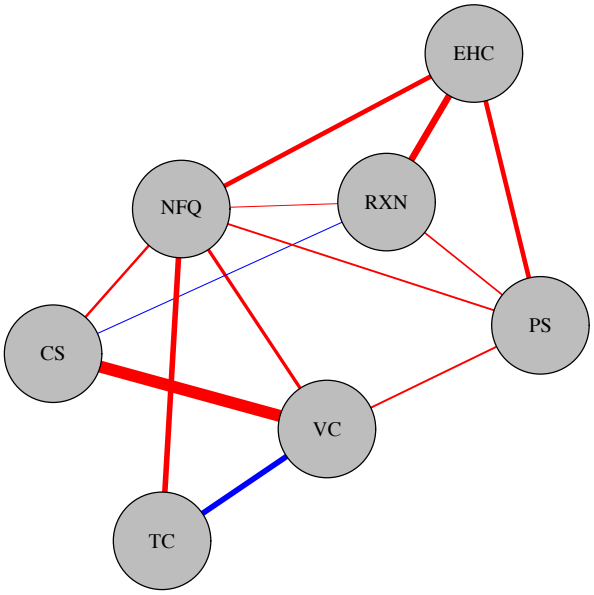


FIGURE C.2: Conditional dependence graph describing patterns of conditional associations among the tasks. An edge is present if the 95% credible interval for the associated regression coefficient does not contain zero. Positive relationships are denoted by red edges, while negative relationships are denoted by blue edges. The relative width of an edge represents the relative strength of the corresponding conditional relationship.

players, we instead place a standard normal prior on each A_i. Because we assume that increased ability corresponds to improved performance for each game statistic, we impose a half-Cauchy prior on γ_j, as well as conjugate Normal/Gamma priors for α_j and τ. For the initial on-base percentage model, we use the following specification and priors:

$$\begin{aligned} \text{OB}_{ij} &\sim \text{Binomial}(N_{ij}, p_{ij}). \\ \text{logit}(p_{ij}) &\sim N(\alpha_j + \gamma_j A_i, \tau^{-1}). \\ A_i &\sim N(0, 1). \end{aligned} \tag{C.3}$$

$$
\begin{aligned}
\alpha_j &\sim N(0, \tau_\alpha^{-1}). \\
\tau_\alpha &\sim TN(0, 1, 0, \infty). \\
\gamma_j &\sim TN(0, , \tau_\gamma^{-1}, 0, \infty). \\
\tau_\gamma &\sim TN(0, 1, 0, \infty). \\
1/\sigma^2 &\sim G(0.5, 0.5). \\
\tau &\sim G(0.5, 0.5),
\end{aligned}
\tag{C.4}
$$

where $TN(\mu, \sigma^2, a, b)$ is the normal distribution truncated to a lower bound a and upper bound b. The initial models for BB% and K% have similar specifications, with the exception that γ_j in the K% model is bounded above by zero, since players with higher ability strike out less frequently.

For the slugging percentage equivalence model, we use the following specification and priors:

$$
\begin{aligned}
\text{SLG}_{ij} &\sim N(\mu_{ij}, \sigma^2/N_{ij}). \\
\mu_{ij} &\sim N(\alpha_j + \gamma_j A_i, \tau^{-1}). \\
A_i &\sim N(0, 1). \\
\alpha_j &\sim N(0, \tau_\alpha^{-1}). \\
\tau_\alpha &\sim TN(0, 1, 0, \infty). \\
\gamma_j &\sim TN(0, , \tau_\gamma^{-1}, 0, \infty). \\
\tau_\gamma &\sim TN(0, 1, 0, \infty). \\
1/\sigma^2 &\sim G(0.5, 0.5). \\
\tau &\sim G(0.5, 0.5),
\end{aligned}
\tag{C.5, C.6}
$$

The specification for the FIP equivalence model is the same as that for the SLG equivalence model, with the exception that γ_j in the FIP model is bounded above by zero, since players with higher ability record lower values for FIP. For each equivalence model, we draw 10,000 samples of the parameters from the joint posterior

distribution via Gibbs sampling after an initial burn-in of 500 iterations. We compute the posterior means and variances for all parameters, including α_j and γ_j. The posterior means and variances of these quantities are used to construct the priors used in the visual-motor models, as described in C.6.

