The Pennsylvania State University
The Graduate School

NOVEL METHODS AND APPLICATION OF BAYESIAN HIERARCHICAL REGRESSION MODELS

A Dissertation in
Statistics and Social Data Analytics
by
Amy Xiang Zhang

© 2021 Amy Xiang Zhang

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

December 2021

The dissertation of Amy Xiang Zhang was reviewed and approved by the following:

Le Bao
Professor of Statistics
Dissertation Advisor
Chair of Committee

David Hunter
Professor of Statistics

Xiaoyue Niu
Professor of Statistics

C. Lee Giles
Professor of Information Sciences and Technology

Michael J. Daniels
Professor of Statistics, University of Florida
Special Member

Ephraim Hanks
Professor of Statistics
Chair of Graduate Studies

Abstract

We present two novel methods and one application study for Bayesian hierarchical regression models (BHRMs). BHRMs are a large, flexible class of probabilistic models and range in complexity from one-way linear regressions to Gaussian Markov random fields. We demonstrate their efficacy on an application study of HIV prevalence in certain key sub-populations that are particularly vulnerable to HIV. The data are heavily imbalanced and often have small sample sizes. We show that a BHRM which pools information across key populations to model the HIV prevalence trend over time reduces predictive error in comparison to modeling the HIV prevalence trend separately within key populations. We compare different levels of pooling, with partial pooling providing the greatest flexibility while also maintaining a low predictive error under leave-a-cluster-out cross-validation, where all observations related to a unit are removed in turn as in, for example, models with repeated measures.

Although information pooling is often credited for the good performance of Bayesian hierarchical models, it has not been explicitly quantified except in the simplest of one-way settings. We propose a novel method which explicitly quantifies information pooling and shrinkage for all regression models, via what we call the borrowing factors. The borrowing factors decompose regression model estimates into weights placed on clusters of data. The weights are informed only by the model definition and data availability and thus can be used to explicitly link the effects of data imbalance and model assumptions to actual model estimates. We also provide a metric to identify point estimates which rely heavily on specific clusters of data, called SSBF. We present theoretical properties of SSBF and the borrowing factors and demonstrate their usage on two examples.

BHRMs can be computationally expensive to fit. Cross-validation (CV) is therefore not a common practice to evaluate the predictive performance of BHRMs. We present a novel method which circumvents the need to re-run computationally costly estimation methods for each cross-validation fold and makes CV more feasible for large BHRMs. By conditioning on the estimated variance-covariance parameters using all the data, we shift the CV problem from probability-based sampling to a simple and familiar optimization problem. In many cases, this produces estimates which are equivalent to full CV. This method is an extension of the borrowing factors to cross-validation. We provide theoretical results, a diagnostic for finite sample sizes, and demonstrate its efficacy on publicly available data.

Table of Contents

List of Figures

List of Tables

List of Symbols

p	Success probability for Binomial distribution, p. 7
$f_k(t)$	k^{th} spline function evaluated at time t, p. 7
CRPS	Continuous ranked probability score, p. 8
F_p	Posterior predictive distribution, p. 8
F	Cumulative distribution function for the standard normal, p. 13
MSS	Mean sum squares, p. 13
$\boldsymbol{Y}$	Vector of response data, p. 22
N	The dimension of Y, p. 22
$\boldsymbol{\beta}$	Vector of coefficients in a regression, p. 22
$\boldsymbol{\beta_1}$, $\boldsymbol{\beta_2}$	Vector of fixed and random effects, respectively, p. 22
P	The dimension of $\boldsymbol{\beta}$, p. 22
P_1, P_2	The dimensions of $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, respectively, p. 22
X	Design matrix of a regression, p. 22
X_1, X_2	Portion of X corresponding to the fixed and random effects, respectively, p. 22
Σ, σ^2	Covariance (Σ) or variance (σ^2, if i.i.d.) of $\boldsymbol{\beta}_2$, p. 22
Φ, ϕ^2	Diagonal matrix of variances for $\boldsymbol{Y}$ (Φ), p. 22
$\mathbf{1}$	The vector of ones, p. 22
V	The conditional posterior covariance of β given $\Sigma, \Phi, \boldsymbol{Y}$, p. 23
$\hat{\Sigma}$	The full-data posterior mean $E[\Sigma\|Y]$, p. 23

$\hat{\Phi}$ The full-data posterior mean $E[\Phi|Y]$, p. 23

$\hat{Y}$ Fitted values for regression model estimates, $E[X\beta|Y]$, p. 24

W The matrix of individual borrowing factors, p. 24

w_{ij} The $(i,j)^{th}$ entry of W, p. 24,27

b_{iJ} The borrowing factor for $\hat{Y}_i$ over $\boldsymbol{Y}_J$, p. 24

x_i' The i^{th} row of design matrix X, p. 25

ϕ_i' The i^{th} entry in the diagonal of Φ, p. 25

B_i The borrower cluster for point estimate $\hat{Y}_i$, p. 25

b_{iB_i} The shrinkage factor for point estimate $\hat{Y}_i$, p. 25

L_i The lenders for point estimate $\hat{Y}_i$, p. 25

b_{iL_i} The pooling factor for point estimate $\hat{Y}_i$, p. 25

SSBF_i The sum square of borrowing factors for $\hat{Y}_i$, p. 25

PSSBF_{iJ} Partial SSBF for $\hat{Y}_i$ over J, p. 26

λ_i^* Historically, the pooling factor for one-way models, p. 28

n_i Dimension of $\boldsymbol{Y}_i$, e.g. p. 28

ρ_{ij} Borrowing factor over Y_j for $\hat{Y}_i$, in the one-way model, p. 28

τ_j Contribution $\bar{Y}_j$ to a_0, in the one-way model, p. 28

S_i Peña's influence analysis metric, p. 32

b_{kj} The shrinkage factor for the Radon example, p. 32

$b_{k'j}$ Borrowing factor over same-county lenders for the Radon example, p. 32

$b_{kj'}$ Borrowing factor over same-basement lenders for the Radon example, p. 32

$b_{k'j'}$ Borrowing factor over different-basement, different-county lenders for the Radon example, p. 32

$\hat{\mu}_{kj}$ The Radon data point estimate, conditional on a_1, $\hat{Y}_{kj} - \hat{a}_1 u_j$, p. 34

A Adjacency matrix, p. 35

η Parameter for Poisson generalized linear model, p. 35

ρ_J Spatial dependence parameter, p. 35

ρ_T Temporal dependence parameter, p. 35

Q Covariance matrix function, p. 35

j_0 Lenders in the same intermediate geography (IG), p. 36

j_1 Lenders in the neighboring IGs, p. 36

j_{2+} Lenders not in the same or neighboring IG, p. 36

t_0 Lenders in the same year, p. 36

t_1 Lenders in adjacent years, p. 36

t_{2+} Lenders 2 or more years away, p. 36

$\boldsymbol{Y}_{-j}$ The vector of response data, excluding $\boldsymbol{Y}_j$, p. 44

X_{-j} The design matrix, excluding rows corresponding to $\boldsymbol{Y}_j$, p. 44

Φ_{-j} Diagonal matrix of variances for $\boldsymbol{Y}_{-j}$, p. 44

$\hat{Y}_j^{\text{AXE}}$ AXE approximation for $E[Y_j|Y_{-j}]$, p. 44

f_η Link function for a generalized linear model, p. 45

ν Variance for a generalized linear model, p. 45

$\tilde{Y}$ Transformed response vector $f_\eta Y$, p. 45

θ Vector of random intercepts which defines the LCO-CV folds, p. 46

μ Vector of fitted values without θ, $X\beta - \theta$, p. 46

$\tilde{\Sigma}$ Posterior mean for Σ over the training data, $E[\Sigma|Y_{-j}]$, p. 46

$\tilde{\phi}$ Posterior mean for ϕ over the training data, $E[\phi|Y_{-j}]$, p. 46

LRR_j Log RMSE ratio for j^{th} CV fold, p. 48

$\tilde{\theta}$ Approximation for sample from $\theta|Y_{-j}$ under Ghosting, p. 48

$w_j^{(s)}$ Importance sample weight for MC sample s and test data Y_j, p. 49

α Data scaling factor, p. 51

δ Test data proportion, p. 52

Acknowledgments

There are many people whose advice and support helped me get to this point. Without the guidance of Dr. Le Bao and Dr. Michael Daniels, this thesis may not exist. Their discerning eyes improved the work many times over and they taught me how to make an academic argument. I am thankful as well to Drs. Maggie Niu, David Hunter, and Lee Giles for their feedback, both during my comprehensives examination and since.

Many collaborators made the work present here and in other publications possible. In particular I would like to thank Dr. Maggie Niu, whose work on the sub-national area models for HIV prevalence is where this thesis began; Dr. Diane Felmlee, who taught me how to work across disciplines; and Dr. Keith Sabin, who hosted my internship at UNAIDS.

I am very thankful to the faculty and staff at the Department of Statistics at Penn State for welcoming me at a time when I knew very little about statistics. I am indebted to my fellow graduate students for their patience in the early years, answering my questions and explaining homework problems. It is a huge credit to them and the department that I was able to thrive here. I am also very thankful to Dr. Burt Monroe and Dr. Bruce Desmarais of the Social Data Analytics program at Penn State for accepting me into their incredible dual-title program. I learned so much and I am grateful for my time in the program.

This work was supported by the National Institutes of Health (NIH) under grants R56AI120812-01A1 and R01 AI136664-01, and the National Science Foundation (NSF) under IGERT grant DGE-1144860, Big Data Social Science. Findings and conclusions do not necessarily reflect the view of the NIH or NSF.

Dedication

To my family who suffered through holidays with me on my laptop; to my dog Dunphy who reminded me to take breaks from my laptop; to my partner Parth whose advice was always to follow my gut (and who always ensured my gut was well-fed)—thank you for your love and support.

Chapter 1 Introduction

Bayesian hierarchical models (BHMs) are often used for their ability to model complex dependence structures while producing probabilistic uncertainty estimates. Their flexible framework, in many cases requiring only that posterior samples can be drawn via, e.g., Gibbs sampling, means that they can be used in situations that may be less feasible under maximum likelihood methods, such as Bayesian additive regression trees or L_0 norms. By jointly modeling the response Y and data X, $f(Y, X)$, they allow for the generation of new data which reflects the properties and total variation of both X and Y, and so BHMs are often referred to as generative models in the machine learning literature (Bernardo et al., 2007). Historically, BHMs have been shown to perform well in situations where data are scarce and imbalanced, a property that is typically attributed to pooling of information and shrinkage of model estimates to some global mean, as with Stein's estimator (Efron and Morris, 1975; Morris, 1983; Stein, 1956).

Bayesian hierarchical regrssion models (BHRMs) are likely the most commonly used Bayesian model and consist of a wide variety of methods, from simple linear regressions where observations $Y \sim N(X\beta, \phi^2)$ for model matrix X and coefficient vector β, and Naïve Bayes classification to Gaussian processes and Gaussian Markov random fields where $\theta \in \beta \sim N(0, \Sigma)$, for covariance matrix Σ. They include all models where Gaussian hyperpriors are placed on mean parameters and as such have more recently been called "latent Gaussian variable models" by Vehtari et al. (2016b). Here, we present one application and two novel methods for BHRMs, with accompanying R packages and a Shiny app.

Chapter 2 contains an application of BHRMs to estimating HIV prevalence within key hard-to-reach sub-populations. In most mid- and low-income countries, data for the general population are relatively rich, but for key sub-populations that are often disproportionately impacted by HIV (e.g. sex workers, intravenous drug users, and men

who have sex with men), data tend to be sparse and highly imbalanced. Without sufficient data for these key populations, designing and implementing efficient interventions to reduce HIV prevalence is difficult. By pooling information across the key populations, we can improve the estimation of HIV prevalence trends. We compare three BHRMs with different levels of information pooling: no pooling (the model is fit separately within each key population), partial pooling, and complete pooling. The models are evaluated using leave-a-cluster-out cross-validation.

Although information pooling is typically credited for the good performance of BHRMs, it has only been explicitly quantified in the simplest of one-way settings, where the data $Y_i \sim N(\alpha_i, \phi_i^2)$ and $\alpha_i \sim N(a_0, \sigma^2)$, where all parameters α_i, ϕ_i, a_0, and σ are scalar. We may have some intuition as to how information is pooled to obtain model estimates; for example, if the data are imbalanced, we expect those clusters with less data to borrow more and for that borrowing to come largely from clusters with more data, but this has not been explicitly quantified for any other model. Typically, the effects of data imbalance on model estimates are understood through simulation, rather than through any property of the model. In Chapter 3, we propose to quantify information borrowing (and therefore information pooling) by decomposing regression model estimates into a vector of weights placed on each data point. This offers the intuitive interpretation that estimates are formed by "borrowing" information from other sets of points, with the weight being the amount borrowed. As such, we call the weights "borrowing factors". The borrowing factors explicitly quantify the impact of data imbalance on model estimates, the impact of high-leverage points, and the degree and manner of information borrowing between point estimates. By understanding at a more granular level how model estimates are produced, we can better diagnose issues with model fit, subject matter considerations, data properties, and model assumptions.

Historically, information borrowing has been understood through the lens of the James-Stein estimator. Given observed data $Y \sim N(\theta, \sigma^2 I)$, Stein (1956) developed a biased estimator which improves upon the unbiased ordinary least squares (OLS) estimator for $\theta \in \mathbb{R}^P$ $P \geq 3$, under squared loss. This result was later improved by James and Stein (1992) and the estimator was dubbed the "James-Stein estimator". Its relation to Bayesian models was first established by Morris (1983), who showed that the James-Stein estimator is one of a class of empirical Bayesian methods which dominate the OLS estimator. Given data $Y_i \sim N(\theta_i, 1)$, the James-Stein estimator is

$$\hat{\theta}_i^{\text{James-Stein}} = \mu_i + \frac{1 - P - 2}{S}(Y_i - \mu_i),$$

where μ is any initial guess at θ and $S = \sum(Y_i - \mu_i)^2$. In an empirical Bayes context, where $\theta_i \sim N(\mu_i, \sigma^2)$, the empirical Bayes estimate is

$$\hat{\theta}_i^{\text{EB}} = \mu_i + 1 - (1 + \sigma^2)^{-1}(Y_i - \mu_i).$$

As $E[(k-2)/S] = 1/(1+\sigma^2)$, this can be reduced to the James-Stein estimator. For μ_i, James and Stein use a single global mean $\mu = \bar{Y}$. In other words, the James-Stein estimator shrinks estimates for θ towards some global mean μ, which produces biased estimates but reduces the variance of the estimator, lowering error. To obtain estimators for Y_i, some algebra and simplification results in the form $\hat{Y}_i = \lambda\mu + (1-\lambda)\bar{Y}_i$, where $\lambda \in [0, 1]$ and was dubbed by Gelman and Pardoe (2006) as the "pooling factor", representing the degree to which information is pooled, i.e. the proportion of $\hat{Y}_i$ that is informed by global mean μ versus the data mean $\bar{Y}_i$. Empirical cross-validation studies, some of which were tabulated by Morris (1983), demonstrated the efficacy of the empirical Bayes estimator under data imbalance and data scarcity. In this paper and in later papers, the good performance of BHRMs has been attributed to information pooling and shrinkage based on intuition provided by the James-Stein estimator. However, the degree of information pooling λ has only been quantified in the one-way case where $Y_i \sim N(\alpha_i, \phi^2)$, $\alpha_i \sim N(a_0, \sigma^2)$, which has limited our understanding of Bayesian models. Gelman and Pardoe (2006) developed a Bayesian R-squared metric based on the pooling factor, but its limitations led to the more recent development of an alternative R-squared metric in Gelman et al. (2019).

This lack of understanding means that, in order to quantify the impacts of data imbalance on model estimates, extensive simulation studies often need to be performed. Eager and Roy (2017) showed in simulation the improvement of fully Bayesian approaches over Frequentist mixed models under data imbalance for models commonly used in linguistics. McCarron et al. (2011) performed multiple simulations to assess the performance of a specific set of Bayesian hierarchical models in combining evidence from randomised and non-randomised controlled studies. Thabtah et al. (2020) studied the impact of class imbalance on classification accuracy for a large class of models, including Naive Bayes. By explicitly quantifying the impacts of data imbalance on model estimates, we can reduce the need for simulation studies.

Another area where understanding Bayesian model estimates can be helpful is in influence analysis. Influence analysis examines those data points which may have a strong effect on the model fit, without which model parameters could be significantly

different. This can be determined through cross-validation, withholding small sets of individual data points at a time. Most famously, Cook (1977) proposed an influence analysis metric called "Cook's distance", which uses the leave-one-out predictive error $E[Y_i|Y_{-i}]$ as the basis for quantifying the overall influence that point Y_i has on model estimates. See Belsley et al. (2005) and Chatterjee and Hadi (2009) for additional procedures and diagnostic criteria. After identifying potentially influential points, the information borrowing weights can help determine exactly which model estimates are the most impacted by these potentially influential points and in what manner. This is much easier to translate to domain knowledge and, as such, domain knowledge can more easily inform whether to omit the potentially influential points.

By examining the degree and manner of information borrowing, we can better address whether the model reflects domain knowledge of the underlying data-generating mechanism. Incorporation of domain knowledge has most recently been an active area of research in the machine learning literature. Tsang et al. (2017) proposed the Neural Interaction Detection (NID) framework, which identifies statistical interactions between features in neural networks. Yan et al. (2019) propose the grouping-based interpretable neural network model, GroupINN, which simultaneously identifies brain sub-networks which best help with classification of function magnetic resonance imaging (FMRI)-derived brain graphs and extracts graph features. Both methods can be used to validate the model in terms of domain knowledge; for example, medical researchers can compare the interactions identified by NID against the existing medical literature. Roscher et al. (2020) gives further examples of explainable machine learning methods and provides a taxonomy for different categories and levels of model transparency.

The borrowing factors are informed only by the data availability and model specification. They also determine how model estimates are formed for Bayesian hierarchical regression models and so can be used to simulate the effects of different levels of data imbalance on model estimates. One particularly salient application of this is cross-validation (CV), which is essentially obtaining the model estimates after removing a certain number of points. In Chapter 4, we propose to use the borrowing factors to approximate cross-validated mean estimates for Bayesian hierarchical regression models. This shifts the CV problem from probability-based sampling to a simple and familiar optimization problem. In many cases, this produces estimates which are equivalent to full CV. We provide theoretical results, a diagnostic for finite sample sizes, and demonstrate its efficacy on a variety of publicly available data.

Chapter 4 is motivated by reducing the computational cost of cross-validation BHRMs,

which can be prohibitive. A fully Bayesian implementation of a model can be expensive to fit even once, which has led to many papers which either present new methods to approximate the posterior density or attempt to make current methods more efficient. Kingma and Welling (2013) proposed a stochastic variational Bayes method which is scalable to large data sets such that sampling-based methods like Monte Carlo would be infeasible. Variational Bayes approximates the intractable posterior $f(\theta|Y)$ with a simpler density that minimizes some dissimilarity function, typically Kullback-Leibler divergence. Lewis and Raftery (1997) proposed the Laplace-Metropolis estimator for approximating model selection. Laplace approximations are used to estimate $f(Y|M_0)$ and $f(Y|M_1)$, where M_0 and M_1 denote competing models, and to calcualte the Bayes factor, $f(Y|M_0)/f(Y|M_1)$. Rue et al. (2009) propose approximating latent Gaussian models using integrated nested Laplacian approximations (INLA), which now has a large software base using the R package R-INLA. Other methods propose variations on Markov Chain Monte Carlo (MCMC) estimation methods by fitting the model to a subset of the data, either through sub-sampling or partitioning the data into K clusters and combining the K posterior densities to approximate the full-data posterior density. Bardenet et al. (2017) provides a survey and comparison of such approaches. More recently, Quiroz et al. (2019) proposed sub-sampling MCMC using an unbiased esteimator of the log-likelihood for the n observations. Korattikara et al. (2014) replace the often-used Metropolis-Hastings step for MCMC with a sequential hypothesis test that allows the algorithm to accept or reject proposal steps using a portion of the full data set.

Cross-validation is particularly expensive when the number of folds increases with the size of the data, N. This is commonly the case under leave-one-out cross-validation or in models with complex dependence structures. Random K-fold cross-validation is the least computationally expensive, requiring re-fitting the model only K times, but when the data are not independence across CV folds, K-fold cross-validaiton selects models which overfit because of the high correlation between test and training data. Opsomer et al. (2001) showed this to be the case in the selection of bandwidth tuning parameters for longitudinal data. Arlot et al. (2010) discuss these and other issues in their survey of cross-validation methods.

The methods proposed here can be time-consuming to code and analyze. To reduce the amount of time spent on implementing these methods, we provide two R packages and a Shiny app which are described in Chapter 5.

Chapter 2
HIV prevalence estimation

Availability of data on HIV prevalence varies widely by country and sub-population. HIV surveillance data are often richer for the general population and sparser for key populations. Bayesian hierarchical modeling allows for pooling of data across different sub-populations or sub-national areas to improve estimation of HIV prevalence. The assumption this rests on is that different populations within a country will experience a similar pattern of HIV prevalence over time.

In section 2.1 we briefly cover published joint work with Dr Xiaoyue Niu which pools data across sub-national areas to improve HIV prevalence estimates Niu et al. (2017). In section 2.2, we introduce a model that pools data between key high-risk populations whose data tend to be sparse. Modeling HIV prevalence within key populations is complicated by two factors: the effects of the key populations are crossed with surveillance sites (rather than sites being nested within sub-national areas); and many of the key populations are related to each other.

2.1 HIV prevalence of sub-national areas

Prior related work is on pooling data across sub-national areas to improve HIV prevalence estimates Niu et al. (2017). Many countries collect data on HIV prevalence within the general population through routine antenatal testing at specific clinics within the country. However, availability of data often varies widely depending on the location of the clinic—in general, data are richer for clinics in urban areas and more sparse for clinics in rural areas.

Typically when fitting HIV prevalence models to sub-national areas, countries use a software called Spectrum/ Estimation and Projection Package (EPP), which fits a model separately within each area, independent of the rest. When data are scarce, this can lead

to poor estimates with large uncertainty. A Bayesian hierarchical model, in this case, can lead to more sensible estimates.

The Bayesian hierarchical model is fit across the full country's data and used to generate an auxiliary site within each sub-national area. The EPP model is then fit within each sub-national area to both the observed data and the auxiliary data.

This method is validated by comparing out-of-sample predictive accuracy of the EPP model with and without the auxiliary data.

2.1.1 Data

Four countries have made their data available through UNAIDS. The data were collected through routine testing at antenatal clinics. As table 2.1 shows, clinics in rural locations generally had fewer clinics with fewer years of data.

Table 2.1. Summary of data availability (table 1 in Niu et al. (2017))

		# Clinics	Years of Data	Average N per Clinic
Angola	Urban	38	6	488
	Rural	10	4	475
Liberia	Urban	22	7	405
	Rural	10	4	444
Ghana	Urban	23	22	427
	Rural	17	21	440
Swaziland	Hhohho	5	9	228
	Manzini	5	9	213
	Shiselweni	5	9	182
	Lubombo	6	9	202

2.1.2 Model

We use the model in Bao et al. (2016), shown below:

$$y_{iat} \sim Binomial(n_{iat}, p_{iat}) \tag{2.1}$$
$$\text{logit}(p_{ijt}) = a_0 + \sum_{k=1}^{4} a_k f_k(t) + a_{0j} + \sum_{k=1}^{4} \alpha_{jk} f_k(t) + a_{0i},$$

where $y_{ia}(t)$ is HIV prevalence at site i, for sub-area j, at time t; $f_k(t)$ is the k^{th} spline basis function spanning the length of years with data; a_k is the coefficient for the $f_k(t)$ with a diffuse normal prior; α_{jk} is an area-specific coefficient for $f_k(t)$ with a normal prior

and informative shrinkage hyperprior on the variance; a_{0j} is a group-specific intercept with a diffuse normal prior; a_{0i} is a site-specific intercept with a normal prior and informative hyperprior on the variance.

Taken together, $a_0 + \sum_{k=1}^{4} a_k f_k(t)$ in the model can be thought of as a fixed time trend that is shared among all groups, while $a_{0j} + \sum_{k=1}^{4} \alpha_{jk} f_k(t)$ is a group-specific random variation from the fixed time trend.

We compared $y_{iat} \sim$ Binomial with $y_{iat} \sim$ Beta-binomial, to account for overdispersion. The last 5 years of data are removed as test data.

Once model 2.1 is fitted to the training data, it is used to impute the values for an unknown site within each sub-area, which we call the auxiliary site. Sample sizes for each auxiliary site are determined through calculation from the posterior variance and posterior mean, or are set to a fixed value (such as 100). Comparisons are made of the different sample sizes.

The EPP model is then fit to both the auxiliary data and the observed training data and used to generate predictions for the test data, to determine the prediction error.

2.1.3 Results

Results are compared in terms of mean absolute error (MAE) and continuous ranked probability score (CRPS) over the test set. CRPS measures both prediction error and the width of the prediction interval; as such, a smaller CRPS is better. It is defined as

$$\text{CRPS}(F_p, y) = E_{F_p}|Y - y| - \frac{1}{2} E_{F_p}|Y - Y'|,$$

where y is the observed prevalence in the test set, F_p is the posterior predictive distribution, and Y, Y' are samples from F_p.

Table 2.2 shows that the EPP model fitted to both the auxiliary data and the observed training data is generally an improvement over the EPP model fitted to just the observed training data for data-scarce areas. For data-rich areas, there is little difference between the two. The EPP fitted to just the observed data is preferred only in two areas in Swaziland.

2.2 HIV prevalence of key populations

While data on HIV prevalence for the general population are often fairly accessible, data on hard-to-reach populations that are particularly vulnerable to HIV are harder to come

Table 2.2. Summary of results, table 2 in Niu et al. (2017)

	Sample Size	MAE	CRPS	MAE	CRPS
		Rural (1 year, 10 sites)		Urban (3 years, 25 sites)	
Angola	0	1.144	0.711	1.097	0.813
	4508 (GLMM)	1.058	0.677	1.099	0.814
	10	1.058	0.679	1.101	0.815
	100	1.067	0.682	1.094	0.811
	1000	1.047	0.675	1.101	0.813
		Rural (2 years, 4.5 sites)		Urban (5 years, 8 sites)	
Liberia	0	2.866	2.228	1.559	1.033
	560 (GLMM)	2.379	1.919	1.589	1.053
	10	2.466	1.948	1.535	1.022
	100	2.418	1.931	1.547	1.025
	1000	2.377	1.915	1.561	1.037
		Rural (16 years, 7.9 sites)		Urban (17 years, 20.7 sites)	
Ghana	0	0.753	0.507	0.925	0.647
	5653 (GLMM)	0.748	0.504	0.924	0.648
	10	0.741	0.499	0.928	0.650
	100	0.746	0.503	0.924	0.647
	1000	0.750	0.504	0.930	0.651
		Hhohho (6 years, 2 sites)		Manzini (6 years, 2 sites)	
Swaziland	0	4.565	3.665	6.019	3.991
	1272 (GLMM)	4.452	3.530	6.157	4.100
	10	4.278	3.389	6.657	4.680
	100	4.420	3.483	6.427	4.392
	1000	4.451	3.530	6.273	4.202
		Shiselweni (6 years, 2 sites)		Lubombo (6 years, 2.2 sites)	
	0	6.066	4.812	4.851	3.555
	1272 (GLMM)	6.117	4.982	4.610	3.328
	10	6.151	5.076	4.490	3.179
	100	6.134	5.040	4.506	3.194
	1000	6.130	5.012	4.601	3.31

by. These hard-to-reach populations are termed key populations, and include female sex workers, clients of female sex workers, intravenous drug users, and men who have sex with men.

Modeling HIV prevalence within key populations is complicated by two factors: multiple key populations can be present at one site, and a key population is present at multiple sites, so the effects of the key populations are crossed, whereas our previous work dealt with nested sub-national areas. Additionally, many of the key populations are related to each other, thus group effects may be correlated.

2.2.1 Data for HIV prevalence estimation among key populations

Ukraine, Jamaica, Morocco, and Pakistan have gathered data on HIV prevalence for key populations and at certain surveillance sites and have made them available through UNAIDS. Also available are data on HIV prevalence within the remaining female population, gathered through routine antenatal testing. Figure 2.1 gives an idea of the data availability—for the majority of countries, data on key key populations is not available until later years and many sites provide only a few years of data. Moreover, the majority of the time, sites have data on only one high risk group per year.

Figure 2.1. Availability of data on key populations (excluding general population). Each row gives the number of key populations available at one site across the time span present in the data.

The amount of data gathered varies widely for each country and high risk group, as shown in table 2.3. HIV prevalence is generally low in the remaining female population across all countries, but average HIV prevalence within other key populations varies per country.

For this analysis, we only consider key populations present at more than one site and with at least 5 data points.

Table 2.3. Summary statistics for HIV prevalence within key populations.

Country	Group	N	Year Range	Avg Prev	SD	Avg SS
Ukraine	Clients	30	2000 – 2014	0.030	0.023	286
	FSW	114	2000 – 2015	0.123	0.097	148
	PWIDs	160	1999 – 2015	0.256	0.162	285
	MSM	96	2007 – 2015	0.064	0.058	220
	Remaining female	320	2004 – 2015	0.007	0.005	17382
Jamaica	Clients	96	1990 – 2014	0.036	0.022	509
	FSW	16	1994 – 2011	0.119	0.066	291
	Homeless	6	2005 – 2013	0.066	0.037	306
	MSM	6	1995 – 2011	0.321	0.017	308
	Prison Inmates	4	1996 – 2011	0.037	0.021	710
	Remaining female	95	1997 – 2014	0.013	0.008	544
Morocco	Clients	34	2006 – 2014	0.009	0.011	799
	FSW	52	2002 – 2014	0.021	0.020	617
	PWIDs	9	2008 – 2015	0.097	0.129	181
	MSM	24	2005 – 2015	0.030	0.025	476
	Remaining female	77	1999 – 2014	0.002	0.002	3109
Pakistan	Clients	1	2001 – 2001	0.001	NA	1000
	Migrant workers	5	1996 – 2013	0.004	0.003	5400
	FSW	16	2005 – 2011	0.010	0.006	353
	PWIDs	44	2003 – 2011	0.191	0.144	339
	MSW	22	2005 – 2011	0.021	0.015	268
	Remaining female	1	2011 – 2011	0.000	NA	26000
	Transgender	33	2004 – 2011	0.040	0.057	238

2.2.2 Spline function selection

As the HIV epidemic began in 1980, peaked in the 1990s, and then gradually decreased since then, we specify 2 knots for the cubic spline functions which more closely matches the course of the HIV epidemic than the 3 knots in 2.1.

To identify the ideal set of knots, we consider all pairs of knot locations that have at least 3 years of data between the knots and compare the effects of the spline functions using by-site cross-validation error. We compare mean absolute error (MAE) and use INLA to fit the models (Rue et al. (2009), Lindgren and Rue (2015)). This is repeated with the addition of 0 prevalence at 1975. In every case, the addition of a 0 prevalence point at 1975 for each site improves the MAE.

Selecting knot locations through cross-validated MAE is expensive. Figure 2.2 shows

that the cross-validated MAE seems to be a smooth function, which we found to be the case across all countries. Thus, we can investigate whether there exists a general rule for choosing knot locations that is consistently close to the ideal combination.

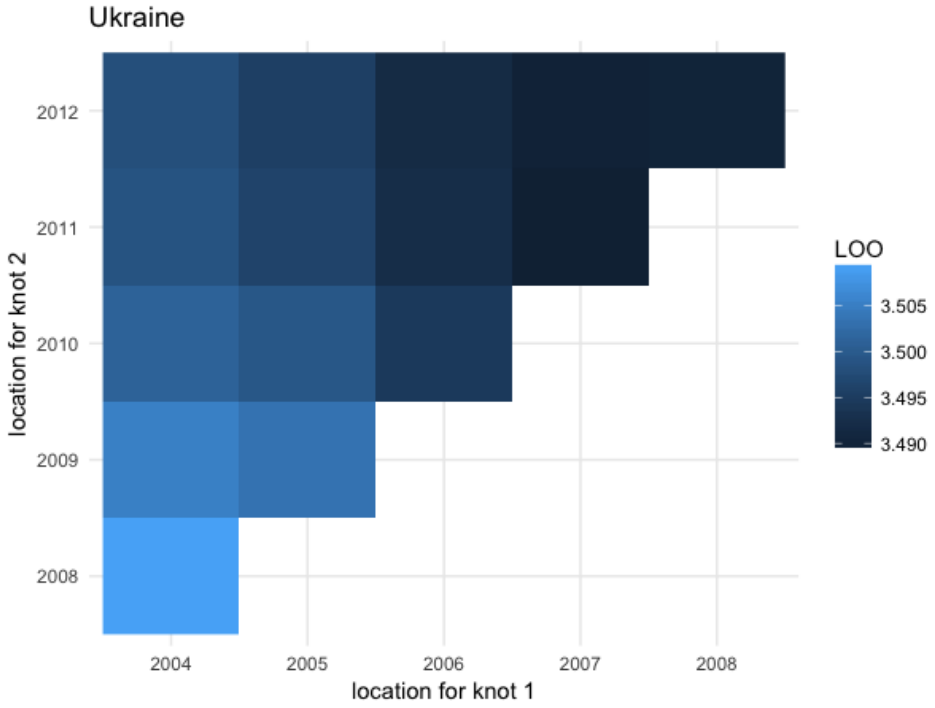

Figure 2.2. Graph of leave-one-out (LOO) MAE based on knot location for Ukraine.

When choosing knot locations, common practice is to space the knots evenly across data points. One alternative is to space the knots evenly across years of data, ignoring how many data points there are at any given year. Figure 2.3 compares the performance of these two approaches relative to the ideal knot locations identified through by-site cross-validation. Between the two approaches, spacing knot locations evenly across the data points is consistently in areas with lower leave-one-out MAE (darker tiles).

