

Methodological Advances for Multi-group Data

by

Elizabeth Bersson

Department of Statistical Science
Duke University

Date: March 5, 2024
Approved:

Peter D. Hoff, Supervisor

Amy H. Herring

Surya T. Tokdar

Elizabeth L. Turner

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
2024

ABSTRACT

Methodological Advances for Multi-group Data

by

Elizabeth Bersson

Department of Statistical Science
Duke University

Date: March 5, 2024
Approved:

Peter D. Hoff, Supervisor

Amy H. Herring

Surya T. Tokdar

Elizabeth L. Turner

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
2024

Copyright © 2024 by Elizabeth Bersson
All rights reserved except the rights granted by the
Creative Commons Attribution-Noncommercial Licence

Abstract

This dissertation focuses on improving inference in analyses of multi-group data, that is, data obtained from non-overlapping subpopulations such as across counties in a state or for various socio-economic groups. Precise and accurate group-specific inference based on such data may be encumbered by small within-group sample sizes. In such cases, inference may be improved by making use of auxiliary information. In this work, we present two streams of methodological development aimed at improving group-specific inference for multi-group data that may feature small within-group sample sizes for some or all of the groups. First, we detail methodology that constructs frequentist-valid prediction regions based on indirect information. We show such prediction regions may feature improved precision over those constructed with standard approaches. To this end, we present methods that result in accurate and precise prediction regions for multi-group data based on a continuous response in Chapter 2 and a categorical response in Chapter 3. We develop straightforward computational algorithms to compute the regions and detail empirical Bayesian estimation procedures that allow for information to be shared across groups in the construction of the prediction regions. In Chapter 4, we present work that improves covariance estimation for structured multi-group data with shrinkage estimation that allows for robustness to structural assumptions. In particular, for multi-group matrix-variate data, we describe a hierarchical prior distribution that improves covariance estimate accuracy by flexibly allowing for shrinkage within

groups towards a Kronecker structure and across groups towards a pooled covariance estimate. We illustrate the utility of all methods presented with simulation studies and data applications.

Dedication

In memory of my grandmother, Dr. Louise Valine, or, “Honey”

Contents

Abstract	iv
List of Tables	x
List of Figures	xi
List of Abbreviations and Symbols	xiii
Acknowledgements	xv
1 Introduction	1
1.1 Prediction for multi-group data	1
1.2 Covariance estimation for multi-group data	3
2 Optimal Conformal Prediction for Small Areas	5
2.1 Introduction	5
2.2 Bayes Optimal Conformal Prediction	10
2.2.1 Review of Conformal Prediction	10
2.2.2 Conformal Prediction via a Normal Working Model	11
2.2.3 Prediction Region Properties and Computation	14
2.3 Numerical Comparisons	19
2.4 FAB Small Area Prediction	23
2.4.1 Information Sharing via a Working Model	23
2.4.2 FAB Conformal Parameter Estimation Procedure	24
2.5 Radon Data Example	26
2.6 Discussion	31

3	Prediction Sets for Species Abundance using Indirect Information	33
3.1	Introduction	33
3.2	Methodology	35
3.2.1	Background and Notation	35
3.2.2	Order-based prediction for a single area	36
3.2.3	Order-based prediction for multiple areas	39
3.2.4	Empirical Bayes estimation of indirect information	40
3.3	Simulation Study	41
3.4	Summarising eBird species abundance data	44
3.4.1	Order-based prediction in Robeson County	47
3.4.2	Inference among species with tied counts in Haywood County	49
3.5	Discussion	51
4	Covariance Estimation for Multi-group, Matrix-variate Data	53
4.1	Introduction	53
4.2	Methodology	57
4.2.1	Partial-pooling shrinkage for multi-group data	57
4.2.2	Kronecker shrinkage for matrix-variate data	58
4.2.3	Flexible shrinkage for multi-group matrix-variate data	60
4.2.4	Interpretation of Parameters	61
4.3	Parameter Estimation	63
4.3.1	Latent Variable Representation	63
4.3.2	Posterior Approximation	64
4.3.3	Hyperparameter specification	70
4.4	Simulation Studies	71
4.4.1	Analysis of computational expense	77
4.5	Examples	78
4.5.1	Classification of spoken-word audio data	79
4.5.2	Analysis of TESIE chemical exposures data	83
4.6	Discussion	87
5	Conclusion	89

A	Supplementary Material for Chapter 2	91
A.1	Derivation of Marginal Likelihood	91
A.2	Notation Simplification	92
A.3	Proofs	93
B	Supplementary Material for Chapter 3	104
B.1	Maximization of the marginal multinomial-Dirichlet likelihood	104
B.2	Proofs	105
C	Supplementary Material for Chapter 4	108
C.1	Metropolis-Hastings Algorithm for SWAG	108
	Bibliography	112
	Biography	118

List of Tables

3.1	Comparison of birds observed in Robeson County and neighboring counties.	47
3.2	Comparison of birds observed in Haywood County and neighboring counties.	49
4.1	Population covariance assumptions used in the simulation study. . . .	72
4.2	Oracle estimators of the population covariances used in the simulation study.	73
4.3	Results from a simulation study comparing accuracy of five procedures.	74
4.4	Average run time of the SWAG Metropolis-Hastings sampler.	77
4.5	Rates of correct classification on audio test dataset.	82

List of Figures

2.1	Frequentist coverage rate of classical pivot and Bayesian prediction intervals.	8
2.2	Schematic of the process to obtain the FAB conformal region.	16
2.3	Results of a numerical analysis comparing the FAB prediction interval to alternatives for a population with $n = 3$	20
2.4	Results of a numerical analysis comparing expected interval width of the FAB and DTA procedures.	22
2.5	Results of a numerical analysis comparing expected interval width of the FAB and EB procedures.	23
2.6	Results of applying the FAB procedure to the EPA radon data.	28
2.7	Results comparing the interval width of FAB and various prediction procedures to EPA radon data.	30
3.1	Results of a numerical analysis comparing expected set cardinality of the FAB and direct procedures.	42
3.2	Within-county log sample size of eBird data in North Carolina. Sample sizes range from 8 to approximately 50,000.	45
3.3	Results of applying two prediction set procedures to the eBird data in North Carolina.	45
3.4	Comparison of species inclusion in prediction sets constructed using two procedures in Robeson County.	47
4.1	Schematic of the SWAG hierarchical model.	62
4.2	Results from a simulation study comparing accuracy of five procedures for a regime with $J = 4, p = 12$	76

4.3	Confusion matrices resulting from classification from the various covariance estimates.	80
4.4	MCMC samples for 4 elements selected at random from the set of covariance matrices Σ	81
4.5	MCMC samples for four elements selected at random from the set of covariance matrices $\{\Sigma_{LHS}, \Sigma_{HS}, \Sigma_C\}$	84
4.6	Approximations to the posterior distributions of key parameters λ, ν, η , and ξ for the TESIE data example.	86
4.7	Covariance estimates from the SWAG parameter estimation for the TESIE data example.	87

List of Abbreviations and Symbols

Symbols

\mathbb{R}	The set of real numbers.
\mathbb{R}^n	The set of n -tuples of real numbers.
\mathcal{S}_p^+	The set of non-singular p -dimensional covariance matrices.
\otimes	Kronecker product.
z_q	The q th quantile of the standard normal distribution.
$N(\mu, \sigma^2)$	The normal distribution with mean μ and variance σ^2 .
$N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	The p -dimensional multivariate-normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.
$N_{p_1 \times p_2}(\boldsymbol{M}, \boldsymbol{\Sigma})$	The $(p_1 \times p_2)$ -dimensional matrix-normal distribution with mean \boldsymbol{M} and covariance matrix $\boldsymbol{\Sigma}$.
$W_p(\boldsymbol{\Sigma}, \nu)$	The Wishart distribution with parameter $\boldsymbol{\Sigma}$ and degrees of freedom ν .
$G(a, b)$	The gamma distribution with shape parameter a and rate parameter b .
χ_ν^2	The standard Chi-squared distribution with ν degrees of freedom.
$MN_K(\boldsymbol{\theta}, N)$	The multinomial distribution with N trials and success probabilities $\boldsymbol{\theta}$ that lie on the K -dimensional simplex.
$Dirichlet_K(\boldsymbol{\gamma})$	The Dirichlet distribution of order K with parameter $\boldsymbol{\gamma}$.
$\Gamma(a)$	The gamma function evaluated at a .
$\mathbb{1}(\cdot)$	An indicator function equal to 1 if the argument in the parenthesis is true and 0 if the argument is false.

- $\mathbf{1}_n$ A n -dimensional vector consisting of all entries equal to 1.
- \mathbf{I}_n A $(n \times n)$ -dimensional matrix with entries along the diagonal equal to 1 and all else equal to 0.

Abbreviations

CPU	Central Processing Unit
DTA	Distance To Average
EB	Empirical Bayes
EPA	Environmental Protection Agency
FAB	Frequentist And Bayes
GB	Giga-byte
MCMC	Markov Chain Monte Carlo
MH	Metropolis-Hastings
MLE	Maximum Likelihood Estimate
NC	North Carolina
QDA	Quadratic Discriminant Analysis
RAM	Random Access Memory
SVOC	Semi-Volatile Organic Compound
SWAG	Shrinkage Within and Across Groups
TESIE	Toddlers Exposure to SVOCs in Indoor Environments

Acknowledgements

First and foremost, I'd like to thank my advisor, Dr. Peter Hoff, for his guidance, support, and time over these past few years. He taught me, among other things, the importance of intentionality and precision in research, and therefore life, for which I am grateful.

Thank you also to my committee members: Amy Herring, Surya Tokdar, and Elizabeth Turner. Their feedback throughout the various stages of the development of this research has been invaluable. Thanks to Kate Hoffman and Heather Stapleton for providing the TESIE data in Chapter 4.

Thanks too to Merlise Clyde, Amy Herring (x2!), Alexander Volfovsky, David Dunson, Mike West, Yue Jiang, and Jason Xu for distinctive mentorship, and special thanks to Joan Combs Durso for guiding me, often by example, to be a better mentor and teacher myself. Thank you also to Lori Rauch for constant support.

Lastly, thank you to all of the graduate students and postdoctoral fellows I have had the pleasure of interacting with during my time at Duke. The thoughtful discussions concerning life and science have been a joy. I am particularly appreciative of my office mates and dear friends, Joseph Lawson, Michael Christensen, and David Buch.

More generally, I am indebted to the entire Duke Statistical Science department for curating such a wonderful environment. I can not conceive of many more enriching ways to spend five years of one's life, and I am grateful to have had such a privilege.

Introduction

Multi-group data consisting of responses obtained from non-overlapping subpopulations or groups, for example, counties in a state or labeled observations, are increasingly common in many fields. As such, precise and accurate group-specific inference for such data is an important goal. This task is often encumbered by small within-group sample sizes for some or all of the groups. To this end, in this dissertation, we develop methodologies that improve group-specific inference in analyses of multi-group data.

1.1 Prediction for multi-group data

Existing methods for constructing prediction regions or sets for multi-group data involve a trade-off between guaranteeing nominal group-specific frequentist coverage rates and improving precision via the incorporation of indirect information. Chapters 2 and 3 detail methodologies aimed at mitigating this trade-off through making use of indirect information to improve precision of a prediction region while maintaining frequentist coverage guarantees of each region. Specifically, the methodologies developed in these chapters each utilize a conformal prediction framework to construct

small prediction regions that are constructed using indirect or prior information and maintain finite sample non-parametric guarantees of frequentist coverage. An R package to implement the prediction methods developed in these two chapters is provided at <https://github.com/betsybersson/fabPrediction> and includes a vignette illustrating usage.

Chapter 2 details work from the article Bersson and Hoff (2022). In this work, we develop a method for obtaining an optimally narrow prediction region for a numeric response and detail a straightforward algorithm to efficiently compute the prediction region. The resulting prediction region is proven to be an interval that contains a standard Bayesian point prediction estimator. This implies a coherent method for classically Bayesian point prediction with narrow uncertainty quantification that maintains frequentist notions of error. We extend this approach to a multi-group or small area regime without altering the frequentist coverage guarantee by incorporating prior information estimated from external populations in an empirical Bayesian manner. This method is illustrated through simulation studies and an application to radon survey data collected by the EPA.

In Chapter 3, based on the work in Bersson and Hoff (2023b), we detail methodology used to construct prediction sets for categorical data. This work is motivated in part by the task of summarizing citizen science data that consist of observations from volunteer-led sampling efforts. Such data often feature unequal sampling efforts across some domain such as a geographic region. An interpretable comparison may be made across subregions, e.g., counties in a state, with prediction sets that maintain the same coverage rate for each subregion regardless of its size or composition. To this end, we develop a nonparametric method for categorical data that constructs a small prediction set from a given group with the incorporation of auxiliary data from other groups and maintains a finite-sample frequentist coverage guarantee. We detail a simple algorithm to compute such a prediction set where the computation

time does not depend on the sample size and scales well with number of categories. The usefulness of this method is demonstrated in summarizing avian species abundance in North Carolina with data from the popular eBird database, a citizen science database of species observations supported by the Cornell Lab of Ornithology.

1.2 Covariance estimation for multi-group data

Chapter 4 details work motivated by improving the accuracy of covariance estimates for multi-group data based on the work in Bersson and Hoff (2023a). In general, accurate covariance estimation is necessary for many statistical tasks including classification, principle component analysis, and multivariate regression analysis, among others. To obtain accurate group-specific covariance estimates, shrinkage estimation methods that shrink an unstructured, group-specific covariance either across groups towards a pooled covariance or within each group towards some structure have been developed. In many applications, however, it is unclear which approach will result in more accurate covariance estimates, and, as such, inference may not be much improved over that which results from a naive approach. Flexibly incorporating both shrinkage approaches allows for robustness to such misspecification.

With this motivation, we present a flexible covariance estimation method for multi-group, matrix-variate data in Chapter 4. We detail a hierarchical prior distribution for covariance matrices that flexibly allows for shrinkage across groups towards a pooled covariance or within groups towards a Kronecker structured covariance. The flexible framework yields more accurate covariance estimates than standard methods in situations where simplifying structural assumptions are unknown. The utility of this method is demonstrated in a high dimensional audio file classification application and an environmental health analysis. In the classification application, our approach features superior classification performance over competitors, and in the health analysis, the interpretability of our approach is demonstrated.

Codes to implement the methodologies presented along with replication codes for all simulation studies and most data applications are available at <https://github.com/betsybersson>.

Optimal Conformal Prediction for Small Areas

2.1 Introduction

Precise and accurate inference on a sample obtained from non-overlapping subpopulations, referred to as areas or domains, is an important goal in a wide range of fields where localized inference for various socio-demographic groups or refined geographic regions is of interest. Analyses in economics (Janicki et al., 2022; Molina et al., 2014), ecology (Sinha and Rao, 2009), health (Nandram and Choi, 2010), and other fields rely on survey samples where it is common to have small area-specific sample sizes. As a motivating example, we consider survey data of household radon concentrations across counties in the Midwestern United States, in which 50% of within-county sample sizes are 8 or less. These sample sizes present challenges in making both precise and valid area-specific inferences, which are particularly important in applications where results often have policy implications.

Direct statistical methods that only make use of within-area samples can be unbiased and achieve target frequentist coverage rates for all areas, but don't take advantage of indirect information, and so may be inefficient. As a result, researchers

often turn to indirect or model-based methods that allow information to be shared across areas. Borrowing information across areas may decrease variability of point estimates and volume of confidence and prediction regions, but doing so can introduce bias and thus alter within-area frequentist coverage rates from their nominal level (Carlin and Gelfand, 1990). To address this, parametric ‘frequentist and Bayes’ (FAB) methods have recently been developed for confidence intervals that maintain within-area frequentist error rate control and allow for information sharing across areas (Yu and Hoff, 2018; Burriss and Hoff, 2020). See Bryan and Hoff (2023); McCormack and Hoff (2023) for other FAB-motivated works.

In this chapter, we focus on the task of obtaining a prediction region for a response in each area. In some small area applications, a prediction region for the units in a small area may be as or more useful than a confidence region for the area-specific mean. For example, we illustrate the methods in this chapter using data on household indoor radon concentrations across multiple counties. Policy makers at the county-level may be interested in obtaining a plausible range of household radon levels within their county, instead of or in addition to a confidence interval for the county-level mean. We develop a non-parametric FAB prediction method that has within-area frequentist coverage rate control while incorporating indirect information to improve prediction region precision. The proposed method is Bayesian in a decision-theoretic sense (Lehmann and Casella, 1998) in that a FAB prediction region has minimum Bayes risk if working model assumptions are accurate.

To illustrate the limitations of commonly used parametric prediction methods, consider a study that includes a simple random sample $\mathbf{Y}_j = \{Y_{1,j}, \dots, Y_{n_j,j}\}$ from area j with population mean θ_j , for instance, household radon concentrations within county j , for $j = 1, \dots, J$. We wish to obtain a prediction region A_j for an unsampled response $Y_{n_j+1,j}$ from population j that is accurate, in the sense that it ideally

maintains exact $(1 - \alpha)100\%$ frequentist coverage,

$$\Pr(Y_{n_j+1,j} \in A_j) = 1 - \alpha, \quad (2.1)$$

and precise, in that the expected volume is comparatively small. For normally-distributed data, a commonly used direct method is the classical normal or t pivot prediction interval. To see what can go wrong, consider the simple case where the population variance is known. In this case, the usual prediction interval is

$$\bar{y}_j \pm z_{1-\alpha/2} (\sigma_j^2(1 + 1/n_j))^{1/2}, \quad (2.2)$$

where \bar{y}_j is the sample mean of area j , σ_j^2 is the population variance, and z_q is the q th quantile of the standard normal distribution. If the parametric distributional assumptions hold true, this interval will have the desired frequentist coverage rate. As this interval may be prohibitively wide due to a small sample size, researchers often turn to a Bayesian interval, guaranteed to be narrower than the pivot interval:

$$\hat{\theta}_j \pm z_{1-\alpha/2} \left(\sigma_j^2 \left(1 + (1/\tau^2 + n_j)^{-1} \right) \right)^{1/2}, \quad (2.3)$$

where $\hat{\theta}_j := (\mu/\tau^2 + \bar{y}_j n_j) / (1/\tau^2 + n_j)$, for parameters μ and τ^2 that may be empirical Bayes. This interval approximately achieves the specified coverage rate on average across areas, but the frequentist coverage rate for a specific area j declines from greater than $1 - \alpha$ when $\theta_j = \mu$ to zero as $|\theta_j - \mu|$ increases (see the solid lines in Figure 2.1). Furthermore, if the parametric assumptions underlying either (2.2) or (2.3) are not accurate, the coverage rates can differ from their nominal levels, even on average across groups. For example, the dashed blue line in Figure 2.1 gives the coverage rate of the normal prediction interval 2.2 when the true area-level distribution has variance one and support on two points equidistant from θ_j . In this case, the actual coverage rate of the 75% pivot interval is only 50%. To summarize Figure 2.1, the standard prediction intervals (2.2) and (2.3) can have lower than desired

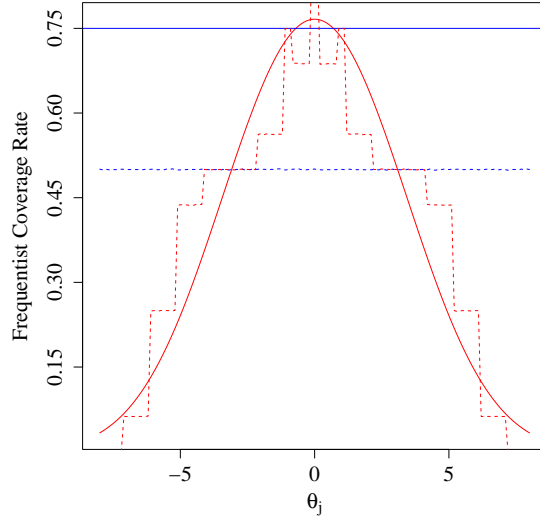


FIGURE 2.1: Frequentist coverage rate of classical pivot prediction (blue) and Bayesian prediction (red) for sample size $n = 3$, known $\sigma^2 = 1$, and Bayesian prior parameters $\mu = 0, \tau^2 = 1/2$, for a target coverage rate of 0.75. Results under a normal (solid lines) and non-normal (dashed lines) within-area distribution.

frequentist coverage rate in the case of either inaccurate parametric assumptions or incorrect prior values.

There is an extensive literature on estimating (or “predicting”) within-area random effects (see, for a review, Skrondal and Rabe-Hesketh (2009) or Pfeiffermann (2013)), but less work has been done on predicting the unit-level responses within an area. Afshartous and De Leeuw (2005) offer a review of parametric point prediction methods. The accuracy and theoretical guarantees of these methods rely on modeling assumptions. Other prediction methods such as those presented in Vidoni (2006), as well as empirical or fully Bayesian prediction methods produce precise prediction intervals (Gelman, 2006), but do not maintain the desired coverage level at each area. For more on small area inference, see Rao and Molina (2015), or, for information on multilevel modeling more broadly, see Gelman and Hill (2006).

Conformal prediction, introduced in Gammerman et al. (1998) and further de-

veloped in Vovk et al. (2005), is a non-parametric method that relies solely on the assumption of exchangeability to produce prediction regions with finite-sample coverage guarantees. A brief review of conformal prediction is in Section 2.2.1. Conformal prediction has primarily focused on methods for a single population (Lei and Wasserman, 2014; Papadopoulos et al., 2011; Vovk et al., 2019). These methods could be used to construct “direct” conformal prediction regions for each area separately, but doing so could be inefficient, as information is not shared across groups. Closely related to our method is the recent work of Dunn et al. (2022) on conformal prediction for two-layer models. They primarily focus on predicting an observation from a new population, or area, using data from previously observed areas and a data-based non-conformity measure. For a specified area, their method results in a prediction region with guaranteed $1 - \alpha$ coverage for that area, on average over data from all of the areas. Furthermore, their prediction approach for a new observation from an observed area is unsupervised. In contrast, our prediction approach focuses on the observed areas, and guarantees $1 - \alpha$ frequentist coverage for each area across data in that area with an approach that allows for supervised prediction via the incorporation of area-specific covariates.

In this chapter, we develop a FAB prediction method that has non-parametric within-area frequentist coverage rate control while incorporating indirect information to improve prediction region precision. Specifically, we build on the generic result shown in Hoff (2023) that conformal prediction regions obtained under the posterior predictive distribution as the conformity measure are optimally precise. We show how this result can be used to obtain narrower prediction regions than standard methods by incorporating indirect information. When the proposed conformity measure is constructed under a normal working model, we prove the resulting prediction region is an interval that contains a standard Bayesian point prediction estimator. This implies a coherent method of classically Bayesian point prediction while providing

uncertainty quantification which maintains frequentist coverage. While many conformal prediction methods for complex conformity measures rely on algorithms which result in approximate prediction regions, we develop a computationally straightforward procedure that makes full use of the data to obtain the exact FAB conformal prediction region.

In Section 2.2, we briefly review the generic conformal prediction method and detail the motivation, properties, and computation of the Bayes-optimal, or FAB, conformity measure for a single population when indirect information is available. Numerical results on the FAB prediction interval’s precision are provided in Section 2.3. In Section 2.4 we extend the FAB conformal algorithm to the small area regime, and in Section 2.5 we apply the proposed prediction method to an EPA radon dataset. We conclude with a discussion in Section 2.6. All supplementary materials, including proofs, are contained in Appendix A.

2.2 Bayes Optimal Conformal Prediction

2.2.1 Review of Conformal Prediction

Conformal prediction is a method of obtaining a prediction region for a new observation Y_{n+1} based on an exchangeable sample $\mathbf{Y} = \{Y_1, \dots, Y_n\}$ from the same population. Having observed $\mathbf{Y} = \mathbf{y}$, a candidate value y_{n+1} of Y_{n+1} is included in the conformal prediction region if it sufficiently “conforms” to the sample, as measured by a conformity measure $C : \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}$ (Vovk et al., 2005). The conformal prediction region can be constructed to have the desired frequentist coverage rate by including only those y_{n+1} -values with corresponding conformity score c_{n+1} greater than or equal to that of some fraction of the conformity scores of the observed elements of the sample, $\{c_1, \dots, c_n\}$. Specifically, a $100(1 - \alpha)\%$ prediction region for Y_{n+1} can be constructed as follows: To determine if a candidate value y_{n+1} is included in the prediction region,

1. compute $c_i(y_{n+1}) := C(\{y_1, \dots, y_n, y_{n+1}\} \setminus \{y_i, y_i\})$ for $i = 1, \dots, n + 1$;
2. set $p_y := \frac{\#\{i = 1, \dots, n + 1 : c_i(y_{n+1}) \leq c_{n+1}(y_{n+1})\}}{n + 1}$.

The value y_{n+1} is included in the region if $p_y > \alpha$. Note that each conformity score $c_i(y_{n+1})$ is a function of the candidate y_{n+1} . More compactly, the conformal prediction region may be expressed as

$$A^c(\mathbf{Y}) = \left\{ y_{n+1} \in \mathcal{Y} : \frac{\#\{i = 1, \dots, n + 1 : c_i(y_{n+1}) \leq c_{n+1}(y_{n+1})\}}{n + 1} > \frac{k}{n + 1} \right\}, \quad (2.4)$$

where $k = \lfloor \alpha(n + 1) \rfloor$. The resulting prediction region $A^c(\mathbf{Y})$ has conservative coverage greater than or equal to $1 - \alpha$ and may have exact coverage if $\alpha = l/(n + 1)$ for some integer $l \in \{0, 1, 2, \dots, n + 1\}$ under some regularity conditions (Dunn et al., 2022; Tibshirani et al., 2019). The frequentist coverage guarantee follows from the exchangeability assumption as all permutations of the collection of random variables $\{Y_1, \dots, Y_{n+1}\}$ are equiprobable, and thus, all permutations of conformity scores $\{c_1, \dots, c_{n+1}\}$ are equiprobable (Balasubramanian et al. (2014) §1.3). A feature of the conformal procedure is that the frequentist coverage guarantee of the conformal method holds regardless of both the true distribution of the random variables and the choice of conformity measure. Thus, a thoughtfully chosen conformity measure can yield a precise prediction region which maintains the desired coverage rate.

2.2.2 Conformal Prediction via a Normal Working Model

Two main criteria of the usefulness of a prediction region are validity and precision. As frequentist validity is guaranteed by the conformal method, we focus on constructing an optimally precise prediction region through a conformity measure that takes advantage of indirect or prior information. This information will enter the conformity measure through a *working model*, that is, a model we use to quantify

different sources of information, but one that does not have to be correct for the inferences to be valid. In this section, we derive a Bayes-optimal conformity measure C_B for a single area using a normal working model:

$$\begin{aligned} Y_1, \dots, Y_n &\sim N(\theta, \sigma^2) \\ \theta &\sim N(\mu, \tau^2 \sigma^2) \\ 1/\sigma^2 &\sim G(a/2, b/2), \end{aligned} \tag{2.5}$$

where Y_1, \dots, Y_n are the measurements on a sample of n units from a single area. Hoff (2023) provides a generic result that the Bayes-optimal conformity measure having the smallest expected volume is the posterior predictive density:

$$C_B(\mathbf{y}, y_{n+1}) = p(y_{n+1}|\mathbf{y}) = \int_{\Theta} p(y_{n+1}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}, \tag{2.6}$$

where $p(y_{n+1}|\boldsymbol{\theta})$ is the probability density of candidate y_{n+1} given model parameters $\boldsymbol{\theta}$, and $p(\boldsymbol{\theta}|\mathbf{y})$ is the posterior density of $\boldsymbol{\theta}$ conditional on data \mathbf{y} . We expand on this result to develop a method which makes use of indirect information to obtain narrower prediction regions than possible under standard methods and derive specific results on FAB conformal prediction under the normal working model.

We proceed with the derivation of the Bayes-optimal conformity measure under a normal working model. A standard calculation shows that the posterior predictive density of the normal working model (2.5) is a non-standard, non-central t density:

$$p(y_{n+1}|\mathbf{y}) = \frac{\Gamma\left(\frac{a_\sigma+1}{2}\right)}{\sqrt{a_\sigma\pi}\Gamma\left(\frac{a_\sigma}{2}\right)} \left(\frac{1}{\sqrt{\sigma_t^2}} \left(1 + \frac{1}{a_\sigma} \frac{(y_{n+1} - \mu_\theta)^2}{\sigma_t^2} \right)^{-(a_\sigma+1)/2} \right), \tag{2.7}$$

where

$$\begin{aligned} \tau_\theta^2 &= (1/\tau^2 + n)^{-1} & \mu_\theta &= (\mu/\tau^2 + \mathbf{1}_n^T \mathbf{y})\tau_\theta^2 \\ a_\sigma &= a + n & b_\sigma &= b + \mathbf{y}^T \mathbf{y} + \mu^2/\tau^2 - (\tau_\theta^2)^{-1} \mu_\theta^2 \end{aligned}$$

and $\sigma_t^2 = \frac{b_\sigma}{a_\sigma}(1 + \tau_\theta^2)$. For now, we assume all hyperparameters of the working model, $\{\mu, \tau^2, a, b\}$, are known. As will be discussed in Section 2.4.2, in practice the parameters may be estimated from indirect information, such as the data of other areas. Given working model hyperparameters, a FAB conformal prediction region, denoted $A^{fab}(\mathbf{Y})$, can be constructed for an area by taking (2.7) as the conformity measure. Regardless of the accuracy of the working model, the resulting prediction region will maintain the specified coverage rate, and if the working model assumptions are accurate, the prediction region will have minimum expected volume.

Given a conformity measure, it is straightforward to test any individual candidate value for inclusion in a prediction region by implementing the conformal algorithm. In practice, testing every candidate prediction value in the sample space may be infeasible due to an infinite number of candidates, or otherwise computationally expensive to do at a meaningful granule. Alternatively, we may make use of the form of the chosen conformity measure to circumvent evaluating the typical conformal algorithm at each candidate value and obtain a more computationally tractable algorithm. As we will show, computations may be facilitated by considering alternative representations of a conformity measure, or ECMs:

Definition 1 (equivalent conformity measure). Two conformity measures are called equivalent conformity measures (ECM) if the resulting conformal prediction regions are equivalent.