C.6 Prior Specifications for On-Field Performance Models

As mentioned in Sections 4.3 and C.5, for each game statistic, we fit a league equivalence model to obtain concentrated priors for α_j, γ_j, τ, and σ^2. We use the posterior means and variances from the equivalence models to form prior distributions in the corresponding on-field performance models. For each equivalence model, let $\hat{\tau}$ and τ_{SD} be the posterior mean and standard deviation of τ, respectively. We use analogous notation for the posterior means and standard deviations for the other parameters. We emphasize that each outcome variable has its own set of parameters, although we use a common notation for convenience. The final on-base percentage model used to produce the results in 4.11, including all prior distributions, is as follows:

$$\begin{aligned} \text{OB}_{ij} &\sim \text{Binomial}(N_{ij}, p_{ij}). \\ \text{logit}(p_{ij}) &\sim N(\alpha_j + \gamma_j A_i, \tau^{-1}). \\ A_i &= \boldsymbol{X}\boldsymbol{\beta}. \end{aligned} \tag{C.7}$$

$$\begin{aligned} \alpha_j &\sim N(\hat{\alpha}_j, \alpha_{j,SD}). \\ \gamma_j &\sim TN(\hat{\gamma}_j, \gamma_{j,SD}, 0, \infty). \\ \tau &\sim TN(\hat{\tau}, \tau_{SD}, 0, \infty). \\ \boldsymbol{\beta} &\sim N(\mathbf{0}, 10^2\boldsymbol{I}). \\ X_{i,miss} &\sim N(0, 1). \end{aligned} \tag{C.8}$$

The specifications for the BB% and K% models are identical, with the exception that γ_j in the K% model is bounded above by zero, since players with higher ability

strike out less frequently. The final slugging percentage model used to produce the results displayed in Table C.2, including all prior distributions, is as follows.

$$\begin{aligned} \text{SLG}_{ij} &\sim N(\mu_{ij}, \sigma^2/N_{ij}). \\ \mu_{ij} &\sim N(\alpha_j + \gamma_j A_i, \tau^{-1}). \\ A_i &= \boldsymbol{X}\boldsymbol{\beta}. \end{aligned} \tag{C.9}$$

$$\begin{aligned} \alpha_j &\sim N(\hat{\alpha}_j, \alpha_{j,SD}). \\ \gamma_j &\sim TN(\hat{\gamma}_j, \gamma_{j,SD}, 0, \infty). \\ \tau &\sim TN(\hat{\tau}, \tau_{SD}, 0, \infty). \\ \sigma &\sim TN(\hat{\sigma}, \sigma_{SD}, 0, \infty). \\ \boldsymbol{\beta} &\sim N(\mathbf{0}, 10^2\boldsymbol{I}). \\ X_{i,miss} &\sim N(0, 1). \end{aligned} \tag{C.10}$$

The specification for the FIP model is the same as that for the SLG model, with the exception that γ_j in the FIP model is bounded above by zero, since pitchers with higher ability record lower values for FIP.

C.7 Posterior Distribution for SLG and FIP Models

Table C.2: Mean coefficients, standard deviations, and 95% credible intervals for each model variable are shown for (A) Slugging Percentage (SLG) and (B) Fielder Independent Pitching (FIP). Values for which the 95% credible interval excludes zero are bolded.

	(A) SLG				(B) FIP			
	Mean	SD	2.5%	97.5%	Mean	SD	2.5%	97.5%
Visual Clarity	-0.01	0.15	-0.30	0.28	-0.24	0.25	-0.71	0.26
Contrast Sensitivity	-0.14	0.14	-0.41	0.15	0.07	0.24	-0.40	0.54
Depth Perception	0.18	0.14	-0.08	0.45	**-0.68**	**0.22**	**-1.12**	**-0.26**
Near-Far Quickness	0.00	0.14	-0.26	0.27	0.25	0.23	-0.20	0.70
Target Capture	0.15	0.14	-0.12	0.43	-0.11	0.21	-0.52	0.29
Perception Span	**0.29**	**0.14**	**0.01**	**0.58**	0.21	0.23	-0.25	0.66
Eye-Hand Coordination	0.05	0.15	-0.25	0.34	-0.18	0.20	-0.58	0.20
Reaction Time	0.09	0.15	-0.21	0.40	-0.07	0.20	-0.47	0.34
Age	**0.44**	**0.15**	**0.15**	**0.72**	**0.55**	**0.21**	**0.15**	**0.96**
Infield	**-0.60**	**0.28**	**-1.13**	**-0.05**				
Catcher	**-1.40**	**0.41**	**-2.21**	**-0.60**				
Intercept	0.07	0.20	-0.33	0.45	-0.32	0.21	-0.74	0.07

Bibliography

Abernethy, B., Burgess-Limerick, R., and Parks, S. (1994a), "Contrasting Approaches to the Study of Motor Expertise," *Quest*, 46, 186–198.