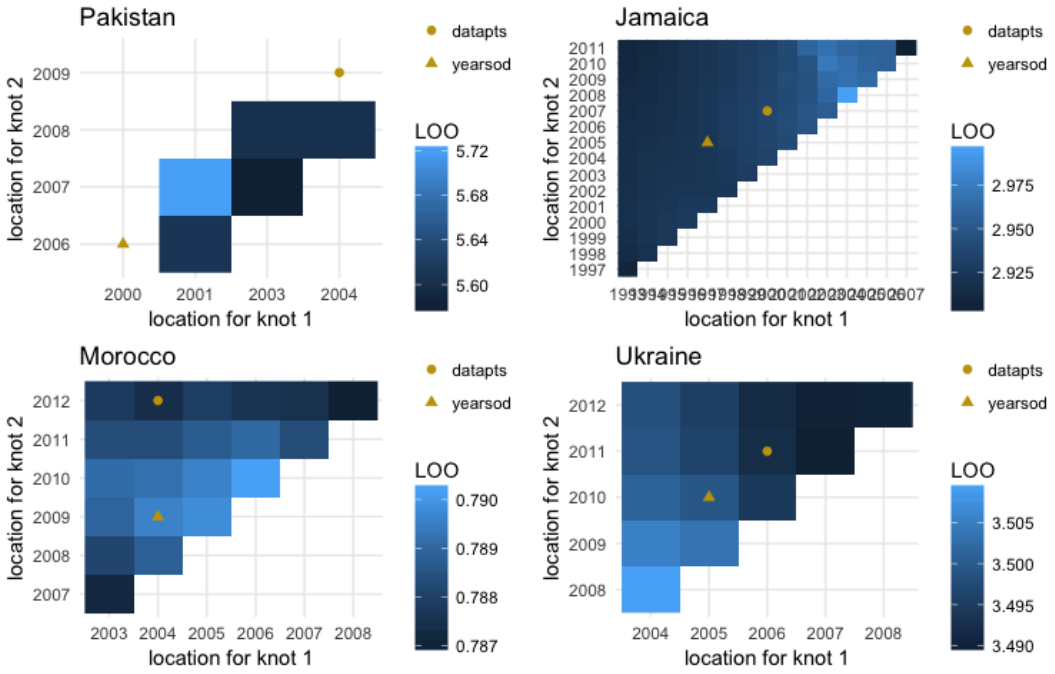

Figure 2.3. Comparison of the two methods to space knot locations against by-site cross-validated MAE of possible knot locations. Empty spaces are due to estimation failure of INLA package.

2.2.3 Estimating HIV prevalence for multiple key populations

We first consider a model that only includes the additive group, site and time effects:

$$F^{-1}(p_{igt}) = \mu_g + \mu_i + \sum_k \mu_k f_k(t) + \epsilon_{igt} \tag{2.2}$$

where F is the cumulative distribution function for the standard normal $N(0,1)$, g indexes groups, i indexes sites, t indexes years, k indexes the spline functions $f_k(t)$, and p_{igt} is the HIV prevalence for population g at site i and time t. The spline functions are chosen as described in section 2.2.2.

We estimate the variances of each component in model 2.2 using ANCOVA. The mean sum squares (MSS) estimates the variation of p_{igt}'s coming from different sources. From table 2.4, we can see that the residual MSS is fairly small suggesting that the simple additive model is promising. Moreover the order of MSS is consistent within countries – $\text{MSS}_{\text{Group}} > \text{MSS}_{\text{Spline}} > \text{MSS}_{\text{Site}} > \text{MSS}_{\text{Residual}}$.

The additive model assumes that the heterogeneity of HIV prevalence between key populations are fully captured by the group effect, μ_g, and there are no group-time or group-site interactions. This may be true for some but not all pairs of key populations.

Table 2.4. Mean sum squares for ANCOVA within each country

Country	Group	Splines	Site	Residuals
Ukraine	99.91	2.00	1.83	0.10
Jamaica	8.15	0.67	0.26	0.04
Morocco	9.22	1.19	0.78	0.07
Pakistan	8.17	1.81	0.79	0.17

To gain information on the relationships between specific pairs of key populations and identify possible correlations, we also fit the model in equation 2.2 to pairs of key populations within each country.

The results of fitting the ANCOVA within pairs of key populations are given in table 2.5. Note that MSS for the splines is small whenever the risk group pair includes the remaining female population. We believe this is due to HIV prevalence rates staying below 1% for the remaining female population of these countries. This may indicate a different time trend for the remaining female population, where a small variance is placed on the coefficients of the spline functions.

There are also a number of risk group pairs where $\text{MSS}_{\text{Group}}$ is much lower than in the ANCOVA across all risk group pairs for that country. This suggests that those risk group pairs are more similar than others within that country. Additionally, in some of these cases $\text{MSS}_{\text{Group}}$, $\text{MSS}_{\text{Spline}}$, and MSS_{Site} are all small in magnitude, which suggests that the risk group pair is both close in average HIV prevalence and that their HIV prevalence curves are relatively flat. For these cases, it may be worthwhile to allow for the groups to be correlated. Pairs of key populations exhibiting one of these behaviors and who have over 15 points in common (same site and year) are bolded in table 2.5: in Ukraine, clients and remaining female, female sex workers and intravenous drug users, female sex workers and men who have sex with men; in Morocco, clients and female sex workers, clients and remaining female, female sex workers and men who have sex with men; in Pakistan, MSW and transgender people.

Inspired by the above exploratory analysis, we fit to each country a simple additive model with a group effect, a time trend fitted using splines, and a site effect,

$$\begin{aligned} Y_{igt} &\sim Bin(n_{igt}, p_{igt}) \\ F^{-1}(p_{igt}) &= a_g + \sum_{k=1}^{3} a_k f_k(t) + \alpha_i, \end{aligned} \tag{2.3}$$

where Y_{igt} is the number of respondents who are HIV positive at site i, in group g, time

Table 2.5. Mean squares for ANCOVA within pairs of key populations, where key populations have at least 10 points in common (same site and year), with comparison to ANCOVA within each country (all key populations).

Country	Risk group pair	Group	Splines	Site	Residuals
Ukraine	All key populations	99.91	2.00	1.83	0.10
	Clients - FSW	10.24	3.48	0.49	0.10
	Clients - PWIDs	37.00	3.03	1.19	0.10
	Clients - Remaining female	10.41	0.65	0.97	0.01
	FSW - PWIDs	20.40	7.35	1.25	0.12
	FSW - MSM	6.60	4.01	0.55	0.12
	FSW - Remaining female	136.14	0.50	1.24	0.06
	PWIDs - MSM	49.61	3.90	0.94	0.15
	PWIDs - Remaining female	355.74	0.60	1.66	0.06
	MSM - Remaining female	62.08	0.36	1.03	0.04
Jamaica	All key populations	8.15	0.67	0.26	0.04
	Clients - Remaining female	7.78	0.56	0.25	0.04
Morocco	All key populations	9.22	1.19	0.78	0.07
	Clients - FSW	2.04	1.22	0.44	0.06
	Clients - MSM	3.50	1.27	0.13	0.07
	Clients - Remaining female	5.87	1.01	0.28	0.05
	FSW - MSM	0.55	0.72	0.46	0.07
	FSW - Remaining female	20.58	1.00	0.56	0.04
	MSM - Remaining female	18.20	0.42	0.23	0.07
Pakistan	All key populations	8.17	1.81	0.79	0.17
	PWIDs - MSW	16.83	1.79	0.59	0.22
	PWIDs - Transgender	16.14	3.44	0.68	0.20
	MSW - Transgender	0.28	0.58	0.46	0.10

t. The priors for parameters in 2.3 are

$$\begin{aligned} a_g &\sim N(0, 100) \\ a_k &\sim N(0, 100) \\ \alpha_i &\sim N(0, \sigma^2) \\ \sigma^2 &\sim \text{half-Cauchy}(0.5). \end{aligned} \tag{2.4}$$

We call this the shared trend model.

For the variance σ^2, we use a half-Cauchy prior as they have been shown to perform better than the inverse-gamma. Half-Cauchy priors perform well near 0 but are also

diffuse enough so that they do not overly shrink large effects (Gelman (2004), Polson et al. (2012)).

As $\text{MSS}_{\text{Group}}$ and $\text{MSS}_{\text{Spline}}$ are the largest sources of variation in the two ANCOVAs, we also fit a model with an interaction term between the spline function and the group effect,

$$\begin{aligned} Y_{igt} &\sim Bin(n_{igt}, p_{igt}) \\ F^{-1}(p_{igt}) &= a_g + \sum_{k=1}^{3} a_k f_k(t) + \sum_{k=1}^{3} \alpha_{gk} f_k(t) + \alpha_i \end{aligned} \tag{2.5}$$

The priors remain the same for a_g, a_k, and α_i. On the group-specific spline coefficients α_{gk} we place a normal prior

$$\begin{aligned} \alpha_{gk} &\sim N(0, \sigma_k^2) \\ \sigma_k^2 &\sim \text{half-Cauchy}(0.75) \end{aligned} \tag{2.6}$$

We call this the random trend model.

For this analysis, we choose the scale parameter for the half-Cauchy hyperpriors such that the 0.95 quantile is approximately equal to the largest MSS for the effect given in table 2.5. For the splines, this is scale parameter 0.75 and for the sites it is 0.5. We plan to perform a sensitivity analysis to quantify the effect of this hyperparameter.

We compare models 2.3 and 2.5 to the baseline model, which is a spline time trend and a site-specific effect, given in equation 2.7. We compare mean absolute error and 95% prediction interval coverage with cross-validation over unique combinations of sites and groups. That is, we remove in turn one group from one site, fit the remaining data to the model, and evaluate the metrics above on the held-out data.

$$\begin{aligned} Y_{it} &\sim Bin(n_{it}, p_{it}) \\ F^{-1}(p_{it}) &= \sum_{k=1}^{3} a_k f_k(t) + \alpha_i \end{aligned} \tag{2.7}$$

Many of the sites did not start collecting data until well after the start of the HIV epidemic. Due to scarcity of the data, we found that it was possible for spline functions to fit to the data in such a way that the spline functions would have a decreasing slope at the earliest years of data, rather than increasing, as would match the HIV epidemic. One solution we explore is to impose a point at 1975 with 0 HIV prevalence within each

site, where the sample size of the point at 1975 is equal to the maximum sample size of the data at that site. When evaluating the models, we compare results both with and without the imposed 0 HIV prevalence at 1975.

2.2.4 Results

Table 2.6 summarizes the results of fitting the baseline, shared trend, and random trend models.

Table 2.6. Cross-validated MAE for each high risk group

		Baseline		Shared Trend		Random Trend	
Country	Group	W/o 0	With 0	W/o 0 0	With 0	W/o 0	With 0
Ukraine	Clients	2.160	2.214	1.919	1.755	1.911	2.023
	FSW	7.604	7.751	7.036	7.057	5.701	5.587
	PWIDs	13.269	13.225	10.934	11.293	10.096	10.087
	MSM	5.282	5.450	4.259	4.319	4.394	4.394
	RFP	0.503	0.505	0.394	0.398	0.413	0.417
Jamaica	Clients	1.782	1.711	1.614	1.635	1.656	1.654
	FSW	6.074	6.483	5.542	5.606	5.031	5.010
	MSM	7.846	10.044	8.412	9.625	6.042	8.539
	RFP	0.546	0.541	0.714	0.755	0.698	0.742
Morocco	Clients	0.593	0.602	0.568	0.572	0.581	0.578
	FSW	1.557	1.648	0.950	0.977	0.993	1.032
	PWIDs	35.415	30.367	10.621	10.710	10.840	10.527
	MSM	1.833	2.872	1.795	1.760	1.974	1.631
	RFP	0.140	0.134	0.105	0.104	0.102	0.104
Pakistan	FSW	0.466	0.442	0.659	0.666	0.654	0.673
	PWIDs	10.744	10.831	11.479	11.510	11.713	11.739
	MSW	2.239	1.932	1.335	1.354	1.432	1.416
	Trans	3.777	3.706	3.579	3.511	3.863	3.919

The results in 2.6 show little difference between including the imposed 0 prevalence point and excluding it; as such we recommend excluding it as this point has high leverage.

Comparing MAE results for models fit without the zeros, we can see that both the shared and random trend models improve upon the baseline model. This level of improvement is similar and occasionally favors the random trend model, as in MSM within Jamaica and and FSW within Ukraine.

The baseline model is preferred in three of the cases: RFP (remaining female population) within Jamaica, FSW within Pakistan, and PWIDs within Pakistan. This

may indicate that these sub-populations have trends that are quite different from the remaining sub-populations. As the random trend and shared trend MAEs are similar, this may indicate that the level of shrinkage in the random trend model is too strong and so it is behaving similarly to the shared trend model, making it unable to sufficiently fit the groups where the baseline model is preferred. Adjusting the splines in the random trend model to have a different level of smoothness per group may alleviate this issue.

We note also that the MAE within PWIDs is consistently higher than other key populations, even within the baseline model. This may be because HIV prevalence is generally higher within PWIDs (0.2-0.8) versus other key populations (<0.1), thus on the prevalence scale the MAE is also high. Calculating MAE on the probit or logit scale may reduce this difference. It also may be because MAE does not take sample size into consideration, thus an alternative measure such as log pseudo-marginal likelihood (LPML) may be more helpful in comparing model fits across groups.

Based on these results, we believe that the random trend model is the most promising as it serves as a mid-way point between the baseline and shared trend models. This model also has the option of fitting level of smoothness of splines to each group, which may allow it to perform better than the baseline in the three cases where the baseline was preferred.

2.3 Discussion

We propose to allow for a correlation between risk group pairs where the MSS for groups much smaller than the $\text{MSS}_{\text{Group}}$ estimated across all key populations, and whose $\text{MSS}_{\text{Group}}$, $\text{MSS}_{\text{Spline}}$, and MSS_{Site} are all small, as shown in 2.5. This would be done by separately estimating a correlation matrix and a variance common to all of the key populations, which is less restrictive than using an inverse-Wishart prior Barnard et al. (2000). One option is to place a uniform prior over the space of correlation matrices; however, given the sparsity of the data, this may lead to poorly-estimated correlations. We propose then to allow for strong correlations between certain key populations and, to avoid unnecessarily constraining the space of correlation matrices, placing a strong prior on the remaining correlations to shrink them towards 0.

Table 2.5 also shows that the prevalence curves may behave differently for the remaining female population than for the key populations. This is intuitive as HIV prevalence within the general population is relatively stable and below 1%. Thus we also propose placing a stronger prior on the variance of the spline function coefficients for the

remaining female population to reflect this.

The last extension we propose is to take a group-specific penalized spline approach within the random trends model. Rather than σ_k^2 in equation 2.6, we would have a σ_g^2. Combined with a roughness penalty matrix and a larger number of knots, this would be a penalized spline approach where the level of smoothness of each group's mean HIV prevalence curve would be estimated separately. One advantage of this is it may allow the random trends model to better capture the curves in which the baseline model performed better. Another is that, due to varying data availability, there are key populations whose data do not span both spline knots within the shared and random trend models; increasing the number of knots would allow all groups to span multiple knots of data.

Chapter 3
Information borrowing in regression models

3.1 Introduction

Model development is often an iterative process, particularly in challenging settings with high-dimensional feature sets and/or complex dependency structures. Model fit, subject matter considerations, data properties, and model assumptions are all factors which are taken into consideration for the final model and multiple models may be evaluated and compared to each other. To diagnose issues with any of these factors, it can be helpful to understand regression model estimates at a more granular level. We propose to understand regression model estimates by expressing them as a function of a vector of weights placed on each data point. This offers the intuitive interpretation that estimates are formed by "borrowing" information from other sets of points, with the weight being the amount borrowed. As such, we call the weights "borrowing factors".

This granular decomposition of regression model estimates can be particularly helpful for Bayesian hierarchical models where a shared hyperprior is placed on model parameters to pool information between them and improve model estimates (Efron and Morris, 1975; James and Stein, 1992; Stein, 1956). Information pooling has historically been understood through the lens of the James-Stein estimator. Given observed data $Y_i \sim N(\alpha_i, \phi^2)$, $j = 1, \ldots, J$, Stein (1956) developed a biased estimator which improves upon the unbiased ordinary least squares (OLS) estimator for $\alpha \in I\!R^P$, $P \geq 3$, under squared loss. This result was later improved by James and Stein (1992) and dubbed the "James-Stein

estimator". Given data $Y_j \sim N(\alpha_j, 1)$, the James-Stein estimator is

$$\hat{\alpha}_j^{\text{JS}} = \mu_j + \frac{1 - P - 2}{S}(Y_j - \mu_j), \quad S = \sum (Y_j - \mu_j)^2,$$

where μ_j is any initial guess at α_j. For μ_j, James and Stein use the global data mean $\mu = \bar{Y}$. This was shown by Morris (1983) to be equivalent to the estimator under an empirical Bayes framework. They additionally proved that the James-Stein estimator is one of a class of empirical Bayesian methods which dominate the OLS estimator under squared loss by shrinking estimates for α towards some global mean μ, producing biased estimates but reducing the variance of the estimator and lowering error. This shrinkage towards the mean is referred to as information pooling.

Information pooling has been quantified for simple one-way models where $Y_i \sim N(\alpha_i, \phi_i^2)$, $\alpha_i \sim N(a_0, \sigma^2)$. Assuming a_0, ϕ, and σ are known, some algebra and simplification results in the empirical Bayes estimator

$$\hat{Y}_i = \lambda a_0 + (1 - \lambda)\bar{Y}_i, \quad \lambda \in [0, 1],$$

where λ was called the "pooling factor" by Gelman and Pardoe (2006). Bayesian hierarchical models have been shown to perform well in several empirical studies, particularly when data are imbalanced or scarce (Gelman and Hill, 2007; Morris, 1983), and information pooling is often cited as the reason. However, information pooling has not been explicitly quantified in scenarios outside of the one-way setting, which has limited its use in applications; one of few examples is Gelman and Pardoe (2006). Our method quantifies information pooling for all regression models and can be used to identify patterns of information borrowing, for example, whether information is pooled evenly or unevenly. We can then confirm whether model estimates are in accordance with domain knowledge, which is often the deciding factor between models which perform similarly well based on metrics like mean squared error.

Other cases where granular examination of regression model estimates are useful include when the data are highly imbalanced, which can lead to biased estimates (Gelman and Hill, 2007). The effects of data imbalance are often evaluated through extensive simulation studies, some recent examples of which include Eager and Roy (2017), McCarron et al. (2011), and Thabtah et al. (2020). As the borrowing factors capture all information on how data availability impacts the model estimates, our approach directly quantifies the impact of data imbalance on the model estimates without simulation.

Another is when a model contains points with high leverage. Researchers may

determine that those points have a large degree of influence on the model fit as a whole and may remove them if they are determined to be outlying (Belsley et al., 2005; Chatterjee and Hadi, 2009). This is typically done through a combination of domain knowledge and influence analysis metrics, which are typically based on removing the influential point and assessing the change in model fit. The borrowing factors explicitly identify which points estimates are impacted the most by high-leverage observations and to what degree, as well as the model's overall information borrowing patterns, which can be more easily validated by domain knowledge than using an influence analysis metric. If an observation is highly influential, but its influence is mostly limited to a small and specific subset of related observations, subject matter and model considerations can then inform whether to remove or include the point.

In Section 3.2, we formally define the borrowing factors. We also introduce a metric which we call the sum squares of borrowing factors (SSBF), which is a summary of the information borrowing pattern for each point, as well as some useful terminology. In Section 3.3, we describe theoretical properties of the borrowing factors and of SSBF. We show that the borrowing factors are connected to the pooling factor and demonstrate SSBF's connection to two influence analysis metrics. In the next two sections, we illustrate how the borrowing factors and SSBF can be used to link the effects of model assumptions and data availability to model estimates using two example data sets. Section 3.4 shows how we can explicitly quantify the effects of data imbalance using the Radon data set (Gelman and Hill, 2007). Section 3.5 shows how model assumptions can be linked to model estimates and how the borrowing factors and SSBF can be used to provide context to influence analysis and quantify the impact of influential points on model estimates.

3.2 Quantifying shrinkage and information borrowing

In this section, we provide an overview of our approach, with detailed discussion of theoretical properties in Section 3.3. We discuss the Bayesian setting first. Let $\boldsymbol{Y} \in \mathbb{R}^N$ denote a continuous response vector that follows

$$\begin{aligned} \boldsymbol{Y}|\boldsymbol{\beta}, \Phi &\sim N(X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2, \Phi), \\ \boldsymbol{\beta}_1 \sim N(\alpha_1, C), \quad \boldsymbol{\beta}_2|\Sigma &\sim N(\alpha_2, \Sigma), \\ \Sigma \sim f(\Sigma), \quad \Phi &\sim f(\Phi), \end{aligned} \tag{3.1}$$

where $X \coloneqq \begin{bmatrix} X_1 & X_2 \end{bmatrix} \in I\!R^{N \times P}$ is the design matrix, $\boldsymbol{\beta}_1 \in I\!R^{P_1}$, $\boldsymbol{\beta}_2 \in I\!R^{P_2}$ s.t. $\boldsymbol{\beta} \coloneqq \begin{bmatrix} \boldsymbol{\beta}_1' & \boldsymbol{\beta}_2' \end{bmatrix}' \in I\!R^P$, $C \in I\!R^{P_1 \times P_1}$ is positive-definite and typically a diagonal matrix, $\Sigma \in I\!R^{P_2 \times P_2}$ is positive-definite, and $\Phi \in I\!R^{N \times N}$ is diagonal and positive-definite. We assume $\mathbf{1} \in \text{span}(X_1)$, where $\mathbf{1}$ is the N-length vector of ones, which is satisfied when the fixed effects include a global intercept or set of intercepts which partition the data. We take $\alpha_1 = \alpha_2 = 0$ throughout this paper, without loss of generality. C is treated as fixed, often with large variances, and thus $\boldsymbol{\beta}_1$ are referred to as the fixed effects. Random variance hyperparameters such as Σ reflect the dependency among the $\boldsymbol{\beta}_2$; the effect is to pool information among related units and shrink them towards a common mean, thus the $\boldsymbol{\beta}_2$ are referred to as random effects. Σ can take many forms, as long as it is positive-definite.

When modeling data as in (3.1), the posterior mean for $X\boldsymbol{\beta}$ conditioned on variance parameters has the form

$$E[X\boldsymbol{\beta}|\Sigma, \Phi, \boldsymbol{Y}] \sim XVX'\Phi^{-1}\boldsymbol{Y}, \quad V = \left(X'\Phi^{-1}X + \begin{bmatrix} C^{-1} & 0 \\ 0 & \Sigma^{-1} \end{bmatrix} \right)^{-1}, \tag{3.2}$$

where C^{-1} is taken as the matrix of 0s, which corresponds to the assumption that the fixed effects have infinite variance. Kass and Steffey (1989) show that the posterior mean $E[\boldsymbol{\beta}|\boldsymbol{Y}] = E[\boldsymbol{\beta}|\boldsymbol{Y}, \hat{\Sigma}_{\text{EB}}](1 + \mathcal{O}\left(P_2^{-1}\right))$, where $\hat{\Sigma}_{\text{EB}}$ denotes the Empirical Bayes estimates and $\hat{\Sigma}_{\text{EB}}$ in turn approximates posterior mean $\hat{\Sigma} = E[\Sigma|\boldsymbol{Y}]$ with order $\mathcal{O}\left(P_2^{-1}\right)$. Conditioning on variance parameters and using the posterior means $\hat{\Sigma}$ and $\hat{\Phi}$ as plug-in estimates in (3.2) then produces estimates which approximate the posterior mean. The accuracy of this approximation is simple to determine by comparing the conditional expectation $E[X\boldsymbol{\beta}|\hat{\Sigma}, \hat{\Phi}, \boldsymbol{Y}]$ to the posterior expectation $E[X\boldsymbol{\beta}|\boldsymbol{Y}]$.

In the frequentist setting, the coefficients $\boldsymbol{\beta_1}$ and variance parameters Φ and Σ are fixed at their estimates, with

$$\begin{aligned} \boldsymbol{Y} &= X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}, \\ \boldsymbol{\epsilon} &\sim N(0, \Phi), \quad \boldsymbol{\beta}_2 \sim N(0, \Sigma). \end{aligned}$$

So, (3.2) directly expresses the fitted values for the frequentist regression model and is not an approximation.

In the case of a generalized linear model, we approximate the non-linear data-level model with a normal distribution having the same moments. This is known as expectation propagation with a Gaussian approximating density (Minka, 2013) and was shown by Daniels and Kass (1998) to be a Laplace approximation with the same asymptotic error.

Expectation propagation has been shown to work well (Vehtari et al., 2016a) and the accuracy of the approximate is straightforward to determine by numerically comparing $E[X\boldsymbol{\beta}|\boldsymbol{Y}]$ to its normal approximate.

Equation (3.2) expresses mean estimates $\hat{Y}_i$, $i = 1, \ldots, N$, as a weighted average of the response data $\boldsymbol{Y}$, where the $N \times N$ matrix of weights is

$$W \coloneqq XVX'\Phi^{-1} \tag{3.3}$$

and is informed only by the model specification and data availability, not the response. How data availability and model specification impact model estimates can then be wholly determined by examining W, and an entry w_{ij} in the i^{th} row and j^{th} column of W can be thought of as the amount of information borrowed from Y_j for point estimate $\hat{Y}_i$. This allows us to explicitly quantify the amount of information borrowing for all model estimates.

How to interpret W such that we can clearly link data availability or model assumptions to model estimates? We aggregate over w_{ij}'s to determine the amount borrowed from a set of points $J \subset \{1, \ldots, N\}$. We refer to both w_{ij} and $\sum_{j \in J} w_{ij}$ as "borrowing factors", with the latter denoted as b_{iJ}. The borrowing factors can then be linked to data availability, model covariates, or other quantities of interest. This can help to identify higher-level patterns of information borrowing and determine which lenders are the most impactful for any specific point estimate $\hat{Y}_i$. After understanding how model assumptions and the data availability lead to point estimates, researchers can verify whether model estimates are generated in ways that are in line with subject matter considerations. For instance, in a model of standardized test scores with school, class, and age as covariates, the borrowing factors can determine whether the estimated standardized test score of a student borrows more from students of the same school, students of the same class, or students of the same age group (younger v.s. older).

When $i = j$, w_{ij} is the amount of information borrowed from a point estimate's own data. It is helpful to separately consider such cases—let x_i' denote the i^{th} row of the model matrix X, and let $B_i = \{j \in 1, \ldots, N : x_j = x_i, \phi_j = \phi_i\}$ indicate rows that have the identical design covariates and variance with the i^{th} row, where ϕ_i^2 is the i^{th} diagonal entry of Φ. We denote the cardinality of B_i as n_i. Note that $w_{ij} = w_{ii}$ for all $j \in B_i$, thus any of the Y_{B_i} can be exchanged with each other and obtain the same model estimates. We call the set of indices B_i the borrowers or the borrower cluster. The shrinkage factor

is the total weight placed on the borrower cluster,

$$b_{iB_i} = \sum_{j \in B_i} w_{ij}. \tag{3.4}$$

All other points are referred to as the lenders, $L_i = \{j \in 1, \ldots, N : x_j \neq x_i\}$. The pooling factor is the total weight placed on lenders,

$$b_{iL_i} = 1 - b_{iB_i} = b_{iL_i}. \tag{3.5}$$

If a point estimate $\hat{Y}_i$ has lower pooling factor, then its value will be closer to $\bar{Y}_i$.

The terms shrinkage and pooling factors originate from the Bayesian literature for simple one-way models, $Y_i \sim N(\alpha_i, \phi_i^2)$, $\alpha_i \sim N(a_0, \sigma^2)$ (Efron and Morris, 1975; Gelman and Pardoe, 2006) and the definitions we present here extend the definition to all regression models, as we show in Section 3.3.1. They help to summarize how similar a point estimate $\hat{Y}_i$ is to its data mean $\bar{Y}_i$ versus how much is borrowed, which by itself can be helpful for understanding model estimates. However, they do not contain information on which lenders are borrowed from the most and thus cannot explain what higher-level patterns of information borrowing exist. We may have some intuition; for example, if the data are imbalanced, we expect those clusters with less data to borrow more and for that borrowing to come largely from clusters with more data, but this has not been explicitly quantified for any model in the literature.

We also propose a metric that summarizes total borrowing in each row of W, the sum squares of borrowing factors (SSBF), where

$$\text{SSBF}_i = \sum_{j \in L_i} w_{ij}^2. \tag{3.6}$$

SSBF is similar to the pooling factor in that it aggregates over the borrowing factors of the lenders but it uses their squared values. Point estimates will thus have higher SSBF if they place high individual weight on lenders and low SSBF if no lender has particularly large weight; in fact, SSBF is proportional to the sample variance of borrowing factors (3.10). Thus points with high SSBF have more distinct borrowing patterns, with some lenders having high individual borrowing factors, based on a relationship they share with the borrower cluster. Understanding how SSBF changes with data availability, model covariates, or other metrics of interest can help identify borrowing patterns. SSBF is also related to both the retrospective value of sample information (Parsons and Bao, 2018)

and Peña (2005)'s metric S_i in the influence analysis literature and can be thought of as the total influence of all lenders due solely to the data availability. In some scenarios, it can also be interpreted as model uncertainty for estimate $\hat{Y}_i$. We show these properties and discuss them in more detail in Section 3.3.2.

To identify borrowing patterns for a borrower cluster B_i, it is often helpful to partition the lenders L_i into a set of relationship groups, where the groups are determined based on the lenders' similarity to the borrower. For models with clustered data, a good starting point to define relationship groups is to examine the locations of non-zero entries of $x_i x_j'$ and to group together those points j that have the same non-zero locations. Zero values of x_i indicate that the corresponding entry in $\boldsymbol{\beta}$ does not contribute to $\hat{Y}_i$; the non-zero entries in $x_i x_j'$ then correspond to coefficients which contribute to both $\hat{Y}_i$ and $\hat{Y}_j$. For example, given a nested model with $E[Y_{ljk}|a_0, \alpha_j, \alpha_{jk}] = a_0 + \alpha_j + \alpha_{jk}$, where Y_{ljk} is the standardized test score of student l from school k of school district j, $l = 1, \ldots, n_{jk}$ represent the borrower cluster that have the same point estimate, $\hat{Y}_{jk}$, a_0 is a global mean parameter, α_j corresponds to school-district-level random effects, and α_{jk} corresponds to school-level random effects, the relationship groups for a point estimate $\hat{Y}_{jk}$ could consist of two clusters: 1) lenders in the same school district but different schools $Y_{jk'}$ (with a_0 and α_j in common); and 2) lenders in different school districts $Y_{j'k'}$ (with only a_0 in common). The most helpful partition will vary, depending on the model and data.

To identify which lenders contribute most to SSBF and have the highest individual weight placed on them, it can be helpful to decompose the SSBF into the sum of square borrowing factors over a set of lenders, denoted by J, which we call the partial SSBF (PSSBF),

$$\text{PSSBF}_{iJ} = \sum_{j \in J} w_{ij}^2. \tag{3.7}$$

As SSBF is additive, the sum of partial SSBFs over all relationship groups is the SSBF. PSSBF offers a more granular interpretation of SSBF and a scatter plot of partial SSBF against SSBF, colored by relationship group, can identify which group of lenders contribute the most to SSBF and thus have the most distinct borrowing patterns, an example of which is in Figure 3.2.

Table 3.1 repeats and summarizes the definitions for each of the terms listed above. Each is a different way of summarizing information borrowing for a given point estimate $\hat{Y}_i$. When referred to without the subscript i, all terms except for B_i and L_i in the table refer to their N-length vector counterparts, where the i^{th} entry is, for example, b_{iJ} or b_{iL_i}.

Table 3.1. Summary of term definitions and notation for borrowing factors and SSBF of a given point estimate $\hat{Y}_i$, for $i \in \{1, \ldots, N\}$.

Notation	Term	Definition
w_{ij}	individual borrowing factor	$(i,j)^{th}$ entry of W
b_{iJ}	aggregate borrowing factor	$\sum_{j \in J} w_{ij}$, for $J \subset \{1, \ldots, N\}$
B_i	borrowers, borrower cluster	$\{j \in 1, \ldots, N : x_j = x_i, \phi_j = \phi_i\}$
L_i	lenders	$\{1, \ldots, N\} \setminus B_i$
b_{iB_i}	shrinkage factor	$\sum_{j \in B_i} w_{ij}$
b_{iL_i}	pooling factor	$\sum_{j \in L_i} w_{ij}$
SSBF_i	SSBF	$\sum_{j \in L_i} w_{ij}^2$
PSSBF_{iJ}	partial SSBF over J	$\sum_{j \in J} w_{ij}^2, J \subset L_i$

Exploring the partial SSBF, SSBF, and borrowing factors can help link model assumptions and the data availability to point estimates. Depending on the relationship groups, the data, and the model, comparing borrowing factors directly to measures of interest can become quite complex, so we propose a two-stage process. First we determine what contributes the most to changes in SSBF. We compare SSBF to the data availability, model covariates, or some other metric of interest, such as partial SSBF. We then decompose model estimates into borrowing factors over relationship groups and compare them to SSBF, typically as a scatter plot. We can then interpret the change in borrowing factors as SSBF increases or decreases as being due to the data availability, model covariates, or some other metric of interest.We have found this approach to be helpful across different models and data sets and demonstrate it in Sections 3.4 and 3.5.

3.3 Theoretical properties

We illustrate theoretical properties of the borrowing factors and SSBF. Section 3.3.1 presents properties relevant to the borrowing factors while Section 3.3.2 presents properties of SSBF.

3.3.1 Properties of the borrowing factors

We first show that the borrowing factors are connected to the shrinkage and pooling factors of the Bayesian literature. Currently within the literature, shrinkage and pooling factors are restricted to one-way models. Given data $\boldsymbol{Y}_i \in I\!R^{n_i} \sim N(\alpha_i, \phi_i^2)$, $\alpha_i \sim N(a_0, \sigma^2)$, $i = 1, \ldots, J$, where a_0, ϕ_i, and σ are known, then it can be shown that the posterior

mean, $\hat{Y}_i$, is a balance between the data mean $\bar{Y}_i$ and global mean parameter a_0,

$$\hat{Y}_i = \lambda_i^* \bar{Y}_i + (1 - \lambda_i^*) a_0, \quad \lambda_i^* = \frac{\phi_i^2}{n_i \sigma^2 + \phi_i^2}, \tag{3.8}$$

where λ_i^* is referred to as the pooling factor and $1 - \lambda_i^*$ as the shrinkage factor (Efron and Morris, 1975; Gelman and Pardoe, 2006; Morris, 1983). This has been used to understand information pooling in Bayesian hierarchical models; clusters with less noise ϕ_i^2 or more data (n_i is large) borrow less from other points, while those that borrow more are shrunk towards the shared global mean a_0. However, this understanding is limited in at least two ways: 1) it is limited to the one-way setting and cannot take into account information borrowing for models with multiple levels and 2) in most cases, a_0 is not known and is also informed by $\bar{Y}_i$, so the shrinkage factor in this setting underestimates the total weight placed on $\bar{Y}_i$. As (3.8) shows that all point estimates are shrunk towards the global mean a_0, it is of interest to understand with more granularity how the data availability or ϕ_i^2 affects the estimation of a_0.