The idea of an ECM and how it may be used to simplify computation of the prediction region have been discussed before in the conformal literature. For example, standard computation of the prediction region resulting from the popular distance to the average (DTA) non-conformity measure,

$$C_{avg}(\mathbf{y}, y_{n+1}) = |y_{n+1} - \bar{\mathbf{y}}|, \quad (2.8)$$

seems to require computing the mean of $n + 1$ sets during the execution of the

conformal algorithm. As discussed in Shafer and Vovk (2008), this can be avoided by using an ECM, $C_{avg}(\{\mathbf{y}, y_{n+1}\}, y_{n+1})$. Considering this representation in place of the classically defined measure allows each conformity score to be defined in terms of the sample mean, an element of the sample, and the unknown candidate y_{n+1} . This in turn simplifies the implementation of the conformal algorithm. Under more complex conformity measures such as the Bayes-optimal measure, the computational gain obtained from constructing an algorithm under an ECM may be substantial. It turns out that under the normal working model, the Bayes-optimal conformity measure has the same ECM property as the DTA measure:

Theorem 1. *If Model 2.5 is the working model, $C_B(\mathbf{y}, y_{n+1})$ and $C_B(\{\mathbf{y}, y_{n+1}\}, y_{n+1})$ are ECM.*

Consideration of the ECM $C_B(\{\mathbf{y}, y_{n+1}\}, y_{n+1})$ greatly simplifies the computation of the FAB prediction region. In the evaluation of the conformal algorithm under this measure, the conformity scores corresponding to each element of the sample and the candidate are each a t density with the same parameters. As such, this form of the conformity measure is more convenient to work with. In what follows, we will derive properties and an efficient computational algorithm for the FAB conformal prediction region under this normal working model.

2.2.3 Prediction Region Properties and Computation

By making use of properties of the form of C_B under the normal working model (2.5), we show that the exact conformal prediction region can be obtained by a procedure that involves evaluating a simple function of the sample. Additionally, we prove that the FAB conformal prediction region under the normal working model is an interval that contains the posterior mean estimator of the population mean, $\tilde{\theta} := (\mu/\tau^2 + \sum_{k=1}^n y_k)(1/\tau^2 + n)^{-1}$.

The FAB conformal prediction region can be obtained via a two step process. First, for each $i = 1, \dots, n + 1$, find the sub-region of acceptance,

$$S_i := \{y_{n+1} \in \mathbb{R} : c_{B,i}(y_{n+1}) \leq c_{B,n+1}(y_{n+1})\}. \quad (2.9)$$

Then, information in the set $\{S_1, \dots, S_{n+1}\}$ can be summed over the domain \mathbb{R} to obtain the number of $i = 1, \dots, n + 1$ such that $c_i \leq c_{n+1}$ at each point in the domain. As made clear by the representation of a generic conformal prediction region given in (2.4), this information fully classifies the conformal prediction region for a given error rate α .

This process is visualized for an example sample of size $n = 4$ in Figure 2.2. For clarity, the conformity scores for each value in the sample and the candidate prediction are plotted as a function of the candidate prediction in panel (a). The corresponding sub-regions of acceptance are the regions in the sample space where each conformity score is less than or equal to the conformity score of the candidate. Under the normal working model, each sub-region of acceptance, plotted in panel (b), is an interval that contains $\tilde{\theta}$. This information can be directly translated to the number of conformity scores less than or equal to the candidate conformity score over the sample space. Dividing this value by $(n + 1)$ yields the conformal p -value, p_y (shown in panel (c)). From Figure 2.2(c), it is easy to see for a prediction error rate of, for example, $\alpha = 0.2$, the resulting conformal prediction region is the region where $\#(i : c_i \leq c_{n+1}) > 1$, which is $[-3.1, 2.4]$.

Given the standard form of the Bayes-optimal conformity measure (2.7), the sub-regions of acceptance are difficult to obtain analytically as the candidate y_{n+1} appears non-linearly in each $c_{B,1}(y_{n+1}), \dots, c_{B,n+1}(y_{n+1})$. Upon consideration of the equivalent representation of C_B given in Theorem 1, the regions S_1, \dots, S_n can be expressed in closed form. This allows for efficient and exact computation of the prediction region and, in turn, can be used to prove the FAB prediction region is an interval under the

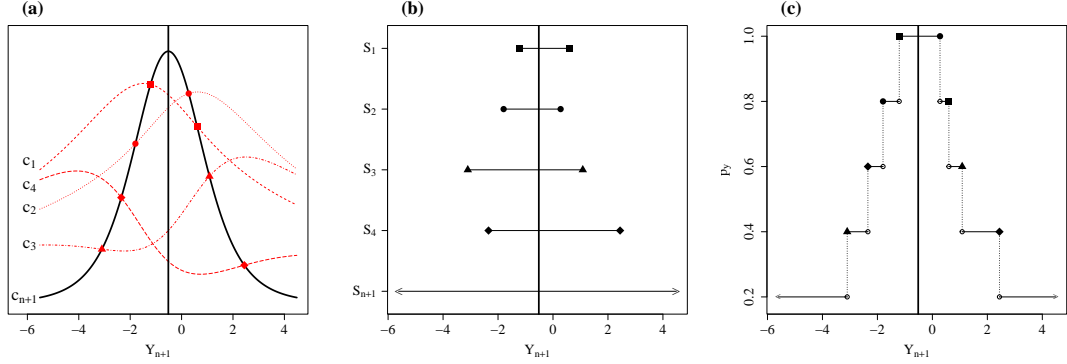


FIGURE 2.2: Schematic of the process to obtain the FAB conformal region: (a) conformity scores for each value in the sample (red dashed curves) and the candidate (thick black curve) over the sample space; (b) sub-regions of acceptance corresponding to the conformity scores; (c) number of conformity scores less than or equal to the candidate conformity score over the sample space; vertical black lines are drawn at $\tilde{\theta}$.

normal working model. These results are formalized below.

We first present two general lemmas (1 and 2) used to prove a conformal prediction region is an interval. If the conformal p -value is a step function over the domain \mathbb{R} with a symmetric number of unit steps to and from $1/(n+1)$ and 1, as in Figure 2.2(c), then the prediction region will be an interval. The following two lemmas may be used to prove this is the case.

Lemma 1. *In the conformal algorithm, if, for each $i = 1, \dots, n+1$, the region*

$$\{y_{n+1} \in \mathbb{R} : c_i(y_{n+1}) \leq c_{n+1}(y_{n+1})\}$$

is an interval and contains some common value, then

$$f(y_{n+1}) = \#\{i \in \{1, \dots, n+1\} : c_i \leq c_{n+1}\}$$

is a step function with ordered output $1, 2, \dots, n, n+1, n, \dots, 2, 1$ over the domain \mathbb{R} .

Lemma 2. *In the conformal algorithm, if $f(y_{n+1})$ is a step function with ordered output $1, 2, \dots, n, n+1, n, \dots, 2, 1$ over the domain \mathbb{R} , then the resulting prediction region is an interval.*

It turns out the hypothesis of Lemma 1 holds for the normal working model case under C_B . Specifically, by utilizing the equivalent form of C_B given by Theorem 1, we are able to conclude that each S_i is an interval for $i \in \{1, \dots, n\}$ and obtain a closed form expression of the bounds:

Lemma 3. *If Model 2.5 is the working model, the region*

$$S_i = \{y_{n+1} \in \mathcal{Y} : C_B(\{\mathbf{y}, y_{n+1}\} \setminus \{y_i, y_i\}) \leq C_B(\{\mathbf{y}, y_{n+1}\} \setminus \{y_{n+1}, y_{n+1}\})\},$$

for each $i = 1, \dots, n$, is an interval $[\min\{y_i, g(y_i)\}, \max\{y_i, g(y_i)\}]$ where

$$g(y_i) := \frac{2 \left(\mu/\tau^2 + \sum_{k \in \{1:n\}} y_k \right) (1/\tau^2 + n + 1)^{-1} - y_i}{1 - 2(1/\tau^2 + n + 1)^{-1}}.$$

Additionally, the posterior mean estimator of the population mean, $\tilde{\theta}$, is contained in each sub-region of acceptance S_1, \dots, S_{n+1} . Not only is this result useful for proving the prediction region is an interval, but it also suggests that using $\tilde{\theta}$ as the estimator for a new prediction and taking $A^{fab}(\mathbf{Y})$ as the prediction interval is a coherent method to predict the unknown value in a classically Bayesian manner while providing uncertainty quantification that maintains the desired frequentist coverage rate:

Lemma 4. *For each $i = 1, \dots, n + 1$, the interval*

$$S_i = \{y_{n+1} \in \mathcal{Y} : C_B(\{\mathbf{y}, y_{n+1}\} \setminus \{y_i, y_i\}) \leq C_B(\{\mathbf{y}, y_{n+1}\} \setminus \{y_{n+1}, y_{n+1}\})\},$$

contains the posterior mean estimator of the population mean,

$$\tilde{\theta} := \left(\mu/\tau^2 + \sum_{k=1}^n y_k \right) (1/\tau^2 + n)^{-1}.$$

In total, these results can be used to prove our main theorem concerning properties of the FAB conformal prediction region constructed under the normal working model:

Theorem 2. *Using conformity measure C_B with Model 2.5 as the working model, the α -level conformal prediction region based on a sample \mathbf{Y} is the k th and $(2n - k + 1)$ th order statistic of \mathbf{v} :*

$$A^{fab}(\mathbf{Y}) = (\mathbf{v}_{(k)}, \mathbf{v}_{(2n-k+1)})$$

for $\mathbf{v} = [y_1 \ \cdots \ y_n \ g(y_1) \ \cdots \ g(y_n)]^T$ and $k = \lfloor \alpha(n + 1) \rfloor$. Furthermore, the conformal prediction region is an interval which contains $\tilde{\theta}$.

The nature of C_B constructed with the normal working model suggests a simple, efficient algorithm to obtain $A^{fab}(\mathbf{Y})$. In particular, as p_y is an incremental step function over the sample space characterized by a symmetry in the number of steps on either side of $p_y = 1$ (e.g. Figure 2.2(c)), the $1 - k/(n + 1)$ prediction region can be obtained by taking the k th ordered step location from either end of the collection of step locations, where $k = \lfloor \alpha(n + 1) \rfloor$. Specifically, to obtain $A^{fab}(\mathbf{Y})$ for known values of working model hyperparameters μ, τ^2 ,

1. for each $i = 1, \dots, n$, compute $g(y_i)$, the critical values of S_i ;
2. set $\mathbf{v} = [y_1 \ \cdots \ y_n \ g(y_1) \ \cdots \ g(y_n)]^T$ and $k = \lfloor \alpha(n + 1) \rfloor$;
3. acquire the bounds of the prediction region, the k th and $(2n - k + 1)$ th order statistics of \mathbf{v} .

Then, the Bayes-optimal conformal prediction region with $(1 - k/(n + 1))100\%$ coverage is

$$A^{fab}(\mathbf{Y}) = (\mathbf{v}_{(k)}, \mathbf{v}_{(2n-k+1)}).$$

If $\alpha(n + 1) = k$, then the prediction interval will have exact coverage, a result which will hold regardless of the values used for $\{\mu, \tau^2\}$, however, as will be discussed extensively in Section 2.3, these values will impact the expected length of the prediction interval. A brief note that if $y_i = g(y_i)$ for at least one $i \in \{1, \dots, n\}$, the resulting prediction region may be a point, dependent on the specified error rate.

2.3 Numerical Comparisons

To demonstrate properties of the FAB prediction method, we numerically evaluate expected prediction interval widths of the FAB prediction regions. We compare the FAB method to a simpler conformal method that does not use indirect information, the distance to average (DTA) conformal prediction method, and the empirical Bayesian (EB) approach described in the Introduction. In particular, DTA (Eqn 2.8) is a popular approach that quantifies non-conformity of a new observation as the distance from the sample average. As this is a non-conformity measure, as opposed to a conformity measure, computation of the DTA prediction region requires determining if the candidate is “too different” from the sample. This corresponds to flipping the direction of the inequality in Step (2) of the generic conformal algorithm.

Both the DTA and FAB methods provide nonparametric frequentist coverage guarantees; the main difference between these two methods is the ability to utilize prior information in the construction of the prediction interval. An EB prediction interval is constructed via incorporation of prior information, but, in contrast to the FAB method, frequentist coverage of an EB prediction interval relies on, among other things, the accuracy of this information. While the conformal prediction methods may be applied to non-normal populations, in this section, we explore their behavior when obtained from a normal sample of size n , $Y_1, \dots, Y_n \sim N(\theta, 1)$. As shown in Figure 2.3 (a), the FAB method shifts prediction bounds away from the true population mean and towards the incorporated prior information. When this prior information well-informs the true mean, the FAB interval is narrower than the alternatives. As the accuracy of the prior information declines, the FAB interval widens, but the coverage guarantee holds. Comparatively, when the prior information is inaccurate, the EB intervals remain narrow, but the corresponding frequentist coverage rate falls below the nominal level (Figure 2.3 (b)).

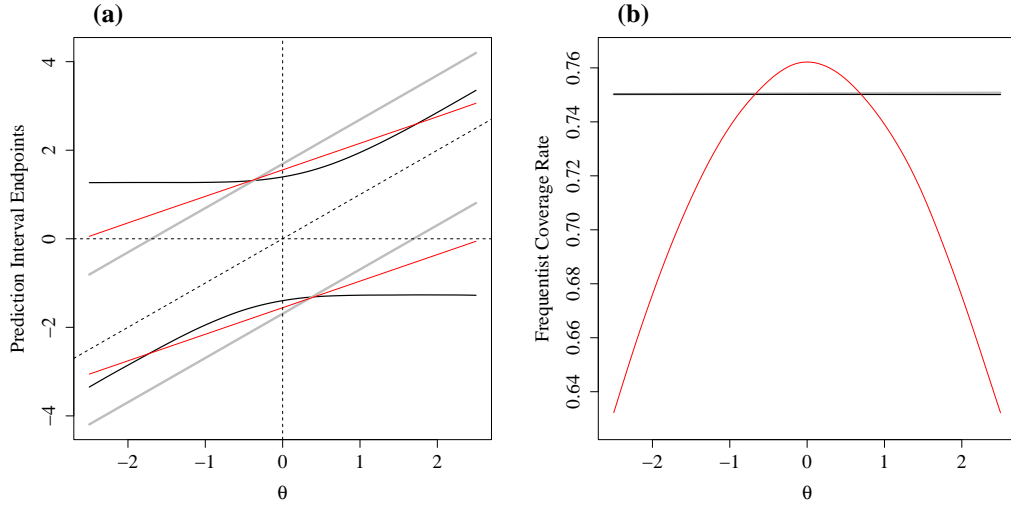


FIGURE 2.3: (a) Expected prediction interval end points and (b) estimated frequentist coverage (dashes) with 95% Wilson confidence intervals. Results plotted for FAB (solid black lines), DTA (thick grey lines), and EB (red lines) prediction methods. For $n = 3, \theta \in [-2.5, 2.5], \mu = 0, \tau^2 = 1/2, \alpha = .25$.

The prior parameters used in the FAB conformal method are $\{\mu, \tau^2\}$ and respectively represent the prior expected population mean and confidence in this expectation. To assess the effect of these prior parameters, we allow them to vary in this study. We consider a prediction error rate of 0.25 and compare numerical results for various sample sizes $n \in \{3, 7, 11, 15, 19\}$, chosen such that the conformal methods will result in regions with exact coverage. In general, the FAB conformal method outperforms the standard DTA method in terms of precision when there is concentrated and accurate prior information regarding the mean of the population, but a limited amount of information in the sample itself. More specifically, the FAB conformal interval can be expected to produce narrower intervals than standard methods when $|\theta - \mu|$ is small, τ^2 is small, and n is small, or a combination of these properties. The EB approach often results in narrower prediction intervals than the FAB approach, but this comes at a cost to the frequentist coverage rate.

The ratio of expected interval widths of the FAB prediction method relative

to the distance to average method are displayed in Figure 2.4. Recall that the FAB method incorporates prior information while the DTA method does not. We compute the expected interval widths via Monte Carlo approximation using 25,000 independently generated replications for each combination of values of θ , τ^2 , and n . The effects of sample size and prior variance of the population mean are the focus of Figure 2.4(a). This figure plots the ratio of Bayes expected interval widths (Bayes risk) of the FAB conformal to distance to average conformal intervals, where the expectations are taken with respect to \mathbf{Y} and θ under the sampling model $Y_1, \dots, Y_n \sim N(\theta, 1)$ and prior $\theta \sim N(\mu, \tau^2)$. The Bayes risk of the conformal intervals does not depend on μ . As informed by the theoretical Bayes-optimality of FAB conformal prediction, the Bayes risk of the FAB interval is smaller than that of the DTA interval, with the overall deviation between the methods' expected interval widths decreasing as the sample size increases. The FAB interval is substantially narrower than the distance to average interval for a wide range of τ^2 values under very small sample sizes. Intuitively, for small sample sizes, even a low level of confidence around the prior value of the population mean μ is useful information and will translate to narrower prediction intervals if utilized in the construction of C_B . More confidence, as conveyed through a smaller τ^2 value, translates to a more substantive increase in precision. Even for larger sample sizes, a nontrivial gain in precision occurs under small (less than about 0.5) τ^2 values, representing very informative prior information about the population mean through a prior with tight concentration around μ .

For a given concentration level τ^2 , it is natural to consider how $|\theta - \mu|$ affects the resulting FAB prediction interval width. Figures 2.4(b)-(c) display the expected interval width ratio of the FAB conformal to the distance to average conformal method for varying population means and sample sizes when $\mu = 0$ and $\tau^2 = 1/2, 2$. Under this set-up, by the Bayes-optimal property of the conformal measure, the FAB interval will outperform alternatives when $|\theta - \mu| \approx 0$ and τ^2 is small. The numerical

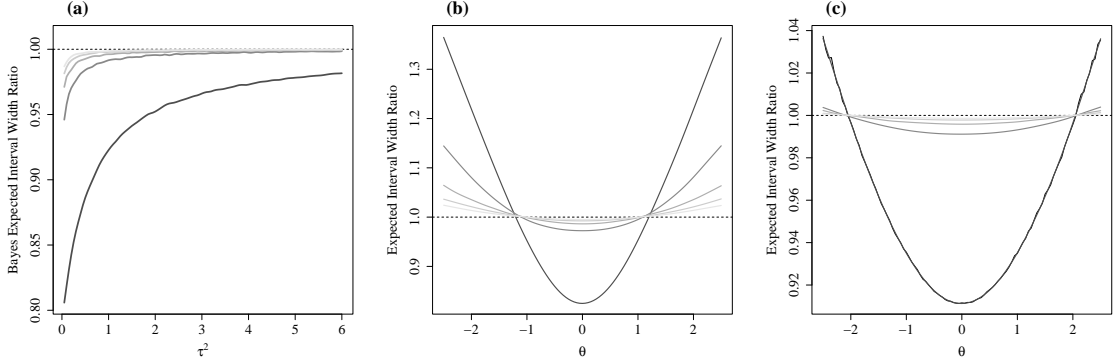


FIGURE 2.4: Expected width ratio of FAB conformal interval to DTA conformal interval for increasing $n \in \{3, 7, 11, 15, 19\}$ in decreasing darkness: (a) expectation taken with respect to \mathbf{Y} and θ ; (b) expectation taken over \mathbf{Y} conditional on θ for $\mu = 0, \tau^2 = 1/2$; (c) same as (b) for $\tau^2 = 2$.

results match this conclusion, and, as the distance between the prior mean and the population mean increases in absolute value, the FAB intervals become wider, and thus the benefit of utilizing this type of prior information declines. As seen in panel (b), for a moderately small τ^2 value, and for the smallest sample size considered, the FAB conformal method results in an interval width that is 17.6% narrower than the standard when $\theta = \mu$ exactly. For larger τ^2 , as seen in the panel (c), there is a less substantial benefit to utilizing this indirect information in C_B , but some benefit is seen nonetheless for a wider range of θ divergences from μ .

A comparison between the FAB approach and the EB approach is presented in Figure 2.5. As seen in panel (a), the EB method often has narrower Bayes risk, but, as discussed, it is not guaranteed to have the desired frequentist coverage. In comparing expected interval width ratios (panels (b-c)), the EB intervals may be narrower, but frequentist coverage relies on the accuracy of the prior information, among other features. The FAB intervals will have the desired validity regardless of the accuracy of either the working model or the prior information.

Overall, when accurate prior information is available, FAB prediction intervals

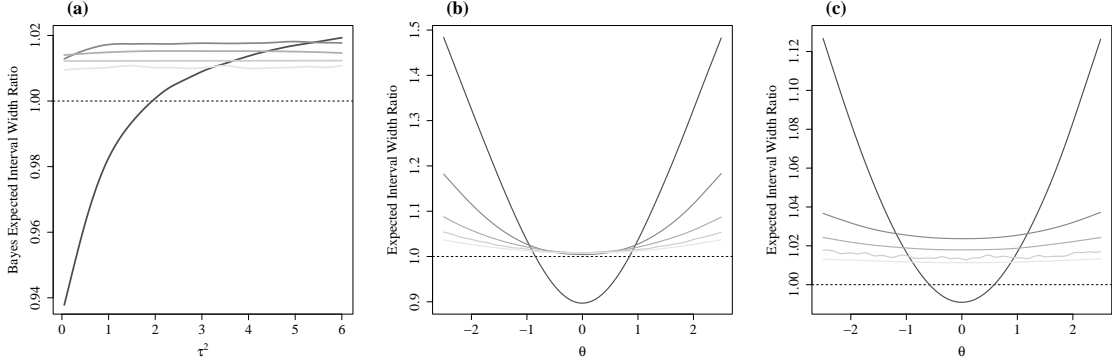


FIGURE 2.5: Expected width ratio of FAB conformal interval to parametric EB interval for increasing $n \in \{3, 7, 11, 15, 19\}$ in decreasing darkness: (a) expectation taken with respect to \mathbf{Y} and θ ; (b) expectation taken over \mathbf{Y} conditional on θ for $\mu = 0, \tau^2 = 1/2$; (c) same as (b) for $\tau^2 = 2$.

outperform commonly utilized conformal prediction intervals in terms of precision. The benefit is particularly large for small sample sizes. Regardless of the accuracy of the prior information, the frequentist coverage rate of FAB prediction intervals is guaranteed. This desirable feature is lost for parametric indirect methods such as the EB approach.

2.4 FAB Small Area Prediction

2.4.1 Information Sharing via a Working Model

In inference on small areas, utilizing indirect methods that share information across areas has been shown to improve precision compared with direct methods, particularly for areas with small sample sizes (Gelman and Hill, 2006). With this motivation, we extend the FAB conformal prediction method to a small area regime. In the construction of the Bayes-optimal conformity measure, information is shared across areas via a multilevel working model in order to increase prediction region precision while maintaining area-level frequentist coverage guarantees.

For each area $j \in \{1, \dots, J\}$, we observe an exchangeable sample $(Y_{1,j}, \dots, Y_{n_j,j}) = (y_{1,j}, \dots, y_{n_j,j})$ of length n_j such that the samples are independent across areas. Sup-

pose a reasonable working model for the populations is a spatial Fay-Herriot model (Fay and Herriot, 1979) that allows for heterogeneous area-specific variances. Specifically,

$$Y_{1,j}, \dots, Y_{n_j,j} \sim N(\theta_j, \sigma_j^2), \text{ independently for } j = 1, \dots, J \quad (2.10)$$

$$\boldsymbol{\theta} \sim N_J(\mathbf{X}\boldsymbol{\beta}, \eta^2\mathbf{G})$$

$$1/\sigma_1^2, \dots, 1/\sigma_J^2 \sim G(a/2, b/2),$$

where \mathbf{G} is a spatial covariance matrix such as that which results from the popular simultaneous (SAR) autoregressive model $\mathbf{G} = [(\mathbf{I} - \rho\mathbf{W})(\mathbf{I} - \rho\mathbf{W}^T)]^{-1}$ (Singh et al., 2005). The matrix \mathbf{W} is a distance matrix among areas that is typically row-standardized to sum to 1, and $\rho \in (-1, 1)$ is a spatial correlation parameter. For more on spatial modeling, see Banerjee et al. (2014). This flexible set-up allows for inclusion of an array of indirect information including area-level covariates \mathbf{X} and spatial relationships in the linking model for the population means which can be exploited to improve precision of prediction regions.

For population j , a FAB conformal prediction interval may be constructed as follows. First, unknown parameters in the working model (Eqns 2.10) can be estimated indirectly, from data independent of area j , $\mathbf{Y}_{-j} := \{\mathbf{Y}_1, \dots, \mathbf{Y}_J\} \setminus \{\mathbf{Y}_j\}$. Then, these values may be used to estimate a mean μ and variance ratio τ^2 of the population mean θ_j , to be used in the Bayes optimal conformity measure (2.7) along with the sample from area j , \mathbf{Y}_j . The resulting prediction region is constructed from a measure that shares information across areas and will maintain the desired frequentist coverage rate for each area.

2.4.2 FAB Conformal Parameter Estimation Procedure

All that remains to implement the FAB conformal method for area j is to acquire values for the unknown prior parameters $\{\mu, \tau^2\}$, which are the prior mean and variance ratio, respectively, of area j 's population mean θ_j used in the Bayes-optimal

conformity measure (Eqn 2.7). Estimates of these prior parameters are used along with the j th sample \mathbf{Y}_j to obtain the critical values $g(\cdot)$ in the proposed FAB algorithm. If the parameters of the working model (Eqns 2.10) are known, we can take $\{\mu, \tau^2\}$ to be the conditional mean and conditional variance proportion of θ_j and proceed with implementation of the Bayes-optimal conformal algorithm. In practice, of course, these values are not known, but they may be estimated from indirect data via standard techniques.

We propose an empirical Bayesian approach whereby values of the prior parameters are obtained for each area j using samples from all other areas, \mathbf{Y}_{-j} . As an overview, for area j , our estimation procedure proceeds with first computing estimates of unknown parameters in the working model using \mathbf{Y}_{-j} . Then, given these estimates, we obtain the conditional mean of θ_j and the proportion of the conditional variance of θ_j to an estimate of area j 's population variance, which are labeled as μ_j and τ_j^2 , respectively.

In more detail, we first obtain the maximum likelihood estimates (MLEs) of a, b by maximizing the marginal density of $\{M_k^2\}_{k \in K}$ for $K = \{1, \dots, J\} \setminus j$ where $M_k^2 = \sum_{i=1}^{n_k} (\mathbf{Y}_{ik} - \bar{\mathbf{Y}}_k)^2$. Under the assumptions of the working model (2.10), this marginal density can be shown to be

$$p(\{m_k^2\}_{k \in K} | a, b) = \prod_{k \in K} f(m_k^2) \frac{\Gamma\left(\frac{a+n_k-1}{2}\right) \left(\frac{b}{2}\right)^{a/2}}{\Gamma\left(\frac{a}{2}\right) \left(\frac{b+m_k^2}{2}\right)^{(a+n_k-1)/2}} \quad (2.11)$$

for a function f that does not depend on the hyperparameters a, b . We use the resulting MLEs to obtain empirical Bayes estimates of each area's population variance. That is,

$$\begin{aligned} \hat{\sigma}_k^2 &= \text{Mode}[\sigma_k^2 | m_k^2, \hat{a}, \hat{b}] = (\hat{b} + m_k^2) / (\hat{a} + (n_k - 1) + 2), \quad \text{for } k \in K, \text{ and} \\ \hat{\sigma}_j^2 &= \text{Mode}[\sigma_j^2 | \hat{a}, \hat{b}] = \hat{b} / (\hat{a} + 2). \end{aligned}$$

See Appendix A for details on the derivations of the marginal density and the conditional modes. Taking $\{\hat{\sigma}_k^2\}_{k \in K}$ as plug-in values of the population variances, we obtain MLEs $\{\hat{\boldsymbol{\beta}}, \hat{\eta}^2, \hat{\rho}\}$ of the mean prior hyperparameters $\{\boldsymbol{\beta}, \eta^2, \rho\}$ through standard REML or ML procedures (Pratesi and Salvati, 2008) with data \mathbf{Y}_{-j} . These MLEs may then be used to obtain an estimate $\hat{\boldsymbol{\theta}}_{-j}$ of $\boldsymbol{\theta}_{-j}$. Finally, empirical Bayes estimates of the prior parameters of area j are the conditional mean and the proportion of conditional variance of θ_j obtained given $\{\hat{\boldsymbol{\beta}}, \hat{\eta}^2, \hat{\rho}, \hat{\boldsymbol{\theta}}_{-j}, \{\hat{\sigma}_k^2\}_{k \in K}, \hat{\sigma}_j^2\}$. By properties of the conditional normal distribution, that is:

$$\mu_j = E[\theta_j | \boldsymbol{\theta}_{-j} = \hat{\boldsymbol{\theta}}_{-j}, \boldsymbol{\beta} = \hat{\boldsymbol{\beta}}, \eta = \hat{\eta}, \rho = \hat{\rho}] \quad (2.12)$$

$$= \mathbf{x}_j^T \hat{\boldsymbol{\beta}} + \mathbf{V}_{[j,-j]} \mathbf{V}_{[-j,-j]}^{-1} \left(\hat{\boldsymbol{\theta}}_{-j} - \mathbf{X}_{[-j]} \hat{\boldsymbol{\beta}} \right)$$

$$\tau_j^2 = \text{Var}[\theta_j | \boldsymbol{\theta}_{-j} = \hat{\boldsymbol{\theta}}_{-j}, \boldsymbol{\beta} = \hat{\boldsymbol{\beta}}, \eta = \hat{\eta}, \rho = \hat{\rho}] / \hat{\sigma}_j^2 \quad (2.13)$$

$$= \left(\mathbf{V}_{[j,j]} - \mathbf{V}_{[j,-j]} \mathbf{V}_{[-j,-j]}^{-1} \mathbf{V}_{[-j,j]} \right) / \hat{\sigma}_j^2$$

where $\mathbf{V} = \hat{\eta}^2 [(\mathbf{I} - \hat{\rho} \mathbf{W})(\mathbf{I} - \hat{\rho} \mathbf{W}^T)]^{-1}$.

Given these prior values, $\{\mu_j, \tau_j^2\}$, obtained from information independent of area j , the conformal algorithm proceeds as described in Section 2.2.3 based on the sample from area j . For each area, the algorithm yields an interval that may have improved precision over other methods as a result of information sharing and maintains the specified frequentist coverage rate.

2.5 Radon Data Example

Radon is an odorless, colorless natural gas that is a known carcinogen. Exposure to radon is the leading cause of lung cancer among non-smokers and is responsible for tens of thousands of deaths in the U.S. each year. To evaluate radon exposure risk, the U.S. Environmental Protection Agency (EPA) conducted an extensive national survey (US Environmental Protection Agency, 1992) of indoor radon concentrations

measured from a random sample of households across 9 at-risk states. In this section, we compare prediction regions for household radon levels within a county using data collected in Minnesota and North Dakota, which consists of 2515 observations in total throughout the states' 138 counties. The within-county sample sizes range from 1 to 172, with a median of about 8 households.

Price et al. (1996) modeled household radon concentrations at the county level in Minnesota with a goal of improving estimated county-level means, and accurate county-specific predictions are cited as being of particular interest. Due to the small within-county sample sizes, these are difficult tasks. Given the abundance of indirect information, including county-level covariates and apparent spatial relationships among radon concentrations across counties, incorporating indirect information in the construction of confidence or prediction intervals is a natural tool to improve inferential precision. In what follows, we compare prediction intervals resulting from FAB, DTA, and EB methods. As county-specific predictive inference is of primary interest, an ideal prediction interval will have the desired $(1 - \alpha)100\%$ frequentist coverage for every county while maintaining an interval width that is practically informative.