Abernethy, B., Neal, R., and Koning, P. (1994b), "Visual-Perceptual and cognitive differences between expert, intermediate and novice snooker players," *Applied Cognitive Psychology*, 8, 185 – 211.

Akande, O., Barrientos, A., and Reiter, J. P. (2018), "Simultaneous Edit and Imputation For Household Data with Structural Zeros," *Journal of Survey Statistics and Methodology*.

Andridge, R. and Little, R. J. A. (2010), "A Review of Hot Deck Imputation for Survey Non-response." *International statistical review*, 78, 40–64.

Appelbaum, L. G. and Erickson, G. B. (2016), "Sports vision training: a review of the state-of-the-art in digital training techniques," *International Review of Sport and Exercise Psychology*, pp. 1–30.

Appelbaum, L. G., Lu, Y., Khanna, R., and Detwiler, K. (2016), "The Effects of Sports Vision Training on Sensorimotor Abilities in Collegiate Softball Athletes," *Athletic Training and Sports Health Care*.

Arnold, B. C., Castillo, E., and Sarabia, J. M. (2001), "Conditionally Specified Distributions: An Introduction," *Statistical Science*, 16, 249–265.

Bahill, A. T. and LaRitz, T. (1984), "Why can't batters keep their eyes on the ball?" *American Scientist*, 72, 249–253.

Banerjee, S., Carlin, B., and Gelfand, A. (2014), *Hierarchical Modeling and Analysis for Spatial Data*, CRC Press.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015), "Fitting Linear Mixed-Effects Models Using lme4," *Journal of Statistical Software*, 67, 1–48.

Box, G. E. P. and Draper, N. R. (1986), *Empirical Model Building and Response Surfaces*, John Wiley & Sons, Inc., New York, NY, USA.

Brewer, M. J. and Nolan, A. J. (2007), "Variable smoothing in Bayesian intrinsic autoregressions," *Environmetrics*, 18, 841–857.

Burris, K. and Hoff, P. (2019a), "Bayesian Hot Deck Imputation for Multivariate Numeric Data," in preparation.

Burris, K. and Hoff, P. (2019b), "Exact adaptive confidence intervals for small areas," *Journal of Survey Statistics and Methodology*.

Burris, K., Vittetoe, K., Ramger, B., Suresh, S., Tokdar, S. T., Reiter, J. P., and Appelbaum, L. G. (2018), "Sensorimotor abilities predict on-field performance in professional baseball," *Scientific Reports*, 8, 116.

Burris, K., Liu, S., and Appelbaum, L. (2019), "Visual-motor expertise in athletes," *Journal of Sports Sciences*, under revision.

Casanova, F., Oliveira, J., Williams, M., and Garganta, J. (2009), "Expertise and perceptual-cognitive performance in soccer: a review," *Rev. Port. Clen. Desp*, 9.

Chib, S. and Greenberg, E. (1998), "Analysis of Multivariate Probit Models," *Biometrika*, 85, 347–361.

Classe, J., Semes, L., M Daum, K., Nowakowski, R., J Alexander, L., Wisniewski, J., A Beisel, J., Mann, K., Rutstein, R., Smith, M., and Bartolucci, A. (1997), "Association between visual reaction time and batting, fielding, and earned run averages among players of the Southern Baseball League," *Journal of the American Optometric Association*, 68, 43–9.

Cochran, W. G. (1977), *Sampling Techniques, 3rd Edition.*, John Wiley and Sons, Inc.

Davids, K. (2002), *Interceptive Actions in Sport: Information and Movement*, Routledge.

Davies, M. A. and Basco, D. (2010), "The many flavors of DIPS: History and Overview," *Baseball Research Journal*, 39, 41–50.