By conditioning only on the variance parameters, we obtain the borrowing factors, defined in (3.3), and can decompose λ_i^* into weights on each of the data cluster means $\bar{Y}_j$,

$$\begin{aligned} \hat{Y}_i &= \left(\frac{n_i \sigma^2}{n_i \sigma^2 + \phi_i^2} + \rho_{ii} \right) \bar{Y}_i + \sum_{j \neq i} \rho_{ij} \bar{Y}_j, \\ \rho_{ij} &= \frac{\phi_i^2}{n_i \sigma^2 + \phi_i^2} \frac{\tau_j}{\sum_{j=1}^J \tau_j}, \quad \tau_j \coloneqq \frac{n_j}{n_j \sigma^2 + \phi_j^2}. \end{aligned} \tag{3.9}$$

It shows that $\lambda_i^* = \sum_{j=1}^J \rho_{ij}$. For derivation, see Appendix A.1.1. Instead of one weight placed on the mean parameter a_0, which might not be known, we have J borrowing factors which are placed on sample data means $\bar{Y}_j$. This allows us to more closely examine the contribution of $\bar{Y}_j$ to the global mean a_0, and thus to $\hat{Y}_i$. This contribution is summarized by τ_j, which is monotonically increasing in n_j for $n_j \geq 1$ and monotonically decreasing in ϕ_j. So, the more informative $\bar{Y}_j$ is, with larger n_j or lower noise variance ϕ_j^2, the closer τ_j is to its limit, σ^{-2}. Note $\tau_j \to \sigma^{-2}$ as σ^2 increases; and, as τ_j is finite, as J increases, the input of individual τ_j lessens in comparison to $\sum_j \tau_j$ and $\tau_j / \sum_j \tau_j \to 0$. So, when σ^2 or J is large, the contribution of any individual $\bar{Y}_j$ to the global mean a_0 is low, even under data imbalance.

Note that, as defined in (3.4), the shrinkage factor from (3.9) is the total weight placed on the borrower cluster mean $\bar{Y}_i$ which is $(1 - \lambda_i^*) + \rho_{ii}$ and the pooling factor is $\lambda_i^* - \rho_{ii}$. The ρ_{ii} term accounts for the contribution of $\bar{Y}_i$ to the estimation of global

mean parameter a_0 and so is moved from λ_i^* to $1 - \lambda_i^*$, where λ_i^* and $1 - \lambda_i^*$ are the shrinkage factor the pooling factor when a_0 is known.

Having shown that the borrowing factors are equivalent to λ_i^* and $1 - \lambda_i^*$ in the one-way case, we now show that the properties of the shrinkage and pooling factors in the one-way case generalize to all regression models. In Theorem 3.3.1, we show that all weights sum to 1. As the borrowing factors can be negative, we additionally show in Theorem 3.3.2, that both the shrinkage and pooling factors are always positive and less than 1 for all regression models.

Theorem 3.3.1. *Let response vector $\boldsymbol{Y} \in \mathbb{R}^N$ of a hierarchical linear regression follow a normal distribution as in (3.1), where the N-length vector of ones is in the column span of X_1, $\mathbf{1} \in span(X_1)$. In the Bayesian setting, we assume $f(\Sigma)$ and $f(\phi)$ are some prior densities such that the posterior is proper. The $N \times N$ matrix of borrowing factors, W, is as defined as in (3.3). Then the sum of borrowing factors $\sum_{j=1}^{N} w_{ij}$ for a point estimate $\hat{Y}_i$ is 1 for all $i = 1, \ldots, N$, i.e. $W\mathbf{1} = \mathbf{1}$.*

Theorem 3.3.2. *Under the same setting as in Theorem 3.3.1, let the shrinkage factor be defined as in (3.4). Then given a point estimate $\hat{Y}_i$, $0 < b_{iB_i} \leq 1$ and likewise $0 <= b_{iL_i} < 1$, where b_{iB_i} is the shrinkage factor and b_{iL_i} the pooling factor.*

For proofs, see A.1.2 and A.1.3, respectively. Point estimates $\hat{Y}_i$ can then be seen as balancing the proportion of information coming from the borrower cluster, b_{iB_i}, with the proportion of information from the lenders, b_{iL_i}.

3.3.2 Properties of SSBF

We propose to summarize a point estimate's pattern of information borrowing from lenders using the sum squares of borrowing factors (SSBF), defined in (3.6). SSBF has a number of properties that make it suitable for this purpose. It is lower-bounded by a function of the pooling factor and its relationship to the sample variance of borrowing factors helps to understand higher-level patterns of information borrowing. SSBF and PSSBF are also related to both model uncertainty and metrics of influence analysis. So, they can be thought of as a more granular version of leverage that summarizes the influence lenders have on a particular $\hat{Y}_i$ due to only the data availability. In this section, we illustrate and discuss each of these properties.

SSBF linearly increases with the sample variance of borrowing factors. Let b_{iL_i} be the pooling factor for $\hat{Y}_i$ and n_{L_i} the number of lenders, then the sample variance is

$(n_{L_i} - 1)^{-1} \sum_{j \in L_i} (w_{ij} - b_{iL_i}/n_{L_i})^2$ and

$$\text{SSBF}_i = \sum_{j \in L_i} (w_{ij} - b_{iL_i}/n_{L_i})^2 + \frac{(n_{L_i} - 1)}{n_{L_i}} b_{iL_i}^2. \tag{3.10}$$

For a fixed b_{iL_i}, a larger sample variance indicates more distinctive patterns of information borrowing, where some subset of lenders have higher individual borrowing factors than others. SSBF is lowest when a point estimate borrows equally from all lenders. By splitting SSBF into a set of partial SSBFs, as defined in (3.7), we can identify which groups of lenders have consistently high individual borrowing factors. In extreme cases, disproportionately large individual weight may be placed on a few lenders, meaning a large portion of the point estimate is derived from a handful of lenders. As such, researchers may wish to examine such point estimates with high SSBF more closely.

SSBF has a lower bound. As the first term in (3.10) is non-negative, the second term represents a lower bound for the SSBF. Thus SSBF increases as the pooling factor, b_{iL_i}, increases. When all borrowing factors are non-negative, as in (3.8), SSBF has an upper bound. Using the triangle inequality,

$$\sum_{j \in L_i} w_{ij} = b_{iL_i} \implies \sum_{j \in L_i} w_{ij}^2 \leq b_{iL_i}^2.$$

SSBF is also related to uncertainty for $\hat{Y}_i$. Let response vector $\boldsymbol{Y}$ follow a normal linear regression as in (3.1) and let $\boldsymbol{Y}$ be grouped into clusters $\boldsymbol{Y}_i \in I\!R^{n_i}$, $i = 1, \ldots, J$ such that $\boldsymbol{Y}_i \sim N(x_i'\boldsymbol{\beta}, \phi_i^2)$. Then the point estimate $\hat{Y}_i$ is a weighted sum over clusters of data means,

$$\hat{Y}_i = \sum_{i=1}^{J} b_{ij} \bar{Y}_j,$$

as the same weight is placed on all individual points in $\boldsymbol{Y}_j$. Let $w_{ij} = b_{ij}/n_j$ denote the individual weight placed on a point in cluster j. Knowing only w_{ij} and $\bar{Y}_j$, the central limit theorem states that, for large n_j, $\bar{Y}_j \approx N(x_j'\boldsymbol{\beta}, \psi_j^2/n_j)$, for some variance ψ_j^2. The variance of $\hat{Y}_i$ is then

$$\sum_j w_{ij}^2 n_j \psi_j^2 = b_{iB_i}^2/n_i\psi_i^2 + \sum_{j \neq i} \text{PSSBF}_{ij}\psi_j^2.$$

Higher SSBF then indicates higher uncertainty surrounding $\hat{y}_i$. This is intuitive when linked to how SSBF is proportional to the sample variance and so larger SSBF values

indicate that $\hat{Y}_i$ is borrowing heavily from a relatively small number of lenders. $\hat{Y}_i$ is then more dependent on a smaller set of data points and thus has larger uncertainty. Note, however, that the standard error for $\hat{Y}_i$ also depends on ψ_j, thus SSBF is not a direct measurement of uncertainty but summarizes the uncertainty that is due to the data availability.

SSBF summarizes the total influence, due to data availability, of all lenders on a point estimate. Influence analysis examines those data points which may have a strong effect on the model fit, without which model parameters could be significantly different. This can be determined through cross-validation, withholding small sets of individual data points at a time. Metrics of influence analysis that are based on single-case-deletion cross-validated estimators have been developed, such as Cook's distance (Cook, 1977). Here we discuss two more recent influence analysis metrics in the literature, Parsons and Bao's (2018) retrospective value of sample information and Peña's (2005) influence metric S_i, and their relationship to PSSBF.

Value of information is an approach to outlier and influence analysis within the Bayesian literature that quantifies the value of sample information Y_j using the reduction in loss that results from including Y_j v.s. excluding it. Let response vector $\boldsymbol{Y}$ follow a normal linear regression as in (3.1) and let $\boldsymbol{Y}_j \in I\!R^{n_j} \sim N(x_j'\boldsymbol{\beta}, \phi_j^2)$. The retrospective value of sample information (RVSI) of Y_j on $\hat{Y}_i$ can be approximated as the product of the sum of squared residuals and PSSBF,

$$\text{RVSI}(Y_j|Y_{-j}; \hat{Y}_i) = \frac{\text{PSSBF}_{ij}}{b_{jL_j}^2} \frac{n_j(\hat{Y}_j - \bar{Y}_j)^2}{\phi^4}(1 + O(P_2^{-1})). \tag{3.11}$$

For derivation, see Appendix A.1.4. Partial SSBF is then the portion of the total influence Y_j has on $\hat{Y}_i$ that is due only to the data availability and model definition only, scaled by the squared pooling factor, $b_{jL_j}^2$.

SSBF has a similar relationship to Peña's S_i in the Frequentist literature. Peña's S_i is the squared norm of the standardized vector $\boldsymbol{s}_i = (\hat{Y}_i - \hat{Y}_{i(1)}, \ldots, \hat{Y}_i - \hat{Y}_{i(N)})'$, where $\hat{Y}_{i(j)} = E[Y_i|Y_{-j}]$. S_i has been shown to be able to identify clusters of high-leverage outliers that can be difficult to detect using the usual influence statistics, such as in large high-dimensional data sets. S_i is the sum total of impact all points (lenders and borrowers) have on a point estimate $\hat{Y}_i$. If $\boldsymbol{Y}_j \in I\!R^{n_j} \sim N(x_j'\boldsymbol{\beta}, \phi_j^2)$, S_i can be written as

a linear combination of Cook's distances multiplied by the PSSBF,

$$S_i = \sum_j \frac{\text{PSSBF}_{ij}}{w_{ii}w_{jj}} \bar{D}_j, \quad \bar{D}_j = \frac{\bar{e}_j^2}{ps^2} \frac{w_{jj}}{(1-w_{jj})^2},$$

where D_j is the average Cook's distance for $\boldsymbol{Y}_j$, $\boldsymbol{e} = \boldsymbol{Y} - X\hat{\boldsymbol{\beta}}$, $\bar{e}_j^2 = \boldsymbol{e}_j'\boldsymbol{e}_j/n_j$, $s^2 = \boldsymbol{e}'\boldsymbol{e}/(N-P)$, and P is the dimension of $\boldsymbol{\beta}$. See Appendix A.1.5 for derivation.

Note the similarity between RVSI and S_i–both can be decomposed into a component describing influence of Y_j due to the data availability and a component describing influence due to squared error e_j. For both, PSSBF has a similar role as leverage does to Cook's distance, except it describes the influence of a lender on point estimates. This can be seen by noting that in S_i, PSSBF_{ij}/w_{ii} replaces the leverage term w_{jj} that is in $\bar{D}_j$ and in RVSI, if $n_j = 1$, then $b_{jL_j}^2 = (1-w_{jj})^2$.

These properties make SSBF and PSSBF helpful metrics for summarizing how a point estimate $\hat{Y}_i$ borrows from its lenders. Higher SSBF indicates the point estimate may borrow more from a small number of lenders and therefore has more distinct borrowing patterns. Examining those points with high SSBF can help researchers identify borrowing patterns that are crucial for model estimates.

3.4 Example: Radon

We demonstrate how SSBF and the borrowing factors can explain the impact of data imbalance on model estimates and information borrowing. The Radon data measures the log radon level of 919 houses in Minnesota and contains data on the house's county, the average level of uranium in the county, and whether the house contains a basement. The data are included as part of the `rstanarm` package (Gabry and Goodrich, 2016) via Gelman and Hill (2007).

We model the log radon level of houses in county j and basement status k with a fixed effect intercept a_{0k} based on basement status, fixed effect coefficient a_1 using the log uranium value, and county-specific random intercept α_j:

$$\begin{aligned} \boldsymbol{Y}_{kj} \sim & N(a_{0k} + a_1 u_j + \alpha_j, \phi^2) \\ & a_{0k} \sim N(0, c_k^2), a_1 \sim N(0, c^2) \\ & \alpha_j \sim N(0, \sigma^2), \sigma \sim f(\sigma), \phi \sim f(\phi), \end{aligned} \tag{3.12}$$

where u_j is the `log uranium` value for county j, c_k and c are fixed scalar values $\in \mathbb{R}^+$, rep-

resenting the variances of a_{0k} and a_1 respectively, and α_j denotes county-specific random effects. The model was fit using `rstanarm`, using the default priors and hyperparameters for `stan_lmer`, under which $c_k = 2$, $c = 5.5$, $\phi \sim \text{Exp}(1)$, and $\sigma \sim \text{Exp}(1)$.

The data are imbalanced across counties and basement status. There are 85 total counties with a mean of 10.8 houses per county, a median of 5, and inter-quartile range from 3 to 10. The eight counties with the most houses make up 50% of the data set. Two of the counties contain data on over 100 houses, each making up over 11% of the data. 766 of the houses (83%) do not have a basement and 153 (17%) do. Intuitively, one would expect that counties with fewer houses borrow more from the counties with a larger number of houses. The borrowing factors allow us to explicitly quantify the amount of borrowing for each county and link this to the data availability. For this example, we partition the observations into the following relationship groups:

- the borrower cluster $\boldsymbol{Y}_{kj} \in I\!R^{n_{kj}}$,
- same-county lenders $\boldsymbol{Y}_{k'j} \in I\!R^{n_{k'j}}$,
- same-basement lenders $\boldsymbol{Y}_{kj'} \in I\!R^{n_{kj'}}$,
- lenders in a different county with a different basement status $\boldsymbol{Y}_{k'j'} \in I\!R^{n_{k'j'}}$,

We first compare SSBF to measures of data availability. Figure 3.1A is a contour plot of SSBF with the borrower cluster size n_{kj} and the number of same-county lenders $n_{k'j}$ on the x- and y-axes. As $n_{k'j}$ increases, SSBF increases, which implies that lenders in the same county have large individual weights placed on them. As n_{kj} decreases, SSBF increases, showing that more is borrowed from same-county lenders to compensate for low borrower cluster size. When $n_{k'j} = 0$, SSBF is low regardless of n_{kj}, indicating that none of the remaining lenders has particularly high individual weight placed on them. Borrowing within the same county is then the most distinctive pattern of borrowing that changes with the data availability and is the main contributor to the change in SSBF across data points.

Next, we examine the borrowing factors for the three relationship groups defined earlier. As we are mainly interested in the effects of data availability which corresponds to the basement status and the county effects, we consider the point estimates conditional on a_1. Let $\hat{\mu}_{kj} \coloneqq E[a_{0k} + \alpha_j | a_1, \boldsymbol{Y}] = \hat{Y}_{kj} - \hat{a}_1 u_j$, b_{kj} be the shrinkage factor for $\hat{\mu}_{kj}$, $b_{k'j}$ be the total amount borrowed from $\boldsymbol{Y}_{k'j}$, and $b_{kj'}$ be the total amount borrowed from $\boldsymbol{Y}_{kj'}$.

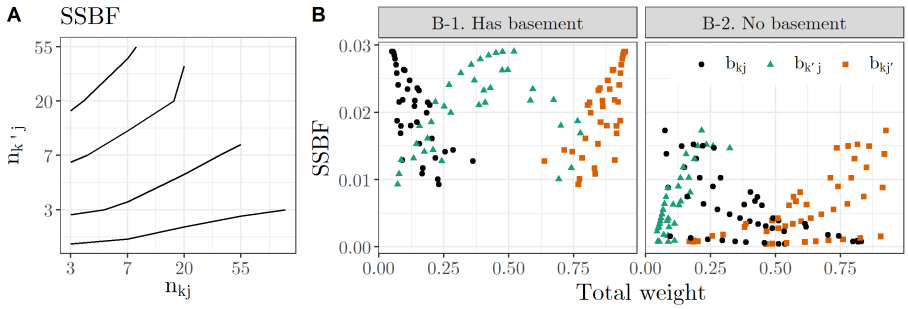

Figure 3.1. For the Radon data, modeled as in (3.12). Panel A is a contour plot of SSBF; contours are based on the mean SSBF for each unique combination of borrower cluster size n_{kj} and same-county lender size $n_{k'j}$. Panel B is a scatter plot of SSBF against the shrinkage factor (b_{kj}) and two borrowing factors corresponding to lenders in the same county and different basement status ($b_{k'j}$) and lenders with the same basement status ($b_{kj'}$).

We notice that $b_{k'j} = -b_{k'j'}$ and only present the borrower cluster and the first two relationship groups. Appendix A.2.1 provides intuition for why $b_{k'j} = -b_{k'j'}$. Figure 3.1B compares the shrinkage factor b_{kj} and borrowing factors $b_{k'j}$ and $b_{kj'}$ to SSBF for all point estimates $\hat{\boldsymbol{\mu}}$. Note that b_{kj} and $b_{kj'}$ are reflections of each other across a vertical line at 0.5 and thus $(b_{kj} + b_{kj'}) = 1$ for all data points. This is because $(b_{kj} + b_{kj'} + b_{k'j} + b_{k'j'}) = 1$ and the summation of the last two terms is zero, as noted earlier. As n_{kj} increases, $b_{kj} \to 1$ and $b_{kj'} \to 0$, and vice versa. Borrowing via the county intercept occurs through relationship groups with the same j and is represented by $b_{kj} + b_{k'j}$. This quantity is typically less than 1. When the number of houses in the county, n_j, is large, $b_{kj} + b_{k'j} \to 1$ and when n_j is small, $b_{kj} + b_{k'j}$ is small, i.e., the model will shrink the amount of borrowing via the county intercept.

In Figure 3.1A, we saw that $n_{k'j}$ is closely related to the SSBF and typically increases as SSBF increases. In Figure 3.1B, we can see that relationship in more detail. As $b_{k'j}$ increases towards 0.5, $n_{k'j}$ increases and so does SSBF. For higher values of $n_{k'j}$ and $b_{k'j}$, SSBF begins to decrease again as the larger number of data points means no single data point gets a large weight.

By examining specific point estimates in Figure 3.1, we can explicitly decompose model estimates into their borrowing factors and link them to both model assumptions and the data availability. One model assumption is that all houses in a county are equally informative of the county-specific effect. As such, the borrowing factors weight both Y_{kj} and $Y_{k'j}$ nearly equally—for the point in panel A with highest SSBF, $b_{k'j} = 0.5$ and $b_{kj} = 0.05$, while $n_{k'j} = 12$ and $n_{kj} = 1$. (The slight difference is because Y_{kj} is also informative for the floor effect, but as there are many other points to inform the floor

effect, it is not necessary to place high additional weight on Y_{kj}.) In other words, $Y_{k'j}$ has much higher total weight placed on it than the borrower cluster's own data, Y_{kj}. This is the case for many of the points in panel B-1, where the shrinkage factor is typically under 0.25 but most $b_{k'j}$s are over 0.25. This is due to the data availability, where fewer houses have basements and so more information is borrowed from those that do. It follows that the reverse is the case in panel B-2, where n_{kj} is typically larger than $n_{k'j}$ and, as such, many of the point estimates have shrinkage factor over 0.25 with most $b_{k'j}$s are under 0.25. Overall, the point estimates with low shrinkage factor and high $b_{k'j}$ are the most affected by this model assumption and are also the counties with the highest data imbalance across basement status.

By comparing SSBF to the data availability in Figure 3.1A, we determined that the number of lenders in the same county is the main contributor to the change in borrowing patterns across data points. By comparing SSBF to the borrowing factors in Figure 3.1B, we were able to link the data availability and model assumptions to patterns of information borrowing. Much of this was intuitive. The borrowing factors simply allow us to place explicit numbers on the degree to which point estimates are affected. In scenarios with more complex models or more severe data imbalance, the intuition may not be so readily available, but the borrowing factors and SSBF can still tell us which point estimates borrow the most from others and which points they borrow from.

3.5 Example: Scottish respiratory disease

Here, we examine a more complex Bayesian hierarchical generalized linear model with spatio-temporal conditional auto-regressive (CAR) intercepts. In Section 3.5.1, we identify the data properties which contribute to higher SSBF and high-level patterns of information borrowing. In Section 3.5.2, we demonstrate how this understanding of model estimates can be used to provide context to influence analysis.

The Scottish respiratory disease data consists of annual observed respiratory-related hospital admissions in the $J = 271$ Intermediate Geographies (IG) of the Greater Glasgow and Clyde health board from 2007 - 2011; the yearly average modelled concentrations of particulate matter less than 10 microns ($\texttt{PM}_{10}$); the average property price in hundreds of thousands of pounds (`Property`); the proportion of the working age population who receive an unemployment benefit called the Job Seekers Allowance (`JSA`); the expected number of hospital admissions, E_{tj}, which is modeled as an offset-term; and the adjacency matrix A, where $A_{ii} = 0$, $A_{ij} = A_{ji} = 1$ if j and i are neighboring districts, and 0 otherwise.

It is available through the `CARBayesST` package in R.

We use the spatio-temporal auto-regressive model in Rushworth et al. (2014), where observed hospital admissions for a year t and IG j are modelled with a Poisson density,

$$\begin{aligned} Y_{tj} &= \text{Poisson}(\eta_{tj} E_{tj}) \\ \log(\eta_{tj}) &= x'_{tj}\boldsymbol{a} + \alpha_{tj}, \end{aligned}$$

where x_{tj} is a vector containing $\texttt{PM}_{10}$, `Property`, and `JSA` values for that year t and IG j; and $\boldsymbol{a}$ is the vector of fixed effects. Within each year, spatial dependence among the corresponding vector of random effects $\boldsymbol{\alpha}_t = (\alpha_{t1}, \ldots, \alpha_{tJ})'$ is modeled with covariance matrix $\sigma^2 Q(\rho_J, A)^{-1}$, where

$$Q(\rho_J, A)^{-1} = \rho_J(\text{diag}(W\mathbf{1}) - A) + (1 - \rho_J)I_J, \quad \rho_J \in [0, 1),$$

which induces spatial auto-correlation and is a special case of a CAR model. Temporal auto-correlation is introduced among the α_t by the conditional density of $\alpha_t|\alpha_{t-1}$:

$$\alpha_t|\alpha_{t-1} \sim N(\rho_T\alpha_{t-1}, \sigma^2 Q(\rho_J, A)^{-1}), j \in \{2, \ldots, T\}.$$

The model is fit using the `ST.CARar()` function in `CARBayesST` with the default priors $\boldsymbol{a} \sim N(0, 100,000)$, $\sigma \sim IG(1, 0.001)$, $\rho_T \sim U(0, 1)$, $\rho_J \sim U(0, 1)$. The resulting posterior means for spatial dependence parameter ρ_J and temporal dependence parameter ρ_T are 0.57 and 0.76, respectively.

As the data are modeled with a Poisson GLMM, the normal priors are not conjugate and the analytical form of (3.2) is no longer available. We instead approximate the data-level Poisson model with a normal distribution having equivalent moments, as described in Daniels and Kass (1998), maintaining conjugacy and a closed-form solution for the borrowing factors. Sample sizes within this data set were large enough that the normal approximation produced closely similar estimates when we compared the normal approximation to actual posterior means (see Appendix A.2.2). In this case, our approximating normal density is

$$\log(Y_{tj}) - \log(E_{tj})|\eta_{tj}, E_{tj} \approx N\left(\log(\eta_{tj}), \eta_{tj}^{-1}\right) \tag{3.13}$$

and we can obtain SSBF along with borrowing factors as described in (3.6). We

derive the joint density of $\boldsymbol{\alpha} = (\alpha_1', \ldots, \alpha_T')'$:

$$s \sim N(0, \sigma^2[(I - \rho_T H)\text{blockdiag}(Q(\alpha, W))(I - \rho_T H)]^{-1})$$
$$H = \begin{bmatrix} \mathbf{0}_{J\times J(T-1)} & \mathbf{0}_{J\times J} \\ \boldsymbol{I}_{J(T-1)} & \mathbf{0}_{J(T-1)\times J} \end{bmatrix},$$

where $I_{J(T-1)} \in I\!R^{J(T-1)\times J(T-1)}$ is the identity matrix, and $\mathbf{0}$ are matrices of 0s with dimensions such that $H \in I\!R^{JT\times JT}$ accounts for the temporal auto-correlation.

For this model, we aggregate the borrowing factors and partial SSBF based on how close the lender is to the borrower, which can be defined both temporally and spatially. The relationship groups are combinations of three spatial and three temporal categories, where the spatial categories are

- the lender is in the same IG, denoted with subscript j_0,
- the lender is in a neighboring IG (j_1),
- or the lender is farther away (j_{2+}),

and the temporal categories are

- the lender is in the same year, denoted with subscript t_0,
- the lender is in 1 year away (t_1),
- the lender is 2 or more years away (t_{2+}),

resulting in 9 total relationship groups.

3.5.1 High-level information borrowing patterns

From the posterior means for spatial dependence parameter ρ_J and temporal dependence parameter ρ_T ($\hat{\rho}_T > \hat{\rho_J}$), we may have some intuition that for point estimate $\hat{Y}_{tj}$, $Y_{t_1,j}$ may have higher weight than Y_{t,j_1}, but it is not clear how other lender groups affect $\hat{Y}_{tj}$ and whether, for example, Y_{t_1,j_1} has noticeable impact on $\hat{Y}_{tj}$ or not. In this section we quantify and compare borrowing across each of the relationship groups to understand which lenders have the most impact on point estimates.

First, we identify what has the largest impact on SSBF and the borrowing patterns. Figure 3.2 illustrates this in two ways. The first, in panel A, is a contour plot of SSBF against two properties of the data, the number of neighbors and the year (this is similar

to the contour plot in Section 3.4, Figure 3.1, which links data availability to the SSBF). The second, in panel B, is a scatter plot of SSBF vs PSSBF which helps to identify which borrowing factors contribute the most to the change in SSBF.

The contour plot links data properties to SSBF and shows that SSBF is the highest for those points at year 2010 with around 90 neighbors. Those points have more potential lenders to borrow from, with a large number of neighboring IGs and two neighboring time points. The scatter plot is a high-level summary of the borrowing patterns and identifies which borrowing factors change the most with SSBF. If PSSBF has a large positive correlation with SSBF, then it is likely that the lenders in that relationship group have high individual weight placed on them. We can see that the borrowing factors for $\boldsymbol{Y}_{t_1 j_0}$ (Figure 3.2B center panel, black points), $\boldsymbol{Y}_{t_0 j_1}$ (Figure 3.2B left panel, green points), and $\boldsymbol{Y}_{t_{2+} j}$ (Figure 3.2B right panel, black points) contribute the most to the change in SSBF, in decreasing order of impact. Correlations between PSSBF and SSBF are $0.94, 0.47$ and 0.36 respectively for each of the relationship groups.

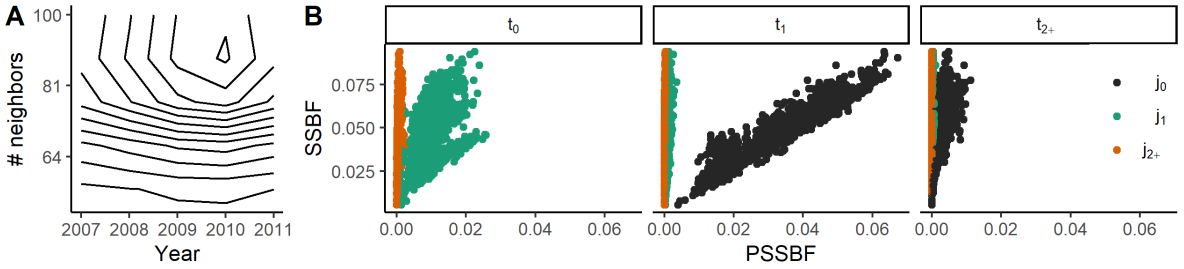

Figure 3.2. Panel A is a contour plot of smoothed SSBF values against the year and number of neighbors for each point. Smoothing is conducted with a Nadaraya-Watson type kernel estimator. Panel B is a scatter plot of SSBF against partial SSBF, where each panel represents a different temporal relationship group (t_0, t_1, t_{2+} for same year, adjacent year, other years, respectively) and colors represent different spatial relationship groups (black for j_0, green for j_1, orange for j_{2+}, corresponding to same IG, neighboring IG, and farther IGs, respectively).

By comparing SSBF to data properties in Figure 3.2A, we determined that point estimates with the highest SSBF values were typically those with a large number of neighbors near the year 2010. The model induces positive correlations on points in neighboring IGs or neighboring years, thus those points that have more neighbors to borrow from have more distinct information borrowing patterns and higher SSBF. We identified which lenders contribute the most to the change in SSBF and thus likely have the highest individual weights placed on them using Figure 3.2B. These relationships may not be readily apparent when examining the posterior mean estimates and the data

alone, but can be determined by examining the borrowing factors which quantify the relative amounts of information borrowing for each of the relationship groups.

More detailed investigation of the relative magnitude of the borrowing factors for each relationship group can be determined by comparing SSBF to the borrowing factors, as in the `ssbf` package Shiny app. A plot of SSBF against borrowing factors is included in the supplementary material, Appendix A.2.2.

3.5.2 Impact of influential points

Influence analysis examines those data points which may have a strong effect on the model fit, without which model parameters could be significantly different. After identifying influential points through the use of a metric such as Cook's distance, RVSI, or S_i, a decision is often made on whether they are outlying, typically based on subject matter considerations and their degree of influence. By examining which point estimates rely the most on these influential points, we can add more context to subject matter considerations of whether to keep or discard the influential points and contextualize their degree of influence on other point estimates. Using SSBF and the borrowing factors, we can understand exactly how an influential point Y_i affects other model estimates $\hat{\mu}_j$ and thus identify those estimates that are most impacted by Y_i.

We identified a set of 11 potentially influential points s using PCA-decomposition of the log case-deletion importance sampling weights, as described in Thomas et al. (2018), which captures both global case influence of an individual point, in terms of distance from the full-data and the case-deleted posterior, and local case influence, through perturbations to the likelihood. Any method which produces estimates of influence for all data points $\boldsymbol{Y}$ can be used.

A point may be influential because of the data availability; in these cases, the covariates corresponding to the point are unique in some way, such as belonging to a rare category or having extreme values. This is most commonly summarized via leverage, essentially the square root of diagonal values of W, where higher values indicate the point has higher impact on model estimates. The point may also be influential because the response value is unexpected in some way under the model. In either case, the points that are most impacted by an influential point are those for which the borrowing factor is higher.

Figure 3.3 consists of boxplots of individual borrowing factors on the 11 influential points, for all model estimates. The boxplots show that the influential points have the most impact on neighboring time points that are in the same IG, with median borrowing

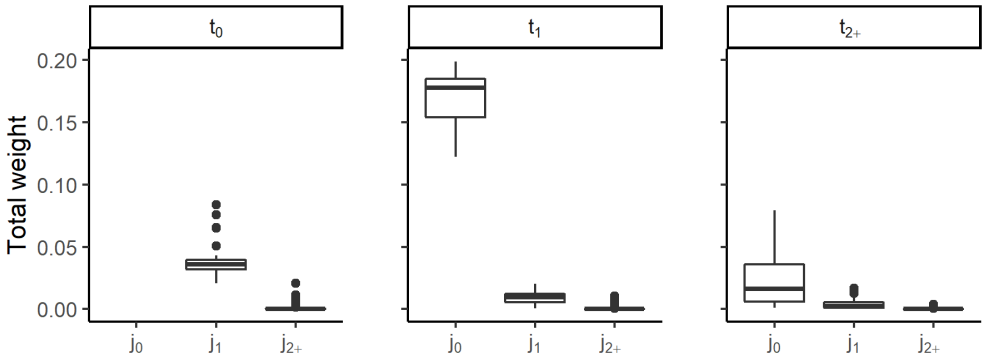

Figure 3.3. Boxplots of total (absolute) weight placed on 11 influential points, split into temporal (t_0, t_1, t_{2+}) and spatial (j_0, j_1, j_{2+}) relationship groups. The plots do not include the shrinkage factor, hence no boxplot for $b_{t_0 j_0}$.

factor near 0.18. The influential points also have a noticeable impact on point estimates for neighboring IGs in the same year and those in the same IG, but more than 1 year away. Both typically have borrowing factors under 0.05. Other relationship groups are less affected, with borrowing factors generally near 0. This is in line with the SSBF vs PSSBF plot in Figure 3.2B, which shows that individual borrowing factors are low for neighboring IGs at the same year. Part of this could be because the temporal dependence is larger than the spatial dependence, based on posterior samples, but a large part of this is likely due simply to data availability. Plots of SSBF against the borrowing factors show that borrowing factors for neighboring IGs at the same year and neighboring years at the same IG are similar in magnitude (see Appendix A.2.2). There are typically a large number of neighboring IGs to borrow from, so less individual weight is placed on each neighbor, lessening the impact of any individual point. There are only one or two neighboring time points that are at the same IG, which leads to higher individual weight placed on those timepoints. We can conclude that although both spatial and temporal dependence in the model is high, influential points will have much greater impact on point estimates from neighboring time points because of the data availability. This can be confirmed by obtaining the weights if the posterior means for ρ_T and ρ_J are switched so that $\rho_J = 0.76$ and $\rho_T = 0.57$, which results in a similar boxplot (see Appendix A.2.2).