We construct prediction intervals for the log radon concentration in a randomly sampled household in each county. Exploratory analyses suggest log radon values are approximately normally distributed, so we utilize the normal working model (2.10) in the construction of the FAB conformal prediction intervals, and follow the prior parameter estimation procedure detailed in Section 2.4.2. The same prior parameters are used in the construction of both the EB intervals. Specifically, we include a county-wide soil uranium measurement as a covariate, incorporate a shared county-wide prior intercept, and allow for spatial effects under the (row-standardized) squared exponential distance matrix \mathbf{W} between county centroids. That is, before row-standardization, the matrix entries are $\{w_{lk}\} = e^{-d(x_l, x_k)^2}$ for counties l, k where

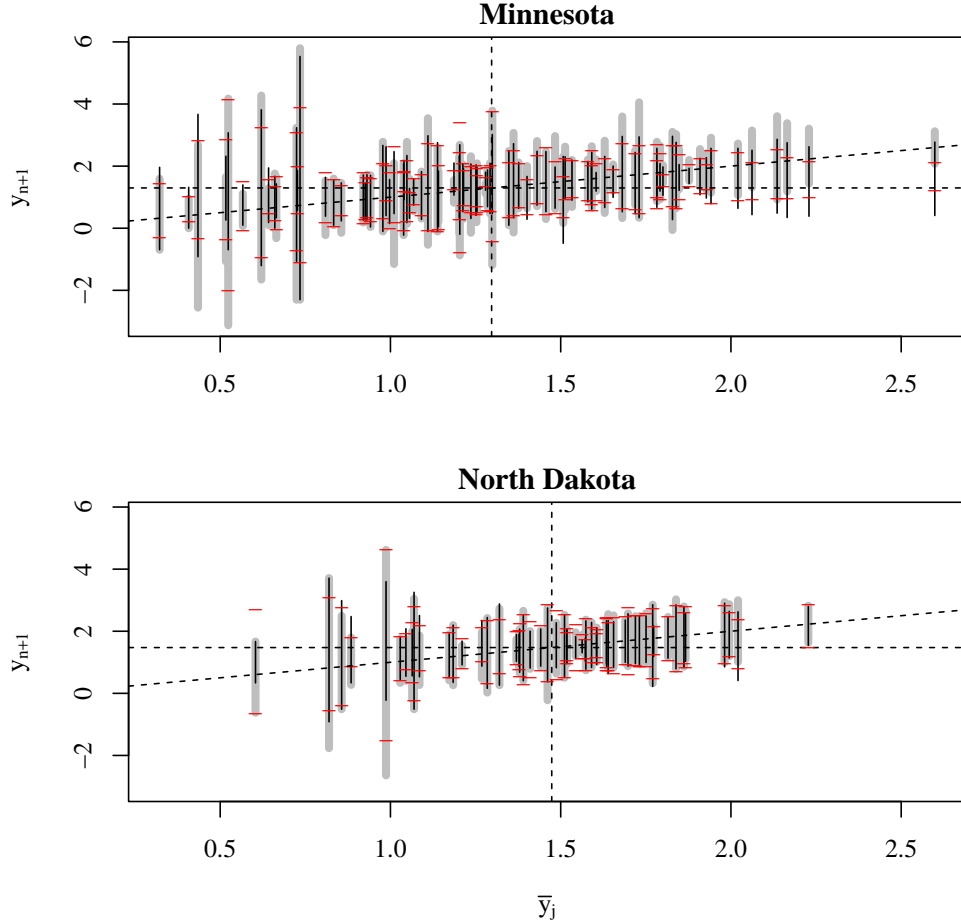


FIGURE 2.6: County-level radon prediction intervals for FAB (solid black lines), DTA (thick grey lines), and EB end points (red dashes) methods. Dashed lines drawn at the state-wide sample mean $\sum \bar{y}_j / J$ and 45° line.

d is a distance function. Preliminary analyses of county-level radon sample means indicate that utilizing this distance metric may be more beneficial for the counties in Minnesota than in North Dakota. In particular, the Geary's C (Geary, 1954) value, a measure of spatial autocorrelation, is further from 1 in Minnesota (0.727) than in North Dakota (0.896), indicating there may be stronger spatial autocorrelation in Minnesota than in North Dakota.

FAB, DTA, and EB prediction intervals are obtained for counties with sample sizes greater than 1 under a county-specific error rate $\alpha_j = \lfloor \frac{1}{3}(n_j + 1) \rfloor / (n_j + 1)$ to

allow for exact $1 - \alpha_j$ frequentist coverage of prediction intervals for each county j . The prediction intervals for each county are plotted in Figure 2.6. In summary, including relevant indirect information in the construction of prediction intervals via the Bayes-optimal conformal procedure results in improved overall interval precision. Specifically, the FAB prediction intervals are narrower than the DTA intervals in 62.2% of counties. At a state level, the FAB intervals are narrower for a higher percentage of counties in Minnesota (65.9%) than in North Dakota (56.6%). The FAB intervals are only narrower than the EB intervals for 42.2% of counties, but recall that the EB method relies on parametric assumptions. If the normality assumption does not hold, or if the prior parameter estimates are inaccurate, the EB intervals may under-cover.

The FAB and EB intervals exhibit classical ‘Bayesian’ or shrinkage behavior in that they are shifted towards the shared region-wide sample mean, while each DTA interval is centered near the respective county-specific sample mean. By the Bayes-optimality property of the FAB prediction method, FAB prediction intervals are narrower than all alternative prediction regions for counties where the working model assumptions are accurate. As such, in this case, the FAB intervals perform best in terms of precision for counties where the heterogeneity across county-specific mean radon values is well described by the spatial and covariate information.

While the FAB prediction intervals are often narrower than the DTA intervals, they are expected to be wider in counties where the utilized prior information is not accurate (Figure 2.4). Practically, this corresponds to outlier counties where available indirect information does not accurately inform that within-county population mean. Furthermore, this behavior is exacerbated by small sample sizes. The relationship between sample size and prediction interval width in the radon data is visualized in Figure 2.7, which plots the ratio of FAB prediction interval width to DTA interval width (panel (a)) and FAB prediction width to EB interval width (panel (b)). As

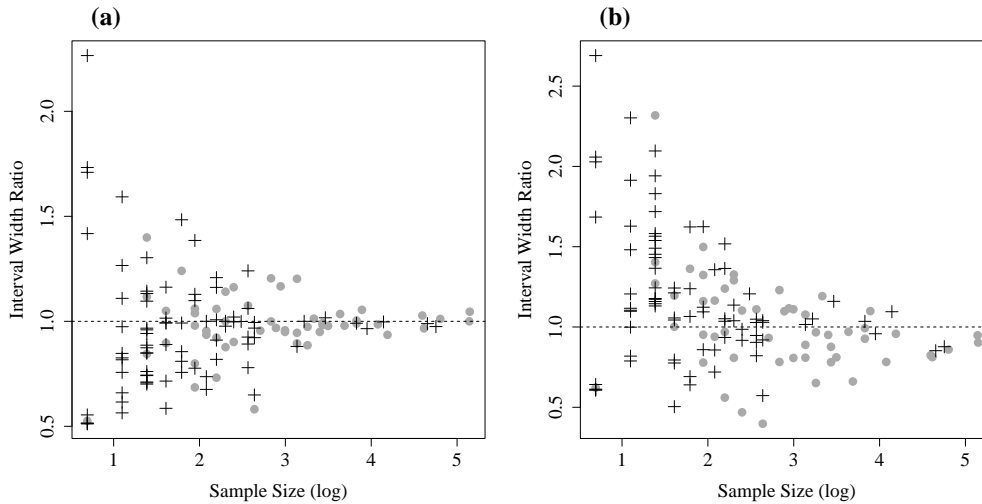


FIGURE 2.7: County-level radon prediction interval width ratio of (a) FAB to DTA and (b) FAB to EB. Black cross marks represent ratios obtained from samples in counties in Minnesota and grey circles represent ratios obtained from samples in counties in North Dakota.

the sample from Minnesota has a higher frequency of small sample sizes than that from North Dakota, there is more opportunity for a substantial gain in county-specific interval precision over the DTA method, and consequently there is also the opportunity for a substantial loss, as seen in the black crosses in Figure 2.7 (a) corresponding to counties with very small sample sizes. As expected, as sample sizes increase, all methods perform similarly in terms of precision. Similar patterns are seen in panel (b) where the EB intervals tend to be narrower than FAB in counties with small sample sizes.

Overall, sharing information via the FAB conformal method can result in narrower prediction intervals than standard non-parametric methods. Standard parametric methods like the EB approach may provide narrower intervals, but the validity of such methods is not guaranteed. As these data were obtained by a simple random sample, each county's FAB and DTA prediction intervals maintain the desired validity guarantee based on the theoretical conformal prediction result which relies

solely on exchangeability of the samples.

2.6 Discussion

The FAB conformal prediction method introduced in this chapter, which utilizes the posterior predictive density of some working model as a conformity measure, produces precise and accurate prediction regions. If the working model assumptions are accurate, FAB prediction regions have minimum expected volume, and, regardless of the accuracy of the working model, they have guaranteed frequentist coverage rate control for each small area. When constructed under a normal working model, FAB conformal prediction regions contain the standard posterior mean Bayes estimator $\tilde{\theta}$. This implies a coherent method of Bayesian point prediction, through $\tilde{\theta}$, where uncertainty quantification maintains the specified frequentist coverage rate. Furthermore, exact FAB conformal prediction regions may be obtained in a straightforward manner.

In practice, FAB conformal intervals are notably narrower than standard prediction intervals when accurate prior information for the population mean is available, especially in the presence of small sample sizes. The FAB conformal prediction method leverages this accurate indirect information to improve interval width precision. As such, this method is particularly useful for small area applications where within-area frequentist coverage guarantees are desirable but sample sizes are small. While commonly-used small area prediction methods offer a trade-off between accuracy and precision, the FAB prediction method maintains a guarantee of accuracy while allowing for increased precision via across-area information sharing.

This chapter focuses on FAB prediction intervals constructed under a normal working model, but the Bayes-optimal conformity measure can be constructed under alternative working models that may be more appropriate for different types of data. Relatedly, our approach is well suited for simple random sample study designs, and

we aim to consider extensions to more complicated study design in future work. The framework presented in Section 2 may be utilized to aid in derivation of efficient computation of conformal prediction regions for either of these extensions.

Prediction Sets for Species Abundance using Indirect Information

3.1 Introduction

Understanding species abundance across heterogeneous spatial areas is an important task in ecology. Citizen science databases that consist of observations of species counts gathered by volunteers are increasingly regarded as one of the richest sources of data for such a task. One of the largest such data sources is the eBird database in which citizen scientists throughout the world input counts of bird sightings (Sullivan et al., 2009). In addition to its use for describing avian species abundance, eBird is a principal resource for understanding global biodiversity and is widely used in constructing and implementing conservation action plans (Sullivan et al., 2017).

More generally, analyses from such databases may be used for informing policy, conservation efforts, habitat preservation, and more, for which understanding species prevalence for non-overlapping geographic areas, such as counties across a state or country, is important. In practice, species abundance from citizen science data are commonly summarised within areas such as counties by empirical proportions from

a sample, as in, e.g., Arnold et al. (2021); Camerini and Groppali (2014). Such proportions can be used to construct a prediction set for each county that provides a description of species prevalence for that county with guaranteed frequentist coverage.

Given the impact on policy design, corresponding uncertainty quantification is of particular import (Lele, 2020), and so it is desirable that precise prediction sets maintain a target coverage rate regardless of the county’s size or composition. This is challenging as a common feature of citizen science data is unequal sampling efforts that results in some counties with large amounts of data information and others with very little. Using direct procedures that only make use of within-county information, a prediction set may be imprecise in these counties with low sampling efforts. This suggests using indirect information such as data from neighboring counties to improve prediction set precision for a given county.

In this chapter, we describe species abundance across sampling areas such as counties with frequentist-valid prediction sets that are constructed to contain an unobserved bird with $1 - \alpha$ probability. That is, a valid prediction set for a given county is a set of avian species such that an unobserved bird will belong to one of those species with $1 - \alpha$ probability in a frequentist sense. We develop a valid nonparametric prediction method that allows for information to be shared across counties. Specifically, our approach results in prediction sets with guaranteed frequentist coverage for each county that are constructed with the incorporation of indirect or prior information. We detail and provide code for an empirical Bayes procedure to estimate such prior information from auxiliary data such as neighboring counties. If this indirect information used to construct the prediction sets is accurate, the prediction sets will be smaller than direct sets that only make use of within-county information.

In Section 3.4, we detail the usefulness of the proposed approach in summarising

the eBird citizen science data. Sharing information across counties generally results in smaller prediction sets as compared to direct prediction approaches, particularly so in counties with low sampling efforts. Moreover, the prediction sets provide a useful summary of the data that may be used to compare information across areas and better inform policy.

3.2 Methodology

3.2.1 Background and Notation

For county $j \in \{1, \dots, J\}$, let \mathbf{X}_j be a vector of length K where $X_{j,i} = x_{j,i}$ is the observed count of species i over some set sampling period that may vary across counties. We model \mathbf{X}_j with a K -dimensional multinomial distribution with $N_j = \sum_{i=1}^K x_{j,i}$ trials and population proportions vector $\boldsymbol{\theta}_j$,

$$\mathbf{X}_j \sim MN_K(\boldsymbol{\theta}_j, N_j). \quad (3.1)$$

We construct a prediction set $A_\alpha(\mathbf{X}_j)$ for an observation of a new bird arising from the same distribution, $\mathbf{Y}_j \sim MN_K(\boldsymbol{\theta}_j, 1)$ where $\mathbf{Y}_j \in \mathcal{Y}$ for $\mathcal{Y} = \{(y_1, \dots, y_K) : \sum_{i=1}^K y_i = 1, y_i \in \{0, 1\} (i = 1, \dots, K)\}$. Let $\mathbf{y}_j^{(k)} \in \mathcal{Y}$ denote a prediction of category k , that is, let $\mathbf{y}_j^{(k)}$ be a vector of length K with a one at index k and zeros elsewhere. In particular, we are interested in a prediction set for \mathbf{Y}_j that maintains frequentist validity for some error rate α . Formally, we refer to this as an α -valid prediction set:

Definition 2 (α -Valid Prediction Set). An α -valid prediction set for a predictand $\mathbf{Y}_j \in \mathcal{Y}$ is any subset A_α of the sample space \mathcal{Y} that contains \mathbf{Y}_j with probability greater than or equal to $1 - \alpha$,

$$P_\theta(\mathbf{Y}_j \in A_\alpha(\mathbf{X}_j)) \geq 1 - \alpha, \quad \forall \theta, \quad (3.2)$$

where the probability is taken with respect to \mathbf{Y}_j and \mathbf{X}_j .

Additionally, small or precise α -valid prediction sets are of particular interest, where prediction set size is measured by expected cardinality, that is, expected number of the K categories in the sample space included in the prediction set.

3.2.2 Order-based prediction for a single area

A standard approach to construct α -valid prediction sets for each county or area is with a direct method that only makes use of within-area information. As such, we first consider construction of a prediction set for a single area j , using only data from area j . For ease of notation, we drop the area-identifying subscript in this subsection.

For multinomial data in general, if the event probability vector $\boldsymbol{\theta}$ is known, an α -valid prediction set is any combination of categories such that their event probabilities cumulatively sum to be greater than or equal to $1 - \alpha$. Equivalently stated, an α -valid prediction set may be constructed by excluding categories such that the cumulative sum of the excluded categories' event probabilities is less than α . Such a prediction set may be constructed by admitting categories in some prespecified order into the prediction set until the cumulative sum of their event probabilities is at least $1 - \alpha$. The resulting prediction set will have $1 - \alpha$ coverage regardless of the ordering used to admit categories. In fact, the class of all α -valid prediction sets may be constructed by following this procedure for non-strict total orderings of categories.

Perhaps intuitively, constructing such a prediction set by including categories with the largest event probabilities will result in the smallest α -valid prediction set. In the terminology of ordering, this corresponds to constructing a prediction set based on an ordering of categories that matches the ordering of the elements in $\boldsymbol{\theta}$. We refer to this optimal ordering as the oracle ordering:

Theorem 3 (Oracle order-based prediction). *Let $\mathbf{Y} \sim MN_K(\boldsymbol{\theta}, 1)$ for $\boldsymbol{\theta}$ known. Then,*

(a) the class of all α -valid prediction sets for a given $\boldsymbol{\theta}$ consists of prediction sets of the form,

$$A_{\alpha}^{\boldsymbol{\theta}, \boldsymbol{o}} = \left\{ \mathbf{y}^{(k)} \in \mathcal{Y} : \left[\sum_{l=1}^K \mathbb{1}(o_k \geq o_l) \theta_l \right] > \alpha \right\}, \quad (3.3)$$

for some vector $\boldsymbol{o} \in \mathbb{R}^K$, and

(b) the **oracle ordering** is that which corresponds to the increasing order statistics of $\boldsymbol{\theta}$,

$$\boldsymbol{o}^{\theta} = \{ \boldsymbol{o} : \theta_m < \theta_n \Rightarrow o_m < o_n \forall m, n \in \{1, \dots, K\}, m \neq n \},$$

and $A_{\alpha}^{\boldsymbol{\theta}, \boldsymbol{o}^{\theta}}$ has the smallest cardinality among all orderings.

In practice, $\boldsymbol{\theta}$ is unknown, but a prediction set may be constructed based on an observed sample $\mathbf{X} = \mathbf{x}$. It turns out, in fact, that any conditional α -valid prediction set can be written similarly to the previous construction (Equation 3.3) where the cumulative sum is computed with respect to the empirical proportions given by \mathbf{x} and \mathbf{y} . This is a generalization of the conformal prediction framework, a popular machine learning approach to construct prediction regions based on measuring conformity (or non-conformity) of a predictand to an observed sample (Vovk et al., 2005).

Theorem 4 (α -valid order-based prediction). *Let $\mathbf{X} \sim MN_K(\boldsymbol{\theta}, N)$, $\mathbf{Y} \sim MN_K(\boldsymbol{\theta}, 1)$. Then, every conformal α -valid prediction set based on observed data \mathbf{x} can be written*

$$A_{\alpha}(\mathbf{x}) = \left\{ \mathbf{y}^{(k)} \in \mathcal{Y} : \left[\sum_{l=1}^K \mathbb{1}(o_k \geq o_l) \frac{x_l + y_l^{(k)}}{N + 1} \right] > \alpha \right\}, \quad (3.4)$$

for some vector $\boldsymbol{o} \in \mathbb{R}^K$.

Note that the prediction set depends on the vector \boldsymbol{o} only through the order of its elements.

For any ordering of the K categories, constructing a prediction set following Theorem 4 results in a prediction set with guaranteed finite-sample $1 - \alpha$ frequentist coverage. The choice of ordering, however, will impact prediction set precision, that is, the set’s cardinality. For inference for a single area, a natural approach is to order the categories with respect to their empirical proportions. The empirical proportions are unbiased for population proportions, so, if the area has a large sample size, an ordering based on the empirical proportions will approximate the oracle ordering well. It turns out this approach is well-motivated by classical prediction approaches. Specifically, a standard direct prediction method constructs a prediction set separately for an area based on an area-specific conditional pivotal quantity (Faulkenberry, 1973; Tian et al., 2022). For a multinomial population, $\mathbf{Y}|\mathbf{X} + \mathbf{Y}$ is such a quantity that follows a multivariate hypergeometric distribution which does not depend on the event probability vector. See Thatcher (1964) for work on prediction sets of this type for binomial data. A prediction set constructed to contain species belonging to a highest mass region of this pivotal distribution is obtained by including species with the largest empirical counts until their cumulative proportion sum exceeds $1 - \alpha$,

$$A_{\alpha}^D(\mathbf{x}) = \left\{ \mathbf{y}^{(k)} \in \mathcal{Y} : \left[\sum_{l=1}^K \mathbb{1} \left(\binom{x_k + y_k^{(k)}}{x_l + y_l^{(k)}} \geq \binom{x_l + y_l^{(k)}}{x_l + y_l^{(k)}} \frac{x_l + y_l^{(k)}}{N + 1} \right) \right] > \alpha \right\}. \quad (3.5)$$

This direct prediction set based on an ordering of the empirical proportions is appealing as it is easy to interpret and has finite-sample guaranteed $1 - \alpha$ frequentist coverage. For an area with low sampling effort, though, the empirical proportions will not precisely estimate the true proportions. As a result, a prediction set may have prohibitively large cardinality such that it is not practically useful. For such an area, incorporating indirect information from neighboring counties can improve the

estimates of the county proportions and thereby increase the precision of a prediction set.

3.2.3 Order-based prediction for multiple areas

In general, in analyzing small area data, that is, areal data featuring small within-area sample sizes in some areas, it is common to utilize indirect methods that share information across areas (Rao and Molina, 2015). The eBird database is a rich data source, and inference in any given county may be improved upon by taking advantage of auxiliary data using an indirect method. In this subsection, we detail how information from neighboring counties may be used in estimating an ordering of categories to improve prediction set precision.

As opposed to a direct prediction set based on an ordering corresponding to within-county empirical proportions, an indirect prediction set can be constructed similarly whereby species are admitted into the prediction set based on an ordering corresponding to empirical posterior proportions estimated from a hierarchical model. Such an estimate may be obtained based on a conjugate Dirichlet prior distribution parameterized with a common concentration hyperparameter for the J areas,

$$\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_J \sim \text{Dirichlet}_K(\boldsymbol{\gamma}). \quad (3.6)$$

Given a hyperparameter $\boldsymbol{\gamma} \in \mathbb{R}^K$, the posterior expectation of the proportions $\boldsymbol{\theta}_j$ in county j is $\tilde{\boldsymbol{x}}_j / (N_j + \sum_{i=1}^K \gamma_i)$ where $\tilde{\boldsymbol{x}}_j = \boldsymbol{x}_j + \boldsymbol{\gamma}$. In this way, $\tilde{\boldsymbol{x}}_j$ may be interpreted as a posterior vector of counts for county j . Then, an α -valid prediction set based on $\tilde{\boldsymbol{x}}_j$ is,

$$A_\alpha^I(\boldsymbol{x}_j) = \left\{ \boldsymbol{y}^{(k)} \in \mathcal{Y} : \left[\sum_{l=1}^K \mathbb{1} \left(\left(\tilde{x}_{j,k} + y_k^{(k)} \right) \geq \left(\tilde{x}_{j,l} + y_l^{(k)} \right) \right) \frac{x_{j,l} + y_l^{(k)}}{N_j + 1} \right] > \alpha \right\}. \quad (3.7)$$

By Theorem 4, $A_\alpha^I(\mathbf{x}_j)$ is an α -valid procedure, and it is constructed based on prior information. Specifically, it differs from the direct set given in Equation 3.5 in that categories are admitted into the prediction set based on an ordering determined by posterior counts that incorporate indirect information $\boldsymbol{\gamma}$, as opposed to an ordering based on the observed sample. Moreover, it has been shown that if the indirect information used is accurate, $A_\alpha^I(\mathbf{x}_j)$ may be more precise than a direct prediction set with the same coverage rate (Hoff, 2023; Bersson and Hoff, 2022).

In total, $A_\alpha^D(\mathbf{x}_j)$ and $A_\alpha^I(\mathbf{x}_j)$ are both α -valid prediction procedures. They differ in the order in which species are admitted into the prediction sets, as species are admitted into the direct set in terms of decreasing empirical proportions and into the indirect set in terms of decreasing posterior counts. As a result, for an area with a small sample size, incorporating accurate prior information can result in an ordering used to construct a prediction set that more accurately approximates the oracle ordering as the empirical proportions might be too unstable. Of note, these two approaches are equivalent for a uniform prior $\boldsymbol{\gamma} = c\mathbf{1}$, for any constant c . This includes, for example, a standard noninformative prior $c = 1$, a standard objective Bayes Jeffrey’s prior $c = 1/2$, and an improper prior $c = 0$.

3.2.4 Empirical Bayes estimation of indirect information

To obtain an α -valid indirect prediction set for county $j \in \{1, \dots, J\}$, all that is required is an estimate of the prior concentration parameter $\boldsymbol{\gamma}$. We propose an empirical Bayesian approach whereby values of $\boldsymbol{\gamma}$ to be used for county j are estimated from data collected in neighboring counties. Specifically, we use the maximum likelihood estimate of the marginal likelihood based on the conjugate hierarchical model

given by Equations 3.1 and 3.6,

$$\begin{aligned} \gamma_j &= \arg \max_{\gamma} \log p \left(\bigcup_{l \in L} \mathbf{X}_l \mid \gamma \right) \\ &= \arg \max_{\gamma} \log \prod_{l \in L} \left[\frac{\Gamma(\sum_{i=1}^K \gamma_i)}{\Gamma(\sum_{i=1}^K x_{l,i} + \gamma_i)} \times \prod_{i=1}^K \frac{\Gamma(x_{l,i} + \gamma_i)}{\Gamma(\gamma_i)} \right], \end{aligned} \quad (3.8)$$

where $L_j \subseteq \{1, \dots, K\} \setminus \{j\}$ is a non-empty set containing the indices of counties neighboring county j . Information is shared across neighboring counties to inform an estimate of the prior for county j , and, when estimated in this way, the prior concentration represents an across-county pooled prior concentration. This optimization problem can be solved numerically with a Newton-Raphson algorithm. See Appendix B.1 for details and derivation of such an algorithm. Code to implement this procedure in the R Statistical Programming language is available online, see Section 3.5.

When γ_j is estimated using data independent of area j and used to construct $A_{\alpha}^I(\mathbf{x}_j)$, the finite sample coverage guarantee of $A_{\alpha}^I(\mathbf{x}_j)$ holds regardless of the accuracy of the estimated prior hyperparameter. If the estimated vector γ_j is accurate, then $A_{\alpha}^I(\mathbf{x}_j)$ may also be more precise than direct prediction approaches.

3.3 Simulation Study

To illustrate how the incorporation of indirect information can affect precision of prediction sets, we compare expected set cardinality obtained from the indirect and direct prediction methods for a single simulated area. In contrast to the eBird data, for example, the analysis of this section corresponds to that of one county. Because citizen science data such as these often feature unequal sampling efforts across counties, we are particularly interested in demonstrating the difference in cardinality between these two approaches for a range of sample sizes $N = 10, 100, 1000$.

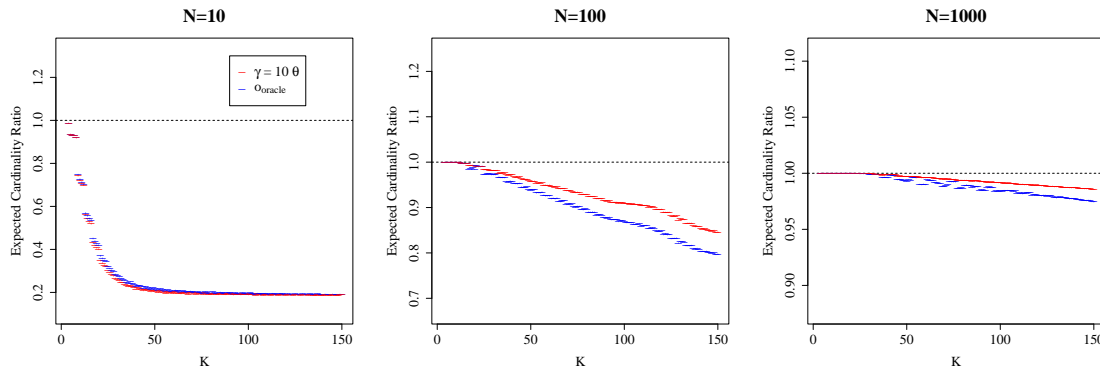


FIGURE 3.1: Monte Carlo approximations (± 1 standard deviation) of the expected cardinality ratio of (red) indirect to direct methods and (blue) α -valid prediction set given the oracle ordering to direct method.

Moreover, we compare results for varying number of categories K . Throughout, we consider a low entropy regime in which $\lceil K/4 \rceil$ categories unequally split nearly all of the probability mass, and the rest of the categories have nearly probability 0. While we do not necessarily expect real populations in practice to have such a distribution, it is chosen to clearly demonstrate the benefit of including indirect information in the construction of prediction sets that maintain frequentist coverage.

In one construction of indirect prediction sets, we consider a prior based on full information with moderate prior precision $\gamma = \theta \times 10$. We compare with direct prediction sets given by Equation 3.5, or, equivalently, indirect prediction sets constructed with a uniform prior $\gamma = c\mathbf{1}$. Finally, we compare the approaches to α -valid order-based prediction sets obtained based on an oracle ordering. Results comparing Monte Carlo approximations of the expected prediction set cardinality ratios between the various approaches obtained from 25,000 replications are displayed in Figure 3.1.

As all methods considered are α -valid procedures, the crucial difference between them is the incorporation of indirect information. Utilizing accurate prior information in the construction of prediction sets generally results in prediction sets distinctly smaller than direct sets, particularly so if there are a large number of categories rel-

ative to the sample size. This is evidenced by the red dashes in Figure 3.1 showing the expected cardinality ratios of the indirect to direct prediction sets are always at or below a value of 1. An accurate prior may be one that approximates the true probability mass vector well with large precision relative to sample size, as seen in the left plot of Figure 3.1 for sample size $N = 10$. More generally, though, all that is needed is a prior that results in posterior counts that accurately approximate the oracle ordering of categories. We discuss the three sample size regimes in detail below.

For a small sample size of $N = 10$, the prior γ used to construct the indirect prediction sets is an informative prior with strong precision in that the scale used is equal to the sample size in this case. As a result, the posterior distributions contain notably more information than what is in each simulated dataset. As a result, the ordering of categories induced by the posterior counts, used to construct the indirect prediction sets, are accurately approximating the oracle ordering of categories. This is evidenced by the nearly identical behavior of the two cardinality ratios explored. In conjunction with the instability of the direct method in the presence of such a small sample size, this results in notably smaller cardinality of the indirect set as compared to the direct set, even for relatively small total numbers of categories. At its best, the indirect prediction set is about 80% smaller than the direct set.

For a moderate sample size of $N = 100$, the prior precision used to construct the indirect prediction sets is not overwhelming as compared to the sample size, and hence the posterior counts do not approximate the oracle ordering as well as in the regime with a smaller sample size. This is evidenced by the divergence of the red and blue dashes in the middle plot of Figure 3.1. Still, particularly as the number of categories increases for fixed N , the benefit of utilizing prior information of this type is highlighted by the decline of the cardinality ratio of the indirect to direct prediction sets (red lines). For example, in the case of $N = 100$ and $K = 150$,

the indirect prediction set constructed with γ is about 15% smaller than the direct prediction set.