De Waal, T. (2017), "Imputation Methods Satisfying Constraints," .

DHHS (2015), "Secretarys Advisory Committee on Human Research Protections. Attachment A: human subjects research implications of big data studies," .

Draper, L. R. and Winkler, W. (1997), "Balancing and Ratio Editing with the New SPEER System," .

Eccles, D. W., Walsh, S. E., and Ingledew, D. K. (2006), "Visual attention in orienteers at different levels of experience," *J Sports Sci*, 24, 77–87.

Efron, B. and Morris, C. (1977), "Stein's Paradox in Statistics," *Scientific American - SCI AMER*, 236, 119–127.

Elmurr, P. (2011), "The Relationship of Vision and Skilled Movement: A General Review Using Cricket Batting," *Eye & Contact Lens*, 37, 164–166.

Erickson, G., Citek, K., Cove, M., Wilczek, J., Linster, C., Bjarnason, B., and Langemo, N. (2011), "Reliability of a computer-based system for measuring visual performance skills," *Optometry*, 82, 528–542.

Fay, R. and Herriot, R. (1979), "Estimates of income for small places: An application of James-Stein procedures to census data," *J. Amer. Statist. Assoc.*, 74, 269–277.

Gathercole, S. E., Pickering, S. J., Ambridge, B., and Wearing, H. (2004), "The structure of working memory from 4 to 15 years of age," *Developmental Psychology*, p. 190.

Gelman, A. and Hill, J. (2006), *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Analytical Methods for Social Research, Cambridge University Press.

Gelman, A., Hwang, J., and Vehtari, A. (2013), "Understanding predictive information criteria for Bayesian models," *Statistics and Computing*, 24.

Gelman, A., Carlin, J. B., Stern, H. S., Rubin, D., and Dunson, D. (2014), *Bayesian Data Analysis*, CRC Press.

Geweke, J. (1991), "Efficient Simulation from the Multivariate Normal and Student-t Distributions Subject to Linear Constraints and the Evaluation of Constraint Probabilities," .

Ghosh, M., Natarajan, K., Walter, L., and Kim, D. (1999), "Hierarchical Bayes GLMs for the analysis of spatial data: An application to disease mapping," *Journal of Statistical Planning and Inference*, 75, 305–318.

Gregory, R. L. (1997), *Eye and brain: Tthe psychology of seeing*, Princeton University Press.

Helsen, W. and Starkes, J. (1999), "A Multidimensional Approach to Skilled Perception and Performance in Sport," *Applied Cognitive Psychology*, 13, 1 – 27.

Hitzeman, S. A. and Beckerman, S. A. (1993), "What the literature says about sports vision," *Optom Clin*, 3, 145–169.

Hoff, P. (2018), *sbgcop: Semiparametric Bayesian Gaussian Copula Estimation and Imputation*, R package version 0.980.

Hoff, P. D. (2007), "Extending the rank likelihood for semiparametric copula estimation," *Ann. Appl. Stat.*, 1, 265–283.

Hoff, P. D. (2009), *A First Course in Bayesian Statistical Methods*, Springer Publishing Company, Incorporated, 1st edn.

Hoffman, L., Polan, G., and Powell, J. (1984), "The relationship of contrast sensitivity functions to sports vision," *Journal of the American Optometric Association*, 55, 747–52.

Huttermann, S., Memmert, D., and Simons, D. J. (2014), "The size and shape of the attentional "spotlight" varies with differences in sports expertise," *J Exp Psychol Appl*, 20, 147–157.

Ishwaran, H. and James, L. F. (2001), "Gibbs Sampling Methods for Stick-Breaking Priors," *Journal of the American Statistical Association*, 96, 161–173.

Kato, T. and Fukuda, T. (2002), "Visual search strategies of baseball batters: Eye movements during the preparatory phase of batting," *Perceptual and motor skills*, 94, 380–6.

Kim, H. J., Reiter, J. P., Wang, Q., Cox, L. H., and Karr, A. F. (2014), "Multiple Imputation of Missing or Faulty Values Under Linear Constraints," *Journal of Business & Economic Statistics*, 32, 375–386.

Kim, H. J., Cox, L. H., Karr, A. F., Reiter, J. P., and Wang, Q. (2015), "Simultaneous Edit-Imputation for Continuous Microdata," *Journal of the American Statistical Association*, 110, 987–999.