By decomposing model estimates using the borrowing factors, we explicitly quantify which point estimates are the most and least impacted by by the 11 identified influential points. We determined that those are the point estimates that are next to an influential point in time, with median borrowing factor around 0.18, followed by those point estimates that are in neighboring IGs, with median borrowing factor under 0.05. Based on the

conclusions from Section 3.5.1, we determined that the relatively low borrowing factors on neighboring IGs was due to the data availability.

3.6 Discussion

Borrowing factors explicitly quantify how the data availability and model specification impact model estimates. We demonstrated this with two examples. In the Radon example, we used both borrowing factors and SSBF over same-county lenders to quantify the impact of data availability on model estimates. In the SRD example, we showed how the number of neighboring lenders affected point estimates and used this understanding to identify lenders that are most impacted by influential points. In both cases, the borrowing factors allowed us to place explicit quantities on relationships that could previously be assumed but would be difficult to verify.

We examined the properties of borrowing factors for point estimates, $\hat{Y}_i$. Researchers may also use the borrowing factors to examine particular coefficients. In this case, the weight matrix W would then be taken as $VX'\Phi^{-1}Y$.

As the dimension of W is often large, we encourage graphical summaries to understand the borrowing factors and SSBF. Graphs can be used to identify both high-level patterns among point estimates as well as providing granular information on a single point estimate. We have found that we can understand model estimates by comparing SSBF to the borrowing factors, partial SSBF, measures of data availability, and model covariates. We provide an R package for creating these plots and an interactive Shiny app for simultaneously displaying multiple plots. Users can select points in any plot, which will then be highlighted and annotated with information across all plots.

With its focus on examining the mechanisms of regression models, philosophically, our method most resembles methods in the explainable machine learning literature, particularly those which allow for integrating domain knowledge (Tsang et al., 2017; Yan et al., 2019). See (Roscher et al., 2020) for a survey and taxonomy of explainable machine learning. The borrowing factors themselves bear the most resemblance in the literature to the pooling factor which, to our knowledge, is the only method in the literature which derives an explicit quantity that describes information borrowing.

Chapter 4
Approximate cross-validated mean estimates for Bayesian hierarchical regression models

4.1 Introduction

Bayesian hierarchical models (BHMs) are often used for their ability to model complex dependence structures while producing probabilistic uncertainty estimates. Except for the simplest of models, BHMs require computationally expensive methods such as Markov Chain Monte Carlo (MCMC) to obtain the posterior density. This has led to many papers which either present new methods to approximate the posterior density (e.g., Kingma and Welling, 2013; Lewis and Raftery, 1997; Rue et al., 2009) or attempt to make current methods more efficient (Bardenet et al., 2017; Korattikara et al., 2014; Quiroz et al., 2019).

The computational cost of a BHM increases by an order of magnitude when cross-validation (CV) is used, where it then becomes necessary to repeat the posterior density estimation process for each CV fold. This can be mitigated by reducing the number of folds—however, when the data are not independent of each other, random K-fold cross-validation selects models which overfit because of the high correlation between test and training data (Arlot et al., 2010; Opsomer et al., 2001). In models with complex dependency structures, it is widely preferred to use a CV scheme that could reduce this correlation, often resulting in a number of folds far beyond standard 10-fold CV.

We address the computational cost of cross-validation by introducing a novel procedure for obtaining Bayesian hierarchical model (BHM) posterior mean estimates under cross-

validation. We focus on regression models (also called latent Gaussian variable models by Vehtari et al. (2016b)) as they are widely used. They include all models where Gaussian hyperpriors are placed on model parameters, such as Gaussian processes and Gaussian Markov random fields. We refer to our procedure as AXE, an abbreviation for (A)pproximate (X)cross-validation (E)stimates. The procedure can be applied to many CV schema, e.g. K-fold, leave-one-out (LOO), and leave-one-cluster-out (LCO). What matters is that the training data is sufficient such that the variance parameters can be well-estimated.

4.2 Approximate cross-validation estimates using plug-in estimators (AXE)

Let $Y \in \mathbb{R}^N$ denote a continuous response vector that follows

$$\begin{aligned} Y|\beta, \Phi &\sim N(X_1\beta_1 + X_2\beta_2, \Phi), \\ \beta_1 \sim N(\alpha_1, C), \quad \beta_2|\Sigma &\sim N(\alpha_2, \Sigma), \\ \Sigma \sim f(\Sigma), \quad \Phi &\sim f(\Phi), \end{aligned} \tag{4.1}$$

where $X \coloneqq \begin{bmatrix} X_1 & X_2 \end{bmatrix} \in \mathbb{R}^{N \times P}$ is the design matrix, $\beta_1 \in \mathbb{R}^{P_1}$, $\beta_2 \in \mathbb{R}^{P_2}$ s.t. $\beta \coloneqq \begin{bmatrix} \beta_1' & \beta_2' \end{bmatrix}' \in \mathbb{R}^P$, $C \in \mathbb{R}^{P_1 \times P_1}$ is positive-definite and typically a diagonal matrix, $\Sigma \in \mathbb{R}^{P_2 \times P_2}$ is positive-definite, and $\Phi \in \mathbb{R}^{N \times N}$ a diagonal positive-definite matrix. We assume $\mathbb{1} \in \text{span}(X_1)$, where $\mathbb{1}$ is the N-length vector of ones, which is satisfied when the fixed effects include a global intercept or set of intercepts which partition the data. We take $\alpha_1 = \alpha_2 = 0$ throughout this paper, without loss of generality. C is treated as fixed, often with large variances, and thus β_1 are referred to as the fixed effects. Random variance hyperparameters such as Σ reflect the dependency among the β_2; the effect is to pool information among related units and shrink them towards a common mean, thus the β_2 are referred to as random effects. Σ can take many forms, so long as it is positive-definite, e.g. models with Gaussian Markov random fields sample from the space of all possible Σ's such that a variable in β_2 is independent of all others, given its neighborhood, while Gaussian processes sample from Σ such that the correlation depends on the (fixed) distance between individual values of β_2.

Posterior means $E[\Sigma|Y]$ and $E[\Phi|Y]$ are denoted as $\hat{\Sigma}$ and $\hat{\Phi}$, respectively. Given a vector Y or matrix X and set of indices $j \subset \{1, \ldots, N\}$, $n_j = |j|$ refers to the cardinality

of set j, $Y_j \in \mathbb{R}^{n_j}$ refers to the vector of entries in Y indexed by j, and $X_j \in \mathbb{R}^{n_j \times P}$ refers to the rows of X indexed by j. For example, j may contain the indices of the test data; then Y_j is the test data response vector. $Y_{-j} \in \mathbb{R}^{N-n_j}$ refers to the vector Y without the entries indexed by j, and likewise for $X_{-j} \in \mathbb{R}^{N-n_j \times P}$, the matrix without the rows indexed by j.

4.2.1 AXE for linear mixed models

When modeling data as in (4.1), the posterior mean for $X\beta$ conditioned on variance parameters has the form

$$E[X\beta|\Sigma, \Phi, Y] \sim XVX'\Phi^{-1}Y, \quad V = \left(X'\Phi^{-1}X + \begin{bmatrix} C^{-1} & 0 \\ 0 & \Sigma^{-1} \end{bmatrix}\right)^{-1}, \tag{4.2}$$

which is analogous to Frequentist linear regression. C^{-1} is taken as the matrix of 0s, which corresponds to the assumption that the fixed effects have infinite variance. Kass and Steffey (1989) show that, for a one-way model, the posterior mean $E[\beta|Y] = E[\beta|Y, \hat{\Phi}_{\text{EB}}, \hat{\Sigma}_{\text{EB}}](1 + \mathcal{O}\left(P_2^{-1}\right)$, where $\hat{\Phi}_{\text{EB}}, \hat{\Sigma}_{\text{EB}}$ denote the Empirical Bayes estimates. They note that $\hat{\Sigma}_{\text{EB}}$ in turn approximates $\hat{\Sigma}$ with order $\mathcal{O}\left(P_2^{-1}\right)$. Conditioning on variance parameters then produces estimates which approximate the posterior mean. The accuracy of this approximation is simple to determine by comparing the conditional expectation $E[X\beta|\hat{\Sigma}, \hat{\Phi}, Y]$ to the posterior expectation $E[X\beta|Y]$.

$$\hat{Y}_j^{\text{AXE}} = E[X_j\beta|Y_{-j}, \hat{\Sigma}, \hat{\Phi}] = X_j \left(X'_{-j}\hat{\Phi}_{-j}^{-1}X_{-j} + \begin{bmatrix} 0 & 0 \\ 0 & \hat{\Sigma}^{-1} \end{bmatrix}\right)^{-1} X'_{-j}\hat{\Phi}_{-j}^{-1}Y_{-j}. \tag{4.3}$$

This shifts the CV problem from probability-based sampling and density estimation methods to maximum likelihood and is likewise $\mathcal{O}\left(N^2P + P^3\right)$ in time for each CV fold. In comparison, Gibbs sampling of the same problem (when available) is $\mathcal{O}\left(SN^3P + SNP^2 + SP^3\right)$ in time for each CV fold, where S is the number of MCMC sampling iterations.

4.2.2 AXE for GLMMs

When the first stage of a BHRM is modeled with a non-normal distribution, the normal priors on the coefficients β are no longer conjugate and the analytic solution from (4.2) is

not available. We can instead approximate the first stage of the GLMM using a normal distribution with equivalent moments, maintaining conjugacy and a closed-form solution for AXE. Take as example a one-way GLMM with stratified response data Y_j, where j denotes the j^{th} group, n_j the size of the group, and $t = 1, \ldots, n_j$ indexes values within the j^{th} group. The data Y_{jt} have some probability density function π and are modeled as a regression through link function f_η so that

$$\begin{aligned} Y_{jt}|X_j\beta &\sim \pi(f_\eta^{-1}(X_j\beta)), \\ \beta|\Sigma \sim N(0,\Sigma), &\quad \Sigma \sim f(\Sigma), \end{aligned}$$

where $E[Y_{jt}|X_j\beta] = f_\eta^{-1}(X_j\beta)$ and $v_j \coloneqq -1/\pi''(f_\eta^{-1}(X_j\beta)) = \text{var}(Y_{jt})$. Then taking the normal approximation with equivalent moments converges as n_j becomes large:

$$\begin{aligned} Y_{jt} &\approx N(f_\eta^{-1}(X_j\beta), v_j) \\ f_\eta(Y_{jt}) &\to N\left(X_j\beta, \frac{v_j}{(f_\eta^{-1})'(X_j\beta)^2}\right) \quad \text{as } n_j \to \infty. \end{aligned} \tag{4.4}$$

Denote in (4.4) the transformed response $f_\eta(Y_{jt})$ as $\tilde{Y}_{jt}$, the variance as $\phi_j(\beta)^2$, the diagonal matrix of $\phi_j(\beta)^2$'s as Φ_β, and the normal approximation as $f_N(\tilde{Y}_{jt}|X_j\beta, \Phi_\beta)$. When Σ and β are unknown, we plug in the marginal posterior mean $\hat{\Sigma} = E[\Sigma|Y]$ and $\hat{\beta} = E[\beta|Y]$. Then,

$$\begin{aligned} X\beta|\tilde{Y}_{jt},\hat{\Sigma},\Phi_{\hat{\beta}} &\sim N_p(m,V) \\ m = XVX'\Phi_{\hat{\beta}}^{-1}\tilde{Y}, &\quad V = \left(X'\Phi_{\hat{\beta}}^{-1}X + \begin{bmatrix} C^{-1} & 0 \\ 0 & \hat{\Sigma}^{-1} \end{bmatrix}\right)^{-1}. \end{aligned}$$

This is also known as expectation propagation where the approximating distribution is Gaussian (Minka, 2013) and has been shown to perform well by Vehtari et al. (2016b). Daniels and Kass (1998) showed that Gaussian expectation propagation is Laplace's method centered at the maximizer of $f_N(\tilde{Y}|X\hat{\beta}, \Phi_{\hat{\beta}})$, while the typical Laplace approximation centers the normal approximation at the mode of the $f_N(\tilde{Y}|\beta, \Phi_{\hat{\beta}})f(\beta|\Sigma)$. Both have $O(\sum_j n_j^{-1})$ error (Daniels and Kass, 1998). To confirm the normal approximation is accurate enough, it is straightforward to numerically compare $E[X_j\beta|\tilde{Y}, \Sigma, \Phi_{\hat{\beta}}]$ to $E[X_j\beta|Y]$.

4.3 Convergence

AXE can be applied to any CV design, of which LOO-CV and LCO-CV are both popular choices. LOO-CV is often used because it maintains as much similarity to the full data model posterior as possible, while still evaluating its ability to fit to new data. LCO-CV is commonly used in models with complex dependency structures, such as spatio-temporal models or models with repeated measures. For both methods, the number of CV folds can grow with the amount of data, which can be computationally expensive, thus they typically benefit the most from CV approximation methods. We focus on proving AXE convergence under LCO-CV, which generalizes to LOO-CV.

Researchers often choose to use a non-random, structured CV design to evaluate BHRMs. In complex models, the dependency among the data can cause K-fold CV to select models which overfit (Arlot et al., 2010; Opsomer et al., 2001). Reducing the amount of correlation between the training and test data typically results in non-random, structured CV. LCO-CV is one example of such a CV design, where the data are partitioned based on the unique values of one or more random intercepts. This is common for mixed-effects models with repeated measures. Under LCO-CV, all repeated measures for a given unit are in the test data, which provides a more realistic estimate of a model's predictive capability for a new unit.

Without loss of generality, we take θ_j as corresponding to a group of random effect intercepts in β_2, defining J total CV folds. Each unique value of θ_j specifies membership in a group. This corresponds to a mixed-effects model with repeated measures. To focus on the random effects which pertain to the CV design, we rewrite $X_1\beta_1 + X_2\beta_2$ in (4.1) as

$$\mu \in I\!R^N, \theta \in I\!R^N \text{ s.t. } X_1\beta_1 + X_2\beta_2 = \mu + \theta, \tag{4.5}$$

where $\mu = X\beta - \theta$ contains all other fixed and random effects which do not pertain to the CV design. The test data consist of those Y_j which are distributed as $N(\mu_j + \theta_j, \phi_j^2)$. Thus θ_j cannot be estimated from the training data, relying instead on samples drawn from hyperparameter Σ.

Even if θ_j cannot be estimated from the training data, $J-1$ other θs remain and the model retains most of the information relevant to the estimation of the variance parameters. This is the mechanism which AXE relies on: the relative stability of the estimation of variance parameters across CV folds, which is formally stated in Theorem 1.

Theorem 4.3.1. *Let response vector $Y \in \mathbb{R}^N$ of a hierarchical linear regression follow a normal distribution as in (4.1), where $\Phi = \phi^2 I$, $\mathbb{1} \in span(X)$, and prior densities $f(\Sigma)$ and $f(\phi)$ such that the resulting posterior is proper. $X\beta = \mu + \theta$ as in (4.5), where θ has J unique values and $\theta_1, \ldots, \theta_J$, and s_j denotes the set of indices such that $\theta_{s_j} = \theta_j \mathbb{1}$, and X_j is made up of identical rows. V is defined as in (4.2). Then $|\operatorname{argmax}_{\Sigma,\phi} f(\Sigma, \phi|Y_{-\theta_j}) - \operatorname{argmax}_{\Sigma,\phi} f(\Sigma, \phi|Y)| \to 0$ as $J \to \infty$.*

Remark 1. One of the appeals of Bayesian modeling is its flexibility and for regression this often occurs through placing different priors on Σ, e.g. variable selection or Bayesian penalized splines. There is also a growing preference for half-t priors over the conjugate inverse-gamma (Gelman et al., 2006; Polson et al., 2012). These considerations make it pragmatic to provide proofs for unspecified $f(\phi^2)$ and $f(\Sigma)$, as long as the resulting posterior is proper.

Corollary 4.3.2. *Let $\hat{\Sigma}$ denote the full-data posterior mean, $E[\Sigma|Y]$, and $\tilde{\Sigma}$ the CV posterior mean over the training data $E[\Sigma|Y_{-j}]$ for CV fold j. Likewise let $\hat{\phi} = E[\phi|Y]$ and $\tilde{\phi} \coloneqq E[\phi|Y_{-j}]$. Under the same conditions as 4.3.1, $E[X\beta|Y_{-j}, \hat{\Sigma}, \hat{\phi}] = E[X\beta|Y_{-j}, \tilde{\Sigma}, \tilde{\phi}](1 + \mathcal{O}(J^{-1}))$.*

For proofs, see Appendix C.1.

4.3.1 Finite sample performance.

In general,will not deviate much from their full-data posterior estimates and AXE will be accurate. This is also when LCO-CV is the most expensive and AXE provides the most benefit. In our experience with AXE, even when posterior densities are wide, we have found AXE can perform well (e.g. Radon subsets data, see Figure 4.1) and the cases where AXE is inaccurate are typically those with a low number of groups or severe data imbalance such that the test data are critical for the estimation of Σ or ϕ.

Intuitively, when the number of groups is low, then running MCV may not be of as much concern. If one needs to reassure the validate of AXE, we suggest running MCV on a randomly selected subset of CV folds, stratified into groups. By comparing the MCV point estimates to the AXE approximation, we can determine how similar the conclusions derived from the AXE approximation are to those using the MCV estimates. One common model evaluation criteria is the root mean square error (RMSE), where the error is the difference between $E[Y_j|Y_{-j}]$ and Y_j. We take the log ratio of AXE-approximated

RMSE to ground-truth MCV RMSE,

$$\text{LRR}_j = \log\left(\sum_{i=1}^{n_j} \frac{(\hat{Y}_{ij}^{(\text{AXE})} - Y_{ij})^2}{(\hat{Y}_{ij}^{(\text{MCV})} - Y_{ij})^2}\right), \tag{4.6}$$

where LRR_j is calculated separately within each CV fold j to obtain more fine-grained comparisons. We also use LRR to compare AXE approximations to other LCO methods in Section 4.6 by replacing $\hat{Y}_{ij}^{(\text{AXE})}$ in (4.6) with the alternative LCO method's point estimate for Y_{ij}.

If exchangeability is assumed between groups, then the sample mean and variance of LRR_j's can provide inference for the expected accuracy of the approximation. When the variance parameters Σ and ϕ are not well-estimated, they are less likely to be consistent across CV folds; in these cases, we expect a large standard deviation among the AXE LRR_j's across CV folds due to the instability of the variance parameter posterior means. In practice, if the mean or standard deviation of LRR_j's is over 0.25 (a roughly 25% difference between AXE and MCV RMSEs), we recommend running MCV. We refer this additional validation step in our calculations of time cost as AXE+ in Section 4.6.

4.4 Existing leave-cluster-out CV approximation methods

We compare AXE to two types of LCO methods: ghosting (GHOST) (Marshall and Spiegelhalter, 2003) and integrated importance sampling (iIS) (Li et al., 2016; Vanhatalo et al., 2013; Vehtari et al., 2016b). We use the notation for LCO-CV in (4.5), where groupgro-specific random effects are denoted as θ_j and μ denotes the remainder of the mean estimate. All LCO methods are summarized in Table 4.1.

Ghosting draws $\tilde{\theta}_j^{(s)}$ from $f(\theta_j|\theta_{-j}^{(s)}, \mu^{(s)}, \Sigma^{(s)}, Y)$ for each posterior sample s. The S total "ghost" samples are then used as an approximation of $f(\theta_j|Y_{-j})$. If θ_j and θ_i, $i \neq j$, are independent given Σ, then the ghost samples $\tilde{\theta}_j^{(s)}$ are simply drawn from $\theta_j|\Sigma^{(s)}$. This mimics the effect of treating the held-out test data Y_j as an unknown groupgro. Using the full-data posterior densities for μ and Σ directly does introduce bias; if $\theta_j \sim N(0, \sigma^2)$, for some parameter σ^2, the ghosting estimate for $E[Y_j|Y_{-j}]$ is $E[\mu|Y]$ rather than $E[\mu|Y_{-j}]$.

iIS methods integrate out the groupgro-specific effects θ_j using the importance sampling weights of Gelfand et al. (1992). Importance sampling (IS) approximates the target training-data posterior density β, ϕ by re-weighting full-data posterior samples.

These weights can then be used to obtain mean estimates by averaging over the weighted posterior samples. The weights $w_j^{(s)}$ for a sample s are proportional to a ratio of the two densities,

$$w_j(s) = \frac{1}{f(Y_j|\phi^{(s)}, \beta^{(s)})} = \frac{1}{f(Y_j|\phi^{(s)}, \beta^{(s)}, Y_{-j})} \propto \frac{f(\beta^{(s)}, \phi^{(s)}|Y_{-j})}{f(\beta^{(s)}, \phi^{(s)}|Y)}.$$

The equality follows from the independence of the Y_j's given β, ϕ. Mean estimates are obtained by averaging over the posterior samples, weighted by $w_j^{(s)}$, $\hat{Y}_j^{IS} = \left(\sum_{s=1}^S w_j^{(s)}\right)^{-1} \sum_{s=1}^S X\beta^{(s)} w_j(s)$. Dividing over the sum of w_j's is a correction for knowing the ratio only up to a constant (Gelfand, 1996).

We differentiate between two iIS methods in the literature: iIS which integrates out only θ_j (Li et al., 2016; Vanhatalo et al., 2013) and iIS which integrates over all mean parameters μ and θ (Vehtari et al., 2016b). We refer to the former as iIS-C, for "CV-specific effects only" integration and the latter as iIS-A for integrating over "all coefficients". The weights and estimates under iIS-C are

$$\begin{aligned}
&\text{iIS-C: } w_j^{\text{iIS-C}}(s) = \frac{1}{f(Y_j|\theta_{-j}^{(s)}, \phi^{(s)}, \mu^{(s)}, \Sigma^{(s)})}, \quad \hat{Y}_j^{\text{iIS-C}} = \frac{\sum_{s=1}^S w_j^{\text{iIS-C}}(s) E[X\beta|\theta_{-j}^{(s)}, \phi^{(s)}, \Sigma^{(s)}, \mu^{(s)}]}{\sum_{s=1}^S w_j^{\text{iIS-C}}(s)}. \\
&\text{iIS-A: } w_j^{\text{iIS-A}}(s) = \frac{1}{f(Y_j|\phi^{(s)}, \Sigma^{(s)}, Y_{-j})}, \quad \hat{Y}_j^{\text{iIS-A}} = \frac{\sum_{s=1}^S w_j^{\text{iIS-A}}(s) E[X\beta|\phi^{(s)}, \Sigma^{(s)}, Y_{-j}]}{\sum_{s=1}^S w_j^{\text{iIS-A}}(s)}.
\end{aligned} \tag{4.7}$$

Both IS and iIS methods are asymptotically unbiased for LCO CV, but in practice, IS tends to perform worse (Li et al., 2016; Merkle et al., 2019), as the posterior is a biased approximation of the target density $f(\mu, \theta, \phi|Y_{-j})$, and the support of the posterior density in finite samples can be much smaller than that of the target density (Li et al., 2016). The larger the difference between the two densities, the larger the variance of importance sampling weights and estimates (Owen, 2013). In extreme cases, large importance sampling weights on only a few points may dominate the estimate, leading to an unreliable estimate.

The variance of importance sampling weights also increases when the target density has heavier tails than the proposal, which is the case for cross-validation. The more test data is removed, the heavier the tail for $f(\Sigma, \phi|Y_{-j})$ in comparison to $f(\Sigma, \phi|Y)$. If Y_j is particularly important for the estimation of $f(\mu|Y)$, then removing Y_j can mean $f(\mu|Y_{-j})$ is quite different from $f(\mu|Y)$, which may again result in high variance for iIS-C. Conversely, ghosting produces biased mean estimates, but may have smaller variance

than importance sampling. When $E[\theta_j|\theta_{-j}, \Sigma] = 0$ and $E[\mu|Y]$ is similar to $E[\mu|Y_{-j}]$, ghosting performs well at approximating LCO-CV mean estimates regardless of the amount of test data, n_j, where ghosting in many cases performs better than iIS.

In the GLMM case, analytical forms for the importance weights are no longer available. We use the normal approximation in (4.4) to obtain approximate log-likelihood values. For iIS-C, we additionally used Monte Carlo simulation to obtain the log-likelihood values by drawing 200 samples from $\theta_j|\theta_{-j}, \Sigma$ and averaging the likelihood of $f(Y_j|\theta_j, \phi)$ (Li et al., 2016). We found little difference in the results, confirming the accuracy of expectation propagation as noted by Vehtari et al. (2016b).

We use Pareto-smoothed importance sampling (PSIS, Gelman et al. (2014), `R` package `loo`) to stabilize the importance weights and ensure finite variance for both iIS methods. In general, we found that Pareto-smoothed iIS performed similarly to or improved upon iIS without smoothing.

Table 4.1 summarizes the differences in assumptions among the LCO methods for approximating mean estimates. While AXE relies only on the expected values of Σ and ϕ being similar, other LCO methods rely on the densities being the same or similar. Note that GHOST has the strongest assumptions, but if $E[\theta_j|\theta_{-j}, \Sigma] = 0$, the sole assumption it relies on is that $E[\mu|Y] = E[\mu|Y_{-j}]$. However, GHOST produces biased estimates, while AXE and iIS do not.

Table 4.1. Posterior distribution assumptions and computational complexity of approximating $E[Y_j|Y_{-j}]$ for each LCO method. Cost of Gibbs sampling for equivalent MCV problem is $\mathcal{O}\left(S(N^3P + NP^2 + P^3)\right)$, where N = total number of data points Y, P = number of coefficients β, S = number of MC samples, n_j = size of test data for j^{th} CV fold.

Method	$f(\Sigma, \phi\|Y)$ vs $f(\Sigma, \phi\|Y_{-j})$	$f(\mu\|Y)$ vs $f(\mu\|Y_{-j})$	Bias	Time
AXE	$E[\Sigma, \phi\|Y] = E[\Sigma, \phi\|Y_{-j}]$	N/A	No	$\mathcal{O}\left(J(N^2P + P^3)\right)$
GHOST	$f(\Sigma, \phi\|Y) = f(\Sigma, \phi\|Y_{-j})$	$f(\mu\|Y) = f(\mu\|Y_{-j})$	Yes	$\mathcal{O}\left(SJP^3\right)$
iIS-C	$f(\Sigma, \phi\|Y) \approx f(\Sigma, \phi\|Y_{-j})$	$f(\mu\|Y) \approx f(\mu\|Y_{-j})$	No	$\mathcal{O}\left(SJP^3\right)$
iIS-A	$f(\Sigma, \phi\|Y) \approx f(\Sigma, \phi\|Y_{-j})$	N/A	No	$\mathcal{O}\left(SJ(N^2P + P^3)\right)$

Table 4.1 also includes each method's computational complexity, with derivations in Appendix C.2. Note that all methods except AXE require separate calculations for all S MC samples, thus AXE is often the fastest LCO method. Manual cross-validation (MCV), when using Gibbs sampling of the same problem (if available), is the most expensive with complexity $\mathcal{O}\left(S(N^3P + NP^2 + P^3)\right)$. However, many of the examples in Section 4.6 were fit using STAN, which uses Hamiltonian Monte-Carlo (Carpenter et al., 2017; Girolami and Calderhead, 2011), where instead of the proposal distribution

being a Gaussian random walk, proposal samples are generated along the gradient of the joint density. This allows for more efficient sampling and much shorter run times. In our examples, iIS-A was often the method which took the longest to run, not MCV. iIS-C and GHOST can also be computationally expensive, typically due to inversion of Σ that is necessary to obtain $E[\theta_j|\theta_{-j}, \Sigma]$ which contributes the $\mathcal{O}(P^3)$ term. This cost can be reduced greatly if this expectation is $E[\theta_j|\theta_{-j}, \Sigma] = 0$ and $\Sigma = \sigma^2 I$ for some hyperparameter σ, $\mathcal{O}(SNJ)$, which is the case when the θ_j, $j = 1, \ldots, J$, are independent and identically distributed. This brings both methods closer to AXE in terms of computing cost; however, as we show in Section 4.6, AXE is in general faster and more robust.

4.5 Example data sets and models

We use publicly available data to compare AXE and the LCO methods described in Section 4.4 to manual cross-validation (MCV). This section describes each of the data sets and models in detail, with all results compiled and described together in Section 4.6.

4.5.1 Eight schools

The eight schools data comes from a meta-analysis conducted by Rubin (1981) on the effects of coaching on verbal SAT scores and appears frequently in the literature. The data consist of mean y_j and standard error t_j of treatment effects from school j, with a total of eight schools. The data are modeled as a one-way linear mixed effects model,

$$y_j \sim N(\mu + \theta_j, t_j^2), \quad \theta_j \sim N(0, \sigma^2), f(\mu) \propto 1, f(\sigma) \propto 1$$

where μ and σ are scalar values with improper uniform priors. In this one-way model, μ corresponds to $X_1\beta_1$ where X_1 is a vector of 1's, while θ_j corresponds to $X_2\beta_2$.

We re-create a scenario derived from Vehtari et al. (2016a), where the eight y_j are multiplied by a data scaling factor α—since t_j are given and fixed, this has the effect of increasing the variance Σ and decreasing the amount of data pooling. The model then becomes:

$$\alpha y_j \sim N(\alpha\mu + \alpha\theta_j, t_j^2), \quad \alpha\theta_j \sim N(0, \alpha^2\Sigma), f(\mu) \propto 1, f(\alpha\Sigma) \propto 1.$$

For each scaling factor $\alpha \in \{0.1, 0.2, \ldots, 3.9, 4.0\}$, cross-validation is conducted by

with-holding each of the y_j in turn. The CV design is then both LCO-CV and LOO-CV because we observe one mean estimate per school. We fit the models in STAN, running 4 chains of 2000 samples each, with the first 1000 as burn-in. For three of the α values $(2.2, 2.4, 2.8)$, MC diagnostics indicated that the chains may not have converged and we increased the number of burn-in to 4000, for 5000 total samples.

4.5.2 Radon

The Radon data measures the log radon level of 919 houses in Minnesota and contains data on location (`county`), the level of uranium in the county (`log uranium`), and whether the house contains a basement (`basement`). It is included as part of the `rstanarm` package (Gabry and Goodrich, 2016) via Gelman and Hill (2007).

We examine three models where all three define response vector Y as the `log radon` level of the house and the `county` covariate as a random effect with $\texttt{county} \sim N(0, \Sigma)$.

$$\begin{aligned}
&\text{Model 1: } Y \sim N(a_0 + \texttt{county}, \phi^2),\\
&\text{Model 2: } Y \sim N(\texttt{basement} + \texttt{county}, \phi^2),\\
&\text{Model 3: } Y \sim N(\texttt{basement} + \log \texttt{uranium} + \texttt{county}, \phi^2).
\end{aligned}$$

Models were fit using the `rstanarm` package Gabry and Goodrich (2016), using the default priors for `stan_lmer`. Cross-validation was performed over counties, with each loop removing all houses within one county as test data. Then using the notation from (4.5), θ corresponds to the `county` random effect, which are also $X_2\beta_2$, and μ corresponds to $X_1\beta_1$, which correspond to a_0 for Model 1, `basement` for Model 2, and `basement` + `log uranium` for Model 3. There are 85 counties in the data with a median of 5 houses per county. Two of the counties contain data on over 100 houses, each making up over 11% of the data.

4.5.3 Radon subsets

The Radon subsets data is a set of simulations where the test data are fixed as the 23 samples from the county of Olmsted and the training data are a randomly selected subset of counties such that the total number of counties J is $\{3, 4, 6, 9, 12\}$ (including Olmsted) and the training data size is approximately $\{77, 58, 46, 38, 32\}$, which corresponds to approximate test data proportions $\delta = \{0.3, 0.4, 0.5, 0.6, 0.7\}$. For all combinations of J and δ, we derive the AXE and MCV values for at most 60 different iterations of training

data (for $J = 3$ and $\delta = 0.3$, there are only 35 iterations available due to the data availability), using the three models in subsection 4.5.2. Models were again fit using the `rstanarm` package with the default priors for `stan_lmer`.

The data imbalance among counties can lead to certain counties being over-represented across the sets of training data. To mitigate this, each training data set is restricted to have a unique combination of county sizes. The combinations are such that they are within 10% of the target training data size $n_{\text{target}} \coloneqq 23(1-\delta)$, meaning that n_{-i} is within $n_{\text{target}} \pm 0.1 n_{\text{target}}$. The sample is inversely weighted by how far the training data set size is from n_{target}. Within each of the 60 combinations, all matching counties are found and one final combination is sampled.

Model 3 contains both the fixed `floor` effect and the continuous fixed effect for `log uranium`, which is defined at the county level only. For the full Radon data, that means there are 84 unique values of `log uranium` in the training data; for the Radon subsets data, the number of unique values ranges from 2 to 11.

4.5.4 Esports players (ESP)

The Esports player data consist of counts of professional player statistics from the popular video game "League of Legends" for players in the North American League Championship Series from January 2020 - June 2020. In a game, two teams of five players compete to capture the opposing team's base. The data include the player's name (`player`); the player's team name (`team`); the player's position on the team (`position`); the name of the player's character in-game (`champion`); log earned gold per minute (`log_egpm`, continuous); log damage per minute (`log_dpm`, continuous), and the player's kills in the game (`Y`). There are 73 unique players, 10 unique teams, 5 unique positions, and 108 unique champions. The data are publicly available at `oracleselixir.com`, which also contains many other in-game statistics.