A similar but less pronounced pattern is seen in the presence of a larger sample size of $N = 1000$. For this sample size with $K \leq 150$, all methods considered perform relatively similarly. However, as the number of categories increases, there is a distinct gain in prediction set precision given the input of indirect information in prediction set construction.

3.4 Summarising eBird species abundance data

In this section, we describe avian species abundance in North Carolina, USA from eBird data obtained from citizen-uploaded complete checklists of species observations in the first week of May 2023. Across the 99 counties, 393 unique species were identified. Some species such as the Northern Cardinal, Carolina Wren, and American Robin were identified frequently. Many others like the Northern Saw-whet Owl and the Solitary Sandpiper were rarely seen; in fact, 50% of species were seen fewer than 100 times each across the entire state. Moreover, within-county sample sizes vary drastically (Figure 3.2) from approximately 50,000 individual birds identified in Wake County, one of the most populous counties in NC that contains the state’s capital, to only 8 in Pasquotank County, a small coastal county consisting of about 1/30th of the human population of Wake County.

As motivated in the Introduction, describing such data with α -valid prediction sets for each county provides a useful summary with unambiguous statistical interpretation. That is, with at least probability $1 - \alpha$, an unobserved bird in a given county will belong to a species contained in the specified prediction set, where the probability is taken with respect to the random sample and the predictand. Here, we demonstrate the usefulness of this approach in gaining better understanding of species abundance. Moreover, we elaborate on the benefit of utilizing indirect infor-

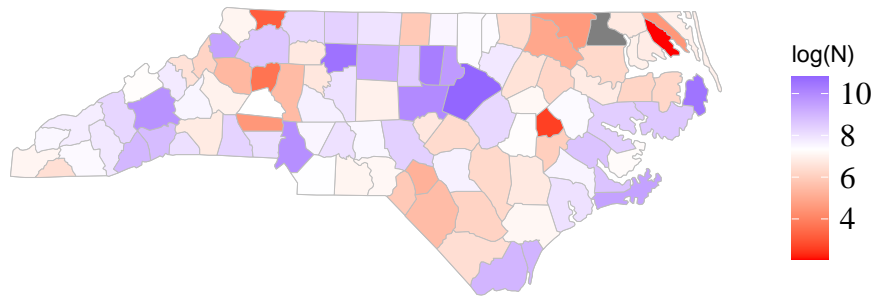


FIGURE 3.2: Within-county log sample size of eBird data in North Carolina. Sample sizes range from 8 to approximately 50,000.

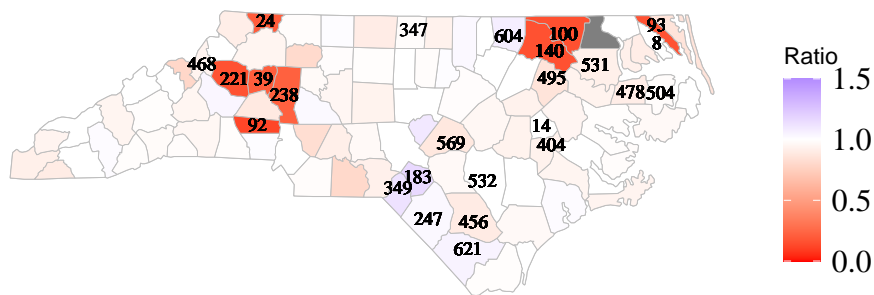


FIGURE 3.3: Cardinality ratio of indirect to direct prediction sets. Prior hyperparameters estimated with an empirical Bayesian procedure based on five nearest neighbors for each county. The lowest quantile sample sizes are overlaid on their respective counties.

mation in the construction of practically useful sets that are precise, particularly for counties with small within-county sample sizes.

For each county in NC, we construct an indirect prediction set based on a prior hyperparameter estimated from data in the five nearest neighboring counties, following the procedure described in Section 3.2.4. The eBird data consist of independent samples collected across the state, so samples are independent across counties. As

a result of this independence, finite-sample coverage of the indirect prediction approach is guaranteed. We compare the cardinality of these indirect prediction sets to that of direct prediction sets, both of which maintain at least 95% coverage for each county. The cardinality ratios of the indirect to direct prediction sets across the counties in NC are plotted in Figure 3.3. To highlight the impact of within-county sample size, the lower quantile sample sizes are overlaid on their respective county.

In general, the incorporation of indirect information in the construction of prediction sets results in notably smaller cardinality of the indirect prediction sets as compared with that of the direct prediction sets. Of the 99 counties in NC, indirect sets have smaller cardinality in 65, and the two approaches result in the same cardinality in 20 counties. The improvement in cardinality is particularly conspicuous in counties with small to moderate sample sizes, as evidenced by the sample sizes of counties with the brightest shade of red in Figure 3.3. Moreover, ten counties have trivial direct sets consisting of all K species, while only two counties with the smallest within-county sample sizes, 8 and 14, have trivial indirect prediction sets. For the county with the third smallest sample size (24), the indirect prediction set only includes 80 species, or about 20% of all possible species, while the direct prediction set is the trivial set.

Overall, even in counties with larger sample sizes, it is most common for the indirect and direct prediction sets to contain a different set of species. In fact, the indirect and direct prediction sets disagree for nearly every county in NC. They are equivalent for only six counties where they aren't both trivial sets. Commonly, this discrepancy corresponds with smaller indirect sets, and hence highlights the benefit of inclusion of indirect information in the construction of prediction sets.

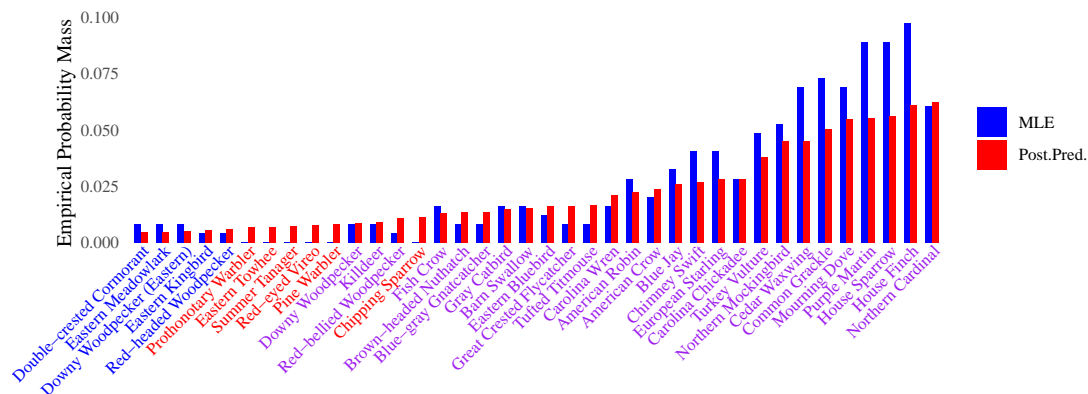


FIGURE 3.4: Empirical probability masses of species included in only the indirect (red text), only the direct (blue text), or both (purple text) prediction sets, sorted by the posterior proportion. MLE plotted in blue, and the posterior proportion based on a prior estimated from data in nearest five neighboring counties.

Table 3.1: Percentage of all birds observed within each respective county, for select species included in either the indirect set (red text) or the direct set (blue text). Estimated prior hyperparameter γ for Robeson County recorded in the last row.

	D. Cormorant	E. Kingbird	Pine Warbler	C. Sparrow
Robeson	0.81%	0.4%	0.00%	0.00%
NC-017	0.00%	2.85%	1.97%	2.63%
NC-047	0.00%	0.00%	1.29%	0.64%
NC-051	0.00%	0.68%	2.62%	5.59%
NC-093	0.00%	0.00%	1.09%	0.55%
NC-165	0.00%	0.86%	2.58%	5.16%
γ	0.00	1.33	3.42	4.69

3.4.1 Order-based prediction in Robeson County

To further compare the two approaches and elucidate the role of the ordering of the species, we elaborate on the construction of indirect and direct prediction sets for Robeson County. Robeson is located near the southeastern border of NC and features a moderately small within-county sample size of 247 birds observed, with species-specific observation counts ranging from zero to ten. The two prediction sets have nearly the same cardinality but contain differences in species inclusion. Specifically, the indirect prediction set contains 33 species, and the direct set contains 32, with

an overlap of 27 species.

To illustrate the role of the ordering used in the construction of α -valid prediction sets, the empirical proportions based on the observed sample (MLE) and posterior proportions (Post.Pred) are plotted in Figure 3.4 for the union of included species in the two sets. In the figure, the species are sorted by increasing posterior proportions. The indirect and direct sets include species based on the posterior and empirical distributions, respectively. Discrepancies between the indirect and direct sets occur when these two distributions disagree. From Figure 3.4, it is easy to see the indirect prediction set consists of the species with the 33 largest posterior predictive proportions. In contrast, the direct set consists of species with the largest sample probability mass. Naturally, the ordering of these two estimates agree for species common to the region, and, as such, there is a fair amount of overlap of species inclusion.

As a result of our estimation procedure for the prior hyperparameter γ for Robeson County, the disparity between inclusion or exclusion of a species among the two prediction set methods is further elucidated by examining species presence in neighboring counties. In short, species with more frequent occurrence in neighboring counties will have a larger estimated prior count than those seen rarely in neighboring counties. Species occurrences in neighboring counties are displayed in Table 3.1 for a select few species along with the estimated γ for Robeson County, obtained by solving Equation 3.8 using data in these neighboring counties.

Intuitively, species that are seen in neighboring counties with some relative frequency, such as the Chipping Sparrow or Pine Warbler, are probably also present in Robeson County, and hence should be included in a prediction set. In practice, these species have a comparatively high estimated prior of about 5 and 4, respectively, and hence are included in the indirect prediction set even though they weren't recorded as being observed in Robeson County in the dataset. Alternatively, consideration of

Table 3.2: Percentage of all birds observed within each respective county for species included in either both prediction sets (purple text) or only the direct set (blue text). Estimated prior hyperparameter γ for Haywood County recorded in the last row. Species are sorted by posterior proportion.

	L. Flycatcher	R. Hawk	C. Yellowthroat	E. Kingbird	Bobolink
Haywood	0.21%	0.24%	0.21%	0.24%	0.24%
NC-021	0.06%	0.29%	0.18%	0.43%	0.14%
NC-099	0.00%	0.08%	0.16%	0.00%	0.00%
NC-115	0.07%	0.14%	0.21%	0.28%	0.00%
NC-173	0.18%	0.18%	0.66%	0.09%	1.41%
NC-175	0.00%	0.30%	1.00%	0.54%	0.84%
γ	0.4	1.29	2.43	1.49	2.15

indirect information yields the conclusion that species like the Eastern Kingbird and Cormorant may be rare in the area in general, as reflected by small γ values, and thus these species are not included in the indirect prediction set.

3.4.2 Inference among species with tied counts in Haywood County

In species abundance data, particularly for areas or counties with small sample sizes, it is common for multiple species to have the same observed count. A feature of the construction of the direct order-based prediction approach as presented is that species with the same observed counts will either be jointly included or excluded from the prediction set. As a result, a direct prediction set constructed from a sample with tied species counts may have increased cardinality over an indirect prediction set that does not necessarily jointly admit all species with tied observed counts. If the direct set has increased cardinality for this reason, the direct set will also have increased coverage over the indirect set.

When constructing a prediction set based on the empirical proportions without consideration of indirect information, as in the construction of the direct set, this may commonly occur, and there is no clear approach to choose among the species with tied counts without further information than what is provided in the sample in

that county. One could randomly choose to include one of the species from the set of species with tied counts, for example, but a more principled manner is to utilize indirect information to determine which species should be included. This is the mechanism used by the indirect prediction approach when the prior hyperparameter is a real valued vector estimated from indirect information. As such, a more nuanced benefit of utilizing indirect information in the construction of a prediction set is the capacity to include a select few categories with tied empirical proportions.

To demonstrate, we elaborate on species inclusion in the indirect and direct prediction sets in Haywood County. Haywood is popular destination in the Blue Ridge Mountains, located near the western border of North Carolina. It features a moderately large within-county sample size of roughly 4000 birds observed. In Haywood County, the indirect prediction set contains 70 species, and the larger direct set contains 74. In the construction of these prediction sets, the ordering of species with regards to the posterior proportions and the empirical proportions agree for most species. As a result, all 70 species included in the indirect set are also included in the direct set. The disparity in species inclusion occurs primarily as a result of tied counts of species occurrence in the sample.

Empirical proportions in Haywood and neighboring counties are reported in Table 3.2 for the five species included in Haywood County’s prediction sets with the smallest posterior proportions. The species with the four smallest posterior proportions are included only in the direct set, and the other species, the Bobolink, is included in both the indirect and direct sets. The Bobolink was observed 9 times in the sample from Haywood County, or about 0.24% of the Haywood sample. For an ordering determined by either the empirical counts or the posterior counts, this species is required to be included in the order-based prediction set to guarantee $1 - \alpha$ coverage. Two of the other species, the Red-shouldered Hawk and Eastern Kingbird, were each also observed 9 times in the sample from Haywood, and, by construction of the order-

based prediction approach, must also be included in the direct set. When admitting the species into a prediction set by posterior counts based on the real-valued prior hyperparameter γ estimated from data in neighboring counties, as in the indirect approach considered, the ‘tie’ among these three species is broken, and only one, the Bobolink, is included in the indirect prediction set.

3.5 Discussion

Species abundance data collected across heterogeneous areas is increasingly important in understanding biodiversity. Some of the largest sources of such data are citizen science databases for which volunteers spearhead the data collection. As a result of the civilian-led scientific effort, such data often feature unequal sampling across a spatial domain where some areas have large within-area sample sizes and others have much smaller within-area sample sizes.

In this chapter, we propose summarising species abundance data of this type with valid prediction sets that are constructed by sharing information across areas. Utilizing indirect information may result in smaller prediction sets than otherwise achievable with direct methods. Meanwhile, maintaining validity of the prediction sets for each area allows for an accessible interpretation that enables a straightforward comparison across areas. In particular, maintaining interpretable statistical guarantees on a descriptor of such data is important as analyses from such data often have far reaching policy implications. Smaller prediction sets may be attainable based on Bayesian inference of a spatial hierarchical model such as that presented in Tang et al. (2023), for example, but these approaches introduce bias and a resulting prediction set would not retain the nominal frequentist coverage rate guarantee for each county.

The usefulness of our approach for summarising citizen science data is motivated in part to combat the common problem of varying sampling efforts across areas. We

detail how α -valid prediction sets can be constructed with the incorporation of indirect information to improve within-county prediction set precision and propose an empirical Bayes procedure to do so. Incorporation of accurate indirect information results in a narrower prediction set for a given county than a direct prediction set by exploiting data in nearest neighboring counties. The proposed empirical Bayes procedure is based on a standard hierarchical model that is straightforward to understand, and the authors provide code for implementation.

There may, however, be a benefit to utilizing a more structured prior that incorporates indirect information in a more complex manner such as a prior that weights data from different parts of the state differently. For example, a model based on a learned intrinsic distance between counties was shown in Christensen and Hoff (2022) to fit a subset of the eBird data better than standard methods based on geographic adjacency structure. In the sample analyzed in Section 3.4, we found an indirect prediction set constructed with a hyperparameter estimated from five nearest neighbors results in overall narrower prediction sets than a direct approach, but it would be valuable to explore if this can be further improved upon with a more detailed prior. More broadly, different applications may warrant an alternative information sharing prior if, for example, there is no notion of spatial distance across the different areas. For example, it may be of interest to compare species abundance variation across different time frames for a given county.

Covariance Estimation for Multi-group, Matrix-variate Data

4.1 Introduction

Matrix-variate datasets consisting of a sample of n matrices Y_1, \dots, Y_n each with dimensionality $p_1 \times p_2$ are increasingly common in modern applications. Examples of such datasets include repeated measurements of a multivariate response, two-dimensional images, and spatio-temporal observations. Often, a matrix-variate dataset may be partitioned into several distinct groups or subpopulations for which group-level inference is of particular interest. For example, a multi-group dataset may be subdivided by socio-economic population or geographic region.

In analyzing multi-group matrix-variate data of this type, describing heterogeneity across groups is often of particular interest. For instance, in remote-sensing studies, scientists may be interested in understanding variation across classes of land cover from repeated measurements of spectral information, as in Johnson et al. (2012). The information collected at each site may be represented as a $p_1 \times p_2$ matrix where the rows represent p_1 wavelengths and the columns represent p_2 dates when the images

were taken. Accurate multi-group covariance estimation is necessary for such a task. Moreover, in many such applications, the population covariance of matrix-variate data may be near separable in that the covariances across the p_2 columns are near each other or the covariances across the p_1 rows are near each other. Incorporation of this structural information in an estimation procedure can improve estimation accuracy. More generally, accurate covariance estimation of multi-group matrix-variate data is a pertinent task in many statistical methodologies including classification, principal component analysis, and multivariate regression analysis, among others. For example, classification of a new observation based on a labeled training dataset with quadratic discriminant analysis (QDA) requires group-level estimates of population means and covariance matrices. As a result, adequate performance of the classification relies on, among other things, accurate group-level covariance estimates.

One approach to analyzing multi-group matrix-variate data is to vectorize each data matrix and utilize methods designed for generic multivariate data separately for each group. In this way, direct covariance estimates which only make use of within-group samples may be obtained from matrix-variate data by vectorizing each observation and computing the standard sample covariance matrices. While a group's sample covariance may be unbiased, the estimate may have large variance unless the group-specific sample size n_j is appreciably larger than the dimension $p = p_1 p_2$. This is a limiting requirement as modern datasets often consist of many features, that is, often $p \approx n_j$, or even $p \gg n_j$. As a result, more accurate covariance estimates may be obtained via indirect or model-based methods which incorporate auxiliary information. Such methods may introduce bias, but can correspond to estimates with lower variance than unbiased methods.

To improve the accuracy of covariance estimates for multi-group data, researchers may estimate each group's covariance with the pooled sample covariance matrix. This implicitly imposes an assumption of homogeneity of covariances across the pop-

ulations and greatly reduces the number of unknown parameters to be estimated. Such an estimator may be biased, but can have lower error than the population-specific sample covariance. For example, linear discriminant analysis (LDA), which assumes homoscedasticity across groups, has been shown to outperform QDA when sample sizes are small, even in the presence of heterogeneous population covariances (see, for example, Marks and Dunn (1974)). A more robust approach estimates each group’s covariance as a weighted sum of a pooled estimate and the group-specific sample covariance. Such an approach is often referred to as partial-pooling, or, in the Bayesian framework, hierarchical modeling. For a nice introduction to partial pooling and an empirical Bayesian implementation, see Greene and Rayens (1989). More work on the topic is found in Friedman (1989); Rayens and Greene (1991); Brown et al. (1999). Relatedly, there has been work which assumes pooled elements of common covariance decompositions (e.g. pooled eigenvectors across groups (Flury, 1987)) and proposals to shrink elements of decompositions to pooled values (Daniels, 2006; Hoff, 2009b).

Alternatively, as opposed to pooling information across groups, accuracy may be improved by imposing structural assumptions on the covariances separately for each group. Some common structural assumptions include diagonality (Daniels and Kass, 1999), bandability (Wu and Pourahmadi, 2009), and sparsity (Friedman et al., 2008), among others. For matrix-variate data, a separable or Kronecker structure covariance assumption (Dawid, 1981) may be more appropriate. A Kronecker structured covariance represents each $p \times p$ population covariance as the Kronecker product of two smaller covariance matrices of dimension $p_1 \times p_1$ and $p_2 \times p_2$ which respectively represent the across-row and across-column covariances. Again, while a separable covariance estimator may be biased, it may have smaller error than an unstructured covariance estimator. In practice, however, the population covariance may not be well represented by a Kronecker structure. To allow for robustness to misspecifica-

tion, a researcher may proceed in a Bayesian manner and adaptively shrink to the Kronecker structure as in Hoff et al. (2022). Such an estimator can be consistent, but may have stability issues similar to that of an unstructured covariance matrix if the population covariance is not well represented by a Kronecker structure. In this instance, more accurate multi-group covariance estimates may be obtained by pooling across groups rather than shrinking within each group to a separable structure.

More generally, in covariance estimation based on matrix-variate data from multiple populations, it is rarely obvious whether shrinking each unstructured covariance separately towards a Kronecker structure or shrinking all unstructured covariances towards an unstructured pooled covariance will result in more accurate estimates. This is particularly difficult in the presence of small within-group sample sizes as popular classical statistical tests for both homogeneity of covariances (Box, 1949) and accuracy of a Kronecker structure assumption (Lu and Zimmerman, 2005) rely on approximations that require large sample sizes to achieve the desired precision.

To account for this uncertainty, we propose a hierarchical model that adaptively allows for both types of shrinkage. Specifically, in this chapter, we provide a model-based multi-group covariance estimation method for matrix-variate data that improves the overall accuracy of direct covariance estimates. We propose a hierarchical model for unstructured group-level covariances that adaptively shrinks each estimate either within-population towards a separable Kronecker structure, across-populations towards a shared pooled covariance, or towards a weighted additive combination of the two. The model features flexibility in the amount of each type of shrinkage. Furthermore, the proposed model has a latent-variable representation that results in straightforward Bayesian inference via a Metropolis-Hastings algorithm. The proposed model provides robustness to mis-specification of structural assumptions and improved stability if assumptions are wrong while maintaining coherence and interpretability.

This chapter proceeds as follows. In Section 4.2 we motivate our method and detail the proposed hierarchical model. We describe a Bayesian estimation algorithm in Section 4.3 and demonstrate properties of the proposed method via a simulation study in Section 4.4. The flexibility of the proposed method is shown in two examples in Section 4.5. In the first example, we demonstrate the usefulness of inference under the proposed model in speech recognition. In the second example, we perform inference on a chemical exposure data set where understanding heterogeneity across socio-demographic groups is of key interest. We conclude with a discussion in Section 4.6.

4.2 Methodology

In this section we introduce the *Shrinkage Within and Across Groups* (SWAG) covariance model, a hierarchical model developed for simultaneous covariance estimation for multi-group, matrix-variate data. We are particularly motivated by improving the overall accuracy of group-specific estimates of population covariances when the true covariance structures are unknown and group-specific sample sizes are small relative to the number of features. The proposed model adaptively allows for flexible shrinkage either across groups, within a group to a Kronecker structure, or an additive combination of the two. The SWAG model is constructed from semi-conjugate priors to allow for straightforward Bayesian estimation and interpretable parameters. The section proceeds by introducing motivation in Sections 4.2.1 and 4.2.2, presenting the proposed hierarchical covariance model in Section 4.2.3, and elaborating on parameter interpretation in Section 4.2.4.

4.2.1 *Partial-pooling shrinkage for multi-group data*

As detailed in the Introduction, a common method used to improve a population's covariance estimate is linear shrinkage from the population's sample covariance ma-

trix towards some covariance term which can be estimated with greater precision. One such method for multi-group multivariate data is partial pooling, as detailed in Greene and Rayens (1989, GR). In particular, for population $j \in \{1, \dots, J\}$, let $y_{1,j}, \dots, y_{n_j,j}$ be an i.i.d. random sample of p -dimensional vectors from a mean-zero normal population with unknown non-singular covariance matrix $\Sigma_j \in \mathcal{S}_p^+$,

$$y_{1,j}, \dots, y_{n_j,j} \sim N_p(0, \Sigma_j), \quad \text{independently for } i = 1, \dots, n_j, j = 1, \dots, J.$$

GR use mutually independent inverse-Wishart priors for each population covariance, $\Sigma_j^{-1} \sim W_p(\Psi_0^{-1}/(\nu-p-1), \nu)$, for $j \in \{1, \dots, J\}$, parameterized such that $E[\Sigma_j | \Psi_0, \nu] = \Psi_0$. The Bayes estimator for the covariance of population j under squared error loss partially-pools each group's sample covariance,

$$\hat{\Sigma}_j := E[\Sigma_j | \mathbf{y}, \Psi_0, \nu] = w_1 S_j + (1 - w_1) \Psi_0, \quad (4.1)$$

where $w_1 = n_j/(n_j + \nu - p - 1)$ and $S_j = \sum_{i=1}^{n_j} y_{i,j} y_{i,j}^T / n_j$ is the sample covariance matrix for population j . GR use plug-in estimates for the pooled covariance and degrees of freedom $\{\Psi_0, \nu\}$ which are obtained in an empirical Bayesian manner.

This partially-pooled estimator linearly shrinks a population's sample covariance matrix towards a pooled covariance by a weight w_1 that depends on the degrees of freedom and the group-specific sample size n_j . In this way, the degrees of freedom parameter determines the amount of shrinkage towards the pooled covariance. In particular, if the degrees of freedom ν is large relative to the sample size n_j , the covariance estimate is strongly shrunk towards the pooled value. In populations where the group-specific sample size is large, or if the degrees of freedom estimate is comparatively small, more weight is placed on the sample covariance matrix.

4.2.2 Kronecker shrinkage for matrix-variate data

For a matrix-variate population, the accuracy of a covariance estimate may be improved via linear shrinkage towards a population-specific Kronecker structure. Let

Y_1, \dots, Y_n be an i.i.d. sample of random matrices, each of dimension $p_1 \times p_2$, from a mean-zero normal population with non-singular covariance matrix $\Sigma \in \mathcal{S}_p^+$ where $p = p_1 p_2$,

$$Y_1, \dots, Y_n \sim N_{p_1 \times p_2}(0, \Sigma), \quad \text{independently for } i = 1, \dots, n.$$

Even if p_1 and p_2 are each relatively small, obtaining a statistically stable estimate of the unstructured p -dimensional covariance may require a prohibitively large sample size. As a result, shrinkage towards a parsimonious Kronecker structured covariance $C \otimes R$ may be used, where “ \otimes ” is the Kronecker product, $R \in \mathcal{S}_{p_1}^+$ is a “row” covariance matrix, and $C \in \mathcal{S}_{p_2}^+$ is a “column” covariance matrix. A linear shrinkage estimator that shrinks a population’s sample covariance towards a population-specific Kronecker separable covariance may be obtained from the following prior,

$$\Sigma^{-1} \sim W_p((C \otimes R)^{-1}/(\gamma - p - 1), \gamma),$$

parameterized so that $E[\Sigma_j | C, R, \gamma] = C \otimes R$. Here, the Bayes estimator of the covariance Σ under squared error loss is

$$\hat{\Sigma} := E[\Sigma | \mathbf{Y}, C, R, \gamma] = w_2 S + (1 - w_2)(C \otimes R), \quad (4.2)$$

where $w_2 = n/(n + \gamma - p - 1)$. An empirical Bayesian estimation approach based on shrinkage towards a Kronecker structure is presented in Hoff et al. (2022). In context to that chapter, here, we will take a fully Bayesian approach.

The estimator given in Equation 4.2 linearly shrinks the unstructured sample covariance towards a Kronecker structured covariance by the weight w_2 that depends on the sample size and the estimated degrees of freedom. As with the partially-pooled estimator, this estimator is strongly shrunk towards the Kronecker structure when the degrees of freedom is large relative to sample size. If the degrees of freedom is small, or the sample size is large, more weight is placed on the sample covariance.

4.2.3 Flexible shrinkage for multi-group matrix-variate data

For each group $j = 1, \dots, J$, let $Y_{j,1}, \dots, Y_{j,n_j}$ be an i.i.d. sample of random matrices, each of dimension $p_1 \times p_2$, from a mean-zero normal population with non-singular covariance $\Sigma_j \in \mathcal{S}_p^+$, that is,

$$Y_{j,1}, \dots, Y_{j,n_j} \sim N_{p_1 \times p_2}(0, \Sigma_j), \quad \text{independently for } j = 1, \dots, J. \quad (4.3)$$

As it is often unclear which of the approaches presented is most appropriate in the presence of multi-group matrix-variate data of this type, we propose an approach that combines the two methods of covariance shrinkage discussed. In particular, we propose the *Shrinkage Within and Across Groups* (SWAG) hierarchical prior distribution which linearly combines an estimate shrunk towards a pooled covariance Ψ_0 and an estimate shrunk towards a group-specific Kronecker structure ($C_j \otimes R_j$) by a weight $\lambda \in (0, 1)$ that is estimated from the data.

Specifically, the SWAG prior utilizes an over-parameterized representation of each group's covariance. That is, for population $j \in \{1, \dots, J\}$,

$$\Sigma_j = \lambda \Psi_j + (1 - \lambda) \Lambda_j, \quad (4.4)$$

where each Ψ_j is shrunk towards a common covariance, and each Λ_j is shrunk towards a group specific Kronecker covariance. Each covariance Ψ_j is shrunk across groups towards a common covariance matrix using the prior distribution

$$\Psi_j^{-1} \sim W_p(\Psi_0^{-1}/(\nu - p - 1), \nu), \quad \text{independently for } j = 1, \dots, J, \quad (4.5)$$

parameterized such that $E[\Psi_j | \Psi_0, \nu] = \Psi_0$. When Ψ_0 is estimated from data across all groups, this term is interpreted as a pooled covariance matrix. As we are interested in obtaining a covariance matrix estimate based on matrix-variate data, each Λ_j term is shrunk towards a group-specific Kronecker structured covariance,

$$\Lambda_j^{-1} \sim W_p((C_j \otimes R_j)^{-1}/(\gamma - p - 1), \gamma), \quad \text{independently for } j = 1, \dots, J, \quad (4.6)$$

where $E[\Lambda_j | R_j, C_j, \gamma] = (C_j \otimes R_j)$. Here, as before, R_j is a $p_1 \times p_1$ row covariance matrix from population j and C_j is the corresponding $p_2 \times p_2$ column covariance matrix. Furthermore, to more clearly separate these two notions of shrinkage (within population or across populations), a Wishart prior on the across-population covariance Ψ_0 allows for flexible shrinkage of this pooled term towards an across-group Kronecker covariance:

$$\Psi_0 \sim W_p((P_2 \otimes P_1)/\xi, \xi), \quad (4.7)$$

parameterized such that $E[\Psi_0 | P_1, P_2, \xi] = (P_2 \otimes P_1)$ where $P_1 \in \mathcal{S}_{p_1}^+$ and $P_2 \in \mathcal{S}_{p_2}^+$. In this way, the weight λ is interpreted as partially controlling the amount of shrinkage towards homogeneity versus towards heterogeneity of covariances across groups. A visualization of the proposed SWAG hierarchy is given in Figure 4.1. In summary, the SWAG model combines a standard hierarchical model on the across-group shrunk Ψ_j covariances with Bayesian shrinkage towards a separable structure on the within-group shrunk Λ_j covariances and the pooled covariance Ψ_0 .

We note that the SWAG model is primarily motivated by the need to obtain more accurate group-specific covariance estimates, so, while there is redundancy in this parameterization, the group-specific covariances $\Sigma_1, \dots, \Sigma_J$ are identifiable. As a result, this over-parameterization will not affect inference on estimation of group-level covariances, estimation of mean effects, imputation of missing data, or response prediction.