Klemish, D., Ramger, B., Vittetoe, K., Reiter, J., Tokdar, S., and Appelbaum, L. (2017), "Visual Abilities Distinguish Pitchers from Hitters in Professional Baseball," *Journal of Sports Sciences*.

Kottas, A., Muller, P., and Quintana, F. (2005), "Nonparametric Bayesian Modeling for Multivariate Ordinal Data," *Journal of Computational and Graphical Statistics*, 14, 610–625.

Kovar, J. and Whitridge, P. (1990), "Generalized Edit and Imputation System; Overview and Applications," *Revista Brasileira de Estadistica*, 51, 85–100.

Krasich, K., Ramger, B., Holton, L., Wang, L., Mitroff, S. R., and Appelbaum, L. G. (2016), "Sensorimotor learning in a computerized athletic training battery," *Journal of Motor Behavior*, 48, 401–412.

Laby, D. M., Rosenbaum, A. L., Kirschen, D. G., Davidson, J. L., Rosenbaum, L. J., Strasser, C., and Mellman, M. F. (1996), "The visual function of professional baseball players," *Am J Opthalmol*, 122, 476–485.

Lebeau, J.-C., Liu, S., Sáenz-Moncaleano, C., Sanduvete-Chaves, S., Moscoso, S., Jane Becker, B., and Tenenbaum, G. (2016), "Quiet Eye and Performance in Sport: A Meta-Analysis," *Journal of Sport & Exercise Psychology*, 38, 441–457.

Little, R. J. A. (1988), "Missing-Data Adjustments in Large Surveys," *Journal of Business & Economic Statistics*, 6, 287–296.

Little, R. J. A. and Rubin, D. B. (1986), *Statistical Analysis with Missing Data*, John Wiley & Sons, Inc., New York, NY, USA.

Liu, J. S. and Wu, Y. N. (1999), "Parameter Expansion for Data Augmentation," *Journal of the American Statistical Association*, 94, 1264–1274.

Liu, S., Eklund, R., and Tenenbaum, G. (2015), "Time pressure and attention allocation effect on upper limb motion steadiness," *Journal of motor behavior*, 47, 271–281.

Mann, D. T., Williams, A. M., Ward, P., and Janelle, C. M. (2007), "Perceptual-cognitive expertise in sport: a meta-analysis," *Journal of sport exercise & psychology*, 29, 457–478.

Manrique-Vallier, D. and Reiter, J. P. (2017), "Bayesian Simultaneous Edit and Imputation for Multivariate Categorical Data," *Journal of the American Statistical Association*, 112, 1708–1719.

Maples, J. J. (2017), "Improving small area estimates of disability: combining the American Community Survey with the Survey of Income and Program Participation," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180, 1211–1227.

Mohammed, S. and Dumville, B. C. (2001), "Team mental models in a team knowledge framework: expanding theory and measurement across disciplinary boundaries," *Journal of Organizational Behavior*, 22, 89–106.

Molia, L. M., Rubin, S. E., and Kohn, N. (1998), "Assessment of stereopsis in college baseball pitchers and batters," *Journal of AAPOS : the official publication of the American Association for Pediatric Ophthalmology and Strabismus / American Association for Pediatric Ophthalmology and Strabismus*, 2, 86–90.

Molina, I. and Marhuenda, Y. (2015), "sae: An R Package for Small Area Estimation," *The R Journal*, 7, 81–98.

Morris, T. P., White, I. R., and Royston, P. J. (2014), "Tuning multiple imputation by predictive mean matching and local residual draws," in *BMC medical research methodology*.

Murray, J. S. (2018), "Multiple Imputation: A Review of Practical and Theoretical Findings," .

Murray, J. S. and Reiter, J. P. (2016), "Multiple Imputation of Missing Categorical and Continuous Values via Bayesian Mixture Models With Local Dependence," *Journal of the American Statistical Association*, 111, 1466–1479.

Murray, J. S., Dunson, D. B., Carin, L., and Lucas, J. E. (2013), "Bayesian Gaussian Copula Factor Models for Mixed Data," *Journal of the American Statistical Association*, 108, 656–665.

Nicholson, K. G. and Kimura, D. (1996), "Sex differences in speech and manual skill," *Perceptual and motor skills*, 82, 3–13.