We model the number of kills a player p achieves in a game g on champion c as a Poisson GLMM. Let design matrix X_1 corresponding to the fixed effects consist of a vector of 1's, binary indicator vectors for `team`, binary indicator vectors for `position`, `log_dpm`, and `log_egpm`, and let $X_1(pgc)$ correspond to the row in X_1 with player p, game g, and champion c:

$$\begin{aligned} Y_{pgc} &\sim \text{Poisson}(\eta_p) \\ \log(\eta_{pgc}) &= X_1(pgc)\boldsymbol{a} + \alpha_p + \alpha_c, \end{aligned}$$

where $\boldsymbol{a}$ is the vector of fixed effect coefficients, α_p corresponds to a player-specific random intercept, and α_c is a champion-specific random intercept which is crossed with players. Players are typically nested within `position` and `team`, although 5 players within the data set are represented with more than one team.

The model was fit in `rstanarm` using default priors for `stan_glmer`. Cross-validation folds are defined by the player-specific random intercept α_p. Using the notation in (4.5), μ corresponds to $X_1\boldsymbol{a} + \alpha_c$. There are 73 total players with a median number of 33 games within the data is 33, a minimum of 2, and a maximum of 56. The AXE approximation is as described in (4.4), where $\tilde{Y} = \log(E[Y_j|Y])$ and $\Phi = \frac{1}{E[X\beta|Y]}$.

4.5.5 Scottish Lip Cancer (SLC)

The Scottish Lip Cancer data consist of total observed male lip cancer counts collected over the time period 1975-1980 in $J = 56$ districts of Scotland; the number of expected cases, E_j, calculated based on standardization of "population at risk" across different age groups; the percent of population employed in agriculture and forestry m_j; and an adjacency matrix A, where $A_{jj} = 0$, $A_{ji} = 1$ if j and i are neighboring districts, and $A_{ji} = 0$ otherwise. It is available through the `CARBayesST` package (Lee et al., 2018) in `R`.

The number of expected cases E_j is used as an offset in a Poisson GLMM of male lip cancer counts where

$$Y_j|\eta_j, E_j \sim \text{Poisson}(\eta_j E_j)$$

and $\log(\eta_j)$ contains the fixed and random effects of the GLMM. In this scenario, the fixed effects consist of a grand mean intercept and term linear in m_j. The random effects consist of district-level random intercepts, which are modeled such that each θ_j is dependent on the values of its neighbors:

$$\begin{aligned}\theta_j|\theta_{-j} &\sim N(\rho_J \sum_{i=1}^{n} A_{ij}\theta_i, \sigma^2), \quad \rho_J \in [0,1] \\ \log(\eta_j) &= a_0 + a_1 m_j + \theta_j.\end{aligned}$$

The value of ρ_J controls the spatial dependence among neighboring districts, where 0 indicates no spatial dependence. The joint distribution of the θ_j simplifies to:

$$\theta_j \sim N(0, \sigma^2(\text{diag}(A\mathbb{1}) - \rho_J A)^{-1}).$$

The covariance of the θ_j is a simplified version of what is called a proper conditional auto-regression (CAR). As the θ_j, given neighboring θ_k, are independent of all remaining θ, this is also a Gaussian Markov random field.

The cross-validation folds are defined by the districts, thus for this example, LCO-CV is equivalent to LOO-CV. θ_j here corresponds to θ_j using the notation of (4.5) and μ_j corresponds to $a_0 + a_1 m_j$. The AXE approximation is as described in the normal approximation of (4.4), where $\tilde{Y} = \log(E[\eta_j|Y])$, to account for the additional offset term which is not modeled by the GLMM.

4.5.6 Scottish respiratory disease (SRD)

We use the model and data as described in Chapter 3.5, where observed hospital admissions for a year t and intermediate geography (IG) j are modelled as a Poisson GLMM:

$$\begin{aligned} Y_{tj} &= \text{Poisson}(\eta_{tj} E_{tj}) \\ \log(\eta_{tj}) &= x'_{tj}\boldsymbol{a} + \alpha_{tj}, \end{aligned}$$

where x_{tj} is a vector containing $\texttt{PM}_{10}$, `Property`, and `JSA` values for that year t and IG j and $\boldsymbol{a}$ is the corresponding vector of fixed effect coefficients. The joint density of $\boldsymbol{\alpha} = (\alpha'_1, \ldots, \alpha'_T)'$ is

$$\begin{aligned} \boldsymbol{\alpha} &\sim N(0, \sigma^2[(I - \rho_T H)\text{blockdiag}(Q(\rho_J, W))(I - \rho_T H)]^{-1}) \\ H &= \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \boldsymbol{I}_{J(T-1)} & \mathbf{0} \end{bmatrix}, \end{aligned}$$

where ρ_T is the temporal dependence parameter, ρ_J the spatial dependence parameter, $I_{J(T-1)} \in I\!R^{J(T-1)\times J(T-1)}$ is the identity matrix, and $\mathbf{0}$ are matrices of 0s with dimensions such that $H \in I\!R^{JT\times JT}$ accounts for the temporal auto-correlation.

Cross-validation is conducted along the $J = 271$ IGs. The AXE procedure is as described in the normal approximation of (4.4), where $\tilde{Y} = \log(E[\eta_j|Y])$, to account for the additional offset term which is not modeled by the GLMM. To save computation time and avoid inverting $V_{-j} \in I\!R^{JT\times JT}$ for each cross-validation fold, we numerically solve for V given the full data and use the Sherman-Morrison matrix equations to obtain

V_{-j}:

$$\begin{aligned} V_{-j} &= (V^{-1} - X_j' \Phi_j^{-1} X_j)^{-1} \\ &= V + V X_j' (\Phi_j - X_j V X_j')^{-1} X_j V. \end{aligned}$$

4.6 Results

Table 4.2 is a high-level summary of the example data sets and models. The first three examples are linear mixed models (LMMs); the last three are generalized linear mixed models (GLMMs). The first four examples include models where θ are independent and identically distributed (i.i.d.) and Σ is a diagonal matrix, while θ in the last two examples are Gaussian Markov random fields (GMRFs).

Table 4.2. Summary of data set and model properties. J = number of CV folds and dimension of θ, N = dimension of response vector Y, n_j = size of test data in CV fold.

Section	Data	Model	θ	J	N	Max n_j
4.5.1	Eight schools	LMM	i.i.d.	8	8	1
4.5.2	Radon	LMM	i.i.d.	85	919	116
4.5.3	Radon subsets	LMM	i.i.d.	3 - 12	59 - 100	23
4.5.4	Esports (ESP)	GLMM	i.i.d.	73	2160	56
4.5.5	Lip cancer (SLC)	GLMM	GMRF	56	56	1
4.5.6	Respiratory disease (SRD)	GLMM	GMRF	271	1355	5

The Eight schools data includes transformations of the response designed to reduce information borrowing as a data scaling factor, α, increases. There are 40 unique α and all results are over the 40 different transformed data sets. The data can be found directly in the original paper (Rubin, 1981). Results for the Radon subsets data are also aggregated over multiple simulations which were designed to examine the performance of AXE and other LCO methods under varying number of groups or percentage of test data. Using the Radon data, we fixed a particular group as the test data and randomly selected $J-1$ other groups as the training data, $J \in \{3, 4, 6, 9, 12\}$, such that the size of test data was some proportion δ of the full data, $n_j = \delta N$, $\delta \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$. For each combination of J and δ, multiple subsets are selected and evaluated for a total of 1,475 simulated subsets. The Radon data are available under a GPL(>=3) license. ESP is publicly available under a CC-by-SA 3.0 license; the data we used are included in

the supplemental material, along with Radon subsets. SLC and SRD are available under a GPL-3 license. All data sets and models are described in detail in Appendix 4.5.

We compare across LCO methods by calculating LRR_j for each method and CV fold j, as in (4.6).

Figure 4.1 contains boxplots of LRR for each LCO method and data set. LRR for AXE approximations are consistently near 0, with lower variance than other LCO methods. Each of the other methods performs well for one or more data sets, but none perform well consistently across all data sets.

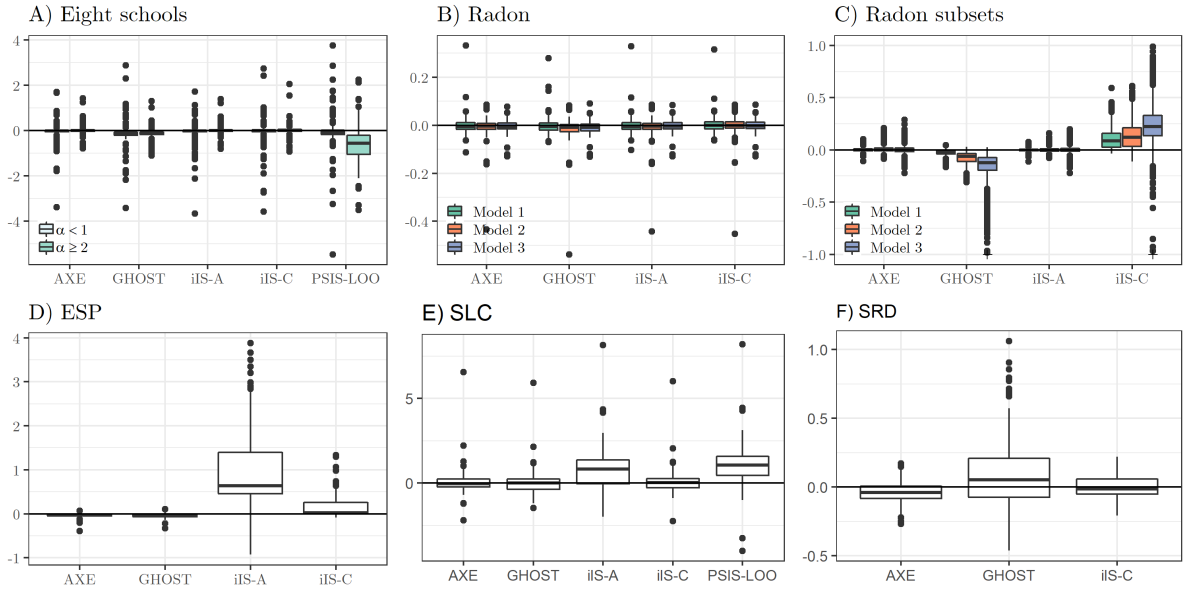

Figure 4.1. Box plots of log RMSE ratios (LRRs) for each LCO method and data set. LRRs, defined in (4.6), are calculated for each CV loop. Horizontal bars are the 25%, 50%, and 75% percentiles of LRR. Vertical lines span the remainder of the data up to 1.5 times the height of the box. Points outside this span are individually annotated with dots. In panel C, a cross ('+') is placed at -1 for those models and LCO methods with a trailing tail of large negative LRRs.

Panel A of Figure 4.1 includes a comparison to PSIS-LOO, as $n_j = 1$ for each CV fold. For larger α values over 2, PSIS-LOO's accuracy decreased, while the LCO approximation methods stayed consistently accurate All LCO methods do have a few cases with high LRR for the Eight Schools data; for the LCO methods, these typically correspond to cases with low MCV RMSE and often a low data scaling factor.

Panel B of Figure 4.1 contains results for the full Radon data, which is a relatively rich data set in comparison to the number of parameters to estimate, and all LCO methods perform well here, with the majority of LRRs varying between -0.2 and 0.2 (note that

the range of y-axis varies among panels). A few points have relatively large |LRR|; these correspond to the same two CV folds across all LCO methods and are cases where the data are particularly informative for the random effect variance. Model 1 is a simple one-way LMM (see Appendix 4.5 for details). Models 2 and 3 each add new fixed effects; model 2 adds an indicator covariate and model 3 adds a continuous covariate.

Panel C of Figure 4.1 contains results for the Radon subsets data. For Model 3, GHOST and iIS-C produce LRRs < -1 which are not presented on the figure: 1.1% for GHOST and 0.3% for iIS-C. Both methods have larger absolute LRR than AXE and iIS-A. In short, this is due to differences between $E[\mu|Y]$ and $E[\mu|Y_{-j}]$ and data imbalance. As the support of samples for $\mu|Y$ does not sufficiently match the support for $\mu|Y_{-j}$, both GHOST and iIS-C perform worse for models 2 and 3. It is interesting to note that GHOST tends to under-estimate RMSE for the Radon subsets data, while iIS-C over-estimates RMSE. As GHOST uses the full-data posteriors directly, it follows that it produces over-optimistic results.

All methods except iIS-A perform similarly in panels D and E of Figure 4.1. Diagnostics indicated that the instability of iIS-A is due to the heavier tails of the training data density $f(\Sigma|Y_{-j})$ that lead to a large variance of the importance weights. iIS-C is not as impacted because it includes information from $\mu|Y$, which in this scenario produced more stable estimates. The relatively good performance of AXE and GHOST in panel D suggests that posterior expectations for μ and Σ were relatively stable across CV folds. In panel E, the variance of LRRs for each method is larger than preceding panels (see Table C.1, in Appendix C.3). This is in part due to small sample sizes, where many points have counts under 5, so the normal approximation is less accurate. All methods except PSIS-LOO use the normal approximation; however, all methods except iIS-A still performed better than PSIS-LOO, which does not use the normal approximation.

In panel F of Figure 4.1, LRRs for iIS-A are omitted due to the amount of time running iIS-A would have required (see Table 4.3, with total computation time for all methods). Among the three remaining LCO approximation methods, GHOST performs the worst, with approximate RMSE values up to 2.7 times that of MCV. Here, $E[\theta_j|\theta_{-j}, \Sigma] \neq 0$ and is a function of both Σ and θ_{-j}. Thus any difference in the posterior densities $f(\Sigma|Y)$ and $f(\theta_{-j}|Y)$ versus their training data counterparts can increase error in the GHOST estimate. As iIS-C takes into account the difference between $f(\Sigma|Y)$ and $f(\Sigma|Y_{-j})$, it is able to perform better than GHOST.

Figure 4.2A contains scatter plots of the AXE approximation for each data point Y_{ji} against actual MCV values for all examples. The vast majority of points lie on or near

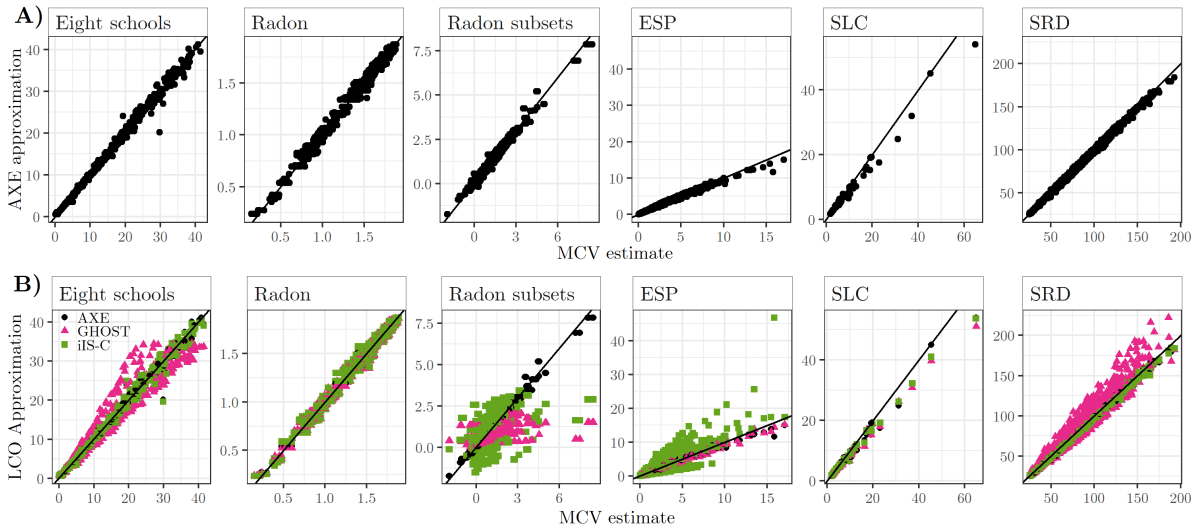

Figure 4.2. Scatter plots comparing the LCO approximation to ground-truth MCV estimate for each data point, model, and data set. Panels in row A compare the AXE approximation $\hat{Y}_{ji}^{\text{AXE}}$ against the MCV estimate for $E[Y_{ji}|Y_{-j}]$. Panels in row B add points with GHOST (pink triangle) and iIS-C (green square) approximations, along with AXE (black circle). Each point in a grid represents one point in the corresponding data set(s) and model(s).

the 45-degree line. In many cases the AXE approximation is point-by-point equivalent to MCV estimates. This is in contrast to the scatter plots in Figure 4.2B, which contain AXE approximations as well as GHOST and iIS-C, both of which performed relatively well, based on Figure 4.1. iIS-C has much higher variance than AXE for the Radon subsets and ESP data sets, with approximations that lie farther from the 45-degree line. GHOST has higher variance in the Eight schools, Radon subsets, and SRD data sets. All three methods perform similarly in the SLC and Radon data sets.

Some of the largest deviations from the 45 degree line are in row B, the Radon subsets data, where the number of groups ranges from 3 to 12. In general, we found that the accuracy of AXE was not particularly impacted as the proportion of test data increased, but it improved greatly as the number of groups increased. This is intuitive, as the estimation of Σ depends largely on group-specific intercepts, so when there are more groups, $E[\Sigma|Y_{-j}]$ is more stable across CV folds.

Table 4.3 provides total computation time on an Intel i7-8700k CPU with 40 Gb of memory. Except for the SRD data, models were fit using `STAN` or the `rstanarm` package, with 4 chains of 2000 samples and a 1000-sample burn-in, resulting in 4000 total samples. For the SRD data, the `CAR.ar()` function from the `CARBayesST` package was used, for four chains of 220,000 samples each, with a 20,000 sample burn-in and thinned to produce

4000 total posterior samples. In general, either AXE or GHOST takes the least amount of time. iIS-C is typically faster than MCV, but in some cases not by much, as with the Eight schools, Radon subsets, and SRD data.

Table 4.3. Total computing time for each method in seconds, excluding time to fit to the full data. Times with "h" are in hours. Times to fit the model to the full data are included for comparison. Table times with an asterisk (*) are approximates; to reduce computation time, a subset of 1000 MC samples was selected uniformly at random. Actual times can be obtained by dividing the table time by 4.

Data	Time						
	Full data	AXE	AXE+	GHOST	iIS-C	iIS-A	MCV
Eight schools	110	0.3	507	204	539	204	1113
Radon	41	4.0	278	0.6	541	56h*	1.1h
ESP	170	60	0.5h	0.6	720	10h*	6.1h
SLC	6	0.0	77	0.17	223	214*	0.9h
SRD	640	404	0.8h	102h*	102h*	102h*	37.0h

For AXE+, we randomly sampled and compared LRR_j's for six CV folds for Radon, ESP, SLC, and SRD data and 4 CV folds for the Eight schools data. The Eight schools data are included for completeness, as with only 8 groups, MCV would likely be used. AXE+ identified 20 out of 40 α values for which running MCV would be recommended, which included all 15 cases where AXE LRRs were over 0.15. For Radon, ESP, and SRD, posterior densities of variance hyperparameters were narrow, indicating they were well-estimated and AXE+ likewise found the mean LRR and standard deviations to be low (under 0.1). For SLC, one of the hyperparameters for Σ had a wide posterior density, indicating a large degree of uncertainty and a higher likelihood that its estimate could change across CV folds. AXE+ found the mean LRR was 0.18 with standard deviation 1.34. The high standard deviation suggests that MCV is likely preferable for this case.

GHOST has low computation times in all examples except for SRD. In the first four examples, the group-specific random effects θ_j are i.i.d., thus the ghosting estimate consists of simple draws from $N(0, \sigma^{2(s)})$. In both the SLC and SRD examples, the ghosting samples were drawn from the conditional multivariate normal density $\theta_j|\theta_{-j}, \Sigma$. An analytical solution to the conditional variance of $\theta_j|\theta_{-j}, \Sigma$ was available for the SLC data, because the LCO-CV scheme was equivalent to LOO-CV in this scenario, and was used to keep computation time low. In the SRD data, it was necessary to solve for $\Sigma^{(s)}$ by inverting the precision matrix first, which is computationally expensive. GHOST does make the strongest assumptions (Table 4.1) and evaluating the applicability of those

assumptions would require comparing to MCV and a subsequent increase in computation time, similar to AXE+.

To date iIS-A has been used only in the LOO-CV case with normal approximations (Vanhatalo et al., 2013; Vehtari et al., 2016b). In our examples, when CV folds contained more than one point of data, iIS-A took longer than MCV. In the Eight schools and SLC data, LCO-CV was equivalent to LOO-CV and we were able to avoid repeated matrix inversions using Sherman-Morrison, with significant computational gains.

4.7 Discussion

AXE is a fast and stable approximation method for obtaining cross-validated mean estimates. Our empirical results show that AXE consistently performed as well or better than more computationally expensive LCO methods. This is because AXE relies on weaker assumptions than competing methods and any large change in the estimation of $E[\Sigma, \phi|Y_{-j}]$ across CV folds also implies a large change in $f(\Sigma, \phi|Y_{-j})$ across CV folds. We show that AXE is more accurate when the number of CV folds is large, which is also when it saves the most time. When variance parameters are not well-estimated, we recommend running MCV for a small sample of folds and checking that the mean and standard deviation of LRRs is low. We hope that AXE will make cross-validation more accessible and ease the typical computational burden associated with BHRMs.

The LCO methods we compared AXE to were ghosting and integrated importance sampling. There is also a third category of LCO methods, which have been called integrated or marginal information criteria in the literature, the most notable method of which is the integrated Watanabe-Akaike Information Criterion (Li et al., 2016; Merkle et al., 2019). Integrated information criteria are used to approximate the expected log predictive density (ELPD), $\sum_{j=1}^{J} \log f(Y_j|Y_{-j})$. They do not produce estimates for $E[Y_j|Y_{-j}]$ and so are omitted from comparison.

AXE is a generic approximation approach for Bayesian hierarchical regression models that makes cross-validation more accessible. The examples we present here are publicly available and as such, the results we present do not have any negative social impact.

Chapter 5
R packages

5.1 Introduction

We propose the development of R packages to allow for easy use of the methods presented in Chapters 3 and 4. The `ssbf` package will include convenience functions for defining relationship groups, calculating borrowing factors, as well as creating a standard set of plots that can be interactively displayed in a Shiny app. The `axe` package will include a set of convenience functions for obtaining AXE estimates, if supplied with the model information and the cross-validation structure.

As both `axe` and `ssbf` require the calculation of the weight matrix W, maintenance of the two R packages is simplified if they are both built on top of one R package which calculates the weights, which will be called `hatmatrix`. In addition, the code to generate annotatable plots within a Shiny app is not specific to examining SSBF and may be useful in other applications and so will be separated into a fourth R package, `annoshiny`. This will also create cleaner code as each R package has a specific, single purpose, making it easier to maintain.

Table 5.1 lists the proposed functions for the R packages and their current stage of completeness. Most of the base code was written as a part of generating the examples for previous chapters and needs to be re-shaped and organized so that it is accessible for users. Most items that do not yet have base code are convenience functions that are extensions of the core functionality, such as calculating the weight matrix W when given an `rstanarm` object.

After this, documentation of the code and example vignettes are all that is needed before submitting the R packages to CRAN.

Section 5.2 describes the `ssbf` package and gives examples of its usage.

Table 5.1. Functions of the R packages and their stages of completeness. B: Base code written. C: Checks and warning or error messages written. T: Unit tests written.

Package	Functions	Completed
`hatmatrix`	• Calculation of W given X, Σ, Φ	B, C, T
	• Calculation of Φ given family and link function	
	• Calculation of W given `rstanarm` or `brms` object	
`axe`	• Cross-validation given X function	B, C
`annoshiny`	• Create `AnnoPlot` objects through manual specification	B, C
	• Convenience functions for `AnnoPlot` objects	B
	• Functions with Shiny modules for `AnnoPlot` objects	B
`ssbf`	• Calculate SSBF and pooling factor	B, C
	• Define relationship groups based on supplied data	B, C
	• Calculate borrowing factors and partial SSBF	B, C
	• Auto-generate a subset of SSBF `AnnoPlot` objects	
	• Shiny app configuration and launch	B, C

5.2 Example usage: The `ssbf` R package

There is a set of steps that are fairly similar across examples for decomposing regression model estimates. The steps can be summarized into four parts, which are

1. Calculating the matrix of borrowing factors (W), the shrinkage factor, and SSBF.
2. Defining relationship groups
3. Calculating borrowing factors and partial SSBF over the relationship groups.
4. Creating plots and comparing across them, optionally using a Shiny app.

Each of the steps above often requires more than a few lines of code. As the process itself is typically fairly similar across examples, the 'ssbf' package can in most cases reduce the amount of time spent coding. The core functionality of the interactive plots and Shiny app has been implemented. What is left is designing the architecture of the code so that it is easy to use and minimizes the amount of code written by the user.

In Section 5.2.1, we present examples of usage for creating the plots. In Section 5.2.2, we describe the Shiny app and outline its usage with examples.

5.2.1 Creating annotatable plots

Many of the plots used within the Shiny app allow you to select points to highlight or annotate within the plot itself. This is done by creating `AnnoPlot` objects and supplying them to the Shiny app. The following pre-defined functions will be used to create `annoplot` objects:

- `ap_scatter` produces a scatter plot. Selected points can be highlighted in a different color and/or labeled with text. Alternatively, all other points that have not been selected can be faded into the background. This is useful if, for example, the scatter plot already contains color and thus highlighting in color is not distinctive enough.
- `ap_density` produces a density plot. Selected points are displayed with vertical lines.
- `ap_tile` produces a heat map. Selected points can be overlaid in different colors and/or labeled with text.

Figure 5.1 contains examples of each, using the 'mtcars' data set. The first row shows the output of the `AnnoPlot` objects without any selected points and the second row shows the annotated plots with points selected. The functions will be created as part of the `annoshiny` package and usage details and examples will be available as a vignette.

Creating annotatable plots that are not in one of the above categories can be done by manually specifying the base plot, plot data, and annotation function and supplying both to the `annoplot` function. The base plot is the plot that is displayed when no points are selected. For example, we can re-create the effect of the `ap_scatter`, as shown below

```
library(ggplot2)
library(purrr)
library(ssbf)
# specify the base plot
p_base <- ggplot(mtcars, aes(x = mpg, y = cyl)) + geom_point()

# specify the annotation function
af_highlightred <- function(plots, pointdf) {
    map(plots, ~. + geom_point(data = pointdf, color = 'red') )
}
```

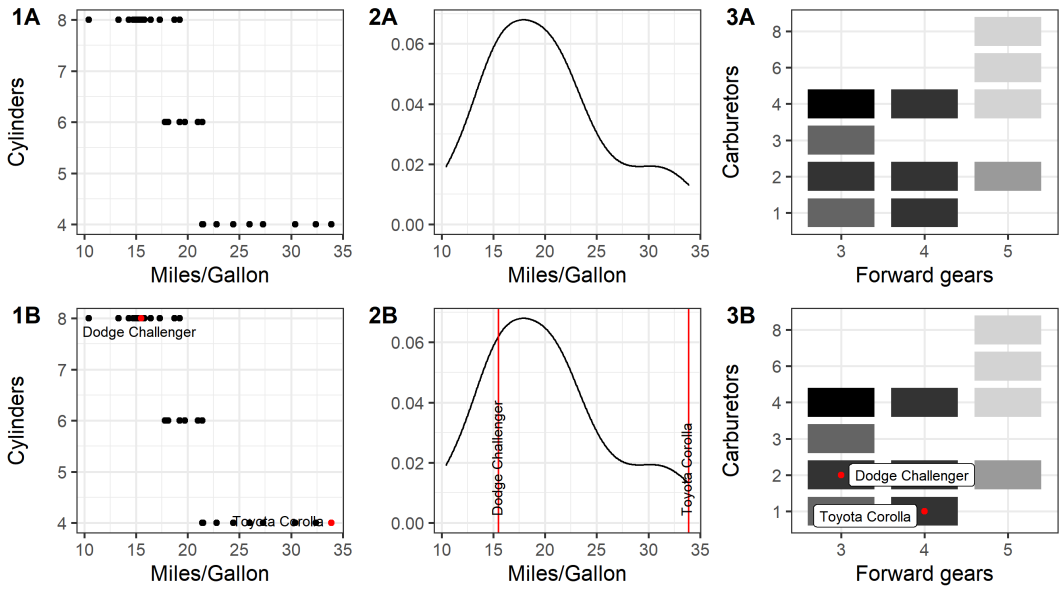

Figure 5.1. Example output from `ap_scatter` (1A and 1B), `ap_density` (2A and 2B), and `ap_tile` (3A and 3B), using the 'mtcars' data set for illustration. In the first row, plots display all data without any specific points highlighted. In the second row, the selected points are highlighted and annotated with user-specified labels.

```
# create annotatable plot
ap_mtcars <- annoplot(plots = list(p_base),
                      plotdata = mtcars,
                      annofun = af_highlightlabel)
```

Here, `p_base` is the base plot and `af_highlightred` is the annotation function. `p_base` must be either a `ggplot` object or a base R plot. The `annoplot` function returns an `AnnoPlot` object with a specific structure that is recognized by the Shiny app. The functions to create `AnnoPlot` objects and the Shiny modules that can read them will be available as a separate R package called "annoshiny". The code here is already developed as a part of the R package.

The annotation function must take two arguments as input. They are a list of the base plots and a dataframe of the selected points. The dataframe must contain all information necessary to create the highlights (e.g. if the plot is `mpg` against `cyl`, the dataframe must contain both `mpg` and `cyl` as columns. In addition to the base plot and annotation function, `annoplot` accepts many other arguments which can be used to fine-tune the plots' appearance within the Shiny app or specify how users can interact with them. These are detailed in the documentation and more examples will also be available as a vignette.

5.2.2 Launching and interacting with plots within the Shiny app

The advantage of the Shiny app is that it allows users to easily obtain granular information as well as examining high-level patterns of information borrowing. This is done in two main ways: one, within the Shiny app, users can hover over specific points of an `AnnoPlot` object to get more detailed information on it; and two, users can select specific points which will then be highlighted across all plots of the Shiny app. Understanding regression model estimates often requires multiple pieces of information—the relationship group size, borrower cluster size, the total borrowing factor, and the partial SSBF. An interactive Shiny app can allow users to explore the relationship across all of these dimensions from a standard set of simple plots.

The Shiny app consists of four main sections, three of which are dedicated to plots and one which displays information on the selected data. Figure 5.2 gives an example of the interface upon launching the Shiny app, with each of the different areas highlighted.

Brushing over or clicking on the points in any of the plots in panels 2, 3, or 4 will select points to highlight. These are displayed in panel 1, which gives the number of points selected in panel 1A. Panel 1B tabulates properties of the selected points. The properties displayed are specified by the user. In this case, the properties are the county of the point, whether it has a basement, the point's SSBF, and id number. The id number is used to synchronize the selection of points across all points and is specified upon creation of the `AnnoPlot` object. The `Go` button in 1A causes all plots to update based on the user's seleected points. The `Reset` button clears all selections and returns the app to the launch state.

The Shiny app is configured and launched using the `ssbf_shiny` function which creates an `SSBFApp` object. The function accepts just four arguments: the dataframe to display in panel 1B and three lists of `AnnoPlot` objects to display in panels 2, 3, and 4. Users can also choose to not display a plot in any of the panels by supplying `NULL` values instead of a list of `AnnoPlot` objects. The pseudocode below gives an example of configuring and launching the Shiny app.

```
ssbfapp_obj <- ssbf_shiny(
    radon, # data that displays in panel 1B
    top = list(annoplot1), # Panel 3
    center = list(annoplot2, annoplot3), # Panel 4
    side = list(annoplot4) # p anel 2
)
```

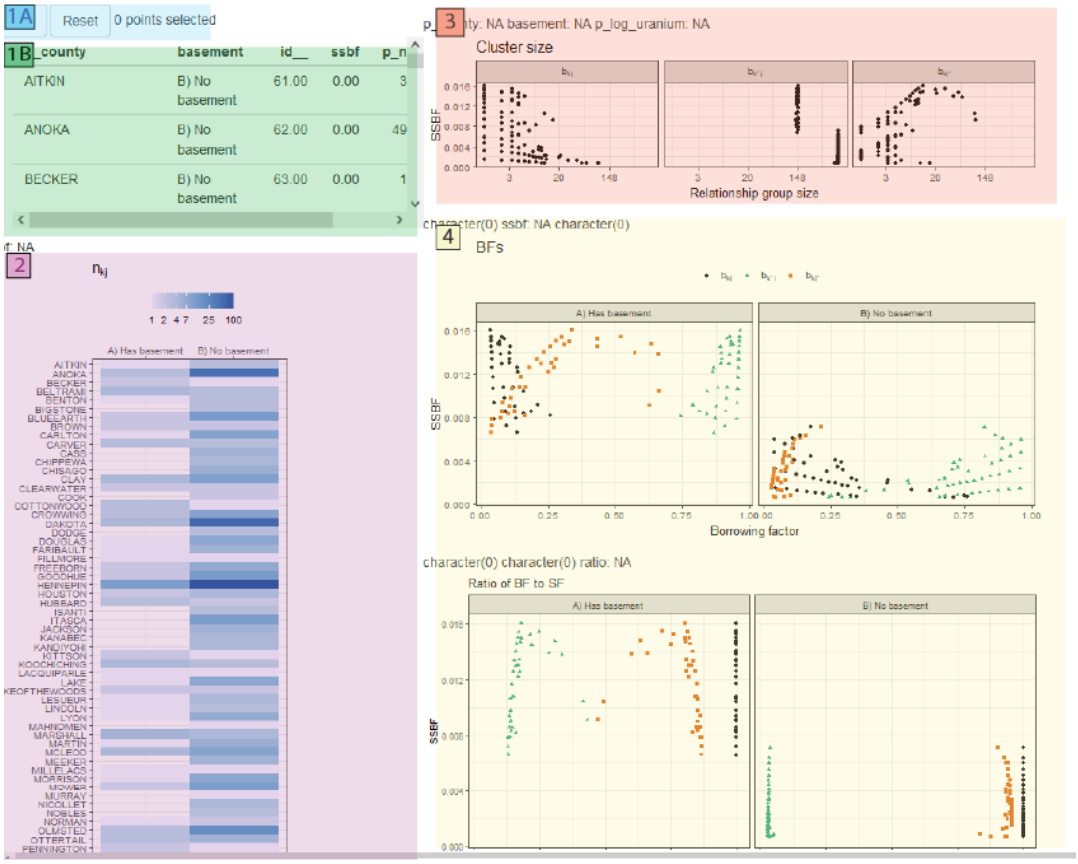

Figure 5.2. Areas of the Shiny app. 1A contains buttons to annotate or clear annotations from the plots. 1B tabulates the data properties, as selected by the user. Panels 2, 3, and 4 are different areas where plots can be placed, as specified by the user. Placement is largely dictated by the size and shape of the plots. The example uses the Radon data (Section 4.5.2.)

```
launch(ssbfapp_obj)
```

The core code to configure and launch the Shiny app is already developed as a part of the R package. We hope to refine the aesthetic of the Shiny app, potentially using Javascript.