4.2.4 Interpretation of Parameters

In this section, we highlight properties of the proposed SWAG covariance priors under the normal sampling model given in Equation 4.3. A priori, regardless of the specified sampling model, the marginal expectation under the SWAG prior is a weighted sum of a pooled covariance and the heterogeneous Kronecker separable

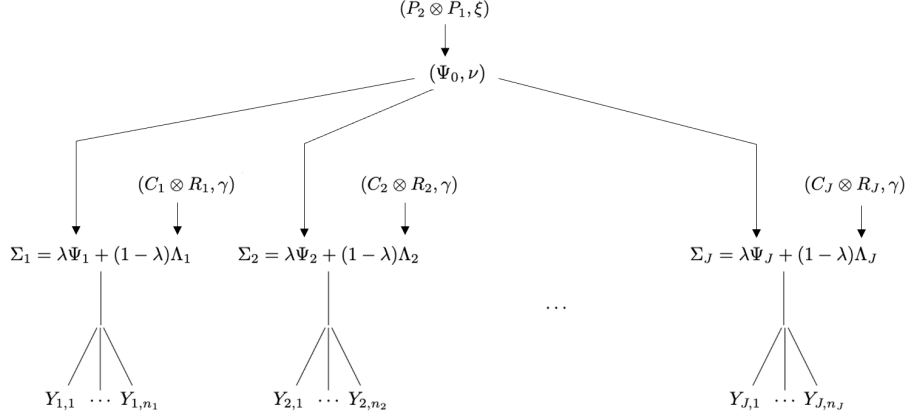


FIGURE 4.1: A graphical representation of the SWAG hierarchical model. The unstructured covariance terms Ψ_1, \dots, Ψ_J are shrunk across groups towards a shared covariance Ψ_0 . The covariance terms $\Psi_0, \Lambda_1, \dots, \Lambda_J$ are each individually shrunk towards a Kronecker covariance.

covariance:

$$E[\Sigma_j | \Psi_0, R_j, C_j, \lambda] = \lambda \Psi_0 + (1 - \lambda) (C_j \otimes R_j), \quad (4.8)$$

for $j \in \{1, \dots, J\}$, where the weight λ quantifies the prior weight on each structure. As a result, the SWAG prior presents a flexible approach to combine shrinkage across groups towards a pooled value and shrinkage within groups towards a separable structure.

The role of the weight λ is further elucidated at the extremes of its sample space. Given $\lambda = 1$, the SWAG model reduces to a Bayesian analogue of the partially-pooled estimators given in Greene and Rayens (1989), as given in Equation 4.1. At the weight's opposite extreme, when $\lambda = 0$, the SWAG model is equivalent to the Kronecker structure shrinkage model presented in Section 4.2.2 applied separately to each group. To alleviate some of the potential ambiguity in interpretation of λ , we further allow Ψ_0 to be flexibly shrunk to a Kronecker structure. In this way, $\lambda = 1$ represents hierarchical shrinkage across populations towards a pooled covariance and $\lambda = 0$ represents within-population shrinkage.

Moreover, under the SWAG prior, the prior marginal expected value of a popula-

tion’s covariance is a weighted sum of a pooled covariance and a population-specific separable covariance (Equation 4.8). Under normality, as the degrees of freedom ν and γ increase, the marginal sampling model of a random matrix $Y_{i,j}$ converges to a normal distribution with this prior expectation as the covariance. That is, when the degrees of freedom parameters ν and γ are large, the unstructured covariances Ψ_j and Λ_j are each strongly shrunk towards their respective prior expected value, and, therefore, the population covariances are approximately represented by the weighted sum of the pooled covariance and population-specific Kronecker structure.

4.3 Parameter Estimation

4.3.1 Latent Variable Representation

The SWAG model has a latent variable representation that allows for straightforward Bayesian inference. Specifically, consider the following representation of the proposed SWAG model:

$$\begin{aligned} \text{vec}(Y_{i,j}) &= \lambda^{1/2}U_{i,j} + (1 - \lambda)^{1/2}E_{i,j}, \quad \text{for } i = 1, \dots, n_j & (4.9) \\ U_{i,j} &\sim N_p(0, \Psi_j) \\ E_{i,j} &\sim N_p(0, \Lambda_j), \end{aligned}$$

independently for each population $j \in \{1, \dots, J\}$ where the priors on each Ψ_j and Λ_j are as given in Equations 4.5 and 4.6. That is, each matrix $Y_{i,j}$ is represented as a weighted sum of one factor ($U_{i,j}$) which partially pools covariances across populations and another factor ($E_{i,j}$) which flexibly shrinks the population covariance towards a group-specific Kronecker structure. Marginal with respect to the factors, this latent variable representation is equivalent to the sampling model proposed in the SWAG model,

$$\text{vec}(Y_{i,j})|\lambda, \Psi_j, \Lambda_j \sim N_p(0, \lambda\Psi_j + (1 - \lambda)\Lambda_j).$$

Conditioning on the factor $U_{i,j}$ results in closed form full conditionals of the covariance parameters of interest, as detailed in subsequent subsections.

4.3.2 Posterior Approximation

In this section, we detail a Metropolis-Hastings algorithm for parameter estimation based on the latent variable representation of the SWAG model. Label $\mathbf{Y} = \{Y_{i,j} : i \in \{1, \dots, n_j\}, j \in \{1, \dots, J\}\}$, $\mathbf{U} = \{U_{i,j} : i \in \{1, \dots, n_j\}, j \in \{1, \dots, J\}\}$, $\mathbf{\Psi} = \{\Psi_1, \dots, \Psi_J\}$, $\mathbf{\Lambda} = \{\Lambda_1, \dots, \Lambda_J\}$, $\mathbf{R} = \{R_1, \dots, R_J\}$, and $\mathbf{C} = \{C_1, \dots, C_J\}$. Then, Bayesian inference is based on the joint posterior distribution, with density

$$p(\lambda, \mathbf{\Psi}, \mathbf{\Lambda}, \Psi_0, \nu, \mathbf{R}, \mathbf{C}, \gamma, P_1, P_2, \xi | \mathbf{Y}),$$

and a Monte Carlo approximation to this posterior distribution is available via a MetropolisHastings algorithm. Based on the latent variable representation presented in Equations 4.9, nearly all of the parameters in the SWAG model maintain semi-conjugacy leading to a straightforward Metropolis-Hastings algorithm which constructs a Markov chain in

$$\boldsymbol{\theta} = \{\lambda, \mathbf{\Psi}, \mathbf{\Lambda}, \mathbf{U}, \Psi_0, \nu, \mathbf{R}, \mathbf{C}, \gamma, P_1, P_2, \xi\}.$$

Such Bayesian inference provides estimates and uncertainty quantification for arbitrary functions of the parameters. While we focus on the mean-zero case, the algorithm presented may be trivially extended to include estimation of mean parameters.

For a full Bayesian analysis, priors must be specified for all unknown parameters. For simplicity, a straightforward $beta(\alpha, \beta)$ prior may be used to describe prior expectations of behavior of λ . On the degrees of freedom parameters ν, γ , and ξ , negative binomial distributions with the appropriate support may be used, that is, $NegBin_{[p+2, \infty)}(r_0, p_0)$, parameterized by size r_0 , success probability p_0 , and lower

bound $p + 2$. Semi-conjugate priors on the remaining covariance parameters are proposed to facilitate computation,

$$R_1, \dots, R_J \sim W_{p_1} (R_0/\eta_1, \eta_1)$$

$$C_1, \dots, C_J \sim W_{p_2} (C_0/\eta_2, \eta_2)$$

$$P_1^{-1} \sim W_{p_1} (P_{01}^{-1}/(\eta_3 - p_1 - 1), \eta_3)$$

$$P_2^{-1} \sim W_{p_2} (P_{02}^{-1}/(\eta_4 - p_1 - 1), \eta_4).$$

A discussion of hyperparameter specification is provided in Section 4.3.3.

A Metropolis-Hastings sampler proceeds by iteratively generating new sets of model parameters based on their full conditional distributions. When iterated until convergence, this procedure will generate a Markov chain that approximates the joint posterior distribution $p(\boldsymbol{\theta}|\mathbf{Y})$. The sampling steps are now detailed.

Sampling of population model parameters

The full conditionals of sampling model covariances and factors, as well as the key shrinkage controlling parameters λ, ν, γ are discussed in this subsection. To reduce dependence along the Markov chain, we propose to sample the weight λ and the latent factor \mathbf{U} from their joint full conditional distribution, $p(\lambda, \mathbf{U}|\mathbf{Y}, \boldsymbol{\theta}_{-\mathbf{U}, -\lambda})$, where, to streamline notation, we let $\boldsymbol{\theta}_{-(\cdot)}$ denote the set containing all variables in $\boldsymbol{\theta}$ except for (\cdot) . A sample from $p(\lambda|\mathbf{Y}, \boldsymbol{\theta}_{-\mathbf{U}, -\lambda})$ must first be obtained where

$$p(\lambda|\mathbf{Y}, \boldsymbol{\theta}_{-\mathbf{U}, -\lambda}) \propto \prod_{j=1}^J [|\lambda\Psi_j + (1-\lambda)\Lambda_j|^{-n_j/2} e^{tr(-Y_j(\lambda\Psi_j + (1-\lambda)\Lambda_j)^{-1}Y_j^T)}] \lambda^{\alpha-1} (1-\lambda)^{\beta-1}.$$

As the full conditional distribution of λ is not available in closed form, a sample of λ may be obtained from a Metropolis step that proceeds by first obtaining a proposed value λ^* from a reflecting random walk around the previous value of λ in the Markov chain (Hoff, 2009a). That is, an initial value is sampled from $\lambda^* \sim$

Uniform($\lambda - \delta_\lambda, \lambda + \delta_\lambda$), and it is reflected across the appropriate bound to retain the correct support, $\lambda \in (0, 1)$:

$$\lambda^* = \begin{cases} \lambda^* & \text{if } \lambda^* \in (0, 1), \\ |\lambda^*| & \text{if } \lambda^* \leq 0, \\ 2 - \lambda^* & \text{if } \lambda^* \geq 1. \end{cases}$$

Then, the proposal λ^* is accepted as an updated value for λ with probability $r = p(\lambda^*|\mathbf{Y}, \boldsymbol{\theta}_{-\mathbf{U}, -\lambda})/p(\lambda|\mathbf{Y}, \boldsymbol{\theta}_{-\mathbf{U}, -\lambda})$. The full conditional of each latent factor U_j for population j is independently $N_{n_j \times p}(M_j, S_j \otimes I_{n_j})$ where

$$S_j = \left(\Psi_j^{-1} + \frac{\lambda}{1-\lambda} \Lambda_j^{-1} \right)^{-1}$$

$$M_j = \frac{\lambda^{1/2}}{1-\lambda} Y_j \Lambda_j^{-1} S_j.$$

Again, to reduce dependence along the Markov chain, we propose to sample the degrees of freedom ν and covariances Ψ as well as γ and Λ from their joint full conditional distribution. In particular, the joint full conditional of (ν, Ψ) is

$$p(\nu, \Psi | \mathbf{Y}, \boldsymbol{\theta}_{-\Psi, -\nu}) = p(\Psi | \mathbf{Y}, \boldsymbol{\theta}_{-\Psi}) \times p(\nu | \mathbf{Y}, \boldsymbol{\theta}_{-\Psi, -\nu})$$

where ν may be sampled from a reflecting random walk Metropolis step. In this case, a proposal ν^* may be obtained from a reflecting random walk based on the previous iteration's value of ν , that is, sample an initial value from $\nu^* \sim \text{Uniform}(\nu - \delta_\nu, \nu + \delta_\nu)$ and utilize the following reassignment schema to ensure the sample has the correct support:

$$\nu^* = \begin{cases} \nu^* & \text{if } \nu^* \geq p + 2 \\ (p + 2) + (p + 2 - \nu^*) & \text{if } \nu^* < p + 2. \end{cases}$$

The proposal ν^* is accepted as an updated value for ν with probability

$$r = \prod_{j=1}^J \frac{p(U_j | \Psi_0, \nu = \nu^*) p(\nu = \nu^* | r_0, p_0)}{p(U_j | \Psi_0, \nu = \nu) p(\nu = \nu | r_0, p_0)}$$

where $U_j|\Psi_0, \nu \sim T_{n_j \times p}(\nu - p + 1, 0, \Psi_0(\nu - p - 1) \otimes I_{n_j})$. Then, sample each Ψ_j from its full conditional distribution,

$$\Psi_j^{-1}|\mathbf{Y}, \boldsymbol{\theta}_{-\Psi_j} \sim W_p((U_j^T U_j + (\nu - p - 1)\Psi_0)^{-1}, \nu + n_j)$$

for each $j \in \{1, \dots, J\}$. Samples from the joint full conditional distribution of $(\gamma, \boldsymbol{\Lambda})$, $p(\boldsymbol{\Lambda}|\mathbf{Y}, \boldsymbol{\theta}_{-\boldsymbol{\Lambda}}) \times p(\gamma|\mathbf{Y}, \boldsymbol{\theta}_{-\boldsymbol{\Lambda}, -\gamma})$ are obtained similarly. A proposal sample γ^* is obtained from a reflecting random walk based on an initial value drawn from $\text{Uniform}(\gamma - \delta_\gamma, \gamma + \delta_\gamma)$ and accepted with probability

$$r = \prod_{j=1}^J \frac{p(Y_j|\lambda, U_j, R_j, C_j, \gamma = \gamma^*) p(\gamma = \gamma^*|r_0, p_0)}{p(Y_j|\lambda, U_j, R_j, C_j, \gamma = \gamma) p(\gamma = \gamma|r_0, p_0)}$$

where $Y_j|\lambda, R_j, C_j \sim T_{n_j \times p}(\gamma - p + 1, \lambda^{1/2}U_j, (1 - \lambda)(C_j \otimes R_j)(\gamma - p - 1) \otimes I_{n_j})$, and the full conditional of each Λ_j^{-1} is independently

$$W_p((\tilde{Y}_j^T \tilde{Y}_j / (1 - \lambda) + (C_j \otimes R_j)(\gamma - p - 1))^{-1}, \gamma + n_j)$$

where $\tilde{Y}_j = (Y_j - \lambda^{1/2}U_j)$.

In addition to facilitating computation, these full conditionals contribute to the interpretation of parameters. In particular, for population $j \in \{1, \dots, J\}$, the full conditional means of Ψ_j and Λ_j resemble shrinkage estimators towards a pooled covariance and a Kronecker structured covariance, respectively. Specifically, the full conditional density of the across-group shrunk covariance Ψ_j , given all other model parameters and the observed data matrices \mathbf{Y} , is inverse-Wishart with mean

$$E[\Psi_j|\mathbf{Y}, \boldsymbol{\theta}_{-\Psi_j}] = w_1 U_j^T U_j / n_j + (1 - w_1) \Psi_0, \quad (4.10)$$

where $w_1 = n_j / (n_j + \nu - p - 1)$. That is, the prior on Ψ_j shrinks the sample covariance of the latent factor U towards the pooled covariance by a shrinkage factor determined by the degrees of freedom ν and the within-group sample size n_j . Similarly, the full

conditional density of the within-group shrunk Λ_j is inverse-Wishart with mean

$$E[\Lambda_j | \mathbf{Y}, \boldsymbol{\theta}_{-\Lambda_j}] = w_2 \tilde{Y}_j^T \tilde{Y}_j / n_j + (1 - w_2) (C_j \otimes R_j), \quad (4.11)$$

where $\tilde{Y}_j = (Y_j - \lambda^{1/2} U_j) / \sqrt{1 - \lambda}$ and $w_2 = n_j / (n_j + \gamma - p - 1)$. In this case, the prior on Λ_j shrinks the sample covariance of the data residual towards a separable covariance by a weight which depends on the degrees of freedom γ and n_j .

Full conditionals of Kronecker shrinkage parameters

The unstructured covariances $\mathbf{\Lambda}$ are each shrunk towards a population-specific Kronecker structured covariance. The derivations of the full conditionals of these Kronecker covariances make use of a few Kronecker product properties, namely,

$$\text{tr}(B^T A_1 B A_2^T) = \text{vec}(B)^T (A_2 \otimes A_1) \text{vec}(B)$$

and $|A_2 \otimes A_1| = |A_1|^{p_2} |A_2|^{p_1}$ for A_1 of dimension $p_1 \times p_1$ and A_2 of dimension $p_2 \times p_2$. Then, it is straightforward to derive the full conditional of each population j 's row covariance,

$$R_j \sim W_{p_1} \left(\left((\gamma - p - 1) \sum_{k=1}^p L_k C_j L_k^T + R_0^{-1} \eta_1 \right)^{-1}, \eta_1 + \gamma p_2 \right)$$

for $L_k = \text{vec}^{-1}(l_k)$ from $\Lambda_j^{-1} = \tilde{L} \tilde{L}^T = \sum_{k=1}^p l_k l_k^T$ and column covariance

$$C_j \sim W_{p_2} \left(\left((\gamma - p - 1) \sum_{k=1}^p L_k^T R_j L_k + C_0^{-1} \eta_2 \right)^{-1}, \eta_2 + \gamma p_1 \right)$$

for $L_k = \text{vec}^{-1}(l_k)$ from $\Lambda_j^{-1} = \tilde{L} \tilde{L}^T = \sum_{k=1}^p l_k l_k^T$.

Full conditionals of pooled shrinkage parameters

The unstructured covariances $\mathbf{\Psi}$ are shrunk across-groups towards a pooled unstructured covariance. The full conditional of the pooled covariance is $\Psi_0 | \mathbf{Y}, \boldsymbol{\theta}_{-\Psi_0} \sim$

$W_p(((\nu - p - 1) \sum_{j=1}^J \Psi_j^{-1} + (P_2 \otimes P_1)^{-1} \xi)^{-1}, \xi + J\nu)$. The final level of the hierarchy allows for shrinkage of this pooled unstructured covariance towards a pooled Kronecker structured covariance. The full conditional for the pooled row covariance is

$$P_1^{-1} \sim W_{p_1} \left(\left(\xi \sum_{k=1}^p L_k P_2^{-1} L_k^T + P_{01}(\eta_3 - p_1 - 1) \right)^{-1}, \eta_3 + \xi p_2 \right)$$

for $L_k = \text{vec}^{-1}(l_k)$ from $\Psi_0^{-1} = \tilde{L}\tilde{L}^T = \sum_{k=1}^p l_k l_k^T$, and the full conditional for the pooled column covariance is

$$P_2^{-1} \sim W_{p_2} \left(\left(\xi \sum_{k=1}^p L_k^T P_1^{-1} L_k + P_{02}(\eta_4 - p_2 - 1) \right)^{-1}, \eta_4 + \xi p_1 \right)$$

for $L_k = \text{vec}^{-1}(l_k)$ from $\Psi_0^{-1} = \tilde{L}\tilde{L}^T = \sum_{k=1}^p l_k l_k^T$. The corresponding degrees of freedom term may be sampled from a Metropolis step, similar to ν, γ . Specifically, a proposal sample may be obtained from a reflecting random walk based on an initial value drawn from $\text{Uniform}(\xi - \delta_\xi, \xi + \delta_\xi)$. Then, ξ^* is accepted as an updated value for ξ with probability

$$r = \prod_{j=1}^J \frac{p(\Psi_0 | P_1, P_2, \xi = \xi^*) p(\xi = \xi^* | r_0, p_0)}{p(\Psi_0 | P_1, P_2, \xi = \xi) p(\xi = \xi | r_0, p_0)}.$$

A note on computational expense

The computational complexity of the proposed Metropolis-Hastings algorithm is at least $\mathcal{O}(\max\{Jp^3, p^2 \sum_j n_j, p \sum_j n_j^2\})$. However, while Bayesian computation of the SWAG model may appear cumbersome, we note that many of the computationally expensive steps may be run in parallel across group. Specifically, sampling the random effects $\{U_j\}_{j=1}^J$, the covariance terms $\{\Psi_j\}_{j=1}^J$ and $\{\Sigma_j\}_{j=1}^J$, and the Kronecker covariance terms $\{R_j\}_{j=1}^J$ and $\{C_j\}_{j=1}^J$ may each be computed in parallel across the J

groups. In this way, the proposed algorithm may scale nicely with number of groups, depending on computational resources.

4.3.3 *Hyperparameter specification*

In the absence of meaningful external or prior information, weakly informative or non-informative priors on all unknown variables can be considered. On the weight λ , the prior hyperparameters α, β may be selected in a way that weakly encourages favoring one of the types of shrinkage, within or across groups, by setting $\alpha = \beta = 1/2$.

Specifying weakly informative hyperparameters for the degrees of freedom priors requires slightly more scrutiny as the impact of ν, γ , and ξ on the Wishart prior distributions will depend on, among other things, covariance dimension p . In a Wishart distribution, the degrees of freedom parameter controls the concentration of the distribution around the prior mean. A value that is large relative to the covariance dimension p can correspond to considerable concentration, and a value near the lower bound $p+2$ corresponds to limited concentration. As such, we suggest using a weakly informative prior for a degree of freedom by specifying hyperparameters r_0, p_0 that allow for nontrivial prior mass on a range from $p+2$ to values large relative to p . In practice, we found it worked well to set hyperparameters such that a large majority of the prior mass is placed on values in the range $[p+2, 2p]$. The size parameter r_0 may be set such that the degrees of freedom prior mean is a value near the first quantile of this range. The prior success probability may be set such that there is a fair amount of dispersion around the mean. In analyses with moderate dimension, we use $p_0 = 0.2$ which corresponds to a degrees of freedom prior variance of 5 times the prior mean. In analyses with large dimension, we use $p_0 = 0.01$ to allow for a larger prior variance. In both cases, such a prior tends to be right skewed, whereby more prior mass is placed on small to moderate values in the parameter space while

still incorporating nonnegligible prior mass on moderate to large values.

The hyperparameters on the covariance parameters, $\mathbf{R}, \mathbf{C}, P_1$, and P_2 , require special attention due to the over-parameterization of the proposed model and the scale ambiguity property of the Kronecker product. That is, for a scalar c and matrices A, B , $(cA \otimes B) = (A \otimes cB)$. To deal with these potential ambiguities, we propose standardizing the data in a pre-processing step before estimating model parameters and setting the scale hyperparameters such that the prior mean of these parameters is the identity matrix. The implications of this hyperparameter choice result in an a priori homoscedastic marginal expectation of the within-group covariances, $E[\Sigma_j] = I_p$ for each $j \in \{1, \dots, J\}$. The corresponding degrees of freedom hyperparameters are taken as $\eta_1 = \eta_3 = p_1 + 2$ and $\eta_2 = \eta_4 = p_2 + 2$ to represent diffuse distributions that maintain finite first moments. For example, a Wishart prior of this type, $W_{p_1}(I_{p_1}/(p_1 + 2), p_1 + 2)$, on R_1 corresponds to a diffuse distribution with prior expected value I_{p_1} . Furthermore, this choice of prior suggests weak shrinkage to an isotropic covariance matrix at the lowest level of the hierarchy. Weakly shrinking to such a matrix has motivations in ridge regression and the regularization discriminant analysis literature, as in Friedman (1989).

4.4 Simulation Studies

We demonstrate the performance of the proposed SWAG model by comparing the accuracy of various covariance estimators obtained under four different population covariance regimes. In particular, each regime features either homogeneous (Ho) or heterogeneous (He) covariances across groups, and the covariances in each regime are either Kronecker structured (K) or not Kronecker structured (N). Our goal in this section is to generally explore results when the true group-specific covariance matrices have off-diagonal values relatively far from zero within a given group. Specifically, in unstructured regimes, each group's true covariance is an exchangeable correlation

matrix of dimension $p \times p$ with a fixed correlation randomly generated between 0.35 and 0.9. In Kronecker structured regimes, each group’s true covariance is the Kronecker product of a $p_2 \times p_2$ exchangeable correlation matrix and a $p_1 \times p_1$ exchangeable correlation matrix. For a given regime and parameter size combination, the true covariance matrices do not vary within the simulation. Details of the within-group true covariances $\{\Sigma_1, \dots, \Sigma_J\}$ for each regime are contained in Table 4.1. While we do not necessarily expect the truth in real-world scenarios to be one of these extreme cases, this study provides insight into the behavior of the SWAG model.

Table 4.1: Population covariance assumptions for each of the four regimes considered. Notionally, $Z^{(p)}$ is an exchangeable correlation matrix of dimension $p \times p$ where the correlation is a fixed value in $[.35, .9]$.

	Ho	He
K	$\Sigma_1 = \dots = \Sigma_J = Z^{(p_2)} \otimes Z^{(p_1)}$	$\Sigma_1 = Z_1^{(p_2)} \otimes Z_1^{(p_1)}, \dots, \Sigma_J = Z_J^{(p_2)} \otimes Z_J^{(p_1)}$
N	$\Sigma_1 = \dots = \Sigma_J = Z^{(p)}$	$\Sigma_1 = Z_1^{(p)}, \dots, \Sigma_J = Z_J^{(p)}$

In a simulation study under each regime, we compare the loss of various covariance estimators under a range of dimensionality and number of groups. Specifically, we consider $p_1 \in \{2, 4, 8\}$, $p_2 = 3$, and $J \in \{4, 10\}$. As we are particularly motivated by the small sample size case, we consider within-group sample sizes $n_1 = \dots = n_J = p + 1$. For each regime, we take the MLEs of the covariances under the simplest correctly specified model as the oracle estimator. In total, this includes the standard sample MLE for each group’s covariance $\mathbf{S} = \{S_1, \dots, S_J\}$, the pooled sample MLE \hat{S}_p , the separable MLE for each group $\hat{\mathbf{K}} = \{\hat{K}_1, \dots, \hat{K}_J\}$, and the pooled separable MLE \hat{K}_p . These oracle estimators for each group’s covariance under the four regimes considered are specified in Table 4.2. We note that the most accurate estimator of the He,U regime may vary depending on the problem dimension and the number of the groups due to high variance of the sample covariance estimator when sample size

is small relative to number of features, as in this study.

Table 4.2: Oracle estimators of the population covariances in the four regimes considered.

	Ho	He
K	\hat{K}_p	$\hat{K}_1, \dots, \hat{K}_J$
N	\hat{S}_p	S_1, \dots, S_J

The simulation study proceeds as follows. For each permutation of regime, dimension, and number of groups, we sample 50 data sets each from a mean-zero matrix normal distribution of the appropriate dimension with population covariance as given in Table 4.1. For each data set, we compute the reference covariance estimates \mathbf{S} , \hat{S}_p , $\hat{\mathbf{K}}$, and \hat{K}_p . Additionally, we run the proposed Metropolis-Hastings sampler for the SWAG model for 28,000 iterations removing the first 3,000 iterations as a burn-in period and saving every 10th iteration as a thinning mechanism. From the resulting 2,500 Monte Carlo samples, we obtain the Bayes estimate under an invariant loss, Stein’s loss, of each group’s covariance, $\hat{\Sigma} = \{\hat{\Sigma}_1, \dots, \hat{\Sigma}_J\}$ where $\hat{\Sigma}_j = E[\Sigma_j^{-1}|Y_j]^{-1}$. Then, we compute Stein’s loss averaged across the populations $\bar{L}(\Sigma, \hat{\Sigma}) = \frac{1}{J} \sum_{j=1}^J L_S(\Sigma_j, \hat{\Sigma}_j)$ where $L_S(\Sigma_j, \hat{\Sigma}_j) = tr\left(\Sigma_j^{-1}\hat{\Sigma}_j\right) - \log\left|\Sigma_j^{-1}\hat{\Sigma}_j\right| - p_1p_2$. We report the average of the 50 \bar{L} values to approximate frequentist risk for each scenario considered. In our analysis of results, we refer to \bar{L} as the loss and do not discuss group-specific Stein losses.

In general, we expect inference with the ‘oracle’ estimator to outperform that of the SWAG model in each regime considered. However, as knowledge of true structural behavior is rare in practice, the oracle estimator of one regime obtained from a correctly specified model may perform arbitrarily poorly in a different regime. In contrast, due to the flexibility of the proposed SWAG model, we expect the SWAG

Table 4.3: \bar{L} values averaged over 50 iterations for J populations and problem dimension $p = p_1 p_2$. The oracle estimator for each regime has a grey background. For each case, the two smallest average losses are in bold font.

	$\hat{\Sigma}$	S	\hat{S}_p	\hat{K}	\hat{K}_p	$\hat{\Sigma}$	S	\hat{S}_p	\hat{K}	\hat{K}_p
	Homogeneous, Kronecker					Heterogeneous, Kronecker				
J = 4, p = 6	1.36	6.82	0.85	1.77	0.31	1.66	6.82	2.54	1.77	1.95
J = 4, p = 12	2.40	13.42	1.69	1.36	0.28	2.73	13.42	5.33	1.36	3.74
J = 4, p = 24	3.21	25.77	3.47	1.81	0.44	4.38	25.77	10.76	1.81	7.42
J = 10, p = 6	1.14	7.12	0.33	1.82	0.12	1.54	7.12	2.21	1.82	1.98
J = 10, p = 12	2.32	13.48	0.65	1.39	0.12	2.67	13.48	4.98	1.39	4.37
J = 10, p = 24	2.55	25.70	1.27	1.81	0.17	4.43	25.70	10.39	1.81	9.14
	Homogeneous, not Kronecker					Heterogeneous, not Kronecker				
J = 4, p = 6	1.49	6.82	0.85	4.81	2.42	1.36	6.82	1.96	6.27	4.53
J = 4, p = 12	2.65	13.42	1.69	7.07	5.79	2.77	13.42	4.21	9.04	10.58
J = 4, p = 24	4.63	25.77	3.47	22.02	20.27	5.03	25.77	8.63	31.70	31.47
J = 10, p = 6	1.47	7.12	0.33	4.64	2.12	1.54	7.12	1.48	6.14	4.44
J = 10, p = 12	2.51	13.48	0.65	7.20	5.60	2.63	13.48	3.14	9.66	10.97
J = 10, p = 24	3.45	25.70	1.27	22.32	20.27	4.52	25.70	6.48	33.55	33.17

estimator to perform nearly as well as each regime’s oracle estimator, and outperform the other estimators considered. Specifically, the SWAG model is correctly specified for all four cases as each regime corresponds to particular limiting choices of parameters in the SWAG model. Furthermore, given that we consider problem dimension size similar to each population’s sample size, we expect the SWAG estimator to outperform the sample covariance estimators in all cases.

The results of the simulation study are presented in Table 4.3. In the table, the oracle estimator for each regime has a grey background, and the smallest two average losses are in bold font. In summary, the SWAG estimator performs best or second best in all cases considered except those in the Ho, Kr regime. In nearly all cases where the SWAG estimator has the second smallest loss, it is beat by the respective regime’s oracle estimator. Therefore, given that the oracle is unknown in practice, we conclude the SWAG model is particularly effective in accurate population covariance

estimation. While the overarching conclusion of the performance of the SWAG model remains the same across most cases considered, the dynamics differ across regime. To gain a better understanding of the variation around the average losses displayed in the table, Figure 4.2 displays the empirical densities of the 50 \bar{L} values for each regime in the case where $J = 4, p = 12$.