Pannekoek, J. and Zhang, L.-C. (2015), "Optimal adjustments for inconsistency in imputed data," *Survey methodology*, 41, 127–144.

Pfeffermann, D. (2013), "New important developments in small area estimation," *Statistical Science*, 28, 40–68.

Pitt, M., Chan, D., and Kohn, R. (2006), "Efficient Bayesian Inference for Gaussian Copula Regression Models," *Biometrika*, 93, 537–554.

Plummer, M. (2003), "JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling," .

Pratesi, M. and Salvati, N. (2008), "Small area estimation: the EBLUP estimator based on spatially correlated random area effects," *Statistical Methods and Applications*, 17, 113–141.

Pratt, J. W. (1963), "Shorter Confidence Intervals for the Mean of a Normal Distribution with Known Variance," *Ann. Math. Statist.*, 34, 574–586.

Price, P., Nero, A., and Gelman, A. (1996), "Bayesian prediction of mean indoor radon concentrations for Minnesota counties," *Health Physics*, 71, 922–936.

Raghunathan, T. E., Lepkowski, J., H. Van Hoewyk, J., and W. Solenberger, P. (2000), "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models," *Survey Methodology*, 27.

Rao, J. N. K. and Molina, I. (2015), *Small Area Estimation*, John Wiley and Sons, Inc.

Rao, V., Lin, L., and Dunson, D. B. (2016), "Data augmentation for models based on rejection sampling," *Biometrika*, 103, 319–335.

Reichow, A. W., E Garchow, K., and Y Baird, R. (2011), "Do Scores on a Tachistoscope Test Correlate With Baseball Batting Averages?" *Eye & contact lens*, 37.

Reiter, J. P. and Raghunathan, T. E. (2007), "The Multiple Adaptations of Multiple Imputation," *Journal of the American Statistical Association*, 102, 1462–1471.

Robert, C. P. and Casella, G. (2005), *Monte Carlo Statistical Methods (Springer Texts in Statistics)*, Springer-Verlag, Berlin, Heidelberg.

Rossi, P. E., Allenby, G., and McCullouch, R. (2005), *Bayesian statistics and marketing*, Wiley.

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*, Wiley.

Rubin, D. B. (2003), "Discussion on Multiple Imputation," *International Statistical Review*, 71, 619–625.

Schalen, L. (1980), "Quantification of tracking eye movements in normal subjects," *Acta Otolaryngol*, 90, 404–413.

Sethuraman, J. (1994), "A Constructive Definition of the Dirichlet Prior," *Statistica Sinica*, 4, 639–650.

Sievert, C. (2015), "pitchRx: Tools for Harnessing MLBAM Gameday Data and Visualizing pitchfx," R package version 1.8.2.

Singer, R. N. (2000), "Performance and human factors: considerations about cognition and attention for self-paced and externally-paced events," *Ergonomics*, 43, 1661–1680.

Singh, B., Shukla, G., and Kundu, D. (2005), "Spatio-temporal models in small-area estimation," *Survey Methodology*, 31, 183–195.

Smith, M. S. and Khaled, M. A. (2012), "Estimation of Copula Models With Discrete Margins via Bayesian Data Augmentation," *Journal of the American Statistical Association*, 107, 290–303.

Solomon, H., Zinn, W., and Vacroux, A. (1988), "Dynamic stereoacuity: a test for hitting a baseball?" *Journal of the American Optometric Association*, 59, 522–6.

Spaniol, F., Cruz, J., Alves, M., Cochran, S., Hicks, B., and Lawson, B. (2015), "The relationship between convergence, divergence, recognition, and tracking skills and batting performance of professional baseball players," .

Starkes, J. L. and Ericsson, K. A. (2003), *Expert performance in sports: Advances in research on sport expertise*, Human Kinetics.

Szymanski, D., Light, T., Voss, Z., and Greenwood, M. (2015), "Relationships between vision performance scores and offensive statistics of college baseball players," Paper presented at the American College of Sports Medicine, Southeast Regional Chapter meeting, Jacksonville, Florida.

Tempelman, C. (2007), "Imputation of restricted data: applications to business surveys," Ph.D. thesis, University of Groningen.