Figure 5.3 gives an example of how the Shiny app may update after a user selects points to highlight. In the figure, three points were selected from panel A of the center SSBF plots. Those same three points were then highlighted across all of the plots in the Shiny app.

Above each of the plots is a string of text. This updates when users hover over specific points and contains information on the point. The text displayed is specified by the user. Figure 5.4 gives an example of how the text above the plot may change.

The interactive components of the Shiny app described above allow for intuitive

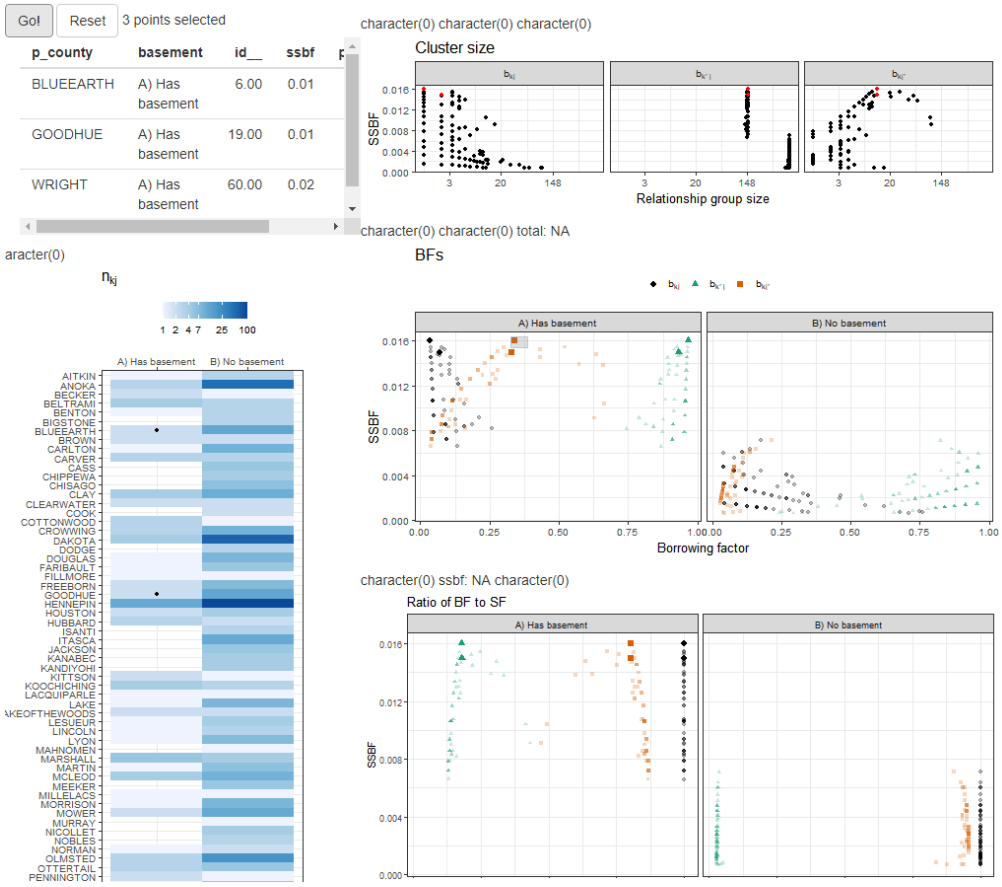

Figure 5.3. Shiny app with three points selected and highlighted across all plots, using the Radon data and model of Section 3.4.

exploration of the many quantities that may be of interest when decomposing regression model estimates. Fully-developed, the R package will automatically generate many of the plots that we have consistently found to be useful across examples, requiring less configuration of the `AnnoPlot` objects and of the Shiny app. The Shiny app itself can additionally be augmented by user-supplied plots, as previously described.

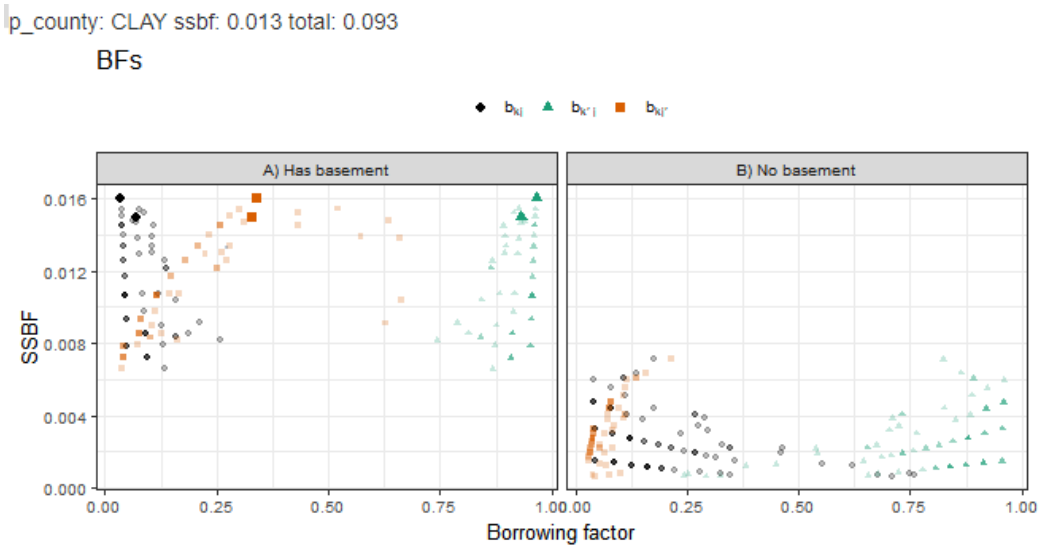

Figure 5.4. Shiny app when the user hovers over a point. The text above the plot displays informaiton on the point, as specified by the user (default is the x- and y-coordinates). This example uses the Radon data and model of Section 3.4

Chapter 6
Conclusion

We presented one application study of Bayesian hierarchical regression models and two novel methods.

In Chapter 2, we compared three different models of HIV prevalence within key populations, across four different countries. The models represented three different levels of information pooling. The data were heavily imbalanced across key populations and spatial and temporal dimensions, which presented a unique set of challenges as the models needed to efficiently capture information related to all three dimensions. Additional future workk could include evaluating the impacts of data imbalance on the model estimates as well as considering other model frameworks, such as a conditional auto-regressive spatio-temporal model.

In Chapter 3, we presented a novel method for decomposing regression model estimates into borrowing factors over a set of relationship groups. We demonstrated theoretical properties of the borrowing factors and show how they explicitly quantify information pooling and shrinkage. We also introduced a metric called SSBF, which can be used to identify those points with the most distinctive borrowing patterns, borrowing heavily from a small subset of lenders. SSBF is related to influence analysis metrics in the literature and can be thought of as the portion of a point's influence which is due just to the data availability. We demonstrated the application of the borrowing factors and SSBF in two examples. The first example quantified the effects of data imbalance on model estimates. The second example identified points which are the most impacted by influential points and could be used to provide context to influence analysis.

Potential applications for the borrowing factors go beyond the examples we presented. One is to help with the collection of new data, which can be expensive and labor-intensive. The diagonal entries of $XVX' = W\Phi$, which are the variances for $X\beta$ conditioned on Σ and Φ, can be interpreted as squared standard errors for the point estimates $X\hat{\beta}$. By

collecting new data which minimizes the standard errors, we can efficiently allocate new resources to improve model estimates. Another potential application is model comparison and evaluation, by comparing the borrowing factors to cross-validated error. We can determine if there are any distinctive information borrowing patterns that are associated with high error and, if so, identify modeling assumptions or data availability issues which contribute to the higher CV error.

In Chapter 4, we presented a novel method, called "AXE',' for obtaining cross-validated mean estimates. This was essentially another application of the borrowing factors, changing the data availability to obtain the cross-validated borrowing factors and thus the cross-validated mean estimates. We showed that AXE relies on the assumption that the posterior means of variance parameters are relatively stable across cross-validation folds, and provided theoretical results. The error bound for the AXE estimate is $\mathcal{O}\left(J^{-1}\right)$, where J is the number of CV folds. AXE had the weakest assumption of all LCO methods we evaluated and therefore produced more robust estimates with lower error than other methods across our six sets of examples.

The current work evaluated AXE to other methods using root-mean-square-error as the cross-validation metric. Future work could adapt AXE to methods which use the expected log posterior density as a metric. This would require accounting for the change in posterior densities across CV folds to approximate the value of the posterior predictive density over the test data. There is also room to improve upon the diagnostic for finite samples, as it requires running cross-validation for a subset of CV folds. The better-estimated the variance parameters are, the less likely they are to change across CV folds. One potential method is based on the width of the full-data posterior densities for the variance hyperparameters. In our set of examples, wider posterior densities indicated that the AXE estimates may be less reliable, but more extensive study would be needed to establish this as a potential diagnostic in general.

In Chapter 5, we proposed the development of R packages to support the adoption of the two novel methods in this thesis. The borrowing factors in particular can require extensive programming, thus the development of an R package and Shiny app with sensible defaults for the plots can save a large amount of time. The chapter outlined the work remaining to be done and provided examples of how the Shiny app can be used to interactively explore the borrowing factors.

Appendix A
Proofs and supplementary material for the borrowing factors and SSBF

This appendix contains all proofs and supplemental material related to Chapter 3. A.1 contains all proofs related to theoretical properties of SSBF and the borrowing factors, described in Chapter 3.3. A.2 contains all supplemental materials related to the examples in Chapter 3.4 and 3.5.

A.1 Proofs related to theoretical properties

A.1.1 contains the derivation for the borrowing factors under a one-way model. A.1.2 contains the proof for Theorem 3.3.1. A.1.3 contains the proof for Theorem 3.3.2. A.1.4 derives the relationship between SSBF and RVSI, described in Chapter 3.3.2. A.1.5 derives the relationship to S_i.

A.1.1 Borrowing factors for one-way models

Here we provide the calculations for the borrowing factors in the one-way setting, shown in (3.9). Given data $\boldsymbol{Y}_i \in I\!R^{n_i} \sim N(\alpha_i, \phi_i^2)$, $\alpha_i \sim N(\mu, \sigma^2)$, where $\mu \in I\!R$, $\alpha_i \in I\!R$, and $i = 1, \ldots, J$. In this scenario, it is possible to analytically solve for the borrowing factors in (3.3).

We begin by solving for V. Defining X_1 and X_2 as in (3.2), we can write V^{-1} as a

block matrix

$$V^{-1} = \begin{bmatrix} X_1'\Phi^{-1}X_1 & X_1'\Phi^{-1}X_2 \\ X_2'\Phi^{-1}X_1 & X_2'\Phi^{-1}X_2 + \Sigma^{-1} \end{bmatrix} \tag{A.1}$$

and obtain a solution for V using the rules for block matrix inversion. Starting in the upper-left quadrant and moving clockwise, let us refer to the corresponding blocks of V as A, B, C, D, such that

$$V = \begin{bmatrix} A & B \\ C & D \end{bmatrix},$$

and $A \in I\!R^{P_1 \times P_1}, B \in I\!R^{P_1 \times P_2}, C = B' \in I\!R^{P_2 \times P_1}, D \in I\!R^{P_2 \times P_2}$.

In this scenario, $X_1 = \mathbf{1}_N$, the vector of ones, and X_2 is the binary matrix of indicator variables where the i^{th} column indicates membership in the i^{th} cluster. Then the form of each block is as follows,

$$V^{-1} = \left[\begin{array}{c|ccc} \sum_{i=1}^{P} \frac{n_i}{\phi_i^2} & \frac{n_1}{\phi_1^2} & \frac{n_2}{\phi_2^2} & \cdots & \frac{n_J}{\phi_J^2} \\ \hline \frac{n_1}{\phi_1^2} & \frac{n_1}{\phi_1^2} + \sigma^{-2} & 0 & \cdots & 0 \\ \frac{n_2}{\phi_2^2} & 0 & \frac{n_2}{\phi_2^2} + \sigma^{-2} & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 & \\ \frac{n_J}{\phi_J^2} & 0 & \cdots & 0 & \frac{n_J}{\phi_J^2} + \sigma^{-2} \end{array}\right]$$

, where the vertical and horizontal lines enclose each of the four blocks in (A.1).

We can now solve for A, using the rules for block matrix inversion,

$$\begin{aligned} A &= \left(\sum_{j=1}^{J} \frac{n_j}{\phi_j^2} - \sum_{j=1}^{J} \frac{n_j^2/\phi_j^4}{n_j/\phi_j^2 + \sigma^{-2}} \right)^{-1} \\ &= \left(\sum_j \left(\frac{n_j}{\phi_j^2} \left(1 - \frac{n_j \phi_j^{-2}}{n_j \phi_j^{-2} + \sigma^{-2}} \right) \right) \right)^{-1} \\ &= \left(\sum_j \left(\frac{n_j}{\phi_j^2} \frac{\sigma^{-2}}{n_j \phi_j^{-2} + \sigma^{-2}} \right) \right)^{-1} \\ &= \left(\sum_j \left(\frac{n_j}{n_j \sigma^2 + \phi^2} \right) \right)^{-1} \\ &= \left(\sum_j \tau_j \right)^{-1}, \end{aligned}$$

where $\tau_j \coloneqq n_j/(n_j\sigma^2 + \phi_j^2)$ as in (3.9).

We derive the remaining block matrices of V in terms of τ_j and A.

$$\begin{aligned} B &= \left\{-\tau_j \sigma^2 A\right\}_{1\times J}, \\ D &= \operatorname{diag}\left(\frac{\phi_j^2\sigma^2}{n_j\sigma^2+\phi_j^2}\right) + \left\{\tau_j\sigma^2 A \tau_{j'}\sigma^2\right\}_{J\times J} \\ &= \operatorname{diag}\left(\frac{\phi_j^2\sigma^2}{n_j\sigma^2+\phi_j^2}\right) + \left\{B_j A^{-1} B_{j'}\sigma^2\right\}_{J\times J} \end{aligned}$$

With V known, with some algebra, we can derive the final result,

$$\begin{aligned} \hat{y}_i &= x_i' V X' \Phi^{-1} \boldsymbol{Y} \\ &= (A + B_i)\sum_j \frac{n_j}{\phi_j^2}\bar{Y}_j + \sum_j \frac{n_j}{\phi_j^2}\bar{Y}_j Bj\left(1 + B_i A^{-1}\right) + \tau_i\sigma^2\bar{Y}_i \\ &= A\frac{\phi_i^2}{n_i\sigma^2+\phi_i^2}\sum_j \frac{n_j}{\phi_j^2}\bar{Y}_j + \sum_j \frac{n_j}{\phi_j^2}\bar{Y}_j Bj\frac{\phi_i^2}{n_i\sigma^2+\phi_i^2} + \tau_i\sigma^2\bar{Y}_i \\ &= A\frac{\phi_i^2}{n_i\sigma^2+\phi_i^2}\sum_j \frac{n_j}{\phi_j^2}\bar{Y}_j\frac{\phi_j^2}{n_j\sigma^2+\phi_j^2} + \tau_i\sigma^2\bar{Y}_i \\ &= \frac{\phi_i^2}{n_i\sigma^2+\phi_i^2}\sum_j \frac{\tau_j}{\sum_j \tau_j}\bar{Y}_j + \tau_i\sigma^2\bar{Y}_i. \end{aligned}$$

A.1.2 Proof of Theorem 3.3.1

We re-state Theorem 3.3.1 below for reference:

Let response vector $\boldsymbol{Y} \in \mathbb{R}^N$ of a hierarchical linear regression follow a normal distribution as in (3.1), where the N-length vector of ones is in the column span of X_1, $\mathbf{1} \in \operatorname{span}(X_1)$. In the Bayesian setting, we assume $f(\Sigma)$ and $f(\phi)$ are some prior densities such that the posterior is proper. The $N \times N$ matrix of borrowing factors, W, is as defined as in (3.3). Then the sum of borrowing factors $\sum_{j=1}^N w_{ij}$ for a point estimate $\hat{Y}_i$ is 1 for all $i = 1, \ldots, N$, i.e. $W\mathbf{1} = \mathbf{1}$.

Proof. Defining X_1 and X_2 as in (3.2), we can write V^{-1} as a block matrix

$$V^{-1} = \begin{bmatrix} X_1'\Phi^{-1}X_1 & X_1'\Phi^{-1}X_2 \\ X_2'\Phi^{-1}X_1 & X_2'\Phi^{-1}X_2 + \Sigma^{-1} \end{bmatrix}$$

and obtain a solution for V using the rules for block matrix inversion. Starting in the upper-left quadrant and moving clockwise, let us refer to the corresponding blocks of V

as A, B, C, D, such that

$$V = \begin{bmatrix} A & B \\ C & D \end{bmatrix},$$

and $A \in I\!R^{P_1 \times P_1}, B \in I\!R^{P_1 \times P_2}, C \in I\!R^{P_2 \times P_1}, D \in I\!R^{P_2 \times P_2}$.

Let $M \coloneqq (X_2' \Phi^{-1} X_2 + \Sigma^{-1})^{-1}$; $H_2 \coloneqq X_2 M X_2' \Phi^{-1}$; and $\tilde{\Phi}^{-1} \coloneqq \Phi^{-1}(I - H_2)$. We solve for each of the blocks in V and write the solutions in terms of M, H_2, and $\tilde{\Phi}^{-1}$:

$$\begin{aligned} A &= (X_1' \Phi^{-1} X_1 - X_1' \Phi^{-1} H_2' \Phi^{-1} X_1)^{-1} = \left(X_1' \tilde{\Phi}^{-1} X_1\right)^{-1} \\ B &= -A X_1' \Phi^{-1} X_2 M \\ C &= B' \\ D &= M + M X_2' \Phi^{-1} X_1 A X_1' \Phi^{-1} X_2 M. \end{aligned} \tag{A.2}$$

Let $H \coloneqq X_1 (X_1' \tilde{\Phi}^{-1} X_1)^{-1} X_1' \tilde{\Phi}^{-1}$ and $H_1 \coloneqq X_1 A X_1' T^{-1}$. Note that $H = H_1(I - H_2)$. The weight matrix can be re-written in terms of H and H_2 using (A.2),

$$\begin{aligned} W = XVX'^{-1}\Phi^{-1} &= X_1 A X_1' \Phi^{-1} + X_1 B X_2' \Phi^{-1} + X_2 C X_1'^{-1} \Phi^{-1} + X_2 D X_2' \Phi^{-1} \\ &= H_1 - H_1 H_2 - H_2 H_1 + H_2 + H_2 H_1 H_2 \\ &= (I - H_2)(H_1 - H_1 H_2) + H_2 \\ &= (I - H_2) H + H_2 \\ &= H + H_2 (I - H). \end{aligned}$$

From Sherman-Morrison, $\tilde{\Phi} = (\Phi + X_2 \Sigma X_2')^{-1}$ is positive-definite. Then H is a projection matrix onto the column space of X_1, with inner product $\tilde{\Phi}^{-1}$, and as $\mathbf{1} \in \text{span}(X_1)$, $H\mathbf{1} = \mathbf{1}$ and $(I - H)\mathbf{1} = 0$. The result follows.

□

A.1.3 Proof for Theorem 3.3.2

We re-state Theorem 3.3.2 below for reference:

Under the same setting as in Theorem 3.3.1, let the shrinkage factor be defined as in (3.4). Then given a point estimate $\hat{Y}_i$, $0 < b_{iB_i} \leq 1$ and likewise $0 <= b_{iL_i} < 1$, where b_{iB_i} is the shrinkage factor and b_{iL_i} the pooling factor.

The proof here is based on our earlier work, Lemma 1 in the supplementary material for Zhang et al. (2020), and is re-created below for reference.

Proof. $b_{iB_i} > 0$: $b_{iB_i} = n_i w_{ii}$. V non-singular and T, Σ positive-definite imply V is positive-definite and XVX' is positive semi-definite. Then the diagonal entries of XVX' are non-negative and $w_{ii} = (XVX')_{ii} T_{ii}^{-1} > 0$.

$b_{iB_i} \leq 1$: Let V_{-i} as in (3.2), where the subscript $_{-i}$ indicates using the design matrix without the borrower cluster, X_{-B_i}, in place of X. We can solve for V as a function of V_{-i} using the Sherman-Morrison formula,

$$\begin{aligned} V &= (V_{-i} + x_i x_i' \phi_i^{-2})^{-1} \\ &= V_{-i} - \frac{n_i}{\phi_i^2} \frac{1}{1 + \frac{n_i}{\phi_i^2} x_i' V_{-i} x_i} V_{-i} x_i x_i' V_{-i}. \end{aligned} \tag{A.3}$$

As V_{-i} is positive-definite and $\frac{n_i}{\phi_i^2} x_i' V_{-i} x_i \geq 0$, (A.3) implies that $V_{-i} - V$ is positive semi-definite. Now, solving for V_{-i} as a function of V yields

$$\begin{aligned} V_{-i} &= (V - x_i x_i' \phi_i^{-1})^{-1} \\ &= V + \frac{n_i}{\phi_i^2} \frac{1}{1 - \frac{n_i}{\phi_i^2} x_i' V x_i} V x_i x_i' V, \end{aligned} \tag{A.4}$$

through another application of Sherman-Morrison. As $(1 - \frac{n_i}{\phi_i^2})^{-1} x_i' V x_i V x_i x_i' V$ is positive semi-definite and $b_{iB_i} = n_i \phi^{-2} x_i' V x_i > 0$, b_{iB_i} must be ≤ 1.

$0 <= b_{iL_i} < 1$: Theorem 3.3.1 and $0 < b_{iB_i} \leq 1$ implies $0 <= b_{iL_i} < 1$. □

A.1.4 Relationship between RVSI and SSBF

Value of information is an approach to outlier and influence analysis within the Bayesian literature that quantifies the value of sample information Y_j using the reduction in loss that results from including Y_j vs excluding it. For example, if $a_{Y_{-j}}$ is the estimator based on all data excluding Y_j and a_{Y_{-j},Y_j} is the estimator for Y_i based on all data, then the retrospective value of sample information (RVSI) under squared loss is

$$\text{RVSI}(Y_j | Y_{-j}; Y_i) = (a_{Y_{-j}} - a_{Y_{-j},Y_j})'(a_{Y_{-j}} - a_{Y_{-j},Y_j}). \tag{A.5}$$

This can be explicitly written in terms of partial SSBF. Let response vector $\boldsymbol{Y}$ follow a normal linear regression with model design matrix X as in (3.1) and let $\boldsymbol{Y}_j \in \mathbb{R}^{n_j} \sim N(x_j' \boldsymbol{\beta}, \phi_j^2)$.

Zhang et al. (2020) showed that, for Bayesian hierarchical regression models, $E[x_j'\boldsymbol{\beta}|Y_{-j}, \hat{\Sigma}, \hat{\phi}] = E[x_j'\boldsymbol{\beta}|Y_{-j}](1 + O(P_2^{-1}))$, for posterior means $\hat{\Sigma}$ and $\hat{\phi}$. Taking as our estimators $a_{Y_{-j}} = E[x_i'\boldsymbol{\beta}|Y_{-j}, \hat{\Sigma}, \hat{\phi}]$ and $a_{Y_{-j},Y_j} = E[x_i'\boldsymbol{\beta}|\boldsymbol{Y}, \hat{\Sigma}, \hat{\phi}]$ then approximates RVSI in (A.5) with $O(P_2^{-1})$ error.

Applications of the Sherman-Morrison formula and some algebra show that

$$E[\boldsymbol{\beta}|Y_{-j}, \hat{\Sigma}, \hat{\Phi}] = E[\boldsymbol{\beta}|\boldsymbol{Y}, \hat{\Sigma}, \hat{\Phi}] + \frac{n_j}{\phi_j^2} \frac{\hat{Y}_j - \bar{Y}_j}{1 - \frac{n_j}{\phi_j^2} x_j' V x_j} V x_j, \tag{A.6}$$

and the difference in our estimators can then be written as the product of the average residual for Y_j and their borrowing factor $n_j w_{ij}$,

$$a_{Y_{-j}} - a_{Y_{-j},Y_j} = \frac{w_{ij}}{b_{jL_j}} \frac{n_j(\hat{Y}_j - \bar{Y}_j)}{\phi^2}, \tag{A.7}$$

where b_{jL_j} denotes the pooling factor for $\hat{Y}_j$.

Combining (A.5) and (A.7), RVSI can be written as the product of the sum of squared residuals and PSSBF,

$$\text{RVSI}(Y_j|Y_{-j}; Y_i) = \frac{\text{PSSBF}_{ij}}{b_{jL_j}^2} \frac{n_j(\hat{Y}_j - \bar{Y}_j)^2}{\phi^4} (1 + O(P_2^{-1}).$$

A.1.5 Relationship between S_i and SSBF

Peña's S_i is the squared norm of the standardized vector $\boldsymbol{s}_i = (\hat{Y}_i - \hat{Y}_{i(1)}, \ldots, \hat{Y}_i - \hat{Y}_{i(N)})'$, where $\hat{Y}_{i(j)} = E[Y_i|Y_{-j}]$. S_i can be re-written as a linear combination of Cook's distances, D_j,

$$S_i = \frac{\boldsymbol{s_i}'\boldsymbol{s}_i}{pv\hat{a}r(\hat{Y}_i)} = \sum_{n=1}^{N} \frac{w_{in}^2}{w_{ii}w_{nn}} D_n, \quad D_n = \frac{e_n^2}{ps^2} \frac{w_{nn}}{(1 - w_{nn})^2}$$

where D_n is the Cook's distance for Y_n, $\boldsymbol{e} = \boldsymbol{Y} - X\hat{\boldsymbol{\beta}}$, $e_n = (Y_n - x_n'\hat{\boldsymbol{\beta}})$, and $s^2 = \boldsymbol{e}'\boldsymbol{e}/(n - P)$, where P is the dimension of $\boldsymbol{\beta}$.

If $\boldsymbol{Y}_j \in I\!R^{n_j} \sim N(x_j'\boldsymbol{\beta}, \phi_j^2)$, then $w_{ik} = w_{ik'}$ for all $i \in \{1, \ldots, N\}$ and all $k, k' \in j$, and we can aggregate over the clusters of data $\boldsymbol{Y}_j$ to obtain

$$S_i = \sum_j \frac{\text{PSSBF}_{ij}}{w_{ii}w_{jj}} \bar{D}_j, \quad \bar{D}_j = \frac{\bar{e}_j^2}{ps^2} \frac{w_{jj}}{(1 - w_{jj})^2}.$$

A.2 Supplemental calculations and figures for SSBF examples

A.2.1 Borrowing factors for the Radon example

For $b_{k'j}$ to be the borrowing factor for the contrast in data means $\bar{Y}_{k'j} - \bar{Y}_{k'j'}$, it is necessary to show that for all lenders g, g' corresponding to $\boldsymbol{Y}_{k'j'}$, $x_i' V x_g = x_i' V x_{g'}$.

For the model in (3.12), let N denote the dimension of $\boldsymbol{Y}$. Under a balanced data scenario, the number of houses in any county j with any basement status k is $n := N/(2J)$. As we are conditioning on the continuous covariate u_j, we note that

$$V^{-1} = \begin{bmatrix} \frac{N}{2\phi^2} & 0 & \frac{n}{\phi^2}\mathbf{1}_J' \\ 0 & \frac{N}{2\phi^2} & \frac{n}{\phi^2}\mathbf{1}_J' \\ \frac{n}{\phi^2}\mathbf{1}_J & \frac{n}{\phi^2}\mathbf{1}_J & \left(\frac{n}{\phi^2} + \frac{1}{\sigma^2}\right) I_J \end{bmatrix}, \tag{A.8}$$

where $\mathbf{1}_J \in \mathbb{R}^J$ is the vector of ones and $I_J \in \mathbb{R}^{J \times J}$ is the identity matrix.

For Y_g and $Y_{g'}$ within the same relationship group (e.g., same-county lenders, same-basement lenders, or others), the only difference between x_g and $x_{g'}$ is the indicator variable for the county-specific effect. Now let $P = J+2$, the number of columns in V and let $M_{gg'} \in \mathbb{R}^{P \times P}$ be the permutation matrix such that $M_{gg'} x_{g'} = x_g$. Then $\tilde{X} := X M_{gg'}$ is the model design matrix with the columns corresponding to indicator variables for counties g and g' switched. As the data are balanced, using $\tilde{X}$ instead of X still results in (A.8) and so

$$V = \left(\tilde{X}'\tilde{X}/\phi^2 + \begin{bmatrix} 0 & 0 \\ 0 & \frac{1}{\sigma^2} I \end{bmatrix} \right)^{-1} = \left(X'X/\phi^2 + \begin{bmatrix} 0 & 0 \\ 0 & \frac{1}{\sigma^2} I \end{bmatrix} \right)^{-1}. \tag{A.9}$$

This implies that

$$V x_{g'} = V M_{gg'} x_{g'} = V x_g. \tag{A.10}$$

Since $x_i' V x_g = x_i' V x_{g'}$ for all lenders g, g' in the same relationship group, we can formulate the point estimate $\hat{u}_{kj}$ as a weighted sum of relationship group means,

$$\hat{\mu}_{kj} = b_{kj}\bar{Y}_{kj} + b_{kj'}\bar{Y}_{kj'} + b_{k'j}(\bar{Y}_{k'j} - \bar{Y}_{k'j'}, \tag{A.11}$$

where b_{kj} is the shrinkage factor and $b_{kj'}$ is the pooling factor. When J is large, this contrast in means, given $\boldsymbol{\beta}$ and a, has expected value of a_j. Then $b_{k'j} = -b_{k'j'}$ isolates

the county-specific effect a_j and represents borrowing from lenders due to a_j. Similarly, $b_{kj'}$ represents borrowing due to the basement intercept.

A.2.2 Supplemental figures for Scottish respiratory disease example

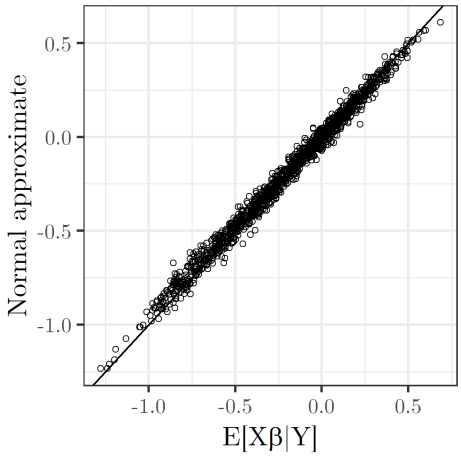

Figure A.1. Scatterplot of point estimates obtained through normal approximation (y-axis) versus actual posterior means $E[X\beta|Y]$ (x-axis) for the Scottish respiratory disease data. The normal approximation used is (3.13).

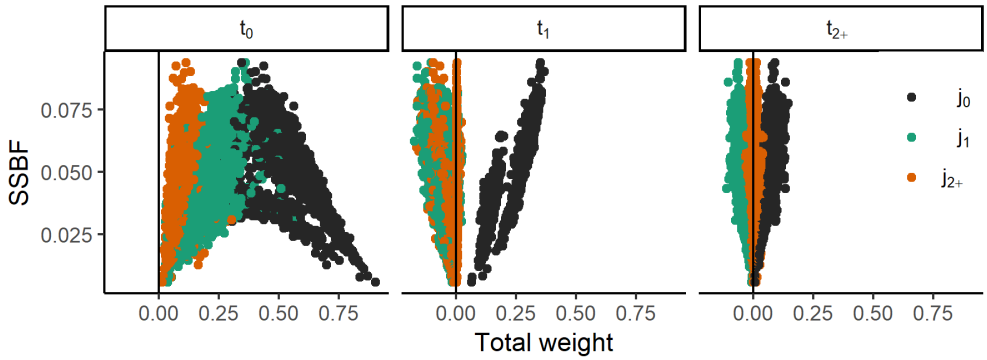

Figure A.2. A scatter plot of SSBF against the total weight applied to lender relationship groups, where each panel represents a different temporal relationship group (t_0, t_1, t_{2+} for same year, adjacent year, other years, respectively) and colors represent different spatial relationship groups (black for j_0, green for j_1, orange for j_{2+}, corresponding to same IG, neighboring IG, and farther IGs, respectively).

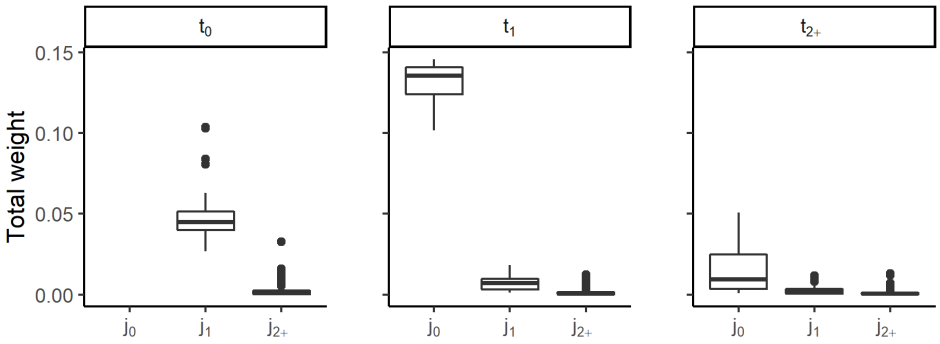

Figure A.3. Boxplots of total (absolute) weight placed on 11 influential points when $\hat{\alpha} = 0.57$ and $\hat{\rho} = 0.76$. Box plots are split into temporal (t_0, t_1, t_{2+}) and spatial (j_0, j_1, j_{2+}) relationship groups. The plots do not include the shrinkage factor, hence no boxplot for $b_{t_0 j_0}$.