In the regime where population covariances are homogeneous across population and Kronecker structured, the oracle estimator \hat{K}_p has the smallest average loss in the cases explored, as expected. Interestingly, for a given J , the average loss corresponding to \hat{K}_p and $\hat{\mathbf{K}}$ for $p = 12$ is less than that for $p = 6$. This is a consequence of the scaling used within Stein’s loss function as, in this case, the difference between the first two terms of the loss increases by less than the increase in p . In general, the pooled MLE and group-specific Kronecker MLEs perform well in this regime, which is not surprising given the variety of estimators considered and the overlap in their underlying assumptions. The SWAG estimator tends to have a similar \bar{L} to these estimators, and a notably smaller average loss than the sample covariance. Furthermore, in most cases considered, the empirical density of the loss corresponding to these three estimators ($\hat{\Sigma}$, \hat{S}_p , and $\hat{\mathbf{K}}$) overlap, as seen in Figure 4.2.

Given population covariances homogeneous across population and not Kronecker structured, the pooled MLE and SWAG estimators have the two smallest average losses. All other estimators have comparatively large average loss. This is particularly apparent in Figure 4.2 where the densities of the losses corresponding to $\hat{\Sigma}$ and \hat{S}_p are smaller than and distinct from the densities corresponding to the other estimators considered. In high dimension cases in particular, incorrectly assuming a Kronecker covariance in this regime results in estimators nearly as inaccurate as the sample covariances.

In the heterogeneous and Kronecker structured regime, the oracle estimator and

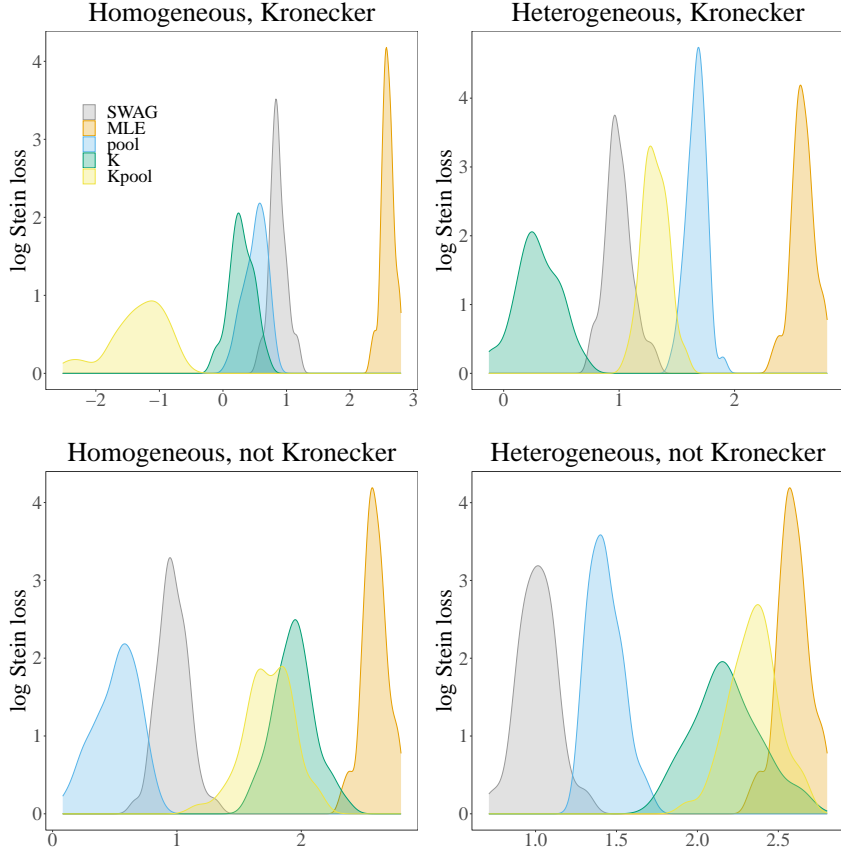


FIGURE 4.2: Empirical densities of the 50 $\log(\bar{L})$ values for each estimator given $J = 4, p = 12$.

the SWAG estimator have the two smallest average losses. Overall in this regime, these two estimators behave similarly and are notable improvements over the other estimators considered, as seen clearly in Figure 4.2. Again, in this regime, the average loss corresponding to the oracle estimator $\hat{\mathbf{K}}$ for $p = 12$ is less than that for $p = 6$ as a result of the scaling used in Stein’s loss function.

In the fourth, and perhaps most realistic, regime considered, where population covariances are heterogeneous across population and not Kronecker structured, the SWAG estimator has the smallest average loss in all cases except when $J = 10, p = 6$, where it has the second smallest average loss. In this regime, the oracle estimator is outperformed by all other estimators considered. This is not surprising given the

Table 4.4: SWAG Metropolis-Hastings sampler run time in minutes for $S = 28,000$ iterations, averaged over 50 replications for each simulation setting comprised of J groups and problem dimension $p = p_1 p_2$.

	Ho, K	Ho, N	He, K	He, N
$J = 4; p = 6$	10.58	12.49	9.89	14.32
$J = 4; p = 12$	13.83	16.23	14.20	18.71
$J = 4; p = 24$	24.46	28.36	25.07	32.59
$J = 10; p = 6$	23.13	27.92	24.41	32.37
$J = 10; p = 12$	31.12	36.96	31.63	42.61
$J = 10; p = 24$	55.92	65.04	56.37	74.33

combination of the simplicity of the exchangeable population covariances in conjunction with large instability of the sample covariance due in part to the small sample sizes considered for each regime. Again, we see estimators based on an incorrect assumption of separability result in distinctly larger loss than the more flexible SWAG estimator (Figure 4.2).

In total, these results support the conclusion that the SWAG model outperforms standard alternatives given the true structure of the population covariances is unknown. The flexibility of the SWAG model is particularly useful given the large error under estimators based on incorrect assumptions. In particular, wrongly assuming a Kronecker structure can result in an error nearly as large as that obtained under the sample covariance. While the extreme regimes considered do not necessarily reflect the truth in real-world scenarios, this simulation study highlights the flexibility of the SWAG model.

4.4.1 Analysis of computational expense

Table 4.4 displays the wall-clock run time of the Metropolis-Hastings sampler for 28,000 iterations, averaged across 50 replications. We sample all parameters as described in Section 4.3.2 and do not use any parallelization in the sampling of parameters. The algorithm was implemented for a given regime with code written with the

R statistical programming language on the Duke University Compute Cluster on a single thread with 32 CPUs and 228 GB of RAM. The smallest parameter space considered ($J = 4, p = 12$) takes about 10-15 minutes to run, and the largest parameter space considered ($J = 10, p = 24$) takes about 50-80 minutes. In summary, doubling of the dimension p from 6 to 12 results in about a 30% increase in computation time. Doubling p from 12 to 24 results in about a 70% increase in computation time. For a given p , an increase in the number of groups from 4 to 10 results in an increase in computation time of approximately 125%. As discussed in Section 4.3.2, the increase in computation time for large populations may be mitigated by parallelizing sampling across populations in Metropolis-Hastings algorithm.

4.5 Examples

We demonstrate the usefulness of the SWAG model for estimating covariance matrices in multi-group matrix-variate populations by analyzing two data sets. In the first example, we perform a speech recognition task on a publicly available spoken-word audio dataset. In the second example, we analyze chemical exposure data which features small group-specific sample sizes.

In general, a data matrix for a single group Y of dimension $n \times p$ can be decomposed into two orthogonal matrices such that one may be used for mean estimation and the other, based on centered data, for covariance estimation. As such, while a mean estimation step could be included in the proposed SWAG Metropolis-Hastings algorithm, we will estimate the covariance matrices based on centered data matrices throughout the applications. To elaborate, first, define the n -dimensional centering matrix $\mathcal{C} = \mathbf{I}_n - \mathcal{P}_1$ where $\mathcal{P}_1 = \mathbf{1}_n \mathbf{1}_n^T / n$ is a rank-1 idempotent projection matrix and \mathcal{C} is a rank- $(n - 1)$ idempotent projection matrix (Christensen, 2011). Then, note that,

$$Y = \mathbf{I}_n Y - \mathcal{P}_1 Y + \mathcal{P}_1 Y = \mathcal{C} Y + \mathbf{1}_n \bar{y}^T$$

where \bar{y} is the length p vector of column means of Y and $\mathcal{C}Y$ is the residual matrix. Then, for $Y \sim N_{n \times p}(\mathbb{1}_n \mu^T, \Sigma \otimes I_n)$, $\bar{y} \sim N_p(\mu, \Sigma/n)$ and $nS \sim W_p(\Sigma, n-1)$ where $S = (\mathcal{C}Y)^T \mathcal{C}Y/n$ is the sample covariance matrix. \bar{y} and $\mathcal{C}Y$ are uncorrelated, so \bar{y} and S are uncorrelated (Mardia et al., 1979). In this way, Σ may be estimated using centered data $\mathcal{C}Y$.

4.5.1 Classification of spoken-word audio data

Classification of a new observation based on a labeled training dataset consisting of n_j matrices, each with common dimensionality $p_1 \times p_2$, observed from each of $j \in \{1, \dots, J\}$ populations is an important statistical task for speech recognition. To illustrate the utility of the SWAG covariance estimator, we analyze classification of spoken-audio samples of words “yes”, “no”, “up”, “down”, “left”, “right”, “on”, “off”, “stop”, and “go” from a dataset consisting of 1-second WAV files where each word has a sample size ranging from 1,987 to 2,103 (Warden, 2017, 2018). Audio data such as these are commonly described by mel-frequency cepstral coefficients (MFCCs) which represent the power spectrum of a sound across time increments. Accordingly, we represent each audio sample as a $p_1 \times p_2$ feature matrix of the first $p_1 = 13$ MFCCs across $p_2 = 99$ time bins (Ligges et al., 2018).

One popular classification method for generic multivariate data is quadratic discriminant analysis (QDA). QDA is based on the result that, assuming normality and equal a priori probabilities of group membership, the probability of misclassification is minimized by assigning an unlabeled matrix $Y \in \mathbb{R}^{p_1 \times p_2}$ to group $j \in \{1, \dots, J\}$ which minimizes the discriminant score function (Mardia et al., 1979),

$$D_j(Y) = (\text{vec}(Y) - \mu_j)^T \Sigma_j^{-1} (\text{vec}(Y) - \mu_j) + \log |\Sigma_j|, \quad (4.12)$$

where $\mu_j \in \mathbb{R}^p$ and $\Sigma_j \in \mathcal{S}_p^+$, $p = p_1 p_2$, are respectively the mean vector and covariance matrix for population j and $\text{vec}(\cdot)$ is the vectorization operator that stacks the

columns of a matrix into a column vector. In practice, of course, these parameters are unknown and thus are estimated from a training data set. As a result, adequate performance of the classification relies on, among other things, accurate group-level covariance estimates.

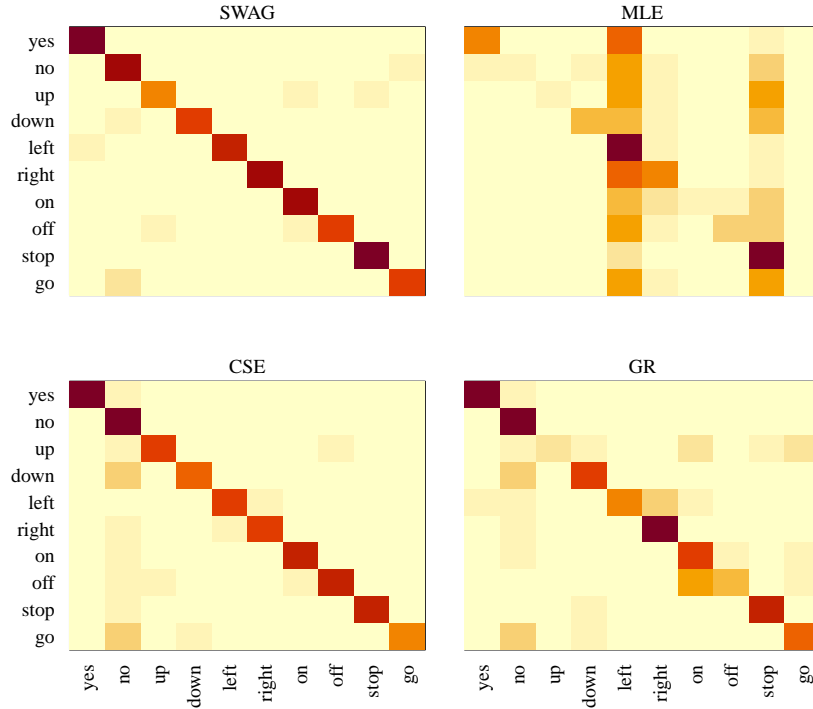


FIGURE 4.3: Confusion matrices resulting from classification from the various covariance estimates. Rows correspond to target words and columns correspond to predictions.

We display the utility of the multiple population SWAG estimators over standard estimators by comparing correct classification rates resulting from QDA in a leave-some-out comparison. To do this, we retain a random selection of 100 observations from each population as a testing dataset, and the remaining observations constitute the training dataset. For the discriminant analysis, as we are interested in comparing covariance estimation approaches, we use the sample mean to estimate each μ_j and a variety of covariance estimates for Σ_j , all computed from the training

dataset. Specifically, we obtain the SWAG covariance estimates $\hat{\Sigma}$ from output from the proposed Metropolis-Hastings algorithm run for 5,100 iterations with a burn-in of 300 and a thinning mechanism of 25. We compare this with the unstructured sample covariance obtained separately for each word \mathbf{S} , labeled MLE in this section. Additionally, we consider the partially pooled empirical Bayesian covariance estimate outlined in Greene and Rayens (1989) (GR), and the core shrinkage estimate (Hoff et al., 2022) which partially shrinks each word’s sample covariance matrix towards a separable covariance (CSE). Prior to computing the various covariance estimates, the data for each word is standardized and centered. The scale is then re-introduced to the covariance estimates prior to conducting the classification analysis.

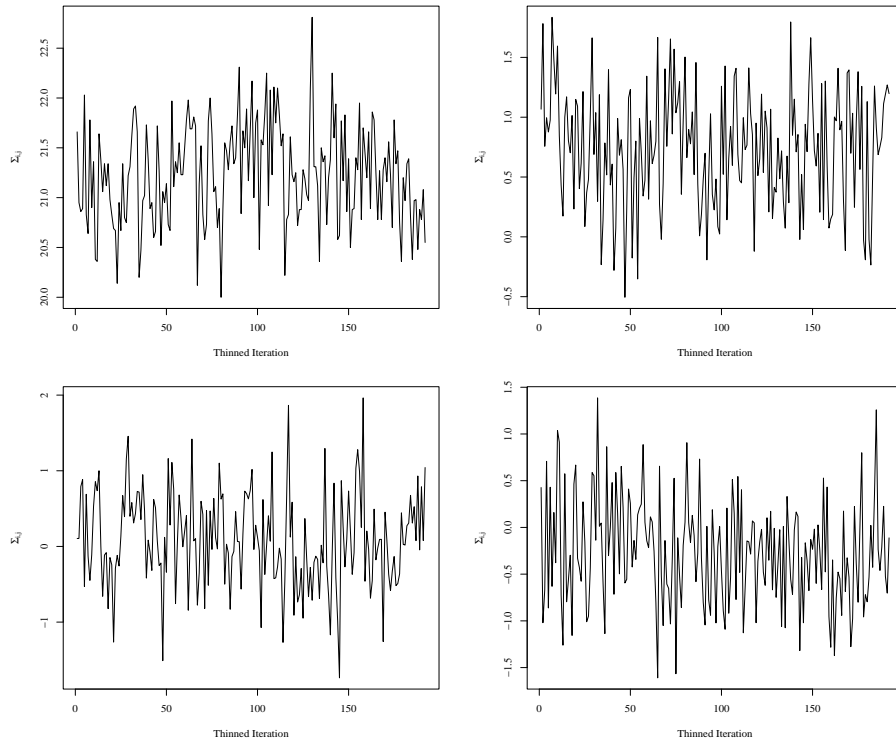


FIGURE 4.4: MCMC samples for 4 elements selected at random from the set of covariance matrices Σ .

To obtain the SWAG estimate, we run the Metropolis-Hastings sampler in an implementation written in the Julia programming language. The code is run on the

Duke University Compute Cluster using a single thread with 64 CPUs and 600 GB of RAM. We do not use any parallelization in the sampling of parameters. One implementation of the algorithm for 5,100 iterations took approximately 25 hours to run, including compilation of the code and saving the large MCMC output files. To assess convergence of the Markov chain, we analyze four randomly selected elements of the within-group covariances Σ in detail. Trace plots corresponding to each of the four elements are plotted in Figure 4.4. The maximum lag-10 autocorrelation among the four elements was 0.10 in absolute value, and the effective sample sizes of the thinned chains corresponding to each element were 117.48, 150.72, 123.87, and 192 (out of 192 thinned iterations).

Table 4.5: Rates of correct classification on audio test dataset from discriminant analysis under different covariance estimators. The average across all words for each method is displayed in the final row.

	SWAG	MLE	CSE	GR
yes	0.94	0.42	0.85	0.76
no	0.80	0.10	0.82	0.80
up	0.48	0.08	0.60	0.17
down	0.69	0.27	0.54	0.57
left	0.74	0.80	0.63	0.41
right	0.86	0.43	0.58	0.77
on	0.80	0.13	0.64	0.60
off	0.65	0.25	0.64	0.32
stop	0.88	0.80	0.70	0.63
go	0.64	0.05	0.47	0.47
<i>average</i>	0.75	0.33	0.65	0.55

Classifications for the test observations are made using each covariance estimate, and results are summarized in confusion matrices displayed in Figure 4.3, with the true word classes along the rows and predicted word classes along the columns. The correct classification rates, or, the values of the diagonal elements in the confusion matrices, are contained in Table 4.5. In general, the SWAG classifier outperforms

the other estimates. The SWAG estimate has a higher correct classification rate averaged over all words, and it features notably larger word-specific classification rates for the majority of words.

When comparing the SWAG estimate with the MLE, it features a significantly higher correct classification rate for every word except two, “left” and “stop”. Upon further inspection, however, the two large correct classification rates for the MLE are a feature of this classifier nearly always choosing one of these two words, as displayed in the confusion matrix. The correct classification rates obtained from the SWAG classifier are greater than or equal to those from the partially pooled estimate GR for every word, oftentimes by a large margin, which corresponds with a greater overall correct classification rate. Moreover, linear discriminant analysis, which uses a pooled covariance estimate \hat{S}_p for each population covariance, performs even worse with an across-word average correct classification rate of 0.27. The most convincing competing classifier is the core shrinkage estimate. It has a larger correct classification rate for for the word “up” and correctly classifies two more observations for the word “no” than the SWAG estimate. While the CSE features slightly better performance for these two words, though, the SWAG classifier performs better over all populations. Furthermore, the SWAG classifier performs much better overall than the separable MLEs obtained separately for each population, $\hat{\mathbf{K}}$, which has an across-word average correct classification rate of 0.49. On the whole, the SWAG classifier outperforms the other classifiers considered with respect to accurate classification across all populations, and this example showcases the benefit of allowing for shrinkage both within and across populations.

4.5.2 Analysis of TESIE chemical exposures data

Many recent medical and environmental studies are concerned with understanding differences among socio-economic groups from repeated measurements of chemical

exposures (James-Todd et al., 2017). In such an application, researchers may be interested in understanding within and across group heterogeneity. Additionally, researchers are often interested in understanding covariate effects and appropriately handling missing data. For all such tasks, accurate covariance modeling is critical.

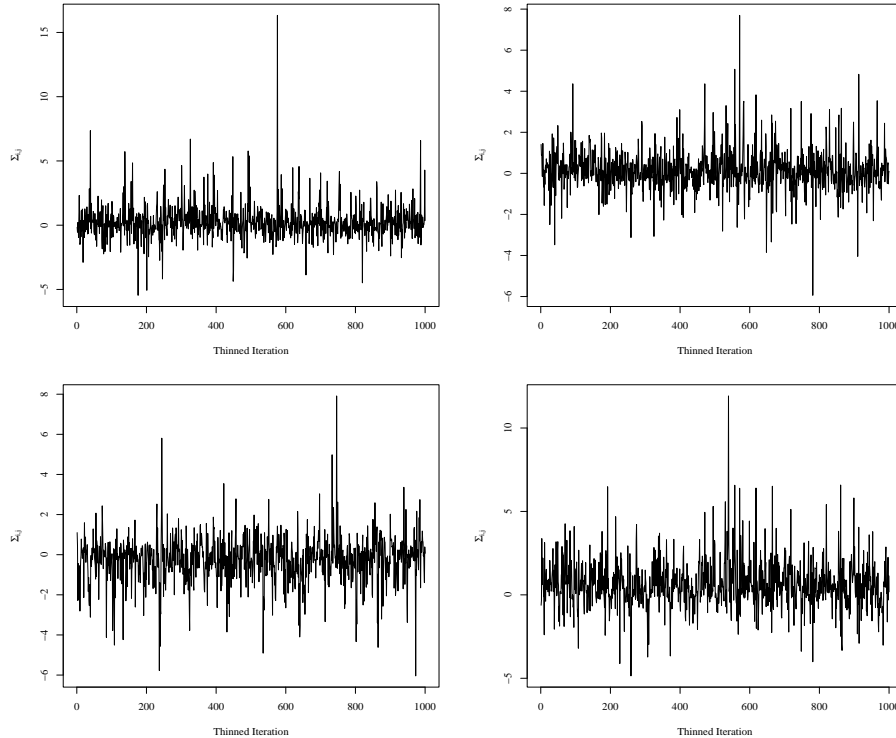


FIGURE 4.5: MCMC samples for four elements selected at random from the set of covariance matrices $\{\Sigma_{LHS}, \Sigma_{HS}, \Sigma_C\}$.

In this section, we analyze a sample gathered in the Toddlers Exposure to SVOCs in Indoor Environments (TESIE) study (Hoffman et al., 2018). In this study, biomarkers of various semi-volatile organic compounds (SVOCs) were extracted from paired samples of urine, blood, and silicone wristbands (see, e.g., Hammel et al. (2018)). In particular, each observation is a $p_1 \times p_2$ matrix where the rows represent biomarkers from $p_1 = 5$ SVOCs obtained from the $p_2 = 3$ sources. Additionally, socio-economic covariates were collected for each individual in the study including highest education level attained and race, among others. We will analyze the $p_1 p_2 \times p_1 p_2$ covariance

matrices across education level as a proxy for different socio-economic populations. Specifically, the three education levels considered are less than high school (LHS), high school degree or GED (HS), and some college or college degree (C). The sample sizes are 30, 19, and 24 for the three populations defined by education levels LHS, HS, and C, respectively.

We proceed with simultaneously estimating each group’s covariance with the SWAG model. While we remove the mean effect, the SWAG model can be extended to include a regression on covariates of interest, for example, with the addition of sampling step for a regression coefficient in the proposed MCMC sampler. We run the Metropolis-Hastings sampler for 33,000 iterations, remove the first 3,000 iterations as a burn-in period, and retain every 30th sample as a thinning mechanism. The 33,000 iterations were completed in 3 minutes in an implementation of the sampler using the R statistical programming language on a personal machine with an Apple Silicon processor and 8 GB of RAM. Mixing of the Markov chain for model parameters was good. The autocorrelation for the thinned chains corresponding to each of the elements in Σ was low, with a maximum lag-10 autocorrelation among all 360 elements of 0.11 in absolute value. Furthermore, the average effective sample size of the thinned chains was 773.73, with a range of 250.40 to 1000 (out of 1000 thinned iterations). For reference, see Figure 4.5 for trace plots of four randomly selected elements of Σ .

Approximations to the posteriors of shrinkage-controlling parameters λ, ν, γ , and ξ are plotted in Figure 4.6. The posterior density of the weight on within versus across population shrinkage, λ , is concentrated around the upper bound of one, indicating that most of the weight is being placed on shrinking across groups towards a pooled covariance. Additionally, the posteriors of the degrees of freedom for across-group shrinkage ν and ξ are concentrated around large values indicating strong shrinkage from an unstructured covariance towards a pooled Kronecker structured covariance.

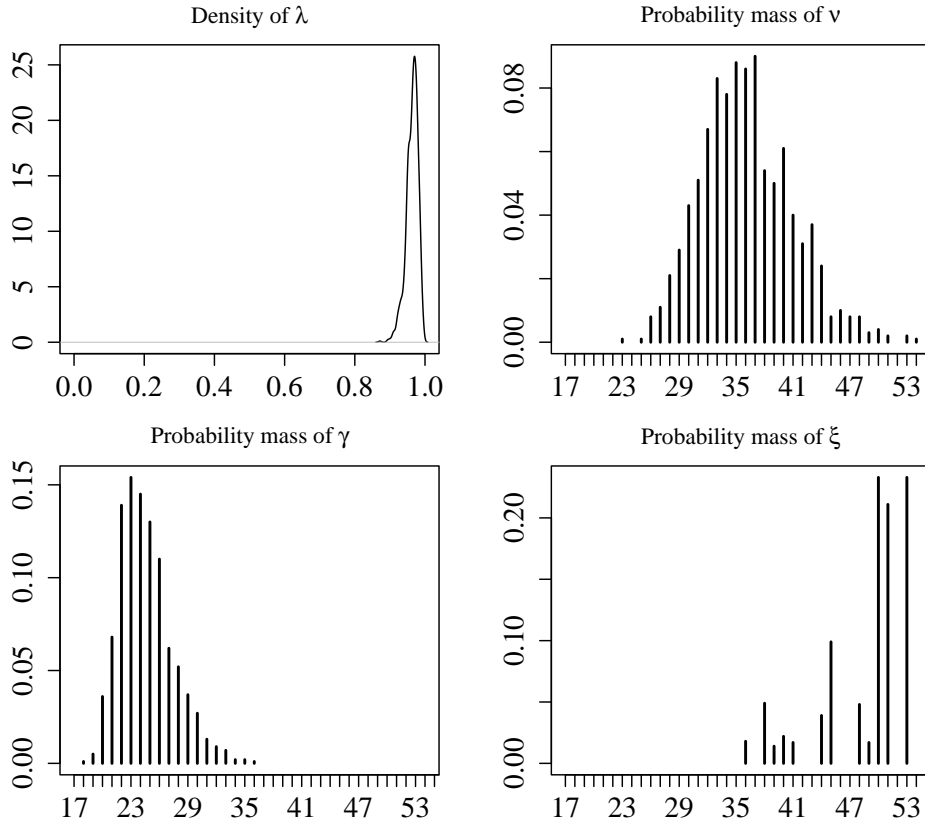


FIGURE 4.6: Approximations to the posterior distributions of key parameters λ , ν , η , and ξ for the TESIE data example.

In this way, interpretation of summarizing across-row and across-column covariance estimates is straightforward.

This shrinkage behavior is evident upon comparing the Bayes estimates of the group-specific covariances $\hat{\Sigma}_{LHS}$, $\hat{\Sigma}_{HS}$, and $\hat{\Sigma}_C$ to the Bayes estimates of the pooled Kronecker covariances \hat{P}_1 and \hat{P}_2 (Figure 4.7). Here, \hat{P}_1 is the across-SVOC covariance estimate and \hat{P}_2 is the across-source covariance estimate. The shrinkage towards a pooled separable structure is recognizable throughout the three group-specific covariance estimates. In particular, the general pattern of across-SVOC heterogeneity seen in \hat{P}_1 is roughly present in each 5x5 block in the group-specific covariances, seen in the top row of Figure 4.7. However, a benefit of the SWAG model is the ability to allow for divergences from this separable structure, a feature also discernible in

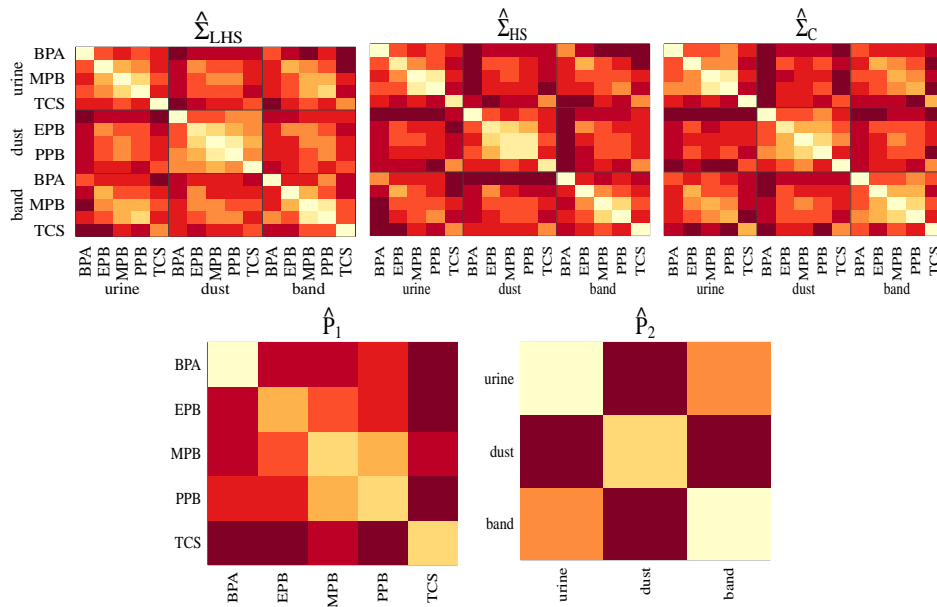


FIGURE 4.7: Bayes estimates under Stein’s loss of the covariances for the three populations LHS, HS, and C are plotted on the top row. Bayes estimates under Stein’s loss of the pooled Kronecker covariance are plotted on the bottom row.

the group-specific estimates. For example, within a given population, say, the LHS population, the pattern among the covariances between BPA and the other SVOCs from samples obtained from urine differs across measurement source by more than a single factor. Moreover, this pattern differs across populations which reflects heterogeneity across populations. In total, the output of the SWAG model allows for interpretation of a row covariance and a column covariance, shared across groups, while allowing for deviations from this structure at the group level.

4.6 Discussion

In this chapter, we propose a flexible model-based covariance estimation procedure for multi-group matrix-variate data. The SWAG hierarchical model provides a coherent approach to combining two common types of shrinkage, within populations towards a Kronecker structure and across populations towards a pooled covariance. Bayesian inference of model parameters is straightforward with a Metropolis-Hastings

algorithm and allows for uncertainty quantification of covariance estimates. In simulation studies, we show the flexibility of the proposed method results in covariance estimates that outperform standard estimates in a wide array of settings in terms of loss.

The flexibility of the SWAG model is developed specifically for matrix-variate data, but the model and estimation procedure can be adapted for other types of data or applications. If the data being analyzed are not matrix-variate, different structures can be utilized in place of the Kronecker product. Additionally, the shrinkage towards a pooled covariance can be replaced to represent more complex relationships across the populations such as, for example, an autoregressive relationship.