Uchida, Y., Kudoh, D., Higuchi, T., Honda, M., and Kanosue, K. (2012), "Dynamic Visual Acuity in Baseball Players Is Due to Superior Tracking Abilities," *Medicine and science in sports and exercise*, 45.

van Buuren, S. (2018), *Flexible imputation of missing data*, CRC Press.

van Buuren, S. and Groothuis-Oudshoorn, C. (2011), "MICE: Multivariate Imputation by Chained Equations in R," *Journal of Statistical Software*, 45.

Vehtari, A., Gelman, A., and Gabry, J. (2016), "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC," *Statistics and Computing*, 27.

Voss, M., Kramer, A., Basak, C., Parkash, R., and Roberts, B. (2010), "Are Expert Athletes Expert in the Cognitive Laboratory? A meta-analytic Review of Cognition and Sport Expertise," *Applied Cognitive Psychology*, 24, 812–826.

Wall, M. M. (2004), "A close look at the spatial structure implied by the CAR and SAR models," *Journal of Statistical Planning and Inference*, 121, 311 – 324.

Wang, J., Loong, B., H. Westveld, A., and Welsh, A. (2017), "A Copula-based Imputation Model for Missing Data of Mixed Type in Multilevel Data Sets," .

Wang, L., Krasich, K., Bel-Bahar, T., Hughes, L., Mitroff, S. R., and Appelbaum, L. G. (2015), "Mapping the structure of perceptual and visual-motor abilities in healthy young adults," *Acta Psychol (Amst)*.

Wang, W. C., DeLang, M. D., Vittetoe, K., Ramger, B., and Appelbaum, L. (2018), "Laterality Preferences in Athletes: Insights from a Database of 1770 Male Athletes," *American Journal of Sports Science*, 6, 20–25.

Ward, P. and Williams, A. (2003), "Perceptual and Cognitive Skill Development in Soccer: The Multidimensional Nature of Expert Performance," *Journal of Sport and Exercise Psychology*, 25, 93–111.

Wickham, H. (2016), *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag New York.

Williams, A. M. and Ericsson, K. A. (2005), "Perceptual-cognitive expertise in sport: some considerations when applying the expert performance approach," *Hum Mov Sci*, 24, 283–307.

Williams, A. M. and Ford, P. R. (2008), "Expertise and expert performance in sport," *International Review of Sport and Exercise Psychology*, 1, 4–18.

Williams, A. M., Ford, P. R., Eccles, D. W., and Ward, P. (2011), "Perceptual-cognitive expertise in sport and its acquisition: Implications for applied cognitive psychology," *Applied Cognitive Psychology*, 25, 432–442.

Yarrow, K., Brown, P., and Krakauer, J. (2009), "Inside the brain of an elite athlete: the neural processes that support high achievement in sports," *Nature Reviews Neuroscience*, 10, 585–596.

You, Y. and Chapman, B. (2006), "Small area estimation using area level models and estimated sampling variances," *Survey Methodology*, 32, 97–103.

Yu, C. and Hoff, P. D. (2018), "Adaptive multigroup confidence intervals with constant coverage," *Biometrika*, 105, 319–335.

Zellner, A. (1986), "On assessing prior distributions and Bayesian regression analysis with g-prior distributions," *Bayesian inference and decision techniques: essays in honor of Bruno de Finetti*, pp. "233–243.

Biography

Kyle Christian Burris graduated from Trinity Academy in Wichita, Kansas in May of 2012, enrolling at Wheaton College (IL) the ensuing fall. He graduated from Wheaton College in May of 2015, completing a Bachelor of Arts in Economics and a Bachelor of Science in Mathematics.

Upon graduating from Wheaton College, Kyle studied at Duke University in pursuit of a PhD in statistics. During this time, he married his college sweetheart Rachel Marie Sloan. Under the direction of Peter Hoff, Kyle worked to develop statistical methodology for small-area confidence interval construction and multiple imputation for mixed data. Kyle also developed an interest in the application of statistics to baseball, working with Greg Appelbaum in the Department of Psychiatry and Behavioral Sciences to study the relationship between visual-motor expertise and athletic performance. He presented his sports-related work at the 2017 New England Symposium for Sports in Statistics (NESSIS) annd 2018 MIT Sloan Sports Analytics conference. In the summer of 2019, Kyle began working as a quantitative analyst for the Cleveland Indians baseball club. He is expected to complete his PhD in Statistical Science in September 2019.