Appendix B
Proof and calculations for AXE

B.1 Proof for Theorem 4.3.1

Lemma B.1.1. *Let response vector $Y \in \mathbb{R}^N$ of a hierarchical linear regression follow a normal distribution as in (4.1) and let s_j be the set of indices of θ_j in θ such that X_{s_j} is made up of identical rows. As shorthand X_{s_j} is referred to as X_j, and x_j' is a row in X_j. V is defined as in (4.2). We list several facts pertaining to V and V_{-j}.*

1. $V_{-j} = V + \frac{n_j}{\tau^2}\frac{1}{1-\frac{n_j}{\tau^2}x_j'Vx_j}Vx_jx_j'V$
2. $V \preceq V_{-j}$, *where* $V \preceq V_{-j}$ *indicates that* $V_{-j} - V$ *is positive semi-definite*
3. $\frac{n_j}{\tau^2}x_j'Vx_j \leq 1$
4. $x_j'V_{-j}x_j = x_j'Vx_j\frac{1}{1-\frac{n_j}{\tau^2}x_j'Vx_j} \geq x_j'Vx_j$
5. $\det(V_{-j}) = \det(V)\frac{1}{1-\frac{n_j}{\tau^2}x_j'Vx_j} \geq \det(V)$
6. $\frac{1}{\tau^4}Y_{-j}'X_{-j}V_{-j}X_{-j}'Y_{-j} = \frac{1}{\tau^4}Y'XVX'Y - \frac{n_j}{\tau^2}\bar{y}_j^2 + \frac{\bar{e}_j^2}{\tau^2/n_j}\frac{1}{1-\frac{n_j}{\tau^2}x_j'Vx_j}$

Proof. 1. This follows from using the Sherman-Morrison formula on $V_{-j} = (V^{-1} - \frac{n_j}{\tau^2}x_jx_j')^{-1}$.

2. Using a similar process as in 1., if we re-write V in terms of V_{-j}, we obtain $V = V_{-j} - \frac{n_j}{\tau^2}\frac{1}{1+\frac{n_j}{\tau^2}x_j'V_{-j}x_j}V_{-j}x_jx_j'V_{-j}$. The difference, $\frac{n_j}{\tau^2}\frac{1}{1+\frac{n_j}{\tau^2}x_j'V_{-j}x_j}V_{-j}x_jx_j'V_{-j}$, is positive semi-definite and so $V_{-j} \succeq V$, where $A \succeq B$ is defined as $A - B$ is positive semi-definite.

3. From 2., we note that $V_{-j} \succeq V$ and from 1., $V_{-j} - V$ is $\frac{n_j}{\tau^2}\frac{1}{1-\frac{n_j}{\tau^2}x_j'Vx_j}Vx_jx_j'V$. V positive-definite implies $\frac{n_j}{\tau^2}x_j'Vx_j > 0$, then for $V_{-j} - V$ to be positive semi-definite, $\frac{n_j}{\tau^2}x_j'Vx_j \leq 1$ must be true.

4. Using 1., $x_j'V_{-j}x_j = x_j'Vx_j + \frac{n_j}{\tau^2}\frac{1}{1-\frac{n_j}{\tau^2}x_j'Vx_j}(x_j'Vx_j)^2 = x_j'Vx_j(1-\frac{n_j}{\tau^2}x_j'Vx_j)^{-1}$. The inequality follows from 3. , as $1 - \frac{n_j}{\tau^2}x_j'Vx_j \leq 1$.

5. As the held-out data correspond to identical rows of X, we have a closed-form solution for the determinant of V_{-j} in terms of V:

$$\begin{aligned}
\det(V^{-1} - \tau^{-2}X_j'X_j) &= \det(V^{-1})\det(1 - \frac{n_j}{\tau^2}x_j'Vx_j) \qquad \textit{Sylvester's det. theorem} \\
&= \det(V^{-1})(1 - \frac{n_j}{\tau^2}x_j'Vx_j) \\
\implies \det(V_{-j}) &= \det(V)\frac{1}{1-\frac{n_j}{\tau^2}x_j'Vx_j} \\
&\geq \det(V) \qquad \textit{from } 3.
\end{aligned}$$

6. We first show that $\frac{1}{\tau^2}V_{-j}X_{-j}'Y_{-j} = \frac{1}{\tau^2}VX'Y + \frac{n_j}{\tau^2}\frac{\bar{e}_j}{1-\frac{n_j}{\tau^2}\nu_j}Vx_j$. Let $\nu_j \coloneqq x_j'Vx_j$:

$$\begin{aligned}
\frac{1}{\tau^2}V_{-j}X_{-j}'Y_{-j} &= \frac{1}{\tau^2}[V + \frac{n_j}{\tau^2}\frac{1}{1-\frac{n_j}{\tau^2}\nu_j}Vx_jx_j'V](X'Y - X_j'Y_j) \qquad \textit{from } 1. \\
&= \frac{1}{\tau^2}VX'Y + Vx_j\left[-\frac{n_j}{\tau^2}\bar{y}_j + \frac{n_j}{\tau^2}\frac{1}{1-\frac{n_j}{\tau^2}\nu_j}\left(\tilde{y}_j - \frac{n_j}{\tau^2}\nu_j\bar{y}_j\right)\right] \\
&= \frac{1}{\tau^2}VX'Y - Vx_j\left[\frac{n_j}{\tau^2}\frac{1}{1-\frac{n_j}{\tau^2}\nu_j}\bar{y}_j - \frac{n_j}{\tau^2}\frac{1}{1-\frac{n_j}{\tau^2}\nu_j}\tilde{y}_j\right] \\
&= \frac{1}{\tau^2}VX'Y + \frac{n_j}{\tau^2}\frac{\bar{e}_j}{1-\frac{n_j}{\tau^2}x_j'Vx_j}Vx_j.
\end{aligned}$$

Following a similar procedure, we can re-write $\frac{1}{\tau^4}Y_{-j}'X_{-j}V_{-j}X_{-j}'Y_{-j}$ in terms of its full-data counterpart, $\frac{1}{\tau^4}Y'XVX'Y$, along with an additional difference term. Let $(\tilde{Y})_{-j}$ refer to $X_{-j}\tilde{\beta}$, the conditional expected value based on the full data Y.

$$\begin{aligned}
\frac{1}{\tau^4}Y_{-j}'X_{-j}V_{-j}X_{-j}'Y_{-j} &= \frac{1}{\tau^4}Y_{-j}'X_{-j}VX'Y + \frac{n_j}{\tau^4}\frac{\bar{e}_j}{1-\frac{n_j}{\tau^2}\nu_j}Y_{-j}'X_{-j}Vx_j \\
&= \frac{1}{\tau^2}Y_{-j}'(\tilde{Y})_{-j} + \frac{n_j}{\tau^2}\frac{\bar{e}_j}{1-\frac{n_j}{\tau^2}\nu_j}\left(\frac{1}{\tau^2}Y'XVx_j - \frac{1}{\tau^2}Y_j'X_jVx_j\right) \\
&= \frac{1}{\tau^2}Y_{-j}'(\tilde{Y})_{-j} + \frac{n_j}{\tau^2}\frac{\bar{e}_j}{1-\frac{n_j}{\tau^2}\nu_j}\left(\tilde{y}_j - \frac{n_j}{\tau^2}\nu_j\bar{y}_j\right)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\tau^2} Y'_{-j}(\tilde{Y})_{-j} + \bar{e}_j \frac{n_j}{\tau^2} \left[\frac{1}{1 - \frac{n_j}{\tau^2}\nu_j} \left(\tilde{y}_j - \frac{n_j}{\tau^2}\nu_j \bar{y}_j \right) + \bar{y}_j - \bar{y}_j \right] \\
&= \frac{1}{\tau^2} Y'_{-j}(\tilde{Y})_{-j} + \bar{e}_j \frac{n_j}{\tau^2} \left[\bar{y}_j + \frac{\bar{e}_j}{1 - \frac{n_j}{\tau^2}\nu_j} \right] \\
&= \frac{1}{\tau^2} Y'_{-j}(\tilde{Y})_{-j} + \bar{e}_j \frac{n_j}{\tau^2} \left[\bar{y}_j + \frac{\bar{e}_j}{1 - \frac{n_j}{\tau^2}\nu_j} \right] - \frac{n_j}{\tau^2}\bar{y}_j\tilde{y}_j + \frac{n_j}{\tau^2}\bar{y}_j\tilde{y}_j \\
&= \frac{1}{\tau^4} Y'XVX'Y - \frac{n_j}{\tau^2}\bar{y}_j^2 + \frac{\bar{e}_j^2}{\tau^2/n_j} \frac{1}{1 - \frac{n_j}{\tau^2} x'_j V x_j}.
\end{aligned}
$$

□

Lemma B.1.2. *Let response vector $Y \in \mathbb{R}^N$ of a hierarchical linear regression follow a normal distribution as in (4.1) and let s_j be the set of indices of θ_j in θ such that X_{s_j} is made up of identical rows. V is defined as in (4.2). The difference Δ_j between the log densities $\ell(\Sigma, \tau | Y_{-j})$ and $\ell(\Sigma, \tau | Y)$ is*

$$
\Delta_j := C + n_j \log \tau - \frac{1}{2} \log \left(1 - \frac{n_j}{\tau^2} x'_j V x_j \right) - \frac{n_j}{\tau^2}\bar{y}_j^2 + \frac{\bar{e}_j^2}{\tau^2/n_j} \frac{1}{1 - \frac{n_j}{\tau^2} x'_j V x_j},
$$

where $C \in \mathbb{R}$ is a constant that does not involve Σ or τ.

Proof. The log-likelihood is as follows:

$$
\ell(\Sigma, \tau | Y) \propto \ell(\tau) + \ell(\Sigma) - N \log(\tau) - \frac{1}{2} \log \det(\Sigma) + \frac{1}{2} \log \det(V) + \frac{1}{2\tau^4} Y'XVX'Y.
$$

Using Theorem C.1.1, items 5 and 6, we can directly obtain the difference between the log-likelihoods $\ell(\Sigma, \tau | Y_{-j}) - \ell(\Sigma, \tau | Y)$ as stated. □

Lemma B.1.3. *Let response vector $Y \in \mathbb{R}^N$ of a hierarchical linear regression follow a normal distribution as in (4.1) and let V defined as in (4.2). Let $B := [\ I_{P_2}] \in \mathbb{R}^{P_2 \times P}$, where I_{P_2} is the P_2-dimensional identity matrix. The partial derivative $\frac{\partial}{\partial \Sigma} a'Vc$ is $\Sigma^{-1}BVac'VB'\Sigma^{-1}$ for vectors $a \in \mathbb{R}^P$, $c \in \mathbb{R}^P$.*

Proof.

$$
\begin{aligned}
\frac{\partial}{\partial \Sigma_{ij}} V &= -V \left(\frac{\partial}{\partial \Sigma_{ij}} V^{-1} \right) V \\
&= VB'\Sigma^{-1}\delta_i\delta'_j\Sigma^{-1}BV,
\end{aligned}
$$

where $\delta_i \in I\!R^{P_2}$ is the binary vector with a 1 only at the i^{th} index. Then,

$$\begin{aligned}\frac{\partial}{\partial \Sigma_{ij}} a'Vc &= \frac{\partial}{\partial \Sigma_{ij}} \operatorname{tr}(a'Vb) \\ &= \operatorname{tr}\left(ba' \frac{\partial}{\partial \Sigma_{ij}} V\right) \\ &= \operatorname{tr}\Big(ba'VB'\Sigma^{-1}\delta_i\delta_j'\Sigma^{-1}BV\Big) \\ &= \operatorname{tr}\Big(\delta_j'\Sigma^{-1}BVca'VB'\Sigma^{-1}\delta_i\Big) \\ &= \delta_j'\Sigma^{-1}BVca'VB'\Sigma^{-1}\delta_i,\end{aligned}$$

as the trace is invariant under cyclic permutations.

Then

$$\frac{\partial}{\partial \Sigma} a'Vc = \Sigma^{-1}BVac'VB'\Sigma^{-1}.$$

□

We re-state Theorem 4.3.1 below:

Let response vector $Y \in I\!R^N$ of a hierarchical linear regression follow a normal distribution as in (4.1), where $\mathbb{1} \in \operatorname{span}(X)$, and prior densities $f(\Sigma)$ and $f(\tau)$ are $\mathcal{O}(1)$. $X\beta = \mu + \theta$ as in (4.5), where θ has J unique values and $\theta_1, \ldots, \theta_J$, and s_j denotes the set of indices such that $\theta_{s_j} = \theta_j \mathbb{1}$, and X_j is made up of identical rows. V is defined as in (4.2). Then $|\operatorname{argmax}_{\Sigma,\tau} f(\Sigma, \tau | Y_{-\theta_j}) - \operatorname{argmax}_{\Sigma,\tau} f(\Sigma, \tau | Y)| \to 0$ as $J \to \infty$.

Proof. The log-likelihood is as follows:

$$\ell(\Sigma, \tau | Y) \propto \ell(\tau) + \ell(\Sigma) - N\log(\tau) - \frac{1}{2}\log\det(\Sigma) + \frac{1}{2}\log\det(V) + \frac{1}{2\tau^4} Y'XVX'Y.$$

We note that, other than the prior densities whose forms are unknown, the log-likelihood consists of terms with P_2 or N summands. As the number of CV folds is J and $J \leq P_2$, $J \leq N$, $J^{-1}\ell(\Sigma, \tau | Y)$ is finite and we can say that for any J, the solution to $J^{-1}\frac{\partial}{\partial \Sigma}\ell(\Sigma, \tau | Y) = 0$ is the same as the solution to $\frac{\partial}{\partial \Sigma}\ell(\Sigma, \tau | Y) = 0$:

$$\operatorname*{argmax}_{\Sigma,\tau} f(\Sigma, \tau | Y) = \operatorname*{argmax}_{\Sigma,\tau} \frac{\ell(\Sigma, \tau | Y)}{J}.$$

The dimensions of Σ, V, and Y are dependent on J, so for a particular J, we denote the corresponding Σ as $\Sigma^{(J)}$, and similarly for $V^{(J)}$ and $Y^{(J)}$. Let a_J be the sequence of

differences between $\frac{\partial}{\partial\Sigma}\frac{\ell(\Sigma^{(J)},\tau|Y^{(J)})}{J}$ and $\frac{\partial}{\partial\Sigma}\frac{\ell(\Sigma^{(J)},\tau|Y^{(J)}_{-j})}{J}$:

$$a_J = \left|\frac{\partial}{\partial\Sigma}\frac{\ell(\tau,\Sigma^{(J)}|Y_{-j})^{(J)}}{J} - \frac{\partial}{\partial\Sigma}\frac{\ell(\tau,\Sigma^{(J)}|Y^{(J)})}{J}\right|.$$

If we show $a_J \to 0$ as $J \to \infty$, then that is equivalent to showing $|\operatorname{argmax}_\Sigma f(\Sigma,\tau|Y_{-s_j}) - \operatorname{argmax}_\Sigma f(\Sigma,\tau|Y)| \to 0$ as $J \to \infty$.

Let Δ_j be the difference in log-likelihoods stated in Theorem C.1.2:

$$\begin{aligned} a_J &= \left|\frac{\partial}{\partial\Sigma}\frac{\ell(\tau,\Sigma^{(J)}|Y^{(J)}_{-j})}{J} - \frac{\partial}{\partial\Sigma}\frac{\ell(\tau,\Sigma^{(J)}|Y^{(J)})}{J}\right| \\ &= \left|\frac{\partial}{\partial\Sigma}\left(\frac{\ell(\tau,\Sigma|Y^{(J)})}{J} - \frac{\Delta_j}{J}\right) - \frac{\partial}{\partial\Sigma}\frac{\ell(\tau,\Sigma^{(J)}|Y^{(J)})}{J}\right| \\ &= \left|\frac{1}{J}\frac{\partial}{\partial\Sigma}\Delta_j\right|. \end{aligned}$$

It remains to show that $\left|\frac{\partial}{\partial\Sigma}\Delta_j\right| \in o(J)$. Of the terms in Δ_j, only those involving $V^{(J)}$ are dependent on J. The derivative is as follows:

$$\frac{\partial}{\partial\Sigma}\Delta_J = \frac{n_j/\tau^2}{1-\frac{n_j}{\tau^2}x_j'V^{(J)}x_j}\left(\frac{1}{2}\left(1+\frac{\bar{e}_j^2}{1-\frac{n_j}{\tau^2}x_j'V^{(J)}x_j}\right)\frac{\partial}{\partial\Sigma}x_j'V^{(J)}x_j + \bar{e}_j\frac{\partial}{\partial\Sigma}x_j'V^{(J)}X'Y\right).$$

Let a, c in Lemma 3 equal x_j, then $\frac{\partial}{\partial\Sigma}x_j'Vx_j = \Sigma^{-1(J)}BV^{(J)}x_jx_j'V^{(J)}B'\Sigma^{-1(J)}$. Let $a = x_j$ and $c = X'Y$, then $\frac{\partial}{\partial\Sigma}x_j'VX'Yj = \Sigma^{-1(J)}B\tilde{\beta}^{(J)}x_j'V^{(J)}B'\Sigma^{-1(J)}$.

From Theorem C.1.1, items 2 and 4, we know that $V^{(J)}$ and $\frac{1}{1-\frac{n_j}{\tau^2}x_j'V^{(J)}x_j}$ decrease as J increases. We also note that $\sqrt{n_j}\bar{e}_j = \mathcal{O}_p(\tau)\,\forall J$. As the non-zero values of x_j are not dependent on J, this similarly implies that the magnitude of $\tilde{\beta}$ is not dependent on J.

Then the partial derivatives decrease as J increases. The increase in dimension does not matter, as the solution also has J dimensions. The remaining terms are $\mathcal{O}_p(1)$, as we take τ fixed. Then $\left|\frac{\partial}{\partial\Sigma}\Delta_J\right| \in \mathcal{O}_p(1)$.

A similar argument as above for τ holds if $\left|\frac{\partial}{\partial\tau}V^{(J)}\right| \in \mathcal{O}_p(1)$. Let $M \in I\!R^{P\times P}$ be a binary matrix where $M_{pq} = 1$ if $\tau^{-2}(X'X)_{pq} \neq 0$. Then:

$$\frac{\partial}{\partial\tau}x_j'Vx_j = -\frac{2}{\tau^3}x_j'VMVx_j.$$

It can be seen that this scalar quantity decreases with J, using Theorem C.1.1, item 1.

□

We re-state Corollary 4.3.2 below:

Let $\hat{\Sigma}$ denote the full-data posterior mean, $E[\Sigma|Y]$, and $\tilde{\Sigma}$ the CV posterior mean over the training data $E[\Sigma|Y_{-j}]$ for CV fold j. Likewise let $\hat{\tau} = E[\tau|Y]$ and $\tilde{\tau} \coloneqq E[\tau|Y_{-j}]$. Under the same conditions as 4.3.1, $E[X\beta|Y_{-j}, \hat{\Sigma}, \hat{\tau}] = E[X\beta|Y_{-j}, \tilde{\Sigma}, \tilde{\tau}](1 + \mathcal{O}(J^{-1}))$.

Proof. Using the same methods as determining $\frac{\partial}{\partial\Sigma}\Delta_J$, it can be shown that $\frac{\partial}{\partial\Sigma}\Delta_J$ is differentiable and, as $XVX \in \mathcal{O}(1)$, the derivative is finite, thus $\frac{\partial}{\partial\Sigma}\Delta_J$ is uniformly continuous. Proof by contradiction shows that $f(\Sigma) = \frac{\partial}{\partial\Sigma}\Delta_J$ is a bijective function. Then as $\frac{1}{J}\frac{\partial}{\partial\Sigma}\Delta_J \to 0$, $E[\Sigma, \tau|Y_{-j}] \to E[\Sigma, \tau|Y]$ and $E[\Sigma, \tau|Y_{-j}] = E[\Sigma, \tau|Y](1 + \mathcal{O}(J^{-1}))$.

Substituting $\hat{\Sigma} + \mathcal{O}(J^{-1})$ and $\hat{\tau} + \mathcal{O}(J^{-1})$ into $\tau^{-2}XVX'Y$ and an application of Sherman-Morrison yields $\tau^{-2}XVX'Y + \mathcal{O}(J^{-1})\,XVX'XVX'Y(\tau^{-2} + \mathcal{O}(J^{-1})) + \mathcal{O}(J^{-1})\,XVB\Sigma^{-1}B'VX'Y$, where $B \in \mathbb{R}^{P\times P_2}$ is a block matrix of 0s in the first P_1 rows and the identity matrix in the remaining P_2. As $XVX', XVX'Y, \Sigma \in \mathcal{O}(1)$, the result follows. □

B.2 Computational complexity calculations

This section contains derivations for computational complexity in Table 4.1. We assume without loss of generality that $\mathcal{O}(P_2) = \mathcal{O}(P)$. For all methods, we assume that the cost of drawing a sample from a specific density is dominated by the calculation of the density's parameters, e.g. if drawing from a multivariate normal density, we assume the calculation of the mean vector and covariance matrix dominates the computational cost.

B.2.1 AXE

We re-state the AXE estimate below:

$$\hat{Y}_j^{AXE} = \frac{1}{\hat{\tau}^2}X_j\left(\frac{1}{\hat{\tau}^2}X'_{-j}X_{-j} + \begin{bmatrix}0 & 0\\ 0 & \Sigma^{-1}\end{bmatrix}\right)^{-1}X'_{-j}Y_{-j}. \tag{B.1}$$

The cost of the matrix inversion in (C.1) is $\mathcal{O}(P^3)$, while the matrix multiplication is $\mathcal{O}(NP^2)$. Conducted over J total cross-validation loops, the computational complexity for AXE is $\mathcal{O}(J(NP^2 + P^3))$.

B.2.2 GHOST

Without loss of generality, let $f(\theta|\Sigma)$ be $N(0, \Sigma)$. Then:

$$\theta_j|\theta_{-j}, \Sigma \sim N(\Sigma_{j-j}\Sigma_{-j-j}^{-1}\theta_{-j}, \quad \Sigma_{jj} - \Sigma_{j-j}\Sigma_{-j-j}^{-1}\Sigma_{-jj}). \tag{B.2}$$

As we assume the dimension of θ_j is fixed, the cost of the matrix inversion in (C.2) is $\mathcal{O}(P^3)$ and the cost of the matrix multiplication is likewise $\mathcal{O}(P^3)$. Conducted over J total cross-validation loops for S samples, the computational complexity for GHOST is $\mathcal{O}(SJP^3)$.

B.2.3 iIS-C

We re-state the iIS-C importance weights $w_j^{(s)}$ and mean estimate for $E[Y_j|Y_{-j}]$:

$$w_j^{(s)} = \frac{1}{f(Y_j|\theta_{-j}^{(s)}, \tau^{(s)}, \mu^{(s)}, \Sigma^{(s)})}, \quad \hat{Y}_j^{\text{iIS-C}} = \frac{\sum_{s=1}^S w_j^{(s)} E\left[Y_j|\theta_{-j}, \tau^{(s)}, \Sigma^{(s)}, \mu^{(s)}\right]}{\sum_{s=1}^S w_j^{(s)}}. \tag{B.3}$$

Let $a_j \coloneqq E[\theta_j|\theta_{-j}, \Sigma]$ and $M \coloneqq \text{Cov}(\theta_j|\theta_{-j}, \Sigma)$, corresponding to the mean and covariance, respectively, of (C.2). Then,

$$Y_j|\theta_{-j}, \tau, \mu, \Sigma \sim N(\mu + a_j, \tau^2 I + M).$$

Note that the cost of obtaining a_j and M for all CV folds and MC samples is the same as ghosting at $\mathcal{O}(SJP^3)$. The cost of obtaining the likelihood in weight $w_j^{(s)}$ for (C.3) is an additional $\mathcal{O}\left(S\sum_{j=1}^J n_j\right) = \mathcal{O}(SN)$ and the cost of the expectation is likewise $\mathcal{O}(SN)$. The total cost is then $\mathcal{O}(SJP^3)$, the same order as ghosting.

B.2.4 iIS-A

We re-state the iIS-A importance weights $w_j^{(s)}$ and mean estimate for $E[Y_j|Y_{-j}]$:

$$\text{iIS-A: } w_j^{(s)} = \frac{1}{f(Y_j|\tau^{(s)}, \Sigma^{(s)}, Y_{-j})}, \quad \hat{Y}_j^{iIS-A} = \frac{\sum_{s=1}^S w_j^{(s)} E\left[Y_j|\tau^{(s)}, \Sigma^{(s)}, Y_{-j}\right]}{\sum_{s=1}^S w_j^{(s)}}. \tag{B.4}$$

$$Y_j|\tau^{(s)}, \Sigma(s), Y_{-j} \tag{B.5}$$

Note that $Y_j|\tau^{(s)}, \Sigma(s), Y_{-j}$ is $N(X_j V_{-j} X'_{-j} P_{-j}^{-1} Y_{-j}, P_j + X_j V_{-j} X'_j)$. Drawing from this density is equivalent to running AXE for every MC sample s and has computational cost $\mathcal{O}(SJ(NP^2 + P^3))$ across all CV folds and MC samples. Obtaining the likelihood in $w_j^{(s)}$ of (C.4) has additional computational cost $\mathcal{O}(SJN)$. The total computational cost is then $\mathcal{O}(SJ(NP^2 + P^3))$.

B.2.5 MCV (Gibbs sampling)

Under the model in (4.1) with prior densities $\Sigma \sim IW(\nu, \Psi)$, $\tau \sim \Gamma^{-1}(a, b)$, $\beta_1 \sim N(0, C)$, $\beta_2|\Sigma \sim N(0, \Sigma)$, one Gibbs sampling scheme is as follows:

$$\beta^{(s)}|\Sigma^{(s-1)}, \tau^{(s-1)}, Y \sim N(\tau^{-2(s-1)} V^{(s-1)} X^T Y, V^{(s-1)}), \tag{B.6}$$

$$V^{(s-1)} = (\Sigma^{-1(s-1)} + \frac{1}{\tau^{2(s-1)}} X^T X)^{-1}$$

$$\Sigma^{(s)}|\beta^{(s)}, \tau^{(s-1)}, Y \sim IW(N + v, \Psi + (\beta_2^{(s)} - \mu)(\beta_2^{(s)} - \mu)^T) \tag{B.7}$$

$$\tau^{2(s)}|\Sigma^{(s)}, \beta^{(s)}, Y \sim \Gamma^{-1}(a + \frac{1}{2}N, b + \frac{1}{2}(Y - X\beta^{(s)})^T (Y - X\beta^{(s)}). \tag{B.8}$$

where s refers to the sth iteration of the Gibbs sampler, IW refers to the inverse-Wishart distribution and Γ^{-1} the inverse-gamma.

We assume that the cost of drawing a sample from the specified densities is dominated by the re-calculation of parameters within each iteration, for example, the cost of drawing $\beta^{(s)}$ is dominated by the calculations of $V^{(s-1)}$ and $V^{(s-1)} X^T Y$.

In eq. (C.6), the inversion of V is $\mathcal{O}(P^3)$, while the multiplication of $VX'Y$ is $\mathcal{O}(NP^2)$. The cost of eq. (C.7) is $\mathcal{O}(P_2^2)$. The cost of eq. (C.8) is $\mathcal{O}(N^2 P)$. For each iteration of the Gibbs sampler, the computational cost is $\mathcal{O}(P^3 + NP^2 + N^2 P)$. Then with M iterations of the Gibbs sampler, the computational complexity is $O(MN^2 P + MNP^2 + MP^3)$.

B.3 Table of LRR mean and standard deviation

Table B.1. Mean and standard deviation (SD) of log RMSE ratio (LRR) for each leave-a-cluster-out CV approximation method and data set. Log RMSE ratios (LRRs), defined in (4.6), are calculated for each CV loop. iIS-A was not applied to the SRD data due to the amount of time it would have taken.

Data	AXE		GHOST		iIS-C		iIS-A	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Eight schools								
$\alpha < 2$	-0.06	0.46	-0.13	0.57	-0.08	0.50	-0.07	0.64
$\alpha \geq 2$	0.02	0.22	-0.12	0.28	0.02	0.22	0.03	0.29
Radon								
Model 1	0.00	0.05	0.00	0.05	0.00	0.05	0.00	0.04
Model 2	-0.01	0.06	-0.02	0.07	-0.01	0.06	-0.01	0.06
Model 3	-0.01	0.03	-0.01	0.03	-0.01	0.03	0.00	0.03
Radon subsets								
Model 1	0.00	0.02	-0.03	0.03	0.00	0.02	0.10	0.10
Model 2	0.01	0.03	-0.08	0.06	0.00	0.02	0.14	0.13
Model 3	0.00	0.05	-0.18	0.20	0.00	0.05	0.23	0.21
ESP	-0.03	0.06	-0.04	0.06	1.05	0.97	0.20	0.33
SLC	0.07	0.71	0.04	0.68	0.76	1.44	0.08	0.67
SRD	-0.04	0.07	0.09	0.25	–	–	0.00	0.08

Appendix C
Proof and calculations for AXE

C.1 Proof for Theorem 4.3.1

Lemma C.1.1. *Let response vector $Y \in \mathbb{R}^N$ of a hierarchical linear regression follow a normal distribution as in (4.1) and let s_j be the set of indices of θ_j in θ such that X_{s_j} is made up of identical rows. As shorthand X_{s_j} is referred to as X_j, and x_j' is a row in X_j. V is defined as in (4.2). We list several facts pertaining to V and V_{-j}.*

1. $V_{-j} = V + \frac{n_j}{\tau^2}\frac{1}{1-\frac{n_j}{\tau^2}x_j'Vx_j}Vx_jx_j'V$
2. $V \preceq V_{-j}$, *where* $V \preceq V_{-j}$ *indicates that* $V_{-j} - V$ *is positive semi-definite*
3. $\frac{n_j}{\tau^2}x_j'Vx_j \leq 1$
4. $x_j'V_{-j}x_j = x_j'Vx_j\frac{1}{1-\frac{n_j}{\tau^2}x_j'Vx_j} \geq x_j'Vx_j$
5. $\det(V_{-j}) = \det(V)\frac{1}{1-\frac{n_j}{\tau^2}x_j'Vx_j} \geq \det(V)$
6. $\frac{1}{\tau^4}Y_{-j}'X_{-j}V_{-j}X_{-j}'Y_{-j} = \frac{1}{\tau^4}Y'XVX'Y - \frac{n_j}{\tau^2}\bar{y}_j^2 + \frac{\bar{e}_j^2}{\tau^2/n_j}\frac{1}{1-\frac{n_j}{\tau^2}x_j'Vx_j}$

Proof. 1. This follows from using the Sherman-Morrison formula on $V_{-j} = (V^{-1} - \frac{n_j}{\tau^2}x_jx_j')^{-1}$.

2. Using a similar process as in 1., if we re-write V in terms of V_{-j}, we obtain $V = V_{-j} - \frac{n_j}{\tau^2}\frac{1}{1+\frac{n_j}{\tau^2}x_j'V_{-j}x_j}V_{-j}x_jx_j'V_{-j}$. The difference, $\frac{n_j}{\tau^2}\frac{1}{1+\frac{n_j}{\tau^2}x_j'V_{-j}x_j}V_{-j}x_jx_j'V_{-j}$, is positive semi-definite and so $V_{-j} \succeq V$, where $A \succeq B$ is defined as $A - B$ is positive semi-definite.

3. From 2., we note that $V_{-j} \succeq V$ and from 1., $V_{-j} - V$ is $\frac{n_j}{\tau^2}\frac{1}{1-\frac{n_j}{\tau^2}x_j'Vx_j}Vx_jx_j'V$. V positive-definite implies $\frac{n_j}{\tau^2}x_j'Vx_j > 0$, then for $V_{-j} - V$ to be positive semi-definite, $\frac{n_j}{\tau^2}x_j'Vx_j \leq 1$ must be true.

4. Using 1., $x_j'V_{-j}x_j = x_j'Vx_j + \frac{n_j}{\tau^2}\frac{1}{1-\frac{n_j}{\tau^2}x_j'Vx_j}(x_j'Vx_j)^2 = x_j'Vx_j(1-\frac{n_j}{\tau^2}x_j'Vx_j)^{-1}$. The inequality follows from 3. , as $1 - \frac{n_j}{\tau^2}x_j'Vx_j \leq 1$.