While an inverse-Wishart prior results in computationally convenient inference and has a practically useful interpretation in that the Bayes estimator of the population covariance under a normal-inverse-Wishart hierarchical model is a linear shrinkage estimator, it is potentially limiting in that, for inference on the covariance of a single group, one degree of freedom parameter controls concentration around the prior for the p standard deviations and the correlation structure. As such, a promising direction for future work is to explore SWAG priors on the correlation decomposition of unstructured covariances, as in Barnard et al. (2000). This could allow for informative priors to be placed on the correlation structure, while using uninformative priors on the variable-specific standard deviations, or, vice-a-versa.

Replication codes are available at <https://github.com/betsybersson/SWAG>.

Conclusion

This thesis details methodology developed to improve inference for a variety of tasks in analyzing multi-group data. In analyzing such data, precise and accurate group-specific inference is difficult to obtain for some groups with small within-group sample sizes. To this end, in Chapters 2 and 3, we detail methodologies to construct prediction regions for numeric and categorical data types. The prediction regions are constructed using indirect information and maintain finite-sample frequentist coverage guarantees. When implemented for multi-group data, these approaches make use of data in auxiliary groups to construct accurate and precise prediction regions for a given group. In Chapter 4, we present methodology to improve covariance estimation based on structured multi-group data. This approach flexibly allows for information to be shared across groups and flexibly make use of structural information to improve the overall accuracy of the covariance estimates. We now briefly summarize directions of future work for both avenues of research.

In the prediction region work, the coverage guarantee of these approaches relies on the assumption of exchangeability. As a result, the prediction region methods presented for multi-group data in Chapters 2 and 3 are applicable to observations

resulting from simple random samples. As such, a natural, and useful, future direction is to extend the approach to account for different sampling schemes. Separately, developing methods for frequentist-valid prediction regions that incorporate indirect information might prove beneficial for different broad application settings such as in online learning.

The covariance estimation work presented in Chapter 4 highlights the benefit of relaxing structural assumptions and instead utilizing a flexible shrinkage approach. With this motivation, in future work, it would be beneficial to extend the framework presented in this chapter to allow for shrinkage to structures more useful for other types of data beyond matrix-variate. Separately, the SWAG approach presented is primarily concerned with covariance estimation. Incorporating shrinkage of an unstructured mean parameter towards a separable structure would make for a flexible holistic model, to be used as a default for multi-group matrix-variate data.

Appendix A

Supplementary Material for Chapter 2

A.1 Derivation of Marginal Likelihood

In this section, we derive the marginal likelihood in (11). Independently for each $j \in \{1, \dots, J\}$, let $m_j^2 | \sigma_j^2 = \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_j)^2 | \sigma_j^2 \sim \sigma_j^2 \chi_{n_j-1}^2$ and $1/\sigma_j^2 \sim G(a/2, b/2)$. Note that,

$$\sigma_j^2 \chi_{n_j-1}^2 =^D \sigma_j^2 G((n_j - 1)/2, 1/2) =^D G((n_j - 1)/2, 1/(2\sigma_j^2)).$$

Then, it is straightforward to obtain the joint marginal likelihood of $\{m_j^2\}_j$:

$$\begin{aligned} p(m_1^2, \dots, m_J^2 | a, b) &= \int \cdots \int p(m_1^2, \dots, m_J^2, \sigma_1^2, \dots, \sigma_J^2 | a, b) d\sigma_1^2 \cdots d\sigma_J^2 \\ &= \prod_{j=1}^J \int p(m_j^2 | \sigma_j^2, a, b) p(\sigma_j^2 | a, b) d\sigma_j^2 \\ &= \prod_{j=1}^J f(m_j^2) \frac{(b/2)^{a/2}}{(a/2)} \int \left(\frac{1}{\sigma_j^2}\right)^{\frac{a+n_j-1}{2}-1} \exp\left\{-\frac{1}{\sigma_j^2} \frac{b+m_j^2}{2}\right\} d\sigma_j^2 \\ &= \prod_{j=1}^J f(m_j^2) \frac{(b/2)^{a/2}}{(a/2)} \frac{\Gamma((a+n_j-1)/2)}{((b+m_j^2)/2)^{(a+n_j-1)/2}} \end{aligned}$$

where $f(\cdot)$ is a function which does not depend on σ^2 , a , or b .

In this derivation, it is easy to see $1/\sigma_j^2 | m_j^2, a, b \sim G\left(\frac{a+n_j-1}{2}, \frac{b+m_j^2}{2}\right)$. Hence, the posterior mode is $\text{Mode}[\sigma_j^2 | m_j^2, a, b] = \frac{b+m_j^2}{a+(n_j-1)+2}$.

A.2 Notation Simplification

We begin by simplifying notation to be used in all proofs hereafter. In working with the Bayes-optimal conformity measure, we often find ourselves making comparisons between densities of two random variables conditional on the other and a shared set of random variables. The information contained in the shared set may be viewed as prior information, which results in fewer variables to keep track of. Specifically, for two random variables of interest, Y_j, Y_k , conditioning on extra data $\mathbf{Y}_{-j,-k}$ simply requires an update of the posterior of model parameters $\boldsymbol{\theta}$. In terms specific to the conformal predictive procedure, Lemma 5 allows us to consider a simplified regime where there is a single data point y_j and a single candidate prediction y_{n+1} .

Lemma 5. *The density $p(Y_j | \{Y_1, \dots, Y_{n+1}\} \setminus Y_j)$ under the model:*

$$Y_1, \dots, Y_{n+1} \sim i.i.d. P_{\boldsymbol{\theta}} \tag{A.1}$$

$$\boldsymbol{\theta} \sim Q,$$

is equivalent to $p(Y_j | Y_k)$ under the model:

$$Y_j, Y_k \sim i.i.d. P_{\boldsymbol{\theta}} \tag{A.2}$$

$$\boldsymbol{\theta} \sim \tilde{Q}.$$

where

$$\tilde{q}(\boldsymbol{\theta}) = p_q(\boldsymbol{\theta} | \mathbf{Y}_{-j,-k}) = \frac{p_{\boldsymbol{\theta}}(\mathbf{Y}_{-j,-k} | \boldsymbol{\theta}) q(\boldsymbol{\theta})}{\int_{\Theta} p_{\boldsymbol{\theta}}(\mathbf{Y}_{-j,-k} | \boldsymbol{\theta}) q(\boldsymbol{\theta}) d\boldsymbol{\theta}}$$

and $\mathbf{Y}_{-j,-k} := \{Y_1, \dots, Y_{n+1}\} \setminus \{Y_j, Y_k\}$ for any $k \in \{1, \dots, n+1\}$, $k \neq j$, and $p_q(\boldsymbol{\theta} | \cdot)$ refers to the posterior density of $\boldsymbol{\theta}$ conditional on (\cdot) under the prior q .

Proof of Lemma 5. We aim to show the distribution of $Y_j|\mathbf{Y}_{-j}$ under (A.1) is equivalent to the distribution of $Y_j|Y_k$ under (A.2). First,

$$\begin{aligned}
p_{\tilde{q}}(\boldsymbol{\theta}|Y_k) &\propto p_{\theta}(Y_k|\boldsymbol{\theta})\tilde{q}(\boldsymbol{\theta}) \\
&= p_{\theta}(Y_k|\boldsymbol{\theta})p_q(\boldsymbol{\theta}|\mathbf{Y}_{-j,-k}) \\
&\propto p_{\theta}(Y_k|\boldsymbol{\theta})p_{\theta}(\mathbf{Y}_{-j,-k}|\boldsymbol{\theta})q(\boldsymbol{\theta}) \\
&= p_{\theta}(\mathbf{Y}_{-j}|\boldsymbol{\theta})q(\boldsymbol{\theta}) \\
&\propto p_q(\boldsymbol{\theta}|\mathbf{Y}_{-j})
\end{aligned}$$

Therefore, $p_{\tilde{q}}(\boldsymbol{\theta}|Y_k) \equiv p_q(\boldsymbol{\theta}|\mathbf{Y}_{-j})$.

Then, for (A.1)

$$\begin{aligned}
p(Y_j|\mathbf{Y}_{-j}) &= \int_{\Theta} p_q(Y_j, \boldsymbol{\theta}|\mathbf{Y}_{-j})d\boldsymbol{\theta} \\
&= \int_{\Theta} p_{\theta}(Y_j|\boldsymbol{\theta})p_q(\boldsymbol{\theta}|\mathbf{Y}_{-j})d\boldsymbol{\theta} \\
&\equiv \int_{\Theta} p_{\theta}(Y_j|\boldsymbol{\theta})p_{\tilde{q}}(\boldsymbol{\theta}|Y_k)d\boldsymbol{\theta},
\end{aligned}$$

which is the definition of $p(Y_j|Y_k)$ for (A.2). □

A.3 Proofs

One straightforward method for proving two conformity measures are ECM is to show, for each $i = 1, \dots, n + 1$, the sub-region of acceptance S_i , $S_i = \{y_{n+1} \in \mathbb{R} : c_i(y_{n+1}) \leq c_{n+1}(y_{n+1})\}$, is the same for both measures. This generic result will be used in the proof of Theorem 1.

Lemma 6. *For conformity measures C, D , if*

$$\{y_{n+1} : c_i(y_{n+1}) \leq c_{n+1}(y_{n+1})\} = \{y_{n+1} : d_i(y_{n+1}) \leq d_{n+1}(y_{n+1})\} \quad \forall i = 1, \dots, n + 1$$

then C and D are ECM.

Proof of Lemma 6.

$$\begin{aligned}
\{y_{n+1} : c_i \leq c_{n+1}\} &= \{y_{n+1} : d_i \leq d_{n+1}\} \quad \forall i = 1, \dots, n+1 \\
&\Rightarrow \#\{i : c_i \leq c_{n+1}\} = \#\{i : d_i \leq d_{n+1}\} \\
&\Rightarrow p_{y,c} = p_{y,d}
\end{aligned}$$

where $p_{y,x}$ is the conformal p -value corresponding to conformity measure x . Thus each candidate prediction value will be treated the same under both conformity measures C, D . \square

Proof of Theorem 1. Based on Lemma 5, we consider the conformity between two values y_1, y_2 . Under the normal working model, in this case, the Bayes-optimal conformity measure is

$$\begin{aligned}
C_B(y_1, y_2) &:= p(y_2|y_1) = \\
&\frac{\Gamma\left(\frac{a_{12}+1}{2}\right)}{\sqrt{a_{12}\pi}\Gamma\left(\frac{a_{12}}{2}\right)} \left(\left(\frac{b_{2|1}}{a_{12}} (1 + \tau_{12}^2) \right)^{-1/2} \left(1 + \frac{1}{a_{12}} \frac{(y_2 - \mu_{2|1})^2}{\frac{b_{2|1}}{a_{12}} (1 + \tau_{12}^2)} \right)^{-(a_{12}+1)/2} \right),
\end{aligned}$$

where

$$\begin{aligned}
\tau_{12}^2 &= (1/\tau^2 + 1)^{-1} \\
\mu_{2|1} &= (\mu/\tau^2 + y_1)\tau_{12}^2 \\
a_{12} &= a + 1 \\
b_{2|1} &= b + y_1^2 + \mu^2/\tau^2 - (\mu/\tau^2 + y_1)^2\tau_{12}^2.
\end{aligned}$$

Now, suppose the conformal algorithm requires that we identify the region of y_1 s.t. $C_B(y_1, y_2) \leq C_B(y_2, y_1)$, or, equivalently, the region where

$$C_B(y_1, y_2) / C_B(y_2, y_1) \leq 1.$$

This region can be shown to be the same as that obtained from $C_B(\{y_1, y_2\}, y_2) \leq C_B(\{y_1, y_2\}, y_1)$, which is the definition of equivalent conformity measures. Well, first

note:

$$\begin{aligned}
& C_B(y_1, y_2) / C_B(y_2, y_1) \\
&= \frac{\frac{\Gamma(\frac{a_{12}+1}{2})}{\sqrt{a_{12}\pi}\Gamma(\frac{a_{12}}{2})} \left(\left(\frac{b_{2|1}}{a_{12}} (1 + \tau_{12}^2) \right)^{-1/2} \left(1 + \frac{1}{a_{12}} \frac{(y_2 - \mu_{2|1})^2}{\frac{b_{2|1}}{a_{12}} (1 + \tau_{12}^2)} \right)^{-(a_{12}+1)/2} \right)}{\frac{\Gamma(\frac{a_{12}+1}{2})}{\sqrt{a_{12}\pi}\Gamma(\frac{a_{12}}{2})} \left(\left(\frac{b_{1|2}}{a_{12}} (1 + \tau_{12}^2) \right)^{-1/2} \left(1 + \frac{1}{a_{12}} \frac{(y_1 - \mu_{1|2})^2}{\frac{b_{1|2}}{a_{12}} (1 + \tau_{12}^2)} \right)^{-(a_{12}+1)/2} \right)} \\
&= \frac{(b_{2|1})^{-1/2} \left(1 + \frac{1}{a_{12}} \frac{(y_2 - \mu_{2|1})^2}{\frac{b_{2|1}}{a_{12}} (1 + \tau_{12}^2)} \right)^{-(a_{12}+1)/2}}{(b_{1|2})^{-1/2} \left(1 + \frac{1}{a_{12}} \frac{(y_1 - \mu_{1|2})^2}{\frac{b_{1|2}}{a_{12}} (1 + \tau_{12}^2)} \right)^{-(a_{12}+1)/2}} \\
&= \left[\frac{(b_{1|2}) \left(1 + \frac{(y_1 - \mu_{1|2})^2}{b_{1|2} (1 + \tau_{12}^2)} \right)^{a_{12}+1}}{(b_{2|1}) \left(1 + \frac{(y_2 - \mu_{2|1})^2}{b_{2|1} (1 + \tau_{12}^2)} \right)^{a_{12}+1}} \right]^{1/2} \\
&= \left[\frac{(b_{1|2}) \left(\frac{b_{1|2}(1 + \tau_{12}^2) + (y_1 - \mu_{1|2})^2}{b_{1|2}} \right)^{a_{12}+1}}{(b_{2|1}) \left(\frac{b_{2|1}(1 + \tau_{12}^2) + (y_2 - \mu_{2|1})^2}{b_{2|1}} \right)^{a_{12}+1}} \right]^{1/2} \\
&= \left[\left(\frac{b_{2|1}}{b_{1|2}} \right)^{a_{12}} \left(\frac{b_{1|2}(1 + \tau_{12}^2) + (y_1 - \mu_{1|2})^2}{b_{2|1}(1 + \tau_{12}^2) + (y_2 - \mu_{2|1})^2} \right)^{a_{12}+1} \right]^{1/2} \tag{A.3} \\
&= \left(\frac{b_{2|1}}{b_{1|2}} \right)^{a_{12}/2} \tag{A.4}
\end{aligned}$$

Since the second term in (A.3) simplifies to 1:

$$\begin{aligned}
& \text{numerator} \left(\frac{b_{1|2}(1 + \tau_{12}^2) + (y_1 - \mu_{1|2})^2}{b_{2|1}(1 + \tau_{12}^2) + (y_2 - \mu_{2|1})^2} \right) \\
& := b_{1|2}(1 + \tau_{12}^2) + (y_1 - \mu_{1|2})^2 \\
& := (b + y_2^2 + \mu^2/\tau^2 - (\mu/\tau^2 + y_2)^2 \tau_{12}^2) (1 + \tau_{12}^2) \\
& \quad + (y_1 - (\mu/\tau^2 + y_2) \tau_{12}^2)^2 \\
& = D + y_2^2 \\
& \quad + y_2 [-2(\mu/\tau^2) \tau_{12}^2 (1 + \tau_{12}^2) + 2(\mu/\tau^2) (\tau_{12}^2)^2] \\
& \quad + y_1^2 \\
& \quad + y_1 [-2(\mu/\tau^2) \tau_{12}^2] \\
& \quad + y_1 y_2 [-2\tau_{12}^2] \\
& = D + y_2^2 \tag{A.5} \\
& \quad + y_2 [-2(\mu/\tau^2) \tau_{12}^2] \\
& \quad + y_1^2 \\
& \quad + y_1 [-2(\mu/\tau^2) \tau_{12}^2] \\
& \quad + y_1 y_2 [-2\tau_{12}^2],
\end{aligned}$$

for some function D that does not depend on y_1 or y_2 . Since the coefficients on the y_1, y_2 terms are the same in (A.5), by symmetry, the denominator of the second term in (A.3) will also be equal to (A.5).

Substituting (A.4) into the original inequality of interest, we are able to show,

$$\begin{aligned}
& C_B(y_1, y_2) \leq C_B(y_2, y_1) \\
& \Leftrightarrow C_B(y_1, y_2) / C_B(y_2, y_1) \leq 1 \\
& \Leftrightarrow \left(\frac{b_{2|1}}{b_{1|2}} \right)^{a_{12}/2} \leq 1 \\
& \Leftrightarrow b_{2|1} - b_{1|2} \leq 0 \\
& := (b + y_1^2 + \mu^2/\tau^2 - (\mu/\tau^2 + y_1)^2 \tau_{12}^2) \\
& \quad - (b + y_2^2 + \mu^2/\tau^2 - (\mu/\tau^2 + y_2)^2 \tau_{12}^2) \leq 0 \\
& \Leftrightarrow \left(y_1 - \mu/\tau^2 \frac{\tau_{12}^2}{1 - \tau_{12}^2} \right)^2 - \left(y_2 - \mu/\tau^2 \frac{\tau_{12}^2}{1 - \tau_{12}^2} \right)^2 \leq 0 \tag{A.6}
\end{aligned}$$

We now show the inequality $C_B(\{y_1, y_2\}, y_2) \leq C_B(\{y_1, y_2\}, y_1)$ simplifies to (A.6).

For any $X \in \mathbb{R}$, the conformity score is:

$$C_B(\{y_1, y_2\}, X) := p(X|y_1, y_2) \tag{A.7}$$

$$= \frac{\Gamma\left(\frac{a_{12'}+1}{2}\right)}{\sqrt{a_{12'}} \pi \Gamma\left(\frac{a_{12'}}{2}\right)} \left(\left(\frac{b_{12}}{a_{12'}} (1 + \tau_{12'}^2) \right)^{-1/2} \left(1 + \frac{1}{a_{12'}} \frac{(X - \mu_{12})^2}{\frac{b_{12}}{a_{12'}} (1 + \tau_{12'}^2)} \right)^{-(a_{12'}+1)/2} \right),$$

where

$$\begin{aligned}
\tau_{12'}^2 &= (1/\tau^2 + 2)^{-1} \\
\mu_{12} &= (\mu/\tau^2 + y_1 + y_2) \tau_{12'}^2 \\
a_{12'} &= a + 2 \\
b_{12} &= b + y_1^2 + y_2^2 + \mu^2/\tau^2 - (\tau_{12'}^2)^{-1} \mu_{12}^2.
\end{aligned}$$

Notice that all parameters of the t distribution that define this chosen conformity measure will be the same regardless of which variable we are obtaining the conformity score for. That is, first note that,

$$\begin{aligned}
& C_B(\{y_1, y_2\}, y_2) \leq C_B(\{y_1, y_2\}, y_1) \\
& \Leftrightarrow (y_1 - \mu_{12})^2 - (y_2 - \mu_{12})^2 \leq 0.
\end{aligned}$$

Well,

$$\begin{aligned}
& (y_1 - \mu_{12})^2 - (y_2 - \mu_{12})^2 \\
:= & (y_1 - (\mu/\tau^2 + y_1 + y_2) \tau_{12'}^2)^2 - (y_2 - (\mu/\tau^2 + y_1 + y_2) \tau_{12'}^2)^2 \\
= & \left[y_1^2 (1 - \tau_{12'}^2)^2 + (\mu/\tau^2 + y_2)^2 (\tau_{12'}^2)^2 - 2y_1 (1 - \tau_{12'}^2) (\mu/\tau^2 + y_2) \tau_{12'}^2 \right] \\
& - \left[y_2^2 (1 - \tau_{12'}^2)^2 + (\mu/\tau^2 + y_1)^2 (\tau_{12'}^2)^2 - 2y_2 (1 - \tau_{12'}^2) (\mu/\tau^2 + y_1) \tau_{12'}^2 \right] \\
= & \left[y_1^2 (1 - \tau_{12'}^2)^2 - y_2^2 (\tau_{12'}^2)^2 - 2y_1 (\mu/\tau^2) (\tau_{12'}^2)^2 - 2y_1 (1 - \tau_{12'}^2) (\mu/\tau^2) \tau_{12'}^2 \right] \\
& - \left[y_2^2 (1 - \tau_{12'}^2)^2 - y_1^2 (\tau_{12'}^2)^2 - 2y_2 (\mu/\tau^2) (\tau_{12'}^2)^2 - 2y_2 (1 - \tau_{12'}^2) (\mu/\tau^2) \tau_{12'}^2 \right] \\
= & \left[y_1^2 (1 - 2\tau_{12'}^2) - 2y_1 (\mu/\tau^2) \tau_{12'}^2 \right] - \left[y_2^2 (1 - 2\tau_{12'}^2) - 2y_2 (\mu/\tau^2) \tau_{12'}^2 \right]
\end{aligned}$$

Furthermore,

$$\begin{aligned}
& \left[y_1^2 (1 - 2\tau_{12'}^2) - 2y_1 (\mu/\tau^2) \tau_{12'}^2 \right] - \left[y_2^2 (1 - 2\tau_{12'}^2) - 2y_2 (\mu/\tau^2) \tau_{12'}^2 \right] \leq 0 \\
\Leftrightarrow & \left(y_1 - \mu/\tau^2 \frac{\tau_{12'}^2}{1 - 2\tau_{12'}^2} \right)^2 - \left(y_2 - \mu/\tau^2 \frac{\tau_{12'}^2}{1 - 2\tau_{12'}^2} \right)^2 \leq 0 \\
\Leftrightarrow & \left(y_1 - \mu/\tau^2 \frac{\tau_{12}^2}{1 - \tau_{12}^2} \right)^2 - \left(y_2 - \mu/\tau^2 \frac{\tau_{12}^2}{1 - \tau_{12}^2} \right)^2 \leq 0
\end{aligned}$$

since

$$\begin{aligned}
\frac{\tau_{12'}^2}{1 - 2\tau_{12'}^2} &:= \frac{(1/\tau^2 + 2)^{-1}}{1 - 2(1/\tau^2 + 2)^{-1}} \\
&= \frac{1}{(1/\tau^2 + 2) - 2(1/\tau^2 + 2)(1/\tau^2 + 2)^{-1}} \\
&= \frac{1}{1/\tau^2} \\
&= \frac{1}{1/\tau^2 + 1 - (\tau_{12}^2)^{-1} \tau_{12}^2} \\
&=: \frac{1}{(\tau_{12}^2)^{-1} - (\tau_{12}^2)^{-1} \tau_{12}^2} \\
&= \frac{\tau_{12}^2}{1 - \tau_{12}^2}.
\end{aligned}$$

So, we have shown

$$\begin{aligned}
C_B(\{y_1, y_2\}, y_2) &\leq C_B(\{y_1, y_2\}, y_1) \\
\Leftrightarrow \left(y_1 - \mu/\tau^2 \frac{\tau_{12}^2}{1 - \tau_{12}^2}\right)^2 - \left(y_2 - \mu/\tau^2 \frac{\tau_{12}^2}{1 - \tau_{12}^2}\right)^2 &\leq 0 \tag{A.8}
\end{aligned}$$

In summary, by (A.6) and (A.8) we have that

$$\begin{aligned}
C_B(y_1, y_2) &\leq C_B(y_2, y_1) \\
\Leftrightarrow \left(y_1 - \mu/\tau^2 \frac{\tau_{12}^2}{1 - \tau_{12}^2}\right)^2 - \left(y_2 - \mu/\tau^2 \frac{\tau_{12}^2}{1 - \tau_{12}^2}\right)^2 &\leq 0 \\
\Leftrightarrow C_B(\{y_1, y_2\}, y_2) &\leq C_B(\{y_1, y_2\}, y_1).
\end{aligned}$$

Then, by Lemma 6, $C_B(\{Y_1, \dots, Y_{n+1}\} \setminus \{Y_i, Y_i\})$ and $C_B(\{Y_1, \dots, Y_{n+1}\}, Y_i)$ are ECD. \square

Proof of Lemma 1. For each $i = 1, \dots, n + 1$, the region $S_i = \{y_{n+1} \in \mathbb{R} : c_i(y_{n+1}) \leq c_{n+1}(y_{n+1})\}$ is an interval and contains a shared value $\gamma \in S_i$, $1 \leq i \leq n + 1$. Clearly $S_{n+1} = \mathbb{R}$ and let $S_i = [l_i, u_i]$ for some $l_i, u_i \in \mathbb{R}$ for each $1 \leq i \leq n$.

Let $x_{(k)}$ denote the k th increasing order statistic of a set $\{x_1, x_2, x_3, \dots\}$. Then, for $y_{n+1} < l_{(1)}$,

$$f(y_{n+1}) = \#\{i \in \{1, \dots, n+1\} : c_i \leq c_{n+1}\} = \#\{(n+1)\} = 1.$$

For $y_{n+1} \in [l_{(1)}, l_{(2)})$,

$$f(y_{n+1}) = \#\{(n+1), \{j : l_j = l_{(1)}\}\} = 2.$$

And so on to $y_{n+1} \in [l_{(n-1)}, l_{(n)})$ where $f(y_{n+1}) = n$. For $y_{n+1} \in [l_{(n)}, u_{(1)}]$, note that $\gamma \in [l_{(n)}, u_{(1)}]$, and $f(y_{n+1}) = n+1$. For $y_{n+1} \in (u_{(1)}, u_{(2)}]$, $f(y_{n+1}) = n$, and so on to $y_{n+1} \in (u_{(n-1)}, u_{(n)}]$ where $f(y_{n+1}) = 2$. Finally, for $y_{n+1} > u_{(n)}$,

$$f(y_{n+1}) = \#\{(n+1)\} = 1.$$

We have shown that, for $y_{n+1} \leq \gamma$, $f(y_{n+1})$ increases stepwise from 1 to $n+1$, and, for $y_{n+1} \geq \gamma$, $f(y_{n+1})$ decreases stepwise from $n+1$ to 1. Therefore, $f(y_{n+1})$ is a step function that takes on ordered values $\{1, 2, \dots, n, n+1, n, \dots, 2, 1\}$ over the domain \mathbb{R} . \square

Proof of Lemma 2. The region in \mathbb{R} where $f(y_{n+1}) > a$ for some $a \in [0, n+1]$ is an interval. Therefore,

$$A(\mathbf{Y}) = \{y_{n+1} \in \mathbb{R} : f(y_{n+1})/(n+1) > \alpha\}$$

is an interval. \square

Proof of Lemma 3. As shown in the proof of Theorem 1,

$$\begin{aligned} C(y_1, y_2) &\leq C(y_2, y_1) \\ \Leftrightarrow C(\{y_1, y_2\}, y_2) &\leq C(\{y_1, y_2\}, y_1) \\ \Leftrightarrow \left(y_1 - \mu/\tau^2 \frac{\tau_{12}^2}{1 - \tau_{12}^2}\right)^2 - \left(y_2 - \mu/\tau^2 \frac{\tau_{12}^2}{1 - \tau_{12}^2}\right)^2 &\leq 0 \end{aligned}$$

Therefore, the inequality reduces to a quadratic function of the unknown candidate y_1 . Label this function h :

$$h(y_1) := \left(y_1 - \mu/\tau^2 \frac{\tau_{12}^2}{1 - \tau_{12}^2} \right)^2 - \left(y_2 - \mu/\tau^2 \frac{\tau_{12}^2}{1 - \tau_{12}^2} \right)^2$$

By standard quadratic theory, we can draw a few conclusions:

1. Notice that $h(y_1)$ is in vertex form. In $h(y_1)$, the leading coefficient is positive, so the parabola will be upward facing.
2. The discriminant obtained via the quadratic formula is $4 \left(y_2 - \mu/\tau^2 \frac{\tau_{12}^2}{1 - \tau_{12}^2} \right)^2$. As the discriminant is non-negative, the solutions to $h(y_1) = 0$ are a repeated real value or 2 unique real values.

By items (1) and (2), the solution to the inequality $h(y_1) \leq 0$ will be an interval. Solving the inequality yields that the interval is of the form

$$[\min\{y_2, g(y_2)\}, \max\{y_2, g(y_2)\}]$$

where $g(y_2) := 2(\mu/\tau^2)\tau_{12}^2(1 - \tau_{12}^2)^{-1} - y_2$. For the remainder of proofs, we will generally assume $y_2 \neq g(y_2)$, as will most likely be the case for continuous data. If $y_2 = g(y_2)$, the FAB conformal prediction region may be a point, depending on the specified error rate. \square

Proof of Lemma 4. Following notation used thus far in proofs, we aim to show

$$\tilde{\theta} \in [\min\{y_2, g(y_2)\}, \max\{y_2, g(y_2)\}]$$

where $\tilde{\theta} := (\mu/\tau^2 + y_2)\tau_{12}^2$ and $g(y_2) := (\mu/\tau^2)\tau_{12}^2(1 - \tau_{12}^2)^{-1} - y_2$. We first consider the case where $y_2 < g(y_2)$.

By the quadratic formula, we conclude the vertex v of the parabola $h(y_1)$ is

$$v := (\mu/\tau^2)\tau_{12}^2(1 - \tau_{12}^2)^{-1}.$$

and $y_2 < v < g(y_2)$. We now prove the result in two steps.

1. First, we show the ordering $y_2 < \tilde{\theta}$:

$$\begin{aligned}
y_2 &< (\mu/\tau^2)\tau_{12}^2(1 - \tau_{12}^2)^{-1} \\
&\Leftrightarrow y_2(1 - \tau_{12}^2) < (\mu/\tau^2)\tau_{12}^2 \\
&\Leftrightarrow y_2 < (\mu/\tau^2)\tau_{12}^2 + y_2\tau_{12}^2 \\
&=: y_2 < \tilde{\theta}.
\end{aligned}$$

2. Additionally, we can show $f(y_2) > \tilde{\theta}$:

$$\begin{aligned}
y_2 &< \tilde{\theta} \\
&:= y_2 < (\mu/\tau^2)\tau_{12}^2 + y_2\tau_{12}^2 \\
&\Leftrightarrow y_2(1 - \tau_{12}^2) < (\mu/\tau^2)\tau_{12}^2 \\
&\Leftrightarrow y_2 < (\mu/\tau^2)\frac{\tau_{12}^2}{(1 - \tau_{12}^2)} \\
&\Leftrightarrow y_2 < (\mu/\tau^2)\left((1 - \tau_{12}^2)^{-1} - 1\right) \\
&\Leftrightarrow y_2 + \mu/\tau^2 < (\mu/\tau^2)(1 - \tau_{12}^2)^{-1} \\
&\Leftrightarrow (y_2 + \mu/\tau^2)\tau_{12}^2 < (\mu/\tau^2)(1 - \tau_{12}^2)^{-1}\tau_{12}^2 \\
&=: \tilde{\theta} < v
\end{aligned}$$

since

$$(1 - \tau_{12}^2)^{-1} - 1 = \frac{1 - (1 - \tau_{12}^2)}{1 - \tau_{12}^2} = \frac{\tau_{12}^2}{1 - \tau_{12}^2}$$

To summarize, given that $y_2 < g(y_2)$, we have shown:

$$y_2 < \tilde{\theta} < v < g(y_2)$$

If $y_2 > g(y_2)$, the ordering of the terms is reversed.