5. As the held-out data correspond to identical rows of X, we have a closed-form solution for the determinant of V_{-j} in terms of V:

$$\begin{aligned}
\det(V^{-1} - \tau^{-2}X_j'X_j) &= \det(V^{-1})\det(1 - \frac{n_j}{\tau^2}x_j'Vx_j) \qquad \textit{Sylvester's det. theorem} \\
&= \det(V^{-1})(1 - \frac{n_j}{\tau^2}x_j'Vx_j) \\
\implies \det(V_{-j}) &= \det(V)\frac{1}{1-\frac{n_j}{\tau^2}x_j'Vx_j} \\
&\geq \det(V) \qquad \textit{from } 3.
\end{aligned}$$

6. We first show that $\frac{1}{\tau^2}V_{-j}X_{-j}'Y_{-j} = \frac{1}{\tau^2}VX'Y + \frac{n_j}{\tau^2}\frac{\bar{e}_j}{1-\frac{n_j}{\tau^2}\nu_j}Vx_j$. Let $\nu_j \coloneqq x_j'Vx_j$:

$$\begin{aligned}
\frac{1}{\tau^2}V_{-j}X_{-j}'Y_{-j} &= \frac{1}{\tau^2}[V + \frac{n_j}{\tau^2}\frac{1}{1-\frac{n_j}{\tau^2}\nu_j}Vx_jx_j'V](X'Y - X_j'Y_j) \qquad \textit{from } 1. \\
&= \frac{1}{\tau^2}VX'Y + Vx_j\left[-\frac{n_j}{\tau^2}\bar{y}_j + \frac{n_j}{\tau^2}\frac{1}{1-\frac{n_j}{\tau^2}\nu_j}\left(\tilde{y}_j - \frac{n_j}{\tau^2}\nu_j\bar{y}_j\right)\right] \\
&= \frac{1}{\tau^2}VX'Y - Vx_j\left[\frac{n_j}{\tau^2}\frac{1}{1-\frac{n_j}{\tau^2}\nu_j}\bar{y}_j - \frac{n_j}{\tau^2}\frac{1}{1-\frac{n_j}{\tau^2}\nu_j}\tilde{y}_j\right] \\
&= \frac{1}{\tau^2}VX'Y + \frac{n_j}{\tau^2}\frac{\bar{e}_j}{1-\frac{n_j}{\tau^2}x_j'Vx_j}Vx_j.
\end{aligned}$$

Following a similar procedure, we can re-write $\frac{1}{\tau^4}Y_{-j}'X_{-j}V_{-j}X_{-j}'Y_{-j}$ in terms of its full-data counterpart, $\frac{1}{\tau^4}Y'XVX'Y$, along with an additional difference term. Let $(\tilde{Y})_{-j}$ refer to $X_{-j}\tilde{\beta}$, the conditional expected value based on the full data Y.

$$\begin{aligned}
\frac{1}{\tau^4}Y_{-j}'X_{-j}V_{-j}X_{-j}'Y_{-j} &= \frac{1}{\tau^4}Y_{-j}'X_{-j}VX'Y + \frac{n_j}{\tau^4}\frac{\bar{e}_j}{1-\frac{n_j}{\tau^2}\nu_j}Y_{-j}'X_{-j}Vx_j \\
&= \frac{1}{\tau^2}Y_{-j}'(\tilde{Y})_{-j} + \frac{n_j}{\tau^2}\frac{\bar{e}_j}{1-\frac{n_j}{\tau^2}\nu_j}\left(\frac{1}{\tau^2}Y'XVx_j - \frac{1}{\tau^2}Y_j'X_jVx_j\right) \\
&= \frac{1}{\tau^2}Y_{-j}'(\tilde{Y})_{-j} + \frac{n_j}{\tau^2}\frac{\bar{e}_j}{1-\frac{n_j}{\tau^2}\nu_j}\left(\tilde{y}_j - \frac{n_j}{\tau^2}\nu_j\bar{y}_j\right)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\tau^2} Y'_{-j}(\tilde{Y})_{-j} + \bar{e}_j \frac{n_j}{\tau^2} \left[\frac{1}{1 - \frac{n_j}{\tau^2}\nu_j} \left(\tilde{y}_j - \frac{n_j}{\tau^2}\nu_j \bar{y}_j \right) + \bar{y}_j - \bar{y}_j \right] \\
&= \frac{1}{\tau^2} Y'_{-j}(\tilde{Y})_{-j} + \bar{e}_j \frac{n_j}{\tau^2} \left[\bar{y}_j + \frac{\bar{e}_j}{1 - \frac{n_j}{\tau^2}\nu_j} \right] \\
&= \frac{1}{\tau^2} Y'_{-j}(\tilde{Y})_{-j} + \bar{e}_j \frac{n_j}{\tau^2} \left[\bar{y}_j + \frac{\bar{e}_j}{1 - \frac{n_j}{\tau^2}\nu_j} \right] - \frac{n_j}{\tau^2}\bar{y}_j\tilde{y}_j + \frac{n_j}{\tau^2}\bar{y}_j\tilde{y}_j \\
&= \frac{1}{\tau^4} Y'XVX'Y - \frac{n_j}{\tau^2}\bar{y}_j^2 + \frac{\bar{e}_j^2}{\tau^2/n_j} \frac{1}{1 - \frac{n_j}{\tau^2} x'_j V x_j}.
\end{aligned}
$$

□

Lemma C.1.2. *Let response vector $Y \in \mathbb{R}^N$ of a hierarchical linear regression follow a normal distribution as in (4.1) and let s_j be the set of indices of θ_j in θ such that X_{s_j} is made up of identical rows. V is defined as in (4.2). The difference Δ_j between the log densities $\ell(\Sigma, \tau | Y_{-j})$ and $\ell(\Sigma, \tau | Y)$ is*

$$
\Delta_j := C + n_j \log \tau - \frac{1}{2} \log \left(1 - \frac{n_j}{\tau^2} x'_j V x_j \right) - \frac{n_j}{\tau^2}\bar{y}_j^2 + \frac{\bar{e}_j^2}{\tau^2/n_j} \frac{1}{1 - \frac{n_j}{\tau^2} x'_j V x_j},
$$

where $C \in \mathbb{R}$ is a constant that does not involve Σ or τ.

Proof. The log-likelihood is as follows:

$$
\ell(\Sigma, \tau | Y) \propto \ell(\tau) + \ell(\Sigma) - N \log(\tau) - \frac{1}{2} \log \det(\Sigma) + \frac{1}{2} \log \det(V) + \frac{1}{2\tau^4} Y'XVX'Y.
$$

Using Theorem C.1.1, items 5 and 6, we can directly obtain the difference between the log-likelihoods $\ell(\Sigma, \tau | Y_{-j}) - \ell(\Sigma, \tau | Y)$ as stated. □

Lemma C.1.3. *Let response vector $Y \in \mathbb{R}^N$ of a hierarchical linear regression follow a normal distribution as in (4.1) and let V defined as in (4.2). Let $B := [\ I_{P_2}] \in \mathbb{R}^{P_2 \times P}$, where I_{P_2} is the P_2-dimensional identity matrix. The partial derivative $\frac{\partial}{\partial \Sigma} a'Vc$ is $\Sigma^{-1}BVac'VB'\Sigma^{-1}$ for vectors $a \in \mathbb{R}^P$, $c \in \mathbb{R}^P$.*

Proof.

$$
\begin{aligned}
\frac{\partial}{\partial \Sigma_{ij}} V &= -V \left(\frac{\partial}{\partial \Sigma_{ij}} V^{-1} \right) V \\
&= VB'\Sigma^{-1}\delta_i\delta'_j\Sigma^{-1}BV,
\end{aligned}
$$

where $\delta_i \in \mathbb{R}^{P_2}$ is the binary vector with a 1 only at the i^{th} index. Then,

$$\begin{aligned}
\frac{\partial}{\partial \Sigma_{ij}} a'Vc &= \frac{\partial}{\partial \Sigma_{ij}} \operatorname{tr}(a'Vb) \\
&= \operatorname{tr}\left(ba' \frac{\partial}{\partial \Sigma_{ij}} V \right) \\
&= \operatorname{tr}\Big(ba'VB'\Sigma^{-1}\delta_i\delta_j'\Sigma^{-1}BV \Big) \\
&= \operatorname{tr}\Big(\delta_j'\Sigma^{-1}BVca'VB'\Sigma^{-1}\delta_i \Big) \\
&= \delta_j'\Sigma^{-1}BVca'VB'\Sigma^{-1}\delta_i,
\end{aligned}$$

as the trace is invariant under cyclic permutations.

Then

$$\frac{\partial}{\partial \Sigma} a'Vc = \Sigma^{-1}BVac'VB'\Sigma^{-1}.$$

□

We re-state Theorem 4.3.1 below:

Let response vector $Y \in \mathbb{R}^N$ of a hierarchical linear regression follow a normal distribution as in (4.1), where $\mathbb{1} \in \operatorname{span}(X)$, and prior densities $f(\Sigma)$ and $f(\tau)$ are $\mathcal{O}(1)$. $X\beta = \mu + \theta$ as in (4.5), where θ has J unique values and $\theta_1, \ldots, \theta_J$, and s_j denotes the set of indices such that $\theta_{s_j} = \theta_j \mathbb{1}$, and X_j is made up of identical rows. V is defined as in (4.2). Then $|\operatorname{argmax}_{\Sigma,\tau} f(\Sigma, \tau | Y_{-\theta_j}) - \operatorname{argmax}_{\Sigma,\tau} f(\Sigma, \tau | Y)| \to 0$ as $J \to \infty$.

Proof. The log-likelihood is as follows:

$$\ell(\Sigma, \tau | Y) \propto \ell(\tau) + \ell(\Sigma) - N\log(\tau) - \frac{1}{2}\log\det(\Sigma) + \frac{1}{2}\log\det(V) + \frac{1}{2\tau^4} Y'XVX'Y.$$

We note that, other than the prior densities whose forms are unknown, the log-likelihood consists of terms with P_2 or N summands. As the number of CV folds is J and $J \leq P_2$, $J \leq N$, $J^{-1}\ell(\Sigma, \tau | Y)$ is finite and we can say that for any J, the solution to $J^{-1}\frac{\partial}{\partial \Sigma}\ell(\Sigma, \tau | Y) = 0$ is the same as the solution to $\frac{\partial}{\partial \Sigma}\ell(\Sigma, \tau | Y) = 0$:

$$\operatorname*{argmax}_{\Sigma,\tau} f(\Sigma, \tau | Y) = \operatorname*{argmax}_{\Sigma,\tau} \frac{\ell(\Sigma, \tau | Y)}{J}.$$

The dimensions of Σ, V, and Y are dependent on J, so for a particular J, we denote the corresponding Σ as $\Sigma^{(J)}$, and similarly for $V^{(J)}$ and $Y^{(J)}$. Let a_J be the sequence of

differences between $\frac{\partial}{\partial\Sigma}\frac{\ell(\Sigma^{(J)},\tau|Y^{(J)})}{J}$ and $\frac{\partial}{\partial\Sigma}\frac{\ell(\Sigma^{(J)},\tau|Y^{(J)}_{-j})}{J}$:

$$a_J = \left|\frac{\partial}{\partial\Sigma}\frac{\ell(\tau,\Sigma^{(J)}|Y_{-j})^{(J)}}{J} - \frac{\partial}{\partial\Sigma}\frac{\ell(\tau,\Sigma^{(J)}|Y^{(J)})}{J}\right|.$$

If we show $a_J \to 0$ as $J \to \infty$, then that is equivalent to showing $|\operatorname{argmax}_\Sigma f(\Sigma,\tau|Y_{-s_j}) - \operatorname{argmax}_\Sigma f(\Sigma,\tau|Y)| \to 0$ as $J \to \infty$.

Let Δ_j be the difference in log-likelihoods stated in Theorem C.1.2:

$$\begin{aligned} a_J &= \left|\frac{\partial}{\partial\Sigma}\frac{\ell(\tau,\Sigma^{(J)}|Y^{(J)}_{-j})}{J} - \frac{\partial}{\partial\Sigma}\frac{\ell(\tau,\Sigma^{(J)}|Y^{(J)})}{J}\right| \\ &= \left|\frac{\partial}{\partial\Sigma}\left(\frac{\ell(\tau,\Sigma|Y^{(J)})}{J} - \frac{\Delta_j}{J}\right) - \frac{\partial}{\partial\Sigma}\frac{\ell(\tau,\Sigma^{(J)}|Y^{(J)})}{J}\right| \\ &= \left|\frac{1}{J}\frac{\partial}{\partial\Sigma}\Delta_j\right|. \end{aligned}$$

It remains to show that $\left|\frac{\partial}{\partial\Sigma}\Delta_j\right| \in o(J)$. Of the terms in Δ_j, only those involving $V^{(J)}$ are dependent on J. The derivative is as follows:

$$\frac{\partial}{\partial\Sigma}\Delta_J = \frac{n_j/\tau^2}{1-\frac{n_j}{\tau^2}x_j'V^{(J)}x_j}\left(\frac{1}{2}\left(1+\frac{\bar{e}_j^2}{1-\frac{n_j}{\tau^2}x_j'V^{(J)}x_j}\right)\frac{\partial}{\partial\Sigma}x_j'V^{(J)}x_j + \bar{e}_j\frac{\partial}{\partial\Sigma}x_j'V^{(J)}X'Y\right).$$

Let a, c in Lemma 3 equal x_j, then $\frac{\partial}{\partial\Sigma}x_j'Vx_j = \Sigma^{-1(J)}BV^{(J)}x_jx_j'V^{(J)}B'\Sigma^{-1(J)}$. Let $a = x_j$ and $c = X'Y$, then $\frac{\partial}{\partial\Sigma}x_j'VX'Yj = \Sigma^{-1(J)}B\tilde{\beta}^{(J)}x_j'V^{(J)}B'\Sigma^{-1(J)}$.

From Theorem C.1.1, items 2 and 4, we know that $V^{(J)}$ and $\frac{1}{1-\frac{n_j}{\tau^2}x_j'V^{(J)}x_j}$ decrease as J increases. We also note that $\sqrt{n_j}\bar{e}_j = \mathcal{O}_p(\tau)\,\forall J$. As the non-zero values of x_j are not dependent on J, this similarly implies that the magnitude of $\tilde{\beta}$ is not dependent on J.

Then the partial derivatives decrease as J increases. The increase in dimension does not matter, as the solution also has J dimensions. The remaining terms are $\mathcal{O}_p(1)$, as we take τ fixed. Then $\left|\frac{\partial}{\partial\Sigma}\Delta_J\right| \in \mathcal{O}_p(1)$.

A similar argument as above for τ holds if $\left|\frac{\partial}{\partial\tau}V^{(J)}\right| \in \mathcal{O}_p(1)$. Let $M \in I\!R^{P\times P}$ be a binary matrix where $M_{pq} = 1$ if $\tau^{-2}(X'X)_{pq} \neq 0$. Then:

$$\frac{\partial}{\partial\tau}x_j'Vx_j = -\frac{2}{\tau^3}x_j'VMVx_j.$$

It can be seen that this scalar quantity decreases with J, using Theorem C.1.1, item 1.

□

We re-state Corollary 4.3.2 below:

Let $\hat{\Sigma}$ denote the full-data posterior mean, $E[\Sigma|Y]$, and $\tilde{\Sigma}$ the CV posterior mean over the training data $E[\Sigma|Y_{-j}]$ for CV fold j. Likewise let $\hat{\tau} = E[\tau|Y]$ and $\tilde{\tau} \coloneqq E[\tau|Y_{-j}]$. Under the same conditions as 4.3.1, $E[X\beta|Y_{-j}, \hat{\Sigma}, \hat{\tau}] = E[X\beta|Y_{-j}, \tilde{\Sigma}, \tilde{\tau}](1 + \mathcal{O}(J^{-1}))$.

Proof. Using the same methods as determining $\frac{\partial}{\partial\Sigma}\Delta_J$, it can be shown that $\frac{\partial}{\partial\Sigma}\Delta_J$ is differentiable and, as $XVX \in \mathcal{O}(1)$, the derivative is finite, thus $\frac{\partial}{\partial\Sigma}\Delta_J$ is uniformly continuous. Proof by contradiction shows that $f(\Sigma) = \frac{\partial}{\partial\Sigma}\Delta_J$ is a bijective function. Then as $\frac{1}{J}\frac{\partial}{\partial\Sigma}\Delta_J \to 0$, $E[\Sigma, \tau|Y_{-j}] \to E[\Sigma, \tau|Y]$ and $E[\Sigma, \tau|Y_{-j}] = E[\Sigma, \tau|Y](1 + \mathcal{O}(J^{-1}))$.

Substituting $\hat{\Sigma} + \mathcal{O}(J^{-1})$ and $\hat{\tau} + \mathcal{O}(J^{-1})$ into $\tau^{-2}XVX'Y$ and an application of Sherman-Morrison yields $\tau^{-2}XVX'Y + \mathcal{O}(J^{-1})XVX'XVX'Y(\tau^{-2} + \mathcal{O}(J^{-1})) + \mathcal{O}(J^{-1})XVB\Sigma^{-1}B'VX'Y$, where $B \in \mathbb{R}^{P\times P_2}$ is a block matrix of 0s in the first P_1 rows and the identity matrix in the remaining P_2. As $XVX', XVX'Y, \Sigma \in \mathcal{O}(1)$, the result follows. □

C.2 Computational complexity calculations

This section contains derivations for computational complexity in Table 4.1. We assume without loss of generality that $\mathcal{O}(P_2) = \mathcal{O}(P)$. For all methods, we assume that the cost of drawing a sample from a specific density is dominated by the calculation of the density's parameters, e.g. if drawing from a multivariate normal density, we assume the calculation of the mean vector and covariance matrix dominates the computational cost.

C.2.1 AXE

We re-state the AXE estimate below:

$$\hat{Y}_j^{AXE} = \frac{1}{\hat{\tau}^2}X_j\left(\frac{1}{\hat{\tau}^2}X'_{-j}X_{-j} + \begin{bmatrix}0 & 0\\ 0 & \Sigma^{-1}\end{bmatrix}\right)^{-1}X'_{-j}Y_{-j}. \tag{C.1}$$

The cost of the matrix inversion in (C.1) is $\mathcal{O}(P^3)$, while the matrix multiplication is $\mathcal{O}(NP^2)$. Conducted over J total cross-validation loops, the computational complexity for AXE is $\mathcal{O}(J(NP^2 + P^3))$.

C.2.2 GHOST

Without loss of generality, let $f(\theta|\Sigma)$ be $N(0, \Sigma)$. Then:

$$\theta_j|\theta_{-j}, \Sigma \sim N(\Sigma_{j-j}\Sigma_{-j-j}^{-1}\theta_{-j}, \quad \Sigma_{jj} - \Sigma_{j-j}\Sigma_{-j-j}^{-1}\Sigma_{-jj}). \tag{C.2}$$

As we assume the dimension of θ_j is fixed, the cost of the matrix inversion in (C.2) is $\mathcal{O}(P^3)$ and the cost of the matrix multiplication is likewise $\mathcal{O}(P^3)$. Conducted over J total cross-validation loops for S samples, the computational complexity for GHOST is $\mathcal{O}(SJP^3)$.

C.2.3 iIS-C

We re-state the iIS-C importance weights $w_j^{(s)}$ and mean estimate for $E[Y_j|Y_{-j}]$:

$$w_j^{(s)} = \frac{1}{f(Y_j|\theta_{-j}^{(s)}, \tau^{(s)}, \mu^{(s)}, \Sigma^{(s)})}, \quad \hat{Y}_j^{\text{iIS-C}} = \frac{\sum_{s=1}^S w_j^{(s)} E\left[Y_j|\theta_{-j}, \tau^{(s)}, \Sigma^{(s)}, \mu^{(s)}\right]}{\sum_{s=1}^S w_j^{(s)}}. \tag{C.3}$$

Let $a_j \coloneqq E[\theta_j|\theta_{-j}, \Sigma]$ and $M \coloneqq \text{Cov}(\theta_j|\theta_{-j}, \Sigma)$, corresponding to the mean and covariance, respectively, of (C.2). Then,

$$Y_j|\theta_{-j}, \tau, \mu, \Sigma \sim N(\mu + a_j, \tau^2 I + M).$$

Note that the cost of obtaining a_j and M for all CV folds and MC samples is the same as ghosting at $\mathcal{O}(SJP^3)$. The cost of obtaining the likelihood in weight $w_j^{(s)}$ for (C.3) is an additional $\mathcal{O}\left(S\sum_{j=1}^J n_j\right) = \mathcal{O}(SN)$ and the cost of the expectation is likewise $\mathcal{O}(SN)$. The total cost is then $\mathcal{O}(SJP^3)$, the same order as ghosting.

C.2.4 iIS-A

We re-state the iIS-A importance weights $w_j^{(s)}$ and mean estimate for $E[Y_j|Y_{-j}]$:

$$\text{iIS-A: } w_j^{(s)} = \frac{1}{f(Y_j|\tau^{(s)}, \Sigma^{(s)}, Y_{-j})}, \quad \hat{Y}_j^{iIS-A} = \frac{\sum_{s=1}^S w_j^{(s)} E\left[Y_j|\tau^{(s)}, \Sigma^{(s)}, Y_{-j}\right]}{\sum_{s=1}^S w_j^{(s)}}. \tag{C.4}$$

$$Y_j|\tau^{(s)}, \Sigma(s), Y_{-j} \tag{C.5}$$

Note that $Y_j|\tau^{(s)}, \Sigma(s), Y_{-j}$ is $N(X_j V_{-j} X'_{-j} P_{-j}^{-1} Y_{-j}, P_j + X_j V_{-j} X'_j)$. Drawing from this density is equivalent to running AXE for every MC sample s and has computational cost $\mathcal{O}(SJ(NP^2 + P^3))$ across all CV folds and MC samples. Obtaining the likelihood in $w_j^{(s)}$ of (C.4) has additional computational cost $\mathcal{O}(SJN)$. The total computational cost is then $\mathcal{O}(SJ(NP^2 + P^3))$.

C.2.5 MCV (Gibbs sampling)

Under the model in (4.1) with prior densities $\Sigma \sim IW(\nu, \Psi)$, $\tau \sim \Gamma^{-1}(a, b)$, $\beta_1 \sim N(0, C)$, $\beta_2|\Sigma \sim N(0, \Sigma)$, one Gibbs sampling scheme is as follows:

$$\beta^{(s)}|\Sigma^{(s-1)}, \tau^{(s-1)}, Y \sim N(\tau^{-2(s-1)} V^{(s-1)} X^T Y, V^{(s-1)}), \tag{C.6}$$

$$V^{(s-1)} = (\Sigma^{-1(s-1)} + \frac{1}{\tau^{2(s-1)}} X^T X)^{-1}$$

$$\Sigma^{(s)}|\beta^{(s)}, \tau^{(s-1)}, Y \sim IW(N + v, \Psi + (\beta_2^{(s)} - \mu)(\beta_2^{(s)} - \mu)^T) \tag{C.7}$$

$$\tau^{2(s)}|\Sigma^{(s)}, \beta^{(s)}, Y \sim \Gamma^{-1}(a + \frac{1}{2}N, b + \frac{1}{2}(Y - X\beta^{(s)})^T (Y - X\beta^{(s)}). \tag{C.8}$$

where s refers to the sth iteration of the Gibbs sampler, IW refers to the inverse-Wishart distribution and Γ^{-1} the inverse-gamma.

We assume that the cost of drawing a sample from the specified densities is dominated by the re-calculation of parameters within each iteration, for example, the cost of drawing $\beta^{(s)}$ is dominated by the calculations of $V^{(s-1)}$ and $V^{(s-1)} X^T Y$.

In eq. (C.6), the inversion of V is $\mathcal{O}(P^3)$, while the multiplication of $VX'Y$ is $\mathcal{O}(NP^2)$. The cost of eq. (C.7) is $\mathcal{O}(P_2^2)$. The cost of eq. (C.8) is $\mathcal{O}(N^2 P)$. For each iteration of the Gibbs sampler, the computational cost is $\mathcal{O}(P^3 + NP^2 + N^2 P)$. Then with M iterations of the Gibbs sampler, the computational complexity is $O(MN^2 P + MNP^2 + MP^3)$.

C.3 Table of LRR mean and standard deviation

Table C.1. Mean and standard deviation (SD) of log RMSE ratio (LRR) for each leave-a-cluster-out CV approximation method and data set. Log RMSE ratios (LRRs), defined in (4.6), are calculated for each CV loop. iIS-A was not applied to the SRD data due to the amount of time it would have taken.

Data	AXE		GHOST		iIS-C		iIS-A	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Eight schools								
$\alpha < 2$	-0.06	0.46	-0.13	0.57	-0.08	0.50	-0.07	0.64
$\alpha \geq 2$	0.02	0.22	-0.12	0.28	0.02	0.22	0.03	0.29
Radon								
Model 1	0.00	0.05	0.00	0.05	0.00	0.05	0.00	0.04
Model 2	-0.01	0.06	-0.02	0.07	-0.01	0.06	-0.01	0.06
Model 3	-0.01	0.03	-0.01	0.03	-0.01	0.03	0.00	0.03
Radon subsets								
Model 1	0.00	0.02	-0.03	0.03	0.00	0.02	0.10	0.10
Model 2	0.01	0.03	-0.08	0.06	0.00	0.02	0.14	0.13
Model 3	0.00	0.05	-0.18	0.20	0.00	0.05	0.23	0.21
ESP	-0.03	0.06	-0.04	0.06	1.05	0.97	0.20	0.33
SLC	0.07	0.71	0.04	0.68	0.76	1.44	0.08	0.67
SRD	-0.04	0.07	0.09	0.25	–	–	0.00	0.08

Bibliography

Sylvain Arlot, Alain Celisse, et al. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79, 2010.

Le Bao, Ben Sheng, Xiaoyue Niu, Yuan Tang, Tim Brown, Peter D Ghys, and Jeff W Eaton. Incorporating hierarchical structure into dynamic systems: An application of estimating hiv epidemics at sub-national and sub-population level. *arXiv preprint arXiv:1602.05665*, 2016.

Rémi Bardenet, Arnaud Doucet, and Chris Holmes. On Markov Chain Monte Carlo methods for tall data. *The Journal of Machine Learning Research*, 18(1):1515–1557, 2017.

John Barnard, Robert McCulloch, and Xiao-Li Meng. Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, pages 1281–1311, 2000.

David A Belsley, Edwin Kuh, and Roy E Welsch. *Regression diagnostics: Identifying influential data and sources of collinearity*, volume 571. John Wiley & Sons, 2005.

JM Bernardo, MJ Bayarri, JO Berger, AP Dawid, D Heckerman, AFM Smith, and M West. Generative or discriminative? getting the best of both worlds. *Bayesian statistics*, 8(3):3–24, 2007.

Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 2017.

Samprit Chatterjee and Ali S Hadi. *Sensitivity analysis in linear regression*, volume 327. John Wiley & Sons, 2009.

R Dennis Cook. Detection of influential observation in linear regression. *Technometrics*, 19(1):15–18, 1977.

Michael J Daniels and Robert E Kass. A note on first-stage approximation in two-stage hierarchical models. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 19–30, 1998.

Christopher Eager and Joseph Roy. Mixed effects models are sometimes terrible. *arXiv preprint arXiv:1701.04858*, 2017.

Bradley Efron and Carl Morris. Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70(350):311–319, 1975.

Jonah Gabry and Ben Goodrich. rstanarm: Bayesian applied regression modeling via Stan. *R package version*, 2(1), 2016.

Alan E Gelfand. Model determination using sampling-based methods. *Markov Chain Monte Carlo in Practice*, pages 145–161, 1996.

Alan E Gelfand, Dipak K Dey, and Hong Chang. Model determination using predictive distributions with implementation via sampling-based methods. Technical report, Stanford Univ. CA Dept. of Statistics, 1992.

Andrew Gelman. Prior distributions for variance parameters in hierarchical models. Technical report, EERI Research Paper Series, 2004.

Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*, volume 1. Cambridge University Press New York, NY, USA, 2007.

Andrew Gelman and Iain Pardoe. Bayesian measures of explained variance and pooling in multilevel (hierarchical) models. *Technometrics*, 48(2):241–251, 2006.

Andrew Gelman, Jessica Hwang, and Aki Vehtari. Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6):997–1016, 2014.

Andrew Gelman, Ben Goodrich, Jonah Gabry, and Aki Vehtari. R-squared for bayesian regression models. *The American Statistician*, 2019.

Andrew Gelman et al. Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian Analysis*, 1(3):515–534, 2006.

Mark Girolami and Ben Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

William James and Charles Stein. Estimation with quadratic loss. In *Breakthroughs in Statistics*, pages 443–460. Springer, 1992.

Robert E Kass and Duane Steffey. Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *Journal of the American Statistical Association*, 84(407):717–726, 1989.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Anoop Korattikara, Yutian Chen, and Max Welling. Austerity in MCMC land: Cutting the Metropolis-Hastings budget. In *International Conference on Machine Learning*, pages 181–189, 2014.

Duncan Lee, Alastair Rushworth, and Gary Napier. Spatio-temporal areal unit modelling in r with conditional autoregressive priors using the carbayesst package. *Journal of Statistical Software*, 84(9), 2018.

Steven M Lewis and Adrian E Raftery. Estimating Bayes factors via posterior simulation with the Laplace—Metropolis estimator. *Journal of the American Statistical Association*, 92(438):648–655, 1997.

Longhai Li, Shi Qiu, Bei Zhang, and Cindy X Feng. Approximating cross-validatory predictive evaluation in Bayesian latent variable models with integrated IS and WAIC. *Statistics and Computing*, 26(4):881–897, 2016.

Finn Lindgren and Håvard Rue. Bayesian spatial modelling with r-inla. *Journal of Statistical Software*, 63(19), 2015.

EC Marshall and DJ Spiegelhalter. Approximate cross-validatory predictive checks in disease mapping models. *Statistics in Medicine*, 22(10):1649–1660, 2003.

C Elizabeth McCarron, Eleanor M Pullenayegum, Lehana Thabane, Ron Goeree, and Jean-Eric Tarride. Bayesian hierarchical models combining different study types and adjusting for covariate imbalances: a simulation study to assess model performance. *PLoS One*, 6(10):e25635, 2011.

Edgar C Merkle, Daniel Furr, and Sophia Rabe-Hesketh. Bayesian comparison of latent variable models: Conditional versus marginal likelihoods. *Psychometrika*, 84(3): 802–829, 2019.

Thomas P Minka. Expectation propagation for approximate Bayesian inference. *arXiv preprint arXiv:1301.2294*, 2013.

Carl N Morris. Parametric empirical Bayes inference: Theory and applications. *Journal of the American Statistical Association*, 78(381):47–55, 1983.

Xiaoyue Niu, Amy Zhang, Tim Brown, Robert Puckett, Mary Mahy, and Le Bao. Incorporation of hierarchical structure into estimation and projection package fitting with examples of estimating subnational hiv/aids dynamics. *AIDS*, 31:S51–S59, 2017.

Jean Opsomer, Yuedong Wang, and Yuhong Yang. Nonparametric regression with correlated errors. *Statistical Science*, pages 134–153, 2001.

Art B Owen. *Monte Carlo Theory, Methods and Examples.* 2013.

Jacob Parsons and Le Bao. The value of information in retrospect. *arXiv preprint arXiv:1806.01458*, 2018.

Daniel Peña. A new statistic for influence in linear regression. *Technometrics*, 47(1): 1–12, 2005.

Nicholas G Polson, James G Scott, et al. On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis*, 7(4):887–902, 2012.

Matias Quiroz, Robert Kohn, Mattias Villani, and Minh-Ngoc Tran. Speeding up MCMC by efficient data subsampling. *Journal of the American Statistical Association*, 114 (526):831–843, 2019.

Ribana Roscher, Bastian Bohn, Marco F Duarte, and Jochen Garcke. Explainable machine learning for scientific insights and discoveries. *IEEE Access*, 8:42200–42216, 2020.

Donald B Rubin. Estimation in parallel randomized experiments. *Journal of Educational Statistics*, 6(4):377–401, 1981.

Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392, 2009.

Alastair Rushworth, Duncan Lee, and Richard Mitchell. A spatio-temporal model for estimating the long-term effects of air pollution on respiratory hospital admissions in greater london. *Spatial and Spatio-Temporal Epidemiology*, 10:29–38, 2014.

Charles Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. Technical report, Stanford University Stanford United States, 1956.

Fadi Thabtah, Suhel Hammoud, Firuz Kamalov, and Amanda Gonsalves. Data imbalance in classification: Experimental evaluation. *Information Sciences*, 513:429–441, 2020.

Zachary M Thomas, Steven N MacEachern, and Mario Peruggia. Reconciling curvature and importance sampling based procedures for summarizing case influence in Bayesian models. *Journal of the American Statistical Association*, 113(524):1669–1683, 2018.

Michael Tsang, Dehua Cheng, and Yan Liu. Detecting statistical interactions from neural network weights. *arXiv preprint arXiv:1705.04977*, 2017.

Jarno Vanhatalo, Jaakko Riihimäki, Jouni Hartikainen, Pasi Jylänki, Ville Tolvanen, and Aki Vehtari. Gpstuff: Bayesian modeling with Gaussian processes. *Journal of Machine Learning Research*, 14(Apr):1175–1179, 2013.

Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 5(27): 1413–1432, 2016a.

Aki Vehtari, Tommi Mononen, Ville Tolvanen, Tuomas Sivula, and Ole Winther. Bayesian leave-one-out cross-validation approximations for Gaussian latent variable models. *The Journal of Machine Learning Research*, 17(1):3581–3618, 2016b.

Yujun Yan, Jiong Zhu, Marlena Duda, Eric Solarz, Chandra Sripada, and Danai Koutra. Groupinn: Grouping-based interpretable neural network for classification of limited, noisy brain data. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 772–782, 2019.

Amy X Zhang, Le Bao, and Michael J Daniels. Approximate cross-validated mean estimates for bayesian hierarchical regression models. *arXiv preprint arXiv:2011.14238*, 2020.

Vita

Amy Xiang Zhang

Doctor of Philosophy (Dec 2021) in Statistics and Social Data Analytics. The Pennnsylvania State University, University Park, PA, USA.

Bachelor of Science (Dec 2012) in Civil and Environmental Engineering. Carnegie Mellon University, Pittsburgh, PA, USA.

Academic Awards

Big Data Social Science - Integrated Graduate Education and Research Training (BDSS-IGERT) Fellow (August 2016)

University Graduate Fellow, Department of Statistics, The Pennsylvania State University (August 2015)

Publications

Amy X. Zhang, Le Bao, and Michael J. Daniels (2021). Explaining regression model estimates through explicit quantification of shrinkage and information borrowing.

Amy X. Zhang, Le Bao, and Michael J. Daniels (2020). Approximate cross-validated mean estimates for Bayesian hierarchical regression models. arXiv preprint arXiv:2011.14238.

Xiaoxiao Li, Amy X. Zhang, Rabah Al-Zaidy, Stefan Baral, Le Bao, and C. Lee Giles (2020). Automating document classification with distant supervision to increase the efficiency of systematic reviews. arXiv preprint arXiv:2012.07565.

Diane Felmlee, Paulina I. Rodis, and Amy X. Zhang (2020). Sexist slurs: Reinforcing feminine stereotypes online. Sex Roles, 83(1), 16-28.

Xiaoyue Niu, Amy X. Zhang, Tim Brown, Robert Puckett, Mary Mahy, and Le Bao (2017). Incorporation of hierarchical structure into estimation and projection package fitting with examples of estimating subnational HIV/AIDS dynamics. Aids, 31(1), S51-S59.

Employment

Graduate Data Science Summer Program Fellow, National Cancer Institute, Bethesda, MD, USA. June 2019 - August 2019.

Intern, UNAIDS, Genève, Switzerland. June 2018 - August 2018.

Technical Writer, MicroStrategy, Vienna, VA. June 2013 - August 2014.