□

Proof of Theorem 2. By Lemmas 3 and 4 each sub-region of acceptance S_i , for $i \in \{1, \dots, n + 1\}$, is an interval which contains $\tilde{\theta}$. Therefore, the hypothesis of Lemma 1 is met, and so, by Lemma 2, the conformal prediction region is an interval which contains $\tilde{\theta}$. Furthermore, by conformal algorithm, the conformal prediction region is

$$A^{fab}(\mathbf{Y}) = \{y_{n+1} \in \mathcal{Y} : \#\{i = 1, \dots, n + 1 : c_i(y_{n+1}) \leq c_{n+1}(y_{n+1})\} > k\},$$

which is the k th and $(2n - k + 1)$ th order statistics of the collection of bounds of the sub-regions of acceptance, $\mathbf{v} = [y_1 \ \cdots \ y_n \ g(y_1) \ \cdots \ g(y_n)]^T$. This result assumes there are no ties in \mathbf{v} . □

Appendix B

Supplementary Material for Chapter 3

B.1 Maximization of the marginal multinomial-Dirichlet likelihood

In this section, we detail a Newton-Raphson algorithm to maximize the marginal log likelihood of a conjugate multinomial-Dirichlet model:

$$\begin{aligned}\mathbf{X}_j &\sim MN_K(\boldsymbol{\theta}_j, N_j), \text{ independently for } j = 1, \dots, J \\ \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_J &\sim \text{Dirichlet}_K(\boldsymbol{\gamma}).\end{aligned}$$

The log likelihood of the marginal likelihood is as follows,

$$\begin{aligned}\mathcal{L}(\boldsymbol{\gamma}) \propto \sum_{j=1}^J \left[\log \Gamma\left(\sum_{i=1}^K \gamma_i\right) - \log \Gamma\left(N_j + \sum_{i=1}^K \gamma_i\right) + \right. \\ \left. \sum_{i=1}^K \log \Gamma(x_{j,i} + \gamma_i) - \sum_{i=1}^K \log \Gamma(\gamma_i) \right].\end{aligned}$$

Define

$$\Psi(s) = \frac{d}{ds} \log \Gamma(s) = -\xi + \sum_{n=0}^{\infty} \left[\frac{1}{n+1} - \frac{1}{n+s} \right],$$

where ξ is the Euler-Mascheroni constant. Then, it is straightforward to obtain the first and second derivatives of the marginal log likelihood,

$$\begin{aligned} \frac{d}{d\gamma_k} &= \sum_{j=1}^J \left[\Psi\left(\sum_{i=1}^K \gamma_i\right) - \Psi\left(N_j + \sum_{i=1}^K \gamma_i\right) + \Psi\left(x_{j,k} + \gamma_k\right) - \right. \\ &\quad \left. \Psi\left(\gamma_k\right) \right] \\ \frac{d}{d\gamma_k^2} &= \sum_{j=1}^J \left[\Psi'\left(\sum_{i=1}^K \gamma_i\right) - \Psi'\left(N_j + \sum_{i=1}^K \gamma_i\right) + \Psi'\left(x_{j,k} + \gamma_k\right) - \right. \\ &\quad \left. \Psi'\left(\gamma_k\right) \right] \\ \frac{d}{d\gamma_k d\gamma_{k'}} &= \sum_{j=1}^J \left[\Psi'\left(\sum_{i=1}^K \gamma_i\right) - \Psi'\left(N_j + \sum_{i=1}^K \gamma_i\right) \right], \end{aligned}$$

where Ψ' is the trigamma function. Let \mathbf{g} be the gradient vector of length K and \mathbf{H} the Hessian matrix. Finally, Newton's method updates γ as follows:

$$\gamma^{(t+1)} = \gamma^{(t)} - \mathbf{H}^{-1}(\gamma^{(t)})\mathbf{g}(\gamma^{(t)}),$$

where the algorithm is iterated until convergence.

B.2 Proofs

Remark 1 (Concerning Theorem 3.). We first elaborate on the construction of an order-based prediction set following Equation 3.3. To test if an element $\mathbf{y}^{(k)}$ in the sample space \mathcal{Y} is included in a prediction set for a given vector \mathbf{o} and known event probability vector $\boldsymbol{\theta}$, the cumulative sum of event probabilities of the categories corresponding to the minimum element of \mathbf{o} up to element k , following the ordering of \mathbf{o} , is computed. If this cumulative sum is greater than the error rate α , then element k is included in the prediction set. As a result, all elements with such

cumulative sums greater than α are included in the prediction set. The elements with such cumulative sums less than or equal to α are not included. Therefore, by construction, $P(\mathbf{Y} \notin A_\alpha^{\theta, \mathbf{o}} | \boldsymbol{\theta}) \leq \alpha$. Consequently,

$$P(\mathbf{Y} \in A_\alpha^{\theta, \mathbf{o}} | \boldsymbol{\theta}) = 1 - P(\mathbf{Y} \notin A_\alpha^{\theta, \mathbf{o}} | \boldsymbol{\theta}) \geq 1 - \alpha,$$

so $A_\alpha^{\theta, \mathbf{o}}$ is α -valid.

Proof of Theorem 3. (1): Note first that \mathbf{o} enters Equation 3.3 only through the ordering of its elements. As such, without loss of generality, consider vectors of the form $\mathbf{o} \in \{0, 1\}^K$. When constructing a prediction set following Equation 3.3 based on such a vector \mathbf{o} , the category space is effectively divided into two disjoint subsets,

$$\mathcal{Y}_0 = \{\mathbf{y}^{(k)} \in \mathcal{Y} : o_k = 0\}$$

$$\mathcal{Y}_1 = \{\mathbf{y}^{(k)} \in \mathcal{Y} : o_k = 1\},$$

such that, by construction,

$$A_\alpha^{\theta, \mathbf{o}} = \begin{cases} \mathcal{Y} & \text{if } \sum_{\{j: \mathbf{y}^{(j)} \in \mathcal{Y}_0\}} \theta_j > \alpha \\ \mathcal{Y}_1 & \text{else} \end{cases}.$$

For a given error rate α , clearly any α -valid prediction set may be constructed under considerations of permutations of the vector \mathbf{o} of the form $\mathbf{o} \in \{0, 1\}^K$.

(2): We wish to show no other ordering results in an α -valid prediction set with strictly smaller cardinality than \mathbf{o}^θ . In words, consider switching the ordering of one element at a time. This will always result in a prediction set with the same cardinality or greater cardinality than that under \mathbf{o}^θ . More formally, let $\tilde{\mathbf{o}} = \{o_1^\theta, o_2^\theta, \dots, o_{k^*}^\theta, o_{k^*-1}^\theta, o_{k^*+1}^\theta, \dots, o_K^\theta\}$, that is, equivalent to \mathbf{o}^θ with the (k^*) th and $(k^* - 1)$ th ordering flipped. But, by construction, $\theta_{k^*} \geq \theta_{k^*-1}$, so $|A_\alpha^{\theta, \mathbf{o}^\theta}| \leq |A_\alpha^{\theta, \tilde{\mathbf{o}}}|$. \square

Proof of Theorem 4. Let $\mathbf{Y}_1, \dots, \mathbf{Y}_{N+1} \sim \text{i.i.d. } MN_K(\boldsymbol{\theta}, 1)$. Then, we wish to construct a conformal prediction set for \mathbf{Y}_{N+1} based on an observation of $\mathbf{X} = \sum_{i=1}^N \mathbf{Y}_i$, and some conformity measure C .

To determine if a candidate category $k \in \{1, \dots, K\}$ is included in a $1 - \alpha$ conformal prediction set, the conformal algorithm proceeds as follows, see Section 2.1 of Bersson and Hoff (2022) for more details:

1. Set $\mathbf{y}_{N+1} = \mathbf{y}^{(k)}$ where $\mathbf{y}^{(k)}$ is a vector of length K with a 1 in the k^{th} index and 0s elsewhere.
2. For $j = 1, \dots, N + 1$, compute conformity scores $c_j = C(\mathbf{x} - \mathbf{y}_j + \mathbf{y}^{(k)}, \mathbf{y}_j)$.

3. Set

$$p_k = \frac{\#\{j \in \{1, \dots, N + 1\} : c_{N+1} \geq c_j\}}{N + 1}$$

More compactly, and by symmetry in the problem, this conformal p -value may be equivalently written as:

$$p_k = \sum_{l=1}^K \mathbb{1}(o_k^* \geq o_l^*) \frac{x_l + y_l^{(k)}}{N + 1},$$

where $o_k^* = c_{N+1}$ and $o_l^* = c_j$ for a $j \in \{1, \dots, N\}$ such that $\mathbf{y}_j = \mathbf{y}^{(l)}$. Then, for prediction mis-coverage rate α , the category k is included in the prediction set if $p_k > \alpha$. A prediction set constructed from this procedure may be concisely written as follows:

$$A_\alpha(\mathbf{x}) = \left\{ \mathbf{y}^{(k)} \in \mathcal{Y} : \left[\sum_{l=1}^K \mathbb{1}(o_k \geq o_l) \frac{x_l + y_l^{(k)}}{N + 1} \right] > \alpha \right\},$$

for some $\mathbf{o} \in \mathbb{R}^K$.

□

Appendix C

Supplementary Material for Chapter 4

C.1 Metropolis-Hastings Algorithm for SWAG

1. Sample $(\lambda, \mathbf{U}) \sim p(w, \mathbf{U} | \mathbf{Y}, \boldsymbol{\theta}_{-\lambda, -\mathbf{U}})$ as follows:

(a) sample $\lambda \sim p(\lambda | \mathbf{Y}, \boldsymbol{\theta}_{-\mathbf{U}, -\lambda})$ using a Metropolis step as follows:

- i. obtain a sample λ^* from a reflecting random walk based on the previous iteration's value of λ , that is, sample an initial value λ^* from $\lambda^* \sim \text{Uniform}(\lambda - \delta_\lambda, \lambda + \delta_\lambda)$, and utilize the following reassignment schema to ensure the sample has the correct support:

$$\lambda^* = \begin{cases} \lambda^* & \text{if } \lambda^* \in (0, 1) \\ |\lambda^*| & \text{if } \lambda^* \leq 0 \\ 2 - \lambda^* & \text{if } \lambda^* \geq 1, \end{cases}$$

- ii. compute the Metropolis acceptance ratio

$$r = \frac{p(\lambda^* | \mathbf{Y}, \boldsymbol{\theta}_{-\mathbf{U}, -\lambda})}{p(\lambda | \mathbf{Y}, \boldsymbol{\theta}_{-\mathbf{U}, -\lambda})} = \prod_{j=1}^J \frac{p(Y_j | \Psi_j, \Lambda_j, \lambda = \lambda^*) p(\lambda = \lambda^*)}{p(Y_j | \Psi_j, \Lambda_j, \lambda = \lambda) p(\lambda = \lambda)},$$

- iii. accept λ^* as the updated value for λ with probability r , and

(b) sample $U_j | \mathbf{Y}, \boldsymbol{\theta}_{-U_j} \sim N_{n_j \times p}(M_j, S_j \otimes I_{n_j})$ for each $j \in \{1, \dots, J\}$, where

$$S_j = \left(\Psi_j^{-1} + \frac{\lambda}{1-\lambda} \Lambda_j^{-1} \right)^{-1}$$

$$M_j = \frac{\lambda^{1/2}}{1-\lambda} Y_j \Lambda_j^{-1} S_j.$$

2. Sample $(\nu, \Psi) \sim p(\nu, \Psi | \mathbf{Y}, \boldsymbol{\theta}_{-\Psi, -\nu})$ as follows:

(a) sample ν from a Metropolis step as follows:

- i. obtain a sample ν^* from a reflecting random walk based on the previous iteration's value of ν , that is, sample an initial value ν^* from $\nu^* \sim \text{Uniform}(\nu - \delta_\nu, \nu + \delta_\nu)$, and utilize the following reassignment schema to ensure the sample has the correct support:

$$\nu^* = \begin{cases} \nu^* & \text{if } \nu^* \geq p+2 \\ (p+2) + (p+2 - \nu^*) & \text{if } \nu^* < p+2, \end{cases}$$

- ii. compute the Metropolis acceptance ratio using a prior distribution q_ν , and

$$r = \prod_{j=1}^J \frac{p(U_j | \Psi_0, \nu = \nu^*) q_\nu(\nu = \nu^*)}{p(U_j | \Psi_0, \nu = \nu) q_\nu(\nu = \nu)}$$

where $U_j | \Psi_0, \nu \sim T_{n_j \times p}(\nu - p + 1, 0, \Psi_0(\nu - p - 1) \otimes I_{n_j})$,

- iii. accept ν^* as the updated value for ν with probability r , and

(b) sample $\Psi_j | \mathbf{Y}, \boldsymbol{\theta}_{-\Psi_j} \sim IW_p((U_j^T U_j + (\nu - p - 1)\Psi_0)^{-1}, \nu + n_j)$ for each $j \in \{1, \dots, J\}$.

3. Sample $(\gamma, \Lambda) \sim p(\gamma, \Lambda | \mathbf{Y}, \boldsymbol{\theta}_{-\Lambda, -\gamma})$ as follows:

(a) sample γ from a Metropolis step as follows:

- i. obtain a sample γ^* from a reflecting random walk based on the previous iteration's value of γ and the initial value drawn from $\gamma^* \sim \text{Uniform}(\gamma - \delta_\gamma, \gamma + \delta_\gamma)$,
- ii. compute the Metropolis acceptance score using a prior distribution q_γ , and

$$r = \prod_{j=1}^J \frac{p(Y_j|\lambda, U_j, R_j, C_j, \gamma = \gamma^*) q_\gamma(\gamma = \gamma^*)}{p(Y_j|\lambda, U_j, R_j, C_j, \gamma = \gamma) q_\gamma(\gamma = \gamma)}$$

where

$$Y_j|\lambda, R_j, C_j \sim T_{n_j \times p}(\gamma - p + 1, \lambda^{1/2} U_j, \\ (1 - \lambda) (C_j \otimes R_j) (\gamma - p - 1) \otimes I_{n_j}),$$

- iii. accept γ^* as the updated value for γ with probability r , and
- (b) sample $\Lambda_j|\mathbf{Y}, \boldsymbol{\theta}_{-\Lambda_j} \sim IW_p((\tilde{Y}_j^T \tilde{Y}_j / (1 - \lambda) + (C_j \otimes R_j) (\gamma - p - 1))^{-1}, \gamma + n_j)$ in parallel for each $j \in \{1, \dots, J\}$ where $\tilde{Y}_j = (Y_j - w^{1/2} U_j)$.
4. Sample $\Psi_0|\mathbf{Y}, \boldsymbol{\theta}_{-\Psi_0} \sim W_p(((\nu - p - 1) \sum_{j=1}^J \Psi_j^{-1} + (P_2 \otimes P_1)^{-1} \xi)^{-1}, \xi + J\nu)$.
5. Sample $R_j \sim W_{p_1}(((\gamma - p - 1) \sum_{k=1}^p L_k C_j L_k^T + R_0^{-1} \eta_1)^{-1}, \eta_1 + \gamma p_2)$ where $L_k = \text{vec}^{-1}(l_k)$ from $\Lambda_j^{-1} = \tilde{L} \tilde{L}^T = \sum_{k=1}^p l_k l_k^T$, for each $j \in \{1, \dots, J\}$.
6. Sample $C_j \sim W_{p_2}(((\gamma - p - 1) \sum_{k=1}^p L_k^T R_j L_k + C_0^{-1} \eta_2)^{-1}, \eta_2 + \gamma p_1)$ where $L_k = \text{vec}^{-1}(l_k)$ from $\Lambda_j^{-1} = \tilde{L} \tilde{L}^T = \sum_{k=1}^p l_k l_k^T$, for each $j \in \{1, \dots, J\}$.
7. sample ξ from a Metropolis step as follows:
 - (a) obtain a sample ξ^* from a reflecting random walk based on the previous iteration's value of ξ and the initial value drawn from $\xi^* \sim \text{Uniform}(\xi - \delta_\xi, \xi + \delta_\xi)$,

- (b) compute the Metropolis acceptance score using a prior distribution q_ξ ,
and

$$r = \prod_{j=1}^J \frac{p(\Psi_0 | P_1, P_2, \xi = \xi^*) q_\xi(\xi = \xi^*)}{p(\Psi_0 | P_1, P_2, \xi = \xi) q_\xi(\xi = \xi)},$$

- (c) accept ξ^* as the updated value for ξ with probability r .

8. Sample $P_1 \sim IW_{p_1} \left(\left(\xi \sum_{k=1}^p L_k P_2^{-1} L_k^T + P_{01}(\eta_3 - p_1 - 1) \right)^{-1}, \eta_3 + \xi p_2 \right)$ where $L_k = \text{vec}^{-1}(l_k)$ from $\Psi_0^{-1} = \tilde{L} \tilde{L}^T = \sum_{k=1}^p l_k l_k^T$.
9. Sample $P_2 \sim IW_{p_2} \left(\left(\xi \sum_{k=1}^p L_k^T P_1^{-1} L_k + P_{02}(\eta_4 - p_2 - 1) \right)^{-1}, \eta_4 + \xi p_1 \right)$. where $L_k = \text{vec}^{-1}(l_k)$ from $\Psi_0^{-1} = \tilde{L} \tilde{L}^T = \sum_{k=1}^p l_k l_k^T$.

Bibliography

- Afshartous, D. and De Leeuw, J. (2005), “Prediction in multilevel models,” *Journal of Educational and Behavioral Statistics*, 30, 109–139.
- Arnold, Z. J., Wenger, S. J., and Hall, R. J. (2021), “Not just trash birds: Quantifying avian diversity at landfills using community science data,” *PLoS ONE*, 16.
- Balasubramanian, V. N., Ho, S.-S., and Vovk, V. (2014), *Conformal Prediction for Reliable Machine Learning: Theory, Adaptations and Applications*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1 edn.
- Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014), *Hierarchical modeling and analysis for spatial data*, Chapman and Hall/CRC, 2 edn.
- Barnard, J., Mcculloch, R., and Meng, X.-L. (2000), “Modeling covariance matrices in terms of standard deviations and correlations, with applications to shrinkage,” *Statistica Sinica*, 10, 1281–1311.
- Bersson, E. and Hoff, P. D. (2022), “Optimal conformal prediction for small areas,” Tech. rep.
- Bersson, E. and Hoff, P. D. (2023a), “Bayesian covariance estimation for multi-group matrix-variate data,” Tech. rep.
- Bersson, E. and Hoff, P. D. (2023b), “Frequentist prediction sets for species abundance using indirect information,” Tech. rep.
- Box, G. E. P. (1949), “A General Distribution Theory for a Class of Likelihood Criteria,” *Biometrika*, 36, 317–346.
- Brown, P. J., Fearn, T., and Haque, M. S. (1999), “Discrimination with many variables,” *Journal of the American Statistical Association*, 94, 1320–1329.
- Bryan, J. G. and Hoff, P. D. (2023), “Smaller p-values in genomics studies using distilled auxiliary information,” *Biostatistics*, 24, 193–208.
- Burris, K. C. and Hoff, P. D. (2020), “Exact adaptive confidence intervals for small areas,” *Journal of Survey Statistics and Methodology*, 8, 206–230.

- Camerini, G. and Groppali, R. (2014), “Landfill restoration and biodiversity: A case of study in Northern Italy,” *Waste Management and Research*, 32, 782–790.
- Carlin, B. P. and Gelfand, A. E. (1990), “Approaches for empirical bayes confidence intervals,” *Journal of the American Statistical Association*, 85, 105–114.
- Christensen, M. F. and Hoff, P. D. (2022), “A nonstationary spatial covariance model for data on graphs,” Tech. rep.
- Christensen, R. (2011), *Plane Answers to Complex Questions*, Springer, New York, fourth edn.
- Daniels, M. J. (2006), “Bayesian modeling of several covariance matrices and some results on propriety of the posterior for linear regression with correlated and/or heterogeneous errors,” *Journal of Multivariate Analysis*, 97, 1185–1207.
- Daniels, M. J. and Kass, R. E. (1999), “Nonconjugate Bayesian Estimation of Covariance Matrices and its Use in Hierarchical Models,” *Journal of the American Statistical Association*, 94, 1254–1263.
- Dawid, A. P. (1981), “Some matrix-variate distribution theory: notational considerations and a Bayesian application,” *Biometrika*, 68, 265–274.
- Dunn, R., Wasserman, L., and Ramdas, A. (2022), “Distribution-Free Prediction Sets for Two-Layer Hierarchical Models,” *Journal of the American Statistical Association*, Ahead-of-p, 1–12.
- Faulkenberry, G. D. (1973), “A method of obtaining prediction intervals,” *Journal of the American Statistical Association*, 68, 433–435.
- Fay, R. E. and Herriot, R. A. (1979), “Estimates of income for small places: an application of James-Stein procedures to census data,” *Journal of the American Statistical Association*, 74, 269–277.
- Flury, B. K. (1987), “Two generalizations of the common principal component model,” *Biometrika*, 74, 59–69.
- Friedman, J., Hastie, T., and Tibshirani, R. (2008), “Sparse inverse covariance estimation with the graphical lasso,” *Biostatistics*, 9, 432–441.
- Friedman, J. H. (1989), “Regularized discriminant analysis,” *Journal of the American Statistical Association*, 84, 165–175.
- Gamerman, A., Vovk, V., and Vapnik, V. (1998), *Learning by Transduction*, Morgan Kaufmann Publishers Inc., Madison, Wisconsin.
- Geary, R. C. (1954), “The Contiguity Ratio and Statistical Mapping,” *The Incorporated Statistician*, 5, 115–127.

- Gelman, A. (2006), “Multilevel (hierarchical) modeling: what it can and cannot do,” *Technometrics*, 48, 432–435.
- Gelman, A. and Hill, J. (2006), *Data Analysis Using Regression and Multi-level/Hierarchical Models*, Cambridge University Press, Cambridge, 1 edn.
- Greene, T. and Rayens, W. S. (1989), “Partially pooled covariance matrix estimation in discriminant analysis,” *Communications in Statistics - Theory and Methods*, 18, 3679–3702.
- Hammel, S. C., Phillips, A. L., Hoffman, K., and Stapleton, H. M. (2018), “Evaluating the Use of Silicone Wristbands To Measure Personal Exposure to Brominated Flame Retardants,” *Environ. Sci. Technol.*, 52, 11875–11885.
- Hoff, P. (2023), “Bayes-optimal prediction with frequentist coverage control,” *Bernoulli*, 29, 901–928.
- Hoff, P., McCormack, A., and Zhang, A. R. (2022), “Core shrinkage covariance estimation for matrix-variate data,” *Journal of the Royal Statistical Society, Series B*.
- Hoff, P. D. (2009a), *A First Course in Bayesian Statistical Methods*, Springer Texts in Statistics, New York.
- Hoff, P. D. (2009b), “A hierarchical eigenmodel for pooled covariance estimation,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71, 971–992.
- Hoffman, K., Hammel, S. C., Phillips, A. L., Lorenzo, A. M., Chen, A., Calafat, A. M., Ye, X., Webster, T. F., and Stapleton, H. M. (2018), “Biomarkers of exposure to SVOCs in children and their demographic associations: The TESIE Study,” *Environment International*, 119, 26–36.
- James-Todd, T. M., Meeker, J. D., Huang, T., Hauser, R., Seely, E. W., Ferguson, K. K., Rich-Edwards, J. W., and McElrath, T. F. (2017), “Racial and ethnic variations in phthalate metabolite concentration changes across full-term pregnancies,” *Journal of Exposure Science and Environmental Epidemiology*, 27, 160–160.
- Janicki, R., Raim, A. M., Holan, S. H., and Maples, J. J. (2022), “Bayesian non-parametric multivariate spatial mixture mixed effects models with applications to american community survey special tabulations,” *The Annals of Applied Statistics*, 16, 144–168.
- Johnson, B., Tateishi, R., and Xie, Z. (2012), “Using geographically weighted variables for image classification,” *Remote Sensing Letters*, 3, 491–499.

- Lehmann, E. and Casella, G. (1998), *Theory of Point Estimation*, Springer, New York, 2 edn.
- Lei, J. and Wasserman, L. (2014), “Distribution-free prediction bands for non-parametric regression,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76, 71–96.
- Lele, S. R. (2020), “How should we quantify uncertainty in statistical inference?” *Frontiers in Ecology and Evolution*, 8.
- Ligges, U., Krey, S., Mersmann, O., and Schnackenberg, S. (2018), “tuneR: Analysis of Music and Speech.” Tech. rep.
- Lu, N. and Zimmerman, D. L. (2005), “The likelihood ratio test for a separable covariance matrix,” *Statistics and Probability Letters*, 73, 449–457.
- Mardia, K., Kent, J., and Bibby, J. (1979), *Multivariate Analysis*, Academic Press, London, 1 edn.
- Marks, S. and Dunn, O. J. (1974), “Discriminant functions when covariance matrices are unequal,” *Journal of the American Statistical Association*, 69, 555–559.
- McCormack, A. and Hoff, P. D. (2023), “Tests of linear hypotheses using indirect information,” *Canadian Journal of Statistics*, 51, 852–876.
- Molina, I., Nandram, B., and Rao, J. N. K. (2014), “Small area estimation of general parameters with application to poverty indicators: A hierarchical Bayes approach,” *Annals of Applied Statistics*, 8, 852–885.
- Nandram, B. and Choi, J. W. (2010), “A Bayesian analysis of body mass index data from small domains under nonignorable nonresponse and selection,” *Journal of the American Statistical Association*, 105, 120–135.
- Papadopoulos, A., Vovk, V., and Gammerman, A. (2011), “Regression conformal prediction with nearest neighbours,” *Journal of Artificial Intelligence Research*, 40, 815–840.
- Pfeffermann, D. (2013), “New important developments in small area estimation,” *Statistical Science*, 28, 40–68.
- Pratesi, M. and Salvati, N. (2008), “Small area estimation: the EBLUP estimator based on spatially correlated random area effects,” *Statistical Methods & Applications*, 17, 113–141.
- Price, P. N., Nero, A. V., and Gelman, A. (1996), “Bayesian prediction of mean indoor radon concentrations for Minnesota counties,” *Health Physics*, 71, 922–936.

- Rao, J. N. K. and Molina, I. (2015), *Small Area Estimation*, John Wiley and Sons, Inc., New York, NY, 2 edn.
- Rayens, W. and Greene, T. (1991), “Covariance pooling and stabilization for classification,” *Computational Statistics and Data Analysis*, 11, 17–42.
- Shafer, G. and Vovk, V. (2008), “A tutorial on conformal prediction,” *Journal of Machine Learning Research*, 9, 371–421.
- Singh, B. B., Kant Shukla, G., and Kundu, D. (2005), “Spatio-temporal models in small area estimation,” *Survey Methodology*, 31, 183–195.
- Sinha, S. K. and Rao, J. N. K. (2009), “Robust small area estimation,” *The Canadian Journal of Statistics*, 37, 381–399.
- Skrondal, A. and Rabe-Hesketh, S. (2009), “Prediction in multilevel generalized linear models,” *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172, 659–687.
- Sullivan, B., Wood, C., Iliff, M., Bonney, R., Fink, D., and Kelling, S. (2009), “eBird: a citizen-based bird observation network in the biological sciences,” *Biological Conservation*, 142, 2282–2292.
- Sullivan, B. L., Phillips, T., Dayer, A. A., Wood, C. L., Farnsworth, A., Iliff, M. J., Davies, I. J., Wiggins, A., Fink, D., Hochachka, W. M., Rodewald, A. D., Rosenberg, K. V., Bonney, R., and Kelling, S. (2017), “Using open access observational data for conservation action: A case study for birds,” *Biological Conservation*, 208, 5–14.
- Tang, B., Clark, J. S., Marra, P. P., and Gelfand, A. E. (2023), “Modeling community dynamics through environmental effects, species interactions and movement,” *Journal of Agricultural, Biological, and Environmental Statistics*, 28, 178–195.
- Thatcher, A. R. (1964), “Relationships between Bayesian and confidence limits for predictions,” *Journal of the Royal Statistical Society. Series B (Methodological)*, 26, 176–210.
- Tian, Q., Nordman, D. J., and Meeker, W. Q. (2022), “Methods to compute prediction intervals: A review and new results,” *Statistical Science*, 37, 580–597.
- Tibshirani, R. J., Foygel Barber, R., Candès, E. J., and Ramdas, A. (2019), “Conformal Prediction Under Covariate Shift,” *33rd Conference on Neural Information Processing Systems*.
- US Environmental Protection Agency (1992), “National residential radon survey: summary report,” Tech. rep., US Environmental Protection Agency, Washington, DC.

- Vidoni, P. (2006), “Response prediction in mixed effects models,” *Journal of Statistical Planning and Inference*, 136, 3948–3966.
- Vovk, V., Gammerman, A., and Shafer, G. (2005), *Algorithmic Learning in a Random World*, Springer US.
- Vovk, V., Shen, J., Manokhin, V., and Xie, M.-G. (2019), “Nonparametric predictive distributions based on conformal prediction,” *Machine Learning*, 108, 445–474.
- Warden, P. (2017), “Speech commands: A public dataset for single-word speech recognition.” .
- Warden, P. (2018), “Speech commands: A dataset for limited-vocabulary speech recognition,” Tech. rep.
- Wu, W. B. and Pourahmadi, M. (2009), “Banding sample autocovariance matrices of stationary processes,” *Statistica Sinica*, 19, 1755–1768.
- Yu, C. and Hoff, P. D. (2018), “Adaptive multigroup confidence intervals with constant coverage,” *Biometrika*, 105, 319–335.

Biography

Elizabeth Bersson completed her undergraduate and masters degrees at The University of Alabama, earning a Bachelor of Science in Mathematics and a Masters of Science in Applied Statistics in May 2017. Following her studies at Alabama, Elizabeth spent two years as a Research Assistant in the Macroeconomic and Quantitative Studies group at The Federal Reserve Board in Washington, D.C. where she engaged in macroeconomic theory research, primarily under the supervision of Dr. Matthias Paustian. In the fall of 2019, she joined the PhD program in the Department of Statistical Science at Duke University. Following graduation in May 2024, Elizabeth will join Dr. Tamara Broderick's group as a Postdoctoral Associate within the Laboratory for Information and Decision Systems at Massachusetts Institute of Technology.